Attentions for Short Duration Speech Classification

Hastin Modi[◇*], Maitreya Patel[◇], Hemant A. Patil
Dhirubhai Ambani Institute of Information and Communication Technology

***Abstract*—Neural attention mechanisms have gained significant popularity and widespread adoption in various applications. These attention mechanisms can be applied to different models, including sequential and spatial models. Recently, the Transformer model, based on multi-head self-attention, and the ResNeSt model, utilizing split-attention, have been developed for sequential and spatial tasks, respectively. However, many attempts to leverage state-of-the-art attention methodologies for speech-related classification problems have been made without thorough analysis. In this paper, we conduct an extensive analysis of various attention-based methods by performing experiments on infant cry classification and speech emotion recognition tasks. Additionally, we evaluate the proposed models on different durations of audio clips from the Baby Chillanto and CREMA-D databases. Our results demonstrate that the Transformer (encoder-only) model significantly outperforms ResNeSt. However, for longer audio clips, ResNeSt exhibits greater robustness compared to the Transformer. Furthermore, both Transformer and ResNeSt surpass the previous state-of-the-art in infant cry classification, achieving recall improvements of 10.9% and 4.3%, respectively.**

***Index Terms*—Attentions, Transformer, ResNeSt, Infant Cry Classification, Emotion Recognition.**

I. Introduction

Recent advancements in deep learning have led to the proposal of various architectures for different domains, including speech technology. These applications encompass a wide range of areas, such as Speech Enhancement [1]–[5], Voice Conversion [6]–[8], and Dysarthria speech severity classification [9], [10]. Convolutional Neural Networks (CNNs) and Long-Short Term Memory (LSTM) brought many advancements in spatial and sequential problems, respectively. However, these models often suffer from limitations and are highly reliant on the available training data, leading to a lack of robustness. To address these limitations and enhance the efficiency and robustness of various tasks in image, text, and speech domains, neural attention mechanisms were introduced [11]–[14]. Attention mechanisms enable models to focus on crucial parts of the input, improving their overall performance. Different types of attention mechanisms have been proposed, including content-based (e.g., cosine), additive, location-based, general, dot-product, and scaled dot-product. Attention can also be categorized into self-attentions, global (soft) attentions, and local (hard) attentions [15]–[20].

One groundbreaking model that has recently emerged for pre-training language models is the Transformer [19]. The Transformer leverages multi-head self-attention to process various parts of the input simultaneously. For example, Dong *et al.* [21] proposed the Speech-Transformer for speech recognition, and Tarantino *et al.* [22] utilized Transformer-based models for speech emotion classification. In addition to sequential models, attention mechanisms have also been applied to CNNs to extract key features from input data. In particular, Zhang *et al.* [23] introduced attention mechanisms on top of CNN for speech emotion classification. However, CNN-based attention approaches have limitations regarding the number of times attention can be applied. To address this, we rely on a more reliable and sophisticated architecture called ResNeSt [24], which utilizes split-attentions at each residual block to enhance the efficiency of computer vision tasks. Furthermore, our analysis reveals that attention-based models outperform regular models.

In this paper, we investigate two major attention-based architectures, namely, the encoder-style Transformer [19] and ResNeSt [24], for short-duration speech classification tasks. We conduct a comprehensive analysis and comparison of their robustness by considering two well-known problems: infant cry classification and speech emotion recognition. We note that the state-of-the-art method for infant cry classification, namely speech commands transfer (sc-transfer) [25], does not leverage attention mechanisms and thus exhibits limited performance. Moreover, the literature on emotion recognition lacks detailed experiments on attention mechanisms [21], [22]. Our findings indicate that the Transformer model outperforms ResNeSt for short-duration audio clips in terms of overall performance. However, for longer audio clips, ResNeSt demonstrates greater robustness and generalizability compared to the Transformer. The main contributions of this work can be summarized as follows:

- We provide the first analysis of the effectiveness and robustness of various attention-based models for short-duration speech classification.
- Our proposed attention-based methods achieve state-of-the-art results for infant cry classification.

The remainder of the paper is organized as follows: Both methodologies are explained in Section 2. Section 3 contains the experimental setup to make the results reproducible. While Section 4 presents all the experiments and results of this study followed by summary and conclusions in Section 5.

II. Approach

In this Section, we explain both the attention-based models, i.e., Transformer and ResNeSt. Furthermore, we explain the details of different attention mechanisms used in both methodologies.

[◇] This work was done while Hastin and Maitreya were at DA-IICT. Currently, they are at Arizona State University.
[*] Reach out at: hmodi5@asu.edu

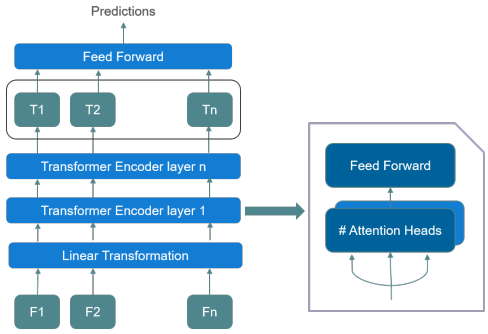

Fig. 1. Architecture design of encoder-only Transformer model. After [19].

