INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Path Signature Approach for Speech Emotion Recognition

Bo Wang[1,3], *Maria Liakata*[3,5], *Hao Ni*[3,4], *Terry Lyons*[2,3], *Alejo J Nevado-Holgado*[1], *Kate Saunders*[1]

[1]Department of Psychiatry, University of Oxford, UK
[2]Mathematical Institute, University of Oxford, UK
[3]The Alan Turing Institute, London, UK
[4]Department of Mathematics, University College London, UK
[5]Department of Computer Science, University of Warwick, UK

bo.wang@psych.ox.ac.uk

Abstract

Automatic speech emotion recognition (SER) remains a difficult task within human-computer interaction, despite increasing interest in the research community. One key challenge is how to effectively integrate short-term characterisation of speech segments with long-term information such as temporal variations. Motivated by the numerical approximation theory of stochastic differential equations (SDEs), we propose the novel use of path signatures. The latter provide a pathwise definition to solve SDEs, for the integration of short speech frames. Furthermore we propose a hierarchical tree structure of path signatures, to capture both global and local information. A simple tree-based convolutional neural network (TBCNN) is used for learning the structural information stemming from dyadic path-tree signatures. Our experimental results on a widely used benchmark dataset demonstrate comparable performance to complex neural network based systems.

Index Terms: speech emotion recognition, path signature feature, convolutional neural network

1. Introduction

Recognising emotions from audio streams in real life has a wide range of commercial applications, especially with the increasing adoption of voice-based assistants such as Alexa and Google Home. Moreover, speech emotion recognition (SER) can be used to assist the detection of psychiatric disorders; e.g., natural expressions of emotion have shown significant positive correlation to heightened mood states in patients with bipolar disorder [1]. Speech emotions can be extracted from both voice characteristics and its linguistic content. In this study, we focus on the acoustic characteristics of the speech signal in order to recognise underlying emotions.

Although the task of recognising speech emotions has attracted significant attention in recent years, it is still a challenging and open research problem. One key step of SER is finding an effective and efficient representation for emotional utterances or speech segments. This is challenging due to the complexity of emotional expressions in speech and the lack of large datasets [2]. Traditional SER systems extract a number of frame-level acoustic features (e.g. fundamental frequency, zero crossing rate, jitter, etc.), known as Low-level Descriptors (LLDs) from utterances of variable lengths, then apply a set of statistical pooling functions (e.g. mean, max, variance, linear regression coefficients, etc.) in order to obtain fixed-size utterance-level features. The role of these high-level statistical pooling functions (HSFs) is to describe the global characteristics of the given utterances, although temporal variations of speech signals are not effectively extracted during this process and important regional information is diluted [3].

To effectively model such temporal information, many recent studies have applied various types of deep learning models, including both convolutional and recurrent neural networks [4, 5, 6, 7]. These models have complex network architectures involving a large number of parameters. As a result they are difficult to build and tune, time-consuming to train and often require expensive computing resources.

Utterances are technically data streams or paths[1]. Path signatures, which were initially introduced in rough path theory as a branch of stochastic analysis, has been successfully applied to various sequence learning tasks, especially for modelling temporal dynamics [8, 9]. In this study, we explore the use of path signatures for modelling temporal sequences of emotional utterances and demonstrate that this method incorporates both the short-term characterisation at the frame-level as well as long-term aggregation at the utterance-level. In addition, path signatures can be applied to input paths of variable length and have the ability to filter redundant information. Dimensionality increases exponentially when higher degree path signature features are used to describe more detailed (local) information in the entire utterance. In order to contain feature dimensionality while capturing both global and local information, we use a hierarchical path structure, namely dyadic path-trees, for representing input utterances and tree-based convolution kernel [10] to capture the underlying structural information.

The contributions of this work are as follows: (1) We show how path signatures can effectively integrate minimally hand-engineered frame-level features (i.e. mel-filterbanks) for SER; (2) We show how tree-based convolutions complement the hierarchical path structure of input utterances; (3) We demonstrate that such a simple CNN model can yield comparable results to complex neural network systems as well as models that make use of a wide range of heavily engineered emotion features.

