INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Attention-enhanced Connectionist Temporal Classification for Discrete Speech Emotion Recognition

Ziping Zhao[1,2], *Zhongtian Bao*[1], *Zixing Zhang*[3], *Nicholas Cummins*[2], *Haishuai Wang*[1], *Björn Schuller*[2,3]

[1]College of Computer and Information Engineering, Tianjin Normal University, China
[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]GLAM – Group on Language, Audio & Music, Imperial College London, UK

zhaoziping@tjnu.edu.cn

Abstract

Discrete *speech emotion recognition* (SER), the assignment of a single emotion label to an entire speech utterance, is typically performed as a sequence-to-label task. This approach, however, is limited, in that it can result in models that do not capture temporal changes in the speech signal, including those indicative of a particular emotion. One potential solution to overcome this limitation is to model SER as a sequence-to-sequence task instead. In this regard, we have developed an attention-based *bidirectional long short-term memory* (BLSTM) neural network in combination with a *connectionist temporal classification* (CTC) objective function (Attention-BLSTM-CTC) for SER. We also assessed the benefits of incorporating two contemporary attention mechanisms, namely component attention and quantum attention, into the CTC framework. To the best of the authors' knowledge, this is the first time that such a hybrid architecture has been employed for SER. We demonstrated the effectiveness of our approach on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) and FAU-Aibo Emotion corpora. The experimental results demonstrate that our proposed model outperforms current state-of-the-art approaches.

Index Terms: speech emotion recognition, connectionist temporal classification, attention mechanism, bidirectional LSTM

1. Introduction

Automatic *speech emotion recognition* (SER), which focuses on the identification of discrete emotion states, can be a challenging task, and relies heavily on the effectiveness of the speech features for classification. Many works in this field treat the task as a typical sequence classification problem, in which each chunk of speech such as an utterance has exactly one label. To predict the emotional label of the chunk such as an utterance, models, such as *long short-term memory recurrent neural networks* (LSTM-RNNs) [1, 2, 3, 4, 5] are built following a sequence-to-label recipe, in which the input is a sequence, and the output is a single emotional label. However, such a conventional sequence-to-label modelling approach for discrete SER modelling is less than ideal. A critical underlying issue is a loss of dynamic temporal information that can strongly reflect a change in emotional state [6].

To address this issue, we herein propose an approach to model the discrete SER problem temporally, utilising sequence-to-sequence learning methods such as *connectionist temporal classification* (CTC) [7]. In this approach, we treated the input and output of the model as sequences. Initial research has highlighted the effectiveness of such an approach [6], and CTC-based models have shown strong performance in tasks such as end-to-end speech recognition systems and social signal detection [8, 9, 10, 11, 12]. To date, however, work exploiting CTC models for discrete SER has been very limited [4, 6, 13, 14].

However, when using CTC, there are two main limitations: the *hard alignment problem* and the *conditional independence constraint* [15]. Conditional independence is of particular concern for discrete SER; it infers that the output predictions are independent given the entire input sequence, which is not the case in SER. We address these issues through the inclusion of *attention* mechanisms directly within the CTC framework. Attention mechanisms enable a model to focus on a subset of its input sequence, and they are widely used in a range of sequence-to-sequence learning tasks, e. g., in speech recognition [16] and *natural language processing* (NLP) [17, 18]. Moreover, the application of attention can improve the performance of SER models [2, 3, 19].

However, there are two key limitations in applying attention in a CTC network: the possible assignment of the same weight to every feature within a given frame [15] and an increase in the number of learnable parameters associated with the inclusion of attention [20]. *Component attention* has recently been proposed as a method to overcome effects relating to spatial uniformity in the learnt weights [15]. The main advantage of this technique is that it enables the assignment of multiple temporal attention weights, one for each spatial component [15]. Concerning an increase in learnable parameters, *quantum attention* [20], which is based on the quantum theory of weak measurement [21], has been demonstrated in NLP to reduce the number of learnable parameters. At the same time, it maintained comparable results to an equivalent system enhanced with a standard attention [20].

