CONSTRAINED DYNAMICAL NEURAL ODE FOR TIME SERIES MODELLING: A CASE STUDY ON CONTINUOUS EMOTION PREDICTION

Ting Dang[1,2] *Antoni Dimitriadis*[2] *Jingyao Wu*[2] *Vidhyasaharan Sethu*[2] *Eliathamby Ambikairajah*[2]

[1]Department of Computer Science and Technology, University of Cambridge, United Kingdom
[2]School of Electrical Engineering and Telecommunications, UNSW Sydney, Australia

ABSTRACT

A number of machine learning applications involve time series prediction, and in some cases additional information about dynamical constraints on the target time series may be available. For instance, it might be known that the desired quantity cannot change faster than some rate or that the rate is dependent on some known factors. However, incorporating these constraints into deep learning models, such as recurrent neural networks, is not straightforward. In this paper, we propose constrained dynamical neural ordinary differential equation (CD-NODE) models, which treat the desired time series as a dynamic process that can be described by an ODE. CD-NODEs model the rate of change of the time series as a function of both itself and the current input features, parameterised as a neural network. We explore the effect of constraining the dynamics of the model by placing explicit restrictions on the rate of change. The proposed model is evaluated on speech-based continuous emotion prediction, where such dynamical constraints are expected, using the publicly available RECOLA dataset. Results suggest that the model achieves performances comparable with the state-of-the-art despite using significantly fewer parameters. Additional analyses reveal that imposing these constraints on the model leads to faster convergence and better performance, especially with smaller training data sets.

1. INTRODUCTION

Time series prediction is of great interest in a wide range of applications, and a significant subset of these applications involve the use of machine learning models. However, the nature and characteristics of these time series can vary significantly. For instance, in the case of emotion state tracking it is known that changes in emotion state may be correlated with specific behavioural events (e.g., laughter) [1], or in the case of health monitoring, biomarkers such as heart rate exhibit characteristic changes in response to changes in the different physical activities (e.g., higher heart rate while running) [2]. Incorporating these known constraints in the prediction model can be expected to allow for both easier training and more accurate predictions.

A variety of different modelling techniques have been employed for time series prediction, ranging from autoregressive (AR) models [3], exponential smoothing [4] and state space and structural models [5, 6] to machine learning techniques such as support vector machines [7] and more recently deep learning techniques such as Recurrent Neural Networks (RNNs) and its variants [8, 9, 10]. Among these, the RNN based models have received the most attention in recent years, achieving state of the art performance in a wide range of applications. However, these recurrent models do not expose the dynamics of the predicted quantity, which in turn makes it difficult to incorporate any prior knowledge of the dynamics or impose constraints on them.

To address these challenges, we propose a novel Constrained Dynamical Neural Ordinary Differential Equation (CD-NODE) model, which explicitly models the dynamics of the time series, and performs predictions using an ODE solver. CD-NODE explores the hypothesis that time series dynamics can be automatically learnt and might even be a simpler modelling task compared to directly modelling the desired time series. Critically, the proposed CD-NODE model incorporates input features as an additional dependent variable. Further, the proposed CD-NODE allows for constraints on the dynamics of the time series, such as limits on the rate of change, to be easily introduced.

We validate the model on a real-world dataset for speech based emotion prediction, as the time-varying emotional state of a person is generally represented as time series of two quantities, arousal (activated to deactivated) and valence (positive to negative) intensity. It can be reasonably expected that the emotional state varies slowly, and significant changes will be correlated with characteristic changes in the modalities used in emotion recognition, such as speech. These observations about emotional state can be easily casted as constraints on the dynamics of emotion state, and therefore serving as a good task to validate the proposed models. The experimental results have shown great promises of CD-NODE compared with state-of-the-art models, and additional analyses have revealed the benefits of the rate constraint, i.e., faster convergence and improved performance.

