INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks

Zixing Zhang[1], *Fabien Ringeval*[2,1], *Jing Han*[1], *Jun Deng*[1], *Erik Marchi*[1], *Björn Schuller*[1,3]

[1]Chair of Complex & Intelligent Systems, University of Passau, Passau, Germany
[2]Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, France
[3]Department of Computing, Imperial College London, London, UK

zixing.zhang@uni-passau.de

Abstract

During the last decade, speech emotion recognition technology has matured well enough to be used in some real-life scenarios. However, these scenarios require an almost silent environment to not compromise the performance of the system. Emotion recognition technology from speech thus needs to evolve and face more challenging conditions, such as environmental additive and convolutional noises, in order to broaden its applicability to real-life conditions. This contribution evaluates the impact of a front-end feature enhancement method based on an autoencoder with long short-term memory neural networks, for robust emotion recognition from speech. Support Vector Regression is then used as a back-end for time- and value-continuous emotion prediction from enhanced features. We perform extensive evaluations on both non-stationary additive noise and convolutional noise, on a database of spontaneous and natural emotions. Results show that the proposed method significantly outperforms a system trained on raw features, for both arousal and valence dimensions, while having almost no degradation when applied to clean speech.

Index Terms: emotion recognition, spontaneous speech, additive and convolutional noises, feature enhancement, autoencoder, LSTM Neural Networks

1. Introduction

Technology for automatic emotion recognition from speech (ERS) has gained increasing commercial attention in the last decade. Rapid progress of this technology has indeed enabled application of ERS in various domains, such as, health care [1], education [2], serious games [3], robotics [4], and call-centers [5]. However, while good performance has been reported in research papers under laboratory conditions [6], or with systems tailored towards specific databases [7], real-life applications of ERS still remain an open challenge. Indeed, various factors make this task highly challenging, which can be grouped into three main categories: (i) the contextual dependencies of the meaning and significance of affective expressions across different speakers, languages and cultures [8], (ii) the presence of varying and degraded acoustic conditions caused by reverberation, background noise, and acoustic properties of the recording devices used, and (iii) the necessity to use distributed systems in a client-server architecture, which introduce some latency and distortion in the data [9].

Stationary, non-stationary, and convolutional noise severely degrade performance of systems, and affect consequently the user experience in real-life conditions [10, 11, 12]. Therefore, many studies have been performed for speech and acoustic feature enhancement (FE), especially for automatic speech recognition (ASR). Recurrent Neural Networks (RNN) are widely used in this field to enhance corrupted features, which is an application of the de-noising autoencoder [13] principle: neural networks are trained to map noisy features to clean features. This method has recently also been exploited for speech enhancement in the time domain [14, 15]. RNN have been also studied for *blind* non-linear source separation, with the aim to enhance the acoustic features by separating noise and speech sources [16, 17]. In the context of speech enhancement, the authors in [14] use deep neural networks to map noisy to clean Mel features, but the network output is synthesised directly into a time domain signal, instead of constructing a filter based on speech and noise magnitudes. A combination of unsupervised noise estimation and Deep Neural Network (DNN) based speech power spectrum estimation is used in [15] to construct a Wiener filter. Supervised training of deep neural networks was performed to predict the ideal ratio mask in an uncertainty decoding framework for ASR [18].

Studies on noise robustness for ERS are much more sparse, despite being necessary for real-life applications of this technology. To the best of our knowledge, only a few studies have addressed this issue so far. Large acoustic feature sets were investigated in [10]. Adaptive noise cancellation was proposed as a front end in [11]. Speech enhancement based on spectral subtraction and masking properties was studied in [12]. Wavelet decomposition [19] and feature selection techniques [20] have also been proposed. Additionally, supervised Nonnegative Matrix Factorization (NMF) was investigated for the robustness of emotion recognition engines [21].

One may note that most existing work on noise robustness for ERS has been performed on acted emotions, which are rarely observable in real-life. Furthermore, only a few of those studies have analysed the impact of reverberated noise, which is known to impact severely the performance of ASR systems [21, 22]. In this light, this present contribution studies the impact of non-stationary additive noise and convolutional noise on the automatic recognition of spontaneous emotions from speech. We propose the use of a FE method based on a memory-enhanced recurrent Denoising Autoencoder (rDA) as a front end, and show that this method can significantly improve the performance, while having almost no degradation when applied to clean speech.

