·Article·

Self-attention transfer networks for speech emotion recognition

Ziping ZHAO[1], Keru Wang[1], Zhongtian BAO[1], Zixing ZHANG[2], Nicholas CUMMINS[3,4], Shihuang SUN[5], Haishuai WANG[5], Jianhua TAO[6*], Björn W. SCHULLER[1,2,3]

1. *College of Computer and Information Engineering, Tianjin Normal University, Tianjin* 300387*, China*
2. *GLAM-Group on Language, Audio & Music, Imperial College London, SW7 2AZ, UK*
3. *Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg,* 86159*, Germany*
4. *Department of Biostatistics and Health Informatics, IoPPN, King's College London, London, SE5 8AF, UK*
5. *Department of Computer Science and Engineering, Fairfield University* 06824*, USA*
6. *National Laboratory of Pattern Recognition, CASIA, Beijing* 100190*, China*

*** Corresponding author,** jhtao@nlpr.ia.ac.cn

Received: 17 September 2020 **Accepted:** 29 December 2020

Supported by the National Natural Science Foundation of China (62071330); the National Science Fund for Distinguished Young Scholars (61425017); the Key Program of the National Natural Science Foundation (61831022); the Key Program of the Natural Science Foundation of Tianjin (18JCZDJC36300); the Open Projects Program of the National Laboratory of Pattern Recognition and the Senior Visiting Scholar Program of Tianjin Normal University; the Innovative Medicines Initiative 2 Joint Undertaking (115902), which receives support from the European Union′s Horizon 2020 research and innovation program and EFPIA.

Citation: Ziping ZHAO, Keru Wang, Zhongtian BAO, Zixing ZHANG, Nicholas CUMMINS, Shihuang SUN, Haishuai WANG, Jianhua TAO, Björn W. SCHULLER. Self-attention transfer networks for speech emotion recognition. Virtual Reality & Intelligent Hardware, 2021, 3(1): 43—54
DOI: 10.1016/j.vrih.2020.12.002

Abstract **Background** A crucial element of human-machine interaction, the automatic detection of emotional states from human speech has long been regarded as a challenging task for machine learning models. One vital challenge in speech emotion recognition (SER) is learning robust and discriminative representations from speech. Although machine learning methods have been widely applied in SER research, the inadequate amount of available annotated data has become a bottleneck impeding the extended application of such techniques (e.g., deep neural networks). To address this issue, we present a deep learning method that combines knowledge transfer and self-attention for SER tasks. Herein, we apply the log-Mel spectrogram with deltas and delta-deltas as inputs. Moreover, given that emotions are time-dependent, we apply temporal convolutional neural networks to model the variations in emotions. We further introduce an attention transfer mechanism, which is based on a self-attention algorithm to learn long-term dependencies. The self-attention transfer network (SATN) in our proposed approach takes advantage of attention transfer to learn attention from speech recognition, followed by transferring this knowledge into SER. An evaluation built on Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset demonstrates the effectiveness of the proposed model.

Keywords Speech emotion recognition; Attention transfer; Self-attention; Temporal convolutional neural networks (TCNs)

1 Introduction

The automatic identification of the emotional states of human speech has long been considered a challenging task for machine learning algorithms. The question of how robust, discriminative, and emotionally salient features can be extracted from the acoustic content of a speech signal remains crucial for researchers in the field of speech emotion recognition.

This challenge has been addressed by a variety of machine-learning approaches. Recently, recurrent neural networks (RNNs) have demonstrated a superior performance in terms of speech emotion recognition (SER)[1–3]. RNNs outperform other models because of their ability to learn sequential speech data and have also performed extremely well on many relevant speech-based problems[4,5].

However, these frameworks are also affected by certain limitations. First, RNNs are extremely difficult to train[6]. Moreover, RNNs can only capture limited temporal information from speech data, making this approach inappropriate for dealing with long-term data[7]. Over the years, machine learning researchers have strived to facilitate the training of RNNs through innovative training architectures and strategies[8]. To address these challenges, temporal convolutional network (TCN) based approaches have recently been proposed. This research avenue is superior to recurrent networks because the training complexity is much lower. In addition, long-term dependencies can be better captured by such methods[9]. However, there are also limitations to TCNs, specifically related to coping with dependencies between long-range patterns in speech[7].

The self-attention mechanism[10], which can help capture long-term contextual dependencies, has demonstrated its ability to capture contextual dependencies in several natural language processing (NLP) tasks[10–13] More recently, they have achieved a state-of-the-art SER performance[14].