Motivated by the above analysis, we have investigated a novel sequence-to-sequence modelling solution, based on attention-BLSTM-CTC, for the task of discrete SER. We combined BLSTM and CTC to align emotional labels to emotionally relevant frames automatically. This set-up should allow the model to cope robustly with long utterances containing both emotional and neutral components. Additionally, we extended the CTC model in three separate versions incorporating local, component, and quantum attention.

Our two main contributions can, therefore, be summarised as follows: (1) We have developed an attention-based BLSTM neural network combined with a unique probabilistic-nature CTC loss function. Our results demonstrate the effectiveness of this sequence-to-sequence modelling solution for discrete SER. (2) We also investigated the use of two contemporary attention mechanisms. These two attention mechanisms allow CTC to be trained using soft, instead of hard, alignments. The presented results indicate the suitability of this approach for discrete SER.

http://dx.doi.org/10.21437/Interspeech.2019-1649

2. Proposed Methodology

In this section, we outline the main steps required to implement model attention directly within the CTC framework. First, we describe the BLSTM and CTC approach used in our proposed model. We then introduce the two attention mechanisms used to form the novel CTC-attention hybrid architecture, which models the discrete SER task in a thorough sequence-to-sequence manner.

2.1. System overview

The architecture of the proposed attention-BLSTM-CTC model consists of four main components (Fig. 1): (i) an *input layer*, where we employ spectrograms as the input of the model, (ii) a *BLSTM layer*, to derive high-level representation from step (i), (iii) an *attention layer*, utilising one of three different attention mechanisms, and (iv) a *CTC layer*, in which the CTC model is used to align emotional labels to emotionally salient frames automatically.

2.2. Bidirectional long short-term memory recurrent neural networks

LSTM-based networks are widely used in the SER literature [22, 23, 1, 2, 3, 4, 5], as they model long-range dynamic dependencies in the data while avoiding the problem of vanishing or exploding gradients during training [24]. As the standard LSTM processes the input only in one direction, the bidirectional long short-term memory recurrent neural network (BLSTM) was proposed to overcome this limitation [25]. In a BLSTM, the input sequence is processed both in the standard order and in a reversed order [26].

2.3. CTC approach

The CTC model uses a loss function for sequence labelling that can account for input and the target label sequence of different lengths, without the need for any pre-segmentation. The key idea of 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).

Given an input sequence $X = (x_1, ..., x_T)$, 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 represents this distribution as a summation of all possible frame-level intermediate representations $\pi = (\pi_1, ..., \pi_T)$ (hereafter referred to as the CTC path):

$$P(l|X) = \sum_{\pi \in \Phi(l)} P(\pi|X), \tag{1}$$

where $\Phi(l)$ denotes the set of CTC paths allowing for the insertion of $Null$ and repetition of non-blank labels to l, i. e., $\Phi^{-1}(\pi) = l$, noting, if $l_u \in L = \{1, ..., K\}$, 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:

$$P(\pi|X) = \prod_{t=1}^{T} y^t_{\pi t}, \tag{2}$$

where y^t_k is the k-th output of the softmax layer at time t, interpretable as the occurrence probability of the corresponding label. The probability distribution $P(l|X)$ can be computed efficiently using the forward-backward algorithm. The detailed CTC training process is described in [6]. Note, we used the CTC model for the work presented in this paper to update the parameters of the BLSTM model. Herein, this combination is denoted as BLSTM-CTC.

