A closed-form solution to the graph total variation problem for continuous emotion profiling in noisy environment

Shaoling Jing, Xia Mao, Lijiang Chen, Maria Colomba Comes, Arianna Mencattini, Grazia Raguso, Fabien Ringeval, Björn Schuller, Corrado Di Natale, Eugenio Martinelli

A Closed-form Solution to the Graph Total Variation Problem for Continuous Emotion Profiling in Noisy Environment

Shaoling Jing, Xia Mao, Lijiang Chen, Maria Colomba Comes, Arianna Mencattini, Grazia Raguso, Fabien Ringeval, Björn Schuller, Corrado Di Natale, Eugenio Martinelli

PII: S0167-6393(18)30142-0
DOI: https://doi.org/10.1016/j.specom.2018.09.006
Reference: SPECOM 2591

To appear in: *Speech Communication*

Received date: 17 April 2018
Revised date: 19 August 2018
Accepted date: 14 September 2018

Please cite this article as: Shaoling Jing, Xia Mao, Lijiang Chen, Maria Colomba Comes, Arianna Mencattini, Grazia Raguso, Fabien Ringeval, Björn Schuller, Corrado Di Natale, Eugenio Martinelli, A Closed-form Solution to the Graph Total Variation Problem for Continuous Emotion Profiling in Noisy Environment, *Speech Communication* (2018), doi: https://doi.org/10.1016/j.specom.2018.09.006

A Closed-form Solution to the Graph Total Variation Problem for Continuous Emotion Profiling in Noisy Environment

Shaoling Jing[a], Xia Mao[a], Lijiang Chen[a], Maria Colomba Comes[b], Arianna Mencattini [b,*], Grazia Raguso[c], Fabien Ringeval[d], Björn Schuller[e,f], Corrado Di Natale[b], Eugenio Martinelli[b]

[a]*School of Electronic and Information Engineering, Beihang University, Beijing 100191, China*
[b]*Department of Electronic Engineering, University of Rome Tor Vergata, Rome 100133, Italy*
[c]*Department of Mathematics, University of Bari Aldo Moro, Bari, 70126, Italy*
[d]*Laboratoire d'Informatique de Grenoble, Universit Grenoble Alpes, Grenoble, France*
[e]*Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany*
[f]*GLAM – Group on Language, Audio & Music, Imperial College London, London, UK*

Abstract

Time-continuous emotion estimation (e. g., arousal and valence) from spontaneous speech expressions has recently drawn increasing commercial attention. However, real-life applications of emotion recognition technology require challenging conditions, such as noise from recording devices and background environments. In this work, we introduce a novel personalized emotion prediction model validated in different noisy environments. It is performed by a three-level noise reduction algorithm: (i) data downsampling, (ii) feature synchronization, and (iii) a modified version of graph total variation. The

*Corresponding author: Department of Electronic Engineering, University of Rome Tor Vergata, Via del Politecnico 1, 00133 Roma.
Email address: mencattini@ing.uniroma2.it (Arianna Mencattini)

approach has been validated on the broadly used RECOLA database with different types of noises, including convolutive and additive noise with different SNRs. The process of feature synchronization improves the concordance correlation coefficient (CCC) absolute values by 0.271 on average for arousal and 0.137 for valence. The proposed denoising approach further improves the values by 0.101 for arousal and 0.086 for valence. Finally, the proposed model considerably improves the CCC values on raw data and all types of noisy data and outperforms the standard denoising methods.

Keywords: continuous emotion profiling from speech, noisy environment, graph total variation denoising

1. Introduction

With the development of artificial intelligence, humans have demanded more and more from affective computing, which facilitates the development of an increasing number of automatic speech emotion recognition (SER) applications and more relevantly dimensional emotion prediction from time-continuous labels Gunes and Schuller (2013); Mencattini et al. (2017); Martinelli et al. (2016); Mariooryad and Busso (2015). Automatic emotion recognition (AER) technology from speech has matured well enough to be applied in some real-life scenarios Vignolo et al. (2016), such as call centers Chen et al. (2012), disease auxiliary diagnosis Schuller et al. (2015), remote education and safe driving. However, these scenarios not only require an almost silent environment to maximize the performance of the system but also need the system to provide the emotional states as accurately as possible. These requirements introduce challenges in emotion recognition from

time-continuous speech, such as reducing the annotation drift (also called Reaction Lag (RL)) Marioryad and Busso (2015) and the environmental noise at the feature level while preserving emotion-related information Chen et al. (2016). These are likely two of the reasons behind the low emotion recognition performance reported in emotion classification studies Meng and Bianchi-Berthouze (2011). Therefore, the goal of this paper is to explore and present a novel personalized emotion prediction model (PEPM), validated in noisy environments. Good performance on SER has been reported in research papers under laboratory environments, but real-life in the wild applications face more complicated conditions, such as reverberation, background noises, and the acoustic properties of the recording devices used. These noises severely degrade the performance of systems and consequently affect the user experience in real-life conditions Tawari and Trivedi (2010); Schuller et al. (2006); Huang et al. (2013); Schuller et al. (2011). Therefore, a fundamental step in this work is to investigate environmental noise reduction via a novel closed-form solution to the graph signal theory-based method, to reduce noise at features level and improve the performance of emotion prediction. Beyond the field of emotion recognition, speech signal is crucial for many applicative contexts: voice activity detection Ariav et al. (2018), indoor speaker recognition D'Arca et al. (2016), vocal folds damage detection Zhong et al. (2016), to mention but a few. Such a widespread applicability has increased the interest towards speech signal denoising in the research community. Many denoising techniques have been proposed for AER systems Liu et al. (2013). Feature enhancement via recurrent neural networks (RNNs) Xia and Bao (2013) and blind source separation via

RNNs Zhang et al. (2014) are two common methods to reduce noises in ASR. To the best of our knowledge, only a few studies addressed the issue of noise and adverse acoustic conditions for automatic speech emotion recognition especially for time-continuous dimensional emotion prediction Zhang et al. (2016); Trentin et al. (2015). For example, in Schuller et al. (2006), the authors first studied the affect estimation under noise conditions. They applied a fast information gain ratio-based feature selection method to select relevant features from a large acoustic feature set. The results indicated that automatic speech emotion suffers from influence of noisy conditions. In Zhang et al. (2018) supervised single-channel technique is applied to speech dereverberation and denoising. In Tawari and Trivedi (2010), the authors utilized a speech enhancement technique based on the adaptive thresholding in the wavelet domain to address noises while in Huang et al. (2013), the authors studied the influence of additive white Gaussian noise on speakers emotion states via a Gaussian mixture model. However, these methods are performed on discrete emotion conditions. The investigations on dimensional emotion states are sparse, and still a challenging work.