Nevertheless, the number of trainable parameters in the models has increased because of the attention mechanisms included in the models, which is suboptimal for smaller (regarding the number of unique examples) emotional corpora[15]. However, it is possible to combine the attention mechanism with transfer learning frameworks[16–19].

In light of the above findings, we propose a novel model that leverages a well-designed self-attention transfer mechanism to facilitate the training of an attention-based TCN paradigm for SER across tasks. The recent success of attention transfer mechanisms has encouraged us to introduce an attention transfer network (ATN)[4,16] for such tasks. In this study, we further explore the contributions of both the TCN-based self-attention model and the teacher-student framework to attention transfer for SER purposes. Moreover, inspired by the positive results achieved by 3D log Mel spectrum features in the SER context[20,21], we also employed log-Mel, deltas, and delta-deltas as 3D inputs to the TCN model in this study. We further compare the effectiveness of both the teacher-student framework and the self-attention mechanism for an attention transfer in this problem.

Our proposed model has a number of advantages: (1) it provides a self-attention transfer mechanism on the basis of a TCN, which is capable of capturing long-term temporal patterns and their dependencies; (2) it automatically transfers self-attention from speech recognition and simultaneously interprets what information should be transferred simultaneously. Given that a teacher network is trained on a similar task, we aim to improve the learning ability of the student network. To the best of our knowledge, this is the first time that such a study was conducted for SER tasks.

2 Related studies

As discussed in the introduction, several existing studies have applied deep learning (e.g., RNNs and

CNNs) to SER tasks[22–24]. The advantages of RNNs have been reported in the context of SER[1–3,25,26].

Recently, owing to their advantageous parallelism, flexible receptive field, and stable gradient, TCNs[27] have proven to be effective at capturing long-range patterns in a wide variety of tasks[28,29].

In addition, attention mechanisms have been widely accepted within the deep learning community. The ultimate objective of applying attention is to enhance the accuracy of the decision-making. Attention mechanisms have been successfully used in speech recognition[30], NLP[31,32], and speech emotion recognition[3,26,33] tasks. The self-attention mechanism has also been found to produce promising results in various tasks including NLP[10–13] and speech-related tasks[34,35].

Unlike traditional machine learning methods, deep learning relies heavily on vast amounts of training data[36]. However, the lack of training data has become an inevitable issue in SER[37]. Accordingly, knowledge transfer, which is a technique leveraged to promote a network performance in cases in which only a small amount of labeled training data are available, has been broadly applied in different settings[38,39]. Most recently, a knowledge transfer mechanism called an attention map[16] has been applied to show that learning smaller "student" networks to simulate larger attention maps and advanced "teacher" network architectures can yield substantial performance improvements in these smaller networks. The authors used attention for cross-domain knowledge transfer from online images to videos[17]. Similarly, Zhuo et al. presented an attention transfer method for traditional domain adaptation[18]. However, both of these studies were based on a CNN, and there are no existing studies on the attention transfer framework for TCNs.

Based on a literature review, we found that existing studies provide compelling evidence suggesting that the addition of self-attention and attention transfer modules is effective. Therefore, the proposed method combines these two modules for SER. To date, no existing studies have integrated these two models for this type of problem.

3 Proposed method

We first provide an introduction of our method on 'cross-task' speech emotion recognition, and then describe the proposed method in detail.

3.1 Framework of the designed method

The proposed method comprises two key tasks: (1) speech recognition, and (2) speech emotion recognition (Figure 1). The aim of the proposed method is to improve the performance in the target domain (i. e., speech emotion recognition) using spatial attention maps in the source task (i.e., speech recognition). Because the amount of target data is insufficient, mapping is learned on data-rich tasks, as high-quality attention is acquired during the training process. Given that the number of data suitable for the training of automatic speech recognition (ASR) systems is significantly larger than the number of SER data,

