Interspeech 2018
2-6 September 2018, Hyderabad, India

Integrating Recurrence Dynamics for Speech Emotion Recognition

Efthymios Tzinis[1,2,†], *Georgios Paraskevopoulos*[1,2,†], *Christos Baziotis*[1], *Alexandros Potamianos*[1,2]

[1]School of Electrical & Computer Engineering, National Technical University of Athens, Greece
[2]Behavioral Signal Technologies, Los Angeles, CA, USA
†Both authors contributed equally to this work

etzinis@gmail.com, geopar@central.ntua.gr, cbaziotis@mail.ntua.gr, potam@central.ntua.gr

Abstract

We investigate the performance of features that can capture nonlinear recurrence dynamics embedded in the speech signal for the task of Speech Emotion Recognition (SER). Reconstruction of the phase space of each speech frame and the computation of its respective Recurrence Plot (RP) reveals complex structures which can be measured by performing Recurrence Quantification Analysis (RQA). These measures are aggregated by using statistical functionals over segment and utterance periods. We report SER results for the proposed feature set on three databases using different classification methods. When fusing the proposed features with traditional feature sets, e.g., [1], we show an improvement in unweighted accuracy of up to 5.7% and 10.7% on Speaker-Dependent (SD) and Speaker-Independent (SI) SER tasks, respectively, over the baseline [1]. Following a segment-based approach we demonstrate state-of-the-art performance on IEMOCAP using a Bidirectional Recurrent Neural Network.

Index Terms: speech emotion recognition, recurrence quantification analysis, nonlinear dynamics, recurrence plots

1. Introduction

Automatic Speech Emotion Recognition (SER) is key for building intelligent human-machine interfaces that can adapt to the affective state of the user, especially in cases like call centers where no other information modality is available [2].

Extracting features capable of capturing the emotional state of the speaker is a challenging task for SER. Prosodic, spectral and voice quality Low Level Descriptors (LLDs), extracted from speech frames, have been extensively used for SER [3]. Proposed SER approaches mainly differ on the aggregation and temporal modeling of the input sequence of LLDs. In utterance-based approaches, statistical functionals are applied over all LLD values of the included frames [1]. These utterance-level statistical representations have been successfully used for SER using Support Vector Machines (SVMs) [4], Convolutional Neural Networks (CNNs) [5] and Deep Belief Networks (DBNs) in a multi-task learning setup [6]. Moreover, segment-based approaches have showcased that computation of statistical functionals over LLDs in appropriate timescales yields a significant performance improvement for SER systems [7], [8]. Specifically, in [8] statistical representations are extracted from overlapping segments, each one corresponding to a couple of words. The resulting sequence of segments representations is fed as input to a Long Short Time Memory (LSTM) unit for SER classification.

Direct SER approaches are usually based on raw LLDs extracted from emotional utterances. CNNs [9] and Bidirectional-LSTMs (BLSTMs) [10] over spectrogram representations reported state-of-the-art performances on Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [11]. LSTMs with attention mechanisms have also been proposed in order to accommodate an active selection of the most emotionally salient frames [12], [13]. To this end, Sparse Auto-Encoders (SAE) for learning salient features from spectrograms of emotional utterances have also been studied [14].

Despite the great progress that has been made in SER, the aforementioned LLDs are extracted under the assumption of a linear source-filter model of speech generation. However, vocal fold oscillations and vocal tract fluid dynamics often exhibit highly nonlinear dynamical properties which might not be aptly captured by conventional LLDs [15]. Nonlinear analysis of a speech signal through the reconstruction of its corresponding Phase Space (PS) lies in embedding the signal in a higher dimensional space where its dynamics are unfolded [16]. Recurrent patterns of these orbits are indicative attributes of system's behavior and can be analyzed using Recurrence Plots (RPs) [17]. Recurrence Quantification Analysis (RQA) provides complexity measures for an RP which are capable of identifying a system's transitions between chaotic and order regimes [18]. A variety of nonlinear features like: Teager Energy Operator [19], modulation features from instantaneous amplitude and phase [20] as well as geometrical measures from PS orbits [21] have been reported to yield significant improvement on SER when combined with conventional feature sets. However, RQA analysis has not yet been employed for SER. In [22] RQA measures have been shown to be statistically significant for the discrimination of emotions but an actual SER experimental setup is missing.