1.1. Main Contribution

The intention of our work is to produce a model that can predict dimensional emotion under various noise conditions. Algorithm 1 summarizes the scheme of the proposed PEPM. First, we propose a three-level noise reduction algorithm consisting of feature down-sampling (the first level), feature synchronization (SYNC, the second level), and a graph total variation regularization (GTVR, the third level) (steps 1-3 in Algorithm 1). Second, we represent the acoustic features as a graph signal at the feature level (steps 4

Algorithm 1 Personalized emotion prediction model (PEPM) in noisy environments

1: Acoustic feature extraction .
2: Pre-processing with downsampling .
3: Reduce the annotation drift via feature synchronization, which calculates the optimal delay lag τ_{opt}^{g} and obtains the synchronized feature matrices .
4: Represent the acoustic features as a graph signal .
5: Define the weighted adjacency matrix $A_{m,n}$ in the graph .
6: Set the optimization formula of the graph signal denoising via graph total variation regularization and obtain the solution .
7: Process the training data and testing data by the solution in step 6 .
8: Feature normalization by Z-scores .
9: Predict the emotional level by Partial Least Square Regression (PLSR) .

and 5 in Algorithm 1). We investigate the utilization of graph signal theory to reduce environmental noise at the feature level. Third, we construct the optimization formula of graph signal denoising by means of GTVR and obtain a closed-form solution (steps 6 and 7 in Algorithm 1). Fourth, a partial least square regression (PLSR) model is trained on the Z-score normalized features and tested on an independent acoustic features partition of the same subject (steps 8 and 9 in Algorithm 1). We perform the model evaluation on a widely used suited spontaneous emotional speech database to compare the performance of our approach with the state-of-the-art methods in a well reproducible and standardized way.

The remainder of the paper is organized as follows. In Section 2, we

introduce the database and the features extracted. In Section 3, we analyze the main procedures in the proposed PEPM and in Section 4, we validate the effectiveness of the proposed denoising method. Finally, discussions and further experiments are included in Section 5 and some conclusive remarks are given in Section 6.

2. Materials

2.1. Database

Our experiment is conducted on the REmote COLlaborative and Affective (RECOLA) database Ringeval et al. (2013), which was recently used for the 6th Audio Visual Emotion Challenge (AVEC 2016) Valstar et al. (2016). The database includes 46 speech sequences, each with five minutes of time-continuous annotations. In order to generate a noisy version of the RECOLA database, we exploited the Audio Degradation Toolbox (ADT) Mauch and Ewert (2013) with various data sets. The overall procedure is summarized in Fig. 1. First, we simulated a recording with the microphone of a smartphone (Google Nexus One ©), using the microphone impulse response (MIR) provided with the ADT toolbox Mauch and Ewert (2013). To study the impact of reverberation noise, we further convolved the obtained signal with different types of room impulse responses (RIR) Xia and Bao (2013), as if the speakers were talking in various (rather large) rooms with a smartphone. The collection of the RIR we used was measured in the Great Hall (multipurpose hall), the Octagon (Victorian building completed with height walls), and a classroom (typical university lecture room) at the Mile End campus of Queen Mary, University of London. In parallel to RIR, we exploited additional

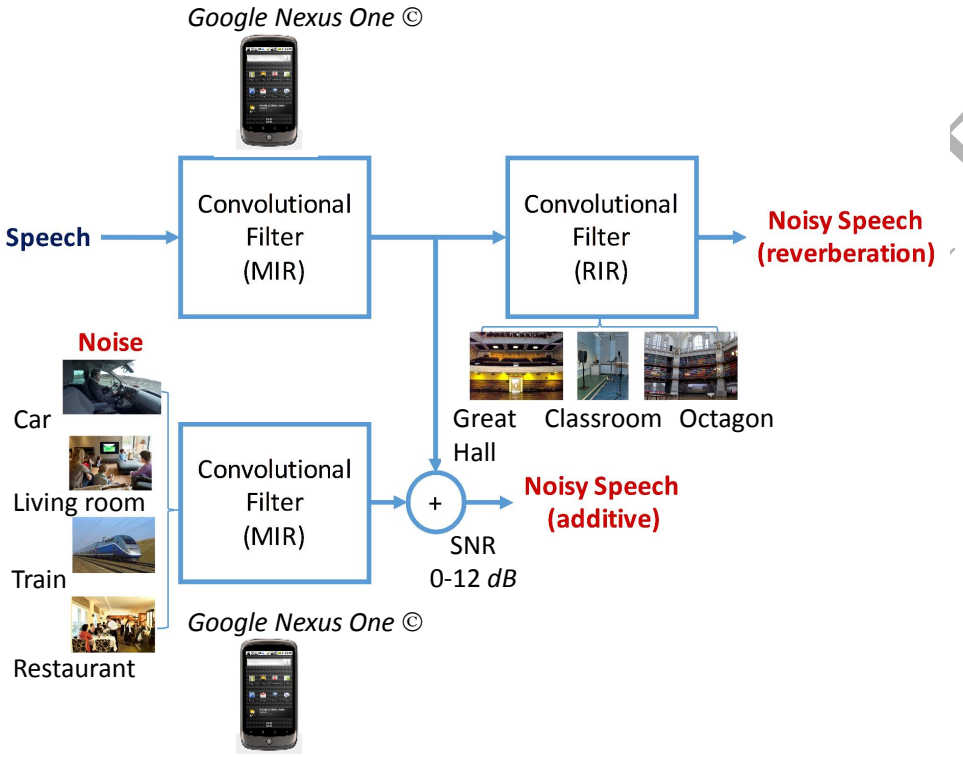

Figure 1: Flowchart of the speech degradation procedure: reverberative (Great Hall, Classroom, Octagon) and additive noises (Car, Living room, Train, Restaurant) are combined with speech recordings as if they were recorded on a smartphone.

datasets to include additive noise, after having been convolved with the MIR of the smartphone using the CHiME corpus Barker et al. (2013), which contains recordings made in domestic environments. Furthermore, other types of noises (car, train, crowed restaurant) with signal-to-noise ratio ranging from 0 dB to 12 dB with a step of 3 dB were investigated.

