INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Temporal Attention Convolutional Network for Speech Emotion Recognition with Latent Representation

Jiaxing Liu[1,3]*, Zhilei Liu*[1,*]*, Longbiao Wang*[1,*]*, Yuan Gao*[1]*, Lili Guo*[1]*, Jianwu Dang*[1,2,3]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3]Pengcheng Laboratory, Shenzhen, China

{jiaxingliu, zhileiliu, longbiao_wang, yuan_gao, liliguo}@tju.edu.cn, jdang@jaist.ac.jp

Abstract

As the fundamental research of affective computing, speech emotion recognition (SER) has gained a lot of attention. Unlike with common deep learning tasks, SER was restricted by the scarcity of emotional speech datasets. In this paper, the vector quantization variational automatic encoder (VQ-VAE) was introduced and trained by massive unlabeled data in an unsupervised manner. Benefiting from the excellent invariant distribution encoding capability and discrete embedding space of VQ-VAE, the pre-trained VQ-VAE could learn latent representation from labeled data. The extracted latent representation could serve as the additional source data to make data abundantly available. While solving data lacking issue, sequence information modeling was also taken into account which was considered useful for SER. The proposed sequence model, temporal attention convolutional network (TACN) was simple yet good at learning contextual information from limited data which was not friendly to complicated structures of recurrent neural network (RNN) based sequence models. To validate the effectiveness of the latent representation, t-distributed stochastic neighbor embedding (t-SNE) was introduced to analyze the visualizations. To verify the performance of the proposed TACN, quantitative classification results of all commonly used sequence models were provided. Our proposed model achieved state-of-the-art performance on IEMOCAP.

Index Terms: speech emotion recognition, temporal convolutional network, sequence modeling

1. Introduction

Speech signals as the most commonly used communication method for humans, not only carry lots of content information but also implicit paralinguistic information about the speakers. Speech emotion recognition (SER) gives valuable information that could improve dialog systems in human-computer interaction. Now, SER has become an attractive research field [1].

SER was a challenging and meaningful task. As we know, the performance of automatic speech recognition (ASR) and image classification tasks had been better than humans, unlike with SER tasks which were still not competitive to trained human listeners. Two main reasons cause this phenomenon. One was the ever-lacking large and naturalistic databases [2]. Compared to general speech datasets, recording and annotating an emotion-related dataset was more time-consuming. The numbers of speakers, recording conditions and the size of the corpus were also much more limited; the other reason was that the extracted emotional features were not efficient and effective enough. Feature extraction was a critical step to bridge the gap between speech signals and speaker emotions. Finding effective emotional feature representation through feature extraction was a direction that required continuous attention [3, 4].

To solve the data lacking issue, one was to integrate these existing emotion datasets and set them in similar annotation space [5]. Due to various factors, the speech signals in different datasets had complex distributions with a high variance which led to this solution being more uncontrollable. Therefore, we focused on the other solution, learning intrinsic expression across these datasets. This strategy usually introduced autoencoder (AE) [6] and denoising autoencoder (DAE) [7] to reconstruct the input to learn representations, and had achieved success in many fields such as image classification [8], speaker identification [9] and speech conversion [10]. Compared with AE and DAE emphasized on input reconstruction, variational autoencoder (VAE) [11] was optimized for latent representation learning. However, the learning representation processing in VAE was continuous. As we know, speech signals were inherently discrete, and typically represented as a sequence of symbols. Therefore, VAE was not a natural fit for modeling speech signals. The Vector Quantised Variational AutoEncoder (VQ-VAE) [12] was a simple yet powerful model. Discrete representation learning of VQ-VAE was more conducive to reasoning and modeling of speech signals. At the same time, the learned representations in VQ-VAE spaned various dimensions as opposed to focusing on local details. This representation learning method was very appropriate for maintaining emotional information which was related to many factors, but not very decisively related to imperceptible details.