In this paper, we extract RQA measures from speech-frames and evaluate them for SER. We test the efficacy of the proposed RQA feature set under both utterance and segment-based approaches by calculating statistical functionals over the respective time lengths. SVMs and Logistic Regression (LR) classifiers are used for the utterance-based approach as well as an Attention-BLSTM (A-BLSTM) for the respective segment-based approach. The performance of the proposed RQA feature set, as well as the fusion of the RQA features with conventional feature sets [1], is reported on three databases and compared with state-of-the-art results for Speaker-Dependent (SD), Speaker-Independent (SI) and Leave One Session Out (LOSO) SER experiments.

2. Feature Extraction

2.1. Baseline Feature Set (IS10 Set)

We use the IS10 feature set [1], in which 1582 features are extracted corresponding to statistical functionals applied on various LLDs. The extraction is performed for both segment and utterance based approaches using the openSMILE toolkit [23].

10.21437/Interspeech.2018-1377

2.2. Proposed Nonlinear Feature Set (RQA Set)

The RQA feature set for a given speech segment or utterance is extracted as described next. First, we break the given speech signal into frames and for each one we reconstruct its PS as shown in Section 2.2.1. For each PS orbit, its respective RP is computed as explained in Section 2.2.2. In order to quantify the complex structures of the RP, a list of RQA measures (described in Section 2.2.3) is extracted; resulting in a 12-dimensional representation of the input speech frame. Representations for speech-segments and utterances containing multiple frames are obtained by applying a set of 18 statistical functionals (listed in Section 2.2.4) over 12-dimensional frame-attributes and their deltas. Thus, a 432-dimensional feature vector is obtained.

2.2.1. *Phase Space Reconstruction*

Given a speech frame with N samples $\{s(i)\}_{i=1}^{N}$ we reconstruct its corresponding PS trajectory by computing m time-delayed versions of the original speech frame by multiples of time lag τ and creating the vectors lying in $\mathbb{R}^m$ as shown next:

$$\mathbf{x}(i) = [s(i), s(i+\tau), ..., s(i+(m-1)\tau)] \quad (1)$$

where m is the embedding dimension of the reconstructed PS and τ is the time lag. If the embedding theorem holds and the aforementioned parameters are set appropriately, then the orbit defined by the points $\{\mathbf{x}(i)\}_{i=1}^{N}$ would truthfully preserve invariant quantities of the true underlying dynamics which are assumed to be unknown [24]. In accordance with [16], parameters τ and m for each speech frame are estimated individually by using Average Mutual Information (AMI) [25] and False Nearest Neighbors (FNN) [26], respectively.

2.2.2. *Recurrence Plot*

Given a PS trajectory $\{\mathbf{x}(i)\}_{i=1}^{N}$ we analyze the recurrence properties of these states by calculating the pairwise distances and thresholding these values in order to compute the corresponding RP [17]. RPs are binary square matrices and are defined element-wise as shown next:

$$\mathbf{R}_{i,j}(\epsilon, q) = \Theta(\epsilon - ||\mathbf{x}(i) - \mathbf{x}(j)||_q) \quad (2)$$

where $\Theta(\cdot)$ is the Heaviside step function, ϵ is the thresholding value, $||\cdot||_q$ is the norm used to define the distance between trajectory points (for $q = 1$, $q = 2$ or $q = \infty$ we compute Manhattan, Euclidean or Supremum norm, respectively). Thus, matrix $\mathbf{R}$ consists of ones in areas where the states of the orbit are close and zero elsewhere. The measure of proximity is defined by threshold ϵ for which multiple selection criteria have been studied [27]. We consider three criteria depending on: 1) a fixed ad-hoc threshold value, 2) a fixed Recurrence Rate (RR) as defined in Table 1 (e.g., For $RR = 0.15$ we set ϵ according to a fixed probability of the pairwise distances of PS's points $P(||\mathbf{x}(i) - \mathbf{x}(j)||_q < \epsilon) = 0.15,\ 1 \leq i, j, \leq N$), and 3) a fixed ratio of the standard deviation σ of points $\{\mathbf{x}(i)\}_{i=1}^{N}$, e.g., $\epsilon = 5\sigma$ [28]. For fixed values of ϵ and q we denote as $\mathbf{R}_{i,j}$ the respective entry of the RP matrix for simplicity of notation.
An L-length diagonal line (of ones) is defined by:

$$(1 - \mathbf{R}_{i-1,j-1})(1 - \mathbf{R}_{i+L+1,j+L+1}) \prod_{k=1}^{k=L} \mathbf{R}_{i+k,j+k} = 1 \quad (3)$$