The following paper is structured as follows: the proposed FE method based on rDA is introduced in Section 2, then extensive experiments on spontaneous emotions are described in Section 3, and a conclusion with future work is given in Section 4.

http://dx.doi.org/10.21437/Interspeech.2016-998

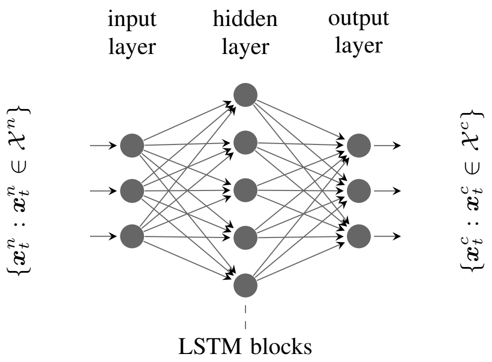

input layer
hidden layer
output layer
$\{\boldsymbol{x}_t^n : \boldsymbol{x}_t^n \in \mathcal{X}^n\}$
$\{\boldsymbol{x}_t^c : \boldsymbol{x}_t^c \in \mathcal{X}^c\}$
LSTM blocks

Figure 1: Structure of a recurrent denoising autoencoder with LSTM neural networks.

2. Feature enhancement

In the past few years, Long Short-Term Memory (LSTM) models have been widely applied to a variety of pattern recognition tasks [23, 24], and show a powerful capability of learning long-range contextual information. In this section, we give a quick overview of such a memory-enhanced RNN, on which the proposed rDA is built.

2.1. Memory-enhanced recurrent neural networks

Compared to conventional RNN, the LSTM-RNN model proposed by Hochreiter and Schmidhuber [25] uses one or multiple LSTM blocks to replace hidden neurons. Every memory block consists of self-connected linear memory cells c and three multiplicative gate units: an input gate i, a forget gate f , and a output gate o, which are responsible for writing, reading, and resetting the memory cell values, respectively. Given an input $\boldsymbol{x}_t$ at the step time t, the activations of the input gate i_t, the forget gate f_t, the memory cell state c_t, and the output gate o_t are separately updated by the following formulas:

$$i_t = f_g(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ci}\boldsymbol{c}_{t-1} + b_i), \tag{1}$$

$$f_t = f_g(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{cf}\boldsymbol{c}_{t-1} + b_f), \tag{2}$$

$$c_t = i_t \cdot f_i(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + b_c) + \boldsymbol{f}_t \cdot \boldsymbol{c}_{t-1}, \tag{3}$$

$$o_t = f_g(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{co}\boldsymbol{c}_t + b_o), \tag{4}$$

$$h_t = o_t \cdot f_o(c_t), \tag{5}$$

where f_g, f_i, and f_o denote the logistic sigmoid, tanh, and tanh activation functions, respectively; $\boldsymbol{W}$ is a weight matrix of the mutual connections; $\boldsymbol{h}_t$ presents the output of the hidden block; b indicates the block bias. From the equations mentioned above, it is observed that the values of all memory cells and block outputs in the previous time step $t-1$ will certainly affect the activations of all three gates, even the input units in the current time step t in the same layer, except for the case between memory cell and output gate. More details about the memory structure can be found in [26].

The main advantage of using such a memory-enhanced block over a traditional neuron in a RNN is that the cell state in a LSTM block sums activations over time. Since derivatives distribute over sums, the backpropagated error does not blow up or decay over time (the vanishing gradient problem) [25, 26].

The general structure of rDA with memory-enhanced neural networks proposed in this paper is illustrated in Fig. 1, which includes an input layer, an output layer, and one or multiple hidden layers that are implemented by the LSTM blocks. In comparison with the conventional DA given in [13] where the DA is modelled with Feedforward Neural Networks (FNN), the presented DA is structured with the above described LSTM-RNN in the hidden layers. Additionally, it also should be noticed that the recurrent autoencoder also differs from the ones described in [24, 27], where an encoder is used to map an input sequence into a fixed length representation, and a decoder is used to decode the target sequence from the representation.

