

Continuous-Time Acoustic Modelling with Neural Controlled Differential Equations

Mattias Cross
Speech and Hearing Group
University of Sheffield
Sheffield, United Kingdom
mcross2@sheffield.ac.uk

Minghui Zhao
Speech and Hearing Group
University of Sheffield
Sheffield, United Kingdom

Anton Ragni
Speech and Hearing Group
University of Sheffield
Sheffield, United Kingdom

Abstract—Text-to-speech (TTS) models commonly address text–speech alignment by expanding phone-level encoder states to frame-level decoder inputs using predicted durations. While this length-regulation step resolves alignment structurally, this use of duration typically changes only where and how often latent states appear, not the values of the states themselves. This paper proposes a continuous-time mechanism for duration-aware acoustic modelling in TTS using neural controlled differential equations (CDEs). We formulate the phone representation as a temporally parameterised control path and use a neural acoustic vector field to produce a continuous-time hidden state whose values evolve with phonetic content and duration-derived timing. The resulting trajectory can be sampled at discrete points and integrated into a standard acoustic decoder pipeline. Objective results contrast CDEs and typical recurrent models. Subjective results suggest that CDE-based models evaluating one phone per step can improve rank-order agreement between synthesised and reference emotion intensity while maintaining comparable emotion-expression quality to a strong baseline. Additional experiments with half-phone step-sizes suggest that temporal resolution changes the trade-off between style tracking and absolute calibration. These results position CDEs as a promising design space for continuous-time and duration-aware style-sensitive TTS.

Index Terms—Text-to-speech, neural CDE, emotional speech

I. INTRODUCTION

Text-to-speech (TTS) is a core component of many speech-centred technologies, including augmentative and alternative communication (AAC), personalised synthetic voices, conversational agents, and data augmentation pipelines for speech and language systems [1], [2]. Synthesised speech must be intelligible, natural, temporally well structured, and adaptable to different speakers, styles, and interaction requirements.

A central problem in TTS is alignment, one text unit represents many acoustic units, and this alignment is *irregular* in the sense that all phones in an utterance do not have equal duration. Further, the duration and timing of each phone can vary widely across different contexts and speakers. A common method to address this is the encoder–decoder architecture, which has seen many formulations [3]–[7]. In

This work was supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

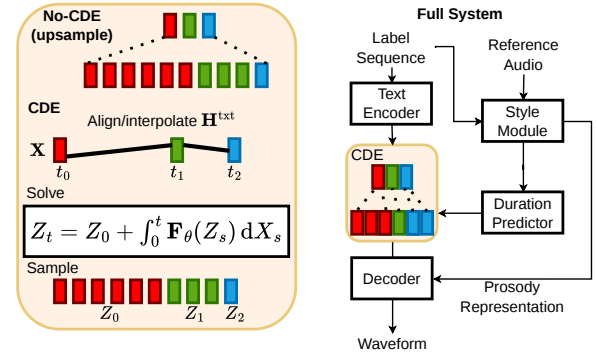


Fig. 1. Proposed CDE TTS model. Without a CDE, H^{txt} is simply aligned by repeating as per the predicted durations, without these durations affecting the values of the decoder input.

this framework, an encoder maps a sequence of linguistic labels to a latent representation. This representation is then upsampled to match the expected temporal resolution of the acoustic target by repeating each encoder frame to its expected duration. The upsampled representation is then processed by a decoder to predict raw waveforms. This formulation resolves alignment structurally, but it makes an imperfect assumption: duration changes where and how often latent states appear in the decoder input, but not the values of the latent states themselves. For example, a lengthened vowel is not simply a short vowel repeated for more frames: its pitch, energy, spectral tilt, and coarticulatory trajectory may evolve across the segment depending on emphasis, phrase position, emotion, and speaking rate. Duration therefore controls not only how long a phone is realised, but also how its acoustic trajectory unfolds. Duration has a key control on how a phone is uttered, and the methods discussed above force the decoder to model this important relationship through frame-wise context and positional cues, rather than introducing duration information into the encoding process. A more aligned approach would be to treat TTS as a transformation between a sequence of phone labels observed at irregular time intervals, and a sequence of regularly-sampled acoustic frames. This form of problem is a fundamental topic in the theory of controlled differential

