Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition[☆]

Ziping Zhao[a,1], Qifei Li[a,1], Zixing Zhang[b], Nicholas Cummins[c,d], Haishuai Wang[e,a], Jianhua Tao[f,g,h,*], Björn W. Schuller[a,b,c]

[a] *College of Computer and Information Engineering, Tianjin Normal University, Tianjin, China*
[b] *GLAM – Group on Language, Audio, & Music, Imperial College London, UK*
[c] *Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany*
[d] *Department of Biostatistics and Health Informatics, IoPPN, King's College London, London, UK*
[e] *Department of Computer Science and Engineering, Fairfield University, USA*
[f] *School of Artificial Intelligence, University of Chinese Academy of Sciences, China*
[g] *National Laboratory of Pattern Recognition, Chinese Academy of Sciences, China*
[h] *CAS Center for Excellence in Brain Science and Intelligence Technology, China*

1. Introduction

The automatic identification of discrete emotional states from human speech has consistently been identified as a challenging task for machine learning algorithms. One considerable challenge currently being faced by researchers in the field of discrete SER is that of how best to extract discriminative, robust, and emotionally salient features from the acoustic content of a speech signal utilising a sequence-to-sequence model. The reasons for this are as follows:

Firstly, previous work on emotion recognition has focused primarily on the extraction of features that are carefully handcrafted and highly engineered. Results from works of this kind have repeatedly demonstrated the importance of discriminative spatio-temporal features in modelling the continual evolutions of different emotions. Moreover, as the amount of both available data and computational power have increased, deep learning methods are rapidly becoming the predominant approach in the SER context. In particular, many recent works in this field have leveraged either recurrent neural networks (RNNs) or

[☆] The work presented in this article was substantially supported by the National Natural Science Foundation of China (Grant No. 62071330,61702370), the National Science Fund for Distinguished Young Scholars, China (Grant No: 61425017), the Key Program of the National Natural Science Foundation of China (Grant No: 61831022), the Key Program of the Natural Science Foundation of Tianjin, China (Grant No. 18JCZDJC36300), the technology plan of Tianjin, China (Grant No: 18ZXRHSY00100).

[*] Corresponding author at: National Laboratory of Pattern Recognition, Chinese Academy of Sciences, China.

E-mail address: jhtao@nlpr.ia.ac.cn (J. Tao).

[1] Both authors contributed equally to this work.

deep convolutional neural networks (DCNNs) as feature extractors to facilitate the learning of discriminative representations, with varying degrees of success (Huang & Narayanan, 2016; Mirsamadi, Barsoum, & Zhang, 2017; Tzinis & Potamianos, 2017; Tzirakis, Trigeorgis, Nicolaou, Schuller, & Zafeiriou, 2017; Wang, Wu, Zhang, & Chen, 2018; Wöllmer, Eyben, Reiter, Schuller, Cox, Douglas-Cowie, & Cowie, 2008). Owing to the success of CNNs and RNNs, there has been increasing interest in incorporating both of these network types into a single architecture in order to capture both long-term and local dependencies (Chen, He, Yang, & Zhang, 2018; Sainath, Vinyals, Senior, & Sak, 2015).

However, such frameworks are affected by some limitations. For example, although long short-term memory (LSTM) recurrent neural networks demonstrate powerful capacity for sequence modelling (Krause, Lu, Murray, & Renals, 2016), the current state cannot be calculated without the results of previous states, making it impossible for these calculations to be conducted in parallel; moreover, it is difficult for LSTM to deal with long-range temporal dependencies, and they converge with a low speed in training. By contrast, the training of CNNs does not depend on the computations of the previous time step, making it possible to implement parallelisation over every element in sequence (Pu, Zhou, & Li, 2018). However, when applying CNNs to SER tasks, a disadvantage of CNNs is that the temporal structure of speech will be gradually lost during this process while the progressive downsampling provides a strong ability to conduct local context modelling and emotion-related pattern detection. As the temporal evolution of speech is assumed to be highly related to emotions, such loss of spatial information may hamper the effectiveness of the SER system (Li, Wu, Jia, Zhao, & Meng, 2019b; Yu, Koltun, & Funkhouser, 2017). Moreover, it has been shown that enlarging the receptive field is an effective means of improving CNN performance (Wang, Sun, & Hu, 2017). Thus, the question of how to better encode spatial relationships and efficiently learn representations efficiently without losing resolutions for CNN-based SER system has become increasingly one.

Recent studies have shown that parallel convolutional layers can be used to extract temporal information at multiple resolutions from the data provided, which can improve the system performance (Latif, Rana, Khalifa, Jurdak, & Epps, 2019). Additionally, the Squeeze-and-Excitation Network (SEnet) has achieved impressive image classification results (Hu, Shen, & Sun, 2018); under this approach, a channel-wise transform is appended to existing DNN building blocks, such as the Residual unit (Hu et al., 2018). Moreover, the representations produced by CNNs can be strengthened through the integration of a Squeeze-and-Excitation (SE) block, which is an architectural unit designed to improve a network's representational power by enabling it to perform dynamic channel-wise feature recalibration into the network that helps to capture the spatial correlations between features (Hu et al., 2018).

Recent successes achieved by Residual Networks (ResNet) (He, Zhang, Ren, & Sun, 2016) approaches on various computer vision tasks prove that ResNet has better image representation capacity than other deep architectures. Furthermore, another approach is the Dilated Residual Network (DRN) (Yu et al., 2017), which utilises dilated convolutions in residual blocks and inherits the properties of a residual network, such that the temporal structure of the network's input signals is maintained. Such a network can also compensate for any reduction in the receptive field, thereby demonstrating its strong ability to model local context with dilation. Recent studies have also indicated that the DRN particularly excels at capturing contextual information, meaning that it can achieve performance that is comparable or superior to that of LSTM across a diverse range of tasks and datasets – including *audio generation* (Oord, Dieleman, Zen, Simonyan, Vinyals, Graves, Kalchbrenner, Senior, & Kavukcuoglu, 2016) and *continuous sign language recognition* (Pu et al., 2018) – while demonstrating longer effective memory (Bai, Kolter, & Koltun, 2018). Moreover, the self-attention mechanism (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, & Polosukhin, 2017), developed in both the encoder–decoder and feed-forward contexts, has led to faster training and state-of-the-art results in several natural language processing (NLP) tasks (Li, Song, Gao, Liu, Huang, He, & Gan, 2019a; Scialom, Piwowarski, & Staiano, 2019; Shen, Zhou, Long, Jiang, Pan, & Zhang, 2018; Vaswani et al., 2017); in addition, and more recently, the application of DRNs combined with a self-attention mechanism has produced promising SER results (Li et al., 2019b; Tarantino, Garner, & Lazaridis, 2019).

