Lyapunov-Inspired LyRIC Activation and GLARE Attention in Chaos-Guided State Space Modeling for EMG-To-Speech Synthesis

Anonymous ACL submission

Abstract

Electromyography-to-Speech (ETS) synthesis attempts to model a non-linear, chaotic dynamical system. Yet, prior works strictly rely on standard reconstruction metrics and parameter-heavy transformers that systematically over-smooth natural acoustic dynamics. To close this gap, we introduce CGS-ETS, a highly compressed, chaos-inspired model featuring the novel encoder – GLAME, which synergizes global Mamba state-space modeling with localized GLARE attention. To explicitly capture the deterministic chaos of human phonation, we propose a physically grounded Lyapunov-derived activation function (LyRIC) alongside two novel loss functions: Lyapunov Exponent Regularization (LER) and Multi-Scale Detrended Fluctuation Analysis (MSDFA). Furthermore, we introduce Post-Vocoder Alignment (PVA), which unlockes comprehensive frame-level acoustic evaluation (LSD, STOI, PESQ) for the first time in ETS literature. CGS-ETS drastically outperforms established baselines, delivering a 4.69× increase in objective intelligibility (STOI: 0.61 vs. 0.13) and a 2.08× improvement in spectral reconstruction (LSD: 1.08 vs. 2.25). Importantly, this massive acoustic superiority is achieved while eliminating 73.49% of total parameters (14.34M vs. 54.10M), establishing a new baseline in the ETS synthesis task. To the best of our knowledge, this is the first work proving that integrating non-linear chaotic physics into network topologies yields superior, compact inductive biases for real-time speech synthesis.

1 Introduction and Related Work

Silent speech interfaces (SSIs) enable speech synthesis without vocalization by decoding articulatory biosignals (Satterlee et al., 2025), thereby supporting individuals with speech impairments and facilitating communication in acoustically challenging environments (Dong et al., 2023). Electromyography (EMG) serves as a highly effective modality for capturing the articulatory muscle activity required for SSI-based synthesis (Satterlee et al., 2025; Kurotaki et al., 2025).

What current ETS systems miss: Despite steady improvements in transcription accuracy, prior ETS architectures (Gaddy and Klein, 2020, 2021; Scheck et al., 2023; Ren et al., 2024; Xie et al., 2025) suffer from three fundamental limitations. ***First***, they approach speech synthesis as a heavily smoothed regression task, optimizing for frame-level reconstruction losses (L1, MSE) while entirely neglecting the *non-linear chaotic dynamics* inherent to human phonation. These non-chaotic objectives inevitably force the network to degenerate into producing artificially muffled, robotic prosody by mathematically penalizing natural aperiodic bursts, jitter, and turbulent aerodynamic variations. ***Second***, state-of-the-art (SOTA) models (Gaddy and Klein, 2020, 2021) rely on brute-force, parameter-heavy Transformer encoders (exceeding 44M parameters). This massive architectural bloat fails to efficiently exploit the highly localized, temporal shift-invariance of articulatory EMG signals, violating the low-latency, real-time constraints required for practical silent speech interfaces. ***Third***, owing to severe vocoder-induced temporal misalignments, prior evaluations are overwhelmingly restricted to Word Error Rate (WER).As frame-count discrepancies preclude the temporal synchronization required for standard acoustic metrics, the true perceptual fidelity, spectral distortion (LSD), and objective intelligibility (STOI) of previous systems have remained uncharacterized in the literature.

To address these gaps, we propose ***CGS-ETS (Chaos-Guided State Space Modeling for EMG-To-Speech)****, which is the first to unify non-linear chaos theory with efficient State Space Model (SSM), achieving unprecedented acoustic fidelity with a drastically reduced parameter footprint.* Our core contributions are manifold: **(1)** We invent and integrate ***Grouped Local Attention with Relative Positional Embeddings (GLARE)*** to the ETS domain. GLARE captures temporal shift-invariance while accelerating inference for long-context continuous EMG sequences by replacing standard Multi-Head Attention (MHA) with Grouped Query Attention (GQA) (Ainslie et al., 2023) and integrating relative positional encodings (Gaddy and Klein, 2021). **(2)** We propose

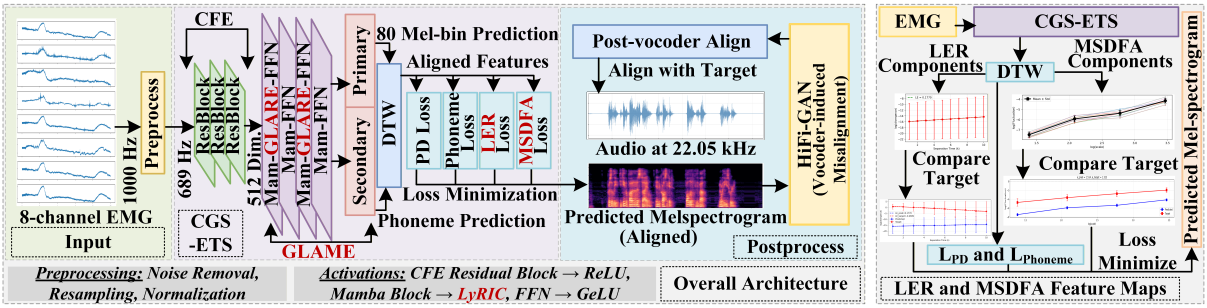

Figure 1: Detailed Architecture of CGS-ETS with LyRIC Activation and GLARE Attention

the ***GLARE-fused Mamba Encoder (GLAME)***, a parameter-efficient SSM architecture, where standard SiLU activations are replaced with our newly designed ***Lyapunov Regularized Implicit Curve*** (LyRIC) activation function to inject chaotic awareness into the state-space trajectory. **(3)** We introduce two novel optimization objectives to constrain the network to the deterministic chaotic invariants of speech: ***LER*** **(Lyapunov Exponent Regularization)** loss which penalizes mismatches in local divergence rates by aligning Lyapunov Exponents (LE) across varying temporal resolutions; and ***MSDFA*** **(Multi-Scale Detrended Fluctuation Analysis)**, which quantifies the fractal-like, long-range temporal chaotic correlations strictly overlooked by standard spectrogram losses. **(4)** We present PVA to resolve vocoder-induced temporal misalignments. *Under **LER** and **MSDFA** supervision, GLAME achieves* $4.69\times$ *higher intelligibility,* $2.08\times$ *better spectral reconstruction, and 73.49% fewer parameters than the SOTA models (Gaddy and Klein, 2021).*

2 Overall Architecture

We show CGS-ETS in Fig. 1. It takes 8-channel EMG data as input, preprocesses it, and generates melspectrogram (mel) bins that are converted to audio at the last stage by a HiFi-GAN vocoder (Kong et al., 2020). It has the following *four innovations for the first time, to the best of our knowledge*. We introduce: **(i)** ***GLARE***, an efficient alternative to standard Multi-Head Attention (MHA) for ETS synthesis. **(ii)** ***LyRIC***, a novel chaos-derived activation function theoretically grounded in local divergence rates. *LyRIC represents the first activation function in the literature derived from chaos theory.* **(iii)** Two novel chaos-guided loss components– ***LER*** and ***MSDFA***, to capture the chaotic temporal dynamics inherent in ETS synthesis. **(iv)** ***GLAME***, a *very lightweight* encoder compared to transformer-based design in (Gaddy and Klein, 2021) for a 77.63% reduced complexity.

2.1 Design of the GLARE

Please note that the Grouped Query Attention (GQA) (Ainslie et al., 2023) is *a generalization of Multi-Query Attention (MQA)* (Shazeer, 2019), and Sliding Window Attention (SWA) captures fine-grained *local context* by sliding over the input sequence (Beltagy et al., 2020) through a sliding window mask (SWM). We adopt GQA within the SWA module rather than employing standard MHA to reduce memory bandwidth overhead during inference *(see Table 10 for ablation on different attention mechanisms)*. Furthermore, we deliberately integrate Relative Positional Embeddings (RPE) from (Gaddy and Klein, 2021) to ensure the attention remains highly robust to natural temporal variations and the precise local alignments required for high-fidelity phonetic decoding. Thus, we propose GLARE, a unified attention mechanism that integrates SWM and RPE with GQA, as illustrated in Fig. 2. *Like optical glare, which achieves maximum perceptual signal through a directed, concentrated beam rather than diffuse ambient light, our GLARE concentrates attention within a sliding local window, groups query computations for efficiency, and encodes positional relationships through relative embeddings — rather than attending globally across all token positions, yielding competitive performance at substantially reduced computational cost compared to full MHA.*

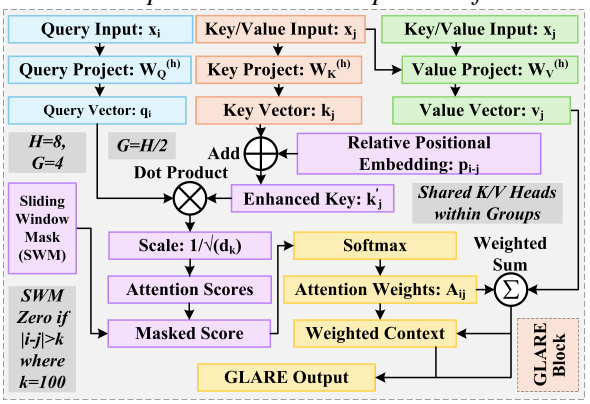

Figure 2: Architecture of GLARE attention.

Governing Equation Design: GLARE is depicted in Fig. 2. Incorporating RPE, we compute the attention weight $A_{ij}^{(h)}$ for head $h \in \{0, \ldots, H-1\}$ between query position i and key position j as:

$$A_{ij}^{(h)} = \frac{\left(W_K^{(g(h))}\mathbf{x}_j + \mathbf{p}_{i-j}\right)^\top \left(W_Q^{(h)}\mathbf{x}_i\right)}{\sqrt{d_k}} \quad (1)$$

where $g(h) = \lfloor \frac{h \cdot G}{H} \rfloor$ and $\mathbf{x}_i \in \mathbb{R}^{d_{model}}$ is the input feature vector at position i, $W_Q^{(h)} \in \mathbb{R}^{d_k \times d_{model}}$ is

the query projection matrix for head h, $W_K^{(g(h))}$ and $W_V^{(g(h))}$ are the shared key and value projection matrices for the key-value group assigned to head h, $\mathbf{p}_{i-j} \in \mathbb{R}^{d_k}$ is a learned relative position embedding indexed by the signed distance $i-j$, and $g(h)$ maps each of the H query heads to one of G shared key-value groups. For $|i-j| > k$, $A_{ij}^{(h)}$ is set to zero, effectively restricting attention to a local window of k steps in each direction. We keep $k = 100$ and $H = 8$. We partition the H query heads into G groups, where each group shares a single key and value head. GQA subsumes both MHA ($G = H$) and MQA ($G = 1$) (Ainslie et al., 2023), and we set $G = \lfloor H/2 \rfloor$ in our implementation, halving the number of key-value heads relative to query heads, which reduces the KV cache size and alleviates the memory bandwidth bottleneck, particularly critical for the long-context EMG sequences in our setting.

2.2 Design of the LyRIC Activation

We mathematically derive the LyRIC activation curve through the expansion of the logarithmic divergence term proposed in the Lyapunov theory (Oseledec, 1968; Wolf et al., 1985).

Logarithmic Divergence Implementation (see Alg.1): The largest Lyapunov Exponent (LE) is computed on non-overlapping windows with time-delay embedding dimension m, and time delay τ. The inputs are first embedded into a higher-dimensional phase space using time-delay embedding, $E_V[i] = [x[i,:], x[i+\tau,:], \ldots, x[i+(m-1)\tau,:]] \in \mathbb{R}^d$ where $x \in \mathbb{R}^{T \times C}$ is the input feature sequence with T time frames and C feature channels, and $d = m \cdot C$ is the embedding dimension (**Rows** ①–④). Then, the Pairwise Distance (PD) matrix is computed, $D[i,j] = \left\| E_V[i] - E_V[j] \right\|_2$ for $i, j \in \{0, 1, \ldots, M-1\}$. Here, the maximum number of embedding vectors to use is $M = T - (m-1)\tau$ (**Rows** ⑤–⑦). Then, the nearest neighbor for each point is calculated, $j^*(i) = \arg\min_{j \in \{1,\ldots,M\}} D[i,j]$ while excluding temporally correlated neighbors using Theiler window (Theiler, 1986) and track the distance evolution for each time offset $k \in \{1, 2, \ldots, k_{max}\}$, $d_k[i] = \left\| E_V[i+k] - E_V[j^*(i)+k] \right\|_2$ (**Rows** ⑨–⑫). Then, ***the logarithmic divergence*** of nearest neighbor pairs over time for each k is estimated as, $S(k) = \frac{1}{N_k} \sum_{i=1}^{M} \log(d_k[i] + \epsilon_1)$, where ϵ_1 prevents logarithm of zero (**Row** ⑭).