2.2. Feature enhancement by an autoencoder

As discussed in Section 1, the speech signal $s(k)$ is easily distorted by the environmental noise and recording devices when facing realistic application scenarios with the Acoustic Impulse Response (AIR) $r(k)$ of finite length T_{60} and the background additional noise $n(k)$. Therefore, the distorted speech signal $\hat{s}(k)$ can be expressed as

$$\hat{s}(k) = s(k) * r(k) + n(k). \tag{6}$$

The signal in the time domain $\hat{s}(k)$ can be approximatedly transformed into the spectrum domain as

$$|\hat{S}(f)|^2 \approx |S(f)|^2 \cdot |R(f)|^2 + |N(f)|^2 \tag{7}$$

by applying a Short-Time Discrete Fourier Transform (STDFT) with three assumptions: 1) T_{60} is shorter than the analysis window size w; 2) The power spectrum of the additive noise in each analysis window w is a slowly varying process, which means that the additive noise is assumed to be stationary in each analysis window; 3) The phase of different analysis windows are non-correlated.

To extract the feature vectors in the ceptral domain such as Mel-Frequency Cepstrum Coefficients (MFCC) for emotion recognition from speech, logarithms and Discrete Cosine Transform (DCT) are performed over the above spectrum. Therefore, Eq. (7) can be further formulated into

$$\begin{aligned} \mathcal{D}(ln|\hat{S}(f)|^2) \approx & \mathcal{D}(ln|S(f)|^2) + \mathcal{D}(ln|R(f)|^2) \\ & + \mathcal{D}(ln(1 + \frac{|N(f)|^2}{|S(f)|^2 \cdot |R(f)|^2})). \end{aligned} \tag{8}$$

From Eq. (8), we can see that the goal of denoising is to eliminate the impact of the last two terms. For the non-stationary noise, however, the cepstrum does not only fluctuate over time, but is also involved with the original speech spectrum which is non-stationary as well. Therefore, the last term in Eq. (8) cannot be simply subtracted due to its non-linear property.

To tackle this non-linear problem, we choose the memory-enhanced rDA as described in Section 2.1 with the purpose of exploiting its advantage of accessing long-range contextual information. The goal of the DA is to reconstruct the features $\boldsymbol{x}^c$ in the clean speech feature domain $\mathcal{X}^c$ from the corresponding features $\boldsymbol{x}^n$ in the corrupted speech feature domain $\mathcal{X}^n$, as shown in Fig. 1. When providing these corrupted features as the input $\boldsymbol{x}^n$ to the first layer, we want the output $\hat{\boldsymbol{x}}^n$ to be highly similar to the clean features $\boldsymbol{x}^c$. To learn the required mapping between noisy and clean features, an objective function – Mean Squared Error (MSE) – is defined to mimise the reconstruction error during training:

$$\mathcal{J}(\theta) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\boldsymbol{x}}_t^n - \boldsymbol{x}_t^c)^2, \tag{9}$$

where T is the number of frames of the training set.

3. Experiments and results

In the following, we firstly describe the selected spontaneous emotion database, then evaluate the performance of the proposed FE method based on the rDA with LSTM neural networks for time- and value-continuous emotion recognition.

3.1. RECOLA and noise database

For the experiments, we chose the REmote COLlaborative and Affective (RECOLA) database [28] which was used as a database for the 5^{th} Audio/Visual^{+} Emotion Challenge (AV^{+}EC 2015) [29]. The motivation of the database collection was to study the complex phenomena, especially emotion, portrayed by humans during social interactions in daily-life.

To generate additive noisy speech, we added the CHiME15 database [30] into the clean (raw or original) speech with various levels of SNR (i.e., 0–12 dB at a step of 3 dB). This database was used for the 3rd CHiME Challenge [30], and was collected in five different locations, such as booth, bus, cafe, pedestrian area, and street junction. The goal of this database is to simulate emotional speech in different places with various additive background noises.