In recent years, deep neural network (DNN) based methods were proposed to overcome the pre-defined limitations of traditional methods [13, 14, 15, 16]. These deep learning-based models, e.g. DNN [17], convolutional neural networks (CNN) [18, 19] achieved competitive results. But the methods ignored the sequence information modeling (contextual information) which was essential to identify the emotional states. In most deep learning models, sequence modeling was synonymous with recurrent networks. Although the works of recurrent neural network (RNN) [20] and bidirectional long short term memory (BLSTM) [21] took into account the problems of sequence modeling, these RNN-based models had complicated structures, slow training speed and could not be fully trained when data was insufficient. The temporal convolutional network (TCN) [22] was based on CNN, and the combination of causal convolutions and dilated convolutions in TCN provided the ability to model the sequence. But this model lacked the

* Corresponding Author

http://dx.doi.org/10.21437/Interspeech.2020-1520

consideration of the position importance in sequence.

To address these two issues, data lacking and sequence modeling, an improved SER model is proposed as shown in Figure 1, which mainly consists of the latent representation learning module and SER considering sequence information module. At first, VQ-VAE was trained on massive unlabeled data to learn pre-trained embedding space. The labeled data was feed to the traditional CNN network and pre-trained VQ-VAE respectively. The labeled data fine-tuned the VQ-VAE, and the representations in the first layer of the decoder were extracted as the latent representations. The latent representations were the additional source of information that concatenated with deep representation learned by CNN to get the fusion representations. We improved the TCN with attention mechanisms and proposed temporal attention convolutional networks (TACN). The fusion representations were feed to TACN to achieve the classification results.

The major contributions of this paper were summarized as 1) An unsupervised VQ-VAE was pre-trained in an unsupervised manner to extract the latent representations of labeled data. 2) A TACN was proposed to model the sequence information for SER.

2. Sequence Modeling integrating Latent Representation

The proposed model as shown in Figure 1 mainly consisted of two modules which were latent representation learning module and sequence modeling SER module. The Latent representation R_l was learned by the first module. The Deep representation R_d was learned by CNN. These two kinds of representation were fused as Fusion representation R_f whcih is followed by a fully connected layer. The output was fed to the proposed TACN to learn the sequence information and which was also set to be the classifier to predict the emotional states.

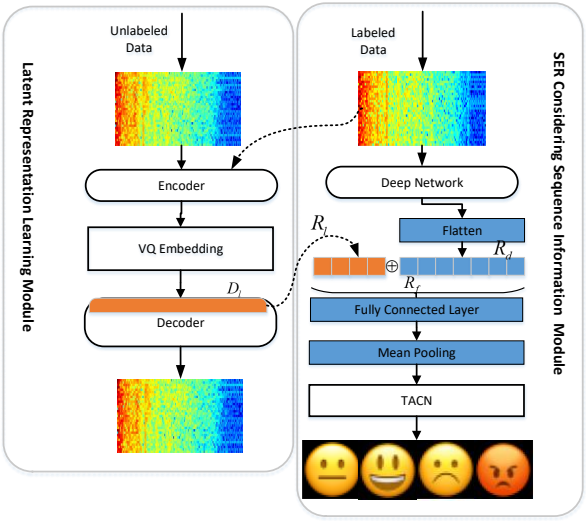


Figure 1: *TACN integrating latent representation.*

2.1. Latent representation learning

The extracted Latent representation as the additional source for SER, an unsupervised model VQ-VAE was introduced. With the good performance of the embedding space in trained VQ-VAE, more rich and robust emotion-related informations were contained directly or indirectly. The pre-training data consisted of EmoV-DB [23] and LibriSpeech which were built for the purpose of emotional speech synthesis and ASR. The VQ-VAE trained by unlabeled data could obtain a good prior latent embedding space and a powerful encoder-decoder structure. The core part of VQ-VAE was VQ embedding as shown in Figure 2. The embedding space had modeled the internal distribution of unlabeled data in pre-trained VQ-VAE. The D_i was the input of the decoder which is calculated as shown in Eq. (1).