Secondly, many works in this field treat the discrete SER task as a typical sequence classification problem in which each chunk of speech (such as an utterance) has exactly one label. However, this type of a conventional sequence-to-label modelling approach is less than ideal for discrete SER. A critical underlying issue is the loss of dynamic temporal information, which can strongly reflect changes in emotional states (Han, Ruan, Chen, Wang, Li, & Schuller, 2018). In order to tackle this problem, the *connectionist temporal classification* (CTC) loss function (Graves, Fernández, Gomez, & Schmidhuber, 2006) with bidirectional long short-term memory (BLSTM) networks, which have been widely investigated in the speech context (Graves & Jaitly, 2014), has been found to be effective in the SER context (Han et al., 2018; Zhao, Bao, Zhang et al., 2019). However, the training speed can be very slow and the training process for BLSTM modelling is difficult.

Accordingly, to solve these issues, some researchers have explored the application of CNNs to CTC for long-range temporal dependencies (Wang, Deng, Pu, & Huang, 2017; Zhang, Pezeshki, Brakel, Zhang, Laurent, Bengio, & Courville, 2016). Although this approach can greatly improve the training speed, CNNs may perform poorly under these circumstances due to the receptive field being insufficiently large. To broaden the receptive fields of CNNs and enhance their sequence modelling ability, moreover, we propose to replace the recurrent layers for CTC with a self-attention Dilated Residual Network in this work.

Motivated by the above observations, this article presents a novel hybrid sequence-to-sequence modelling solution for the task of discrete SER. Our approach is based on the combination of PCN, integrated with SEnet (PCNSE), and self-attention-DRN-CTC, namely PCNSE-SADRN-CTC, in order to retain high temporal structure resolution in feature learning, while employing a similar receptive field size to CNN-based approaches. Meanwhile, we use dilated convolutions with CTC loss to model the dependencies between different frames. When compared with RNN-CTC (Han et al., 2018), we believe that this combination of different networks can further improve the performance of discrete SER while also speeding up training. In this work, inspired by the positive results of 3D log Mel spectrum features in the SER context (Chen, He et al., 2018; Meng, Yan, Yuan, & Wei, 2019), we first employ log-Mel, deltas, and delta–deltas as 3D input to the CNN model. We utilise this 3D input as the delta. Moreover, the delta–delta features are able to effectively capture the effects of emotion in speech (Chen, He et al., 2018), while also being less susceptible to the impact of non-relevant acoustic factors. Log Mel-spectrograms are extracted from the set of acoustic features, after which we calculate deltas and delta–deltas for the log Mel-spectrogram to make up the 3-D data as the input of PCN. Subsequently, three parallel 2-D convolutional layers with different filter sizes are utilised to capture both the long-term and short-term changes from 3-D spectrograms. Finally, we fuse each DRN-CTC block with the self-attention mechanism in order to further improve our model. The features extracted from the PCNSE block are then fed into SADRN-CTC for classification.

Our two main contributions can, therefore, be summarised as follows: (1) We have developed a parallel CNN stacked on a self-attention DRN, paired with CTC loss for discrete SER, which operates on both the time and frequency dimensions. The proposed model can model temporal as well as spectral local correlations and achieve translational invariance in speech signals; (2) the presented results demonstrate the effectiveness of this sequence-to-sequence modelling solution for discrete SER tasks.

2. Related work

SER is a highly active research field, with many novel approaches being proposed and investigated over the past decade. Due to the increases in the amount of available data and computational power, deep learning methods are rapidly becoming the predominant approach in this area (Wang, Cui, Chen, Avidan, Abdallah, & Kronzer, 2018; Wang, Wu, Pan, Zhang, & Chen, 2017; Wang, Zhang, Wu, Pan, & Chen, 2019). In particular, much of the recent research in this domain has explored leveraging of deep neural networks as feature extractors in order to learn discriminative representation (Wang et al., 2018). Moreover, due to their success in many visual recognition tasks, CNNs have been widely adopted for feature representation learning in various speech analysis tasks. For example, Huang et al. (2014) used spectrograms of speech, together with CNN, to perform SER (Huang, Dong, Mao, & Zhan, 2014). Similar work was also presented by Mao, Dong, Huang, and Zhan (2014), in which a CNN was employed to learn affect-salient features from spectrograms. In addition to being successfully applied to automatic speech recognition (ASR) (Abdel-Hamid, Mohamed, Jiang, Deng, Penn, & Yu, 2014) and speaker identification tasks (Nagrani, Chung, & Zisserman, 2017), CNNs have also achieved promising results compared with conventional approaches when applied to SER (Poria, Chaturvedi, Cambria, & Hussain, 2016).

Furthermore, given that context information is crucial to the detection of emotional states, RNN paradigms are widely used in SER to exploit the temporal information inherent in speech signals. LSTM-RNNs, in particular, are frequently employed in SER tasks (Huang & Narayanan, 2016; Mirsamadi et al., 2017; Tzinis & Potamianos, 2017; Tzirakis et al., 2017; Wöllmer et al., 2008).