equations (CDEs), and is the focus of this paper. Neural CDEs model a dynamically evolving hidden state driven by a control signal, such as a sequence of phone embeddings. The concept of a vector-field being *driven* is associated with rough-path theory [8], where changes in a control signal drive/affect a vector-field. Although they have previously been applied to speech data for classification tasks [9], [10], their applicability for TTS remains unknown. To our knowledge, this work is the first to formulate TTS as a neural acoustic vector-field driven by a continuous-time phonetic latent representation. This formulation exposes temporal resolution as a modelling choice: the phonetic path can be evaluated not only at phone boundaries, but also at intermediate points along the trajectory. This allows duration to influence the values of the latent representation, rather than only determining the number of repeated frames. This method should not be confused with continuous-depth models such as Grad- and Matcha-TTS [6], [7], which introduce continuous dynamics in the generation process (e.g., diffusion/flow depth) rather than modelling a continuously parameterised phonetic trajectory that explicitly mediates text–time alignment. To explore the properties of CDEs in TTS, we use a neural CDE for emotion-controlled TTS (Figure 1). This involves predicting duration and other prosodic features from a reference speech sample, then synthesising an utterance in the same style. Results from subjective evaluation show that a CDE processing one phone per step has a significant Spearman rank correlation of emotion intensity between the synthesised speech and the reference style, whilst maintaining the same quality of emotion expression. The properties of continuous-time modelling are further demonstrated by processing a phone-label sequence at a half-phone step-size, a degree of freedom not commonly supported in TTS. This setting offers a less correlated yet more accurate emotion intensity, suggesting that different time resolutions offer different dynamics. These results suggest that continuous-time models offer a useful mechanism for style-sensitive TTS, and that the temporal resolution of the CDE changes the dynamics of style control. Broadly, this work introduces CDEs as a continuous-time modelling framework for TTS, opening a wider design space in which control paths, interpolation schemes, solver choices, sampling resolutions, and stacked dynamics remain to be explored. The remainder of this paper is organised as follows: Section II describes the proposed CDE-based encoder–decoder method. Section III presents the experimental setup. Section IV reports the results and discusses the findings. Section V concludes the paper. Code is publicly available.¹

II. METHOD

A. Encoder–Decoder Architecture

Given a phone-label sequence $\mathbf{w} \in \mathbb{Z}^L$ of length L , the text encoder $\mathbf{E}_\theta$ maps it to a C -dimensional phone-level representation,

$$\mathbf{H}^{\text{txt}} = \mathbf{E}_\theta(\mathbf{w}) \in \mathbb{R}^{C \times L}. \quad (1)$$

The latent representation is then upsampled to the length of the reference number of acoustic frames using a product of an alignment matrix $\mathbf{H}^{\text{aco}} = \mathbf{H}^{\text{txt}} \mathbf{A} \in \mathbb{R}^{C \times T}$, where T is the target number of acoustic frames and the alignment matrix $\mathbf{A} \in \{0, 1\}^{L \times T}$ is obtained from methods such as Montreal Forced Aligner [11], Monotonic Alignment Search [3], and automatic speech recognition (ASR) models supporting alignment [12]. This operation is equivalent to repeating each phone embedding according to its predicted or ground-truth duration. A duration predictor is trained to provide alignments at test time, the acoustic features $\mathbf{H}^{\text{aco}}$ are then used by a decoder or decoder + vocoder to produce waveforms. A duration predictor can either use label feature input, or style features as presented in the next subsection.

B. Style Embeddings

Encoder-decoder architectures can be enhanced with style embeddings, which allow controllable synthesis guided by the style of a short reference clip [13], [14]. This involves producing a prosodic text representation $\mathbf{H}^{\text{pros}} \in \mathbb{R}^{C \times L}$, using approaches such as a phone-level BERT model adopted by StyleTTS 2 [15]. The prosodic representation is then used to condition a diffusion model to predict a style embedding s^{aco} conditioned on the reference audio sample. The style and prosodic features are then used to predict pitch, energy, and duration as well as providing additional information to the decoder. Although this method captures style successfully, it introduces a disjoint pipeline where the acoustic embedding $\mathbf{H}^{\text{aco}}$ is only affected by duration through the length regulated upsampling, whilst the value of the acoustic embedding remains invariant to duration in every other aspect.

C. Continuous-Time Modelling with CDEs

Although length regulation provides a form of frame-level alignment, it represents duration only through repetition: the repeated encoder states remain unchanged across frames and therefore do not explicitly encode position-dependent temporal evolution within the alignment. In particular, two frames may have similar values while originating from phones with different durations or different positions within a phone segment. This weakens the decoder’s access to duration-dependent information such as local speaking rate, phone-boundary proximity, and temporal progression within the utterance.

To make this information available to the decoder, we augment the phone embedding with a time channel derived from the alignment $\mathbf{A}$. Let $\mathbf{t} = (t_1, \dots, t_L)$ denote the frame timestamps induced by the cumulative phone durations, and define the augmented control sequence

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{H}_i^{\text{txt}} \\ t_i \end{bmatrix} \in \mathbb{R}^{C+1}, \quad i = 1, \dots, L. \quad (2)$$

A continuous path $\mathbf{X}$ is then obtained by interpolating the discrete sequence $\{\mathbf{X}_i\}_{i=1}^L$. Typical interpolations are cubic and linear, this work uses linear and leaves other schemes for future work. Since CDEs are sensitive to change in time, and not absolute time [16], the interpolation knots $\mathbf{X}_i$ can be

¹https://github.com/Mattias421/CDE_StyleTTS.git

placed at regular computational intervals while preserving the temporal information through the values of the time channel itself.