An L-length vertical line is described by:

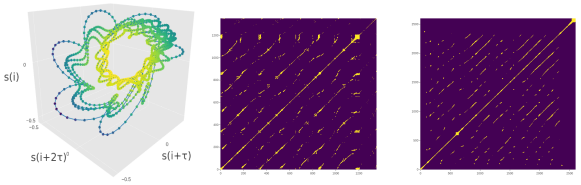
$$(1 - \mathbf{R}_{i,j-1})(1 - \mathbf{R}_{i,j+L+1}) \prod_{k=1}^{k=L} \mathbf{R}_{i,j+k} = 1 \quad (4)$$

An L-length white vertical line (of zeros) is defined as:

$$\mathbf{R}_{i,j-1}\mathbf{R}_{i,j+L+1} \prod_{k=1}^{k=L} (1 - \mathbf{R}_{i,j+k}) = 1 \quad (5)$$

We also denote with $P_d(l)$, $P_v(l)$ and $P_w(l)$ the histogram distributions of lengths of diagonal, vertical and white vertical lines, respectively. Hence, the total number of these lines are correspondingly $N_d = \sum_{l \geq d_m} P_d(l)$, $N_v = \sum_{l \geq v_m} P_v(l)$ and $N_w = \sum_{l \geq w_m} P_w(l)$, where $d_m = 2$, $v_m = 2$ and $w_m = 1$ define the minimum lengths for each type of line [18].

Emerging small-scale structures based on lines of ones or zeros reflect the dynamic behavior of the system. For instance, diagonal lines indicate both similar evolution of states for different parts of PS's orbit and deterministic chaotic dynamics of the system [18]. This is also depicted in Figure 1.



(a) (b) (c)

Figure 1: *(a) Reconstructed PS ($m = 3$, $\tau = 7$) and (b) RP ($\epsilon = 0.15$, Manhattan norm) of 30ms frame corresponding to vowel /e/. (c) RP of Lorenz96 system displaying chaotic behavior [29]*

2.2.3. *Recurrence Quantification Analysis (RQA)*

For each $N \times N$ RP we extract 12 RQA measures using the pyunicorn framework [30]. Following the notation established in Section 2.2.2 we provide an overview of these measures in Table 1; they are comprehensively studied in [18], [31].

Table 1: *Recurrence Quantification Analysis Measures*

Name	Formulation
Recurrence Rate	$\frac{1}{N^2}\sum_{i,j=1}^{N} \mathbf{R}_{i,j}$
Determinism	$\frac{\sum_{l=d_m}^{N} lP_d(l)}{\sum_{l=1}^{N} lP_d(l)}$
Max Diagonal Length	$max(\{l_i\}_{i=1}^{N_d})$
Average Diagonal Length	$\frac{\sum_{l=d_m}^{N} lP_d(l)}{\sum_{l=d_m}^{N} P_d(l)}$
Diagonal Entropy	$\sum_{l=d_m}^{N} \frac{P_d(l)}{N_d} ln(\frac{N_d}{P_d(l)})$
Laminarity	$\frac{\sum_{l=v_m}^{N} lP_v(l)}{\sum_{l=1}^{N} lP_v(l)}$
Max Vertical Length	$max(\{v_i\}_{i=1}^{N_v})$
Trapping Time	$\frac{\sum_{l=v_m}^{N} lP_v(l)}{\sum_{l=v_m}^{N} P_v(l)}$
Vertical Entropy	$\sum_{l=v_m}^{N} \frac{P_v(l)}{N_v} ln(\frac{N_v}{P_v(l)})$
Max White Vertical Length	$max(\{w_i\}_{i=1}^{N_w})$
Average White Vertical Length	$\frac{\sum_{l=w_m}^{N} lP_w(l)}{\sum_{l=w_m}^{N} P_w(l)}$
White Vertical Entropy	$\sum_{l=w_m}^{N} \frac{P_w(l)}{N_w} ln(\frac{N_w}{P_w(l)})$

2.2.4. Statistical Functionals

After the extraction of frame-wise features and their associated deltas we apply 18 statistical functionals: min, max, mean, median, variance, skewness, kurtosis, range, 1_{st}, 5_{th}, 25_{th}, 50_{th}, 75_{th}, 95_{th}, 99_{th} percentile and 3 quartile ranges.

