Adieu recurrence? End-to-end speech emotion recognition using a context stacking dilated convolutional network

Duowei Tang
Dept. of Electrical Engineering
ESAT-STADIUS, KU Leuven
Leuven, Belgium
duowei.tang@kuleuven.be

Peter Kuppens
Faculty of Psychology and Educational Sciences
KU Leuven
Leuven, Belgium
peter.kuppens@kuleuven.be

Luc Geurts
e-Media Research Lab
Dept. of Electrical Engineering
ESAT-STADIUS, KU Leuven
Leuven, Belgium
luc.geurts@kuleuven.be

Toon van Waterschoot
Dept. of Electrical Engineering
ESAT-STADIUS, KU Leuven
Leuven, Belgium
toon.vanwaterschoot@esat.kuleuven.be

***Abstract*—In state-of-the-art end-to-end Speech Emotion Recognition (SER) systems, Convolutional Neural Network (CNN) layers are typically used to extract affective features while Long Short-Term Memory (LSTM) layers model long-term temporal dependencies. However, these systems suffer from several problems: 1) the model largely ignores temporal structure in speech due to the limited receptive field of the CNN layers, 2) the model inherits the drawbacks of Recurrent Neural Network (RNN)s, e.g. the gradient exploding/vanishing problem, the polynomial growth of computation time with the input sequence length and the lack of parallelizability. In this work, we propose a novel end-to-end SER structure that does not contain any recurrent or fully connected layers. By levering the power of the dilated causal convolution, the receptive field of the proposed model largely increases with reasonably low computational cost. By also using context stacking, the proposed model is capable of exploiting long-term temporal dependencies and can be an alternative to RNN. Experiments on the RECOLA database publicly available partition show improved results compare to a state-of-the-art system. We also verify that both the proposed model and the state-of-the-art model learned from short sequences (i.e. 20 s) can make accurate predictions for very long sequences (e.g. $\geqslant 75\,s$).**

***Index Terms*—End-to-end learning, Speech Emotion Recognition, Dilated Causal Convolution, Context Stacking**

I. Introduction

Emotion plays a basic yet important role in people's daily life. Vocal expression is a direct way to express emotions, and thus forms a crucial modality to recognize and interpret emotion by artificial intelligence systems, as well as in human-computer interaction applications. At a very basic level, the realm of emotional experience can be described by means of two orthogonal dimensions, arousal (activation - deactivation) and valence (pleasure - displeasure), see [1].

Early SER systems follow a pipeline of feature extraction, modelling and inference. The definition of emotion-related features in this case becomes a key aspect towards an accurate and robust SER system. Hand-crafted features are widely used because careful feature design results in better generalization to various affectively annotated datasets [2], however, it requires exhaustive selection and experiments. Even if hand-crafted features may yield relatively simple predictive models, the feature selection process suffers from a potentially huge information loss [3], which could be harmful to the SER performance.

With the increasing popularity of Deep Neural Networks (DNNs), the feature extraction for an SER system has shifted to data-driven feature learning. The work in [4] has used "shallow features" (spectrograms) as the input for learning the "deep features" with a DNN structure where CNN layers are used. The CNN is shown to be competent to learn emotion-related features that can achieve comparable performances to the conventional features (i.e. energy related, voicing related and spectral features).

A number of works investigate even "shallower" input data and propose an end-to-end approach to SER [5]–[8]. In the end-to-end SER approach, raw recording samples are used as the input to a DNN consisting of both CNN and different types of RNN layers. In this case, the CNN layers are applied on frames of raw recording samples to produce higher-level features.