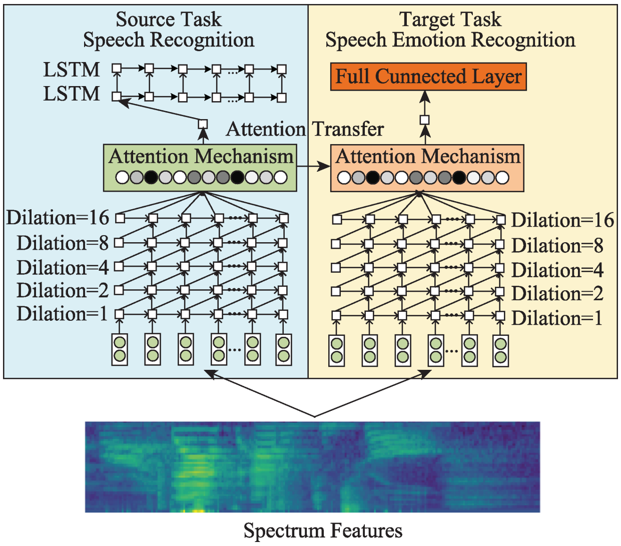

Figure 1 Framework of the designed model. The model first trains a self-attention-based encoder–decoder model. After learning the parameters of hidden layers in the speech recognition network, the model freezes its parameters. In the next step, the model trains the attention weights through a speech recognition task, and then feeds these weights into the speech emotion recognition system.

we can expect trained ASR systems to be more robust to speaker and condition variations. Therefore, as a possible solution, transfer learning using ASR as the source task might be an efficient approach in emotion recognition to leverage knowledge acquired from speech recognition tasks.

From this perspective, three key modules are included in the proposed framework: (1) we train an attention-based encoder-decoder network (or a teacher network) for speech recognition to learn the initial attention maps, (2) the model trains a (shallower) student network for the SER task by applying a self-attention transfer mechanism to simulate the attention maps of the teacher network[16], and (3) the final component of our hybrid model is the speech emotion recognition module, in which TCN modules are combined with the self-attention mechanism.

In the proposed model, TCNs were employed to capture high-level feature representations. Further exploration of the advantages of the self-attention mechanism is described in the latter two sub-sections.

3.2 Standard soft attention

Standard soft attention mechanisms are used to select the relevant encoded hidden vectors via attention weights (an informative sequence of weights) during the decoding phase[32]. At each timestep i, the attention weights $\alpha_{i,j}$are produced by normalizing the scalar values $e_{i,j}$ across the memory using the softmax function:

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{T} \exp(e_{i,k})}, \tag{1}$$

where $e_{i,j}$ is an alignment scoring mechanism used to determine how well the inputs around position j and the output at position i match. This is computed as follows:

$$e_{i,j} = \alpha(s_{i-1}, h_j), \tag{2}$$

where $s_i$denotes the state of the decoder, h_j indicates the j-th entry of the hidden state sequence h = $\{h_{1,\ldots,}h_T\}$, and $\alpha(\cdot)$ is a learnable deterministic "energy function". Typically, a single-layer neural network using tanh nonlinearity is utilized as $\alpha(\cdot)$; however, other functions (such as a simple dot product between $s_{i-1}$and h_j) have also been used[32]. It should be noted that we used tanh as the nonlinear activation function in the present study.

The output of the attention layer, denoted as c_i, is the weighted average of the encoder hidden state sequence h, which is defined as follows:

$$c_i = \sum_{j=1}^{T} \alpha_{i,j}\, h_j, \tag{3}$$

Finally, the decoder state s_i is updated based on s_i, c_i, and the decoder outputs y_i as follows:

$$s_i = f\left(s_{i-1}, y_{i-1}, c_i\right), \tag{4}$$

$$y_i = g(s_i, c_i), \tag{5}$$

where $f(\cdot)$is a TCN, and $g(\cdot)$is a learnable nonlinear function that maps the decoder state to the output space.

In this study, unless otherwise stated, "soft attention" refers to the global attention approach. In global attention, all hidden states of the encoder are considered to enable the derivation of the context vector c_i.

3.3 Temporal convolutional network (TCN)

Similar to WaveNet[9], TCNs were utilized to learn the temporal dynamic representation in this study[40].

Causal Convolutions. Given sequence data x with length T, we denote $x = x_1,\ldots,x_T$ (where x_t represents the data at timestep t) and $y = y_1,\ldots,y_T$ (the prediction at each timestep). According to the causal constraint, the prediction of y_t depends only on past observations $x_{<t}$ and not on future observations. For instance, a bidirectional RNN does not fall under a causal constraint[40].

Dilated Convolutions. Because standard convolutions have a certain filter size, their temporal understanding is fixed. Following[27], dilated convolutions were utilized to enable an exponentially large receptive field. More specifically, given an input $x = x_1,\ldots,x_T \in \mathbb{R}^T$ with length T and a filter $f{:}\{0,\ldots,k - 1\} \rightarrow \mathbb{R}$, the dilated convolution operation F on element s of the sequence can be defined as follows:

$$F(s) = (x *_d f)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{<(T-di)}, \tag{6}$$

where d denotes the dilation factor, k denotes the filter size, and * denotes the convolution operation. When $d = 1$, a dilated convolution is reduced to a standard convolution. Figure 2 presents an illustration of a dilated causal convolution.