Activation Function Design: We approximate $S(k)$ with a singular logarithmic divergence term to seamlessly integrate this dynamical systems metric as an element-wise neural activation. We generalize the local state-space divergence, $d_k[i]$, to the absolute magnitude of the input activation $|x|$, which ensures a valid logarithmic domain ($|x| \geq 0$). Then, we introduce a bounded stability margin of $\epsilon_1 = 1$ to strictly prevent domain violations within the logarithm and avoid singularity collapse. Thus, ϵ_1 offset guarantees zero-centered topological stability (i.e., $f(0) = 0$), making it suitable for stable gradient propagation in deep architectures. This mathematically motivated derivation yields the continuous, non-linear mapping $h(x) = \log(|x| + 1)$. We introduce a learnable scalar parameter α to instantiate this divergence metric as a highly adaptive neural activation that exponentiates the absolute magnitude to dynamically control the local state-space expansion rate. Then, we modulate the logarithmic term with the raw input x to ensure odd symmetry (directional phase preservation), strictly bound the activation magnitude, and apply a denominator-based stability normalization. This theoretically motivated synthesis yields an intermediate function, defined as:

$$g(x) = x \left[\frac{\log(|x|^\alpha + 1)}{|x|^\alpha + 1} \right] \tag{2}$$

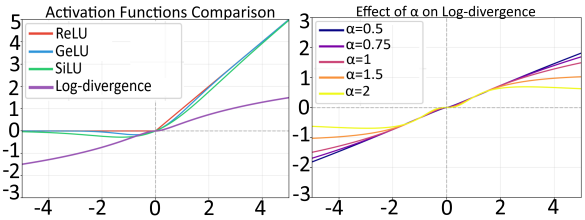

Figure 3: (Left) Log-divergence (Eqn. 2) against standard activation functions. (Right) Effect of α on log-divergence.

This formulation is mathematically guaranteed to preserve critical topological and optimization properties *(see Appendix A.3.1)*, as illustrated by the plots in Fig. 3. However, the direct computation of Eqn. 2 introduces non-trivial inference latency due to an increase in computational overhead. In particular, the nested logarithmic and fractional evaluations demand considerably higher Floating Point Operations (FLOPs) compared to standard, hardware-optimized activations such as ReLU, GeLU, or SiLU *(see computational profiling in Appendix A.3.2, Table 11)*. Therefore, we construct a highly efficient polynomial approximation through *Maclaurin series and geometric series* expansions to strictly mitigate this computational bottleneck while preserving the chaotic dynamic properties. The constituent logarithmic and rational terms from Eqn. 2 decompose as follows:

$$\log(|x|^\alpha + 1) = |x|^\alpha - \frac{|x|^{2\alpha}}{2} + \frac{|x|^{3\alpha}}{3} - \ldots \tag{3}$$

$$\frac{1}{|x|^\alpha + 1} = 1 - |x|^\alpha + |x|^{2\alpha} - |x|^{3\alpha} + \ldots \tag{4}$$

We derive the mathematically equivalent, computationally tractable polynomial representation by computing the Cauchy product of Eqns. 3 and 4, and subsequently multiplying the series by x. Then

the term $g(x)$ in Eqn. 2 becomes:

$$g(x) = x|x|^{\alpha} - \frac{3}{2}x|x|^{2\alpha} + \frac{11}{6}x|x|^{3\alpha} - \dots \quad (5)$$

We truncate the infinite series and isolate the leading-order expansion term, $x|x|^{\alpha}$, as our definitive activation formulation to minimize computational overhead while preserving the essential nonlinear dynamics of Eqn. 2. This streamlined approximation robustly retains the foundational topological properties of the original Lyapunov derived function (Eqn. 2), which are odd symmetry, origin linearity, and adaptive phase-space expansion *(please compare Fig. 3 and 5)* without the prohibitive FLOPs burden of nested logarithmic evaluations. However, arbitrarily parameterizing the expansion coefficient α in $x|x|^{\alpha}$ can induce discontinuities *(i.e.,* $\alpha \in \{-1, -1.5\}$*; see Fig. 4)* across the activation manifold. We therefore constrain $\alpha > 0$, which mathematically guarantees C^1 continuity (continuity at the shared junction between the two curves) at the operational points.

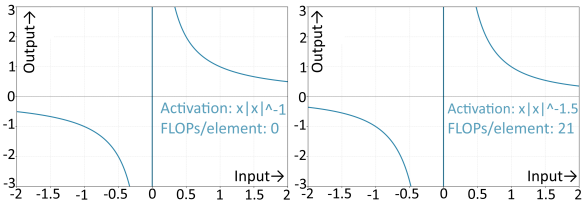

Figure 4: LyRIC Activation with $\alpha \in \{-1, -1.5\}$.

The activation manifolds for different hyperparameter configurations exhibit **analogous topological characteristics** as demonstrated in Fig. 5. Nevertheless, configuring the expansion coefficient to $\alpha = 1$ yields a considerably lower FLOPs footprint than $\alpha \in \{0.5, 1.5, 2, 2.5, 3\}$ *(see Table 1 for FLOPs footprint analysis of hyperparameter* α *and Table 7 for ablation on hyperparameter* α*).* Therefore, we formalize our canonical LyRIC activation function as the following which ensures well-behaved gradient flow while simultaneously achieving optimal hardware-level computational efficiency suitable for high-throughput ETS synthesis:

$$f(x) = x|x|^{\alpha}, \quad \text{where } \alpha = 1 \quad (6)$$

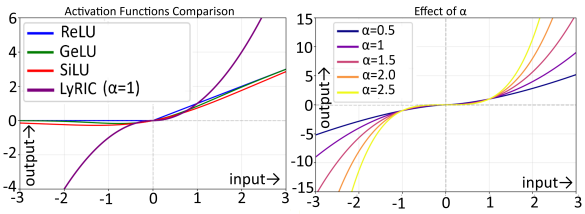

Figure 5: (Left) Comparison of different activation functions. (Right) Effect of α on LyRIC Activation .

Figure 5 compares ReLU, GELU, and SiLU against LyRIC at α=1 (left), and illustrates the effect of varying $\alpha \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$ on the LyRIC activation shape (right). Furthermore, Table 1 shows that fractional values of α in the LyRIC activation (P_0, P_2, P_4) incur 11–13 FLOPs/element due to non-integer exponentiation, while higher-order integers (P_3, P_5) reduce this to 2–3 FLOPs via successive multiplications. The canonical α=1 (P_1) simplifies the leading term to a sign-preserving quadratic map ($x|x|$), eliminating exponentiation entirely and achieving the theoretical minimum of 1 FLOP/element while preserving the topological properties of LyRIC. A rigorous empirical analysis is provided in Appendix A.3.

Table 1: FLOPs footprint analysis of hyperparameter α

	Function	FLOPs/element↓
P_0	$\alpha = 0.5$	11
P_1	$\alpha = 1$	1
P_2	$\alpha = 1.5$	12
P_3	$\alpha = 2$	2
P_4	$\alpha = 2.5$	13
P_5	$\alpha = 3$	3

2.3 Convolutional Feature Extraction (CFE)

CFE block (see Fig. 1) comprises 3 residual blocks. Each block has two 1D convolutional (Conv-1D) layers with ReLU, augmented by a *parameterized* skip connection to guarantee dimensional congruency for the residual connection (Appendix A.1 for details). A uniform channel capacity of $C_D = 512$ is rigorously maintained across all layers. *This configuration is designed to achieve superior EMG synthesis with a considerably reduced parameter footprint (see Table 9 for channel ablation).*

2.4 GLARE-fused Mamba Encoder–GLAME

Succeeding the CFE front-end, we introduce a novel encoder formulated to achieve strict parameter efficiency while robustly modeling the complex, non-linear temporal dependencies of EMG signals. We empirically restrict the encoder depth to four layers (*see Fig. 1 & Table 8 for layer count ablation*). Two of the layers seamlessly integrate Mamba state-space block (Gu and Dao, 2024), our localized GLARE attention mechanism, and a terminal Feed Forward Network (FFN). We interleave the GLARE attention mechanism exclusively within the first and third layers of the encoder. This selective attention routing maximizes local context aggregation while sharply minimizing computational overhead (*see Table 6 for GLARE ablation*). Collectively, we name the four-layered encoder *(see Fig. 1)* as GLAME (GLARE-fused Mamba Encoder). To facilitate multi-task learning, the final latent representations of the GLAME bifurcate into two task-specific projection heads. The primary acoustic head applies a linear transformation, mapping the hidden states directly into an 80-dimensional melspectrogram space for high-fidelity waveform inversion. Concurrently, an auxiliary linguistic head derives phonetic probability distributions that enables explicit phoneme-level supervision during training to rigorously enforce

structural articulatory constraints.

2.5 Design of the LER Loss ($\mathcal{L}_{\text{LER}}$)

Motivation: We introduce a novel Lyapunov Exponent Regularization (LER) loss to capture the rapid, nonlinear fluctuations and sensitivity to initial conditions inherent in articulatory biosignals and dynamic properties that *conventional reconstruction objectives typically overlook.* LER is built on Lyapunov Exponent (LE) (Oseledec, 1968), which fundamentally quantifies the exponential rate of divergence between infinitesimally close state-space trajectories. *As chaos is inherent in speech, we use LE to reconstruct the underlying chaotic attractors of speech by directly penalizing mismatches in the local divergence rates between predicted and target sequences, yielding highly natural, temporally coherent acoustic features that faithfully preserve the complex micro-structure of human phonation.*

Loss Calculation (see Alg. 1): The extracted aligned EMG features from the encoder output are passed to compute the logarithmic divergence $S(k)$ of nearest-neighbor pairs over time for each k, as described in Section 2.2 (**Row** ⑭). The largest LE (λ) is then estimated as the weighted linear slope of $S(k)$: $\lambda = \frac{\sum_{i=1}^{n_k} \tilde{k}_i \tilde{S}_i}{\sum_{i=1}^{n_k} \tilde{k}_i^2 + \epsilon_2}$ where $\epsilon_2 = 10^{-8}$ prevents division by zero. $\tilde{k}$ and $\tilde{S}$ represent the centered values of time offsets and logarithmic divergences, respectively. Then, we define LER loss as $\mathcal{L}_{\text{LER}}(\hat{Y}_S[j], Y_V[i]) = |\lambda^{(\hat{Y}_S[j])} - \lambda^{(Y_V[i])}|$, where $\hat{Y}_S[j]$ and $Y_V[i]$ denote temporally aligned predicted and target audio features, respectively (**Row** ⑮).