From a perspective of modelling sequential data, speech

This research work was carried out at the ESAT Laboratory of KU Leuven. The research leading to these results has received funding from the Chinese Scholarship Council (CSC) (grant no. 201707650021), the KU Leuven Internal Funds C2-16-00449 and VES/19/004, and the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

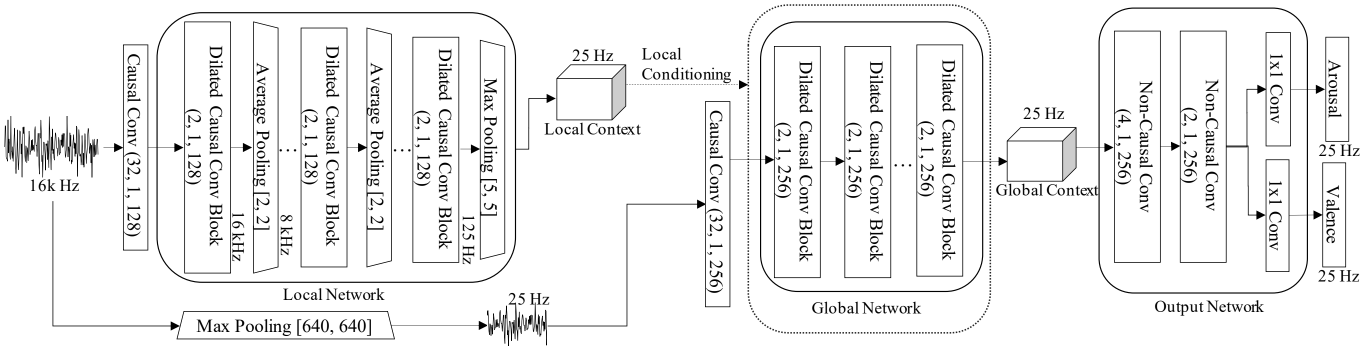

Fig. 1: The proposed dilated convolutional network with context stacking. The convolution filter width, stride and filter depth is listed in round brackets, the pooling width and stride is listed in square brackets.

emotion analysis has some inherent difficulties. Research shows that the duration of people's emotions can range from just a few seconds to over several hours [9], and that the dynamics of emotional experience is regulated by both internal and external excitations [10]. Thus a good model should be able to represent and learn the dependencies or relations over a sufficiently long period. The aforementioned state-of-the-art end-to-end SER systems use LSTM layers for this purpose. A recent work using Gated Recurrent Unit (GRU) layers along with the attention mechanism confirms that the simpler GRU is effective for long-term dependencies modelling [11], and can also be used in end-to-end SER systems as an alternative to LSTM [8].

However, the RNN type of layers used in the state-of-the-art SER systems suffer from several disadvantages: firstly, RNN has a sequential type of processing which results in a polynomial growth of computation time with increasing input sequence length. This type of processing is also not capable to be parallelised. Secondly, RNN suffers from the gradient exploding/vanishing problem especially when long sequences are used. Nevertheless, this problem has been alleviated by the developments in LSTM and GRU [12], [13]. Finally, empirical research shows that convolution structures outperform recurrent networks such as LSTM, suggesting that temporal convolution networks should be preferred over RNN for sequence modelling [14].

In this work, we propose a novel DNN structure in the end-to-end SER framework that maps the raw recording samples to arousal and valence values, without using any recurrent connections and fully connected layers inside the network. The proposed model contains dilated causal convolution blocks adapted from the WaveNet model [15], resulting a pure multi-task regression model. This model consists of two networks which are stacks of dilated causal convolution blocks. First, in a "Local network", it learns local context information (e.g. from 0.04 s length raw inputs). Second, in a "Global network", it learns very wide context information (e.g. from 20 s length raw inputs) with reasonable computational cost. This proposed model, which is trained with 20 s raw audio inputs, can accurately infer on varying length of inputs (e.g. 10 s to 150 s). By injecting input noise during training, the proposed model finally surpasses the state-of-the-art model of [7] for the REmote COLlaborative and Affective (RECOLA) database.

The rest of the paper is organized as follows. Section 2 provides a brief overview about the most recent studies related to our work. In Section 3 we introduce the proposed model structure that will be optimized on the Concordance Correlation Coefficient (CCC) objective function. After the description of the database and experimental settings, we will present the simulation results in Section 4.

II. Related works

The proposed model structure is inspired by the WaveNet model that has shown successful outcomes in Text-to-speech (TTS) problems [15]. The WaveNet model does not contain any recurrent connections, and has a very large receptive field by stacking many dilated causal convolution blocks. In these blocks, both residual connections and gated activations are present, which have also been combined with dilated convolutions in some prior image processing works [16], [17].