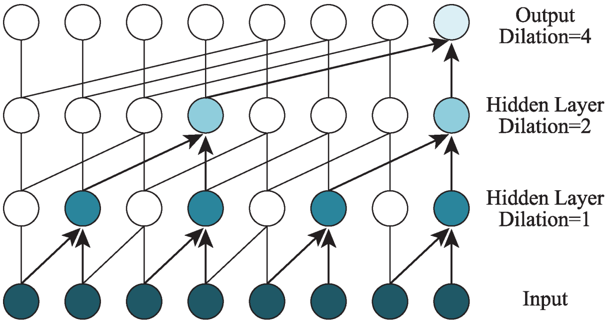

Figure 2 Illustration of a dilated causal convolution.

3.4 Self-attention

Self-attention is defined as an attention technique based on an encoder-decoder structure that does not employ any form of recurrence; instead, it uses weighted correlations between the elements of the input sequence[10]. Under this paradigm, the encoder maps an input sequence into several attention matrices, whereas the decoder uses these matrices to generate a new output token. A *transformer*, a model that utilizes *self-attention*, has been demonstrated to achieve state-of-the-art performance in several NLP tasks, with a computing cost one or two orders of magnitude (depending on the size of the model) lower than that of a conventional RNN[11−13]. It should be noted that this section focuses solely on the implementation of the encoder because our proposed hybrid network does not require a decoder.

Self-attention is used to calculate the queries, keys (properties of the input), and values (the output) for the frames in a given hidden sequence H through a linear transformation of input sequence X, as follows:

$$Q = W_q X\,;\ K = W_k X\,;\ V = W_v X, \tag{7}$$

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

$$Z = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{8}$$

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

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

3.5 Attention transfer network

Most recent studies on attention transfer have focused on computer-vision-related tasks and spatial attention maps designed for CNNs[16,18]. In these methods, the activation maps in a specific convolutional layer for both domains are first computed via *Lp*-norm pooling, after which the domain discrepancy is minimized in the second-order correlation statistics of the attention maps[18]. Based on this concept, we designed a self-attention transfer process for TCNs.

Regarding the activation-based attention model, this section describes the approach applied to define the spatial attention map, along with how attention information is transferred between a teacher network and a student network.

First, given a convolutional layer with its activation tensor $A \in \mathbb{R}^{C \times H \times W}$, where C refers to channels with spatial dimensions $H \times W$, a spatial attention map is denoted by a mapping function F (where A is the input and a spatial attention map as the output of F) ,which is expressed as follows:

$$F\colon \mathbb{R}^{C \times H \times W} \rightarrow \mathbb{R}^{H \times W}. \tag{9}$$

Because the absolute value of a hidden unit represents its importance with respect to the corresponding input, it is possible to design a spatial attention map through the statistical calculation of all absolute values from the channel dimension. Accordingly, we have the following spatial attention mappings:

$$(F(A))_{i,j} = \sum_{k=1}^{C} \left| A_{k,i,j} \right|^{p}, i \in \{1, 2, \cdots, H\} \text{ and} j \in \{1, 2, \cdots, W\}, \tag{10}$$

where i and j are spatial indexes.

Given the spatial attention maps in a teacher network, an attention transfer aims to learn a student network that makes correct predictions and learns the attention maps similar to those in the teacher network.

To maintain the generality, an assumption is made that the transfer losses between the teacher and student attention maps with the same spatial resolution are considered. Attention maps can be interpolated to match the shapes. Therefore, the loss can be defined as follows:

$$L_T = L_{SER} + W_{AT} \times L_{AT}, , \tag{11}$$

where L_{SER} represents the loss in the speech emotion recognition task, W_{AT} represents the proportion of attention transfer, and L_{AT} represents the loss in attention transfer. L_{AT} can be calculated as follows:.

$$L_{AT} = \sum_{j \in I} \left\| Q_{SER}^{j} - Q_{S}^{j} \right\|_{1}, \tag{12}$$

where I represents the attention map indices, Q_{SER}^{j} denotes the j-th attention map pair in speech emotion recognition, and Q_{S}^{j} denotes speech recognition tasks. l_1 normalized attention maps are utilized in the attention transfer process.