2.3. Fused Feature Set (RQA + IS10 Set)

For any emotional speech segment or utterance we extract both feature sets IS10 and RQA as described previously and concatenate them. The final feature vector has 2014 dimensions.

3. Classification Methods

We investigate both utterance-based and segment-based SER as outlined below:

Utterance-based method: For each utterance we obtain its statistical representation by extracting the corresponding feature set as described in Section 2. For emotion classification we employ an SVM with Radial Base Function (RBF) kernel and one-versus-rest LR classifier. Cost coefficient C lies in the interval $[0.001, 30]$ for both SVM and LR models which is the only hyper-parameter to be tuned. Both models are implemented using the scikit-learn framework [32].

Segment-based method: We break each utterance into segments of 1.0 s length and 0.5 s stride in accordance with [8]. For each speech segment we extract the feature sets described in Section 2 and as a result each utterance is now represented by a sequence of statistical vectors corresponding to different time steps. This sequence is fed as an input to a Long Short Time Memory (LSTM) unit for emotion classification. SER can be formulated as a many-to-one sequence learning where the expected output of each sequence of segment features is an emotional label derived from the activations of the last hidden layer [12]. We employ an A-BLSTM architecture [13] where the decision for the emotional label is derived from a weighted aggregation of all timesteps. We implement this architecture in pytorch [33]. In addition, the grid space of hyper-parameters consists of: number of layers $\{1, 2\}$, number of hidden nodes $\{128, 256\}$, input noise $[0.3, 0.8]$, dropout rate $[0.3, 0.8]$ and learning rate $[0.0002, 0.002]$.

4. Experiments and Results

The following databases are used in our experiments:

SAVEE: Surrey Audio-Visual Expressed Emotion (SAVEE) Database [34] is composed of emotional speech voiced by 4 male actors. SAVEE includes 480 utterances (120 utterances per actor) of 7 emotions i.e., 60 anger, 60 disgust, 60 fear, 60 happiness, 60 sadness, 60 surprise and 120 neutral.

Emo-DB: Berlin Database of Emotional Speech (Emo-DB) [35] contains 535 emotional sentences in German, voiced by 10 actors (5 male and 5 female). Specifically, 7 emotions are included i.e., 127 anger, 45 disgust, 70 fear, 71 joy, 60 sadness, 81 boredom and 70 neutral.

IEMOCAP: IEMOCAP database [11] contains 12 hours of video data of scripted and improvised dialog recorded from 10 actors. Utterances are organized in 5 sessions of dyadic interactions between 2 actors. For our experiments we consider 5531 utterances including 4 emotions (1103 angry, 1636 happy, 1708 neutral and 1084 sad), where we merge excitement and happiness class into the latter one [5], [6], [9], [10].

We evaluate our proposed feature set under three different SER tasks described next. We also compare our results with the most relevant experimental setups reported in the literature. For all tasks, we report: Weighted Accuracy (WA) which is the percentage of correct classification decisions and Unweighted Accuracy (UA) which is calculated as the average of recall percentage for each emotional class.

After an extensive study of the RQA configuration parameters described in Section 2.2.2, we conclude that best results on SER tasks are obtained using a frame duration of 20 ms for extracting RPs. In addition, the best performing parameters for the RP configuration seem to be a Manhattan norm with a threshold setting depending on a fixed recurrence rate lying in $[0.1, 0.2]$.

4.1. Speaker Dependent (SD)

We evaluate RQA features on SAVEE and Emo-DB following the utterance-based approach described in Section 3. In this setup we apply per-speaker z-normalization (PS-N) and split randomly utterances in train and test sets. Accuracies using 5-fold cross-validation are summarized on Table 2 for the best performing classifier hyper-parameter values.