Due to the positive results obtained by CNNs and RNNs, there has been increasing interest in incorporating them both into a single architecture. For example, in Sainath et al. (2015), the Convolutional Long Short-Term Memory Deep Neural Network (CLDNN) model for speech recognition was proposed; this approach consists of convolutional layers, LSTM gated recurrent layers, and fully connected (FC) layers. In Chen, He et al. (2018), moreover, a 3-D attention-based convolutional Recurrent Neural Network (ACRNN) was proposed for SER. This model combines CNN with LSTM, while the 3-D spectral features of the segments are employed as input. Another promising network structure that has recently been developed is the end-to-end network architecture, which can automatically and directly extract representations from *raw* (unprocessed) data, thereby removing the need to manually extract hand-crafted features. In addition, the SER approach proposed in Tzirakis et al. (2017) jointly exploited a CNN (to automatically extract suitable representations from raw audio signals) and an LSTM-RNN (to capture the required temporal information). A similar framework was proposed in Ma, Yang, Chen, Huang, and Wang (2016) for the related task of speech-based depression detection. Finally, in Ma, Wu, Jia, Xu, Meng, and Cai (2018), a specially designed neural network structure that accepts speech segments of variable length was proposed for SER; this approach combines CNN-based deep spectrogram representations with an RNN in order to handle the variable-length speech segments.

Recent research results suggest that Dilated Residual Networks have been convincingly shown to achieve both compelling convergence and high accuracy in the computer vision (Yu et al., 2017) and speech analysis contexts (Oord et al., 2016). Successful attempts along this line have also been reported very recently in the SER context (Li et al., 2019b).

Recently, attention mechanisms have seen widespread adoption within the deep learning community. Although the combination of an attention mechanism with RNNs has improved performance in SER tasks, this is limited by the state of the cell (e. g., LSTM), which can contain only a limited amount of information. Meanwhile, this approach is also impacted by the exploding and vanishing gradient problems (Pascanu, Mikolov, & Bengio, 2013). The self-attention mechanism, which can help with the capturing of long-term contextual dependencies, was proposed in Vaswani et al. (2017). This mechanism has been proven capable of capturing contextual dependencies in several NLP tasks (Li et al., 2019a; Scialom et al., 2019; Shen et al., 2018; Vaswani et al., 2017), and, more recently, has produced state-of-the-art SER results (Tarantino et al., 2019).

As an end-to-end acoustic modelling method, CTC based on recurrent (RNNs) or convolutional neural networks (CNNs), has exhibited strong performance in the speech-related tasks such as end-to-end speech recognition systems (Leung, Liu, & Meng, 2019; Shi, Hwang, & Lei, 2019; Wang, Deng et al., 2017). To date, however, work exploiting CTC models for discrete SER has been very limited (Chen, Han et al., 2018; Chernykh & Prikhodko, 2017; Han et al., 2018).

From the literature discussed above, we can see that recent works present strong evidence for the value added by PCN-SEnet and self-attention DRN. Accordingly, our proposed method utilises a combination of these existing approaches, paired with CTC loss, for discrete SER. To the best of our knowledge, no existing work has yet combined these methods for such a task.

3. Methodology

As noted above, CNN and CTC both possess features that make them highly suited to the discrete SER task, although the combination of these two components has not been fully explored. In this section, we outline the main steps required to implement the proposed model. We first describe the parallel 2-D convolutional layers embedded with SEnet to create the feature extraction block. Next, we introduce the stacked multi-layered DRN with self-attention mechanism and the CTC loss function.

3.1. System overview

The architecture of the proposed model comprises four main components (Fig. 1): (i) an *input layer*, where 3-D spectrograms are used as the model input; (ii) a *feature extraction layer*, designed to derive a high-level representation from step (i), using PCN integrated with SEnet; (iii) a *Self-Attention Dilated Residual Network*, in which the SADRN is used to model long-range dependencies; and (iv) a *CTC layer*, in which the CTC model is used to automatically align emotional labels to emotionally salient frames.

3.2. 3D Log-Mels spectrogram generation

In recent years, excellent results have been achieved through the application of CNNs to capture information in the spectrograms for SER (Cummins, Amiriparian, Hagerer, Batliner, Steidl, & Schuller, 2017; Zhao, Bao, Zhao et al., 2019). However, static spectrograms can contain personalised information about the speaker, which can negatively influence the SER performance (Chen, He et al., 2018). Inspired by the successful use of 3D log-Mel spectrograms for SER (Meng et al., 2019), our hybrid system also uses

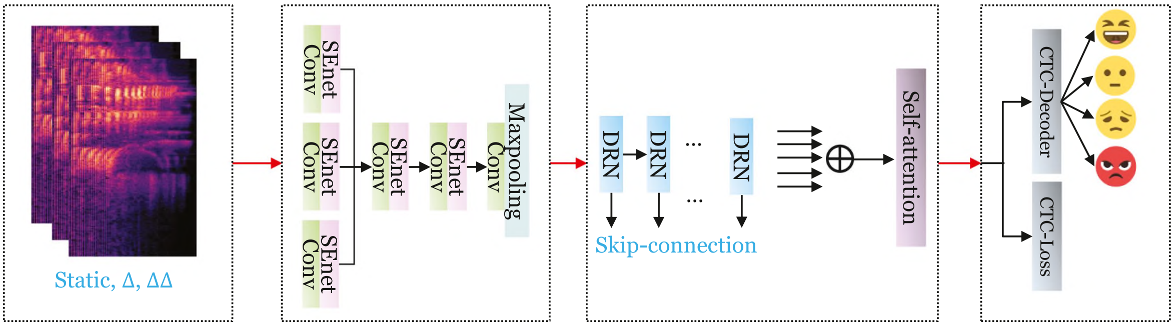


Fig. 1. Framework of the proposed PCNSE-SADRN-CTC model.

log-Mels, together with deltas and delta–deltas, as the inputs to the DCNN.