2. COMBINING NEURAL NETWORKS WITH DYNAMICAL MODELS

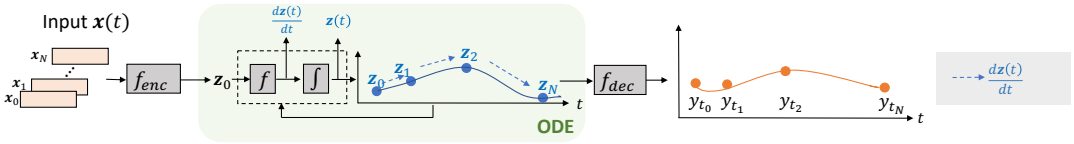
A new family of Neural Ordinary Differential Equations (NODEs) was recently proposed to model a dynamic process described by an ODE with the governing function parameterized by a neural network [11, 12]. This combines the advantages of ODEs in modelling continuous process and the great computational capability of neural networks. One of its variant, latent ODEs [12], was specifically designed for time series prediction. As shown in Figure 1(a), latent ODEs adopt a variational autoencoder to predict a time series $y(t)$ given an input series $\mathbf{x}(t)$. An encoder f_{enc} maps $\mathbf{x}(t)$ to a point $\mathbf{z}_0$ in a latent space, following which a trajectory $\mathbf{z}(t) = [\mathbf{z}_{t_0}, \cdots, \mathbf{z}_{t_N}]$ is obtained by solving an ODE initial value problem:

$$[\mathbf{z}_{t_0}, \cdots, \mathbf{z}_{t_N}] = ODESolve(f, \boldsymbol{\theta}, \mathbf{z}_{t_0}, t_0, \cdots, t_N) \quad (1)$$

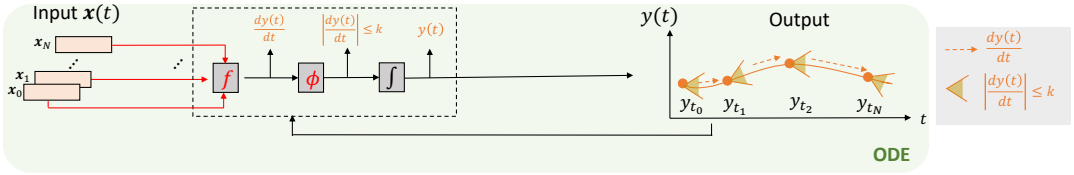
given,

$$\frac{d\mathbf{z}(t)}{dt} = f(\mathbf{z}(t), t; \boldsymbol{\theta}) \quad (2)$$

where, f is a neural network with parameters $\boldsymbol{\theta}$, governing the dynamics of the latent variable $\mathbf{z}$. Finally, a decoder f_{dec} maps the latent space to the output space to obtain time series $y(t)$.



(a) Latent ODEs



(b) Proposed CD-NODEs

Fig. 1: A comparison between latent ODEs and proposed CD-NODEs. (a) Latent ODEs map input sequence $\mathbf{x}_t$ through encoder f_{enc} to an initial point $\mathbf{z}_0$ and evaluate a sequence $\mathbf{z}_t = [\mathbf{z}_0, \cdots, \mathbf{z}_N]$ in the latent space using an ODE. The output y_t is mapped from $\mathbf{z}_t$ through an decoder f_{dec}. (b) Proposed CD-NODE learn the dynamics of y_t (dashed orange arrows) directly using an input-driven ODE, with time-varying $\mathbf{x}_t$ as constraint. Further, additional constraint is incorporated to limit the rate of change $\frac{dy(t)}{dt}$ no larger than k, via function ϕ.

Latent ODEs may suffer from two key limitations. Firstly, the dynamics of $y(t)$ (i.e., $\frac{dy(t)}{dt}$) are not directly learnt, but captured implicitly via $\frac{d\mathbf{z}(t)}{dt}$. It is reasonable to assume that dynamics $\frac{dy(t)}{dt}$ are more associated with the input $\mathbf{x}(t)$ and past observations of $y(t)$ instead of an unknown latent representation, thus modelling $\frac{dy(t)}{dt}$ from $\mathbf{x}(t)$ may be a better solution to capture the underlying dynamics. Secondly, the latent representations $\mathbf{z}(t)$ are learnt in a data-driven manner and do not allow for any known constraints on $\frac{dy(t)}{dt}$ to be imposed. The proposed CD-NODE aims to eliminate the learning in latent space and explicitly model the dynamics $\frac{dy(t)}{dt}$ from $\mathbf{x}(t)$ and $y(t)$, with additional constraints embedded in the model.