The fused set achieves significant performance improvement over the baseline IS10 feature set for both datasets. On SAVEE, WA is improved by 3.1% ($77.1\% \rightarrow 80.2\%$) and UA by 3.4% ($74.5\% \rightarrow 77.9\%$). We also achieve an improvement of 4.9% ($88.4\% \rightarrow 93.3\%$) and 5.7% ($87.2\% \rightarrow 92.9\%$) for WA and UA, respectively on Emo-DB. The feature set used in [36] is extracted over cepstral, spectral and prosodic LLDs similar to the ones used in IS10 [1]. Noticeably, they achieve similar performance to ours when we use only IS10 but our fused set with LR outperforms on both Emo-DB (5% in UA and 4.6% in WA) and SAVEE (4.5% in UA and 3.9% in WA). The proposed combination of features and LR also surpasses a Convolutional SAE approach [14] in terms of WA by 5% on Emo-DB and 4.8% on SAVEE. Presumably, RQA measures contain information closely related to speaker-specific emotional dynamics not captured by conventional features.

Table 2: *SD results on SAVEE and Emo-DB. (ESR) Ensemble Softmax Regression*

Features	Model	SAVEE		Emo-DB	
		WA	UA	WA	UA
IS10	SVM	77.1	74.5	88.4	87.2
	LR	74.4	71.8	87.4	86.3
RQA	SVM	66.0	63.0	81.8	80.4
	LR	64.4	61.1	81.9	79.9
RQA+IS10	SVM	77.3	75.5	90.1	88.9
	LR	**80.2**	**77.9**	**93.3**	**92.9**
[14] Spectrogram	SAE	75.4	-	88.3	-
[36] LLDs Stats	ESR	76.3	73.4	88.7	87.9

4.2. Speaker Independent (SI)

Again, we follow the utterance-based approach described in Section 3 on both SAVEE and Emo-DB datasets but we do not make any assumptions for the identity of the user during training. We use leave-one-speaker-out cross validation, where one speaker is kept for testing and the rest for training. The mean and standard deviation are calculated only on training data and used for z-normalization on all data. From now on we refer to

this normalization as Per Fold-Normalization (PF-N). Table 3 presents accuracies averaged over all folds for the best performing classifier hyper-parameter values.

In comparison with the baseline IS10 feature set, the fused feature set obtains an absolute improvement of 5.5% and 8.2% on SAVEE as well as 2.4% and 3.2% on Emo-DB in terms of WA and UA, respectively. Furthermore, our fused set achieves higher performance on SAVEE (3.5% in WA and 4.5% in UA) and slightly lower in Emo-DB compared to [36]. In [37] Weighted Spectral Features based on Hu Moments (WSFHM) are fused with IS10 on utterance-level which is similar to our approach. In direct comparison using the same model (SVM) we surpass the reported performance in terms of WA by 2.5% and 0.4% on SAVEE and Emo-DB, respectively. In addition, both RQA and IS10 sets achieve quite low performance on SAVEE. However, their combination yields an impressive performance improvement of 5.5% ($48.5\% \rightarrow 54.0\%$) in WA and 10.7% ($43.1\% \rightarrow 53.8\%$) in UA over IS10 when we use LR. Our results suggest that RQA measures preserve invariant aspects of nonlinear dynamics occurring in emotional speech and are shared across different speakers.

Table 3: *SI results on SAVEE and Emo-DB. (ESR) Ensemble Softmax Regression*

Features	Model	SAVEE		Emo-DB	
		WA	UA	WA	UA
IS10	SVM	47.5	45.6	79.7	74.3
	LR	48.5	43.1	76.1	71.9
RQA	SVM	45.6	41.1	70.9	64.2
	LR	47.7	42.3	71.1	67.1
RQA+IS10	SVM	52.5	50.6	82.1	76.9
	LR	**54.0**	**53.8**	80.1	77.5
[36] LLDs Stats	ESR	51.5	49.3	**82.4**	**78.7**
[37] WSFHM+IS10	SVM	50.0	-	81.7	-