First, we split the raw speech signal into short frames with Hamming windows of 25 ms and a 10 ms shift. Next, the power spectrum for each frame is calculated and passed through the Mel-filter bank i to produce the output p_i. A logarithmic operation is then conducted on p_i to obtain the log-Mel spectrogram m_i. Finally, we calculate the m_i^d feature, which is the deltas of m_i obtained via formula (1), while the value of N is set to 3. Similarly, the delta–deltas features m_i^{dd} are calculated by taking the derivative of the deltas, as shown in Eq. (2).

$$m_i^d = \frac{\sum_{n=1}^{N} n(m_{i+n} - m_{i-n})}{2\sum_{n=1}^{N} n^2}, \tag{1}$$

$$m_i^{dd} = \frac{\sum_{n=1}^{N} n(m_{i+n}^d - m_{i-n}^d)}{2\sum_{n=1}^{N} n^2}. \tag{2}$$

After the above calculations are complete, we obtain a three-dimensional feature representation $X \in R^{t\times f\times c}$ for use as the input of the DCNN model, where t denotes the length of frame, while f represents the number of Mel-filter banks. In our work, f is set to 80, while c is 3, representing the static, deltas, and delta–delta log-mel spectrogram respectively.

3.3. PCNSE model

For the feature extraction block, we use parallel convolutional layers with multiple filter lengths to capture both long-term and short-term interactions from the 3-D spectrograms. The PCN in this article consists of three parallel 2-D convolutional layers, followed by three layers of 2-D convolution (Fig. 2a). We then concatenate the outputs of these three pooling layers in order to obtain features with multiple temporal resolutions.

3.3.1. Squeeze-and-excitation block

In our work, all 2-D convolution layers are followed by an SE block. This is done to enhance the model's expression ability by utilising the relationship between the various channels of the convolution feature (Fig. 2b).

The SEnet acts as a computational unit for any transformation, as follows: $F_{tr} : X \to U, X \in \mathbb{R}^{W'\times H'\times C'}, U \in \mathbb{R}^{W\times H\times C}$.

The outputs of F_{tr} are represented as $U[u_1, u_2, \ldots, u_c]$, where:

$$u_c = v_c * X = \sum_{s=1}^{C'} v_c^s * x^s. \tag{3}$$

Here, the convolution operation is denoted by *, while the 2-D spatial kernel is indicated by v_c^s. The single channel of v_c acts on the corresponding channel of X. As outlined in Hu et al. (2018), channel interdependence is simulated in two steps – namely ***squeeze*** and ***excitation*** – in order to adjust the filter response.

The ***squeeze*** operation makes use of a global average pooling to generate channel-wise statistics by utilising the contextual information outside of the local receptive field. The output of the transformation, U, is shrunk through spatial dimensions $W \times H$ to enable the computation of the channel-wise statistics, $z \in \mathbb{R}^c$. The cth element of z is calculated as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H} u_c(i,j). \tag{4}$$

The role of the ***excitation*** operation, moreover, is to aggregate the information obtained by the ***squeeze*** operation in order to capture the dependencies between the channels. In order to achieve this, two full connection layers are employed, as follows:

$$s = F_{ex}(z, W) = \sigma(W_2\delta(W_1, z)). \tag{5}$$

Here, δ and σ denote the Rectified Linear Unit (ReLU) and Sigmoid activation functions, respectively; moreover, $W_1 \in \mathbb{R}^{\frac{c}{r}\times c}$ and $W_2 \in \mathbb{R}^{c\times\frac{c}{r}}$ are all of the training parameters, while r is determined empirically to have a value of 8. s is the output of the ***excitation*** operation and can be regarded as a set of channel weights. Finally, the output of SEnet is represented as follows:

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c \cdot u_c, \tag{6}$$

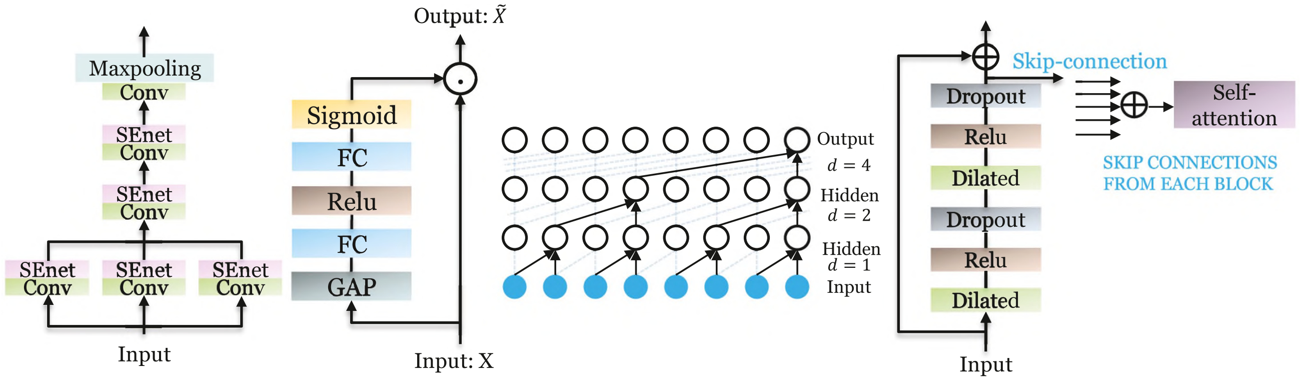
where $\tilde{X} = \left[\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_C\right]$ and $F_{scale}(u_c, s_c)$ refer to the channel-wise multiplication between the scalar s_c and the feature map $u_c \in \mathbb{R}^{H\times W}$.

3.4. Dilated residual network

Inspired by the work presented in Bai et al. (2018), Li et al. (2019b), Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke, and Rabinovich (2015), the Dilated Residual Network employed in this article comprises five groups of 1-D temporal convolutional layers when modelling long-range dependencies (Fig. 2d). By skipping input values with a certain step size, the network can increase the size of the receptive field without the need for a high number of convolutional layers or large filter sizes. In more detail, for a 1-D sequence input $X \in \mathbb{R}^n$ and a filter f of size k, the dilated convolution operation F on element s of the sequence can be defined as follows:

$$F(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d\cdot i}. \tag{7}$$

Here, d is the dilation factor, the value of which is always an exponent of 2; moreover, k is the filter size, while $s - d \cdot i$ represents the previous direction. In other words, we can increase the DRN's receptive field size by increasing the dilation factor and using a larger filter size, where the effective history r of one such layer is $r = (k - 1)d$. Generally speaking, when using the dilated



(a) PCN (b) SEnet (c) Dilated Convolution (d) Self-attention-based DRN

Fig. 2. Architecture of different blocks in our proposed model.

convolution, we increase the dilation factor d exponentially with the depth of the network (i. e., $d = 2^i$ at number i of the DRN).

