HIERARCHICAL ATTENTION TRANSFER NETWORKS FOR DEPRESSION ASSESSMENT FROM SPEECH

Ziping Zhao[1,2]*, Zhongtian Bao*[1]*, Zixing Zhang*[3]*, Nicholas Cummins*[2]
Haishuai Wang[4,1]*, 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] Group on Language, Audio & Music, Imperial College London, UK
[4] Department of Computer Science and Engineering, Fairfield University

ABSTRACT

A growing area of mental health research is the search for speech-based objective markers for conditions such as depression. However, when combined with machine learning, this search can be challenging due to a limited amount of annotated training data. In this paper, we propose a novel cross-task approach which transfers attention mechanisms from speech recognition to aid depression severity measurement. This transfer is applied in a two-level hierarchical network which mirrors the natural hierarchical structure of speech. Experiments based on the Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) dataset, as used in the 2017 Audio/Visual Emotion Challenge, demonstrate the effectiveness of our Hierarchical Attention Transfer Network. On the development set, the proposed approach achieves a root mean square error (RMSE) of 3.85, and a mean absolute error (MAE) of 2.99, on a Patient Health Questionnaire (PHQ)-8 scale [0, 24], while on the test set, it achieves an RMSE of 5.66 and an MAE of 4.28. To the best of our knowledge, these scores represent the best-known speech-only results to date on this corpus.

Index Terms— Depression, Attention Transfer, Hierarchical Attention, Monotonic Attention

1. INTRODUCTION

Major depression disorders are highly prevalent and can cause a substantial burden for associated individuals, families and society. Early interventions, aimed at predicting the onset of clinical depression, represent an essential means to help reduce this burden. To aid the depression diagnosis, the problem of automatically detecting and monitoring depression from speech signals has recently gained considerable attention [1]. By providing a common research platform, the *Audio/Visual Emotion Challenge* (AVEC) series has helped accelerate these research efforts [2, 3, 4, 5, 6].

In the field of depression analysis from speech, *Convolutional Neural Networks* (CNNs) are, at present, the predominant deep-learning architecture for data-driven learning systems [7, 8, 9]. Despite having the advantage over CNNs in that they are capable of modelling the sequential structure of speech, the suitability of *Recurrent Neural Networks* (RNNs), remains understudied. This advantage can be inferred from the related field of *Speech Emotion Recognition* (SER), in which RNN paradigms are frequently employed [10, 11, 12].

A potential reason for the lack of RNN-based approaches in depression analysis could be the structure of the associated databases. Speech depression corpora contain recordings measuring some minutes in length with only one associated label over the whole length, the associated depression score. It has been previously demonstrated that RNNs typically struggle in such learning conditions [13].

The inclusion of attention mechanisms, in particular hierarchical attention mechanisms [14, 15], into the RNN framework is one possible solution to this issue. The benefits of a hierarchical attention mechanism have been demonstrated in *Natural Language Processing* (NLP) [16, 17], as the structure of these networks naturally mirrors the hierarchical structure of documents. However, the inclusion of attention mechanisms increases the number of learnable parameters in the associated models. This increase does not fit with smaller, in terms of the number of unique samples, depression corpora [1]. The attention mechanism can, however, be used in combination with transfer learning paradigms [18, 19, 20].

In this regard, we herein propose a novel depression analysis framework combining a hierarchical attention strategy with an attention transfer mechanism. The proposed *Hierarchical Attention Transfer Network* (HATN) model automatically transfers attentions, learnt from speech recognition, at both the frame and sentence levels. To the best of our knowledge, this is the first time that such a study has been conducted for depression severity measurement. Our key contributions are summarised as follows: i) we introduce an attention transfer process which can transfer attentions across tasks at both the frame and sentence level; ii) extensive experiments demonstrate that the proposed hierarchical attention transfer

model outperforms other state-of-the-art approaches for the task of speech-based depression severity assessment.

2. RELATION TO PRIOR WORK

We demonstrate the advantages of our proposed approach on the *Distress Analysis Interview Corpus – Wizard of Oz* (DAIC-WOZ) [21] database, as partitioned for the depression detection task of AVEC 2017 [5]. A range of audio-only approaches was presented in the challenge; these include knowledge-driven approaches [5, 22], as well as deep learning-based data-driven systems [8, 23].