2.2. Feature extraction

A manifold of information can be extracted from the speech signal, such as features related to the spectrum, voice quality, pitch, loudness and duration. Smaller specific sets of Low Level Descriptors (LLDs) have recently proven to be useful as they can be computed with close to real-time capabilities and even provide better performance Ringeval et al. (2016). In this study, we combined two expert-knowledge feature sets that have shown robustness in emotional speech recognition: eGeMAPS and MFCCs. The eGeMAPS feature set Eyben et al. (2016) includes 25 measures covering loudness, spectrum, pitch, and the first three formants. Features were computed with the open-source openSMILE feature extractor Eyben et al. (2013).

3. Personalized emotion prediction model (PEPM)

3.1. Pre-processing

Each speech sequence from RECOLA dataset includes about 7500 values stored at a frame rate of 40 ms for a total length of 300 s (5 min). After being down-sampled by a factor of 10, the size of the data and the computational time are considerably reduced in the subsequent process, making it possible as an online emotion estimation.

3.2. Annotation drift reduction

To reduce the annotation drift, as already proposed in Mencattini et al. (2017); Martinelli et al. (2016), we use a feature synchronization (SYNC) procedure. Assuming a different RL for each speaker, here, we perform synchronization of each feature sequence with respect to the corresponding output

annotations (for arousal and valence separately). Results show an average RL of 3.75 s (σ=2.59 s) for arousal and 8.75 s (σ=6.45 s) for valence. RLs and standard variations vary with the signal-to-noise ratio. The evaluators have a higher reaction time for valence than arousal during annotation, which is consistent with the experimental results reported in the literature Mencattini et al. (2017). The lower RL values reported in Mencattini et al. (2017) are motivated by the reduced number of speakers used for those experiments (23 vs 46 in the present experiment) and an extended set of acoustic, spectral, and intensity features considered.

3.3. Environmental noise reduction

In Sandryhaila and Moura (2013, 2014); Chen et al. (2014) authors referred to data indexed by the nodes of the graph as the graph signal, and they proposed the theory of discrete signal processing on graphs (DSPG). We briefly review relevant concepts of DSPG in the following sections in order to introduce the novel graph denoising algorithm proposed herein.

3.3.1. Feature representation

Our representation method is based on two fundamental theories intersecting each other: Total Variation Denoising Rudin et al. (1992) and Discrete Signal Processing on Graphs (DSPG) Chen et al. (2014). With the term Total Variation Denoising, we refer to a regularizing criterion with the goal of eliminating spurious details from the noisy signal so that it will be close to the original signal, preserving its meaningful details. On the other hand, DSPG allows to represent the structure of a signal with a graph G, defined by an ordered couple ($\mathcal{V}$, $\mathbf{A}$), where $\mathcal{V}$ is the set of nodes of the graphs, and

A is called the graph shift or also weighted adjacency matrix, representing connections between nodes (edges). Unlike DSPG presented in Sandryhaila and Moura (2013, 2014); Chen et al. (2014), we built an individual graph for each speaker: in total, we had 46 graphs. The reason lies in the need to perform personalized emotional predictions, that is crucial in many different clinical as well as customer satisfaction scenarios. In addition, each acoustic feature of different speakers will show different characteristics, i.e., different time constants depending on which is the information it is capturing, and the optimal adjacency matrix A may be different for each feature. The graph architecture is used to represent the temporal behaviour of the temporal feature vector as well as the degree of similarity/dependency expected at distinct time instants. Assume that the extracted feature sequences are represented by graph signals and that each acoustic feature sequence of length N is written as the vectors $f = (f_1, f_2, \cdot, f_N)^T$. Note that each element f_m is indexed by the node v_m of a given representation graph $G = (\mathcal{V}, \mathbf{A})$. K denotes the nearest neighbour size in the graph with nodes representing features. Each node is connected to its K closest neighbours. When the $N * N$-dimensional adjacency matrix A is used on $N * 1$ (one-dimensional) feature vector, a new $N * 1$-dimensional feature is formed, when A is applied to $N * N_f$ (N_f-dimension) feature matrix, a new $N * N_f$-dimensional feature matrix is formed. Therefore, in this representation, the feature vector collected at time t_n is associated to the node v_n, whereas the element $A_{m,n}$ represents the edge from node v_n to node v_m (feature vector collected at time t_m). It quantifies the desired degree of relationship, similarity or dependency, between feature vectors collected at times t_n and t_m. More specifically, given

a number K (set in the range $1 \leq K \leq 20$), we define the generic weight element $A_{m,n} = P_{m,n}/K$, as

$$P_{m,n} = \begin{cases} K, \text{ if } m = 1,\ n = 1 \\ 1, \text{ if } m - n \equiv t\, mod\, N,\ t = 1, 2, \ldots, K \\ 0, \text{ otherwise.} \end{cases} \quad (1)$$

If, for example, we consider the two situations, $K = 1$, $N = 4$ and $K = 2$, $N = 4$, we will extract the adjacency matrices with associated graphs illustrated in Fig. 2. Note that, a change in the matrix coefficients implies a change in the associated graph architecture and related edges. Moreover,

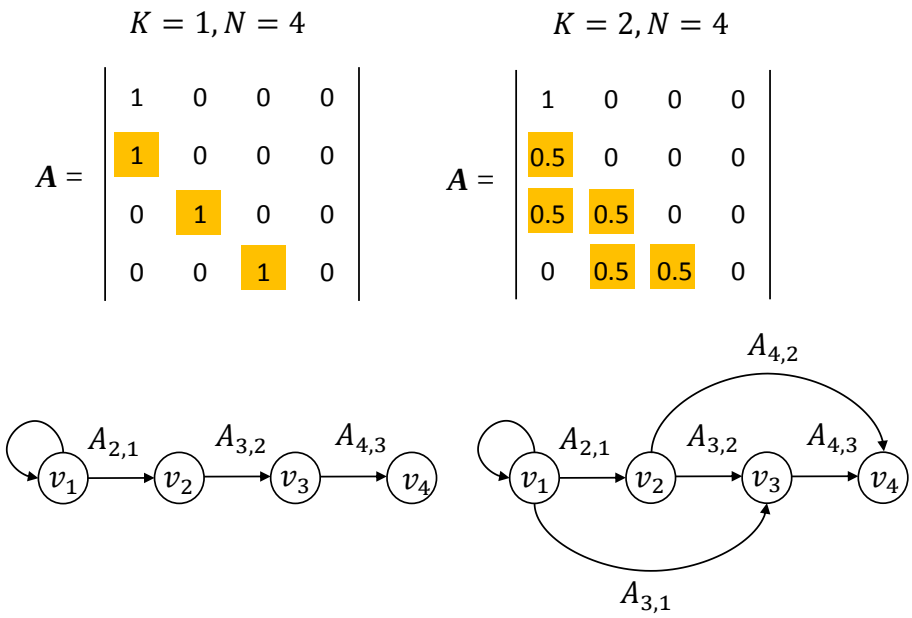

Figure 2: Two examples of adjacency matrix and related graphs: (left) $K = 1$ for $N = 4$ and (right) $K = 2$ for $N = 4$. Coloured elements indicate the non-zero elements, hence, the edge values in the graph.