4.3. Leave One Session Out (LOSO)

In this task, we assume that the test-speaker identity is unknown but we are able to train our model considering other speakers who are recorded in similar conditions. We evaluate on both utterance and segment-based methods (described in Section 3) on IEMOCAP. Given our assumption, we treat each of the 5 sessions as a speaker group [11]. We use LOSO in order to create train and test folds. In each fold, we use 4 sessions for training and the remaining 1 for testing. For the testing session we use one speaker as testing set and the other for tuning the hyper-parameters of our models. We repeat the evaluation by reversing the roles of the two speakers. In the final assessment, we report the average performance obtained in terms of WA and UA obtained from all speakers [5], [6], [10]. In order to be easily comparable with the literature we follow three different normalization schemes. We use the aforementioned PS-N and PF-N schemes as well as Global z-normalization (G-N). In G-N we calculate the global mean and standard deviation from all the available samples in the dataset and perform z-normalization over them. Results on IEMOCAP for the three different normalization schemes are demonstrated on Table 4.

A consistent performance improvement is shown for all combinations of normalization techniques and employed models when the fused set is used instead of IS10. Specifically, for SVM the fused set yields a relative improvement varying from 0.3% to 1.0% in WA and from 0.2% to 0.9% in UA under all normalization strategies. The same applies for LR (in WA from 0.8% to 1.0% and in UA from 0.3% to 1.0%) as well as for A-BLSTM (in WA from 0.1% to 0.7% and in UA from 0.2% to 0.7%). In accordance with our intuition [8], a segment-based approach using A-BLSTM surpasses all utterance-based ones in WA from 3.4% to 8.4% and in UA from 3.8% to 6.8% for all normalization schemes, when the fused set is used.

In [5] low level Mel Filterbank (MFB) features are fed directly to a CNN. In [10] a stacked autoencoder is used to extract feature representations from spectrograms of glottal flow signals and then a BLSTM is used for classification. We surpass both reported results by 0.2% in UA for [5] and by a margin of 8.7% in WA and 8.5% in UA for [10], respectively even with simple models. Compared to a multi-task DBN trained for both discrete emotion classification and for valence-activation in [6], we report 2.0% higher WA and 3.1% higher UA. We also report 4.6% higher UA and 1.9% lower WA compared to CNNs over spectrograms [9]. We assume that this inconsistency in performance metrics occurs because a slightly different experimental setup is followed where the final session is excluded from testing [9].

Table 4: *LOSO results on IEMOCAP. (GFS): Glottal Flow Spectrogram, (SP): Spectrogram.*

Features	Model	PS-N		PF-N		G-N	
		WA	UA	WA	UA	WA	UA
IS10	SVM	58.3	60.9	58.9	60.1	59.2	60.5
	LR	57.5	61.2	54.6	57.9	53.5	57.5
	A-BLSTM	62.0	65.1	62.6	65.0	62.8	65.0
RQA	SVM	52.9	54.6	53.1	53.8	53.1	53.7
	LR	52.2	54.8	52.6	54.0	52.8	54.3
	A-BLSTM	55.6	59.3	56.6	58.3	56.7	58.7
RQA	SVM	59.3	61.8	59.2	60.4	59.5	60.7
+	LR	58.3	62.0	55.6	58.7	54.5	58.7
IS10	A-BLSTM	**62.7**	**65.8**	**63.0**	**65.2**	62.9	**65.5**
[5] MFB	CNN	-	61.8	-	-	-	-
[6] IS10	DBN	-	-	-	-	60.9	62.4
[9] SP	CNN	-	-	-	-	**64.8**	60.9
[10] GFS	BLSTM	-	-	50.5	51.9	-	-

5. Conclusions

We investigated the usage of nonlinear RQA measures extracted from RPs for SER. The effectiveness of these features has been tested under both utterance-based and segment-based approaches across three emotion databases. The fusion of nonlinear and conventional feature sets yields significant performance improvement over traditional feature sets for all SER tasks; the performance improvement is especially large when speaker identity is unknown. The fused data set improves on the state-of-the-art for SER under most testing conditions, classification methods and datasets. Recurrence analysis of speech signals is a promising direction for SER research. In the future, we plan to automatically extract features from RPs using convolutional autoencoders in order to substitute RQA measures.