4 Experiments and results

4.1 Dataset

We performed experiments using the IEMOCAP dataset, which is a well-benchmarked dataset among the speech emotion recognition research community. The IEMOCAP dataset contains transcriptions of dialogs and audio-visual recordings of two actors[41]. Therefore, we were able to perform speech emotion recognition and speech recognition tasks using this dataset. IEMOCAP is split into *script* and *improvise* sections. To avoid any confusion caused by semantic information, only the *improvise* part was used in our experiments. The final number of instances for each emotion category is listed in Table 1.

4.2 Features

In this study, we used a 3D log-mels spectrogram as the input. Our spectrograms were created using the extraction process outlined in [21]. In brief, the output from a 40-dimensional mel-scale log filter bank was used to construct each spectrogram. We computed the features over a frame length and a stride of 25ms

Table 1 Number of samples in four emotion classes (i.e., happy, neutral, angry, and sad) in the IEMOCAP data

Session	*N.*	*H.*	*S.*	*A.*	*Total*
1	223	132	104	62	521
2	217	191	100	22	530
3	198	149	190	90	627
4	174	195	81	84	534
5	287	280	133	31	731
Total	1099	947	608	289	2943

and 10ms, respectively. In the final step, we calculated the delta and delta-deltas of the spectrogram, which reflect the process of emotional change. After the completion of the entire processing, the original 40-dimensional features changes to 120-dimensional features that are capable of providing the classifier with more information.

4.3 Experimental settings and performance measurements

Parameters in the model. All our models were implemented using the PyTorch[1] framework. Because there are restrictions in place concerning time and computational efficiency, we set the number of training iterations to 100.

In our experiment, the models in each experiment were trained for 100 epochs. Once the training of each epoch was complete, the parameters of the model were saved, and the developed set of IEMOCAP was utilized to evaluate the TCN performance. After tuning the model on the development set, we loaded the hyperparameters that obtained the best performance on the development set to perform the final emotion prediction on the test set.

On the IEMOCAP dataset, we performed a five-fold cross-validation using a leave-one-session out strategy, in line with the methodology utilized in previous studies[33,42]. Each training process involved the use of eight speakers from four sessions as training data; the remaining session was divided into two parts, one of which was used as validation data and the other as test data.

To conduct attention transfer, a source task (speech recognition) was trained, and attention maps acquired from a speech recognition task were collected. The IEMOCAP dataset and a self-attention-based encoder-decoder model were used to train the speech recognition model. We used the CMU pronouncing dictionary[43] to decrease the number of states of the original transcription in the speech recognition task. Furthermore, we used a TCN with 5-layer 1D convolutional modules, and dilation factors 1, 2, 4, 8, and 16 were used for the encoder in the speech recognition pre-training task. We used a long short-term memory (LSTM) containing 256 single-memory-cell blocks for the decoder. An Adam optimizer with a fixed learning rate of 10^{-4} was applied to train the encoder-decoder model. After the parameters in the TCN network were learned, we froze the parameters of the network. The model trained the speech emotion recognition model after completing the above training steps. We also applied TCN for training, and the learning rate was 10^{-4}.

For comparison, we used the BLSTM network to replace TCN. In the comparison experiment, we fed the 3D log-mels spectrogram as input into a one-layer BLSTM network, which had 128 forward hidden nodules and 128 backward hidden nodules.

Evaluation metrics. Standard evaluation criteria were used to evaluate the results generated by the two datasets. For the IEMOCAP-generated results, the evaluation metrics used were unweighted accuracy and

[1] https://pytorch.org/

weighted accuracy (UA and WA, respectively).

4.4 Results and discussion

We conducted experiments to validate the performance of the self-attention transfer network (SATN). For the purpose of comparison, Table 2 presents the results for four benchmark models that have been successfully developed on IEMOCAP, which employ global-soft-attention[32] based TCNs and a global-soft-attention transfer network. We also compared the proposed method with a self-attention-based TCN model without an attention transfer.

Table 2 Performance comparison of our proposed model and other selected models on the IEMOCAP dataset

Methods	Dev.		Test	
	WA [%]	UA [%]	WA [%]	UA [%]
Previously reported methods				
DNN+ELM[44,45]	–	–	57.9	52.1
RNN+ELM[45]	–	–	62.9	63.9
Attention+RNN[3]	–	–	63.5	58.8
GMM+HMM[46]	–	–	55.0	60.3
Proposed self-attention models (AT denotes attention transfer)				
BLSTM+soft attention	63.8	62.9	59.6	59.7
BLSTM+self-attention	63.7	64.5	60.0	60.5
BLSTM+soft attention w/AT	65.6	65.6	62.1	62.2
BLSTM+self-attention w/AT	66.9	68.1	63.8	64.5
TCN+soft attention	65.5	66.6	61.8	62.5
TCN+self-attention	67.5	67.4	63.7	64.2
TCN+soft attention w/AT	67.2	67.9	63.4	64.4
TCN+self-attention w/AT	**68.6**	**69.5**	**65.0**	**66.1**

Although RNNs, for example, those with LSTM, have proven their efficiency for speech emotion recognition, the main advantage of HMMs over these emerging LSTMs for SER is that HMM-based architectures are also effective in dynamic modeling. Therefore, the latest approach utilizing HMM-based architectures was also incorporated for comparison.