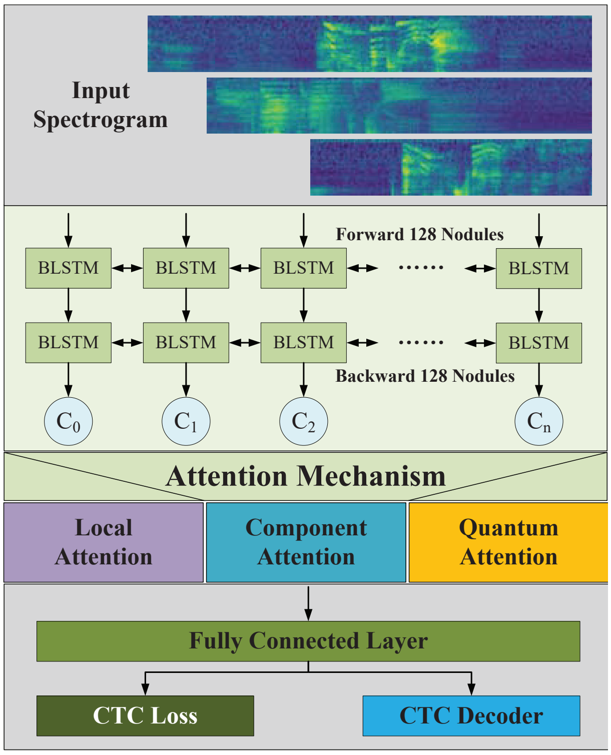

Figure 1: *Framework of our proposed model. First, spectrograms are fed into a BLSTM layer. We then apply one of three different attention mechanisms and use CTC to align labels to emotionally salient input frames.*

2.4. Attention-BLSTM-CTC model

In this section, we introduce the component attention and quantum attention mechanisms to improve the suitability of the BLSTM-CTC-based framework for SER.

2.4.1. *Component Attention*

Similar to [15], the component attention model investigated in this paper considers a small subsequence of hidden feature vectors rather than the entire sequence. Instead of applying the same attention weight to all features extracted from the frame, multiple attention weights are assigned to each frame – one for each spatial component.

In order to compute the weights of each component in the local attention mechanism, rather than the weights of each frame, the weights are calculated as follows:

$$e_{t,f_n} = \tanh(W \times C_{[t-\frac{scope-1}{2}, t+\frac{scope+1}{2}], f_n} + b), \tag{3}$$

where W denotes the learnable weight matrix; C denotes the output of BLSTM network; t, the time step of the input; f_n, the number of features; and *scope*, the length of the component attention mechanism window. After calculating the weights of each component in every frame, we then normalise these weights in feature level axes as follows:

$$\alpha_{t,f_n,i} = \frac{\exp(e_{t,f_n,i})}{\sum_{j=-\frac{scope-1}{2}}^{\frac{scope-1}{2}} \exp(e_{t+j,f_n,i})}, \tag{4}$$

where $\alpha_{t,f_n,i}$ is interpretable as the level of contribution from C, the output of the BLSTM network. Meanwhile, since features are individually treated in component attention, i is used to denote each feature.

The final output of attention mechanism U can be computed as:

$$\begin{aligned} U_t &= \text{Annotate}(\alpha, C) \\ &= \sum_{i=-\frac{scope=1}{2}}^{\frac{scope-1}{2}} \alpha_{t+i} \times C_{\left[t-\frac{scope=1}{2}, t+\frac{scope-1}{2}\right]}. \end{aligned} \tag{5}$$

We added a fully-connected layer after the component attention layer to produce the probability of each label in every time step. Finally, the CTC loss function is used to calculate the gradient, and the RMSProp Optimiser [27] is used to train the whole network.

2.4.2. Quantum Attention

Quantum Theory (QT) is widely employed to explain cognitive activities in psychology and cognition science, and it is widely used in information theory [28]. In an analogy of weak measurement in physics [21], we regard a human's recognition of emotion in speech as a system of Quantum Attention, i. e., the features of the frame are the observable variables, and the importance of the different frames to the emotion of the utterance are the measurement results.