2. Related Work

2.1. Speech Emotion Recognition

Many existing SER models utilise low-level descriptor features from short frames of typically 20 to 60 msec [11], then either apply a set of high-level statistical functions to get an utterance-

[1]Following Rough Path theory notation, a path refers to a continuous function mapping from a compact time interval $J := [S, T]$ to $E := \mathbb{R}^d$.

http://dx.doi.org/10.21437/Interspeech.2019-2624

level feature representation of emotions, fed to a classifier (support vector machine and extreme learning machine are popular choices [12, 13]), or input them to convolutional or recurrent neural networks for learning longer context and salient features [3, 4, 6, 7]. Some recent SER works have proposed directly feeding neural network models with spectrogram bins [14, 2] or even raw waveforms [15, 16], following the trend in other deep representation learning tasks. However, such end-to-end models have a large number of parameters and are prone to overfitting due to the paucity of training data in SER, as demonstrated by [5]. Our proposed path signature approach operates on minimally hand-engineered filter-bank energy features.

Emotions conveyed by speech are inherently sequential, and it is crucial to model such temporally sequential information for SER. Recurrent neural networks (RNN), especially long short-term memory (LSTM), have gained popularity in handling such sequential data. As described in [7], many categorical SER-specific LSTMs essentially perform a sequence-to-label task. In order to learn high-level representations, different pooling strategies are adopted for these recurrent models: *final-pooling*, *mean-pooling* or *weighted-pooling* LSTMs with attention mechanism added [5, 4, 2, 17]. As result of these pooling operations, important information may be lost from successive frames [18, 7]. In contrast, path signature provides a natural and systematic way of modelling sequential data with variable length and in obtaining fixed-size feature representation. Additionally, inspired by [19], we propose a hierarchical tree structure for path signature features and adopt tree-based convolutions for the integration of global, regional and local information as well as for filtering irrelevant and redundant information.

2.2. Path Signatures

The theory of rough paths, originally studied by Chen [20] and developed by Lyons [21], can be thought of as a non-linear extension of classical theory of controlled differential equations driven by very irregular paths. The essential object in rough path theory is the path signature, which provides a pathwise definition to the solution of differential equations driven by very rough signals [21]. In recent years path signatures have been applied in various areas of machine learning. The main idea is to utilise the signature of a path as the basis function for representing a trajectory in the un-parameterised path space. Levin et al. [22] was the first one to propose a non-parametric time series model by combining path signature features and a linear model. Since then using path signatures as features in a suitable neural network model has shown strong discriminate performance in various applications, such as online handwritten Chinese text recognition [23], action recognition in videos [9], financial data analysis [8] and distinguishing psychiatric disorders using self-reported mood scores [24]. For audio processing, Lyons and Sidorova [25] showed that a stereo audio signal can be reconstructed from a truncated version of a path signature. Thus path signatures have been shown to provide an effective and informative representation for a range of sequential data.

3. Proposed Approach

In this section we first describe a hierarchical approach of representing path signatures, which transforms mel-filterbanks of speech frames to utterance-level fixed-length feature representation. Then we describe how tree-based convolution neural networks are applied to the tree representation of the utterance path signature. The overall design of our model is shown in Figure 1.

3.1. Path Signature

An utterance or audio sample is essentially a sequence of multi-dimensional signals that can be embedded into a continuous path. Path signatures operate in such path space[2]. Thus we propose tackling the challenge of extracting utterance-level feature representations by using path signatures[3]. For each utterance, we interpolate a d-dimensional stream of frames P over the time interval $[0, T] \subset \mathbb{R}$, to a continuous map $P : [0, T] \to \mathbb{R}^d$. Each frame is represented by its 40-dimensional filter-bank energy feature (i.e. $d = 40$). The signature $S(P)$ of this path P over time interval $[0, T]$ is the collection of the k folded iterated integrals of P:

$$S(P)_{0,T} = (1, S(P)^1_{0,T}, \dots, S(P)^k_{0,T}, \dots) \tag{1}$$

where $S(P)^k_{0,T} = \int\limits_{\substack{u_1 < \cdots < u_k \\ u_1, \dots, u_k \in [0,T]}} \cdots \int dP_{u_1} \otimes \cdots \otimes dP_{u_k}, \forall k \geq 1,$

$P_t \in R^d, \forall t \in [0, T]$. In practise we use n^{th} degree truncated signature, where the degree of its iterated integrals is no greater than n. This ensures the path signature has finite dimensional representation. Let $TS(P)^n_{0,T}$ denote the truncated signature of P of degree n, i.e.

$$TS_k(P)^n_{0,T} = (1, S(P)^1_{0,T}, \dots, S(P)^{k_n}_{0,T}) \tag{2}$$

The 0^{th} term (i.e. a constant value set to 1) is optional for feature set. Therefore the dimensionality of the truncated path signature is $(d^{n+1} - d)(d - 1)^{-1}$.