The reason for the use of a residual connection here is that a residual network is able to maintain the temporal structure of the inputs from PCNSE. Within a residual block, the DRN contains two dilated convolution layers, which are activated by *ReLU*. Subsequently, a spatial dropout is added after each dilated convolution for regularisation purposes. The output of the top dropout layer has two branches: one is for the skip-connection from each block, while the other is for the next DRN.

3.5. Self-attention

Self-attention is an attention technique based on an encoder–decoder structure. This approach does not employ any form of recurrence; instead, it uses weighted correlations between the elements of the input sequence (Vaswani et al., 2017). In this paradigm, the encoder maps an input sequence into several attention matrices, while the decoder subsequently uses these matrices to generate a new output token. The *Transformer*, the model that utilises *self-attention*, has been demonstrated to achieve state-of-the-art performance in several NLP tasks, and with a computing cost that is one or two orders of magnitude (depending on the size of the model) lower than that of conventional RNNs (Li et al., 2019a; Scialom et al., 2019; Shen et al., 2018). It should be noted here that this section focuses only on the implementation of the encoder, as a decoder is not required by our proposed hybrid network.

Self-attention calculates queries, keys (properties of the input) and values (the output) for the frames in a given hidden sequence H by means of linear transformation of the input sequence X, as follows:

$$Q = W_q X;\ K = W_k X;\ V = W_v X, \tag{8}$$

where the matrices Q, K, and V denote the set of queries, keys and values respectively of an input/output sequence, while W_q, W_k, and W_v represent the learnt linear operations. A scaled dot-product operation is performed on the query and key to obtain the similarity weights, which are then normalised by the softmax function. The attention matrix is calculated as follows:

$$Z = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{9}$$

where d_k is a scaling factor, set as the dimensionality of K.

Moreover, Z is the attention matrix ($N \times d_k$), where N denotes the number of elements in the input sequence.

3.6. CTC approach

The CTC model uses a loss function for sequence labelling that is able to account for the input and the target label sequence of different lengths, without the need for any pre-segmentation. The key concept underpinning CTC is to introduce a blank label, *Null* (meaning the network generates no label). This addition enables the network to suppress frame-wise outputs, including repetitions of the same labels, into the sequence of target outputs (e. g., phonemes or characters).

When fed an input sequence $X = (x_1, \ldots, x_T)$, CTC trains the model to maximise the probability distribution $P(l|X)$ for the corresponding target label sequence l of length $U(\leq T)$. CTC expresses this distribution as a summation of all possible frame-level intermediate representations

$\pi = (\pi_1, \ldots, \pi_T)$ (hereafter referred to as the CTC path):

$$P(l|X) = \sum_{\pi \in \Phi(l)} P(\pi|X), \tag{10}$$

where $\Phi(l)$ denotes the set of CTC paths that allow for the insertion of *Null* and the repetition of non-blank labels to l, i. e., $\Phi^{-1}(\pi) = l$, noting that if $l_u \in L = \{1, \ldots, K\}$, the softmax layer is composed of $|L \cup \{blank\}| = K + 1$ units. Based on the conditional independence assumption, the decomposition of the posterior $P(\pi|X)$ is given by the following:

$$P(\pi|X) = \prod_{t=1}^{T} y^t_{\pi t}, \tag{11}$$

where y^t_k is the kth output of the softmax layer at time t. This can be interpreted as the occurrence probability of the corresponding label. The probability distribution $P(l|X)$ can be efficiently computed using the forward–backward algorithm. The detailed CTC training process is described in Zhao, Bao, Zhang et al. (2019).

4. Experiments and results

To demonstrate the effectiveness of the proposed methods, we performed a set of experiments on the popular interactive emotional dyadic motion capture dataset (IEMOCAP) (Busso, Bulut, Lee, Kazemzadeh, Mower, Kim, Chang, Lee, & Narayanan, 2008) and FAU Aibo Emotion corpus (FAU-AEC) (Steidl, 2009).

4.1. Datasets

IEMOCAP is a well-known corpus containing audio–visual data recordings and transcriptions of dialogues between two actors

Table 1
Instance distribution over four emotion classes – **N**eutral, **H**appy, **S**ad, and **A**ngry – of the IEMOCAP Dataset.

Session	N.	H.	S.	A.	Total
1	223	33	104	62	521
2	217	54	100	22	530
3	198	60	190	90	627
4	174	31	81	84	534
5	287	106	133	31	731
Sum	1099	284	608	289	2280

Table 2
Instance distribution over five emotion classes – **A**ngry, **E**mphatic, **N**eutral, **P**ositive, and **R**est – of the FAU Aibo Emotion Corpus.

	A.	E.	N.	P.	R.	Total
Train	881	2093	5590	674	721	9959
Test	611	1508	5377	215	546	8257
Sum	1492	3601	10967	889	1267	18216

Table 3
Class weights for data balance when using the FAU Aibo Emotion Corpus.

	Angry	Emphatic	Neutral	Positive	Rest
weight	1.1	0.5	0.2	1.5	1.4

(Busso et al., 2008). The corpus is divided into two parts: *improvise* and *script*. In our experiments, we used only the former part in order to reduce the potentially confounding effect of semantic information disturbance. The final number of instances for each emotion class are presented in Table 1.

The second corpus used for evaluation is the FAU Aibo Emotion Corpus, which is made up of spontaneous and emotional German speech samples (Steidl, 2009). The corpus contains 9.2 h of German speech obtained from a total of 51 children at two different schools during their interactions with Sony's pet robot Aibo. Following Schuller, Steidl, and Batliner (2009), we used 9 959 utterances from 26 children (13 males, 13 females) as the training set, and 8 257 utterances from 25 children (8 males, 17 females) as the test set. We further concentrated on the five-class problem, utilising the emotion categories of anger, emphatic, neutral, positive, and rest. The final number of instances of each emotion class are listed in Table 2.