To provide supervision for attention generation, an attention-based encoder-decoder model was implemented for the speech recognition task. For the speech recognition module, we use the word error rate (WER) as the evaluation metric, which is defined as follows:

$$WER\left(label, predict\right) = 100 \times \frac{N_s + N_D + N_I}{N}\%, \tag{13}$$

where N denotes the total number of words, and N_s, N_D, and N_I represent the number of substitutions, deletions, and insertions, respectively. Subsequently, a WER of 47.7% was obtained on the test set of the IEMOCAP utilizing the TCN-based self-attention model.

From the results, we observe that the proposed method outperforms existing methods in terms of WA and UA on the IEMOCAP dataset (Table 2). It can be observed that the best WA (65.0%) and UA (66.1%) on the test set, and the best WA (68.6%) and UA (69.5%) on the development set, were achieved by our novel self-attention-based TCN model with an attention transfer mechanism (Table 2). This shows a significant improvement compared to the baseline DNN-ELM model presented in[45] ($p < 0.05$ in a one-tailed z-test).

We also use the confusion matrix (Figure 3) to analyze the per-class performance of our proposed method on the IEMOCAP dataset. Each fold experiment applied a confusion matrix, and the final confusion matrix was obtained by averaging all the matrices. We observe that *sad* and *happy* are two easily recognized classes and their classification rates are relatively higher than others, particularly for *sad* samples, which reached an accuracy of 70.13% with our proposed method. Intuitively, *sad* samples have obvious voice characteristics, such as a low pitch and slow speed. However, the neutral class suffered more misclassifications during the experiment. Some neutral samples were misclassified as happy. We believe that this is because many people do not react much through their voice when they are happy.

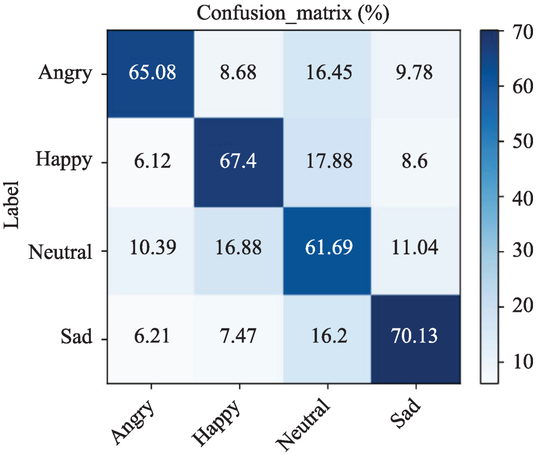

Figure 3 Confusion matrix of our proposed method on the IEMOCAP dataset.

Meanwhile, our novel TCN-based self-attention model outperformed the attention-based BLSTM, regardless of whether any attention strategy or attention transfer mechanism was employed on both the IEMOCAP test and development sets. Moreover, although the performance of the BLSTM-based self-attention transfer model is inferior to that of the TCN-based self-attention transfer model, it is superior to the three baseline models proposed in [3,44−46], considering the UA and WA on the test set of the IEMOCAP dataset.

In terms of the attention transfer mechanism, it was found that the attention model with no attention transfer is less effective than the attention model with an attention transfer mechanism (either with BLSTM or TCNs). Therefore, combining an attention transfer mechanism with an attention-based TCN model can help in handling the SER task. This supports our assumption that learning to stimulate the attention maps of the teacher network is a useful approach.

In addition, the effectiveness of the global soft attention model[32] is less than that of the self-attention model with or without an attention transfer, although it is superior to that of the baseline model proposed in [44] in terms of both UA and WA for both the development and test sets of the IEMOCAP dataset.

5 Conclusion

Our proposed attention-based model, called SATN, combines self-attention with a knowledge transfer for SER tasks. The two major contributions of this model are as follows. First, we build a self-attention transfer method that transfers self-attention to apply SER across tasks. On the one hand, the TCN structure enables the model to learn long-term data with complex spatio-temporal patterns. On the other hand, the temporal attention block captures the dependencies between these patterns. Second, the experimental results on the IEMOCAP dataset reveal the effectiveness of our proposed TCN-based system combination.