To generate convolution noise, we applied (via convolution) the Microphone Impulse Response (MIR) of the Google Nexus One smartphone to the recordings from RECOLA using the Audio Degradation Toolbox (ADT) [31]. The goal is to simulate reverberant speech being recorded with a smartphone. Moreover, other noises are further simulated via the MIR, by applying the Room Impulse Response (RIR) of classroom or grand hall as the second convolutional noise. This aims to simulate the scenarios that someone speaks on the phone in different environments.

Note that, when adding the CHiME noise, each noise recording was firstly normalised to 0 dB peak energy and concatenated according to the type of noise. Then, the recording was cut into three partitions of the same length for the training, the validation, and the test sets, respectively. Finally, for each recording of RECOLA, we randomly chose an excerpt of the concatenated noise signal from the relevant partition, to ensure both speaker and noise independent partitions.

3.2. Experimental setups

At the front-end of the emotion recognition system, 13 Low-Level Descriptors (i.e., MFCCs 0–12) were firstly extracted. In detail, the feature vectors of $\boldsymbol{x}_t^n$ and $\boldsymbol{x}_t^c$ were separately extracted from the distorted speech signals and the original clean speech signals at every 10 ms using a window size of 25 ms. Before training the rDA, the global means and variances were calculated of the noisy and clean speech. Then, standardisation was performed over the network inputs and targets using the means and variances from the corresponding training sets, respectively.

For the rDA, both input and output node numbers are equal to the dimension of the feature vector (13 in our case). Two bidirectional LSTM hidden layers were chosen, and each layer consists of 30 memory blocks. During network training, gradient descent was implemented with a learning rate of 10^{-6} and a momentum of 0.9. Zero mean Gaussian noise with standard deviation 0.1 was added to the input activations in the training phase such as to improve generalisation. All weights were randomly initialised in the range from -0.1 to 0.1. Note that, all these parameters were optimised on the validation set. Finally, the early stopping strategy was used, i.e., training was stopped when no improvement of the MSE on the validation set has been observed during 20 epochs or the predefined maximum number of training epochs (200 in our case) has been executed. Further, to accelerate the training process, we updated the network weights after running every mini batch of 8 sequences for computation in parallel. The training was performed with our CURRENNT toolkit [32].

After the procedure of FE, functionals – mean and variance – were applied over each of the enhanced MFCCs with a window size of 8 s at a step of 0.04 s, which leads to 26 attributes for each window. These statistical features were then fed into the back-end of the system used for emotion recognition, where L2-regularised L2-loss Support Vector Regression (SVR) implemented in the LIBLINEAR toolbox [33] was used. The complexity value of SVR was optimised by the best performance of the validation set, i.e., $C = 5 \cdot 10^{-5}$ for arousal and $C = 5 \cdot 10^{-3}$ for valence in our experiments.

For the performance evaluation, we choose the Concordance Correlation Coefficient (CCC) [34]. Compared to Pearson's Correlation Coefficient (PCC), CCC can estimate not only the linear correlation, but also the difference of the bias between two variables. Formally, CCC is formulated as

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{10}$$

where ρ is the correlation coefficient between the two variables; μ_x and μ_y are the means of the two variables; and σ_x^2 and σ_y^2 are the corresponding variances. Moreover, it is worth noting that the gold standard ratings for all the recordings were shifted back in time with a four seconds delay. This is due to the reaction delay of human for continuous emotion annotation [35].

3.3. Results and Discussion

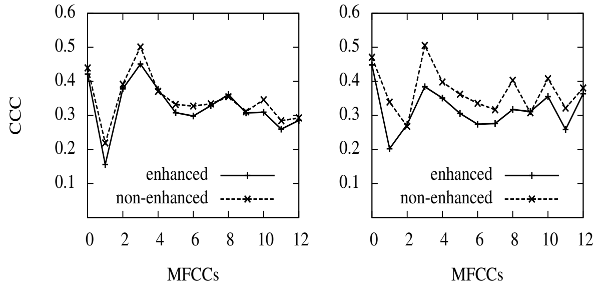
To verify the effectiveness of the rDA with LSTM Neural Networks for FE, we separately performed two experiments: 1) on the non-stationary *additive* noisy speech only, i.e., by adding *CHiME15* noise; 2) on the *smartphone* related *convolutional* noisy speech, i.e., the speech is distorted by applying MIR only (smartphone), or additionally by applying RIR of the classroom/hall (+classroom/hall), or additionally by adding various levels of CHiME noise (+CH).