4.2. Features

Our spectrograms were created using the extraction process described in Chen, He et al. (2018). In brief, each spectrogram was constructed using the output of a 40-dimensional mel-scale log filter bank. The features were computed over frames 25 ms in length and with a 10 ms stride. At the final step, we calculated the delta and delta–deltas of the spectrogram, which reflect the process of emotional change.

4.3. Experimental setup and evaluation metrics

On the IEMOCAP dataset, we performed a 5-fold cross-validation using a leave-one-session out strategy, in line with the methodology outlined in previous work (Zhao, Bao, Zhang et al., 2019; Zhao, Zheng, Zhang, Wang, Zhao, & Li, 2018). Each training process involved the use of eight speakers from four sessions as training data; the remaining session was separated into two parts, one of which was regarded as validation data and the other as test data. For the FAU Aibo Emotion Corpus, moreover, we followed the Interspeech 2009 Emotion Challenge guidelines (Schuller et al., 2009), consequently employing utterances from one school (the "Ohm-Gymnasium") for training and the other (the "Montessori-Schule") for testing.[2]

The proposed PCNSE-SADRN-CTC model has a large number of hyperparameters, a proportion of which were tuned based on recommendations from previous works that had utilised the same database. We also used the same feature extraction block, consisting of three parallel convolutional layers, which provided multiple temporal dependencies for comparison. *ReLU* is used as the activation function. In this article, all models were implemented using the *keras*[3] framework. Furthermore, cross-entropy was selected as the loss function for the PCNSE block, while the Adam optimiser (with an initial learning rate of 3×10^{-4} and a decay of 10^{-6}) was used for training; for the dilated residual network, moreover, CTC loss and the Adadelta optimiser (with an initial learning rate of 4×10^{-3} and a decay of 10^{-6}) were utilised for training. Additional details regarding these hyperparameters are presented in Table 4.

Standard evaluation criteria were used to evaluate the results generated by the two datasets. For the IEMOCAP-generated results, unweighted and weighted accuracies (UA and WA respectively) were used as the evaluation metrics. For the FAU-AEC-generated results, moreover, we considered only unweighted accuracy (UA), since the FAU Aibo Emotion corpus is extremely unbalanced. Furthermore, in order to tackle the problem of unbalanced data, we applied class weights during training (cf. Table 3) as outlined in Zhao, Bao, Zhao et al. (2019).

4.4. Results and discussion

This section presents the results of our experiments, with the aim of verifying the efficiency of our proposed PCNSE-SADRN-CTC model. We first performed an ablation analysis to elucidate the benefits of incorporating PCN, SE block, DRN, self-attention mechanism, and CTC loss into the final proposed model. The effectiveness of our hybrid framework is further highlighted through comparison with other key results obtained in the literature on the IEMOCAP and FAU-AEC datasets (see Table 5). The state-of-the-art models utilised for comparison purposes listed in Table 5 include three methods that have previously achieved good performance on IEMOCAP. The models that employ PCNSE are used alone, while the PCN model without SE block, the PCNSE-DRN model with different pooling strategies, the PCNSE model with self-attention-BLSTM, and our proposed model based on cross entropy loss are also compared with our proposed approach.

From the results, it can be seen that the proposed approach outperforms previous works on the IEMOCAP dataset (cf. Table 5). For IEMOCAP, the best WA (73.1%) and UA (66.3%) were attained by our proposed PCNSE-SADRN-CTC model, which achieved a significant improvement relative to the baseline CNN-BLSTM model presented in Satt et al. (2017) ($p < 0.05$ in a one-tailed z-test). The same system setup achieved a UA of 41.1%, on the FAU-AEC dataset; this is slightly lower than the best performance achieved by the BLSTM-CTC model (Zhao, Bao, Zhang et al., 2019). However, compared with our proposed model, the use of BLSTM combined with CTC loss function has several drawbacks. The sequential dependencies make the computation of BLSTM via parallel GPU acceleration difficult, which leads to slow training and inference when modelling sequences. With the increasing of the model size and the amount of hyper-parameters, more time

[2] All results are entirely reproducible by others. To ease the procedure, upon acceptance, we will provide a URL for a document containing the details of all partitions and seeds.

[3] https://github.com/keras-team/keras

Table 4
Parameters of the proposed PCNSE-SADRN-CTC and the PCNSE-ABLSTM model. SABLSTM denotes the self-attention-based BLSTM.

Network	Layer	Shape	Dilations
PCNSE-SADRN-CTC	Input	500 × 40 × 3	–
	Conv_1(left)	1 × 1 × 64	–
	Conv_2(mid)	3 × 3 × 64	–
	Conv_3(right)	5 × 5 × 64	–
	Concatenate	–	–
	Conv_4	8 × 8 × 64	–
	Conv_5	3 × 3 × 128	–
	Conv_6	3 × 3 × 64	–
	Maxpooling2D	2 × 2	–
	Reshape	243 × 832	–
	DRN	2 × 2 × 128	(1,2,4,8,16,32,64,128)
	Attention	–	–
	Output	128	–
PCNSE-SABLSTM	Input	500 × 40 × 3	–
	Conv_1(left)	1 × 1 × 64	–
	Conv_2(mid)	3 × 3 × 64	–
	Conv_3(right)	5 × 5 × 64	–
	Concatenate	–	–
	Conv_4	8 × 8 × 64	–
	Conv_5	3 × 3 × 128	–
	Conv_6	3 × 3 × 64	–
	Maxpooling2D	2 × 2	–
	Reshape	243 × 832	–
	BLSTM	64 × 2	–
	Attention	–	–
	Output	128	–

Table 5
Performance comparison between the proposed PCNSE-SARDN-CTC model, the PCNSE-ABLSTM model with other models on the IEMOCAP and FAU Aibo Emotion corpus. SABLSTM denotes the self-attention-based BLSTM.