Algorithm 1: Logarithmic Divergence & LER

Require: EMG Feature sequence $x \in \mathbb{R}^{T \times C}$, $m = 10, \tau = 1$
Ensure: Lyapunov exponent λ and LER loss $\mathcal{L}_{\text{Lyap}}$
Compute $M \leftarrow T - (m-1)\tau, d \leftarrow m \cdot C$
for $i = 0$ to $M - 1$ **do**
  $E_V[i] \leftarrow [x[i,:], x[i+\tau,:], \ldots, x[i+(m-1)\tau,:]]$
end for
for $i, j \in \{0, \ldots, M-1\}$ **do**
  $D[i,j] \leftarrow \|E_V[i] - E_V[j]\|_2$
end for
for $i = 0$ to $M - 1$ **do**
  Find $j^*(i)$ minimizing $D[i,j]$ (exclude Theiler window)
  for $k = 1$ to $k_{\max}$ **do**
    $d_k[i] \leftarrow \|E_V[i+k] - E_V[j^*(i)+k]\|_2$
  end for
end for
$S(k) \leftarrow \frac{1}{N_k} \sum_i \log(d_k[i] + \epsilon_1)$, Center k and $S(k)$ to get $\tilde{k}, \tilde{S}$
Calculate $\lambda \leftarrow \frac{\sum \tilde{k}_i \tilde{S}_i}{\sum \tilde{k}_i^2 + \epsilon_2}$ and $\mathcal{L}_{\text{LER}} \leftarrow |\lambda^{(\hat{Y}_S)} - \lambda^{(Y_V)}|$
return λ, $\mathcal{L}_{\text{LER}}$

2.6 Design of the MSDFA Loss ($\mathcal{L}_{\text{MSDFA}}$)

We introduce a Multi-Scale Detrended Fluctuation Analysis (MSDFA) objective (Peng et al., 1994) to quantify the fractal-like, long-range temporal correlations strictly overlooked by conventional frame-level spectrogram losses. *Empirically, omitting the MSDFA objective causes the network to degenerate into generating smoothed, muffled prosody lacking natural dynamic variation, which underscores the critical necessity of long-range chaotic correlations for intelligible speech synthesis.*

Loss Calculation (see Alg. 2): For each channel $c \in \{1, \ldots, C\}$, we compute the mean μ_c and define the profile (integrated features) as the cumulative sum of deviations from the mean $y[k,c]$ where $k \in \{1, \ldots, T\}$ and T is the number of time frames. This sum constructs the windowed profile tensor Y_s (**Rows** ①–③). The design matrix $X_s \in \mathbb{R}^{s \times 2}$ is defined with rows $[1, \tau_j]$ where $\tau_j = (j-1)/(s-1)$ represents the normalized time within the segment for $j \in \{1, \ldots, s\}$ and the projection matrix onto the column space of X_s is $P_s = X_s(X_s^T X_s)^{-1} X_s^T \in \mathbb{R}^{s \times s}$. The detrending kernel can be denoted by $K_s = I_s - P_s \in \mathbb{R}^{s \times s}$ where $I_s \in \mathbb{R}^{s \times s}$ is the identity matrix. The detrended residuals are computed as $R_s[c,i,j] = \sum_{j'=1}^{s} Y_s[c,i,j'] \cdot K_s[j',j]$ (**Rows** ④–⑦). Here, $i \in \{1, \ldots, N_s\}$ and $j \in \{1, \ldots, s\}$ are segment (window) index and position within the segment, respectively. The MSDFA scaling component for channel c is calculated using the **Row** ⑧, where $\epsilon_3 = 10^{-8}$ prevents division by zero. The $\log s[i]$ and $\log F[c,i]$ are centered logarithmic values, which are root-mean-square across all R_s. The s and F are pre-defined set of scales for MSDFA function. Given, temporally aligned $\hat{Y}_S[j]$ and $Y_V[i]$, we define MSDFA loss as $\mathcal{L}_{\text{MSDFA}} = \frac{1}{|\mathcal{V}|} \sum_{c \in \mathcal{V}} \left| \alpha_c^{(\hat{Y}_S[j])} - \alpha_c^{(Y_V[i])} \right|$, where $|\mathcal{V}|$ is cardinality of the valid channel set.

Algorithm 2: MSDFA Loss Computation

Require: EMG Feature sequence $x \in \mathbb{R}^{T \times C}$, scales s
for $c = 1$ to C **do**
  $\mu_c \leftarrow \text{mean}(x[:,c])$
  $Y_s \leftarrow \text{cumsum}(x[:,c] - \mu_c)$
  Construct X_s with rows $[1, \tau_j]$, $\tau_j = (j-1)/(s-1)$
  $P_s \leftarrow X_s(X_s^T X_s)^{-1} X_s^T, \quad K_s \leftarrow I_s - P_s$
  $R_s \leftarrow Y_s K_s$ (detrended residuals)
  Compute $F[c,i]$ as RMS over R_s for each scale
  $\alpha_c \leftarrow \frac{\sum \log s[i] \log F[c,i]}{\sum \log s[i]}^2 + \epsilon_3$
end for
$\mathcal{L}_{\text{MSDFA}} \leftarrow \frac{1}{|\mathcal{V}|} \sum_{c \in \mathcal{V}} |\alpha_c^{(\hat{Y}_S)} - \alpha_c^{(Y_V)}|$

2.7 Total Loss Function

We employ Dynamic Time Warping (DTW), Pairwise Distance (PD) loss ($\mathcal{L}_{PD}$), and phoneme loss ($\mathcal{L}_{Ph}$) from (Gaddy and Klein, 2021), together with our two novel *chaos-inspired* loss functions, $\mathcal{L}_{LER}$ and $\mathcal{L}_{MSDFA}$. The $\mathcal{L}_{LER}$ is designed to capture chaotic temporal irregularities, whereas $\mathcal{L}_{MSDFA}$ models chaotic multi-scale dynamics. The encoder generates 80 melspectrogram bins, which are temporally aligned at the frame level using DTW. These aligned bins are then used to compute the LE and MSDFA. The total loss $\mathcal{L}_{\text{total}}$:

$$\mathcal{L}_{total} = \mathcal{L}_{PD} + \mathcal{L}_{Ph} + \mathcal{L}_{LER} + \mathcal{L}_{MSDFA} \quad (7)$$

3 Comprehensive Analysis

3.1 Evaluation Metrics

We present a comprehensive evaluation of reconstructed audio in ETS synthesis which incorporates Log Spectral Distance (LSD), Short-Time Objective Intelligibility (STOI), Perceptual Evaluation of Speech Quality (PESQ), Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), Non-Intrusive Speech Quality Assessment Mean Opinion Score (NISQA-MOS), Floating-Point Operations (FLOPs), Inference Time (IT) per batch, Word Error Rate (WER), and Real-Time Factor (RTF). *No prior work in the literature has employed such an extensive set of evaluation metrics for ETS systems.*

3.2 Post-Vocoder Alignment (PVA)

Prior ETS frameworks (Gaddy and Klein, 2020, 2021) suffer from vocoder-induced temporal misalignments that prevent them to use standard frame-level acoustic metrics, such as LSD, STOI, and PESQ strictly. To solve this problem, for the first time, we use dynamic time warping (DTW) at the post-vocoder stage. We specifically extract Short-Time Fourier Transform (STFT) magnitude spectrograms, S_g and S_t, from the generated and ground-truth audio, respectively. We compute a cosine distance matrix formulated as $C = \mathbf{1} - \hat{S}_g \hat{S}_t^T$ following frame-wise normalization, where $\hat{S}_g$ and $\hat{S}_t$ denote the L_2-normalized spectral features. The optimal warping path, derived via DTW, is subsequently applied to align the vocoder-synthesized representations before reconstructing the waveform with the inverse STFT (ISTFT).

3.3 Hyperparameter and Configuration

We train for 80 epochs with batch size 32 using Adaptive Moment Estimation with Weight Decay (AdamW) optimizer having a learning rate, $\eta = 10^{-3}$ and weight decay $= 10^{-7}$. Gradients are clipped to norm 0.5. Models are trained in FP32 precision using distributed data parallelism across two NVIDIA RTX 4090 GPUs *(see Appendix A.8 for further hyperparameter & configuration details).*

3.4 Dataset and Preprocessing

We train our model on the SOTA gold-standard open dataset introduced by (Gaddy and Klein, 2020), comprising 19 hours of facial EMG data collected under silent and vocalized speech conditions. We do minimal preprocessing to facilitate stable model optimization *(see Appendix A.5 for detailed preprocessing). Noises are not synthetically added as articulatory EMG inherently contains real-world noises – muscle cross-talk, electrode shifts, and electromagnetic interference. Hence, we use unmodified EMG data for ablations to ensure our model's validity in noisy conditions.*

3.5 Comparative Analysis with Baselines

Table 2 presents a comprehensive quantitative evaluation of our proposed CGS-ETS against two widely established SOTA baselines: (Gaddy and Klein, 2020) (P_0) and (Gaddy and Klein, 2021) (P_1). Historically, prior works (Gaddy and Klein, 2020, 2021; Scheck et al., 2023; Xie et al., 2025) report only WER, and none of them do a comprehensive quality assessment using LSD, STOI, PESQ, SI-SDR, and NISQA-MOS due to temporal misalignment issues in the post-vocoder stage. However, our PVA approach enables us to rigorously compare a full suite of perceptual and structural metrics (LSD, STOI, PESQ, SI-SDR, NISQA).

Comparing our architecture against the strongest baseline (P_1) reveals that CGS-ETS establishes a definitive new standard for acoustic fidelity and signal reconstruction. Our chaos-inspired framework with the GLARE attention delivers a **2.08x improvement** in spectral reconstruction accuracy (LSD = 1.08 vs 2.25), a **4.69x increase** in objective speech intelligibility (STOI = 0.61 vs 0.13), and a **1.28x enhancement** in noise suppression (SI-SDR = -32.61 vs -41.96). CGS-ETS yields an improved WER of 41.36%. Furthermore, CGS-ETS improves over-smoothed prosody characteristic of non-chaotic baselines, yielding a superior perceptual quality (NISQA = 3.33 vs 3.30).

Table 2: Comparison Against SOTA Methods. Here, L=LSD, ST=STOI, P=PESQ, SD=SI-SDR, N=NISQA-MOS, W=WER.

	Model	L↓	ST↑	P↑	SD↑	N↑	W↓	PVA	Para.↓
P_0	(Gaddy and Klein, 2020)	3.23	0.04	1.17	-52.52	3.28	68.00%	×	60.21M
P_1	(Gaddy and Klein, 2021)	2.25	0.13	1.20	-41.96	3.30	42.20%	×	54.10M
P_2	**CGS-ETS (Proposed)**	**1.08**	**0.61**	**1.19**	**-32.61**	**3.33**	**41.36%**	✓	**14.34M**

Please note that CGS-ETS achieves reconstruction improvement with a significant degree of architectural compression. As detailed in Table 2, our model operates with a mere **14.34M parameters**, contrasting with the highly parameter-dense **54.10M** requirement of P_1. This constitutes a remarkable **73.49% reduction** in total model capacity. *These efficiency metrics unequivocally validate our core theoretical hypothesis: explicitly integrating non-linear chaotic dynamics (through LyRIC, LER, and MSDFA) and bounding local context (GLARE) provides a superior but structurally compact inductive bias for ETS synthesis than relying on the brute-force mapping of massive multi-head attention transformers.*

Please note that we use two baselines P_0 and P_1 for comparison, as *recent SOTA works (Scheck et al., 2023; Ren et al., 2024; Xie et al., 2025) do not report model size or computational complexity, and their codebases are not publicly available, precluding direct efficiency comparisons.*

3.6 Computational Complexity

Table 3 compares the computational efficiency of CGS-ETS against the best-performing reproducible baseline (Gaddy and Klein, 2021). CGS-ETS achieves *substantial* parameter reduction across all components while maintaining real-time inference. The encoder parameters decrease from 44.03M to 9.85M, which is a **77.63% reduction** and the CFE parameters decrease from 10.07M to 4.49M (**55.41% reduction**), yielding a total parameter reduction from 54.10M to 14.34M (**73.49% reduction**). CGS-ETS further shows a considerable improvement in computational efficiency, reducing FLOPs from 2.40G to 926.55M (**61.39% reduction**) and inference time from 5.97ms to 5.61ms per batch, maintaining an identical RTF of 0.0057 for real-time processing. These efficiency gains are accompanied by improvements in audio quality and intelligibility *(see Table 2)*.

Table 3: Study on Computational Efficiency.

	Model	CFE	Enc.	Total	IT	FLOPs	RTF
P_0	(Gaddy and Klein, 2021)	10.07M	44.03M	54.10M	5.97ms	2.40G	0.0057
P_1	**CGS-ETS (Proposed)**	**4.49M**	**9.85M**	**14.34M**	**5.61ms**	**926.55M**	**0.0057**

3.7 Ablation Study

3.7.1 PVA and Loss Components

Table 4 isolates the individual and synergistic contributions of PVA and our novel chaotic losses. Comparing the unaligned model (P_1) to the PVA-integrated architecture (P_2) empirically proves that enforcing exact temporal signal synchronization fundamentally resolves catastrophic acoustic degradation. This results in an immediate **4.28x surge** in intelligibility (STOI: 0.14→0.60) and a **44.4% reduction** in spectral distortion (LSD: 1.98→1.10). The introduction of our chaos-theory objectives to the acoustically aligned model (P_2) enforces critical micro-structural constraints. For example, isolated LER (P_3) provides localized gradient regularization, dropping WER by **5.85%** (51.56% → 48.34%). Conversely, MSDFA (P_4) captures macro-level fractal correlations, reducing WER by **11.47%** (51.56%→45.65%) while simultaneously boosting perceptual naturalness (NISQA: 3.31). Ultimately, optimizing both objectives (P_5) achieves the optimal results that mathematically constrain the *deterministic chaotic invariants of speech* to minimize WER to **41.36%** while maximizing absolute acoustic fidelity (STOI: 0.61, LSD: 1.08, SI-SDR: -32.61).

Table 4: Ablation Study on PVA and Chaotic Losses.