$$D_i(x) = e_k \quad \text{where} \quad k = argmin_j \|E_O - e_j\|^2 \tag{1}$$

where the high dimension representation E_o represented the output of the encoder. The processing of extracting D_i was a dictionary learning algorithm. The discrete embedding space was the dictionary, the vector quantization was the query. The query was calculated by a nearest neighbour look-up using the shared embedding space to determine the subscript k. The corresponding embedding vector e_k replaced the vector in vector quantisation map. Finally, E_o was rebuilt to D_i by the pre-trained VQ embedding processing. The red line in Figure 2 represented the gradient during the backward from decoder unaltered to the encoder. Through this pre-trained VQ-VAE model, the various latent information and internal relations in the labeled data would be compressed into D_i which was followed by a Flatten layer to get Latent representation R_l

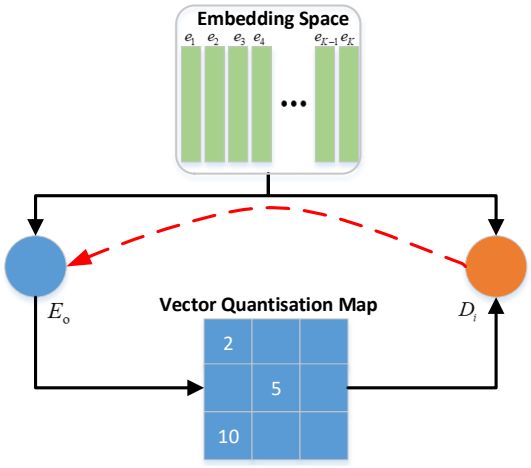


Figure 2: *Vector quantisation embedding.*

2.2. Temporal attention convolutional network

Broadly speaking, sequence modeling was that given an input sequence $[x_0, x_1, x_2...x_{T-2}, x_{T-1}, x_T]$, and wanted to predict the corresponding outputs $[y_0, y_1, y_2...y_{T-2}, y_{T-1}, y_T]$. The key constraint was to predict the output y_t for some time t only could use those inputs that had been previously observed: $[x_0, ..., x_t]$. The future inputs $[x_t + 1, ..., x_T]$ could not be used in the prediction of y_t. The sequence mapping was shown in Eq. (2) and the f was the modeling function.

$$y_0, ..., y_T = f(x_0, ..., x_T) \tag{2}$$

The prediction for later steps in traditional RNN-based methods must wait for their predecessors to complete. The RNN-based methods could not be calculated in parallel. Some RNN-based models, such as LSTM and gated recurrent unit (GRU) used 'gate' to model the sequence, whose structures were complicated and hard to be fully trained in small size data.

Considering the importance of sequence modeling in SER and scarcity of emotional speech, TCN was introduced as the sequence modeling method.

TCN mainly consisted of two parts, a 1D fully-convolutional network (FCN) architecture and causal dilated convolutions. The FCN made each hidden layer in TCN have the same length as the input layer. The causal dilated convolution was an exquisite design that let the convolutional layers have the ability of sequence modeling. Based on TCN, the position importance of hidden layer data was considered and proposed the TACN model as shown in Figure 3.

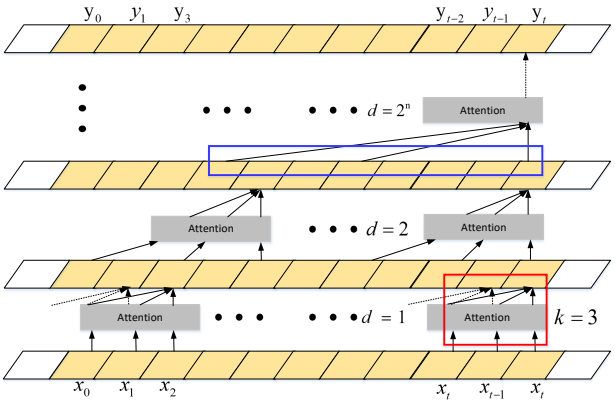


Figure 3: *Temporal attention convolutional network.*

To achieve the constraint mentioned before, future information could not leak to the past. The first basic design was causal convolution as shown in the red box in Figure 3. In this convolution, an output at time t was convolved only with elements from time t and earlier in the previous layer.

