Speaker-invariant Psychological Stress Detection Using Attention-based Network

Hyeon-Kyeong Shin, Hyewon Han, Kyungguen Byun and Hong-Goo Kang

Department of Electrical & Electronic Engineering, Yonsei University, Seoul, South Korea
E-mail: [hkshin, hwhan, piemaker90]@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

***Abstract*—When people get stressed in nervous or unfamiliar situations, their speaking styles or acoustic characteristics change. These changes are particularly emphasized in certain regions of speech, so a model that automatically computes temporal weights for components of the speech signals that reflect stress-related information can effectively capture the psychological state of the speaker. In this paper, we propose an algorithm for psychological stress detection from speech signals using a deep spectral-temporal encoder and multi-head attention with domain adversarial training. To detect long-term variations and spectral relations in the speech under different stress conditions, we build a network by concatenating a convolutional neural network (CNN) and a recurrent neural network (RNN). Then, multi-head attention is utilized to further emphasize stress-concentrated regions. For speaker-invariant stress detection, the network is trained with adversarial multi-task learning by adding a gradient reversal layer. We show the robustness of our proposed algorithm in stress classification tasks on the Multimodal Korean stress database acquired in [1] and the authorized stress database Speech Under Simulated and Actual Stress (SUSAS) [2]. In addition, we demonstrate the effectiveness of multi-head attention and domain adversarial training with visualized analysis using the t-SNE method.**

***Index Terms*—speaker-invariant stress detection, multi-head attention, adversarial multi-task learning**

I. Introduction

Stress is defined as the psychological and/or physical tension experienced when a person has difficulty in adapting to new environments, which may cause anxiety or mental challenges [3] [4]. Stress detection is an important task because it can allow for the provision of more appropriate services to people by improving the understanding of their emotional status. In general, psychological stress can be reliably detected by measuring the changes in biological signals such as heart rate and cortisol hormone levels [5] [6]. However, in daily life, these approaches are difficult to apply because additional devices and inconvenient procedures are needed to acquire the aforementioned bio-signals. As an alternative, speech has been proposed as a means to perform psychological state detection. Notable advantages of speech include the fact that it can be easily acquired by microphones in real environments and that speech characteristics such as pitch or energy information tend to vary depending on the mental state of the speaker [7] [8] [9].

Previous studies focused on modeling relationships between stress and speech characteristics using rule-based methods or statistical models such as Hidden Markov Models (HMMs) [10] [11] [12]. Recently, stress detection methods using deep learning-based algorithms have been proposed [1] [13]. Deep learning-based approaches have shown high detection performance since they can efficiently model high-level relationships between stress labels and speech features using a data-driven approach.

Han et al. [1] proposed a deep learning-based stress detection algorithm to extract stress features from speech utterances. To train their model, they collected a database by recording speech from individuals in both neutral and stressed situations, making subjects perform simple script reading versus performing an interview in a foreign language without prior notice. Their approach used a Long Short-Term Memory (LSTM) network to model frame-wise stress-related features and used mean-pooling to obtain stress features at the utterance-level. Although this work showed that deep learning can be effectively applied to stress detection, the relative simplicity of the model implies that structural improvements can still be made. Since stress may not be revealed in every speech segment (i.e., it may be concentrated at particular parts of a word or an utterance [14]), simply applying average pooling may not be the optimal way to obtain utterance-wise representations related to stress detection.

In this paper, we propose an advanced network architecture that can detect psychological stress from speech signals. We use a spectral-temporal encoder consisting of a convolution-recurrent neural network (CRNN) for automated feature extraction. The CNN layers capture the local relations between adjacent time-frequency (T-F) bins in the acoustic features and bi-directional LSTM (BLSTM) layers find the time-varying global relationships among the CNN outputs. We apply a multi-head attention (MHA) mechanism to efficiently integrate the stress-related information of the encoded vectors extracted at each analysis frame. Using these modules, we can obtain an utterance-wise feature that effectively represents a psychological stress state. We also adopt a domain adversarial training (DAT) technique using a gradient reversal layer (GRL), which provides a constraint to the network to focus on stress detection but not have other domain capabilities such as speaker-dependent discrimination. Experimental results on the Multimodal Korean stress database (described in Section III.A) and the SUSAS-corpus with the proposed architecture showed much higher accuracy than those of previous works [1] [13]. In addition, by investigating the distribution of embeddings in a low-dimensional space using t-SNE, we verify that the

proposed method effectively captures the distinctive characteristics of stressed and non-stressed speech.