There is hardly any work investigating dilated convolutions in the SER framework. In [18], it is suggested that dilated residual CNNs facilitate the reduction of the receptive field and hence yield a strong ability to learn local content. However, LSTM is still used in their model after the dilated residual CNNs for modeling the temporal relationships.

Our proposed DNN architecture is different from the standard WaveNet model where we remove the skip-connections in their residual block. Further we also expand the context stacking idea of [15], to propose a context stacked dilated convolutional network. In the proposed model a "global network" learns the global context conditioned by a "local network" while the "local network" distils local information from the raw input samples.

III. Method

We propose a DNN structure, without any recurrent connection, for end-to-end SER, as shown in Fig.1. The network consists of dilated causal convolutions that are used for increasing the receptive field [15], [17] (Sec III-A), then two sub-networks are stacked to further increase the filter span on the input sequence (Sec III-B). Finally, the model outputs the arousal and valence and is trained to minimize the CCC objective function (Sec III-C).

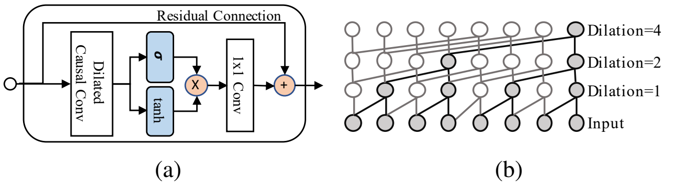

Fig. 2: (a) A dilated causal convolution block and (b) a stack of dilated causal convolution blocks.

A. Dilated causal convolution blocks

The dilated causal convolution block in the proposed model, shown in Fig. 2a, is similar to the one in [15] without the skip-connection because the skip-connections are intended to learn a sample-wise distribution which can be used to generate new samples. However, this is not our objective. Each dilated causal block consists of a residual path and a gated dilated convolution path. The residual path is a direct connection from the input to the output. On the other hand, in the convolution path, the input first goes to two dilated causal convolution layers that different activations are applied on their outputs. The outputs from the activations can be considered as "filter" when $\tanh(\cdot)$ activation is applied and "gate" when $\sigma(\cdot)$ (sigmoid activation) is applied. Element-wise multiplication is applied on these outputs afterwards, where the "filter" output is gated (i.e. selected or not selected) by the "gate" output. Then after going through a 1×1 convolution, the gated dilated causal convolution path is summed together with the residual path to generate the final output of the block. The block is then stacked many times with different dilation number in the network. The dilation number defines the number of skipping steps between two neighbour filter inputs sampled from the input sequence. Fig. 2b shows a stack of dilated causal convolution blocks with filter width 2, and dilation numbers 1, 2 and 4.

B. Context stacking with local conditioning

Local conditioning in [15] refers to adding an extra term to both the "filter" and the "gate" inputs. Consider a raw input sequence containing T samples, $\boldsymbol{x} = \{x_1, \ldots, x_T\}$. If the gated filter output is $\boldsymbol{z}$, a local conditioning is defined as, adapting the notation from [15]:

$$\boldsymbol{z} = \tanh(W_{f,k} * \boldsymbol{x} + V_{f,k} * \boldsymbol{y}) \odot \sigma(W_{g,k} * \boldsymbol{x} + V_{g,k} * \boldsymbol{y}) \quad (1)$$

where $W_{f,k}, V_{f,k}$ and $W_{g,k}, V_{g,k}$ are the learnable "filter" and "gate" parameters in the k^{th} layer, $*$ is the convolution operation, and $\odot$ is the element-wise multiplication. $\boldsymbol{y}$ is another sequence having the same length as $\boldsymbol{x}$, and containing the conditioning information.

Referring to the context stacking idea in [15], we propose a stacked structure using local conditioning for end-to-end SER. The network consists of two sub-networks, where one sub-network has a smaller receptive field that learns local information and is denoted as the "local network". The other sub-network has a wider receptive field that can learn longer dependencies in a long sequence and is denoted as the "global network". The two networks connect by letting the "local network" define the local conditioning on the "global network". We also add pooling layers in the "local network" to downsample the sequence with the aim of reducing computation cost and memory requirement.