The decoder state is parameterised as the solution of a CDE,

$$Z_t = Z_0 + \int_0^t \mathbf{F}_\theta(Z_s) d\mathbf{X}_s, \quad (3)$$

where $\mathbf{Z}$ is the hidden state, and $\mathbf{F}_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times C+1}$ is a neural vector field and Z_0 is predicted with an initial value network on the observed control sequence $\mathbf{X}$. In this formulation, the hidden trajectory evolves in response to changes in both the conditioning sequence $\mathbf{H}^{\text{txt}}$ and the duration-derived time channel. The CDE therefore provides a continuous-time mechanism for integrating duration information into the decoder, rather than relying only on the repeated or interpolated embeddings produced by length regulation. Note that the formulation of equation 3 is a Riemann–Stieltjes integral, which is incompatible with black-box ODE solvers, and must be re-written as

$$Z_t = Z_0 + \int_0^t \mathbf{F}_\theta(Z_s) \frac{d\mathbf{X}}{ds}(s) ds, \quad (4)$$

which is a matrix-vector product between the vector-field $\mathbf{F}_\theta(Z_s)$ and control derivative $\frac{d\mathbf{X}}{ds}(s)$. For an example of how an ODE solver can be used to sample from the continuous vector field, consider Euler’s method

$$Z_{i+1} = Z_i + \mathbf{F}_\theta(Z_i) dX_i h, \quad (5)$$

where h is the selected step size. The formulation of CDEs and their associated ODE solvers is flexible, with several valid choices of control representation and numerical integration scheme. The experiments presented later in this paper use fixed step-size solvers. A step-size $h = 1$ corresponds to processing one phone per solver step with a control derivative equal to the difference in the next control frame and the current $dX_i h = X_{i+1} - X_i$. Smaller values of h increase the temporal resolution of the CDE evaluation at the cost of increasing the total number of function evaluations required to solve the CDE. For example, a step size of $h = 0.5$ introduces an intermediate solver evaluation within each unit phone interval. Since timing information is incorporated into the control path $\mathbf{X}$, the vector field $\mathbf{F}_\theta$ can respond to irregular changes in time, even when it is evaluated in a regular manner.

The resulting hidden state $\mathbf{Z}$ can either be sampled at the frame-level directly or at the phone-level then aligned as $\mathbf{H}^{\text{aco}} = \mathbf{Z}\mathbf{A}$, compatible with the decoder during both training and synthesis. Backpropagation through neural CDEs is commonly performed using the adjoint state method [17], which can reduce memory requirements by avoiding the storage of the full forward trajectory $\mathbf{Z}$, making continuous-time modules attractive relative to recurrent and attention-based alternatives for long sequences due to $O(\frac{1}{h}L + d(C+1))$ memory complexity. Further details on training neural differential equations are provided in [17] and [16].

Beyond these training-time memory considerations, neural differential equations offer a unique design space. In particular,

each CDE layer produces a continuous hidden trajectory rather than a single static representation. This trajectory can itself be treated as a control signal for a subsequent CDE layer, enabling continuous conditioning modules to be stacked efficiently. The next section describes how this can be achieved.

D. Efficient Stacking of CDE Layers

The CDE formulation permits multiple continuous-time conditioning layers to be stacked without explicitly materialising and re-interpolating intermediate trajectories [16], [18]. For example, consider two CDE layers in which the hidden path of the first layer, $\mathbf{Z}_t$, acts as the control for the second layer:

$$d\mathbf{U}_t = \mathbf{G}_\phi(\mathbf{U}_t) d\mathbf{Z}_t, \quad d\mathbf{Z}_t = \mathbf{F}_\theta(\mathbf{Z}_t) d\mathbf{X}_t. \quad (6)$$

By the chain rule for CDEs, the stacked dynamics can be expressed directly with respect to the control path $\mathbf{X}_t$:

$$d\mathbf{U}_t = \mathbf{G}_\phi(\mathbf{U}_t) \mathbf{F}_\theta(\mathbf{Z}_t) d\mathbf{X}_t. \quad (7)$$

Equivalently, defining the augmented state

$$\mathbf{V}_t = \begin{bmatrix} \mathbf{U}_t \\ \mathbf{Z}_t \end{bmatrix}, \quad (8)$$

the two layers may be integrated jointly as

$$d\mathbf{V}_t = \mathbf{H}_{\theta,\phi}(\mathbf{V}_t) d\mathbf{X}_t, \quad \mathbf{H}_{\theta,\phi}(\mathbf{V}_t) = \begin{bmatrix} \mathbf{G}_\phi(\mathbf{U}_t) \mathbf{F}_\theta(\mathbf{Z}_t) \\ \mathbf{F}_\theta(\mathbf{Z}_t) \end{bmatrix}. \quad (9)$$

This avoids a separate first-layer CDE solver, interpolation of the sampled states $\mathbf{Z}$, and a second CDE solver driven by the reconstructed path. Instead, both layers are propagated in a single augmented solve over the same control. The main practical restriction is that the stacked layers share the same numerical solver and integration grid; when different solvers, tolerances, or output resolutions are required, a two-stage interpolation approach remains more flexible.