The rest of this paper is organized as follows. Section II explains the background of the proposed method. Section III introduces the two databases used in this paper by describing the acquisition process of the Multimodal Korean stress database and the SUSAS database widely used in other stress-related research. Section IV describes the proposed algorithm for stress detection specifically. The experiments and analysis results of the attention mechanism and domain adversarial learning are described in Section V. Finally, Section VI concludes the paper.

II. Background

A. Multi-head attention

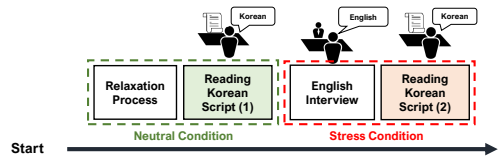
A self-attention mechanism, proposed by Lin et al. [15], finds the internal representations of encoded feature vectors that are related to the specific task of the model by setting key, query, and value, which are separately projected from the same input sequence. To further find these internal features in a more informative manner, the multi-head attention mechanism, proposed by Vaswani et al. [16], computes multiple attention weights in parallel with self-attention layers. Since each attention head captures different representations from the encoded feature vectors, this approach can capture various useful characteristics for the overall task. In our proposed method, we use multi-head attention to find informative internal representations of the stress state of the speaker.

B. Domain adversarial training

Domain adversarial training (DAT), proposed by Ganin et al. [17], is a learning strategy that aims to improve the feature representations for a given target domain by reducing undesirable variances between the source and the target domain distributions.

To perform DAT, a network uses two classifiers: a label classifier for the main task and a domain classifier that determines whether the input sample is from the source or target domain. A gradient reversal layer (GRL) is employed in the domain classifier to reverse the gradient during backpropagation in the training step. By applying multi-task learning for the two classifiers, latent representations are learned for the main task in the target domain.

Since speech includes a wide variety of information that can be applied to various tasks, we need to design an embedding network that emphasizes stress-related factors but de-emphasizes other factors such as phonetic or speaker-related ones. The database that we collected includes common phonetic information but strong speaker characteristic factors in each speech sample. To deal with this, we apply DAT by training the domain classifier on a speaker classification task, which aids in making the latent embeddings effective for stress state detection while reducing their capability for distinguishing speaker characteristics.



Stage	Relaxation Process	Reading Korean Script (1)	English Interview	Reading Korean Script (2)
Identical phonetic	-	O	X	O
Status	Non-stress	Non-stress	Stress	Stress
Survey score	1.92	2.07	4.42	2.85

Fig. 1: Experiment procedure for the Multimodal Korean stress database

TABLE I: Configuration of the Multimodal Korean stress database

State	Train	Test	Total
Non-stress	4.89 hr.	1.21 hr.	6.1 hr.
Stress	4.63 hr.	1.15 hr.	5.78 hr.
Total	9.52 hr.	2.36 hr.	11.88 hr.

III. Datasets

A. Multimodal Korean stress database

Stressed and non-stressed speech were recorded as a subset of the multi-modal stress detection database. A total of 91 native Korean speakers whose ages ranged from the 20s to the 40s participated in the recording process. The database acquisition process is described in detail in [1]; here, we explain the recording process for collecting the stressed and non-stressed speech portion of the database. Fig. 1 illustrates the recording and collection process. Each participant first went through a relaxation process by watching a calm video. Afterwards, the participants read a given script written in Korean without any external pressure. Then, an interviewer came into the recording room and asked questions in English about their personal status and daily life for 5 minutes. After the English interview, participants were made to read the same script that they read before the interview.