	Model Configuration	L↓	ST↑	P↑	SD↑	N↑	W↓	Para.↓
P_0	(Gaddy and Klein, 2021)	2.25	0.13	1.20	-41.96	3.30	42.20%	54.10M
P_1	CGS-ETS w/o PVA+LER+MSDFA	1.98	0.14	1.17	-35.78	3.28	51.56%	14.34M
P_2	CGS-ETS w PVA w/o LER+MSDFA	1.10	0.60	1.18	-32.53	3.30	51.56%	14.34M
P_3	CGS-ETS w PVA+LER w/o MSDFA	1.10	0.60	1.18	-33.70	3.29	48.34%	14.34M
P_4	CGS-ETS w PVA+MSDFA w/o LER	1.09	0.60	1.19	-32.60	3.31	45.65%	14.34M
P_5	**CGS-ETS w PVA+LER+MSDFA**	**1.08**	**0.61**	**1.19**	**-32.61**	**3.33**	**41.36%**	**14.34M**

3.7.2 Activation Functions

Table 5 evaluates activation routing across the Residual blocks in CFE, Mamba, and FFNs. Swapping SwiGLU for GeLU in the terminal FFN ($P_0 \rightarrow P_1$) cuts FLOPs (**1.038G→929.76M**) and improves WER (**51.16%→47.06%**), while misplacing LyRIC in the residual blocks (P_2) or FFNs (P_3) destabilizes acoustic representation. Anchoring LyRIC within the Mamba alone (P_4) yields peak spectral and intelligibility scores (LSD: 1.08; STOI: 0.61) but, when paired with SwiGLU, reintroduces computational bloat (1.035G FLOPs) and stalls WER at 45.71%. The optimal configuration (P_5) – LyRIC-driven Mamba with a GeLU periphery achieves the best phonetic accuracy (**WER: 41.36%**), perceptual fidelity (NISQA: 3.33), and lowest inference (FLOPs: 926.55M; RTF: 0.0057).

Table 5: Ablation Study on Activation Functions.

	CFE	Mam.	FFN	L↓	ST↑	P↑	SD↑	N↑	W↓	IT↓	FLOPs↓	RTF↓
P_0	ReLU	SiLU	SwiGLU	1.10	0.59	1.16	-31.47	3.25	51.16%	7.12ms	1.038G	0.0075
P_1	ReLU	SiLU	GeLU	1.09	0.60	1.18	-31.30	3.28	47.06%	5.92ms	929.76M	0.0059
P_2	LyRIC	SiLU	GeLU	1.16	0.54	1.13	-33.22	2.90	71.86%	5.92ms	929.76M	0.0059
P_3	ReLU	SiLU	LyRIC	1.10	0.59	1.17	-31.85	3.23	52.63%	5.70ms	928.43M	0.0057
P_4	ReLU	LyRIC	SwiGLU	1.08	0.61	1.19	-32.63	3.30	45.71%	6.21ms	1.035G	0.0062
P_5	**ReLU**	**LyRIC**	**GeLU**	**1.08**	**0.61**	**1.19**	**-32.61**	**3.33**	**41.36%**	**5.61ms**	**926.55M**	**0.0057**

3.7.3 GLARE Attention Depth

Table 6 evaluates the optimal depth routing for GLARE attention within the encoder hierarchy. No attention (P_0) confirms that pure Mamba lacks the fine-grained local receptive field needed for coarticulatory decoding (WER: 73.22%), while saturating all layers (P_6) bloats computation (1.006G FLOPs; IT: 7.58ms) and induces representational redundancy (WER: 44.56%). Contiguous or imbalanced placements (P_1–P_5) fail to balance local and global context, stalling WER between 44.91% and 50.25%. The proposed alternating configuration (P_7) – GLARE exclusively at Layers 1 and 3 resolves this bottleneck by synchronizing local attention with global state-space mappings, achieving the lowest WER (41.36%), highest perceptual fidelity (NISQA: 3.33), and minimal inference cost (IT: 5.61ms; params: 14.08M).

Table 6: Ablation Study on GLARE Attention.

	Layers	L↓	ST↑	P↑	SD↑	N↑	WER↓	Para.↓	IT↓	FLOPs↓	RTF↓
P_0	No Attention	1.15	0.53	1.14	-34.42	2.91	73.22%	12.68M	4.23ms	848.33M	0.0046
P_1	$L = 1, 2$	1.08	0.60	1.18	-32.41	3.30	46.14%	14.34M	5.91ms	926.55M	0.0059
P_2	$L = 3, 4$	1.10	0.59	1.17	-31.96	3.25	50.25%	14.34M	5.97ms	926.55M	0.0059
P_3	$L = 1, 4$	1.10	0.59	1.17	-33.54	3.30	49.01%	14.34M	5.91ms	926.55M	0.0061
P_4	$L = 1, 2, 3$	1.08	0.61	1.19	-33.31	3.26	46.38%	15.05M	6.86ms	966.45M	0.0068
P_5	$L = 2, 3, 4$	1.09	0.60	1.19	-32.81	3.28	44.91%	15.05M	7.11ms	966.45M	0.0071
P_6	$L = 1, 2, 3, 4$	1.08	0.61	1.18	-31.94	3.24	44.56%	15.83M	7.58ms	1.006G	0.0077
P_7	$L = 1, 3$	**1.08**	**0.61**	**1.19**	**-32.61**	**3.33**	**41.36%**	**14.34M**	**5.61ms**	**926.55M**	**0.0057**

3.7.4 Hyperparameter α in LyRIC Activation

Table 7 profiles the non-linear expansion coefficient α, which governs phase-space divergence in LyRIC. Setting $\alpha = 0$ (P_0) degenerates LyRIC to a linear identity ($f(x) = x$), ablating chaotic capacity and collapsing WER to 48.22%, while a fractional $\alpha = 1.5$ (P_1) restores intelligibility at the cost of non-integer exponentiation overhead (929.89M FLOPs; RTF: 0.0060). Higher-order expansions

($\alpha \in \{2, 3\}$; P_2, P_3) induce overly aggressive state-space trajectories, destabilizing the gradient manifold and spiking WER to 49.26%. The canonical formulation (P_4, $\alpha = 1$) simplifies LyRIC to a lightweight sign-preserving quadratic ($x|x|$), optimally balancing chaotic dynamics with topological stability and achieving the lowest WER (41.36%), highest perceptual naturalness (NISQA: 3.33), and minimal inference (IT: 5.61ms; RTF: 0.0057).

Table 7: Ablation Study on Different Values of Hyperparameter α of LyRIC Activation

	$x\|x\|^\alpha$	L↓	ST↑	P↑	SD↑	N↑	W↓	IT↓	FLOPs↓	RTF↓
P_0	$\alpha = 0$	1.09	0.60	1.18	-34.14	3.30	48.22%	5.98ms	926.95M	0.0058
P_1	$\alpha = 1.5$	1.09	0.61	1.18	-30.79	3.26	44.73%	5.95ms	929.89M	0.0060
P_2	$\alpha = 2$	1.08	0.60	1.18	-32.25	3.30	45.40%	5.95ms	927.20M	0.0059
P_3	$\alpha = 3$	1.09	0.60	1.18	-32.98	3.30	49.26%	6.00ms	927.33M	0.0059
P_4	$\alpha = 1$	**1.08**	**0.61**	**1.19**	**-32.61**	**3.33**	**41.36%**	**5.61ms**	**926.55M**	**0.0057**

3.7.5 Layer Count of GLAME

Table 8 evaluates the hierarchical depth (L_c) of the GLAME's constitutional blocks. A shallow $L_c = 2$ topology (P_0) yields an extreme lightweight module (5.67M params) but underfits coarticulatory transitions, collapsing WER to 55.94%. Extending to $L_c = 5$ (P_2) minimizes WER (41.91%) but violates low-latency constraints through excessive computational bloat (14.18M encoder params; >1.1G FLOPs) and degrades perceptual naturalness (NISQA: 3.25), indicating acoustic oversmoothing. The optimal configuration (P_3, $L_c = 4$) balances hierarchical capacity with compression, saturating spectral metrics (STOI: 0.61; PESQ: 1.19) and achieving the highest perceptual fidelity (NISQA: 3.33) and lowest WER (41.36%) while maintaining sub-10M parameters (9.85M) and real-time viability (RTF: 0.0057).

Table 8: Ablation Study on Layer Count of GLAME

	Count	L↓	ST↑	P↑	SD↑	N↑	W↓	Enc.↓	IT↓	FLOPs↓	RTF↓
P_0	$L_c = 2$	1.10	0.58	1.16	-32.75	3.20	55.94%	5.67M	4.52ms	721.48M	0.0045
P_1	$L_c = 3$	1.08	0.60	1.18	-31.45	3.33	45.18%	8.51M	5.37ms	863.65M	0.0057
P_2	$L_c = 5$	1.08	0.61	1.19	-33.06	3.25	41.91%	14.18M	8.50ms	1.148G	0.0091
P_3	$L_c = 4$	**1.08**	**0.61**	**1.19**	**-32.61**	**3.33**	**41.36%**	**9.85M**	**5.61ms**	**926.55M**	**0.0057**

3.7.6 Channel Dimension of CGS-ETS

Table 9: Ablation Study on Channel Dimension

	Dimension	L↓	ST↑	P↑	SD↑	N↑	W↓	Para.↓	IT↓	FLOPs↓	RTF↓
P_0	$C_D = 128$	1.13	0.57	1.15	-34.35	3.09	58.88%	0.91M	5.53ms	59.64M	0.0055
P_1	$C_D = 256$	1.11	0.59	1.16	-33.09	3.20	52.14%	3.59M	5.59ms	234.030M	0.0056
P_2	$C_D = 1024$	1.08	0.61	1.19	-32.67	3.34	45.10%	56.82M	15.72ms	3.690G	0.0088
P_3	$C_D = 512$	**1.08**	**0.61**	**1.19**	**-32.61**	**3.33**	**41.36%**	**14.34M**	**5.61ms**	**926.55M**	**0.0057**

We evaluate the effect of internal channel capacity (C_D) on continuous EMG signal modeling in Table 9. Small capacities ($C_D \in \{128, 256\}$; P_0, P_1) deliver lightweight models (≤3.59M parameters) but insufficient dimensionality to preserve chaotic trajectories that results in phonetic collapse (WER > 52.14%) and degraded perceptual naturalness (NISQA ≤ 3.20). Contrarily, C_D=1024 (P_2) saturates acoustic metrics (STOI: 0.61, PESQ: 1.19) but incurs excessive overhead (56.82M parameters; 3.69G FLOPs; IT: 15.72ms) and overfitting, stalling WER at 45.10%. The proposed C_D=512 (P_3) balances these extremes, achieving the best WER (a 3.74% absolute reduction over P_2) and peak spectral fidelity (LSD: 1.08) at a compact 14.34M total model parameters and RTF of 0.0057.

3.7.7 Different Attention Mechanisms

Table 10 isolates the impact of attention mechanism choice within the GLAME. Full MHA across all layers (P_0) incurs the highest computational overhead (14.49M; 1.166G FLOPs) and inference latency (7.42ms), yet yields a suboptimal WER of 43.32% which indicates representational redundancy. Restricting MHA to alternating layers (P_1) reduces cost marginally but degrades WER to 44.73%, while MQA variants (P_2, P_3) collapse representational capacity, resulting in a STOI of 0.60 and SI-SDR of -31.81. Our proposed GLARE (P_4) resolves this trade-off by integrating GQA and SWA with relative positional embeddings at $L \in \{1, 3\}$, achieving the best WER (41.36%), peak perceptual naturalness (NISQA: 3.33), fastest inference (5.61ms), and lowest RTF (0.0057) at only 9.85M encoder parameters.

Table 10: Ablation Study on Different Attentions

	Attention Type	L↓	ST↑	P↑	SD↑	N↑	W↓	Enc.↓	IT↓	FLOPs↓	RTF↓
P_0	MHA (All Layers)	1.08	0.61	1.19	-32.43	3.30	43.32%	14.49M	7.42ms	1.166G	0.0074
P_1	MHA ($L = 1, 3$)	1.08	0.61	1.18	-32.36	3.27	44.73%	12.39M	5.92ms	1.061G	0.0059
P_2	MQA (All Layers)	1.07	0.60	1.18	-31.69	3.33	44.85%	10.56M	7.10ms	966.50M	0.0071
P_3	MQA ($L = 1, 3$)	1.08	0.60	1.18	-31.81	3.31	44.90%	9.38M	5.69ms	907.41M	0.0058
P_2	**GLARE** ($L = 1, 3$)	**1.08**	**0.61**	**1.19**	**-32.61**	**3.33**	**41.36%**	**9.85M**	**5.61ms**	**926.55M**	**0.0057**

3.8 Subjective Mean Opinion Score (MOS)

Subjective comparison of CGS-ETS against (Gaddy and Klein, 2021) and unprocessed EMG is conducted by a panel of 10 persons. We use 5-point (1=bad to 5=excellent) MOS ratings for subjective evaluation. The unprocessed EMG has MOS=1, meaning completely unrecognizable. *Our CGS-ETS, having MOS=4.25, outperforms (Gaddy and Klein, 2021), which has MOS=3.98. Therefore, we can comment that CGS-ETS may better reconstruct missing chaotic speech artifacts from the raw EMG data that is neglected by the previous work.* Results provide strong evidence that CGS-ETS consistently generates perceptually higher quality audio, favored by a wide range of listeners. We refer to Appendix A.10 for details on MOS analysis.