In future studies, we will investigate the use of hierarchical self-attention for learning representations and other SER-related tasks, such as mood detection.

Declaration of competing interest

We declare that we have no conflict of interest.

References

1 Wllmer M, Eyben F, Reiter S, Schuller B, Cowie R. Abandoning emotion classes – towards continuous emotion recognition with modelling of long-range dependencies. In: Proceedings INTERSPEECH 2008, 9th Annual Conference

of the International Speech Communication Association, incorporating 12th Australasian International Conference on Speech Science and Technology. 2008, 597‒600

2 Tzirakis P, Trigeorgis G, Nicolaou M A, Schuller B W, Zafeiriou S. End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1301‒1309
DOI:10.1109/jstsp.2017.2764438

3 Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA, IEEE, 2017, 2227‒2231
DOI:10.1109/icassp.2017.7952552

4 Zhao Z P, Bao Z T, Zhang Z X, Deng J, Cummins N, Wang H S, Tao J H, Schuller B. Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(2): 423‒434
DOI:10.1109/jstsp.2019.2955012

5 Zhao Z P, Bao Z T, Zhang Z X, Cummins N, Wang H S, Schuller B. Hierarchical attention transfer networks for depression assessment from speech. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain, IEEE, 2020, 7159‒7163
DOI:10.1109/icassp40776.2020.9053207

6 Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. In: Proceedings of the 30th International Conference on International Conference on Machine Learning(ICML). Atlanta, GA, USA, 2013, 1310‒1318

7 Dai R, Minciullo L, Garattoni L, Francesca G, Bremond F. Self-attention temporal convolutional network for long-term daily living activity detection. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Taipei, Taiwan, China, IEEE, 2019, 1‒7
DOI:10.1109/avss.2019.8909841

8 Bengio S, Vinyals O, Jaitly N, Shazeer N. Scheduled sampling for sequence prediction with recurrent Neural networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1. Montreal, Canada, MITPress, 2015, 1171‒1179

9 van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K. WaveNet: a generative model for raw audio. 2016

10 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA, Curran Associates Inc. 2017, 6000‒6010

11 Scialom T, Piwowarski B, Staiano J. Self-attention architectures for answer-agnostic neural question generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, Stroudsburg, PA, USA, Association for Computational Linguistics, 2019, 6027‒6032
DOI:10.18653/v1/p19-1604

12 Shen T, Zhou T, Long G, Jiang J, Pan S, Zhang C. DiSAN: Directional Self-Attention Network for RNN/CNN-Free Language Understanding. 2017

13 Li X P, Song J K, Gao L L, Liu X L, Huang W B, He X N, Gan C. Beyond RNNs: positional self-attention with Co-attention for video question answering. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 8658‒8665
DOI:10.1609/aaai.v33i01.33018658

14 Tarantino L, Garner P N, Lazaridis A. Self-attention for speech emotion recognition. In: Interspeech. ISCA, 2019
DOI:10.21437/interspeech.2019-2822

15 Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri T F. A review of depression and suicide risk assessment using speech analysis. Speech Communication, 2015, 71: 10‒49
DOI:10.1016/j.specom.2015.03.004

16 Zagoruyko S, Komodakis N. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. 2016

17 Li J, Wong Y, Zhao Q, Kankanhalli M S. Attention transfer from web images for video recognition. In: Proceedings of the 25th ACM international conference on Multimedia. Mountain View, California, USA, Association for Computing Machinery, 2017, 1−9
DOI:10.1145/3123266.3123432

18 Zhuo J B, Wang S H, Zhang W G, Huang Q M. Deep unsupervised convolutional domain adaptation. In: Proceedings of the 25th ACM international conference on Multimedia. Mountain View California USA, New York, NY, USA, ACM, 2017
DOI:10.1145/3123266.3123292

19 Kim J, Park S, Kwak N. Paraphrasing complex network: network compression via factor transfer. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada, Curran Associates Inc, 2018, 2765−2774

20 Meng H, Yan T H, Yuan F, Wei H W. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. IEEE Access, 2019, 7: 125868−125881
DOI:10.1109/access.2019.2938007

21 Chen M Y, He X J, Yang J, Zhang H. 3D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters, 2018, 25(10): 1440−1444
DOI:10.1109/lsp.2018.2860246