3.2. Dyadic Path Signature Feature

In order to capture a finer description of a path, a higher degree of signature has to be used, which leads to the dimension of signatures growing exponentially. However according to Chen's identity [20], which is stated in eq.(3) below, the information provided by the nth degree signature of the entire path can be well approximated by the concatenation of the lower degree signatures over all the partitions of this path:

$$S(P)^n_{0,T} = \sum_{j=0}^{n} S(P)^j_{0,S} \otimes S(P)^{n-j}_{S,T} \tag{3}$$

where $0 \leq S \leq T$ and $\otimes$ denotes a tensor product and the superscript denotes the length of indices. Thus Yang et al. [19] proposes replacing the higher degree of path signatures with the lower degree of the signatures over the dyadic partition of this path. More specifically, a dyadic path signature with the dyadic level of m is the collection of the signatures of M dyadic pieces of the entire path:

$$\begin{aligned} D^m S(P)_{0,T} = (&S(P)_{0,\frac{1}{M}T}, \dots, S(P)_{\frac{i-1}{M}T,\frac{i}{M}T}, \\ &\dots, S(P)_{\frac{M-1}{M}T,T}). \end{aligned} \tag{4}$$

where $M = 2^m, m \in \mathbb{N}$. To combine information from different granularity, the final dyadic path signature is as follows:

$$\begin{aligned} DTS(P)^n_{0,T} = (&D^0 TS(P)^n_{0,T}, D^1 TS(P)^n_{0,T}, \\ &\dots, D^m TS(P)^n_{0,T}) \end{aligned} \tag{5}$$

[2]The ability to embed discrete streams of sequential data into a path space provides the flexibility to describe them in a unified way. This can help with the problem of missing data, sequences of variable length and unequal spaced sampling.

[3]The rigorous introduction of path signatures as a faithful description for un-parameterised paths can be found in [26].

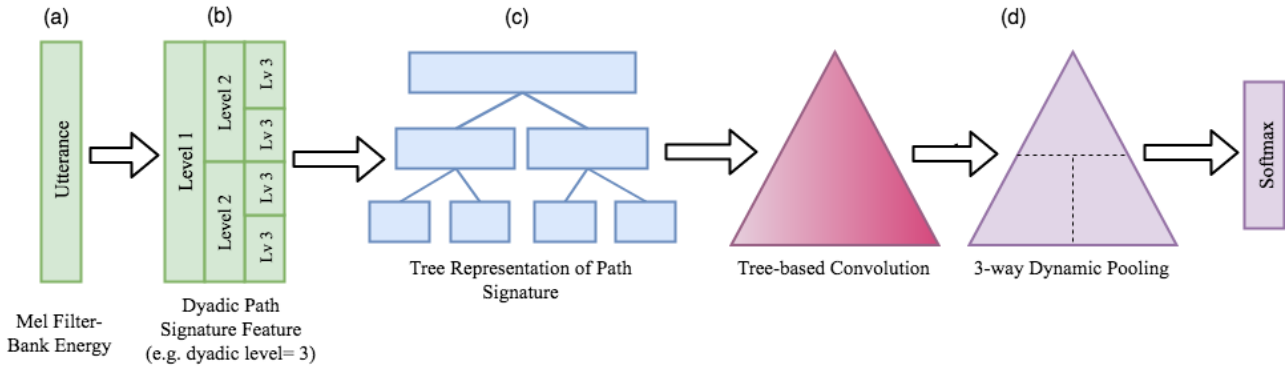

Figure 1: *Overview of the proposed SER system. (a) A stream of frame-level energy features is extracted from an utterance; (b) The entire stream of frames within the utterance is segmented to dyadic paths. A truncated signature is extracted from each dyadic path, and in (c) is transformed to the dyadic path-tree signature representation; (d) Tree-based convolution and dynamic pooling are applied to learn the underlying structure, and an output layer is added for the final classification.*