In theory, causal convolution could model a long effective input through an extremely deep network or very large filters. But when applying the causal convolution in sequence tasks, it costed too much time and memory, neither of the mentioned two methods was particularly feasible. The dilated convolution was shown in the blue box in Figure 3 which solved those mentioned issues.

The dilated convolution operation F on element s of the sequence was defined as:

$$F(s) = (\mathrm{X} *_d f)(s) = \sum_{i=0}^{k-1} f(i) \cdot \mathrm{X}_{s-d \cdot i} \tag{3}$$

where the X was the input sequence, f was the filter, d was the dilation factor, $s - d \cdot i$ represented the direction of the past. When $d = 1$, a dilated convolution reduced to a regular convolution. The d was fixed in the same hidden layer and a multiple difference between adjacent layers. Usually we increased d exponentially with the depth of the network, and $d = 2^n$, n was the n-th hidden layer. With the help of the increase of the dilation factor, the receptive field grew exponentially. But the accompanying problem was ignored. The distance between the convolution elements in the sequence was farther away. Treating these dilated elements equally did not meet the needs of adaptation.

In the sequence model tasks, adding a weight to each input element was a common operation[24, 25, 26] Given a sequence S_n, which was in the n-th hidden layer.

$$S_n = [\hat{x}_{t-k \cdot d}, \cdots \hat{x}_{t-d}, \hat{x}_t] \quad \text{where} \quad d = 2^n \tag{4}$$

where k was the filter size, and $\hat{x}$ were the elements in the hidden layer that had the same importance degrees. We improved this situation through a multi-head self-attention method [27].

$$Att(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{D_K}}\right) V \tag{5}$$

In Eq. (5), the input matrix consists of Q, K, V which represented queries, keys, and values respectively and the dimension of keys was D_K. Instead of performing a single calculation of Q, K, V, it was beneficial to linearly project the queries, keys, and values, h times with different learned linear projections. The h results were concatenated and once projected, resulting in the final output $M_h(Q, K, V)$ as shown in Eq. (6).

$$M_h(Q, K, V) = W(Att_1 + ... + Att_h) \tag{6}$$

when $Q == K == V == S_n$, this attention method was called multi-head self-attention. The S_n was calculated h times without sharing parameters and the h results were projected to $\hat{S}_n$ as shown in Eq. (7).

$$\hat{S}_n = M_h(S_n) \tag{7}$$

In this paper, the number of hidden layers n was 8, the filter size k was 3, the head number h was 8.

3. Experiments and Analysis

3.1. Experimental setup

Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [28] is used in the experiments. The audio data samples at 16HKz with 5,531 utterances which consists of four emotion categories: Neutrality (29%), Anger (20%), Sadness (20%), and Happiness (31%). In this paper, we use the same preprocessing method, segment length, and parameters of CNN with Satt et al. [21]. The time of each segment is 265-ms and the input spectrogram has the following $time \times frequency$: 32×129. We chose cross-entropy as the loss function, Adadelta as the optimizer, and ReLU as the activation. The batch_size is set as 128. The data is randomly split to 80% training set and 20% testing set and still represents to the same proportions in the training/testing sets as in the whole corpus.

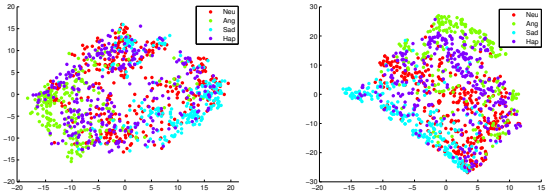
3.2. Experiment results and analysis

To verify the effectiveness of the proposed model, we set up three groups of comparative experiments to confirm the validity of Latent representation, evaluate the performance of proposed TACN, and evaluate the performance of the proposed whole model.