by applying matrix $\mathbf{A}$ to a signal x, we obtain a new signal as a particular linear combination of its element values, i.e., $\mathbf{A}x$, therefore, we also call $\mathbf{A}$ as graph shift Sandryhaila and Moura (2013). According to this scenario, the denoising problem is formulated as an optimization problem Chen et al. (2014). Let us consider a signal in the DSPG representation (graph signal), affected by noise (noisy signal graph)

$$s = x + w, \tag{2}$$

where x is the original signal, without noise, and w the noise term. The goal is to rebuild the original signal x from s, reducing the randomly distributed noise term w. Denoising for graph signals is formulated as the following optimization problem:

$$\tilde{x} = \arg\min \left(\frac{1}{2} \|x - s\|_2^2 + \alpha S_2(x) \right). \tag{3}$$

The first term is a term of signal approximation, the second term is the quadratic form of total variation which refers to the smoothness Sandryhaila and Moura (2014). Its explicit form is

$$S_2(x) = \frac{1}{2} \left\| x - \frac{1}{\lambda_{max}(\mathbf{A})} \mathbf{A}x \right\|^2, \tag{4}$$

where $\lambda_{max}(\mathbf{A})$ is the largest magnitude eigenvalue of $\mathbf{A}$, here normalized to one for simplicity. The parameter α, called regularization parameter, controls the trade-off between the two objective function components. The objective function is a linear combination of two quadratic functions in x. Calculating and setting the derivative to zero Chen et al. (2014), we get the exact solution in a closed form

$$\tilde{x} = \left[\mathbf{I} + \alpha \left(\mathbf{I} - \mathbf{A} \right)^* \left(\mathbf{I} - \mathbf{A} \right) \right]^{-1} s \tag{5}$$

from

$$[\mathbf{I} + \alpha\,(\mathbf{I} - \mathbf{A})^{*}\,(\mathbf{I} - \mathbf{A})]\tilde{x} = s. \tag{6}$$

The implementation of this method is called Graph Total Variation (GTV)). The term denoising indicates the noise elimination or reduction to recover the original signal, without noise, from the one affected by noise. In the literature there are different approaches, among which smoothing with a low-pass filter, moving average filter, and Wavelet Denoising Barclay and Bonner (1997). In this work, we present an approach based on the representation of signals as graphs Sandryhaila and Moura (2013, 2014); Chen et al. (2014), which is entirely innovative in the field of SER. In fact, in the proposed work such approach is applied to the processing of audio signals recorded from speakers of the RECOLA database. Each speech feature signal is schematically shown as a graph, and on it we apply a parametric denoising method by GTV: the method is characterized by a penalty parameter α. In the original approach Sandryhaila and Moura (2014); Chen et al. (2014), α is defined a priori as an input. However, we present a closed-form method for the automatic optimization of the α parameter. That choice of α parameter determines the optimal K value for the construction of the adiacency matrix $\mathbf{A}$.

3.3.2. Parameter choice method: Graph Generalized Cross-Validation

Beyond the innovative application context of the GTV method considered in this work, we present here a novel closed-form to automatize the regularization parameter α: it is an adaptation on graphs of the known Generalized Cross-Validation (GCV) method. Indeed, there is a parallelism between the denoising optimization problem on graphs and the problem at the basis of

the Tikhonov regularization method O'Leary (2001); Golub et al. (1979). The latter problem assumes the following formulation

$$\min_x \left(\|\mathbf{C}x - b\|_2^2 + \alpha \|\mathbf{L}x\|_2^2 \right), \tag{7}$$

while the former is expressed as in Eq. (7). The two equations Eq. (3) and Eq. (7) are comparable term by term except for a multiplicative factor 1/2: for the first term, the following substitutions are needed: $\mathbf{C} \to \mathbf{I}$, $b \to s$, where b represents data affected by noise, and s is the noisy signal input. Regarding the second term, they are both derivative discrete operators. In fact, $S_2(x)$ is by definition Sandryhaila and Moura (2014):

$$S_2(x) = \frac{1}{2} \|x - \mathbf{A}x\|_2^2 = \frac{1}{2} \|(\mathbf{I} - \mathbf{A})x\|_2^2 . \tag{8}$$

Then, we rewrite Eq. (1) as follows:

$$\tilde{x} = \underset{x}{\operatorname{argmin}} \left(\frac{1}{2} \|x - s\|_2^2 + \alpha \frac{1}{2} \|(\mathbf{I} - \mathbf{A})x\|_2^2 \right). \tag{9}$$

Comparing the second terms of Eq. (7) and of Eq. (9), it appears immediately that $\mathbf{L} \to (\mathbf{I} - \mathbf{A})$. Then, α can be determined as a minimum of the functional

$$G(\alpha) = \frac{(\mathbf{I} - \mathbf{A}_\alpha) f}{(trace\,(\mathbf{I} - \mathbf{A}_\alpha))^2}, \tag{10}$$

where $\mathbf{A}_\alpha = (\mathbf{I} + \alpha\,(\mathbf{I} - \mathbf{A}) * (\mathbf{I} - \mathbf{A}))^{-1}$, symbol $*$ denotes matrix multiplication, and $trace\,(\mathbf{I} - \mathbf{A}_\alpha)$ denotes the sum of the elements on the diagonal of the matrix $(\mathbf{I} - \mathbf{A}_\alpha)$. We call this method Graph Generalized Cross-Validation (GGCV). All the properties of the GCV method are true also in this case. The combined technique proposed here is a closed form solution to the problem of feature denoising before emotion prediction: the first