Apart from that, we carried out two FE strategies – 1) *matched* FE: Several FE models are trained separately on the data sets with different noise conditions. For example, when testing on clean speech, the same quality of speech is used, i.e., clean speech is employed to train the rDA; 2) *mixed* FE: One FE model is trained on a data set with mixed noise conditions. Therefore, the distinction between the two FE strategies is based on the noise condition of the training data that can match or not with the one of the testing data. For example, when testing the clean speech, the mixed conditional speech, i.e., all kinds of CHiME noisy speech or the smartphone related noisy speech together with the clean speech, are utilised to train the rDA.

Table 1 shows the performance of the non-enhanced and enhanced speech (with the matched or mixed FE strategies) evaluated on the emotion recognition model trained on the clean speech for both, arousal and valence regression. In almost all cases, the proposed FE method significantly outperforms the system trained on the non-enhanced noisy speech (baseline). Taking the additive noisy speech (CHiME15) for example, the average CCC over the recordings at different levels of SNRs on the test set is boosted from 0.563 to 0.596 and 0.594, respectively, by matched and mixed FE for arousal, and from 0.176 to 0.223 and 0.199, respectively, by matched and mixed FE for

Table 1: Performance (Concordance Correlation Coefficient [CCC]) of the *validation* and *test* sets for the proposed *matched* and *mixed* feature enhancement (FE) model on the *CHiME15* noisy speech only or on the *smartphone related* noisy speech, in the evaluation of *arousal* and *valence* emotional tasks. class.: classroom; $\overline{CH}$.: the average CCC over five different 'smartphone + CHiME' noisy speeches with 0–12 dB of SNRs at a step of 3 dB. The mark of "/": no other noise is added onto the smartphone related noisy speech.

	CHiME15							smartphone +					
CCC	clean	12dB	9dB	6dB	3dB	0dB	mean	clean	/	class.	hall	$\overline{CH.}$	mean
arousal on the *validation* set													
baseline	0.736	0.680	0.657	0.626	0.584	0.526	0.635	0.736	0.726	0.629	0.634	0.436	0.545
matched FE	0.735	0.715	0.710	0.692	0.666	0.627	**0.691**	0.735	0.723	0.641	0.662	0.675	**0.682**
mixed FE	0.693	0.721	0.711	0.691	0.648	0.594	**0.676**	0.690	0.686	0.650	0.651	0.599	**0.630**
arousal on the *test* set													
baseline	0.732	0.628	0.590	0.542	0.480	0.404	0.563	0.732	0.719	0.618	0.609	0.356	0.495
matched FE	0.729	0.658	0.646	0.611	0.510	0.422	**0.596**	0.729	0.713	0.614	0.694	0.682	**0.684**
mixed FE	0.717	0.683	0.651	0.598	0.499	0.418	**0.594**	0.712	0.690	0.612	0.600	0.532	**0.586**
valence on the *validation* set													
baseline	0.402	0.304	0.276	0.246	0.213	0.180	0.270	0.402	0.359	0.306	0.302	0.156	0.239
matched FE	0.383	0.335	0.299	0.259	0.257	0.223	**0.293**	0.383	0.335	0.236	0.331	0.187	**0.246**
mixed FE	0.201	0.227	0.275	0.275	0.262	0.205	0.249	0.253	0.243	0.251	0.253	0.215	0.230
valence on the *test* set													
baseline	0.278	0.190	0.172	0.155	0.139	0.124	0.176	0.278	0.237	0.212	0.205	0.003	0.105
matched FE	0.269	0.227	0.258	0.214	0.200	0.172	**0.223**	0.269	0.211	0.126	0.152	0.140	**0.162**
mixed FE	0.171	0.210	0.217	0.214	0.202	0.179	**0.199**	0.195	0.175	0.208	0.209	0.085	**0.135**



(a) CHiME (b) smartphone-in-the-hall

Figure 2: Concordance Correlation Coefficient (CCC) of 13 Low-Level Descriptors (MFCC 0–12) between the enhanced/non-enhanced speech and the clean speech over the whole test set with the *CHiME* noise (a) or the *smartphone-in-the-hall* noise (b).

valence. Furthermore, it is expected that the matched FE outperforms the mixed FE, since the matched FE uses different FE models for denoising corresponding noisy data, whereas the mixed FE trains only one FE model for denoising all kinds of noisy data.