3. PROPOSED CONSTRAINED DYNAMICAL NODE

Model definition. The proposed CD-NODE is depicted in Figure 1(b), and the model can be written as:

$$\frac{dy(t)}{dt} = \phi(f(\mathbf{x}(t), y(t); \boldsymbol{\theta})) \tag{3}$$

where f denotes the governing function parameterized by a neural network $\boldsymbol{\theta}$, and ϕ denotes a function that imposes constraints on the rate of change of $y(t)$.

The dynamics $\frac{dy(t)}{dt}$ is implicitly modelled by the input-driven governing function, allowing the dynamics to vary with the input features $\mathbf{x}(t)$ and and past observations $y(t)$. An additional advantage of CD-NODE is the possibility of explicitly incorporating constraints on $\frac{dy(t)}{dt}$, represented by ϕ. For instance, arousal and valence trajectories (representing the time-varying emotion state) are generally smooth and the rate of change may not exceed a certain limit. Such a constraint can be easily introduced by applying a function ϕ as in Eq. (3). Specifically, we define ϕ as:

$$\frac{dy(t)}{dt} = \alpha * \tanh(\frac{1}{\alpha} f(\mathbf{x}(t), y(t); \boldsymbol{\theta}) \tag{4}$$

where $\tanh$ limits the derivatives to -1 and 1, and α scales the $\tanh$ outputs as well as stretching (when $\alpha \geq 1$) the 'linear region' of $\tanh$ function. A large α would be akin to an unconstrained CD-NODE, while a small α limits the value of the derivatives, i.e., the rate of arousal/valence change.

Training of CD-NODE. Learning the CD-NODE model parameters involves a forward propagation using an ODE solver to obtain $y(t)$, and a backward propagation via another call of an ODE solver to optimize the parameters. A Runge-Kutta method [13] is used during forward propagation, and the adjoint sensitivity method is adopted for backpropagation [11].

Forward propagation. The solution of $y(t)$ is obtained using the ODE solver as:

$$y(t) = ODESolve(f, \boldsymbol{\theta}, y_{t_0}, \mathbf{x}(t), t_n), t_n \in [t_0, t_N] \tag{5}$$

Specifically, at each time step, y_{t_n} can be obtained as:

$$y_{t_{n+1}} = y_{t_n} + \int_{t_n}^{t_{n+1}} f(\mathbf{x}(\tau), y(\tau), \boldsymbol{\theta}) d\tau \tag{6}$$

Given the initial condition y_{t_0} and the features $\mathbf{x}(t)$ at each time step t_n, CD-NODE can solve the ODE initial value problem to obtain $y(t) = [y_{t_0}, y_{t_1}, \cdots, y_{t_N}]$ by numerical integration (i.e., Eq. 6). To obtain an accurate estimate during forward propagation, an adaptive Runge-Kutta (RK) ODE solver using a small step size is employed As we additionally introduce the conditional variable of features $\mathbf{x}(\tau)$ to estimate the dynamics, we modified the solver with the assumption that features $\mathbf{x}(\tau)$ are kept unchanged for all the internal steps within $[t_n, t_{n+1}]$. This is reasonable since features are generally quasi-stationary during the short time intervals.

Backward propagation. The loss function, $L(T, \boldsymbol{\theta})$, measures the difference between the predictions $\hat{y}(t)$ and true time series $y(t)$ over all time steps $T = [t_0, t_1, \cdots, t_N]$. It should be chosen based on the application and the characteristics of the time series. Any differentiable loss function can be employed with CD-NODE.

To find the gradients of $L(T, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and optimize the model parameters, the adjoint sensitivity method is adopted. This choice lets us reduce memory cost and numerical errors. $\frac{dL}{d\boldsymbol{\theta}}$ is estimated using chain rule with an intermediate adjoint states of $\mathbf{a}(t) = \frac{\partial L}{\partial y(t)}$, and it involves another ODE solver computing the intermedia values backwards from $t = T$ to $t = 0$. Details can be referred to [11, 14]. We still view features $\mathbf{x}_t$ as constant within two consecutive time steps $[t_n, t_{n+1}]$ for the ODE solver.