4 Conclusion

CGS-ETS is the first ETS architecture rooted in nonlinear dynamical systems and deterministic chaos theory. Chaos-guided ***LyRIC*** activation and ***GLARE*** attention formulated within the parameter-efficient GLAME, combined with chaos-inspired objectives ***(LER** and **MSDFA**)* establishes a strong theoretical grounding for articulatory-to-acoustic mapping, while PVA enables frame-level evaluation (LSD, STOI, SI-SDR, PESQ, NISQA-MOS) for the first time in ETS literature. On the Gaddy benchmark (Gaddy and Klein, 2020), CGS-ETS achieves superior fidelity and intelligibility with 73.49% fewer parameters and lower inference FLOPs which validates nonlinear dynamics as an efficient inductive prior for real-time SSI.

5 Acknowledgment

We are thankful to Gaddy et al. (Gaddy and Klein, 2020) for releasing the EMG-to-speech dataset as open-source. We also express gratitude to the participants of the subjective evaluation tests for their contributions to the listening tests. We acknowledge that we have used ChatGPT for debugging codes, and finding grammatical errors to improve the manuscript.

6 Limitations

Our current work establishes a new baseline using EMG-to-speech datatset for English phonetics from (Gaddy and Klein, 2020), which is the gold standard for EMG-to-speech synthesis. Our future work will extend this chaos-driven paradigm to multi-corpus and multi-lingual environments while our current architecture robustly handles intrinsic physiological artifacts across extrinsic acoustic degradation.

7 Potential Risks/Ethical Considerations

While the intention of designing the CGS-ETS model is for speech restoration research purposes, it can be misused for potential secret eavesdropping, impersonation, or deepfake audio generation. This model has the potential for severe privacy and security risks. To avoid these, we have to be very careful to ensure transparency, protect consent, and always follow guidelines.

References

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Penghao Dong, Yizong Li, Si Chen, Justin T Grafstein, Irfaan Khan, and Shanshan Yao. 2023. Decoding silent speech commands from articulatory movements through soft magnetic skin and machine learning. *Materials Horizons*, 10(12):5607–5620.

Dario Farina, Roberto Merletti, and Roger M Enoka. 2014. The extraction of neural strategies from the surface emg: an update. *Journal of applied physiology*, 117(11):1215–1230.

David Gaddy and Dan Klein. 2020. Digital voicing of silent speech. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5521–5530, Online. Association for Computational Linguistics.

David Gaddy and Dan Klein. 2021. An improved model for voicing silent speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 175–181, Online. Association for Computational Linguistics.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Yuta Kurotaki, Shunsuke Yamakoshi, Reitaro Yoshida, Yutaka Isoda, Tamami Takano, Yuji Isano, Yusuke Miyake, Kentaro Kuribayashi, and Hiroki Ota. 2025. Soft active emg interface for machine learning-enabled silent speech recognition.

Edward N. Lorenz. 1963. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141.

Roberto Merletti and Dario Farina. 2016. *Surface Electromyography: Physiology, Engineering, and Applications*. John Wiley & Sons.

V.I. Oseledec. 1968. A multiplicative ergodic theorem: Lyapunov characteristic numbers for dynamical systems. *Trans. Moscow Math. Soc.*, 19:197–231.

P Padmanabhan and S Puthusserypady. 2004. Nonlinear analysis of emg signals: A chaotic approach. *IEEE Transactions on Biomedical Engineering*, 51(6):1064–1067.

C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. 1994. Mosaic organization of dna nucleotides. *Physical Review E*, 49(2):1685–1689.

Ph Ravier, O Buttelli, and P Couratier. 2005. Fractal and chaos analysis of emg signals. *Journal of Electromyography and Kinesiology*, 15(4):400–413.

Zhao Ren, Kevin Scheck, Qinhan Hou, Stefano van Gogh, Michael Wand, and Tanja Schultz. 2024. Diff-ets: Learning a diffusion probabilistic model for electromyography-to-speech conversion. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. IEEE.

Michael T. Rosenstein, James J. Collins, and Carlo J. De Luca. 1993. A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena*, 65(1–2):117–134.

Nicholas Satterlee, Xiaowei Zuo, Kee Moon, Sung Q Lee, Matthew Peterson, and John S Kang. 2025. Sentence-level silent speech recognition using a wearable emg/eeg sensor system with ai-driven sensor fusion and language model. *Sensors*, 25(19):6168.

Kevin Scheck, Darius Ivucic, Zhao Ren, and Tanja Schultz. 2023. Stream-ets: Low-latency end-to-end speech synthesis from electromyography signals. In *Speech Communication; 15th ITG Conference*, pages 200–204. VDE.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*.

Steven H. Strogatz. 1994. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Perseus Books, Reading, MA.

James Theiler. 1986. Spurious dimension from correlation algorithms applied to limited time-series data. *Physical Review A*, 34(3):2427–2432.

Emanuel Todorov. 2004. Optimality principles in sensorimotor control. *Nature neuroscience*, 7(9):907–915.

Alan Wolf, J. B. Swift, Harry L. Swinney, and John A. Vastano. 1985. Determining lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16(3):285–317.

Liang Xie, Yakun Zhang, Hao Yuan, Meishan Zhang, Xingyu Zhang, Changyan Zheng, Ye Yan, and Erwei Yin. 2025. Neural chinese silent speech recognition with facial electromyography. *Speech Communication*, 171:103230.

A Appendix

A.1 Convolutional Feature Extraction (CFE)

CFE block comprises three residual blocks following (Gaddy and Klein, 2021). Each block comprises two 1D convolutional (Conv-1D) layers (kernel size=3) interspersed with ReLU activations, augmented by a parameterized skip connection. The first Conv-1D downsamples inputs by $2\times$ (stride=2) while the second one maintains the same resolution (stride=1). The skip connection utilizes a matching projection as the first Conv-1D (stride=2) to guarantee dimensional congruency for the residual connection. A uniform channel capacity of $C_D = 512$ is rigorously maintained across all layers. *This streamlined configuration is designed to capture the rapid, dynamic EMG patterns of facial articulation, which achieves superior phonetic decoding performance with a considerably reduced parameter footprint.*

A.2 EMG as A Chaotic Dynamical System

Articulatory EMG captures the coordinated, spatio-temporal activation of facial and laryngeal muscles during speech production (Gaddy and Klein, 2020, 2021). This myoelectric generation is a *non-linear dynamical process characterized by deterministic chaos* (Padmanabhan and Puthusserypady, 2004; Ravier et al., 2005). The physiological "source" of the signal is the central neural drive, which recruits populations of motor neurons. These neurons subsequently generate asynchronous Motor Unit Action Potentials (MUAPs) across varying muscle fibers. However, before reaching the surface electrodes, these localized action potentials undergo severe non-linear spatial filtering through the visco-elastic subcutaneous tissue, skin layers, and shifting muscle geometries (Merletti and Farina, 2016). The resulting surface EMG is not a simple linear summation, but rather a highly complex interference pattern characterized by amplitude modulation, quasi-periodic bursting, and phase cancellations (Farina et al., 2014).

Essentially, the articulatory neuromuscular system operates through a continuous, closed-loop sensorimotor framework. Motor unit recruitment is dynamically modulated in real-time by proprioceptive feedback from muscle spindles and Golgi tendon organs (Todorov, 2004). Because the central nervous system rapidly adjusts firing rates based on this non-linear feedback, the resulting muscle contractions exhibit extreme sensitivity to initial conditions. Consequently, the EMG signal frequently undergoes bifurcations– such as sudden motor unit synchronization during rapid coarticulation or fatigue-induced tremors—intermittently driving the signal into *chaotic regimes* rigorously indicated by positive Lyapunov exponents (Padmanabhan and Puthusserypady, 2004).

The bio-signal naturally alternates between quasi-stationary periods (e.g., sustained vowel postures) and highly stochastic, irregular micro-fluctuations (e.g., rapid consonant transitions) during continuous phonetic articulation. Moreover, the transient biochemical kinetics of ion channels (e.g., Na^+/K^+ gradients) and varying action potential conduction velocities introduce continuous micro-temporal jitter into the firing patterns (Merletti and Farina, 2016). This creates a non-stationary, fractal-like time-series where minor physiological perturbations map to massive variations in the recorded interference pattern.

A purely linear regression model or standard sequence-to-sequence transformer basically overlooks these critical bio-physical realities (Gaddy and Klein, 2021). Conventional models mathematically average out the sub-harmonic firing frequencies, bifurcations, and aperiodic MUAP bursts by

optimizing strictly for mean-squared reconstruction. In the context of ETS synthesis, this directly causes the network to output unnaturally over-smoothed, muffled acoustic representations devoid of natural prosody and crisp phonetic transitions. Therefore, to faithfully invert this bio-signal into highly intelligible speech, the predictive architecture must explicitly incorporate non-linear, chaos-inspired topological constraints (e.g., LER, MSDFA) capable of mapping both the structured macroscopic muscle gestures and the deterministic chaotic micro-structures of the electrophysiological source.

A.3 Theoretical and Empirical Analysis of the LyRIC Activation Function

Lyapunov theory provides the mathematical foundation for analyzing the stability and dynamic behavior of complex nonlinear systems without requiring explicit analytical solutions to their underlying differential equations. Modeling continuous biological signals such as EMG and chaotic dynamical systems such as speech through Lyapunov theory relies on two fundamental pillars: **Equilibrium Stability** and **Deterministic Chaos**. Equilibrium Stability theory help us design the **LyRIC activation** function, while Deterministic Chaos behavior underpins the formulation of the **LER loss**. A classical Lyapunov function in nonlinear dynamical system must satisfy the Equilibrium Stability conditions.

Equilibrium Stability: This method assesses whether a perturbed system will return to a state of equilibrium. It introduces a scalar, continuously differentiable "energy-like" function, $V(x)$, to evaluate the system state x. For an equilibrium point at the origin ($x = 0$), the system is considered **asymptotically stable** if a $V(x)$ exists such that:

- **Positive Definite:** $V(x) > 0$ for all $x \neq 0$, and $V(0) = 0$. (The system has positive "energy" when displaced).
- **Negative Definite Derivative:** The temporal derivative along the system's trajectory, $\dot{V}(x) = \nabla V(x) \cdot \dot{x}$, is strictly less than zero ($\dot{V}(x) < 0$ for $x \neq 0$).
- **Radically Unbounded:** $V(x) \rightarrow \infty$ as $||x|| \rightarrow \infty$ for global stability.

A.3.1 LyRIC Activation Properties

In section 2.2, we derived the following equation (Equation 2): $g(x) = x\left[\frac{\log(|x|^\alpha+1)}{|x|^\alpha+1}\right]$. From this equation, we derived our main activation function: $f(x) = x|x|^\alpha$. **Log-divergence** equation (Equation 2) and **LyRIC** activation function (Equation 6) are mathematically guaranteed to preserve the following critical topological and optimization properties, thereby establishing Lyapunov's Equilibrium Stability theory:

- **Odd Symmetry and Origin Linearity:** The explicit multiplication by the raw input x rigorously preserves the directional phase (sign) of the latent features, while smoothly degenerating to a trivial zero map ($g(0) = 0$) at the origin.
- **Asymptotic Attenuation:** The stability normalization divisor, $|x|^\alpha + 1$, guarantees that the activation smoothly decays toward zero for excessively large inputs ($\lim_{|x|\rightarrow\infty} g(x) = 0$), acting as a powerful dynamic regularizer against outlier activations.
- **Adaptive Phase-Space Expansion:** The learnable parameter α empowers the network to dynamically optimize the rate of non-linear expansion directly from the morphological features of the physiological data such as EMG.
- **Deterministic Chaotic Mapping:** The core term, $\log(|x|^\alpha + 1)$, directly models the exponential divergence characteristic of chaotic dynamical systems, compelling the network to maintain acute sensitivity to critical micro-variations in the bio-signal.
- **Smoothness and Gradient Stability:** The function is continuously differentiable (C^1 smooth) across its entire domain, ensuring well-behaved, stable gradient propagation throughout deep architectural hierarchies.