Methods	IEMOCAP		FAU-AEC
[%]	WA	UA	UA
CNN-BLSTM (Satt, Rozenberg, & Hoory, 2017)	68.8	58.4	–
CNN-GRU (Ma et al., 2018)	71.5	64.2	–
BLSTM-CTC (Han et al., 2018)	66.9	65.1	41.4
PCN	68.6	56.8	38.4
PCNSE	69.8	58.5	38.8
PCN-SABLSTM	70.8	62.7	39.8
PCNSE-SABLSTM	72.1	65.4	40.5
PCN-SADRN	71.1	62.5	41.1
PCNSE-DRN w/ Global max-pooling	71.5	62.0	38.1
PCNSE-DRN w/ Global average-pooling	71.2	62.5	39.5
PCNSE-SADRN	72.5	65.0	40.4
PCNSE-SADRN-CTC	**73.1**	**66.3**	41.1

Note : For IEMOCAP, we provide both unweighted and weighted accuracies (UA and WA respectively) as the evaluation metric; for FAU-AEC, moreover, we only adopt UA as the evaluation measure, since this dataset is extremely unbalanced.

is needed in training and hyper-parameters tuning when using BLSTM-CTC.

As for the Squeeze-and-Excitation Network introduced in this work, the performance of the model with no SE block can be observed to be lower than that achieved by the model integrated within the SE block on both the IEMOCAP and FAU-AEC datasets. Moreover, the PCNSE model performs better than the PCN model, while the performance of the PCNSE-SADRN model is better than that of the PCN-SADRN model and the PCNSE-SABSTM model outperformed the PCN-SABLSTM model. From these results, we can conclude that incorporating an SE block into a PCN model such as ours is an effective solution that is well suited for SER applications.

We also compared the performance of PCNSE-SADRN with the PCNSE-DRN model, without self-attention, to determine the benefits of using a self-attention mechanism in our proposed model for SER. As can be seen from Table 5, regardless of which kind of pooling strategy was adopted in the PCNSE-DRN model, the experimental results are all lower than those achieved when a self-attention mechanism was employed on both the IEMOCAP and FAU-AEC datasets.

For the system trained with CTC loss, we can observe that our proposed method yields the best performance in terms of both WA and UA on the IEMOCAP and FAU-AEC (only UA given) datasets, outperforming those models that utilise a cross entropy loss function.

In summary, the present results demonstrate that our proposed model achieves notable performance improvements on both the IEMOCAP and FAU-AEC datasets compared to other existing methods, which demonstrates the effectiveness of our proposed hybrid network. Furthermore, we observed that the performance of the combined PCNSE and self-attention-DRN-CTC is superior to that of either of these two methods when they are used alone. This validates our hypothesis that the combination of PCNSE and self-attention-DRN-CTC results in additional improvement, as well as it reveals that our proposed PCNSE module can effectively extract features from 3D spectrograms, making it well-suited to SER tasks.

We can further observe that, although the performance of the PCNSE-SABLSTM model is slightly inferior to that of the PCNSE-SADRN-CTC, it still surpasses that of most other existing baseline models in terms of WA and UA on both the IEMOCAP and FAU-AEC corpora. In addition, the UA result achieved by the PCNSE-SABLSTM model is even better than that obtained by our proposed PCNSE-SADRN model on the IEMOCAP and FAU-AEC datasets.

5. Conclusion

In this article, we proposed a novel deep CNN architecture, called PCNSE-SADRN-CTC, which leverages PCN integrated with an SEnet combined with a self-attention DRN trained with CTC loss for discrete SER applications. Our proposed model takes full advantage of the long-range dependencies and local information contained in speech sequences. Experimental results indicate that our proposed model, by utilising 3-D spectrograms, achieves state-of-the-art performance on the IEMOCAP and FAU-AEC datasets, and can also be trained more efficiently on long utterances. This suggests that convolutional architectures can act as a replacement for recurrent ones in the speech emotion recognition context. Moreover, the experimental results are representative to reveal the effectiveness of our proposed CTC-based system combination.

In future work, we plan to further explore the potential of our proposed model by determining its suitability for other speech and acoustic recognition tasks.

CRediT authorship contribution statement

Ziping Zhao: Conceptualization, Writing - original draft. **Björn W. Schuller:** Data curation, Writing - original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language processing*, *22*(10), 1533–1545.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335.

Chen, X., Han, W., Ruan, H., Liu, J., Li, H., & Jiang, D. (2018). Sequence-to-sequence modelling for categorical speech emotion recognition using recurrent neural network. In *Proc. 1st AAAC Asian conference on affective computing and intelligent interaction* (pp. 1–6). Beijing, China.

Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, *25*(10), 1440–1444.

Chernykh, V., & Prikhodko, P. (2017). Emotion recognition from speech with recurrent neural networks. arXiv preprint arXiv:1701.08071.

Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., & Schuller, B. (2017). An image-based deep spectrum feature representation for the recognition of emotional speech. In *Proc. 25nd ACM international conference on multimedia* (pp. 478–484). Mountain View, California, USA.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with Recurrent Neural Networks. In *Proc. 23rd international conference on machine learning* (pp. 369–376). Pittsburgh, Pennsylvania, USA.

Graves, A., & Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *Proc. 31st international conference on machine learning* (pp. 1764–1772). Beijing, China.

Han, W., Ruan, H., Chen, X., Wang, Z., Li, H., & Schuller, B. (2018). Towards temporal modelling of categorical speech emotion recognition. In *Proc. INTERSPEECH* (pp. 932–936). Hyderabad, India.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proc. IEEE conference on computer vision and pattern recognition* (pp. 770–778). Las Vegas, NV, USA.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proc. IEEE conference on computer vision and pattern recognition* (pp. 7132–7141). Salt Lake City, Utah.

Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014). Speech emotion recognition using CNN. In *Proc. 22nd ACM international conference on multimedia* (pp. 801–804). Orlando, Florida, USA.

Huang, C. W., & Narayanan, S. S. (2016). Attention assisted discovery of sub-utterance structure in speech emotion recognition. In *Proc. INTERSPEECH* (pp. 1387–1391). San Francisco, California, USA.

Krause, B., Lu, L., Murray, I., & Renals, S. (2016). Multiplicative LSTM for sequence modelling. arXiv preprint arXiv:1609.07959.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., & Epps, J. (2019). Direct modelling of speech emotion from raw speech. In *Proc. INTERSPEECH* (pp. 3920–3924). Graz, Austria.