In the proposed model, this construction allows a hierarchy of CDE-based conditioning blocks to operate on the same duration-augmented control $\mathbf{X}_t$. The lower layer learns a continuous representation of the aligned phone-duration trajectory, while the upper layer integrates this representation into a higher-level conditioning state.

III. EXPERIMENTS

A. Data

LJSpeech: A well-known 24 hour set of audiobook-style narration spoken by a single U.S. speaker [19]. Experiments use the same split as [20]. Despite its popularity, LJSpeech is prosodically neutral, which may not require the duration-variant features that CDEs provide.

Emotional speech dataset: We follow the original work presented in the first version of StyleTTS [5] by using the emotional speech dataset (ESD) [21] to measure emotional speech synthesis quality. The English subset of ESD comprises 350 sentences spoken by 10 speakers in 5 emotions (sad, happy, angry, neutral, surprised) totalling 10 hours. A subset

of 30 mins per speaker was selected as fine-tuning data. A random subset held out for validation and testing containing 100 samples each, split equally across speakers and emotion. For subjective tests, the test split is reduced to 50 samples which are also split equally. The validation and test sets also contain an accompanying reference split which provides each evaluation sample with an utterance from the same speaker and emotion for style prediction. This dataset is challenging because emotional speech relies on subtle prosodic cues, such as dynamic energy and pitch control, which are perceptually salient [22] but only partially captured by TTS training losses. The use of a 5 hour subset further challenges models to capture nuanced patterns from limited data.

B. Training a CDE within Matcha-TTS

Recall that Matcha-TTS is an encoder-decoder architecture without style embeddings [7]. It uses a neural ODE decoder which offers continuous computational depth. To observe how CDEs interact within TTS pipelines, we set up a CDE that uses the Matcha encoder as a control path and Matcha decoder as a readout layer from hidden state to Mel-spectrogram. Only the CDE is trainable, acting as a bridge between text and acoustic representation. This provides a practical test of whether a CDE can be inserted into an existing TTS pipeline without disrupting the pretrained encoder-decoder mapping.

C. Emotional TTS Guided by Reference Style

The core challenge of this paper involves synthesising speech in the same style as a reference sample, with attention to emotion intensity. This system largely follows Figure 1, where an encoder produces a control path for a CDE, which is then sampled and then transformed to a waveform by a decoder. The non-CDE modules are initialised by training StyleTTS 2 on LibriTTS (multi-speaker audiobook narration data, neutral emotion) [13]. StyleTTS 2 introduces speech language model (SLM) training, where an adversarial objective is used with WavLM features [23], an effective yet computationally demanding process. Several memory-efficient tools are used such as mixed precision, max acoustic length of 175/800 frames, batch size of 6 without SLM adversarial training, and a batch size of 4 with SLM training. The model is trained for a total of 5 epochs, with style diffusion being trained from the second epoch and SLM from the fourth. All modules are trainable. This approach allows finetuning on an NVIDIA A100/H100 GPU with less than 70 GB of video memory. For baseline models we use StyleTTS 2 checkpoints before and after finetuning. This setup uses an ASR-based text aligner. For a strong baseline unrelated to StyleTTS, we consider F5 TTS, which has demonstrated strong zero-shot high-quality synthesis [24]. Studies with autoregressive models are left as future work. Neural CDEs can also be interpreted as continuous-time recurrent models [10]. To examine whether the observed behaviour arises from sequential capacity rather than continuous dynamics, we conduct an experiment comparing several CDE configurations with an

LSTM. Both models are trained with the *duration-augmented control path* $\mathbf{X}$ provided as input.