Finally, the outputs from the global network are processed with two traditional CNN layers, and two separated 1×1 convolution layers to generate arousal and valence estimates.

C. Objective function

To reduce the inductive bias, a loss function that has a direct link to the evaluation metric based on the CCC (ρ_c) has been proposed in [5]–[7] and was shown to provide better regression performance on the affective arousal and valence predictions. Given the predicted sequence (denoted by index m) for arousal/valence dimension and its corresponding ground-truth sequence (denoted by index n), the loss is:

$$\mathcal{L}_c = 1 - \rho_c = 1 - \frac{2\sigma_{mn}^2}{\sigma_m^2 + \sigma_n^2 + (\mu_m - \mu_n)^2} \quad (2)$$

where μ_m and μ_n are the sample means, σ_m^2 and σ_n^2 are the sample variances, and σ_{mn}^2 denotes the covariance between the two sequences.

IV. Experiments and results

A. Dataset

To validate the proposed modelling framework and to compare with the state-of-the-art end-to-end SER model, we use the same RECOLA database [19] used in earlier work. This database contains abundant affective data annotated by experts. The content of the data consist of interviews in which people talk about real-life stories. However, since the database is not fully publicly available, we can only acquire a sub-partition that is used in the 2015 and 2016 Audio/Visual Emotion Challenge and Workshop (AVEC) [20], [21]. Although the RECOLA database contains four modalities (audio, video, electrocardiogram (ECG), and electro-dermal activity (EDA)), we only use the raw audio and the corresponding labels provided by the AVEC 2016 challenge. Some characteristics of this sub-partition used in this paper are summarized in TABLE I. It should be noted that this sub-partition is not completely identical to the database used in [5]–[7], so we also tested the state-of-the-art model on the same database partition that our model is tested on. The original database has three parts (train, development, and test), but since the test part is preserved for the challenge, we can not test the model on the test partition and will only report the performance on the development part in Section IV-C. Nevertheless, the results should be representative since the development partition in this database originates from similar recording conditions as the testing partition.

We also pre-process the audio by rescaling the raw audio samples into the amplitude range $[-1, 1]$ and downsampling them to 16 kHz.

TABLE I: RECOLA database sub-partition characteristics

	Training	***Development***
Number of speakers	5 Females, 4 Males	5 Females, 4 Males
Audio duration per speaker	5 minutes	5 minutes
Audio sampling frequency	44.1 kHz	
Label information	Arousal and Valence values	
Label characteristic	Continuous numbers in range [-1, 1]	
Label frequency	25 Hz	
Language	French	

B. Implementation Details

Stride 1 and zero-padding is used across all convolution layers to retain the same sequence length after each operation.

1) Local network: Firstly, we construct the local network, which consists of 30 dilated causal convolution block layers, with filter width 2 and depth equal to 128. The first six layers have dilation numbers of $D_k = 2^k$ where $k = 0 \dots 5$. Then the same dilation numbers are repeated 5 times. The input is first filtered with a causal convolution with filter width 32 and depth equal to 128, before being fed into the local network. There are in total 7 intermediate average-pooling layers in the local network, with pooling width of 2, which progressively downsample the processed data from 16 kHz to 125 Hz. Then an output max-pooling layer with pooling width 5 further downsamples it to 25 Hz.

In parallel to the local network, the input raw audio is fed into a max-pooling layer with pooling width 640, so that its output has a sampling rate of 25 Hz, which is the same as the label sampling rate. Together with the output from the local network which serves as the local conditioning, the downsampled audio is then fed into a causal convolution layer with filter width 32 and depth 128, and then to the global network.

2) Global network: It consists of 47 dilated causal convolution block layers. The first 20 layers have dilation numbers equal to $D_k = 2^k$ ($k = 0 \dots 9$), which are repeated 2 times, while the last 27 layers have dilation numbers $D_k = 2^k$ ($k = 0 \dots 8$), which are repeated 3 times. These numbers are chosen based on validation performance. However, a key aspect is to choose exponentially increasing dilation numbers such as to cover a sufficiently long span (e.g. the global network has dilation number up to 512 which covers 20 s span of the input audio data). These dilated convolution block layers have filter depth 256 and filter width 2. Finally, all the blocks are conditioned on the local context. We implement those context stacking filters (filter "V" in III-B) by normal CNNs with filter depth 256 and width 2.