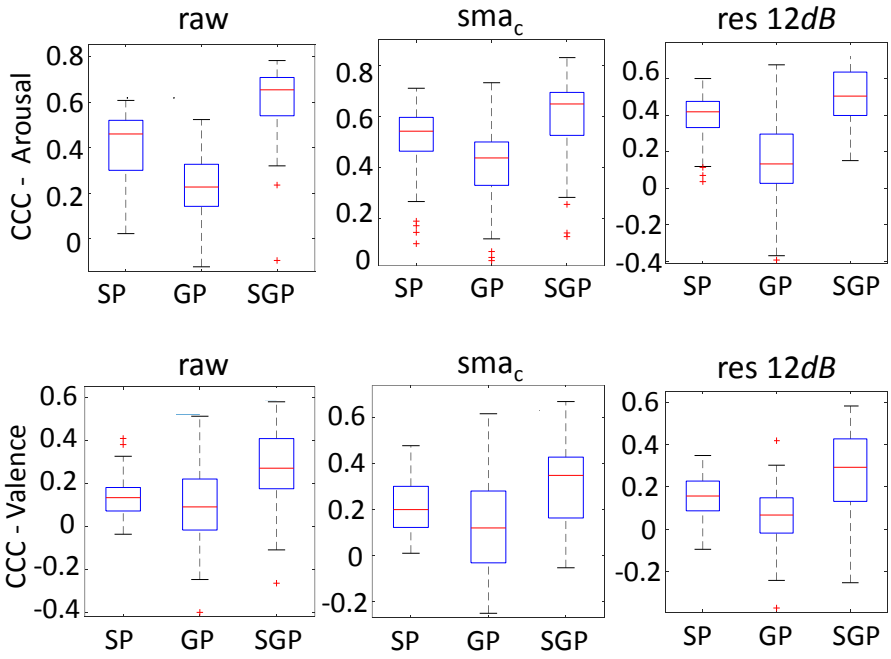

Figure 3: Box-plots of CCC values obtained by using SYNC (labelled as SP), the graph denoising procedure (labelled as GP), and combining these two procedures (labelled as SGP) for arousal (top) and valence (bottom) on raw and noisy data under the conditions of a classroom (sma$_c$) and restaurant with a 12-dB SNR (res 12 dB).

part (GTV), is a parametric method based on concepts of graph and total variation, while the second part (GGCV), is an automatic parameter choice method that selects the penalty parameter α for each subject. Hence, personalized emotion prediction is achieved by profiling the way the system is tuned.

3.4. Feature normalization and regression

The denoised features are normalized by Z-score. A partial least square regression (PLSR) is trained on denoised features to predict arousal and

valence.

3.5. Experimental set-up

The proposed algorithm has been validated on the RECOLA database. Each l-length speech sequence was divided into an $l/2$-length training set and $l/2$-length testing set. To be robust in generalization, we carry out experiments on *raw* data (without adding noise) and noisy data under different noise conditions: smartphone (sma), smartphone_classroom (sma_c), smartphone_hall (sma_h), and smartphone_octagon (sma_o), respectively. Regarding additive noise, we denote as *CHiME*, *cars*, *trains*, and *restaurants* the corresponding noisy feature signal. Performance of the PEPM Mencattini et al. (2017) was evaluated through the concordance correlation coefficient (CCC), the official benchmark measure used in Valstar et al. (2016), denoted by ρ_{ccc}:

$$\rho_{ccc}\left(y_1, y_2\right) = \frac{\rho_{cc}(y_1, y_2)\sigma_{y_1}\sigma_{y_2}}{\sigma_{y_1}^2 + \sigma_{y_2}^2 + \left(\mu_{y_1} - \mu_{y_2}\right)^2}, \tag{11}$$

where ρ_{cc} is the Pearson's correlation coefficient (CC), y_1 and y_2 are two comparative sequences with μ and σ indicating their average and standard deviation values.

4. Results

In order to validate the proposed PEPM we compare three distinct approaches: the first comparative method, labelled as SP, is structured into three distinct steps: down-sampling + SYNC + PLSR; a second comparative approach, labelled as GP, is composed by down-sampling + GTVR + PLSR, and finally the proposed PEPM method, labelled as SGP, is based

on down-sampling + SYNC + GTVR + PLSR. Fig. 3 shows the comparative performance obtained by the three approaches on raw data and different types of noisy data for arousal and valence regression. Results demonstrate the importance of GTVR (comparison SP vs SGP) and of synchronization (comparison GP vs SGP). In order to verify that this statistical comparison is meaningful, we perform a t-test. The returned H and p values are used to evaluate its usefulness. Table 1 lists the p-values of the two-by-two comparative t-tests run on the CCC values for arousal and valence prediction in case of raw, sma_c, and res_{12dB}.

Table 1: p-values of the two-by-two t-tests run on the three methods SP, GP, and SGP for arousal and valence prediction. *ns* stands for not statistically significant at 0.05 confidence level.

Arousal	raw	sma_c	res_{12db}
p_{SP-GP}	< 1.0e-6	< 0.005	< 1.0e-7
p_{SP-SGP}	< 1.0e-7	< 0.005	<0.0005
p_{GP-SGP}	< 1.0e-17	< 1.0e-7	< 1.0e-12
Valence	**raw**	sma_c	res_{12dB}
p_{SP-GP}	ns	< 0.01	< 0.005
p_{SP-SGP}	< 1.0e-4	< 0.001	< 0.0001
p_{GP-SGP}	< 5.0e-5	<5.0e-6	< 1.0e-6

Figure 4 shows two examples of continuous emotion prediction of arousal (top) and valence (bottom) for subjects P60 and P53, respectively. CCC values of 0.83 and 0.67 are achieved.

From Fig. 3, Tab. 1 and Fig. 4, we can conclude the following points: i) the proposed PEPM not only improves the predictive performance on noisy data