The work presented here has focused on the depression analysis by leveraging the hierarchical attention strategy combined with an attention transfer mechanism. While the use of attention mechanisms, specifically hierarchical attention mechanisms [14, 15]. This approach has achieved state-of-the-art performance in various NLP document-based classification tasks [14, 15, 24]. To the best of our knowledge, the advantages of using attention mechanisms in speech-based depression analysis have yet to be established.

We use a transfer learning paradigm to learn the attention weights [18, 19, 20]. Recent works in this regard have generally focused on computer-vision-related tasks, with the developed spatial attention maps being designed for CNNs [18, 19]. Encouraged by the recent success of attention transfer mechanisms and unsupervised learning, the present study presents a novel attention transfer process designed for BLSTM.

3. HIERARCHICAL ATTENTION TRANSFER NETWORK

In this section, we first present an overview of the proposed Hierarchical Attention Transfer Network (HATN) model for cross-task depression severity measurement. We then outline the core technical details of the model.

The HATN model consists of three components (Fig. 1). The first is the *speech recognition model* which we train to generate our initial attention maps The second component is the *attention transfer mechanism*. The workings of this component are inspired by the activation-based attention model presented in [18]. In [18], attention transfer is achieved by training a shallower student network for the target task, which mimics the attention maps of a deeper teacher network. The third component is the *depression recognition module*. In this component, we use a hierarchical attention neural network, which consists of frame-level and sentence-level attentions.

3.1. Monotonic Attention

As highlighted in Section 1, attention-based encoder-decoder models are an effective approach across many sequence-based learning tasks. However, a major drawback of this approach is that the model has to pass over the full input sequence to produce the output sequence [25]. This aspect results in increased computational complexity and does not allow for online decoding. Recently, *monotonic attention* mechanisms have been introduced to alleviate these issues [25, 26]. The issues of quadratic-time complexity and no option for online decoding in conventional soft-attention, are brought about by the trained attention mechanisms needed to inspect every entry of the model's memory at each output time-step (Fig. 2).

For the sake of brevity, we do not fully describe the monotonic attention training process within this paper. The interested reader is referred to [25, 26] for these details.

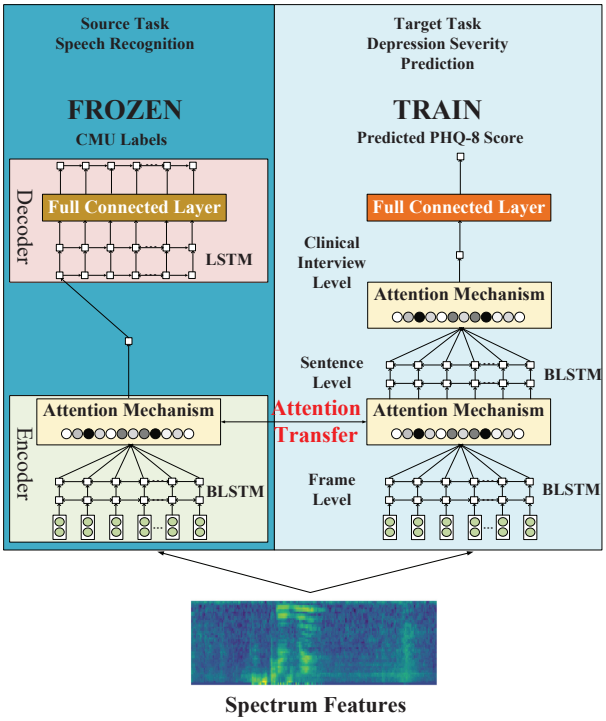
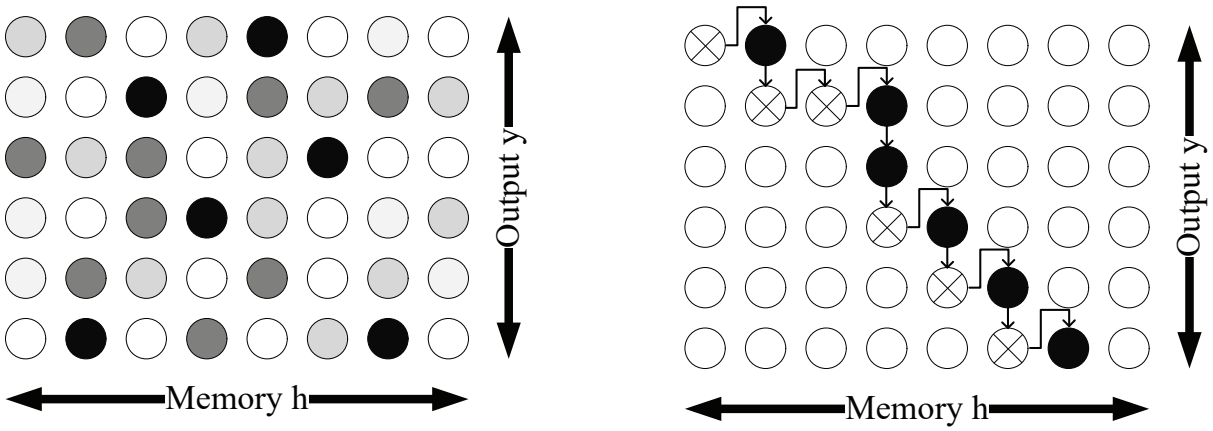