Instead of simply concatenating all the dyadic path signatures horizontally as is done in [19], in this work we propose to present all the dyadic pieces in a tree form, from the entire utterance to only a few short frames[4], namely dyadic path-tree. This way salient features at different granularity along with any structural information can be extracted more naturally and explicitly. As illustrated in Figure 1c (blue), this dyadic path-tree signature constitutes the root node representing the signature of the entire path (i.e. $D^0 TS(P)^n_{0,T}$) and all the subsequent children nodes over each dyadic level.

3.3. Tree-Based Convolutional Neural Network

We use a set of tree-based convolution kernels, which function as fixed-depth[5] feature detectors sliding over the entire dyadic path-tree. To extract positional information of each sub-path-signature, we adopt the continuous tree approach in [10]. For node x_i in a window, its weight matrix is the linear combination of three positional weight matrices W^t_{conv}, W^l_{conv} and W^r_{conv}, and three coefficients η^t_i, η^l_i and η^r_i, (referring to the *left*, *right* and *top* positions). These coefficients are computed according to the relative position of the node in the sliding window and the depth of the window in the tree. We thereafter apply three-way dynamic pooling, where features are pooled according to their position in the tree. Finally, we add a softmax layer for classification into emotions.

The entire parameter set for this tree-based CNN model is $\Theta = \{W^t_{conv}, W^l_{conv}, W^r_{conv}, W_{out}, b_{conv}, b_{out}, \mathrm{vec}(\cdot)\}$, where W^t_{conv}, W^l_{conv}, W^r_{conv} and b_{conv} are the positional weights and bias in the convolutional layer, W_{out} and b_{out} are the weight and bias in the output layer, and $\mathrm{vec}(\cdot)$ is the pre-computed path signature feature for the input utterances. All the weights and biases are initialised randomly from a truncated normal distribution.

[4]Previous work [27] has considered a hierarchical tree structure of speech in the time-frequency scale-space using a composite prosodic signal consisting of energies of utterances in English.

[5]In our experiment, we set this window depth to 2.

4. Experiments and Analysis

4.1. Data

To evaluate the effectiveness of our proposed system, the popular Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [28] is used for all experiments. It comprises approximately 12 hours of audio-visual recordings performed by 10 skilled actors. All recordings are organised in 5 sessions, and each session is composed of two actors, one male and one female. Overall it contains 10039 (manually segmented) utterances with an average duration of 4.5 seconds. The database can be further divided into an improvised speech data set and a scripted data set. To be consistent with previous works [2, 29, 30, 31, 32], we consider 4 emotional categories *Angry*, *Happy*, *Neutral* and *Sad*, and choose the improvised data set since the scripted speech exhibits strong correlation with the manually labelled emotions leading to bias over linguistic content learning.

4.2. Feature Extraction

We use the openSMILE toolkit [33] for extracting 40-dimensional mel-filterbank features from each utterance, and add one more dimension for representing time. A sliding Hamming window of length 25 ms is used for segmenting to frames. Then these features are normalised by z-scores. For computing the path signatures over each stream of frames, we use iisignature[6] Python package and set the degree of signatures to 2.

4.3. Experimental Setup

Following previous works [2, 29, 30, 34], we choose leave-one-speaker-out as evaluation scheme. We use Tensorflow as the deep learning framework for training. We use fixed choices of hyperparameters for all of our experiments to ensure the results are comparable and reproducible. These hyperparameters are shown in Table 1. Cross-entropy loss is used for optimisation. The final SER performance is evaluated using widely adopted metrics: weighted accuracy (WA), which is the overall classification accuracy; and unweighted accuracy (UA), which aver-

[6]https://pypi.org/project/iisignature/

Table 1: *Hyperparameter choices.*

Degree of signature	2
Dyadic level	4
Optimiser	Adam [35]
Batch size	50
Max. epochs	60
Learning rate	0.0001
No. of convolution filters	1
Convolutional layer dim.	100
Activation function	`tanh`

ages accuracy of each emotion category.