A.3.2 Comprehensive Analysis of Log-divergence Equation

Visual Analysis: Figure 3 provides a visual analysis of the Lyapunov-derived Log-divergence equation (Equation 2) manifold. The left panel contrasts its topology against conventional activations (ReLU, GELU, SiLU). While standard functions truncate the negative domain and exhibit unbounded linear growth for positive inputs, the Lyapunov formulation guarantees strict odd symmetry (bilateral phase preservation) and imposes sub-linear asymptotic attenuation, bounding extreme activations to prevent gradient explosion. The right panel empirically demonstrates the parametric influence of the expansion coefficient, α. Variations in α dynamically modulate the non-linear expansion rate with higher values (e.g., $\alpha = 2.0$) enforcing tighter asymptotic bounds while mathematically preserving zero-centered topological stability ($f(0) = 0$) across all configurations.

Output Density Distributions: Figure 6 contrasts the output density distributions of standard activation functions (ReLU, GELU, SiLU) against Log-divergence equation (Equation 6) under a standard normal input sequence ($x \sim \mathcal{N}(0, 1)$). As illustrated, conventional activations inherently suppress negative inputs, culminating in a severe density spike at the origin. This zero-state saturation not only induces catastrophic representational collapse (the "dying neuron" problem) but also artificially inflates the output variance ($\sigma \in [0.56, 0.59]$) through extreme distributional imbalance. The Lyapunov derived function (Equation 6) preserves bilateral phase geometry. By seamlessly extending into the negative domain, it strictly prevents zero-collapse, ensuring continuous gradient flow. Its considerably lower variance ($\sigma = 0.35$) is a direct mathematical manifestation of Lyapunov *Equilibrium Stability*. This function rigorously bounds the state-space trajectory by guaranteeing that latent representations neither collapse to absolute zero ($V(x) > 0$ for $x \neq 0$) nor exponentially explode ($\dot{V}(x) < 0$). *This theoretical constraint yields a tightly controlled, zero-centered distributional footprint, ensuring the uniform and mathematically stable signal utilization required to model the deterministic chaos of continuous EMG sequences.*

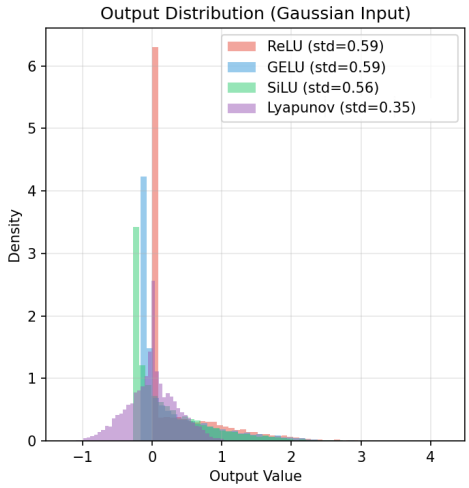

Figure 6: Output Density Distributions of ReLU, GELU, SiLU against Log-divergence Equation (Equation 6)

Computational Profiling: Table 11 quantifies the element-wise computational overhead of the exact log-divergence formulation (Equation 6) against standard nonlinear activations. Hardware-optimized functions such as ReLU (P_0), GELU (P_1), and SiLU (P_2) require minimal floating-point operations (1–22 FLOPs/element). However, direct evaluation of the nested logarithmic and fractional terms in the exact formulation (P_3) incurs 47 FLOPs per element. This severe latency bottleneck-demanding more than double the computational burden of SiLU motivates and mathematically justifies our prior derivation of the streamlined LyRIC polynomial approximation (Equation 6) which ensures robust chaotic dynamic modeling without compromising the high-throughput inference requirements of real-time ETS synthesis.

Table 11: Comparison of Computational Profiling among ReLU, GeLU, SiLU, and Log-divergence Equation (Equation 6)

	Function	FLOPs/element↓
P_0	ReLU	1
P_1	GeLU	14
P_2	SiLU	22
P_3	Log-divergence	47

A.3.3 Comprehensive Analysis of LyRIC Activation

Motivation: As quantified in Table 11, the direct evaluation of the nested logarithmic and fractional operators in Equation 6 imposes a prohibitive FLOPs penalty relative to hardware-optimized baselines (ReLU, GELU, SiLU). To resolve this computational bottleneck, we derive Equation 6 which is a highly efficient, lower-bound polynomial approximation formulated through truncated Maclaurin and geometric series expansions. This mathematically rigorous simplification neutralizes the algorithmic overhead while faithfully preserving the deterministic chaotic expansion properties and topological stability of the original continuous manifold. As mentioned in Table 1, we set α=1 for LyRIC to minimize computational cost (1 FLOP/element).

Topological Sensitivity Analysis: The heat map in Figure 7 visualizes the output magnitude across the input (x) and exponent (α) manifold. The surface exhibits odd symmetry (bilateral phase preservation) and high topological continuity across the entire input domain. The dashed yellow iso-line ($\alpha = 1.0$) identifies the region of optimal gradient stability, where the function achieves a balance between deterministic nonlinear expansion and controlled asymptotic attenuation. The absence of sharp discontinuities or saturation plateaus confirms the activation's suitability for preserving high-entropy, chaotic EMG signal microstructures. To evaluate the operational stability of the proposed LyRIC activation function, we perform a sensitivity analysis on its output surface relative to the expansion coefficient, $\alpha \in [0.6, 2.0]$. As illustrated in Figure 7, the activation function maintains a continuous, smooth gradient landscape across the entire input-exponent manifold. Unlike standard saturating activations that induce hard nonlinearities, the LyRIC surface exhibits gradual, predictable

transitions. This is critical for ETS synthesis, as it ensures that the model remains sensitive to the non-stationary fluctuations of articulatory EMG signals rather than suppressing them via premature saturation. The observed heat map color-coding — transitioning uniformly from dark blue (negative magnitude) to white (origin) to dark red (positive magnitude) — demonstrates the function's strict odd symmetry, ensuring that bidirectional neuromuscular activations are mapped with uniform topological fidelity. The iso-line at $\alpha = 1.0$ (highlighted) represents the selected operational point. This configuration acts as a stable equilibrium for the neural network, providing sufficient nonlinear expansion to model chaotic attractors while avoiding the representational instability associated with higher-order exponentiation (> 1.5). Therefore, the LyRIC activation provides the representational bandwidth to decode the rapid, aperiodic micro-fluctuations characteristic of deterministic chaotic dynamics in human motor control, while maintaining a minimal computational footprint of 1 FLOP/element.

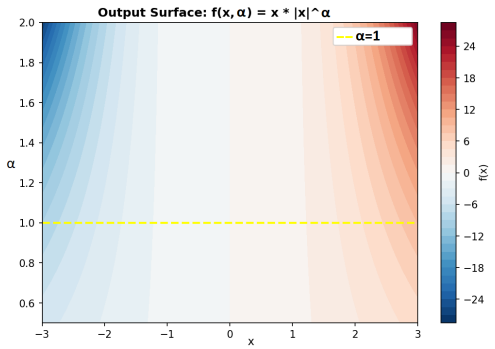

Figure 7: Topological Sensitivity Analysis of the LyRIC Activation Surface ($f(x, \alpha) = x|x|^{\alpha}$).

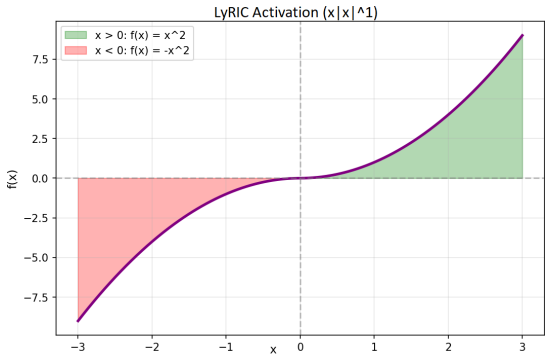

Figure 8: Geometric Representation of the LyRIC Activation Function at the Optimal Expansion Coefficient $\alpha = 1.0$.

Geometric Representation: The LyRIC Activation function can be reduced to a signed square topology ($f(x) = \text{sign}(x) \cdot x^2$). Figure 8 demonstrates continuous differentiability and bilateral phase preservation. As derived in Section 2.2 (Equation 6), our LyRIC formulation simplifies to a signed square topology when $\alpha = 1.0$, expressed as $f(x) = \text{sign}(x) \cdot x^2$. This topology provides a fundamental advantage over standard rectifying activations: it eliminates the zero-state saturation zone ($x < 0$). LyRIC maintains the full representational bandwidth of the EMG source signal by mapping negative signal fluctuations to negative outputs. This bilateral sensitivity ensures that the encoder remains explicitly responsive to the rapidly shifting phase transitions and aperiodic bursts that characterize the chaotic dynamics of neuromuscular motor unit recruitment.

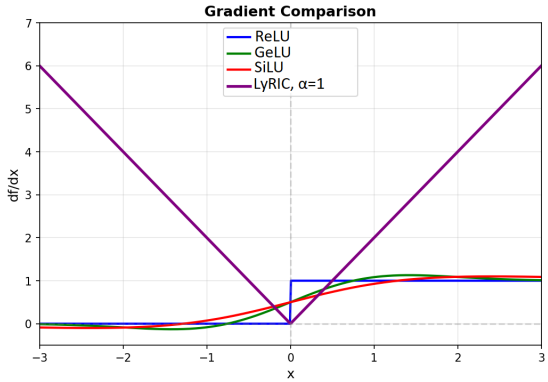

Figure 9: Gradient Landscape Analysis of ReLU, GeLU, SiLU, and LyRIC ($\alpha = 1$).

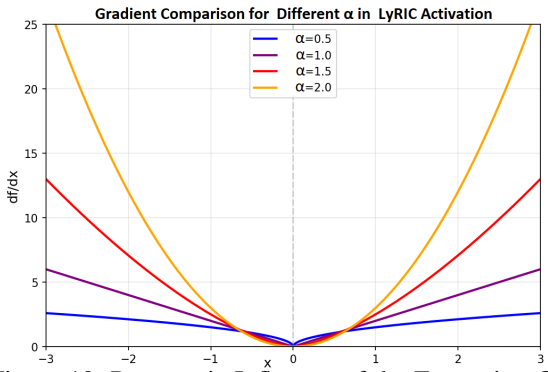

Figure 10: Parametric Influence of the Expansion Coefficient (α) on the Gradient Landscape of the LyRIC Activation.

Gradient Landscape Analysis: The *LyRIC* activation $f(x) = x|x|$ has gradient $f'(x) = 2|x|$, which scales proportionally with input magnitude, as shown in Figure 9. Unlike saturating activations (ReLU, GELU, SiLU) whose gradients converge to 1.0 or 0.0, LyRIC exhibits adaptive gradient expansion, enabling the network to propagate high-magnitude signals without saturation artifacts. This nonsaturating property is mathematically analogous to a positive Lyapunov exponent, ensuring the model retains sensitivity to the high-variance, aperiodic bursts characteristic of chaotic EMG articulatory dynamics throughout the GLAME. Figure 10 illustrates how α controls gradient gain

across the latent manifold to analyze the influence of expansion rate α on the gradient function $f'(x) = (1+\alpha)|x|^{\alpha}$. Lower values ($\alpha = 0.5$) produce a flat gradient profile, risking representational underfitting for subtle EMG fluctuations, while higher values ($\alpha > 1.0$) accelerate gradient magnitude which potentially induce optimization instability. The configuration α=1.0 provides a balanced, linear expansion rate that preserves the high-entropy microstructure of the EMG signal while maintaining stable gradient flow throughout the GLAME.

A.4 Theoretical Analysis of the LER Loss

Deterministic Chaos theory (Strogatz, 1994; Lorenz, 1963) of Lyapunov in nonlinear dynamical systems motivates the formulation of the **LER loss**. It describes complex, non-periodic behavior in nonlinear dynamical systems that are highly sensitive to initial conditions. While such systems are governed by deterministic laws, their long-term evolution is inherently unpredictable due to the exponential amplification of infinitesimal perturbations.

Deterministic Chaos: In Deterministic Chaos theory, the fundamental metric for quantifying the chaotic behavior is the *Lyapunov Exponent* (λ), which measures the mean exponential rate $|\delta x(t)|$ of trajectory divergence in state-space. The spectrum of λ classifies the system dynamics: $\lambda < 0$ denotes a stable, over-damped system converging to a fixed point; $\lambda = 0$ corresponds to marginally stable limit cycles or periodic oscillators; and $\lambda > 0$ serves as the definitive indicator of deterministic chaos. In such chaotic regimes, microscopic variations in the initial state map to macroscopic, divergent variations in the system's output– *a hallmark of human phonation and neuromuscular motor control*.