Fig. 1: The framework of the HATN model. The model firstly learns attentions through a speech recognition task, which are transformed into the hierarchical depression detection system.

3.2. Activation-based Attention Transfer

When performing attention transfer, the goal is to train a student network using the spatial attention maps of a teacher network such that the student network will not only make correct predictions but will also have attention maps that are similar to those of the teacher. The first step in this process is to define the spatial attention maps used in the teacher model. As our teacher model will be a bidirectional-LSTM (BLSTM), we consider a BLSTM layer and its corresponding activation tensor $A \in \mathbb{R}^{C \times H \times W}$ which consists of C ($C = 1$ for BLSTM) channels with spatial dimensions $H \times W$, and a mapping function F that takes the above BLSTM layer activations A (3D tensor) as input and outputs. A spatial attention map can then be defined as:

$$F : \mathbb{R}^{C \times H \times W} \rightarrow R^{H \times W}. \quad (1)$$



(a) Soft Attention

(b) Monotonic Attention

Fig. 2: Schematics diagrams to highlight the difference between soft and monotonic attention. In soft attention (a), the model assigns a probability (the different shading of each node) to each memory entry at each output timestep. The context vector is computed as the weighted average of the memory. Monotonic attention (b), inspects memory entries from left-to-right monotonic manner. It chooses whether to move on to the next memory entry (shown as nodes with $\times$) or stop and attend (shown as black nodes). The context vector is then assigned to the memory entries attended to. Figure adapted from [25, 26].

As the absolute value of a hidden neuron activation indicates the importance of that neuron with respect to a specific input, we can construct a spatial attention map by computing the statistics of these absolute values across the channel dimension. Specifically, we consider the following spatial attention mappings:

$$(F(A))_{i,j} = \sum_{k=1}^{C} |A_{k,i,j}|^p, \tag{2}$$

where $i \in \{1, 2, ..., H\}$ and $j \in \{1, 2, ..., W\}$ are spatial indexes.

During the attention transfer, we assume, without loss of generality, that transfer losses are placed between student and teacher attention maps and are of the same spatial resolution. However, the attention maps can be interpolated to match their shapes if required. We can then define the following total loss:

$$L_T = L_D + W_{AT} \times L_{AT}, \tag{3}$$

where L_D denotes the loss of the depression recognition task and W_{AT} the weight of attention transfer. L_{AT} denotes the loss of attention transfer, which can be computed as:

$$L_{AT} = \sum_{j \in \mathcal{I}} \left\| Q_D^j - Q_S^j \right\|_1, \tag{4}$$

here, $\mathcal{I}$ denotes the indices of the attention map, while Q_D^j and Q_S^j represent the j-th pair of the attention map of the depression recognition and speech recognition tasks respectively. As can be seen, during attention transfer we make use of $l1$-normalized attention maps.

3.3. Hierarchy attention model

Clinical interviews, such as those in the DIAC-WOZ corpora [21], generally consist of questions from a (virtual) therapist and answers from a participant. When attempting to use such interviews within a machine learning framework, we can assume that not all answers, herein referred to as sentences, will contribute equally to the associated depression score. Moreover, the sentences themselves consist of frames of information, with each frame having a different influence on the representation of the corresponding sentence.