Specifically, the performance obtained on the clean speech condition almost does not degrade when executing the matched FE method, but this conclusion is not supported by performing the mixed FE method. This should be mainly due to huge mismatch between the clean speech and the mixed noisy speech. However, this can be easily solved by inserting a noise detector at the front-end to distinguish whether the signal is noisy or clean [36]. If it was clean, the speech signal can directly be fed into the recognition model without any procedures of FE. Moreover, for the convolutional noise of smartphone in the classroom or in the hall, we can see that our proposed method does not work quite well, however, for the convolutional noise of the smartphone with CHiME noise, the proposed method can surprisingly improve the baseline. This may be because LSTM-RNN does not work efficiently for a linear problem, as the T_{60} of the MIR used for these experiments is short and the convolutinal noise can be regarded as a constant value in the spectral domain (see Eq. (8)).

To further investigate the efficiency of the proposed FE, we calculated the CCC between the enhanced (by matched FE)/non-enhanced speech and the clean speech over the whole test set with the CHiME noise or the smartphone-in-the-hall noise, as illustrated in Fig. 2. It can be seen that the enhanced speech could deliver higher correlation coefficients with the speech, which possibly contributes to the better emotion recognition performance.

4. Conclusions

We presented a feature enhancement method based on a denoising autoencoder with Long Short-Term Memory neural networks for spontaneous emotion recognition from speech. Extensive experiments were carried out with non-stationary additive noise and convolutional noise. The results show that, the presented feature enhancement method is significantly superior to the baseline without any enhancement methods. With the fast development of deep learning technologies, there are many possibilities that could be used to further improve the robustness performance of emotion recognition systems from speech. For example, Convolutional Neural Networks are good at reducing spectral variation for the clean speech, which could also be effective for noisy speech. Methods combined with Deep Neural Networks in an end-to-end structure [37] is worth evaluating as well in future. Further, some other traditional denoising approaches, e.g., minimum mean square error [38], may be also of interest in this task.

5. Acknowledgements

The research leading to these results was supported by the EC's 7th Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), the EU's Horizon 2020 Programme through the Innovative Action No. 644632 (MixedEmotions), No. 645094 (SEWA), and the Research Innovative Action No. 645378 (ARIA-VALUSPA), and by the German Federal Ministry of Education, Science, Research and Technology (BMBF) under grant agreement #16SV7213 (EmotAsS). We further thank the NVIDIA Corporation for their support of this research by Tesla K40-type GPU donation.

6. References

[1] D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, C. Setz, G. Troster, and C. Haring, "Activity and emotion recognition to support early diagnosis of psychiatric diseases," in *Proc. of PervasiveHealth*, Istanbul, Turkey, 2008, pp. 100–102.

[2] R. A. Calvo and S. D'Mello, "Frontiers of affect-aware learning technologies," *IEEE Intelligent Systems*, vol. 27, no. 6, pp. 86–89, Nov 2012.

[3] B. Schuller, E. Marchi, S. Baron-Cohen, A. Lassalle, H. O'Reilly *et al.*, "Recent developments and results of asc-inclusion: An integrated internet-based environment for social inclusion of children with autism spectrum conditions," in *Proc. of IDGEI*, Atlanta, GA, 2015, no pagination.

[4] E. Marchi, F. Ringeval, and B. Schuller, "Voice-enabled assistive robots for handling autism spectrum conditions: An examination of the role of prosody," in *Speech and Automata in the Health Care*, A. Neustein, Ed. Walter de Gruyter GmbH & Co KG, 2014, pp. 207–236.

[5] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proc. of Artificial Neural Networks in Engineering*, vol. 710, St. Louis, MO, 1999, pp. 7–10.

[6] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. of ASRU 2009*. Merano, Italy: IEEE, Dec 2009, pp. 552–557.