Initial value for ODEs. To reconstruct the predicted time series $y(t)$, the ODE solver requires an initial value, y_{t_0}. Consequently, we evaluate the model using an additional layer to learn the initial values and then pass it to the ODE solver. The additional layer for predicting y_{t_0} and all other layers of the model are jointly learnt.

4. MODELLING EMOTION DYNAMICS

The development of the CD-NODE model was motivated by the speech based continuous emotion prediction. Evidence suggests that

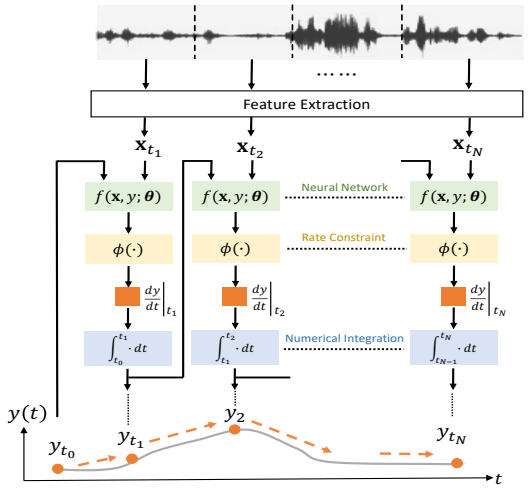


Fig. 2: Proposed CD-NODE for speech emotion prediction. Features $\mathbf{x}(t)$ are first extracted and incorporated as a conditional variable in the model f to learn the dynamics $\frac{dy(t)}{dt}$. Additional constraint on the limits of dynamics is further imposed via ϕ.

emotion change are better perceived compared with numerical ratings of the emotional state [15], and that emotion change might be predicted more accurately than absolute emotion [16].

4.1. Datasets

The RECOLA database is one of the most popular, publicly available emotion dataset, which contains multimodal cues in French including audio, video and physiological signals [17]. Speech collected from 18 speakers was equally partitioned into training and development sets with 9 speakers in each set, identically to the partitions used in [17]. All experiments were conducted using the training dataset to develop the model and optimize the hyperparameters, and evaluated on the development set. The annotation was rated between -1 and 1 by six gender-balanced raters for both arousal and valence sampled every 40ms. The ground truth was obtained as the weighted average value over six raters.

4.2. Experimental setup

An overview of the CD-NODE model employed for speech emotion prediction system is shown in Figure 2.

Features. Bag-of-Audio-Word (BoAW) [18] features were extracted. 39-dimension MFCCs were first extracted from speech on each 25 ms window, with a hop size of 10 ms. Then BoAW is developed using a 100-dimension audio codebook, generated using k-means++ clustering for each 3 seconds window. Readers are referred to [18] for details.

Network parameters. Three fully connected (FC) layers are used to approximate the governing function f in the CD-NODE, with each layer comprised of 64 neurons. $tanh$ activation was employed in the first 2 layers to introduce nonlinearity, while a linear activation was adopted for the 3rd layer in CD-NODE$_{\boldsymbol{\phi}}$ (i.e., without rate constraint ϕ) and a scaled $tanh$ activation for the CD-NODE (i.e., with rate constraint ϕ). The additional layer that predicts the initial values, y_{t_0}, also contains 64 neurons with a $tanh$ activation function. The absolute and relative error tolerances of the ODE solver were chosen as 10^{-13} and 10^{-7} respectively based on preliminary empirical analyses [11].

To determine the proper range of α values (cf. Eq. 4) for rate constraint of CD-NODE, the true rates of change over the test set were calculated by taking successive differences of the labels and dividing by the sample rate of 0.04. The maximum rate of change for arousal and valence were observed to be 6.25 and 3.88 respectively, with 95% quantiles of 0.21 and 0.13. Therefore, we select $\{0.1, 0.25, 0.5, 1, 2, 4, 6, 8\}$ as the set of α values for testing.