22 Mao Q R, Dong M, Huang Z W, Zhan Y Z. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia, 2014, 16(8): 2203−2213
DOI:10.1109/tmm.2014.2360798

23 Huang Z W, Dong M, Mao Q R, Zhan Y Z. Speech emotion recognition using CNN. In: Proceedings of the 22nd ACM international conference on Multimedia. New York, NY, USA, ACM, 2014
DOI:10.1145/2647868.2654984

24 Zhang Y Y, Du J, Wang Z R, Zhang J S, Tu Y H. Attention based fully convolutional network for speech emotion recognition. In: 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Honolulu, HI, USA, IEEE, 2018
DOI:10.23919/apsipa.2018.8659587

25 Tzinis E, Potamianos A. Segment-based speech emotion recognition using recurrent neural networks. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). San Antonio, TX, USA, IEEE, 2017, 190−195
DOI:10.1109/acii.2017.8273599

26 Huang C W, Narayanan S S. Attention assisted discovery of sub-utterance structure in speech emotion recognition. In: Interspeech 2016. ISCA, 2016
DOI:10.21437/interspeech.2016-448

27 Bai S J, Kolter J Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. 2018

28 Lea C, Flynn M D, Vidal R, Reiter A, Hager G D. Temporal convolutional networks for action segmentation and detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, IEEE, 2017, 1003−1012
DOI:10.1109/cvpr.2017.113

29 Du Z Y, Wu S W, Huang D, Li W X, Wang Y H. Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. IEEE Transactions on Affective Computing, 2019
DOI:10.1109/taffc.2019.2940224

30 Chorowski J, Bahdanau D, Serdyuk D, Cho K, Bengio Y. Attention-based models for speech recognition. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada. MIT Press, 2015, 577−585

31 Vinyals O, Kaiser L, Koo T, Petrov S, Sutskever I, Hinton G. Grammar as a foreign language. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada, MIT Press, 2015, 2773−2781

32 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings 3rd International Conference on Learning Representations (ICLR). San Diego, CA, USA, 2014

33 Zhao Z P, Zheng Y, Zhang Z X, Wang H S, Zhao Y Q, Li C. Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition. In: Interspeech 2018. ISCA, 2018
DOI:10.21437/interspeech.2018-1477

34 Li Y C, Zhao T, Kawahara T. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In: Interspeech 2019. ISCA, 2019
DOI:10.21437/interspeech.2019-2594

35 Salazar J, Kirchhoff K, Huang Z H. Self-attention networks for connectionist temporal classification in speech recognition. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom, IEEE, 2019, 7115‒7119
DOI:10.1109/icassp.2019.8682539

36 Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A Survey on Deep Transfer Learning. In: Artificial Neural Networks and Machine Learning‒ICANN. Cham. Springer International Publishing, 2018, 270‒279

37 Deng J, Xu X Z, Zhang Z X, Frühholz S, Schuller B. Semisupervised autoencoders for speech emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(1): 31‒43
DOI:10.1109/taslp.2017.2759338

38 Yim J, Joo D, Bae J, Kim J. A gift from knowledge distillation: fast optimization, network minimization and transfer learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, IEEE, 2017, 7130‒7138
DOI:10.1109/cvpr.2017.754

39 Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. Fitnets: Hints for thin deep nets. In: Proceedings 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2015

40 Haque A, Guo M, Verma P, Li F F. Audio-linguistic embeddings for spoken sentences. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom, IEEE, 2019, 7355‒7359
DOI:10.1109/icassp.2019.8682553

41 Busso C, Bulut M, Lee C C, Kazemzadeh A, Mower E, Kim S, Chang J N, Lee S, Narayanan S S. IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 42(4): 335‒359
DOI:10.1007/s10579-008-9076-6

42 Zhao Z P, Bao Z T, Zhang Z X, Cummins N, Wang H S, Schuller B W. Attention-enhanced connectionist temporal classification for discrete speech emotion recognition. In: Interspeech 2019. ISCA, 2019
DOI:10.21437/interspeech.2019-1649

43 Lenzo K. The CMU pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 2007

44 Han K, Yu D, Tashev I. Speech emotion recognition using deep neural network and extreme learning machine. In: Interspeech. 2014

45 Lee J, Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition. In: Proc. INTERSPEECH. Dresden, Germany, 2015, 1537‒1540

46 Mao S Y, Tao D H, Zhang G Y, Ching P C, Lee T. Revisiting hidden Markov models for speech emotion recognition. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom, IEEE, 2019, 6715‒6719
DOI:10.1109/icassp.2019.8683172