We select a range of neural network models as our baselines, including three bi-directional LSTM models [34], a deep neural network (DNN) followed by an Extreme Learning Machine (ELM) [29], as well as a CNN model using an attention mechanism [4]. Among the LSTM models, COVAREP extracts commonly used speech features for emotion recognition, using the COVAREP toolkit [36], while both LSTM (Speech) and LSTM (Glottal) extract spectrograms from the speech waveforms and glottal flow waveforms respectively. The attentive CNN model [4] uses a range of features including log mel-filterbanks and the hand-crafted eGeMAPS feature set [11]. It cuts and pads to make sure each utterance is 7.5s long. By contrast our path signature approach allows us to deal with utterances of variable length in a natural way.

4.4. Results and Discussion

Table 2 summarises the performance comparison between different methods. As we can see, the best UA score 56.83% and WA score 61.95% are achieved by the attentive CNN model [4]. Despite its simplicity, the performance of our model, named **PTS-CNN** (i.e. Path-Tree-Signature based CNN), is on par with the best results (-3.80% in UA and -3.05% in WA comparing to Attentive CNN). It also outperforms other LSTM and DNN based models which have complex network design. In terms of computation time, our model takes on average 57 seconds to train and 0.7 seconds for inference, per fold. It has a small number of hyperparameters, which makes it very easy to tune[7]. The dyadic path signature of an utterance with dyadic-level $m = 4$ (i.e. contains 15 paths) and signature-degree of $n = 2$, takes on average 0.0061 seconds to compute[8].

Table 2: *Accuracy comparison among different models*

Model	**UA**	**WA**
COVAREP [34]	51.84	49.64
DNN-ELM [29]	52.13	57.91
LSTM (Speech) [34]	51.85	51.94
LSTM (Glottal) [34]	54.56	52.82
Attentive CNN [4]	56.83	61.95
PTS-CNN (our model)	53.03	58.90

The shuffle product identity [21] of signature states that the product of two lower-level signature coefficients can be expressed as a linear combination of some higher-level coefficients. Therefore by transforming filter-bank energy of a stream of frames to a signature of degree 2, we automatically include more nonlinear prior knowledge in our final feature set (at utterance level). The use of dyadic path-tree signature allows us to capture long and short term information of the utterance without having to use higher degree signatures. As a result, this allows the minimal design of our model, as only one convolution layer is used without any fully connected hidden layer.

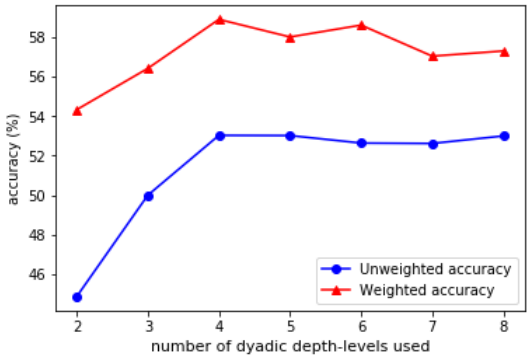

Figure 2: *Model performance with increasing dyadic level*

In the second experiment, we evaluate model performance by increasing the hierarchical level of our dyadic path-tree. As seen in Figure 2, we observe a significant increase in both weighted and unweighted accuracy between level 2 and 4, showing the benefits of having more dyadic partitions and extracting regional and local information. However, little difference in performance is observed after level 4, which suggests level 4 dyadic path signature contains sufficient local information of the utterances. In fact, very small dyadic paths can become sensitive to local noise of sampling points.

5. Conclusions

In this paper, different to all the existing speech emotion works, we presented a path signature approach for speech emotion recognition. Specifically, we propose a dyadic path-tree structure of signature, capturing both global and local information. By using tree-based convolution, we show our path signature approach can effectively integrate speech frames and efficiently learn discriminative features for classification of emotion in utterances. The experiments demonstrate that the proposed PTS-CNN model achieves comparable performance to existing works, while requiring minimal model tuning or manual engineering.

For future work, we plan to evaluate the path signature approach with other LLDs (e.g. pitch, jitter, etc.), and also incorporate the signature transform as part of a neural network with backpropagation, combining speech with other modalities such as transcribed text and video.