Similar to [20], the quantum attention mechanism investigated in this work employs *weak measurement* rather than *standard quantum measurement* to model the process of SER. Moreover, the weak value under the *two-state vector formalism* (TSVF) is used to represent the degree of importance of different frames for SER. The pre-state is the forward memory cell which contains the information of all input features generated in the past; the post-state is the backward memory cell which contains the information of all input features (to be) generated in the future. Thus, the weak value in the two-state vector formalism is:

$$e_{i,t} = \frac{C_i^{fin} \times (C_i^t)^T \times C_i^t \times (C_i^{in})^T}{C_i^{fin} \times (C_i^{in})^T}, \tag{6}$$

where i denotes the instance; t, the time step of the input; fin, the final time step; and in, the first time step. Equation (6) can be regarded as the degree of importance of the different frames to the emotion at the statistical level.

As a result of the weak values produced by the above formula, a tanh layer is applied to the resulting values:

$$\alpha_t = \tanh(e_t). \tag{7}$$

Finally, every feature in each frame is multiplied by w_i thereby applying the quantum attention mechanism. The final output of the attention mechanism U is computed as follows:

$$U = \text{Annotate}(\alpha, C) = \alpha \cdot C. \tag{8}$$

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	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
Sum	1 099	947	608	289	2 943

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	2 093	5 590	674	721	9 959
Test	611	1 508	5 377	215	546	8 257
Sum	1 492	3 601	10 967	889	1 267	18 216

3. 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 (IEMOCAP) [29] and FAU Aibo Emotion corpus (FAU-AEC) [30] databases. The latter was thereby featured in the original Interspeech Emotion Challenge.

3.1. Datasets

IEMOCAP is a well-known corpus made up of audio-visual data with transcriptions of recordings of dialogues between two actors [29]. The corpus is divided into two parts: *improvise* and *script*. In our experiments, we only used the part *improvise* in order to reduce the potentially confounding effect of semantic information disturbance. However, the data distribution of each emotion class is heavily unbalanced. As in [31], we therefore merged the happy and excited utterances into the happy class. This merger results in the use of four emotion categories for training and evaluation: angry, happy, sad, and neutral. The final number of instances of each emotion class are given in Table 1.

We also used the FAU Aibo Emotion Corpus for evaluation, which is composed of spontaneous and emotional German speech samples [30]. The corpus contains 9.2 hours of German speech from a total of 51 children interacting with Sony's pet robot Aibo at two different schools. As per [32], 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 concentrated on the five-class problem with the emotion categories of anger, emphatic, neutral, positive, and rest. The final number of instances of each emotion class are given in Table 2.

3.2. Features

We used the extraction process described in [33] to form our spectrograms. In short, each spectrogram was constructed using the output of a 40-dimensional mel-scale log filter bank. These features were computed over frames of 25 ms length and 10 ms stride and normalised to be in the range [0,1].

3.3. Experimental setup and evaluation metrics

Using the IEMOCAP dataset, we performed a 10-fold cross-validation using a leave-one-out strategy, adopting the methodology of previous work. In each training process, eight speakers from four sessions were used as training data, and the remaining session was separated into two parts: one being regarded as validation data and the other as test data. For the FAU Aibo Emotion Corpus, we followed the Interspeech 2009 Emotion Challenge guidelines [32], employing utterances from one school (the Ohm-Gymnasium) for training and the other (the Montessori-Schule) for testing[1].

The proposed model has several parameters that are tuned based on the recommendations of previous work that utilised the same database. For CTC training, we made use of the *TensorFlow*[2] deep learning framework. The RMSProp optimiser was used to train our model in all the experiments, with a fixed learning rate of 10^{-3}. The mini-batch size was 32. We set 256 as the dimension for the bidirectional LSTM, and the network contains two hidden layers with 256 bidirectional LSTM cells (128 forward nodes and 128 backward nodes). Similar to [6], we also compared the performance when assigning the number of emotional labels as 1) the number of words in x, 2) the number of voiced phonemes in x, and 3) the double number of voiced phonemes in x.

For the evaluation of results generated by the two datasets, we used standard evaluation criteria. For IEMOCAP-generated results, unweighted and weighted accuracies (UA and WA respectively) are used as evaluation metrics. For FAU-AEC-generated results, we consider only unweighted accuracy (UA), since the FAU Aibo Emotion corpus is extremely unbalanced. The instance upsampling strategy was also applied for the FAU-AEC dataset.