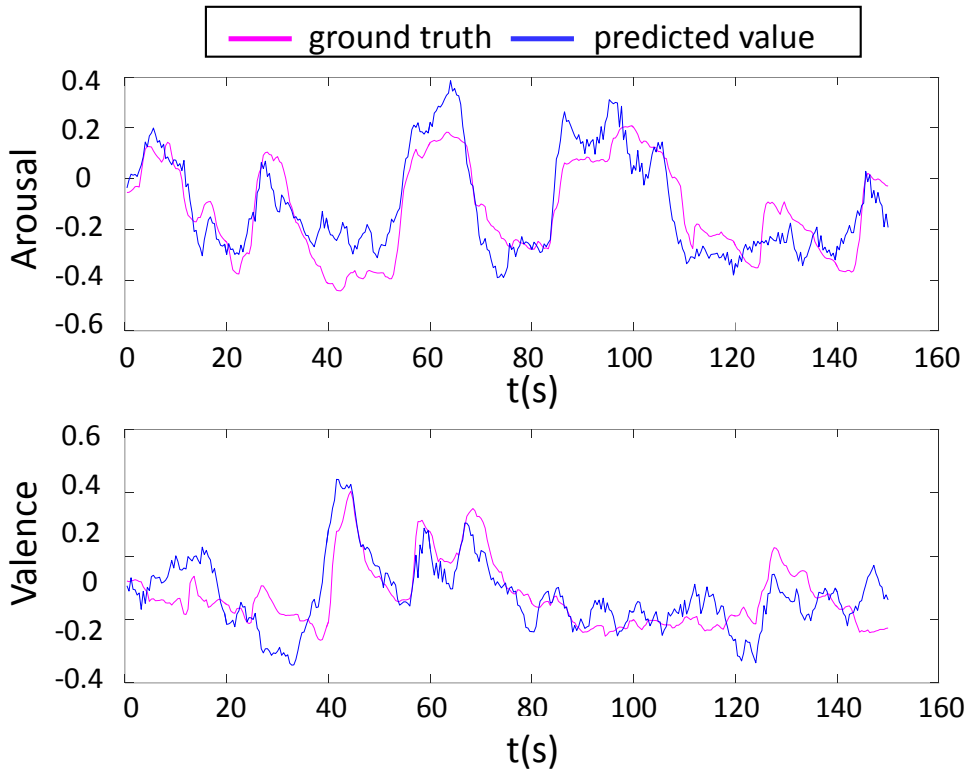

Figure 4: Automatic prediction of arousal (top) and valence (bottom) obtained frame by frame for subject P60 for arousal (CCC=0.83) and P53 for valence (CCC=0.67).

but also shows the higher improvement on raw data, which provides us an opportunity to obtain a more accurate prediction also in controlled scenarios; ii) compared with the use of SYNC or GTVR alone, the combination of SYNC and GTVR (SGP) can improve the system's ability to predict emotions, which illustrates that it is important to reduce the annotation drift and environmental noise before predicting emotions; iii) the p-values of a paired t-test between the CCC values obtained on those three approaches show that the combined approach outperforms the performance reached by each single component; iv) in most of the cases, we obtain $p < 0.001$ between pairwise comparative methods for each dimension, verifying the effectiveness of the GTVR method. Moreover, the results illustrate that the proposed GTVR method is more than a mere denoising approach, but rather a signal approximator able to extract the relevant information from the underlying signal; v) consistently with the literature Mencattini et al. (2017), the results show that arousal is better identified than valence. As an example, Fig. 4 shows predicted arousal for speaker P60 and predicted valence for speaker P53. Very high CCC values of 0.831 for P60 and of 0.671 for P53 have been obtained.

5. Discussion

5.1. Comparison with standard approaches

We further compare the proposed method with alternative noise reduction approaches, such as standard smooth processing and Long Short-Term Memory (LSTM)-based neural network method proposed in Zhang et al. (2016). The comparative approaches have been labelled as SSP and LSTM

respectively. Fig. 5 shows the performance obtained by using the SSP method against the proposed PEPM on raw data and different types of noisy data for arousal and valence prediction. Furthermore, in Table 2 we also reported the average CCC values obtained using LSTM approach for arousal and valence prediction. Average values and standard deviation were computed over convolutive and additive noises separately. Results show that the proposed PEPM significantly outperforms the two comparative approaches, especially in presence of convolutive noises. The improvement is also stronger in valence than in arousal. This is crucial since valence prediction represents a challenging emotion dimension, recently successfully used in very critical clinical applications Mencattini et al. (2018).

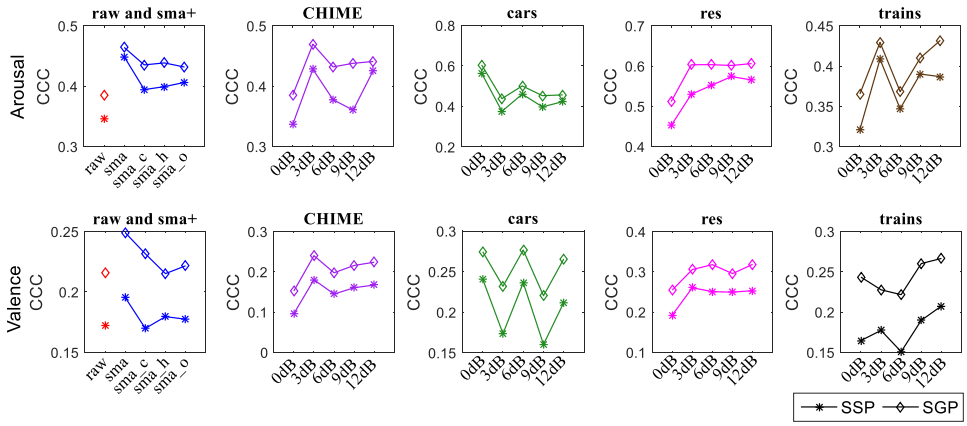

Figure 5: Average CCC values for arousal and valence obtained by comparing the PEPM (SGP) method with the smooth denoising method (SSP).

The generalization capability of the proposed PEPM based on PLSR is additionally validated against an alternative popular regression method based on support vector regression (SVR) Grimm et al. (2007)) with the default

Table 2: Average CCC values and standard deviation computed over convolutive (first two columns) and additive (second two columns) noise contributions, using PEPM against LSTM approach.

	Convolutive		Additive	
	avg CCC	std CCC	avg CCC	std CCC
Arousal - PEPM	0.705	0.014	0.631	0.026
Arousal - LSTM	0.674	0.045	0.626	0.069
Valence - PEPM	0.418	0.004	0.216	0.049
Valence - LSTM	0.232	0.047	0.244	0.035

settings, i.e., a complexity value of $C = 1$ and a Gaussian kernel with $\sigma = \frac{1}{N_f}$, where N_f denotes the number of features. The results are reported in Fig. 6. The two methods show almost the same performance, but the training time of PLSR is shorter than that of SVR, thus confirming the robustness of the proposed architecture.

The proposed method performs better than baselines in the noise-reduction due to the characteristics of the neighborhood matrix A. The weights $A_{m,n}$ are not restricted to being non-negative reals, and they can be arbitrary real or complex values leading to a very flexible weighting procedure among adjacent signal values. Additionally, the weights $A_{m,n}$ can form either a symmetric or asymmetric matrix allowing for a non-trivial nodes combination with respect to standard weighted averaging.