To evaluate the stress levels of each stage, at the end of the experiment, participants evaluated their stress scores in each stage on a scale from 1 (not stressful) to 5 (very stressful). The average stress level score of each stage is shown in Fig. 1. From the survey scores, the English interview stage was the most stressful environment and the script reading after the interview was the second most stressful. To focus solely on the stress-related aspects without phonetic variation, the speech recorded before the interview was labeled $non\text{-}stress$, and the speech recorded after the interview was labeled $stress$. The total dataset we recorded amounts to approximately 12 hours, with about 8 minutes per speaker on average. All the speech was sampled at a rate of 16 kHz. The database is divided into training and evaluation sets in a 4-to-1 ratio for each speaker when conducting the experiments. Table I demonstrates the configuration of the database.

B. SUSAS corpus

The Speech Under Simulated and Actual Stress (SUSAS) corpus [2] is a database recorded in simulated and actual stressful situations such as riding a roller coaster. It consists of 16,000 brief utterances that are 1-2 seconds long, and each sample contains a word from a male or female speaker sampled at a rate of 8 kHz. Since the SUSAS database is often used for analyzing speech under stress, it was adopted as a reference to verify the performance of our proposed model. We evaluate our approach on the same classification tasks as described in [13] to show the robustness of our proposed model. Specifically, three tasks were performed: 2-class (angry, neutral), 4-class (neutral, angry, soft, fast), and 9-class (angry, clear, fast, Lombard effect, loud, neutral, question, slow, soft) where all of the classes are spoken by 9 speakers.

IV. Proposed Algorithm

A. Feature extraction

To extract robust input features for training, several pre-processing steps are applied to the recorded audio. First, Wiener filtering is applied to remove undesired background noise components. Then, a pre-emphasis filter, i.e., a low order high pass filter, is applied to reduce the dynamic range of the frequency spectrum by emphasizing the high frequency regions. Speech segments are obtained from the pre-processed audio by applying voice activity detection (VAD) [18], after which they are converted into spectrograms by the short-time Fourier transform (STFT) at 10 ms intervals with a 25 ms Hann window. 40-dimensional log mel-spectrograms are then computed and normalized to have zero mean and unit variance to get robust features. We then segment the normalized mel-spectrograms into fixed time lengths, taking into account the different average utterance lengths of the two databases (2 seconds for SUSAS and 5 seconds for Multimodal).

B. Network architecture

Fig. 2 depicts a block diagram of the CRNN-Attention architecture used in this work. It consists of a spectral analysis module, a temporal modeling module, and an attention module with two classifiers for adversarial multi-task learning.

1) Spectral-temporal embedding: The CRNN-based embedding network encodes the input log mel-spectrogram $\mathbf{X}$ into the latent embedding $\mathbf{E}$ to make it more interpretable for stress detection. The convolutional neural network is used to effectively capture the relationships of local time-frequency (T-F) bins in the given log mel-spectrogram. For the convolution block, we adopt the VGG-A model (VGG-11) usually used for image classification [19] [20] because the model also showed high performance in speech recognition tasks [21]. To effectively capture the non-linearity features in speech signals, we use a 3x3 kernel with a very small receptive field. We reduce the number of convolutional layers in the original VGG-11 model from 11 to 7 to build a lighter model. A batch normalization layer then follows after applying a rectified linear unit (ReLU) activation function and a max-pooling layer.

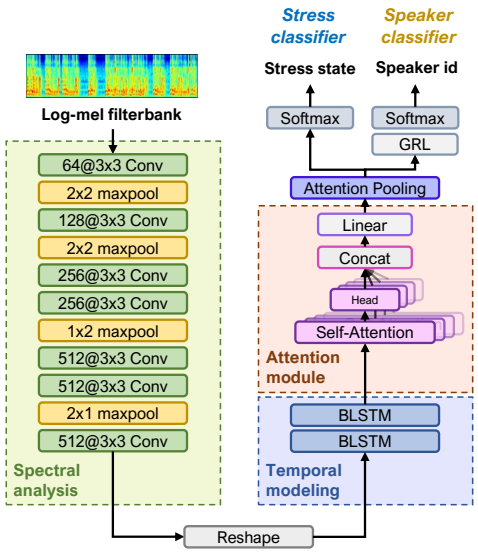


Fig. 2: The proposed multi-task network architecture consists of three components: 1) Spectral analysis; 2) Temporal modeling; 3) Attention module; with two classifiers for stress and speaker classification

The encoded feature from the CNN goes into the temporal modeling module, which consists of two bi-directional LSTM (BLSTM) layers with 256 hidden units each to model the sequential properties of the speech signal. Batch normalization is applied to improve the classification performance by accelerating learning and improving the gradient flow through the network [22]. The output of the temporal modeling module is the latent embedding $\mathbf{E}$.