3.2.1. Validation of the extracted latent representation

To observe the changes brought by Latent representation, t-distributed stochastic neighbor embedding (t-SNE) [29] was introduced to visualize the difference between the Deep representation R_d and Fusion representation R_f. The visualization was shown in Figure 4.

The green points (Anger) performed well in both plots. The blue points (Sadness) in Figure 4(b) performed better, the distance of points were closer to each other. The red points (Neutral) were the same situation with blue points, the distance in



(a) Deep representation (b) Fusion representation

Figure 4: *The t-SNE visualization of Deep and Fusion representation.*

Figure 4 (b) was shorter. The purple points represented Happiness, the points distributed cross everywhere in both plots, the same situation was also reported [30]. Relatively speaking, the addition of Latent representation in Figure 4 (b) let these four emotion points have more distinct distributions.

3.2.2. Validation of the proposed TACN

As shown in Table 1, the proposed model TACN outperformed all the commonly used RNN-based sequence models and TCN. Compared with RNN-based models, the proposed model TACN had absolute increases of more than 4.34% and 4.88% on Deep and Fusion representation. The performance of TACN was also better than TCN and had absolute increases of 2.08% and 2.71% on Deep and Fusion representation.

Table 1: *The results of comparative experiments*

		Deep	Fusion
RNN-based	SimpRNN [20]	57.41	59.76
	GRU	59.58	61.21
	BGRU	62.12	64.20
	LSTM	59.13	61.57
	BLSTM [21]	61.75	64.56
CNN-based	TCN	64.38	66.73
	TACN	**66.46**	**69.44**

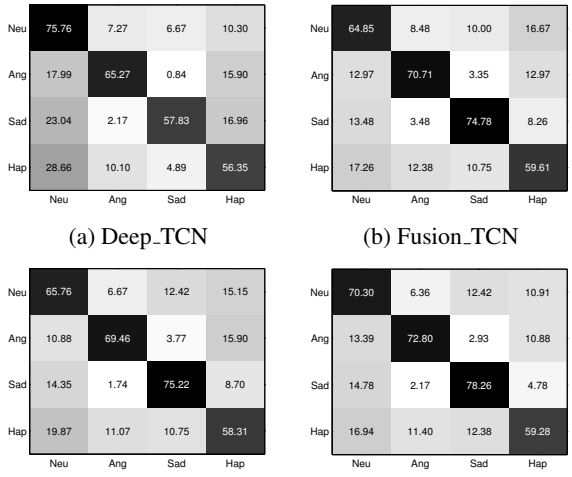
3.2.3. Validation of the proposed model

To quantitatively analyze the proposed whole model, we provided the results of the ablation studies, whose evaluation criteria were F1 score (F1), weighted accuracy (WA), and unweighted accuracy (UA) as shown in Table 2. Also, the four confusion matrices of the experiments were shown in Figure 5.

Table 2: *The results of ablation studies*

Model	F1(%)	WA(%)	UA(%)
Deep_TCN	64.68	66.38	63.80
Fusion_TCN	67.00	66.73	67.48
Deep_TACN	66.98	66.46	67.18
Fusion_TACN (our)	**69.75**	**69.44**	**70.16**

Observing Table 2 and Figure 5, three phenomena could be got. The first one was in Deep_TCN, massive four emotions were rudely identified as Neutrality that led to the worst F1 score. The second one was in Fusion_TCN and Deep_TACN, all four emotions performed more stable and more normal, indicating the effectiveness of the Latent representation addition and the proposed TACN respectively. The third phenomenon was that the proposed model Fusion_TACN outperformed all three models and performed best in three emotions. It was recognized that the classification of Happy emotion in IEMOCAP was difficult [31, 32], due to not only the dataset annotation problems but also other relevant factors that need more research in the future.