Given this observation, our depression detection model has a hierarchical structure, consisting of two levels of attention mechanisms applied at the sentence and frame level. This structure enables the model to attend differentially, to more and less important content when performing its analysis. For a clinical review with m sentences$\{S_1, S_2, ..., S_m\}$, the $i-th$ sentence is S_i, which consists of l_i frames as $S_i = f_1^i f_2^i ... f_{l_i}^i$, and f_t^i is the $i-th$ frame in S_i, $t \in [0, l_i]$.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Corpus

As previously mention, we utilised the (DAIC-WOZ) [21] database, as partitioned for the depression detection task of AVEC 2017 [5]. The task is to predict the exact *Patient Health Questionnaire* (PHQ)-8 depression index score [27] in the range [0,24] associated with the provided files. During our analysis, we split the individual DAIC-WOZ recordings into individual participant turns based on the manual transcriptions provided with the corpus.

4.2. Features

As in [28], we use Mel-spectra as our model input. The spectrograms were 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].

4.3. Model Parameters

In order to implement attention transfer, we first trained an ASR system to acquire the attention maps. The ASR model is trained on the DAIC-WOZ dataset, and we reduced the number of states in the origin transcriptions by leveraging the CMU pronouncing dictionary [29]. We utilised a BLSTM network with 128 blocks, with the learning rate set to 10^{-4}. During model training, we investigate the impact of different attention strategies, namely standard global soft attention, local soft attention [30], and monotonic attention.

Once the above training step of the ASR unit is complete, its parameters are frozen, and the training of the depression recognition model can commence. For this, we also used a two-layer BLSTM, which consists of 128 single-memory-cell LSTM memory blocks in the forward and backward hidden layers. The learning rate is again set to 10^{-4}. Finally, the outputs of the fully connected layers can be regarded as the final predicted PHQ-8 score.

4.4. Results and Discussion

As depression severity prediction is a regression task, the accuracy metric for the challenge was the *Root Mean Square Error* (RMSE) and *Mean Absolute Error* (MAE). When comparing the scores obtained from different variations of our proposed approach (Table 1), we observed that the strongest results, MAE of 2.99 and RMSE of 3.85 on the development set and MAE of 4.28 and RMSE of 5.66 on the test set, were achieved by the hierarchical monotonic attention model combined with the attention transfer mechanism. This result supports our hypothesis that learning to mimic the attention maps of the teacher model can be helpful in depression analysis. It also highlights the advantage of reducing the complexity associated with the attention mechanism.

Furthermore, we observed that, on both the development and the test sets, our proposed model outperformed the AVEC 2017 baseline (Table 1). Moreover, the performance of our hierarchical monotonic attention model with attention transfer on the test set is even better than the multimodal CNN-based approach presented in [23], and almost matches the approach presented in [8]. To the best of the author's knowledge, the obtained results are the best speech-only-based results achieved on this experimental corpus to date.

Table 1: The RMSE and MAE scores of the experiment of (Hier)archical Attention Transfer Network gained on both the dev(elopment) and test sets of the DAIC-WOZ corpus. Note that AT denotes attention transfer

Methods	RMSE	MAE
Development Partition		
Hie. global soft attention	5.26	3.67
Hie. local soft attention (classic)	5.20	3.59
Hie. local soft attention (monotonic)	4.87	3.02
Hie. global soft attention w/AT	4.14	3.65
Hie. local soft attention (classic) w/AT	4.07	3.56
Hie. local soft attention (monotonic) w/AT	**3.85**	**2.99**
AVEC 2017 Audio Baseline [5]	6.74	5.36
AVEC 2017 Audio-Video Baseline [5]	6.62	5.52
DCNN-DNN [23]*	4.65	3.98
Multivariate regression model [8]*	3.09	2.48
Test Partition		
Hierarchical global soft attention	6.54	5.03
Hie. local soft attention (classic)	6.43	4.99
Hie. local soft attention (monotonic)	6.14	4.76
Hie. global soft attention w/AT	5.96	4.78
Hie. local soft attention (classic) w/AT	5.81	4.76
Hie. local soft attention (monotonic) w/AT	**5.66**	**4.28**
AVEC 2017 Audio Baseline [5]	7.78	5.72
AVEC 2017 Audio-Video Baseline [5]	7.05	5.66
DCNN-DNN [23]*	5.97	5.16
Multivariate regression model [8]*	5.40	4.36

* Indicates a multimodal system was utilised.

5. CONCLUSION

In this contribution, we presented a novel hierarchical attention-based model, which transferred attentions from speech recognition, for the task of depression severity detection. This approach is highly suitable for speech-based depression detection, in that it uses hierarchical attention that mirrors the structure of the speech present within the clinical interviews. State-of-the-art experimental results achieved on the DAIC-WOZ verified the suitability of this approach. Future work will explore transferring attention from and to other speech-related tasks.