D. Implementation Details

First, the initial hidden state Z_0 is predicted with an LSTM which takes the control sequence in reverse, and compresses the number of channels. This encodes the full context of the control into the initial state with the initial control state (first phone) X_0 being observed last, $Z_0 = \text{LSTM}(X_L, \dots, X_0)$, providing a strong bias to the CDE. To reduce variance of the control variable $\mathbf{X}$, we normalise the time channel by the total sum of phone durations for the given utterance. The majority of CDE implementations [10] parametrise the vector-field $\mathbf{F}_\theta$ with a multi-layer perceptron and a final hidden dimension dC , which is then reshaped to form a matrix shape satisfying the ODE in equation 4. When the number of control channels is 512 with a hidden dimension of 128, the final linear layer produces a vector of 65536 from 128, a prohibitive expansion. An alternative approach proposed in this paper is to use a U-Net [25] with 1 downsampling, 256 middle, and 512 upsampling channels. The U-Net strides along the hidden dimension d and produces the required matrix to evaluate the CDE (Equation 4). The CDE is integrated with the fixed-step 4th order Runge-Kutta method [26], [27]. Once the hidden sequence has been sampled $Z_{0,\dots,L}$, it is then upsampled to 512 channels with another U-Net striding frame-wise. A gated residual connection $\mathbf{H}^{\text{cde}} = \mathbf{H}^{\text{txt}} + \gamma \mathbf{Z}$, with trainable parameter γ , is used to smoothly integrate the CDE module with the rest of the pretrained model during finetuning. The β_1 hyperparameter of the Adam optimiser is set to 0.9 for the CDE, and 0 elsewhere. The rest of the setup remains unchanged from a typical Matcha-TTS/StyleTTS2 finetuning configuration [7], [13]. Afterwards, the CDE-processed label embeddings are upsampled, ready for the decoder. Although the reparametrisation invariance theorem [16] suggests operating in label-space provides a sufficiently strong model, an interesting question left unanswered in this work is whether a CDE operating one frame per step rather than one phone per step provides a richer temporal progression at the cost of more ODE solver steps.

E. Evaluation

The subjective quality of the models is evaluated with emotion intensity mean opinion score (MOS-EI) through a MUSHRA-type set up on the MTurk crowdsourcing platform. A total of 25 participants were issued to listen to a reference audio, then a list of synthetic and ground truth audio permuted in a random order. The listener is only aware which sample belongs to the reference split. Recall that the reference is given to the TTS model to condition style, but not linguistic content. The participant then rates each clip on a 1-5 integer scale for how strong the emotion is present in the audio e.g. “how happy is this speech?”. This scheme is inspired by the valence-arousal model of emotion [28], where the label, e.g. angry, gives a hint to the target valence-arousal, and the user judges a subjective level of arousal/intensity. The mean opinion of this test reveals

TABLE I
EFFECT OF TRAINING A CDE AS A BRIDGE BETWEEN PRE-TRAINED
ENCODER-DECODER ON LJSPEECH.

System	ODE Steps	MCD ↓	log F_0 RMSE ↓
Matcha-TTS	10	5.36 ± 0.52	0.30 ± 0.08
+CDE	10	5.35 ± 0.49	0.30 ± 0.08
Matcha-TTS	5	5.30 ± 0.52	0.30 ± 0.07
+CDE	5	5.29 ± 0.48	0.29 ± 0.08

how strong different systems express a target emotion, and the correlation of subjective scores to the reference presents how well-calibrated a system is to the intensity of the reference. Overall, each sample was listened to by an average of 10 listeners. Although MUSHRA is effective at evaluating many systems, listeners pay less attention to subtle effects, motivating comparative MOS (CMOS) [13], [29], [30]. CMOS is an A/B test which displays two samples at a time, and prompts the user to rate which one is better. This work constructs emotion quality CMOS (CMOS-EQ) with a +3/-3 integer scale to represent how better expressed the target emotion is in system B vs system A, with negatives meaning system B is worse than system A. One of the systems will always be the candidate, and the other one of the designated baselines, displayed in random order. In both MOS tests, the listeners are prompted to focus on emotional content. Consistency was checked by duplicating 10 comparisons and secretly swapping the A/B order, all participants self-agreed for at least 5 comparisons. A question arising from these studies is how do listeners perceive unlabelled emotional speech? This is answered by measuring emotion accuracy, which prompts the user with a simple data annotation task given both real and synthetic recordings to identify how well intended emotions are expressed by different systems, and how they are perceived. Each sample is classified by 20 participants. Objective metrics such as Mel-cepstral distortion (MCD) and log-F0 root mean squared error (RMSE) are used to measure how different design choices affect model outputs. MCD measures the difference between a generated and ground truth Mel spectrogram, with lower values correlating with audio quality/similarity. Log-F0 RMSE measures the error of the pitch contour of the generated signal, lower values correlate with correct intonation, emotion and prosody.

IV. RESULTS AND DISCUSSION

The results on LJSpeech are displayed in Table I. Given that Matcha-TTS offers continuous computational-depth, we evaluate with both 5 and 10 ODE solver steps. The CDE produces comparable objective scores to the pretrained mapping, with small numerical differences in favour of the CDE conditions. The CDE does not disrupt the encoder-decoder mapping.

Table II shows the objective performance of short-context models (175 frames) over different step-sizes (h) and stacked CDE layers. The LSTM and CDE configurations yield different objective behaviour despite receiving the same duration-augmented inputs, indicating that the CDE is not behaving as

TABLE II
OBJECTIVE EVALUATION RESULTS OF ESD FOR THE BASELINE SYSTEMS
AND CDE (SHORT AND LONG CONTEXT).