Leung, W.-K., Liu, X., & Meng, H. (2019). CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis. In *Proc. 44th international conference on acoustics, speech and signal processing* (pp. 8132–8136). Brighton, UK.

Li, X., Song, J., Gao, L., Liu, X., Huang, W., He, X., et al. (2019). Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proc. 33rd AAAI conference on artificial intelligence* (pp. 8658–8665). Honolulu, Hawaii, USA.

Li, R., Wu, Z., Jia, J., Zhao, S., & Meng, H. (2019). Dilated residual network with multi-head self-attention for speech emotion recognition. In *Proc. 44th international conference on acoustics, speech and signal processing* (pp. 6675–6679). Brighton, UK.

Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., & Cai, L. (2018). Emotion recognition from variable-length speech segments using deep learning on spectrograms. In *Proc. INTERSPEECH* (pp. 3683–3687). Hyderabad, India.

Ma, X., Yang, H., Chen, Q., Huang, D., & Wang, Y. (2016). DepAudioNet: An efficient deep model for audio based depression classification. In *Proc. 6th international workshop on audio/visual emotion challenge* (pp. 35–42). Amsterdam, The Netherlands.

Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, *16*(8), 2203–2213.

Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access*, *7*, 125868–125881.

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using Recurrent Neural Networks with local attention. In *Proc. 42nd international conference on acoustics, speech and signal processing* (pp. 2227–2231). New Orleans, LA, USA.

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). Voxceleb: A large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proc. 30th international conference on machine learning* (pp. 1310–1318). Atlanta, Georgia, USA.

Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016). Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Proc. IEEE international conference on data mining* (pp. 439–448). Barcelona, Spain.

Pu, J., Zhou, W., & Li, H. (2018). Dilated convolutional network with iterative optimization for continuous sign language recognition. In *Proc. 27th international joint conference on artificial intelligence* (pp. 884–891). Stockholm, Sweden.

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. In *Proc. 40th international conference on acoustics, speech and signal processing* (pp. 4580–4584). Brisbane, QLD, Australia.

Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. In *Proc. INTERSPEECH* (pp. 1089–1093). Stockholm, Sweden.

Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *Proc. INTERSPEECH* (pp. 312–315). Brighton, UK.

Scialom, T., Piwowarski, B., & Staiano, J. (2019). Self-attention architectures for answer-agnostic neural question generation. In *Proc. 57th annual meeting of the association for computational linguistics* (pp. 6027–6032). Florence, Italy.

Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., & Zhang, C. (2018). Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proc. 32nd AAAI conference on artificial intelligence* (pp. 5446–5455). New Orleans, Louisiana, USA.

Shi, Y., Hwang, M.-Y., & Lei, X. (2019). End-to-end speech recognition using a high rank LSTM-CTC based model. In *Proc. 44th IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7080–7084). Brighton, UK.

Steidl, S. (2009). *Automatic classification of emotion related user states in spontaneous children's speech*. Logos Verlag, Berlin: University of Erlangen-Nuremberg Erlangen, Germany.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *Proc. IEEE conference on computer vision and pattern recognition* (pp. 1–9). Boston, MA, USA.

Tarantino, L., Garner, P. N., & Lazaridis, A. (2019). Self-attention for speech emotion recognition. In *Proc. INTERSPEECH* (pp. 2578–2582). Graz, Austria.

Tzinis, E., & Potamianos, A. (2017). Segment-based speech emotion recognition using recurrent neural networks. In *Proc. 7th international conference on affective computing & intelligent interaction* (pp. 190–195). San Antonio, Texas, USA.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing, Special Issue on End-to-End Speech and Language Processing*, *11*(8), 1301–1309.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Proc. 31st annual conference on neural information processing systems* (pp. 5998–6008). Long Beach, CA, USA.

Wang, H., Cui, Z., Chen, Y., Avidan, M., Abdallah, A. B., & Kronzer, A. (2018). Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *15*(6), 1968–1978.

Wang, Y., Deng, X., Pu, S., & Huang, Z. (2017). Residual convolutional CTC networks for automatic speech recognition. arXiv preprint arXiv:1702.07793.

Wang, T., Sun, M., & Hu, K. (2017). Dilated deep residual network for image denoising. In *Proc. 29th IEEE international conference on tools with artificial intelligence* (pp. 1272–1279). Boston, MA, USA.

Wang, H., Wu, J., Pan, S., Zhang, P., & Chen, L. (2017). Towards large-scale social networks with online diffusion provenance detection. *Computer Networks*, *114*, 154–166.

Wang, H., Wu, J., Zhang, P., & Chen, Y. (2018). Learning shapelet patterns from network-based time series data. *IEEE Transactions on Industrial Informatics*, *15*(7), 1–14.

Wang, H., Zhang, Q., Wu, J., Pan, S., & Chen, Y. (2019). Time series feature learning with labeled and unlabeled data. *Pattern Recognition*, *89*, 55–66.

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., et al. (2008). Abandoning emotion classes – Towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. INTERSPEECH* (pp. 597–600). Brisbane, Australia.

Yu, F., Koltun, V., & Funkhouser, T. (2017). Dilated residual networks. In *Proc. IEEE conference on computer vision and pattern recognition* (pp. 472–480). Honolulu, Hawaii.

Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Laurent, C., Bengio, Y., et al. (2016). Towards end-to-end speech recognition with deep convolutional neural networks. In *Proc. INTERSPEECH* (pp. 410–414). San Francisco, USA.

Zhao, Z., Bao, Z., Zhang, Z., Cummins, N., Wang, H., & Schuller, B. (2019). Attention-enhanced connectionist temporal classification for discrete speech emotion recognition. In *Proc. INTERSPEECH* (pp. 206–210). Graz, Austria.

Zhao, Z., Bao, Z., Zhao, Y., Zhang, Z., Cummins, N., Ren, Z., et al. (2019). Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. *IEEE Access*, *7*, 97515–97525.

Zhao, Z., Zheng, Y., Zhang, Z., Wang, H., Zhao, Y., & Li, C. (2018). Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition. In *Proc. INTERSPEECH* (pp. 272–276. Hyderabad, India.