3.4. Results and discussion

We observed that the BLSTM-CTC combined with attention mechanisms outperformed the BLSTM-CTC without the attention model on both datasets (cf. Table 3). For IEMOCAP, the best UA (67.0 %) and WA (69.0 %) were achieved by our proposed BLSTM-CTC with component attention, in a significant improvement over the baseline BLSTM-CTC ($p < 0.05$ in a one-tailed z-test). This set-up also achieved the best UA (42.9 %) on FAU-Aibo, which is also a significant improvement over the baseline BLSTM-CTC ($p < 0.05$ in a one-tailed z-test). Overall, the proposed approach outperformed the baseline BLSTM-CTC model. Using IEMOCAP, relative improvements of 3.0 % (UA) and 3.1 % (WA) were observed, while a relative improvement of 3.6 % in UA was observed using FAU Aibo Emotion Corpus.

Furthermore, we observed that the BLSTM-CTC with component attention performed better than BLSTM-CTC with local attention (Table 3). As the component attention is an improvement upon local attention, this higher performance validates our hypothesis that separately weighting each component of a feature is essential in the computation of the attention vector.

Comparing the two attention mechanisms introduced in this work, the performance of BLSTM-CTC with quantum attention is lower than BLSTM-CTC with component attention, but still surpasses the basic BLSTM-CTC model (in a one-tailed z-test, $p < .10$ for IEMOCAP, $p < .10$ for FAU). Therefore, the incorporation of quantum attention can still be considered a better-suited solution for SER.

The performance comparison between the different strategies on how to split an utterance into emotional segments in this work is consistent with the conclusions drawn in [6]; phoneme level segmentation achieves the best performance.

Table 3: *Performance comparison of the baseline BLSTM-CTC model and the improved BLSTM-CTC models enhanced by one of three different attention mechanisms reported on both the IEMOCAP and FAU-AEC datasets.*

Methods	IEMOCAP		FAU-AEC
[%]	UA	WA	UA
baseline BLSTM-CTC models w/o attention [6]			
Word Number	63.5	64.7	40.3
Phoneme Number	65.1	66.9	41.4
Phoneme Number×2	63.6	64.7	40.8
proposed BLSTM-CTC models enhanced w/ attention			
w/ Local Attention	66.3	68.0	41.8
w/ Component Attention	**67.0**	**69.0**	**42.9**
w/ Quantum Attention	65.4	68.0	42.4

Note: for IEMOCAP, we provide both unweighted and weighted accuracies (UA and WA respectively) as the evaluation metric, while for FAU-AEC, we only adopt UA as the evaluation measure, since this database is extremely unbalanced.
The performance of the BLSTM-CTC is evaluated based on assigning the number of emotional labels as the number of voiced phonemes in an utterance.

4. Conclusions

In this paper, we presented an effective hybrid sequence-to-sequence modelling approach for categorical speech-based emotion recognition. The experimental results indicate that our attention-based BLSTM-CTC approach achieves state-of-the-art performance on the IEMOCAP and FAU-AEC datasets. We hypothesise that the attention mechanisms yield improvements in system accuracy by allowing the model to focus on the emotionally salient parts of the speech signal.

In our experiments, CTC attention consistently outperformed the BLSTM-CTC approach. As our proposed hybrid CTC-attention architecture is easily adaptable, future work will focus on demonstrating its suitability in other computational paralinguistics tasks. In other future efforts, we will explore more effective sequence-to-sequence approaches to improve speech emotion recognition.

5. Acknowledgements

The work presented in this paper was substantially supported by the National Natural Science Foundation of China (Grant No. 61702370), the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300), the Open Projects Program of the National Laboratory of Pattern Recognition, and the Senior Visiting Scholar Program of Tianjin Normal University.

This project also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE). We thank Dr. Judith Dineley for her proof-reading work.

[1] We will provide a URL for a doc with details on all partitions and seeds upon acceptance.