6. Acknowledgements

This work was supported by the MRC Mental Health Data Pathfinder award to the University of Oxford [MC_PC_17215], by the NIHR Oxford Health Biomedical Research Centre and by the The Alan Turing Institute under the EPSRC grant EP/N510129/1.

[7] All experiments including signature computation are performed on Microsoft Azure NC6 VM, which has 6 cores of Intel Xeon CPU E5-2690 v3 @ 2.60GHz, 56GB of RAM and a NVIDIA Tesla K80.

[8] Both the signature computation and model inference time are negligible, which makes real-time deployment possible.

7. References

[1] S. Khorram, M. Jaiswal, J. Gideon, M. McInnis, and E. Mower Provost, "The priori emotion dataset: Linking mood to emotion detected in-the-wild," in *Interspeech*, 2018.

[2] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Interspeech*, 2018, pp. 3087–3091.

[3] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *ICASSP 2017*. IEEE, 2017, pp. 2741–2745.

[4] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *Proc. Interspeech 2017*, pp. 1263–1267, 2017.

[5] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *ICASSP 2017*. IEEE, 2017, pp. 2227–2231.

[6] F. Tao and G. Liu, "Advanced lstm: A study about better time dependency modeling in emotion recognition," in *ICASSP 2018*. IEEE, 2018, pp. 2906–2910.

[7] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, and B. Schuller, "Towards temporal modelling of categorical speech emotion recognition," in *Proc. Interspeech 2018*, 2018, pp. 932–936.

[8] L. G. Gyurkó, T. Lyons, M. Kontkowski, and J. Field, "Extracting information from the signature of a financial data stream," *Quantitative Finance*, 2013.

[9] W. Yang, T. Lyons, H. Ni, C. Schmid, L. Jin, and J. Chang, "Leveraging the path signature for skeleton-based human action recognition," *arXiv preprint arXiv:1707.03993*, 2017.

[10] L. Mou, G. Li, L. Zhang, T. Wang, and Z. Jin, "Convolutional neural networks over tree structures for programming language processing," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[11] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[12] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *ICASSP 2011*. IEEE, 2011, pp. 5688–5691.

[13] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[14] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[15] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP 2016*. IEEE, 2016, pp. 5200–5204.

[16] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using dnns," *Proc. Interspeech 2018*, pp. 3097–3101, 2018.

[17] P.-W. Hsiao and C.-P. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *ICASSP 2018*. IEEE, 2018, pp. 2526–2530.

[18] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 517–524.

[19] W. Yang, L. Jin, H. Ni, and T. Lyons, "Rotation-free online handwritten character recognition using dyadic path signature features, hanging normalization, and deep neural network," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 4083–4088.

[20] K.-s. Chen, "Integration of pathsa faithful representation of paths by non-commutative formal power series," *Transactions of the American Mathematical Society*, vol. 89, no. 2, pp. 395–407, 1958.

[21] T. J. Lyons, "Differential equations driven by rough signals," *Revista Matemática Iberoamericana*, vol. 14, no. 2, pp. 215–310, 1998.

[22] D. Levin, T. Lyons, and H. Ni, "Learning from the past, predicting the statistics for the future, learning an evolving system," *arXiv preprint arXiv:1309.0260*, 2013.

[23] Z. Xie, Z. Sun, L. Jin, H. Ni, and T. Lyons, "Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 8, pp. 1903–1917, 2018.

[24] I. P. Arribas, K. Saunders, G. Goodwin, and T. Lyons, "A signature-based machine learning model for bipolar disorder and borderline personality disorder," *Translational Psychiatry*, p. 274, 2018.

[25] T. J. Lyons and N. Sidorova, "Sound compression: a rough path approach," in *Proceedings of the 4th international symposium on Information and communication technologies*. Trinity College Dublin, 2005, pp. 223–228.

[26] T. Lyons, "Rough paths, signatures and the modelling of functions on streams," in *Proceedings of the International Congress of Mathematicians*, 2014, pp. 163–184.

[27] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[28] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[29] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[30] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *INTERSPEECH*, 2017, pp. 1089–1093.

[31] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1771–1775.

[32] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," in *Proc. Interspeech 2018*, 2018, pp. 3683–3687.

[33] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[34] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition." in *Interspeech*, 2016, pp. 3603–3607.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *ICASSP 2014*. IEEE, 2014, pp. 960–964.