To train the model, the loss function was chosen as:

$$L(T, \boldsymbol{\theta}) = 1 - \rho_c(y(t), \hat{y}(t)) \tag{7}$$

where ρ_c represents the Concordance Correlation Coefficient (CCC) (cf. Eq. 8), a measure of similarity between predictions, $\hat{y}(t)$, and the target, $y(t)$. A higher CCC, same as a lower loss, indicates a better predicted time series. Adam optimizer was adopted, and the initial learning rate is optimized to 0.01, with a decaying ratio of 0.9. 60 and 30 epochs were tested for CD-NODE and an LSTM baseline respectively. 2- and 4-second perception delays for arousal and valence were adopted respectively [19]. Three different types of post-processing were tested including mean shift, scaling, and both mean shift and scaling [19], with the best performance reported. It should be noted that the speech utterance was chunked into segments for training to increase the training efficiency, with the length ranging from 1 second to 10 seconds (with a step size of 1 second). The entire test utterance was predicted without segmentation to match the practical scenarios.

Evaluations. All results are reported in terms of CCC [17] as:

$$\rho_c = \frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_y)^2} \tag{8}$$

where σ_y and $\sigma_{\hat{y}}$ are the standard deviation for the ground truth and the predictions, μ_y and $\mu_{\hat{y}}$ are the corresponding mean of the two variables. ρ is the Pearson's correlation coefficient between the two variables [1].

5. EXPERIMENTAL RESULTS

5.1. Comparison with State-of-the-arts

A comparison of the proposed CD-NODE$_\alpha$ with state-of-the-art systems in terms of CCC is shown in Table 1. It can be observed that the proposed systems outperform the baseline for both arousal and valence prediction; and perform comparably, if not somewhat better, than other systems. This is especially true for valence prediction, which is generally the harder of the two for speech-based emotion prediction [19]. This may be because by modelling valence dynamics directly and constraining the valence prediction to be a continuously evolving process, the prediction system is less likely to make noisy predictions. Finally, it should be noted that the proposed system employs a very simple neural network structure comprising of only a few dense layers, while most of the other systems adopt either a complex network structure, a complex training strategy, or both. Specifically, the number of parameters employed in CD-NODE$_\alpha$ is only 19.7% of that in baseline LSTM.

5.2. Impacts of the Rate Constraint

To investigate the effects of the rate constraint, we have compared the model convergence and performance between unconstrained CD-NODE$_{\bar{\alpha}}$ and constrained CD-NODE$_\alpha$ with varying rate constraints. For each α value, the experiment is repeated for 10 different seeds. Both CD-NODE$_{\bar{\alpha}}$ and CD-NODE$_\alpha$ are initialised with the same parameters for a fair comparison.

In terms of the convergence behavior of the models, we first compared the training loss over 60 epochs between CD-NODE$_{\bar{\alpha}}$

[1] Our code is greatly enabled by torchdiffeq, and will be made public available upon publication at Github.

Table 1: Comparison of the proposed CD-NODE$_\alpha$ (bold) and state-of-the-art systems for emotion prediction in terms of Concordance correlation coefficients (CCC). The standard deviation of the CCC computed across all test utterances is reported in paranthesis.

Systems	Features	Arousal	Valence
End-to-end	Raw signals	0.741	0.325
Adversarial	Functionals	**0.797**	0.474
Adversarial^{wd}	Functionals	0.780	**0.501**
Reconstruction	Functionals	0.754	0.378
LSTM	BoAW	0.728(0.098)	0.396(0.145)
CD-NODE$_{\bar{\alpha}}$	BoAW	**0.782**(0.052)	**0.506**(0.119)
CD-NODE$_\alpha$	BoAW	0.778(0.072)	0.491(0.115)

*wd: Wasserstein Distance used in adversarial training.