Deterministic Chaos for Designing LER Loss: Conventional ETS systems optimize for frame-level reconstruction losses (e.g., L1 or MSE loss), which mathematically enforces a linear, stable convergence ($\lambda \leq 0$). This objective forces the network to eliminate natural jitter, aperiodic bursts, and rapid state transitions, resulting in unnaturally smoothed acoustic prosody. To counter this, we introduce the *Lyapunov Exponent Regularization (LER) loss*, designed to preserve the chaotic manifold of the source signals. As formalized in Algorithm 1, LER operates by reconstructing the latent phase-space trajectories of both the predicted EMG synthesized features and the ground-truth targets. We utilize the Rosenstein-based divergence estimation (Rosenstein et al., 1993), which computes the average logarithmic rate of separation between neighboring trajectories. By enforcing a penalty on the distance between the predicted ($\lambda_{\hat{Y}}$) and ground-truth (λ_Y) Lyapunov exponents, we constrain the encoder to replicate the *divergence characteristics* of the source neuromuscular signals. In the context of our EMG systems, this physically grounded supervision ensures the model does not collapse into a stable equilibrium. Instead, it compels the network to represent the deterministic chaos inherent in the electrophysiological source– particularly the sub-harmonic firing patterns, bifurcations, and timing irregularities– which are essential for synthesizing human-like, temporally coherent speech. Therefore, the LER objective serves as a topological constraint, ensuring that the synthesized acoustic features faithfully reflect the complex, non-linear micro-structure of the underlying EMG articulatory gestures.

A.5 Dataset Preprocessing Pipeline

We utilize the established state-of-the-art (SOTA) 2020 ETS dataset (Gaddy and Klein, 2021) to benchmark our proposed CGS-ETS architecture, which comprises parallel (silent and vocalized) and non-parallel vocalized articulatory recordings. Prior to experimentation, we rigorously verified that the corpus is devoid of Personally Identifiable Information (PII) and any abusive or harmful content, adhering to ethical standards for bio-signal research. The corpus is completely partitioned into training, validation, and test splits following the standardized JSON indices provided in the benchmark.

EMG Signal Processing: Each utterance is contextually loaded by concatenating it with its temporally adjacent neighbors ($i-1, i, i+1$) to precisely mitigate transient boundary artifacts during digital filtering which isolates the target sequence only after continuous filtration. We suppress ambient electromagnetic interference by applying a rigorous comb notch filter at 60 Hz and its subsequent seven harmonics (up to 420 Hz). A 3rd-order Butterworth high-pass filter (2 Hz cutoff) is subsequently applied to eliminate bio-electrical baseline drift. The continuous signals are then subsampled to synchronize with the acoustic frame rate (≈689 Hz). Following filtration, we extract a dense, multi-domain feature representation utilizing a sliding window (16-sample length, 6-sample hop). This results in localized time-domain statistics (mean amplitudes, RMS envelopes, zero-crossing rates of the high-frequency residual) alongside frequency-domain STFT magnitudes. The concatenated EMG features are standardized through global mean-variance normalization and tightly bounded using a scaled hyperbolic tangent projection ($8\tanh(x/8)$) to guarantee strict topological stability during gradient propagation.

Acoustic and Linguistic Processing: Ground-truth acoustic waveforms are dynamically loaded, collapsed from stereo to mono configurations to isolate the primary channel, and rigorously resampled to 22.05 kHz. We extract 80-dimensional mel-spectrograms to formulate the primary regression targets utilizing a 1024-point STFT, a 1024-sample Hanning window, and a 256-sample hop size, bounding the spectral representation between 0 and 8,000 Hz. Concurrently, to facilitate our auxiliary linguistic optimization, we parse forced-alignment `TextGrid` annotations to extract frame-level phonetic labels across a 48-token inventory (including silence). Finally, raw text transcripts are stripped of punctuation, lowercased, and normalized through the `JiWER` pipeline to ensure pristine linguistic mapping during multi-task supervision.

A.6 Dataset Class and DataLoader

The aforementioned signal processing and feature extraction pipelines are formalized within a custom PyTorch `Dataset` class (`EMGDataset`). The global utterance indices are deterministically shuffled utilizing a fixed random seed upon initialization to guarantee absolute experimental reproducibility. The total dataset capacity (`__len__`) reflects the number of valid sequences mapped within the predefined JSON splits. During runtime, the `__getitem__` method dynamically retrieves the multi-modal representations which heavily leverages Least Recently Used (LRU) caching to minimize disk I/O latency. Within this method, the raw and structured EMG features are normalized and topologically bounded through scaled hyperbolic tangent projections. For silent articulation samples, the method automatically retrieves and aligns the corresponding parallel vocalized acoustic targets. To maximize PCIe data transfer rates to the GPU, all output tensors are explicitly allocated to page-locked memory (`pin_memory()`). During the optimization phase, we instantiate a `DataLoader` governed by a custom `SizeAwareSampler`. This sophisticated sampler dynamically aggregates sequences based on a cumulative temporal frame limit (`max_len = 256,000`) rather than utilizing naive, fixed instance-count batching. This prevents GPU Out-Of-Memory (OOM) faults across highly variable continuous speech sequences while maximizing parallel processing throughput. Finally, a custom collation function (`collate_raw`) forcibly pads these variable-length multi-modal tensors and seamlessly routes parallel acoustic targets when processing silent EMG inputs.

A.7 Details on Computational Complexity

Table 3 contrasts the inference-time computational complexity of our proposed CGS-ETS architecture against the state-of-the-art baseline (P_0) (Gaddy and Klein, 2021).

Model Size (Parameters): The baseline architecture requires a massive 54.10M parameters. In stark contrast, CGS-ETS strictly compresses the network to just 14.34M parameters. This extraordinary **73.49% reduction** in parameter footprint is achieved by aggressively pruning the CFE block (10.07M $\rightarrow$ 4.49M) and deploying our highly optimized GLAME (44.03M $\rightarrow$ 9.85M). Consequently, CGS-ETS demands considerably less memory bandwidth, making it exceptionally suitable for deployment on limited-resource, edge-device SSIs. A comprehensive layerwise parameter breakdown of CGS-ETS is shown in Table 13.

Computational Cost (FLOPs): The baseline framework demands 2.40G FLOPs per inference pass to resolve the temporal mapping. CGS-ETS requires only **926.55M FLOPs** by replacing brute-force multi-head self-attention with our streamlined, Grouped Query-driven GLARE mechanism and Mamba state-space core. This 61.39% reduction in absolute computational operations mitigates thermal throttling and power consumption constraints on portable hardware. The complexity is bifurcated into two primary operational domains:

1. **CGS-ETS without Activation Functions:** These operations constitute the overwhelming majority of the compute budget, accounting for **924.81M FLOPs** (99.81% of the total). This includes the multi-stage ResBlock feature extraction (CFE Block) and GLAME (the Mamba state-space models, the GLARE attention mechanisms, FFNs, and the terminal output heads).

2. **Activation Overhead:** The non-linear mapping modules utilize a substantially smaller fraction of the computational budget, contributing **1.74M FLOPs** (0.19% of the total). This breakdown includes the following:

   - *ReLU* (6 layers): 0.18M FLOPs.
   - *LyRIC* (8 layers): 0.13M FLOPs.
   - *GELU* (4 layers): 1.43M FLOPs.

As summarized above, the model achieves a total computational footprint of 926.55M (924.81M+1.74M) FLOPs. This analysis confirms that the non-linear activation overhead is negligible which accounts for less than 0.2% of the total FLOPs. Furthermore, the efficiency of our LyRIC activation which operates at a strict lower bound of 1 FLOP per element– ensures that incorporating chaotic dynamic modeling does not impose a prohibitive computational penalty, maintaining the architecture's high-throughput viability for edge-based SSIs.

Real-Time Factor (RTF) and Inference Time: Despite the massive reduction in FLOPs, both architectures maintain an identical, highly competitive RTF of 0.0057, indicating that the baseline's latency is bounded by framework-level I/O constraints rather than pure computation. However, CGS-ETS consistently achieves a faster absolute Inference Time (IT) of 5.61 ms per batch (compared to the baseline's 5.97 ms). Hence, by supervising the GLAME with chaos-inspired topologies during training, CGS-ETS achieves unprecedented performance gains across all acoustic and linguistic metrics while operating at a fraction of the baseline's computational cost.

A.8 Hyperparameters and Configuration

The training and model hyperparameters for the CGS-ETS setup, with use cases and rationale are provided thoroughly in Table 12.

All experiments, architectural modifications, and ablation studies are executed within an isolated Anaconda virtual environment running Python 3.9 to guarantee computational reproducibility and optimal hardware utilization. The core deep learning framework, including the GLAME is implemented utilizing PyTorch ($\geq$2.0.0), integrated with Torchvision and Torchaudio, and hardware-accelerated through the NVIDIA CUDA Toolkit 11.8. For high-fidelity digital signal processing, dynamic temporal filtering, and mel-spectrogram feature extraction, we leverage a standardized suite of scientific libraries, including Librosa ($\geq$0.9.0), SciPy ($\geq$1.7.0), and NumPy ($\geq$1.21.0), while Soundfile ($\geq$0.12.0) manages all rapid waveform I/O operations. Furthermore, to ensure absolute fairness and standardization in our linguistic evaluation pipeline, phonetic forced-alignment parsing is handled by Praat-Textgrids ($\geq$1.4.0), text transcript normalization is processed via Unidecode ($\geq$1.3.0), and all objective WER computations are derived utilizing `JiWER` (v2.2.1) alongside the pre-trained Mozilla DeepSpeech engine (v0.9.3).

All models are trained in FP32 precision on two NVIDIA RTX 4090 GPUs (24 GB VRAM each), hosted on a workstation equipped with an AMD Ryzen 9 7950X3D CPU, 192 GB system RAM, and 10 TB of storage. Training is parallelized across both GPUs via PyTorch Distributed Data Parallel (DDP), using the NCCL backend for high-throughput inter-GPU communication and TCP initialization for process group rendezvous. This configuration supports both the current dual-GPU training regime and straightforward scaling to multi-node environments without architectural modification.

A.9 Inference and Evaluation Workflow

We initialize an inference pipeline that loads the optimized CGS-ETS model weights alongside the fine-tuned HiFi-GAN vocoder to rigorously benchmark the acoustic and linguistic fidelity of our architecture. The script iterates over the designated evaluation split (validation or test) of the EMG corpus. For each continuous sequence, the inference engine executes the following highly optimized workflow:

1. **Signal Retrieval:** Dynamically loads and preprocesses the 8-channel articulatory EMG signals, extracting both the raw physiological time-series and the structured, normalized sliding-window features.
2. **Latent Projection:** Propagates the localized EMG representations through the GLAME to simultaneously predict 80-dimensional mel-spectrogram matrices and auxiliary phoneme probability distributions.
3. **Feature-Level Synchronization:** Computes a pairwise cosine distance matrix between the predicted spectrograms and the ground-truth targets, utilizing Dynamic Time Warping (DTW) to strictly align the temporal resolution of the generated latent features for loss optimization.
4. **Waveform Inversion:** Inverts the aligned mel-spectrogram predictions into high-fidelity, time-domain acoustic waveforms utilizing the pre-trained HiFi-GAN vocoder.
5. **Post-Vocoder Alignment (PVA):** Extracts STFT magnitude spectrograms from both the synthesized and target waveforms, computes frame-wise L_2-normalized cosine distances, and applies a secondary continuous DTW mapping. The generated audio chunks are subsequently warped to guarantee absolute frame-level temporal synchronization with the target signal.
6. **Metric Exportation:** Saves the precisely synchronized, 22.05 kHz audio pairs into structured directories which enables the immediate computation of comprehensive frame-level acoustic metrics (LSD, STOI, PESQ) alongside legacy Automatic Speech Recognition (ASR) evaluations for Word Error Rate (WER). For the computation of PESQ and NISQA-MOS, waveforms are downsampled to 16 kHz to comply with PESQ's fixed wideband/narrowband frequency constraints and NISQA-MOS's strict input sample-rate requirements.