2) Multi-head attention: In the attention layer, the last hidden layer of the temporal modeling module is concatenated with multiplicative multi-head self-attention weights. Although the CRNN is able to extract stress-related features, directly delivering them to the final classifier is not a good approach. This is because even in stressed speech, stress levels are not constant in every moment of the speech, but rather concentrated in certain moments. To efficiently handle this property, we concatenate a multi-head attention mechanism with the BLSTM, thus helping the network learn where to pay attention in the frame-wise output of the BLSTM.

The query, key, and value of the i-th head are denoted by Q_i, K_i, V_i $(i = 1, ..., r)$ with dimension $d_q/r, d_k/r, d_v/r$, where r denotes the number of heads. These matrices obtained by the linear projection of the embedding $\mathbf{E}$ with different weight matrices ${W_i}^Q, {W_i}^K, {W_i}^V$. Then, each attention head is computed as shown in Equation 1.

$$H_i = Softmax\left(\frac{Q_i {K_i}^T}{\sqrt{d_k/r}}\right) \cdot V_i \tag{1}$$

Finally, multi-head attention is computed by multiplying the

TABLE II: Stress recognition performance achieved from various structures of networks on the evaluation set of the Multimodal database

Model structure	Accuracy
LSTM-RNN [1]	65.76%
CRNN (5x5 Conv + LSTM)	66.55%
CRNN (3x3 Conv + LSTM)	70.73%
CRNN (3x3 Conv + BLSTM)	73.08%
CRNN (3x3 Conv + BLSTM) + Multi-Head 1 Att.	73.73%
CRNN (3x3 Conv + BLSTM) + Multi-Head 2 Att.	74.55%
CRNN (3x3 Conv + BLSTM) + Multi-Head 4 Att.	**75.08%**
CRNN (3x3 Conv + BLSTM) + Multi-Head 8 Att.	74.51%

concatenated heads with the learnable weight matrix as shown in Equation 2.

$$MHA = Concat(H_1, H_2, ..., H_n) \cdot W_o \quad (2)$$

After applying the multi-head attention, the output is connected to a binary classifier with a softmax function to determine whether the input speech is stressed or not.

C. Speaker-invariant learning strategy

Although every speech segment is normalized, latent features learned by using only binary cross-entropy loss are also significantly affected by speaker characteristics, as shown in Fig. 3 (a), (b). These characteristics can detract from the stress detection performance in the inference stage. Therefore, we put an additional speaker classifier with a gradient reversal layer to the multi-head attention output and applied adversarial multi-task training to disentangle the speaker and stress-related characteristics.

$$L = (1 - \lambda)L_{stress} - \lambda L_{speaker} \quad (3)$$

As shown in Equation 3, the loss function is defined as the weighted sum of binary cross-entropy loss L_{stress} for the stress classifier and categorical cross-entropy loss $L_{speaker}$ for the speaker classifier with a negative coefficient for gradient reversal. The relative weights of the stress classification and speaker classification losses are controlled by the weight λ.

V. Experiments

To show the effectiveness of the proposed method, we conducted stress detection and classification experiments on the Multimodal database and SUSAS database. We set 20% of the total data for evaluation and the rest of the samples for training. From the training data, 20% was set aside to be used as a validation set. We investigated the performance of various network setups to determine the best model in terms of accuracy on the Multimodal database. We also evaluated the performance of the domain adversarial learning by adjusting the weights between the two losses and applying dropout to find the optimal controlling parameter. We used He initialization [23] and Adam as the optimizer with a learning rate of 10^{-4}.