System	MCD ↓	log-F0 RMSE ↓
<i>Short-context (175 frames)</i>		
Baseline	5.30 ± 0.99	0.40 ± 0.13
LSTM baseline	5.28 ± 0.98	0.41 ± 0.15
$h = 0.50$, 1 layer	5.20 ± 0.95	0.40 ± 0.13
$h = 0.50$, 2 layers	5.22 ± 1.01	0.40 ± 0.14
$h = 1.00$, 1 layer	5.17 ± 0.97	0.42 ± 0.14
$h = 1.00$, 2 layers	5.35 ± 0.96	0.42 ± 0.14
$h = 1.00$, 4 layers	5.50 ± 1.03	0.41 ± 0.15
<i>Long-context (800 frames)</i>		
Baseline	5.18 ± 0.93	0.40 ± 0.13
F5 baseline	6.24 ± 1.46	0.47 ± 0.18
LibriTTS baseline	5.57 ± 0.89	0.41 ± 0.16
$h = 0.50$, 1 layer	5.20 ± 0.97	0.40 ± 0.16
$h = 0.50$, 2 layers	5.15 ± 0.97	0.40 ± 0.13
$h = 0.50$, 4 layers	5.26 ± 0.92	0.39 ± 0.14
$h = 1.00$, 1 layer	5.04 ± 0.94	0.41 ± 0.14
$h = 1.00$, 2 layers	5.07 ± 0.93	0.40 ± 0.13
$h = 1.00$, 4 layers	5.24 ± 0.93	0.40 ± 0.13

a direct replacement for the discrete recurrent baseline. The best, second-best, and worst results are indicated by blue, orange, and red. All models benefitted from longer context. The F5 baseline suffered pacing inaccuracies.

Table III shows the results of the MUSHRA-style listening test. The first section is the Spearman rank correlation between a system’s per-sample rating and the reference, with a macro correlation reported as an aggregate summary across emotions (computed by Fisher z -transforming the per-emotion correlations, averaging in z -space, and transforming back). The second section is the mean absolute error (MAE), between a system’s intensity rating and the reference. The final section is the overall mean rating, where higher numbers indicate a higher intensity of emotion. Recall that maximising emotion intensity does not indicate a better model, as controllable synthesis requires faithfulness to the reference emotion style, which is captured by the correlation and MAE metrics. The results show that no system is the best across all emotions, and all systems tend to under-express emotion relative to the reference, although the degree of under-expression varies across conditions. The ESD finetune baseline correlates well for surprise, is fair for neutral and happy, but negative for anger and sad. The proposed CDE models have different behaviour depending on their configuration. Using $h = 1.0$ and a single layer hinders neutral speech but improves calibration in other emotions, with surprise being worse off. The macro correlation of this CDE (0.53) is higher than the ESD finetuned baseline (0.08). The dynamics of adjusting time step and stacking layers is non-monotonic, but it appears that stacking layers benefits correlation for $h = 0.5$, albeit a single layer has less MAE from the reference, and grounds further experimentation into the dynamics of continuous-time discretisation.

Table IV shows the results of the emotion expression quality comparison for a 1-layer CDE with $h = 1.0$. On average listeners preferred the CDE method over non-CDE baseline

TABLE III
PER-EMOTION REFERENCE-STYLE TRACKING AND CALIBRATION.

System	Neu.	Ang.	Hap.	Sad	Sur.	Macro
<i>Spearman correlation with reference ($\uparrow$) ($p < 0.05$ in bold)</i>						
style reference	1.00	1.00	1.00	1.00	1.00	1.00
ground truth	0.29	-0.15	0.33	0.29	0.14	0.18
F5	0.66	0.16	0.26	0.15	-0.02	0.26
LibriTTS	0.05	0.12	0.05	0.48	0.16	0.18
ESD finetune	0.35	-0.20	0.15	-0.30	0.39	0.08
h=1.0, 1 layer	0.13	0.64	0.77	0.70	0.19	0.53
h=1.0, 2 layers	-0.16	-0.07	0.62	0.30	0.83	0.38
h=0.5, 1 layer	0.15	0.14	0.25	0.55	0.35	0.30
h=0.5, 2 layers	0.64	0.20	0.64	0.18	0.40	0.43
<i>Mean absolute rating error from reference ($\downarrow$)</i>						
style reference	0.000	0.000	0.000	0.000	0.000	0.000
ground truth	0.227	0.499	0.292	0.447	0.268	0.347
F5	0.473	0.429	0.255	0.521	0.576	0.451
LibriTTS	0.305	0.343	0.289	0.289	0.572	0.360
ESD finetune	0.281	0.482	0.288	0.396	0.418	0.373
h=1.0, 1 layer	0.298	0.426	0.214	0.225	0.543	0.341
h=1.0, 2 layers	0.311	0.405	0.262	0.331	0.555	0.373
h=0.5, 1 layer	0.246	0.419	0.355	0.157	0.403	0.316
h=0.5, 2 layers	0.157	0.538	0.356	0.310	0.517	0.376
<i>Absolute emotion intensity rating (MOS-EI)</i>						
style reference	3.59	3.54	3.49	3.43	3.62	3.53
ground truth	3.35	3.30	3.45	3.19	3.65	3.39
F5	3.09	3.51	3.31	2.96	3.10	3.20
LibriTTS	3.50	3.39	3.19	3.14	3.01	3.24
ESD finetune	3.31	3.20	3.20	3.20	3.18	3.22
h=1.0, 1 layer	3.43	3.12	3.28	3.21	3.05	3.21
h=1.0, 2 layers	3.37	3.29	3.33	3.23	3.05	3.25
h=0.5, 1 layer	3.41	3.21	3.18	3.30	3.17	3.25
h=0.5, 2 layers	3.43	3.14	3.20	3.26	3.08	3.22