6. ACKNOWLEDGEMENTS

The work presented in this paper was substantially supported by the the National Natural Science Foundation of China (Grant No: 61702370), the National Science Fund for Distinguished Young Scholars (Grant No: 61425017), the National Natural Science Foundation of China (Grant No. 61702370), the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300), the technology plan of Tianjin (Grant No: 18ZXRHSY00100). This project also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

7. REFERENCES

[1] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, July 2015.

[2] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proc. AVEC*, Barcelona, Spain, 2013, pp. 3–10.

[3] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge," in *Proc. AVEC*, Orlando, Florida, USA, 2014, pp. 3–10.

[4] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. AVEC*, Amsterdam, The Netherlands, 2016, pp. 3–10.

[5] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Proc. AVEC*, Mountain View, California, USA, 2017, pp. 3–9.

[6] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E. Messner, et al., "AVEC 2019 workshop and challenge: State-of-mind, depression with ai, and cross-cultural affect recognition," in *Proc. 9th International Audio/Visual Emotion Challenge and Workshop (AVEC)*, Nice, France, 2019.

[7] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of Biomedical Informatics*, vol. 83, pp. 103–111, May 2018.

[8] L. Yang, H. Sahli, X. Xia, E. Pei, M. Cédric Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," in *Proc. AVEC*, Mountain View, CA, USA, 2017, pp. 45–51.

[9] L. Yang, D. Jiang, and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures," *IEEE Transactions on Affective Computing*, 2018.

[10] 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.

[11] 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.

[12] C. 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.

[13] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, Jakob S. Grue, and J. Nie, "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion," in *Proc. CIKM*, Melbourne, Australia, Oct. 2015, pp. 553–562.

[14] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. NAACL-HLT*, San Diego, California, USA, 2016, pp. 1480–1489.

[15] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. ACL*, Melbourne, Australia, 2018, pp. 2225–2235.

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

[17] 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.

[18] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. ICLR*, Toulon, France, 2017, 13 pages.

[19] J. Zhuo, S. Wang, W. Zhang, and Q. Huang, "Deep unsupervised convolutional domain adaptation," in *Proc. ACM MM*, Mountain View, CA, USA, 2017, pp. 261–269.

[20] J. Kim, S. Park, and N. Kwak, "Paraphrasing complex network: Network compression via factor transfer," in *Proc. NIPS*, Montréal Canada, 2018, pp. 2760–2769.

[21] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. Rizzo, and L.-Philippe Morency, "The distress analysis interview corpus of human and computer interviews," in *Proc. LREC*, Reykjavik, Iceland, 2014, pp. 3123–3128.

[22] Z. Sherhan Syed, K. Sidorov, and D. Marshall, "Depression severity prediction based on biomarkers of psychomotor retardation," in *Proc. AVEC*, Mountain View, California, USA, 2017, pp. 37–43.

[23] L. Yang, D. Jiang, X. Xia, E. Pei, M. Cédric Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proc. AVEC*, Mountain View, California, USA, 2017, pp. 53–59.

[24] L. Stappen, N. Cummins, E. Messner, H. Baumeister, J. Dineley, and B. Schuller, "Context modelling using hierarchical attention networks for sentiment and self-assessed emotion detection in spoken narratives," in *Proc. ICASSP*, Brighton, United Kingdom, 2019, pp. 6680–6684.

[25] C. Raffel, M. Luong, P. J Liu, R. J Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *Proc. ICML*, Sydney, Australia, 2017, pp. 2837–2846.

[26] C. Chiu and C. A. Raffel, "Monotonic chunkwise attention," in *Proc. ICLR*, Vancouver, BC, Canada, 2018, 16 pages.

[27] K. Kroenke, T. W Strine, R. L Spitzer, J. BW Williams, J. T Berry, and A. H Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of affective disorders*, vol. 114, no. 1, pp. 163–173, Apr. 2009.

[28] 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, 2018, pp. 27–33.

[29] K. Lenzo, "The cmu pronouncing dictionary," *URL: http://www.speech.cs.cmu.edu/cgi-bin/cmudict*, 2007.

[30] M. Luong, H. Pham, and C. D Manning, "Effective approaches to attention-based neural machine translation," in *Proc. EMNLP*, Lisbon, Portugal, 2015, pp. 1412–1421.