TABLE III: Stress recognition performance achieved from various loss weights of DAT on the evaluation set

λ	Dropout prob.	Multimodal DB	SUSAS-corpus		
			2-class	4-class	9-class
-	-	75.08%	96.82%	82.53%	72.83%
0.05	-	76.44%	97.61%	89.48%	**93.26%**
0.01	-	76.55%	96.82%	91.07%	93.21%
0.005	-	75.55%	96.42%	91.51%	92.47%
0.003	-	76.48%	96.42%	90.27%	92.70%
0.001	-	75.76%	94.84%	**92.41%**	93.08%
0.05	0.2	76.12%	96.82%	90.52%	92.92%
0.01	0.2	76.37%	95.63%	90.87%	93.02%
0.005	0.2	75.66%	96.82%	90.87%	93.04%
0.003	0.2	**77.30%**	**98.01%**	90.82%	92.89%
0.001	0.2	76.91%	96.03%	90.17%	92.96%

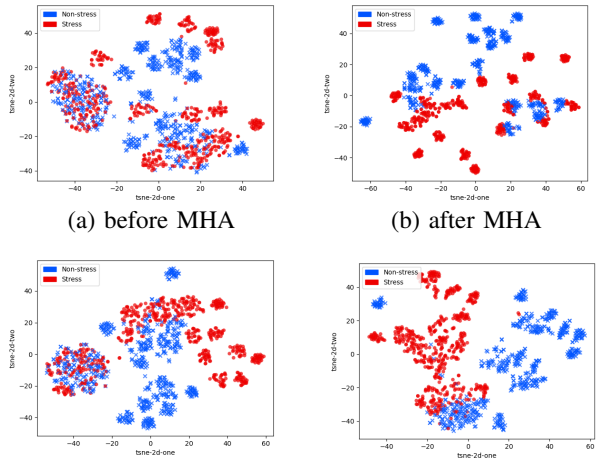
A. Comparison of network architectures

To find the most effective CRNN architecture, we investigated the impact of various factors such as the kernel size of the convolutional layer (5x5 vs. 3x3) and the recurrent layer structure (LSTM vs. BLSTM) at the spectral-temporal embedding stage. We also evaluated the performance of the multi-head attention mechanism with varying numbers of heads. Table II summarizes the stress detection performance depending on the network structure. All implemented CRNN structures achieved higher accuracies than the conventional two-layers LSTM-RNN network [1]. The model with 3x3 kernel size and BLSTM exhibited better performance than ones using a 5x5 kernel or LSTM. For the multi-head attention module, using 4 heads resulted in the best performance, achieving 75.08% stress recognition accuracy. All the rest of the experiments in the paper are based on this architecture.

B. Comparison of various settings of DAT

We explored the appropriate controlling loss weight λ of adversarial multi-task training for the proposed model on the Multimodal database and the SUSAS database. The results are shown in Table III. When the model was trained on the Multimodal database, DAT led to a relative improvement of 0.63-2.95% over a proposed network without it. The model achieved the best performance of 77.30% in the stress recognition task when λ value was set to 0.003 with a dropout probability of 0.2.

We also evaluated our proposed model on the SUSAS database. Here, the network trained with only stress classification loss achieved a recognition performance of 96.82% (2-class), 82.53% (4-class), 72.83% (9-class) for each task on evaluation set. It is clear from the results that it outperforms the previous research [13], which obtained results of 76% (2-class), 71% (4-class), and 70% (9-class). Using DAT resulted in significant further improvements on every task resulting in accuracies of 98.01% (2-class), 92.41% (4-class), and 93.26% (9-class).



(a) before MHA (b) after MHA

(c) before MHA (with DAT) (d) after MHA (with DAT)

Fig. 3: t-SNE visualizations of high-dimensional non-stress embedding (blue dot) and stress embedding (red dot) from the proposed model on randomly chosen 20 speakers from the evaluation set in the Multimodal database; i) trained without DAT: (a), (b); ii) trained with DAT ($\lambda = 0.003$) and dropout: (c), (d)

C. Analysis of stress embeddings

To evaluate the representations learned by the proposed algorithm, we analyzed the embedding vectors obtained by our network. Figure 3 shows t-SNE visualizations of the high-dimensional stress embeddings extracted from the proposed architecture. (a) and (b) show stress embeddings before and after the attention layer, respectively. (c) and (d) are stress embeddings before and after the attention layer when speaker-domain adversarial training is applied. By comparing (a) with (b) and (c) with (d), we can observe that multi-head attention separates the embedding vectors more distinctly.