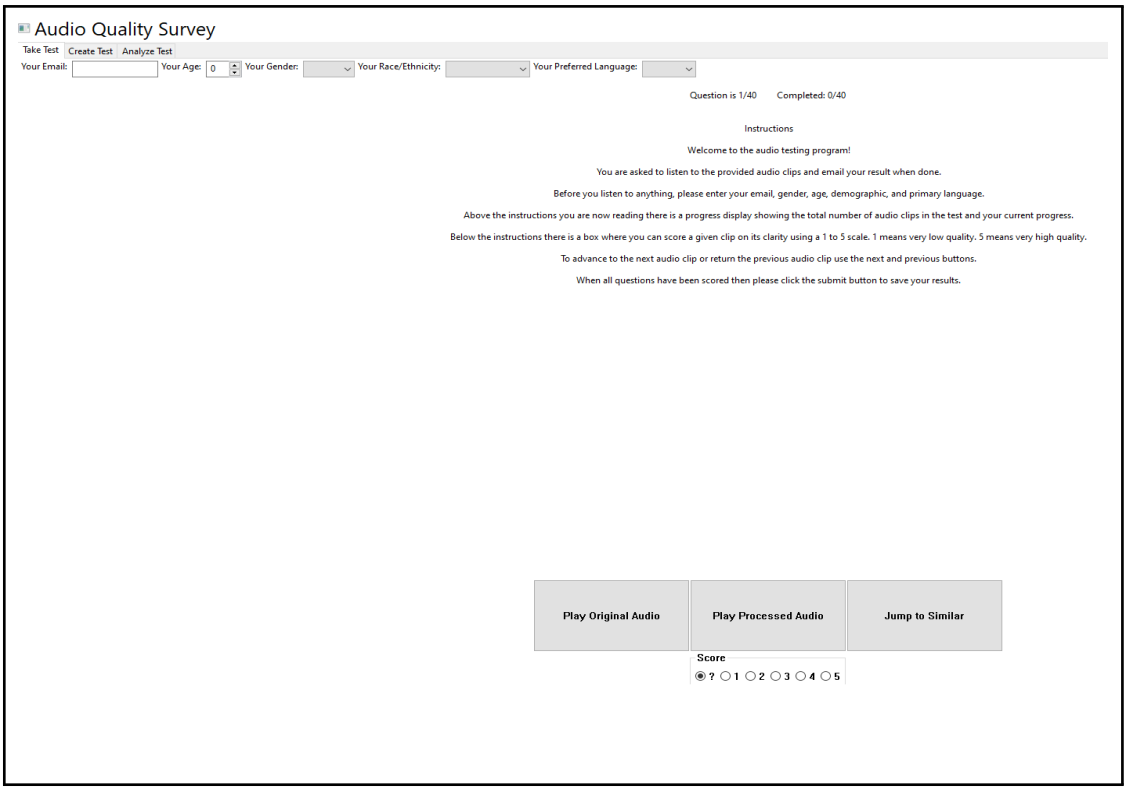

Figure 11: The Software Interface Used in Subjective Tests.

A.10 Subjective Evaluation Details

We conducted a formal subjective listening test to evaluate the perceptual fidelity of the CGS-ETS architecture. The study cohort comprised ten normal-hearing (NH) English-speaking volunteers, balanced for gender (four males, six females) with a mean age of 24. Participants were briefed on the experimental objectives and the underlying research goals. All subjects were trained on a standardized Mean Opinion Score (MOS) rubric—ranging from 1 (lowest quality) to 5 (highest quality) to ensure optimal perceptual accuracy– prior to the evaluation. All testing sessions were conducted using high-fidelity, sound-isolated headsets to eliminate extrinsic acoustic interference. Each participant evaluated 30 distinct utterances, each comprising three randomized versions: **(i)** the baseline reconstruction (Gaddy and Klein, 2021), **(ii)** the output synthesized by our proposed CGS-ETS, and **(iii)** a reference EMG data for each corresponding speech signal. Participants were tasked with assigning a 5-point MOS rating to each sample and selecting their most preferred version among the synthesized candidates. The test set utilized clips derived from the Gaddy benchmark which ensures diverse phonetic and articulatory coverage. The evaluation interface, illustrated in Figure 11, provided a controlled environment for these subjective comparisons. As demonstrated in Figure 12, the speech synthesized by CGS-ETS was consistently and considerably preferred over the established state-of-the-art (SOTA) baseline. In particular, CGS-ETS achieved a **5.52% improvement** in subjective MOS scores. These results empirically validate that our chaos-inspired inductive bias effectively mitigates the unnatural spectral over-smoothing characteristic of prior ETS systems which yields highly intelligible and perceptually natural speech even when decoding from non-stationary, low-SNR articulatory bio-signals.

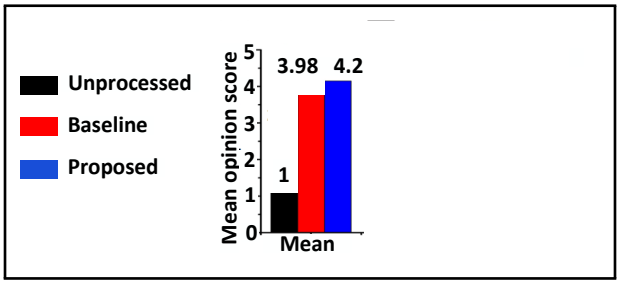

Figure 12: The MOS score.

Hyperparameter	Value	Use Case & Rationale
Optimization & Training Dynamics		
Max epochs	80	Provide sufficient gradient updates for convergence while preventing catastrophic representational over-fitting.
Batch size	32	Mathematically balance gradient variance stabilization with strict GPU VRAM (OOM) constraints.
Initial learning rate (η)	1×10^{-3}	Ensure robust initial gradient descents across the highly non-convex objective manifold.
Learning rate warmup	500 steps	Strictly prevent early gradient spikes and stabilize initial optimization dynamics.
Weight decay	1×10^{-7}	Subtly regularize network weights to prevent over-fitting on specific real-world EMG physiological artifacts.
Gradient clip norm	0.5	Aggressively bound gradient explosion to maintain topological stability during State Space Model (SSM) optimization.
Random seed	0	Guarantee absolute experimental reproducibility across dataset shuffling and network initialization.
GLAME Architecture & Complexity		
Hidden dimension (D)	512	Provide sufficient parameter bandwidth to map complex 8-channel myoelectric signals into acoustic topologies.
GLAME layers (L_c)	4	Establish an optimal equilibrium between deep hierarchical sequence modeling and strict architectural compression.
GLARE window size	32	Constrain the localized attention receptive field to precisely capture rapid, short-term coarticulatory transitions.
SSM state dimension (d_{state})	8	Compress global temporal dependencies efficiently within the Mamba parallel scan core.
SSM expansion factor	1.25	Optimize the internal state-space expansion to capture deterministic features while mitigating computational bloat.
LyRIC exponent (α)	1.0	Balance non-linear chaotic phase-space expansion with optimal hardware-level efficiency (1 FLOP/element).
Dropout probability	0.2	Systematically mitigate feature co-adaptation across the deep feed-forward and attention hierarchies.
Signal Processing & Acoustic Targets		
EMG sampling rate	689.06 Hz	Subsample the filtered physiological sequences to mathematically synchronize with the acoustic frame rate.
Audio sampling rate	22.05 kHz	Establish the high-fidelity wide-band frequency target necessary for HiFi-GAN waveform inversion.
Mel-spectrogram bins	80	Define the dense, perceptually relevant acoustic projection space for the primary regression head.
STFT size (`n_fft`)	1024	Balance frequency resolution against computational throughput for continuous acoustic feature extraction.
STFT hop length	256	Establish temporal overlap resolution to perfectly align predicted acoustic frames with localized EMG segments.
Chaos Configuration Settings		
DTW loss weight	1.0	Enforce strict primary supervision for the macroscopic temporal synchronization of generated spectrograms.
Phoneme loss weight	0.5	Scale the auxiliary linguistic objective to explicitly enforce structural articulatory constraints.
LER embedding dim (m)	10	Construct the multi-dimensional phase-space trajectory required for precise deterministic chaos analysis.
LER time delay (τ)	1	Establish the minimal temporal offset necessary for accurate state-space divergence computation.
MSDFA temporal scales	4, 8, 16, 32	Quantify fractal-like, long-range chaotic correlations across exponentially varying physiological resolutions.
Chaos loss start epoch	10	Allow the network to stabilize its primary acoustic mapping before enforcing complex chaotic topological constraints.

Table 12: Comprehensive Summary of Critical Hyperparameters, Architectural Configurations, and Their Theoretical Justifications within the Proposed CGS-ETS Framework.

Module	Stage	Layer Type	Input→Output	Kernel	Stride	Padding	Params
CFE Block	Block 1	Conv1d (Main)	8→512	3	2	1	12 800
		BatchNorm1d + ReLU	512→512	–	–	–	1 024
		Conv1d (Main)	512→512	3	1	1	786 944
		BatchNorm1d	512→512	–	–	–	1 024
		Conv1d (Residual)	8→512	1	2	0	4 608
		BatchNorm1d + ReLU	512→512	–	–	–	1 024
	Block 2	Conv1d (Main)	512→512	3	2	1	786 944
		BatchNorm1d + ReLU	512→512	–	–	–	1 024
		Conv1d (Main)	512→512	3	1	1	786 944
		BatchNorm1d	512→512	–	–	–	1 024
		Conv1d (Residual)	512→512	1	2	0	262 656
		BatchNorm1d + ReLU	512→512	–	–	–	1 024
	Block 3	Conv1d (Main)	512→512	3	2	1	786 944
		BatchNorm1d + ReLU	512→512	–	–	–	1 024
		Conv1d (Main)	512→512	3	1	1	786 944
		BatchNorm1d	512→512	–	–	–	1 024
		Conv1d (Residual)	512→512	1	2	0	262 656
		BatchNorm1d + ReLU	512→512	–	–	–	1 024
						Total Parameters in CFE Block	**4 486 656**
GLAME Layer (w/ GLARE)	Mamba	Linear (`in_proj`)	512→1280	–	–	–	655 360
		Depthwise Conv1d + LyRIC	640→640	4	1	3	3 200
		Linear (`x_proj`)	640→25	–	–	–	16 000
		SSM Params (`A_log`, `D`)	–	–	–	–	652
		Gating + LyRIC	640→640	–	–	–	0
		Linear (`out_proj`, no bias)	640→512	–	–	–	327 680
		LayerNorm (`norm1`)	512→512	–	–	–	1 024
	GLARE	Linear (`q_proj`)	512→512	–	–	–	262 144
		Linear (`k_proj`)	512→256	–	–	–	131 072
		Linear (`v_proj`)	512→256	–	–	–	131 072
		Relative Positional Bias	–	–	–	–	504
		Linear (`out_proj`)	512→512	–	–	–	262 144
		LayerNorm (`norm2`)	512→512	–	–	–	1 024
	FFN	Linear (no bias) + GELU	512→1024	–	–	–	524 288
		Linear (no bias)	1024→512	–	–	–	524 288
		LayerNorm (`norm3`)	512→512	–	–	–	1 024
						Layer w/ GLARE total	2 841 476
						Layers 1 & 3	**(2× 2 841 476) 5 682 952**
GLAME Layer (w/o GLARE)	Mamba	Linear (`in_proj`, no bias)	512→1280	–	–	–	655 360
		Depthwise Conv1d + LyRIC	640→640	4	1	3	3 200
		Linear (`x_proj`, no bias)	640→25	–	–	–	16 000
		SSM Params (`A_log`, `D`)	–	–	–	–	652
		Gating + LyRIC	640→640	–	–	–	0
		Linear (`out_proj`, no bias)	640→512	–	–	–	327 680
		LayerNorm (`norm1`)	512→512	–	–	–	1 024
	FFN	Linear (no bias) + GELU	512→1024	–	–	–	524 288
		Linear (no bias)	1024→512	–	–	–	524 288
		LayerNorm (`norm3`)	512→512	–	–	–	1 024
						Layer w/o GLARE total	2 053 516
						Layers 2 & 4	**(2× 2 053 516) 4 107 032**
GLAME Terminal Output	Acoustic	Linear (Primary Head, `w_out`)	512→80	–	–	–	41 040
	Linguistic	Linear (Secondary Head, `w_aux`)	512→40	–	–	–	20 520
						Total Terminal Output	**61 560**
			GLAME Summary				
						GLAME Layers 1 & 3 (2× w/ GLARE)	5 682 952
						GLAME Layers 2 & 4 (2× w/o GLARE)	4 107 032
						Terminal Output	61 560
						Grand Total of GLAME	**9 851 544**
			CGS-ETS Architecture Summary				
						CFE Block	4 486 656
						GLAME	9 851 544
						Grand Total of CGS-ETS Architecture (CFE+GLAME)	**14 338 200**

Table 13: Comprehensive Layer-wise Parameter Breakdown of the Proposed CGS-ETS Architecture. The summary block demonstrates the strict architectural compression achieved by our GLAME and its selective application of GLARE attention modules.