TABLE IV
CMOS-EQ RESULTS FOR CDE $h = 1$, 1 LAYER. POSITIVE VALUES INDICATE PREFERENCE FOR CDE.

Baseline	Neutral	Angry	Happy	Sad	Surprise	Overall
ESD finetune	+0.03	+0.37	+0.11	+0.37	-0.28	+0.12
Ground truth	-0.50	-0.90	-1.34	-0.93	-2.01	-1.14

for anger and sadness, no preference for neutral and happy, and preferred the baseline for surprise. When scores were averaged per sample, the proposed and baseline systems were statistically indistinguishable under a Wilcoxon signed-rank test, indicating that the two systems produced comparable emotion-expression quality at the utterance-level. However, when averaged per listener, the results showed a significant ($p < 0.05$) overall preference for the CDE system, suggesting that listeners exhibited a consistent, albeit modest, preference for the proposed system across the evaluation set. In contrast, the CDE is significantly further than ground truth, highlighting room for improvement. Comparing with Table III, the results suggest a partial relationship between calibration and

TABLE V
HUMAN EMOTION CLASSIFICATION ACCURACY $\uparrow$.

System	Neu.	Ang.	Hap.	Sad	Sur.	Macro
style reference	0.602	0.639	0.620	0.662	0.858	0.676
ground truth	0.636	0.780	0.591	0.796	0.653	0.691
LibriTTS	0.606	0.437	0.238	0.398	0.087	0.353
ESD finetune	0.537	0.399	0.304	0.479	0.209	0.386
h=1.0, 1 layer	0.478	0.492	0.245	0.469	0.175	0.372
h=0.5, 2 layers	0.422	0.318	0.294	0.558	0.255	0.369

perceived expression quality. This trend follows for surprise. Neutral-style speech contradicts this but can be considered a special case of non-emotion. Another interesting feature is that the CDE has the lowest intensity rating for anger, yet has +0.37 CMOS-EQ. This could be a compromise where listeners prefer a less intense but more appropriately expressed realisation of anger. Finally, the results for emotion accuracy are in Table V. There is a clear difference between how listeners perceived real speech versus synthetic. Unsurprisingly, LibriTTS overfits to neutral speech, yielding high recall. Between the other systems, the baseline is moderately favoured, with the best accuracy being the 2 layer half step CDE on sadness at 0.558. From a deployment perspective, it seems that the best strategy is to train multiple model configurations and select the optimal model for a given intended style. Together, the results indicate that CDE-based finetuning can improve faithfulness to the reference emotion style and may also improve perceived expressive quality for some emotions, although these benefits remain dependent on emotion category and configuration.

V. CONCLUSION

This paper introduced neural controlled differential equations (CDEs) as a continuous-time mechanism for duration-aware acoustic modelling in text-to-speech. The proposed method augments phone-level representations with duration-derived timing information and uses a neural acoustic vector field to produce hidden states whose values evolve with phonetic content and temporal structure, rather than using duration only for length regulation. Experiments show that CDEs provide a small benefit on neutral audiobook-style speech, but can improve reference-style tracking in emotional TTS. In particular, a single-layer CDE with $h = 1.0$ achieved the strongest macro Spearman correlation with reference emotion-intensity ratings, while CMOS-EQ results indicated comparable utterance-level emotion-expression quality and a modest listener-level preference over the finetuned StyleTTS 2 baseline. The results also show that step size and layer depth affect the trade-off between style tracking, calibration, and perceived emotion quality, suggesting that temporal resolution is a meaningful modelling choice in CDE-based TTS. This work motivates future investigation into alternative interpolation schemes, solver choices, frame-level control paths, and larger-scale evaluation across more diverse speakers, styles, and data resources.

VI. ACKNOWLEDGEMENTS

Thanks to Aaron Fletcher for proof reading. We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing. Portions of the research in this paper used the ESD Database made available by the HLT lab, National University of Singapore, Singapore. The writing of this paper and accompanying code was assisted by ChatGPT and Big Pickle for clearer writing, tables, debugging, and paper/code review.