5.2. Training-Testing partition effect

Furthermore, to study the impact of the training set dimension on the model performance, we conducted experiments by progressively decreasing

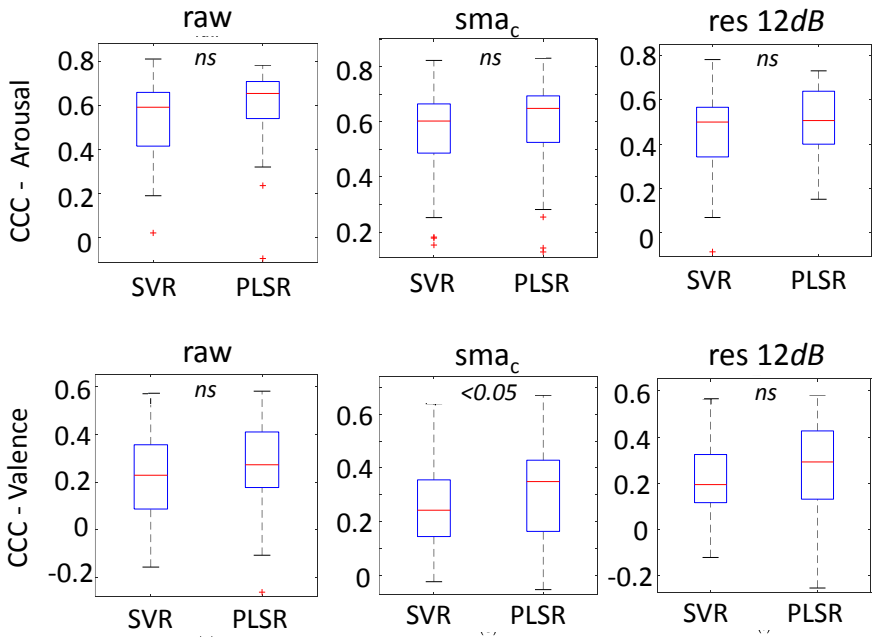

Figure 6: Box-plots of the CCC values on raw data and two types of noisy data for the proposed SGP method using PLSR compared with SVR. *ns* stands for non statistically significant at 0.05 confidence level.

training set partition dimension. As expected, in most of the cases, the performance increases with the growth of the training set. However, the results shown in Fig. 7, indicate that the proposed PEPM still achieves good performance even with a reduced training set.

6. Conclusion

Taking into account that noise seriously degrades the performance of speech emotion prediction in real-life applications, in this paper we focus on the noise reduction method. This study analyzed and proposed a per-

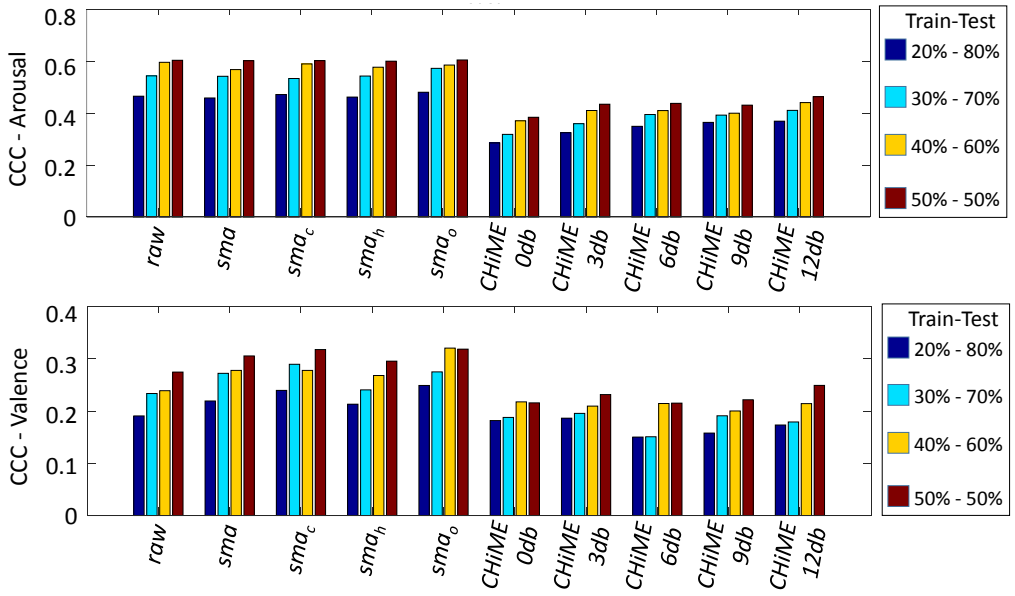

Figure 7: Average CCC values for arousal (left) and valence (right) obtained by changing training-testing partitions.

sonalized emotion prediction model based on a three-level noise reduction algorithm. The algorithm is composed of down-sampling, input-output synchronization, and GTVR. The novel denoising approach, GTVR, is proposed to reduce different types of environmental noise. Finally, the method is applied on raw data and different types of noisy data. The results indicate that the proposed GTVR method not only improves the performance of emotion prediction on noisy data but also yields higher CCC values on raw data. Furthermore, the performance of prediction by the proposed PEPM significantly outperforms the state of the art approaches. Future efforts will focus on the combination of the proposed approaches with other efficient denoising

methods including data from paralinguistic tasks.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China [grant numbers 61603013] and by PainTCare project (University of Rome Tor Vergata, Uncovering Excellence program).

References

Ariav, I., Dov, D., Cohen, I., 2018. A deep architecture for audio-visual voice activity detection in the presence of transients. Signal Processing 142, 69–74.

Barclay, V.J., Bonner, R.F., 1997. Application of Wavelet Transforms to Experimental Spectra: Smoothing, Denoising, and Data set Compression. Analytical Chemistry 69, 78–90.

Barker, J., Vincent, E., Ma, N., et al., 2013. The PASCAL CHiME speech separation and recognition challenge. Computer Speech and Language 27, 621–633.

Chen, L., Mao, X., Xue, Y., et al., 2012. Speech Emotion Recognition: Features and Classification Models. Digital Signal Processing 22, 1154–1160.

Chen, L., Mao, X., Yan, H., 2016. Text-Independent Phoneme Segmentation Combining EGG and Speech Data. IEEE/ACM Transactions on Audio Speech & Language Processing 24, 1029–1037.

Chen, S., Sandryhaila, A., Moura, J.M.F., et al., 2014. Signal Denoising on Graphs via Graph Filtering, in: IEEE Global Conference on Signal and Information Processing (GlobalSIP 2014), IEEE, Atlanta, GA, USA. pp. 872–876.

D'Arca, E., Robertson, N., Hopgood, J., 2016. Robust indoor speaker recognition in a network of audio and video sensors. Signal Processing 129, 137–149.