[2] https://www.tensorflow.org

6. References

[1] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using Recurrent Neural Networks," in *Proc. ACII*, San Antonio, Texas, USA, 2017, pp. 190–195.

[2] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using Recurrent Neural Networks with local attention," in *Proc. ICASSP*, New Orleans, USA, 2017, pp. 2227–2231.

[3] C. W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition," in *Proc. INTERSPEECH*, San Francisco, California, USA, 2016, pp. 1387–1391.

[4] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion recognition from speech with Recurrent Neural Networks," https://arxiv.org/abs/1701.08071, 2017, 18 pages.

[5] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio-temporal representations by integrating attention-based Bidirectional-LSTM-RNNs and FCNs for speech emotion recognition," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 272–276.

[6] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, and B. Schuller, "Towards temporal modelling of categorical speech emotion recognition," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 932–936.

[7] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with Recurrent Neural Networks," in *Proc. ICML*, Pittsburgh, Pennsylvania, USA, 2006, pp. 369–376.

[8] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with Deep Recurrent Neural Networks," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 6645–6649.

[9] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. ASRU*, Scottsdale, Arizona, USA, 2015, pp. 167–174.

[10] K. Rao, A. Senior, and H. Sak, "Flat start training of CD-CTC-SMBR LSTM RNN acoustic models," in *Proc. ICASSP*, Shanghai, China, 2016, pp. 5405–5409.

[11] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. ICML*, New York, NY, USA, 2016, pp. 173–182.

[12] H. Inaguma, K. Inoue, M. Mimura, and T. Kawahara, "Social signal detection in spontaneous dialogue using Bidirectional LSTM-CTC," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1691–1695.

[13] X. Chen, W. Han, H. Ruan, J. Liu, H. Li, and D. Jiang, "Sequence-to-sequence modelling for categorical speech emotion recognition using Recurrent Neural Network," in *Proc. ACII Asia*, Beijing, China, 2018, pp. 1–6.

[14] J. Lee and I. Tashev, "High-level feature representation using Recurrent Neural Network for speech emotion recognition," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 1537–1540.

[15] A. Das, J. Li, R. Zhao, and Y. Gong, "Advancing connectionist temporal classification with attention modeling," in *Proc. ICASSP*, Calgary, Alberta, Canada, 2018, pp. 4769–4773.

[16] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NIPS*, Montreal, Canada, 2015, pp. 577–585.

[17] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Proc. NIPS*, 2015, pp. 2773–2781.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," https://arxiv.org/abs/1409.0473, 2014, 15 pages.

[19] Y. Zhang, J. Du, Z. Wang, and J. Zhang, "Attention based fully convolutional network for speech emotion recognition," in *Proc. APSIPA ASC*, Honolulu, Hawaii, USA, 2018.

[20] X. Niu, Y. Hou, and P. Wang, "Bi-directional LSTM with quantum attention mechanism for sentence modeling," in *Proc. ICONIP*, Guangzhou, China, 2017, pp. 178–188.

[21] B. Tamir and E. Cohen, "Introduction to weak measurements and weak values," *Quanta*, vol. 2, no. 1, pp. 7–17, 2013.

[22] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes – Towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008, pp. 597–600.

[23] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "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*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[24] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[25] M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[26] G. Keren and B. Schuller, "Convolutional RNN: an enhanced model for extracting features from sequential data," in *Proc. IJCNN*, Vancouver, BC, Canada, 2016, pp. 3412–3419.

[27] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[28] G. Chen, Y. Liu, J. Cao, S. Zhong, Y. Liu, Y. Hou, and P. Zhang, "Learning Music Emotions via Quantum Convolutional Neural Network," in *Proc. BI*. Beijing, China: Springer, 2017, pp. 49–58.

[29] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, Nov. 2008.

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

[31] R. Xia and Y. Liu, "DBN-ivector framework for acoustic emotion recognition," in *Proc. INTERSPEECH*, San Francisco, USA, 2016, pp. 480–484.

[32] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 312–315.

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