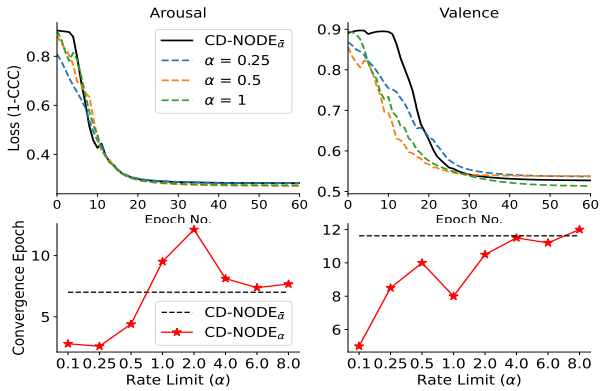


Fig. 3: Comparison of convergence behaviour between CD-NODE$_{\bar{\alpha}}$ and CD-NODE$_\alpha$ with varying values of α. Top: training loss v.s epoch with the best model over 10 seeds. Bottom: convergence epoch averaged across 10 seeds. CD-NODE$_\alpha$ generally shows earlier convergence over CD-NODE$_{\bar{\alpha}}$.

and CD-NODE$_\alpha$. The training loss with the best performed model among 10 different seeds is reported in Figure 3 (top), for arousal and valence respectively. It is observed that the CD-NODE$_\alpha$ with $\alpha = 0.25, 0.5, 1$ converge faster than CD-NODE$_{\bar{\alpha}}$ for both arousal and valence, especially for valence. Further, we compared the number of epochs needed for the training loss to fall under 90% of its initial loss, defined as the convergence epoch (bottom). The value is averaged over 10 seeds. The smaller the convergence epoch, the faster the model convergence rate. The smallest values of α result in earlier convergence on average. On the valence models however, nearly every selected value of α results in an earlier convergence.

Figure 4 shows the model performance in terms of CCC for CD-NODE$_{\bar{\alpha}}$ and CD-NODE$_\alpha$, with different values of α. A significant improvement is observed for arousal at all selected α values. In the case of valence, both models show comparable performance. The poor performance in valence for $\alpha = 0.1$ is to be expected as when the constraints are too tight, it is impossible for the model to adequately capture the dynamics of the signal. Finally, we note that on both arousal and valence, the best performance occurs when $\alpha = 0.5$, where there is a relative improvement of 4.5% and 3.5% over unconstrained CD-NODE$_\alpha$ for arousal and valence.

From these results we conclude that introducing the constraint with a reasonable choice of α results in both improved performance and earlier convergence on both arousal and valence. For both tasks, we found experimentally that $\alpha = 0.5$ yields the best results. Interestingly, this demonstrates that it is not necessary to set the constraints larger than the maximum rates of change seen in the dataset.

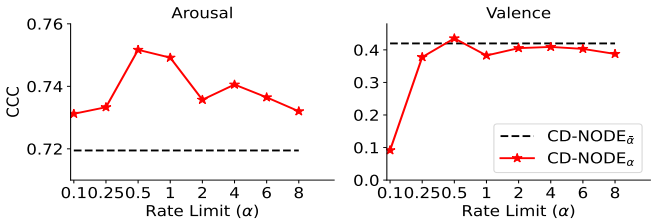


Fig. 4: System performance in terms of CCC for CD-NODE$_{\bar{\alpha}}$ and CD-NODE$_\alpha$ with varying values of α.

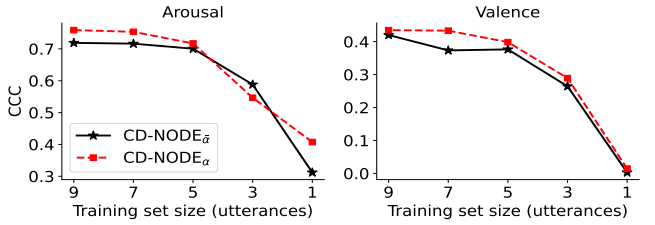


Fig. 5: System performance in terms of CCC with different size of training data for arousal and valence. α is chosen as 0.5 for CD-NODE$_\alpha$.

This is likely because the highest rates of change are due to noise in the labels rather than the underlying signal.

5.3. Smaller Training Sets

We also investigate the impact of the constraints when training data is limited. The size of training data was decreased from 9 training utterances (i.e., 100%) to 1 random utterance (i.e., 11%), with a step size of 2 utterances (i.e., 22%). As the subset of training data is randomly selected, the experiments for each subset were thus conducted for 5 random folds, and the final performance is averaged across 5 folds to reduce the impact of randomness. The experiments were also carried out for 10 different seeds, and the best performance among 10 seeds is reported in Figure 5.