We also checked the effect that DAT had on reducing the inclusion of speaker information in the embeddings by comparing (b) and (d). Without DAT (Figure 3 (b)), the extracted embeddings appear to display speaker information, forming many small clusters within each class. However, when DAT is applied, the class dots are scattered widely and less likely to express speaker information.

To further verify the effect of DAT, we measured the cosine distance between non-stress embeddings and stress embeddings. The cosine distance d_{cos} in an n-dimensional space can be obtained via the following equation:

$$d_{cos} = 1 - \frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}||||\mathbf{v}||} \tag{4}$$

where $\mathbf{u}$ and $\mathbf{v}$ denote embeddings from non-stressed and stressed speech, respectively. To obtain the overall distance between two classes, we calculated and averaged the distance of all possible pairs between the two classes. We obtained cosine distances of 1.006 for (b) and 1.319 for (d). From this result, we can infer that the model can extract more distinctive stress-related features when DAT is applied.

VI. Conclusion

In this work, we proposed a deep learning approach to detect and classify stressful conditions from speech, using the Multimodal Korean stress database and the SUSAS database. The proposed network consists of CRNN and a multi-head attention mechanism to find and utilize stress-related information from each analysis frame. Our model successfully outperformed the baseline LSTM-based model due to more efficient modeling of spectral and temporal information from the acoustic features and effectively captured information related to a speaker's psychological stress state. In addition, by applying domain adversarial training with a gradient reversal layer, we were able to enhance the stress recognition performance while reducing factors related to other speech characteristics.

Acknowledgment

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [2016-0-00562 (R0124-16-0002), Emotional Intelligence Technology to Infer Human Emotion and Carry on Dialogue Accordingly]

References

[1] H. Han, K. Byun, and H.-G. Kang, "A deep learning-based stress detection algorithm with speech signal," in *Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia*. ACM, 2018, pp. 11–15.

[2] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting started with susas: a speech under simulated and actual stress database," in *EUROSPEECH*, 1997.

[3] T. Johnstone, "The effect of emotion on voice production and speech acoustics," 2017.

[4] K. Tomba, J. Dumoulin, E. Mugellini, O. A. Khaled, and S. Hawila, "Stress detection through speech analysis," in *ICETE (1)*, 2018, pp. 560–564.

[5] H. Kurniawan, A. V. Maslov, and M. Pechenizkiy, "Stress detection from speech and galvanic skin response signals," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. IEEE, 2013, pp. 209–214.

[6] M. Salai, I. Vassányi, and I. Kósa, "Stress detection using low cost heart rate sensors," *Journal of Healthcare Engineering*, vol. 2016, 2016.

[7] D. A. Cairns and J. H. Hansen, "Nonlinear analysis and classification of speech under stressed conditions," *The Journal of the Acoustical Society of America*, vol. 96, no. 6, pp. 3392–3400, 1994.

[8] L. J. Rothkrantz, P. Wiggers, J.-W. A. Van Wees, and R. J. van Vark, "Voice stress analysis," in *International conference on text, speech and dialogue*. Springer, 2004, pp. 449–456.

[9] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman, "Stress and emotion classification using jitter and shimmer features," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–1081.

[10] B. D. Womack and J. H. Hansen, "N-channel hidden markov models for combined stressed speech classification and recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 668–677, 1999.

[11] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden markov models," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[12] L. He, M. Lech, S. Memon, and N. Allen, "Recognition of stress in speech using wavelet analysis and teager energy operator," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[13] A. R. Avila, S. R. Kshirsagar, A. Tiwari, D. Lafond, D. O'Shaughnessy, and T. H. Falk, "Speech-based stress classification based on modulation spectral features and convolutional neural networks," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[14] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[15] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

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

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[18] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1181–1185, Aug 2018.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] W. Hartmann, R. Hsiao, T. Ng, J. Z. Ma, F. Keith, and M.-H. Siu, "Improved single system conversational telephone speech recognition with vgg bottleneck features." in *INTERSPEECH*, 2017, pp. 112–116.

[21] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun, "Very deep multilingual convolutional neural networks for lvcsr," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4955–4959.

[22] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.