6. Acknowledgements

This work has been partially supported by the BabyRobot project supported by EU H2020 (grant #687831). Special thanks to Nikolaos Athanasiou and Nikolaos Ellinas for their contributions on the experimental environment setup.

7. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proceedings of INTERSPEECH*, 2010, pp. 2794–2797.

[2] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[4] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2011.

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

[6] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, 2017.

[7] B. Schuller and G. Rigoll, "Timing levels in segment-based speech emotion recognition," in *Proceedings of INTERSPEECH*, 2006.

[8] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 190–195.

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

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

[11] 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.

[12] C.-W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition." in *Proceedings of INTERSPEECH*, 2016, pp. 1387–1391.

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

[14] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[15] H. Herzel, "Bifurcations and chaos in voice signals," *Applied Mechanics Reviews*, vol. 46, no. 7, pp. 399–413, 1993.

[16] V. Pitsikalis and P. Maragos, "Analysis and classification of speech signals by generalized fractal dimension features," *Speech Communication*, vol. 51, no. 12, pp. 1206–1223, 2009.

[17] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *EPL (Europhysics Letters)*, vol. 4, no. 9, p. 973, 1987.

[18] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," *Physics reports*, vol. 438, no. 5-6, pp. 237–329, 2007.

[19] R. Sun and E. Moore, "Investigating glottal parameters and teager energy operators in emotion recognition," in *Affective Computing and Intelligent Interaction*, 2011, pp. 425–434.

[20] T. Chaspari, D. Dimitriadis, and P. Maragos, "Emotion classification of speech using modulation features," in *Proceedings of Signal Processing Conference (EUSIPCO)*, 2014, pp. 1552–1556.

[21] A. Shahzadi, A. Ahmadyfard, A. Harimi, and K. Yaghmaie, "Speech emotion recognition using nonlinear dynamics features," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 23, no. Sup. 1, pp. 2056–2073, 2015.

[22] A. Lombardi, P. Guccione, and C. Guaragnella, "Exploring recurrence properties of vowels for analysis of emotions in speech," *Sensors & Transducers*, vol. 204, no. 9, p. 45, 2016.

[23] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013, pp. 835–838.

[24] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of statistical Physics*, vol. 65, no. 3-4, pp. 579–616, 1991.

[25] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical review A*, vol. 33, no. 2, p. 1134, 1986.

[26] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Phys. Rev. A*, vol. 45, pp. 3403–3411, 1992.

[27] S. Schinkel, O. Dimigen, and N. Marwan, "Selection of recurrence threshold for signal detection," *The european physical journal special topics*, vol. 164, no. 1, pp. 45–53, 2008.

[28] M. Thiel, M. C. Romano, J. Kurths, R. Meucci, E. Allaria, and F. T. Arecchi, "Influence of observational noise on the recurrence quantification analysis," *Physica D: Nonlinear Phenomena*, vol. 171, no. 3, pp. 138–152, 2002.

[29] N. Marwan, J. Kurths, and S. Foerster, "Analysing spatially extended high-dimensional dynamics by recurrence plots," *Physics Letters A*, vol. 379, no. 10, pp. 894 – 900, 2015.

[30] J. F. Donges, J. Heitzig, B. Beronov, M. Wiedermann, J. Runge, Q. Y. Feng, L. Tupikina, V. Stolbova, R. V. Donner, N. Marwan *et al.*, "Unified functional network and nonlinear time series analysis for complex systems science: The pyunicorn package," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, no. 11, p. 113101, 2015.

[31] C. L. Webber Jr and J. P. Zbilut, "Recurrence quantification analysis of nonlinear dynamical systems," *Tutorials in contemporary nonlinear methods for the behavioral sciences*, pp. 26–94, 2005.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[33] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[34] S. Haq and P. Jackson, "Speaker-dependent audio-visual emotion recognition," in *Proceedings Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Norwich, UK*, Sept. 2009.

[35] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[36] Y. Sun and G. Wen, "Ensemble softmax regression model for speech emotion recognition," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 8305–8328, 2017.

[37] Y. Sun, G. Wen, and J. Wang, "Weighted spectral features based on local hu moments for speech emotion recognition," *Biomedical signal processing and control*, vol. 18, pp. 80–90, 2015.