As expected, performance decreases drastically for both CD-NODE$_{\bar{\alpha}}$ and CD-NODE$_\alpha$ as the training set becomes substantially smaller. However, CD-NODE$_\alpha$ generally outperforms the CD-NODE$_{\bar{\alpha}}$ for both arousal and valence, and especially shows a consistent improvement for valence predictions. Specifically, we can observe that when the training set decreases from 9 utterances to 1 utterance, the performance decrease with CD-NODE$_\alpha$ is significantly smaller than CD-NODE$_{\bar{\alpha}}$, with 56% and 46% decrease respectively.

6. CONCLUSION

A novel constrained dynamical neural ODE model has been proposed for time series prediction, which allows for explicit constraints on dynamics of desired time series, such as input-driven nature and the maximum rate of change, to be incorporated into the model. The effectiveness was validated on continuous emotion prediction tasks, and the proposed system was shown to consistently outperform a LSTM based baseline system, and provide prediction accuracy in line with that of state-of-the-art deep learning systems for arousal and valence prediction. The results suggest that modelling emotion dynamics with known constraints is more advantageous than directly modelling the numerical attributes. More importantly, additional analysis has revealed that rate constraint on the model enables a faster convergence and an improved prediction performance for both arousal and valence. In summary, CD-NODE allows for insights from previous empirical studies of time series dynamics to be easily incorporated.

7. REFERENCES

[1] Klaus R Scherer et al., "Psychological models of emotion," *The neuropsychology of emotion*, vol. 137, no. 3, pp. 137–162, 2000.

[2] André E Aubert, Bert Seps, and Frank Beckers, "Heart rate variability in athletes," *Sports medicine*, vol. 33, no. 12, pp. 889–919, 2003.

[3] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung, *Time series analysis: forecasting and control*, John Wiley & Sons, 2015.

[4] Everette S Gardner Jr, "Exponential smoothing: The state of the art," *Journal of forecasting*, vol. 4, no. 1, pp. 1–28, 1985.

[5] James Durbin and Siem Jan Koopman, *Time series analysis by state space methods*, Oxford university press, 2012.

[6] Jan G De Gooijer and Rob J Hyndman, "25 years of time series forecasting," *International journal of forecasting*, vol. 22, no. 3, pp. 443–473, 2006.

[7] Nicholas I Sapankevych and Ravi Sankar, "Time series prediction using support vector machines: a survey," *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 24–38, 2009.

[8] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[9] Bryan Lim, Stefan Zohren, and Stephen Roberts, "Recurrent neural filters: Learning independent bayesian filtering steps for time series prediction," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[10] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[11] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud, "Neural ordinary differential equations," *arXiv preprint arXiv:1806.07366*, 2018.

[12] Yulia Rubanova, Ricky TQ Chen, and David Duvenaud, "Latent odes for irregularly-sampled time series," *arXiv preprint arXiv:1907.03907*, 2019.

[13] John R Dormand and Peter J Prince, "A family of embedded runge-kutta formulae," *Journal of computational and applied mathematics*, vol. 6, no. 1, pp. 19–26, 1980.

[14] Lev Semenovich Pontryagin, *Mathematical theory of optimal processes*, CRC press, 1987.

[15] Georgios N Yannakakis and Hector P Martinez, "Grounding truth via ordinal annotation," in *2015 international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2015, pp. 574–580.

[16] Zhaocheng Huang and Julien Epps, "Prediction of emotion change from speech," *Frontiers in ICT*, vol. 5, pp. 11, 2018.

[17] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 3–10.

[18] Maximilian Schmitt, Fabien Ringeval, and Björn W Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech.," in *Interspeech*, 2016, pp. 495–499.

[19] Zhaocheng Huang, Ting Dang, Nicholas Cummins, Brian Stasak, Phu Le, Vidhyasaharan Sethu, and Julien Epps, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 41–48.