Eyben, F., Scherer, K., Schuller, B., et al., 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing 7, 190–202.

Eyben, F., Weninger, F., Groß, F., et al., 2013. Recent developments in openSMILE, the Munich open-source multimedia feature extractor, in: Proc. of the 21st ACM International Conference on Multimedia (ACM MM), ACM, Barcelona, Spain. pp. 835–838.

Golub, G.H., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21, 215–223.

Grimm, M., Kroschel, K., Narayanan, S., 2007. Support Vector Regression for Automatic Recognition of Spontaneous Emotions in Speech, in: Proc. of the 32th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), Honolulu, HI, USA. pp. 1085–1088.

Gunes, H., Schuller, B., 2013. Categorical and Dimensional Affect Analysis in Continuous Input: Current Trends and Future Directions. Image and

Vision Computing, Special Issue on Affect Analysis in Continuous Input 31, 120–136.

Huang, C., Chen, G., Yu, H., et al., 2013. Speech Emotion Recognition under White Noise. Archives of Acoustics 38, 457–463.

Liu, Y., Xiao, M., Tie, Y., 2013. A Noise Reduction Method Based on LMS Adaptive Filter of Audio Signals, in: Proc. of the 3rd International Conference on Multimedia Technology (ICMT 2013), Springer, Budapest, Hungary. pp. 1001–1008.

Mariooryad, S., Busso, C., 2015. Correcting Time-Continuous Emotional Labels by Modeling the Reaction Lag of Evaluators. IEEE Transactions on Affective Computing 6, 97–108.

Martinelli, E., Mencattini, A., Daprati, E., et al., 2016. Strength is in Numbers: Can Concordant Artificial Listeners Improve Prediction of Emotion from Speech? PLoS ONE 11. E0161752.

Mauch, M., Ewert, S., 2013. The Audio Degradation Toolbox and its Application to Robustness Evaluation, in: Proc. of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013), Curitiba, Brazil. pp. 83–88.

Mencattini, A., Martinelli, E., Ringeval, F., et al., 2017. Continuous Estimation of Emotions in Speech by Dynamic Cooperative Speaker Models. IEEE Transactions on Affective Computing 8, 314–327.

Mencattini, A., Mosciano, F., Comes, M., De Gregorio, T., Raguso, G., Daprati, E., Ringeval, F., Schuller, B., Martinelli, E., 2018. An emotional

modulation model as signature for the identification of children developmental disorders. Scientific Reports .

Meng, H., Bianchi-Berthouze, N., 2011. Naturalistic Affective Expression Classification by a Multi-stage Approach Based on Hidden Markov Models, in: Proc. 4th biannual Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2011), Springer Berlin Heidelberg, Memphis, TN, USA. pp. 378–387.

O'Leary, D., 2001. Near-optimal parameters for tikhonov and other regularization methods. SIAM Journal on scientific computing 23, 1161–1171.

Ringeval, F., Marchi, E., Grossard, C., et al., 2016. Automatic Analysis of Typical and Atypical Encoding of Spontaneous Emotion in the Voice of Children, in: INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, ISCA, San Fransisco (CA), USA. pp. 1210–1214.

Ringeval, F., Sonderegger, A., Sauer, J., et al., 2013. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions, in: Proc. of the 2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013), IEEE, Shanghai, China.

Rudin, L.I., Osher, S., Fatemi, E., 1992. Nonlinear Total Variation based Noise Removal Algorithms. Physica D: Nonlinear Phenomena 60, 259–268.

Sandryhaila, A., Moura, J.M.F., 2013. Discrete Signal Processing on Graphs: Graph Filters. IEEE Transactions on Signal Processing 61, 1644–1656.

Sandryhaila, A., Moura, J.M.F., 2014. Discrete Signal Processing on Graphs: Frequency Analysis. IEEE Transactions on Signal Processing 62, 3042–3054.

Schuller, B., Arsić, D., Wallhoff, F., et al., 2006. Emotion Recognition in the Noise Applying Large Acoustic Feature Sets, in: Proc. 3rd International Conference on Speech Prosody, SP 2006, ISCA. ISCA, Dresden, Germany. pp. 276–289.

Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication 53, 1062–1087.

Schuller, B., Marchi, E., Baron-Cohen, S., et al., 2015. Recent developments and results of ASC-Inclusion: An Integrated Internet-Based Environment for Social Inclusion of Children with Autism Spectrum Conditions, in: Proc. of the of the 3rd International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI 2015), Atlanta, GA.

Tawari, A., Trivedi, M.M., 2010. Speech Emotion Analysis in Noisy Real-World Environment, in: Proc. of the 20th International Conference on Pattern Recognition (ICPR), IEEE, Istanbul, Turkey. pp. 4605–4608.

Trentin, E., Scherer, S., Schwenker, F., 2015. Emotion Recognition from Speech Signals via a Probabilistic Echo-State Network. Pattern Recognition Letters 66, 4–12.

Valstar, M., Gratch, J., Schuller, B., et al., 2016. AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge, in: Proc. of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC'16, co-located with the 24th ACM International Conference on Multimedia, MM 2016, ACM, Amsterdam, The Netherlands. pp. 3–10.

Vignolo, L.D., Prasanna, S.R.M., Dandapat, S., et al., 2016. Feature Optimisation for Stress Recognition in Speech. Pattern Recognition Letters 84, 1–7.

Xia, B.Y., Bao, C.C., 2013. Speech Enhancement with Weighted Denoising Autoencoder, in: Proc. of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013), ISCA, Lyon, France. pp. 3444–3448.

Zhang, L., Xu, X., Chen, H., Chen, J., Ye, Z., 2018. Supervised single-channel speech dereverberation and denoising using a two-stage model based sparse representation. Speech Communication 97, 1–8.

Zhang, Z., Pinto, J., Plahl, C., et al., 2014. Channel Mapping using Bidirectional Long Short-Term Memory for Dereverberation in Hands-Free Voice Controlled Devices. IEEE Transactions on Consumer Electronics 60, 525–533.

Zhang, Z., Ringeval, F., Han, J., et al., 2016. Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks, in: Proc. of the 17th Annual Conference

of the International Speech Communication Association (INTERSPEECH 2016), San Francisco, CA. pp. 3593–3597.

Zhong, Z., Zhang, B., Durrani, T., Xiao, S., 2016. Nonlinear signal processing for vocal folds damage detection based on heterogeneous sensor network. Signal Processing 126, 125–133.