3) Output network: Then in the output network, two non-causal convolution layers are used which have filter depths equal to 256, filter width of 4 for the first layer, and 2 for the second layer. The final outputs are two 1×1 convolution layers that yield the arousal and valence predictions.

4) Data augmentation: To overcome over-fitting, we select the training inputs with overlapping frames from the training set, then the selected frames are shuffled during training. Gaussian noise with zero mean and standard deviation of 0.01 is added to the input frames. Those data augmentation techniques are widely used in neural network practice and may yield better generalization of the model [22]. In validation, we only predict successive frames with neither overlapping nor additional input noise.

5) Training settings: Finally, we use the Adam optimizer to train our model, with a fixed learning rate of 10^{-4}. Batch-size is 3 due to memory constraints. l_2 regularization with a regularization parameter of 10^{-4} is applied. There is no post-processing on the output predictions. Every experiment is repeated for 5 times with randomly initialized model parameters at the beginning of every run. Afterwards, we select the model which has the highest total arousal and valence CCC on the validation set. Finally, the mean CCC across 5 runs is reported.

C. Comparison with the state-of-the-art

In the first experiment, we compare our proposed model with the state-of-the-art end-to-end SER system by Tzirakis et al. [7] (denoted as CNN-LSTM). We implement their model with the model architecture and hyper-parameter settings provided in [7] (input sequence length is 20 s). Then we re-train the model with the same dataset partition that is used for training and evaluating our model. Both models are trained with length 20 s input sequences and evaluated also on length 20 s sequences. The validation results are shown in the 2^{nd} point in Fig. 3 (corresponding to 20 s), where the proposed model gets CCCs equal to 0.782 and 0.499 for arousal and valence separately, whereas the CNN-LSTM model gets 0.763 and 0.516 separately.

Input noise is then added to train both models, and the resulting CCCs are shown in Fig. 3 with suffix "_IN". Regarding generalization, the effectiveness of adding input noise is subtle to both the CNN-LSTM model and the proposed model in arousal predictions, however, it significantly helps our proposed model in valence predictions. The proposed model trained with input noise yields CCCs equal to 0.78 in arousal predictions and 0.518 in valence predictions. As a result, the proposed model, trained with input noise, outperforms the state-of-the-art CNN-LSTM model for input data of length 20 s.

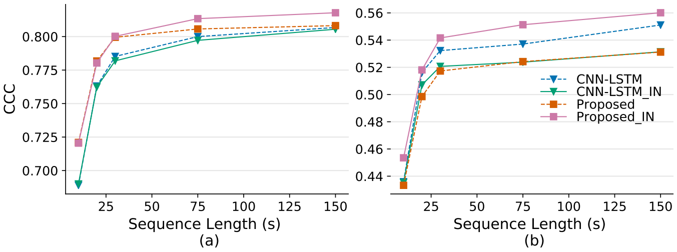

Fig. 3: CCCs for arousal (a) and valence (b) with varying input sequence length, while both models are trained only with 20 s sequences and are applied to predict for varying input sequences length. Suffix "_IN" indicates input noise is added during training.

D. Identification of the influence of input sequence length

To further identify the influence of the input sequence length, we also evaluate the performance with input frames of 10 s, 20 s, 30 s, 75 s, and 150 s. Re-training of the models is not necessary since they accept varying sequence lengths, therefore we use the models trained with 20 s sequences and evaluate them on the aforementioned sequence lengths. The effectiveness of adding input noise is also tested with both models.