[7] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters, Special Issue on Pattern Recognition in Human Computer Interaction*, vol. 66, pp. 22–30, Nov 2015.

[8] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proc. of the National Academy of Sciences of the United States of America (PNAS)*, vol. 107, no. 6, pp. 2408–2412, 2009.

[9] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Distributing recognition in computational paralinguistics," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 406–417, Oct 2014.

[10] B. Schuller, D. Arsić, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," in *Proc. of Speech Prosody*, Dresden, Germany, 2006, no pagination.

[11] A. Tawari and M. Trivedi, "Speech emotion analysis in noisy real-world environment," in *Proc. of ICPR*, Istanbul, Turkey, 2010, pp. 4605–4608.

[12] C. Huang, G. Chen, H. Yu, Y. Bao, and L. Zhao, "Speech emotion recognition under white noise," *Archives of Acoustics*, vol. 38, no. 4, pp. 457–463, 2013.

[13] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. of ICML*, Helsinki, Finland, 2008, pp. 1096–1103.

[14] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 3444–3448.

[15] B.-Y. Xia and C.-C. Bao, "Speech enhancement with weighted denoising auto-encoder," in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 436–440.

[16] Y. Tan, J. Wang, and J. M. Zurada, "Nonlinear blind source separation using a radial basis function network," *IEEE Transactions on Neural Networks*, vol. 12, no. 1, pp. 124–134, 2001.

[17] Z. Zhang, J. Pinto, C. Plahl, B. Schuller, and D. Willett, "Channel mapping using bidirectional long short-term memory for dereverberation in hand-free voice controlled devices," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 3, pp. 525–533, Aug 2014.

[18] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 7092–7096.

[19] J. C. Vasquez-Correa, N. Garcia, J. R. Orozco-Arroyave, J. D. Arias-Londono, J. F. Vargas-Bonilla, and E. Noth, "Emotion recognition from speech under environmental noise conditions using wavelet decomposition," in *Proc. of ICCST*, Taipei, China, 2015, pp. 247–252.

[20] F. Eyben, F. Weninger, and B. Schuller, "Affect recognition in real-life acoustic conditions – a new perspective on feature selection." in *Proc. of INTERSPEECH*, vol. 2013, Lyon, France, 2013.

[21] F. Weninger, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–16, Dec 2011.

[22] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Berlin: Springer, 2005.

[23] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, Jan 2015.

[24] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. of NIPS*, Montreal Canada, 2014, pp. 3104–3112.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, Nov 1997.

[26] A. Graves, *Supervised sequence labelling with recurrent neural networks*. Berlin/Heidelberg, Germany: Springer, 2012, vol. 385.

[27] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," in *Proc. of ICML*, Lille, France, 2015, pp. 843–852.

[28] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. of EmoSPACE (FG)*, Shanghai, China, 2013, pp. 1–8.

[29] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV$^+$EC 2015 – The first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. of AVEC Workshop*, Brisbane, Australia, 2015, pp. 3–8.

[30] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. of ASRU Workshop*, Scottsdale, AZ, 2015, pp. 504–511.

[31] M. Mauch and S. Ewert, "The audio degradation toolbox and its application to robustness evaluation," in *Proc. of ISMIR*, Curitiba, Brazil, 2013, pp. 83–88.

[32] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT: The munich open-source CUDA RecurREnt Neural Network Toolkit," *Journal of Machine Learning Research*, vol. 16, pp. 547–551, Mar 2015.

[33] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[34] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, March, 1989.

[35] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. of AVEC Workshop*, Brisbane, Australia, 2015, pp. 73–80.

[36] N. Garner, P. Barrett, D. Howard, and A. Tyrrell, "Robust noise detection for speech detection and enhancement," *Electronics Letters*, vol. 33, no. 4, pp. 270–271, 1997.

[37] G. Trigeorgis, F. Ringeval, R. Bruckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a Deep Convolutional Recurrent Network," in *Proc. of ICASSP*, Shanghai, China, 2016, pp. 5200–5204.

[38] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," in *Proc. of ICASSP*, Salt Lake City, UT, 2001, pp. 301–304.