REFERENCES

- [1] W.-Z. Leung, M. Cross, A. Ragni, and S. Goetze, “Training Data Augmentation for Dysarthric Automatic Speech Recognition by Text-to-Dysarthric-Speech Synthesis,” in *Interspeech 2024*, 2024, pp. 2494–2498.
- [2] N. Rossenbach, B. Hilmes, and R. Schlüter, “On the relevance of phoneme duration variability of synthesized training data for automatic speech recognition,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.
- [3] J. Kim, S. Kim, J. Kong, and S. Yoon, “Glow-tts: A generative flow for text-to-speech via monotonic alignment search,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 8067–8077.
- [4] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 5530–5540. [Online]. Available: <https://proceedings.mlr.press/v139/kim21f.html>
- [5] Y. A. Li, C. Han, and N. Mesgarani, “Styletts: A style-based generative model for natural and diverse text-to-speech synthesis,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 19, no. 1, pp. 283–296, 2025.
- [6] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, “Grad-tts: A diffusion probabilistic model for text-to-speech,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8599–8608. [Online]. Available: <https://proceedings.mlr.press/v139/popov21a.html>
- [7] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, “Matcha-TTS: A fast TTS architecture with conditional flow matching,” in *Proc. ICASSP*, 2024.
- [8] T. J. Lyons, “Differential equations driven by rough signals,” *Revista Matemática Iberoamericana*, vol. 14, no. 2, pp. 215–310, 1998. [Online]. Available: <http://eudml.org/doc/39555>
- [9] N. Wang and D. Yang, “Speech emotion recognition using fine-tuned wav2vec2.0 and neural controlled differential equations classifier,” *PLOS ONE*, vol. 20, no. 2, pp. 1–13, 02 2025. [Online]. Available: <https://doi.org/10.1371/journal.pone.0318297>
- [10] P. Kidger, J. Morrill, J. Foster, and T. Lyons, “Neural controlled differential equations for irregular time series,” in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020.
- [11] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi,” in *Interspeech 2017*, 2017, pp. 498–502.
- [12] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML ’06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: <https://doi.org/10.1145/1143844.1143891>
- [13] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, “Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 19 594–19 621.
- [14] T. Xie, Y. Rong, P. Zhang, W. Wang, and L. Liu, “Towards controllable speech synthesis in the era of large language models: A systematic survey,” in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 764–791. [Online]. Available: <https://aclanthology.org/2025.emnlp-main.40/>
- [15] Y. A. Li, C. Han, X. Jiang, and N. Mesgarani, “Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [16] P. Kidger, “On neural differential equations,” Ph.D. dissertation, University of Oxford, 2021.
- [17] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, “Neural ordinary differential equations,” in *Advances in neural information processing systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.
- [18] S. Y. Jhin, H. Shin, S. Hong, M. Jo, S. Park, N. Park, S. Lee, H. Maeng, and S. Jeon, “Attentive neural controlled differential equations for time-series classification and forecasting,” in *2021 IEEE International Conference on Data Mining (ICDM)*, 2021, pp. 250–259.
- [19] K. Ito and L. Johnson, “The lj speech dataset,” <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [20] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgianakis, and Y. Wu, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.
- [21] K. Zhou, B. Sisman, R. Liu, and H. Li, “Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 920–924.
- [22] P. Larrouy-Maestri, D. Poeppel, and M. D. Pell, “The sound of emotional prosody: Nearly 3 decades of research and future directions,” *Perspectives on psychological science : a journal of the Association for Psychological Science*, vol. 20, no. 4, p. 623–638, July 2025. [Online]. Available: <https://europepmc.org/articles/PMC12231869>
- [23] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [24] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, “F5-tts: A fairytale that fakes fluent and faithful speech with flow matching,” *arXiv preprint arXiv:2410.06885*, 2024.
- [25] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.
- [26] C. Runge, “Ueber die numerische auflösung von differentialgleichungen,” *Mathematische Annalen*, vol. 46, pp. 167–178, 1895. [Online]. Available: <https://api.semanticscholar.org/CorpusID:119924854>
- [27] W. Kutta, *Beitrag zur näherungsweise Integration totaler Differentialgleichungen*. Teubner, 1901.
- [28] J. POSNER, J. A. RUSSELL, and B. S. PETERSON, “The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology,” *Development and Psychopathology*, vol. 17, no. 3, p. 715–734, 2005.
- [29] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, E. Liu, Y. Leng, K. Song, S. Tang, Z. Wu, T. Qin, X. Li, W. Ye, S. Zhang, J. Bian, L. He, J. Li, and S. Zhao, “NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models,” in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 22 605–22 623. [Online]. Available: <https://proceedings.mlr.press/v235/ju24b.html>
- [30] C. Minixhofer, O. Klejch, and P. Bell, “Ttsds-text-to-speech distribution score,” in *SLT*, 2024.