The results are plotted in Fig. 3. Firstly, an increasing trend in performance is observed when longer sequences are used. This trend is valid for both models, with or without input noise. The best CCCs for the proposed model are obtained when the input sequence length equals 150 s and with input noise. The result increases 4.8% on arousal predictions and 8% on valence predictions compared to the baseline case where 20 s sequences are used [7]. Secondly, adding input noise does help the proposed model to generalize well for longer sequence lengths, whereas it is not the case for the CNN-LSTM model. The input noise is harmful to the CNN-LSTM model in valence predictions when the sequence length is $\geqslant 20\,s$. This maybe due to the fact that the proposed model with a deeper structure and more trainable parameters is more capable to learn a robust mapping, whereas in contrast, the CNN-LSTM model may over-fit to the noise. Finally, the proposed model trained with input noise outperforms the state-of-the-art CNN-LSTM model with 1.4% to 4.4% on arousal CCCs, and 2.2% to 5.4% on valence CCCs for all the testing cases.

V. Conclusions

We proposed a novel end-to-end DNN structure for SER that does not consist of any recurrent and fully connected layers. Simulation results firstly indicated that the proposed model, which is deeper and consists of more learnable parameters than the selected state-of-the-art model, can learn a more robust mapping when input noise is added. Secondly, the results also showed that models learned from 20 s sequences can yield accurate predictions on longer sequences, which seems to indicate that some features or parameters learned from short sequences are also effective in modelling longer sequences. Finally, our experiments show that the proposed model with dilated causal convolution layers outperforms the state-of-the-art model which consists of both CNN and LSTM layers. Future work includes presetting the "local network" with prior knowledge and fine tuning it during training. Moreover, we aim to apply the methods to new, ecologically valid emotional speech data.

References

[1] J. A. Russell, "Core Affect and the Psychological Construction of Emotion," *Psychol. Rev.*, vol. 110, no. 1, pp. 145–172, 2003.

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

[3] Z. Liu, M. Wu, W. Cao, J. Mao, J. Xu, and G. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271 – 280, Jan. 2018.

[4] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li, "Deep spectrum feature representations for speech emotion recognition," in *ASMMC-MMAC'18*, pp. 27–33, 2018.

[5] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '16)*, pp. 5200–5204, 2016.

[6] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, 2017.

[7] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '18)*, pp. 5089–5093, 2018.

[8] Z. Zhang, B. Wu, and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *Proc. 2019 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '19)*, pp. 6705–6709, 2019.

[9] P. Verduyn, L. Van Mechelen, E. Kross, C. Chezzi, and F. Van Bever, "The relationship between self-distancing and the duration of negative and positive emotional experiences in daily life," *EMOTION*, vol. 12, no. 6, pp. 1248–1263, 2012.

[10] P. Kuppens, Z. Oravecz, and F. Tuerlinckx, "Feelings Change: Accounting for Individual Differences in the Temporal Dynamics of Affect," *J. Pers. Soc. Psychol.*, vol. 99, no. 6, pp. 1042–1060, 2010.

[11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS*, 2014.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.

[13] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, 2014.

[14] S. Bai, J. Z.Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv*, 2018.

[15] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *ISCA*, 2016.

[16] A. van den Oord, N. Kalchbrenner, L. Espeholt, K. kavukcuoglu, O. Vinyals, and A. Graves, "Conditional image generation with pixelcnn decoders," in *NIPS '2016*, (Barcelona, Spain), pp. 4790–4798, 2016.

[17] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. - 2017 IEEE Conf. Comput. Vis. Pattern Recognition. (CVPR '2017)*, pp. 636–644, 2017.

[18] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *Proc. 2019 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '19)*, pp. 6675–6679, 2019.

[19] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. of IEEE Face & Gestures 2013*, pp. 22–26, 2013.

[20] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2015 - the 5th international audio/visual emotion challenge and workshop," in *Proc. of the 5th Audio/Visual Emotion Challenge and Workshop (AV+EC'15)*, pp. 3–8, 2015.

[21] M. Valstar, J. Gratch, B. Schuller, F. Ringevaly, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Panticz, "AVEC 2016 - Depression, mood, and emotion recognition workshop and challenge," in *Proc. of the 6th Audio/Visual Emotion Challenge and Workshop (AV+EC'16)*, pp. 3–10, 2016.

[22] L. Holmstrom and P. Koistinen, "Using additive noise in backpropagation training," *IEEE Trans. Neural Netw.*, vol. 3, no. 1, pp. 24–38, 1992.