(a) Deep_TCN (b) Fusion_TCN

(c) Deep_TACN (d) Fusion_TACN

Figure 5: *The confusion matrices.*

4. Conclusions

In this paper, a natural fit for the speech latent representation learning model, VQ-VAE, was introduced in an unsupervised manner. The latent representation involved invariant distribution which was different from deep representation learned by the supervised network. The visualization analysis and quantitative experiments proved its effectiveness. Sequence modeling was often overlooked in SER, in this paper a CNN-based architecture of TACN was proposed. The proposed TACN not only achieved better performance than commonly used RNN-based models but also overcame the shortages of TCN which was unstable in emotion recognition. Detailed experimental results and confusion matrices also verified that the proposed model was outstanding. The proposed model achieved the performance of 66.46% and 69.44% on Deep and Fusion representation. Compared with RNN-based models and TCN, our model had absolute increments of more than 4.88% and 2.71%. With the high sensitivity to all the emotions and stable performance, our model shows great potential to solve the cross-corpus SER tasks and multimodel emotion recognition tasks.

5. Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1305200, the National Natural Science Foundation of China under Grant 61771333 and the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330.

6. References

[1] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.

[2] B. W. Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[3] K. John and R. Saurous, "Emotion recognition from human speech using temporal information and deep learning," *Interspeech 2018*, pp. 937–940, 2018.

[4] S. Ramakrishnan and I. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, 2013.

[5] H. Luo and J. Han, "Cross-corpus speech emotion recognition using semi-supervised transfer non-negative matrix factorization with adaptation regularization," *Proc. Interspeech 2019*, pp. 3247–3251, 2019.

[6] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[7] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[9] J. Villalba, N. Brümmer, and N. Dehak, "Tied variational autoencoder backends for i-vector speaker recognition." in *INTERSPEECH*, 2017, pp. 1004–1008.

[10] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *stat*, vol. 1050, p. 1, 2014.

[12] A. van den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[13] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics*, 2016, pp. 5200–5204.

[14] A. Zakaria and P. Mower, "Using regional saliency for speech emotion recognition," in *IEEE ICASSP*, 2017, pp. 2741–2748.

[15] L. Guo, L. Wang, J. Dang, L. Zhang, H. Guan, and X. Li, "Speech emotion recognition by combining amplitude and phase information using convolutional neural network," in *Proc. Interspeech 2018*, 09 2018, pp. 1611–1615.

[16] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local-global aware deep representation learning," in *Proc. of ICASSP 2020*, 05 2020, pp. 7174–7178.

[17] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[18] L. Guo, L. Wang, and J. Dang, "A feature fusion method based on extreme learning machine for speech emotion recognition," *ICASSP 2018*, pp. 2666–2670, 2018.

[19] L. Guo, L. Wang, J. Dang, Z. Liu, and H. Guan, "Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine," *IEEE Access*, vol. 7, pp. 75 798 – 75 809, 06 2019.

[20] J. Lee and I. Ivan, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. Interspeech*, 2015, pp. 1537–1540.

[21] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. Interspeech*, 2017, pp. 1089–1093.

[22] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[23] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *arXiv preprint arXiv:1806.09514*, 2018.

[24] L. Zhou, J. Zhang, C. Zong, and H. Yu, "Sequence generation: From both sides to the middle," in *IJCAI*, 2019.

[25] D. Ulmer, D. Hupkes, and E. Bruni, "Assessing incrementality in sequence-to-sequence models," in *Workshop on Representation Learning for Nlp*, 2019.

[26] D. Kastaniotis, I. Ntinou, D. Tsourounis, G. Economou, and S. Fotopoulos, "Attention-aware generative adversarial networks (ata-gans)," in *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, 2018.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, and I. Kaiser, Ł.and Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[28] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, and et.al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, p. 335, 2008.

[29] L. V. D. Maaten, "Learning a parametric embedding by preserving local structure," *Journal of Machine Learning Research*, vol. 5, pp. 384–391, 2009.

[30] M. Neumann and T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," 02 2019, pp. 7390–7394.

[31] M. Chen, X. H. andJ. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, pp. 1440–1444, 2018.

[32] J.-H. Tao, J. Huang, Y. Li, Z. Lian, and M.-Y. Niu, "Semi-supervised ladder networks for speech emotion recognition," *International Journal of Automation and Computing*, no. 2, 2019.