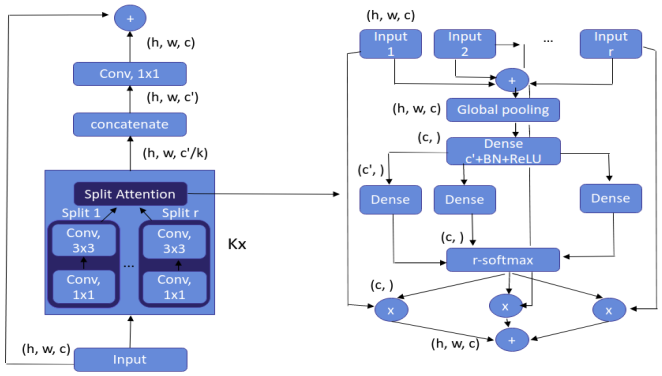

Fig. 2. ResNeSt block and split-attention. After [24].

A. Encoder-only Transformer

Transformer was proposed by Vaswani *et. al.* [19] for various natural language processing (NLP) tasks. There are two main structures in the Transformer architecture, namely, an encoder and a decoder. Within these structures, there are various layers through which the input passes, such as Multi-Head Attention, Normalization, and a Feedforward neural network. Furthermore, Devlin *et al.* [26] introduced Bidirectional Encoder Representations from Transformer (BERT), which has an architecture that uses only the encoder structure of the Transformer to achieve state-of-the-art results on various NLP tasks. This shows that encoders are sufficient to learn the representations and interactions between various input components of the sequential data. Therefore, inspired by BERT, we apply a similar architecture for speech data classification. Fig. 1 shows the architecture used in our study. Before inserting the input into the encoder, a positional encoding is added to the input since the Transformer does not contain any recurrence or convolution and cannot determine the order of the input sequence. We have also inserted a linear layer before passing the input to the encoder, as suggested in Pham *et al.* [27], and our experiments demonstrate that incorporating a linear layer yields superior performance compared to omitting it. From Fig. 1, we can observe that the model is composed of N number of stacked layers.

1) Multi-Head Attention (Self-Attention): The encoder's multi-head attention-based sub-layers perform the scaled dot-product attention operation in parallel a fixed number of times, governed by a hyperparameter. Suppose each sub-layer applies m distinct self-attentions, the model will then have a total of $N \times m$ attentions.

Scaled dot-product attention takes query (q), key (k), and value (v) as inputs. Here, q, k, and v are extracted from the output of previous layer using neural networks. When scaling is applied across all the feature vectors of the given inputs, each attention's result can be represented as follows [19]:

$$Attention(Q, K, V) = softmax\Big(\frac{QK^T}{\sqrt{d_k}}\Big)V, \tag{1}$$

where Q, K, and V are the matrix of queries, keys, and values extracted from the entire input speech signal. Furthermore, for m number of heads of a single layer, Multi-Head attention can be represented as [19]:

$$MultiHead(Q, K, V) = Concate(head_1, ..., head_m)W^0, \tag{2}$$

where $head_i = Attention(QW_i^Q, K_i^K, V_i^V)$.

B. ResNeSt

ResNeSt, proposed by Zhang *et al.* [24] for various computer vision tasks, introduces a modification to the ResNet architecture by incorporating feature map split attention within individual network blocks. In each block, the feature map is divided into groups along the channel dimension, creating finer-grained sub-groups. This unit is referred to as a split-attention block, and ResNeSt consists of multiple such blocks. Fig. 2 provides an illustration of a ResNeSt block in the cardinality-major view, showcasing the split-attention mechanism.

1) Split-attention Block: The split-attention block consists of two operations:

Feature map Group - The features are divided into multiple groups, where the number of groups, denoted as *cardinal* groups, is determined by a hyperparameter K. Within each cardinal group, the feature map is further split into sub-groups, with the number of splits given by another hyperparameter R. Consequently, the total number of feature groups is $G = KR$.

Split attention in Cardinal Groups - Fig. 2 shows split-attention within a cardinal group, where $c = C/K$. Fusing via elementwise summation across splits is used to obtain a combined representation for each cardinal group. The k^{th} cardinal group's representation is given by $\hat{U}^k = \sum_{j=R(k-1)+1}^{Rk} U_j$, where $\hat{U}^k \in \mathbb{R}^{H \times W \times C/K}$ for $k \in 1, 2, ..., K$.

H, W, and C are block output feature map sizes. To capture global contextual information, global average pooling is applied across the spatial dimensions, which is given by $s^k \in \mathbb{R}^{C/K}$. The c^{th} component is calculated as [24]:

$$s_c^k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \hat{U}_c^k(i, j). \tag{3}$$

Channelwise soft attention is used for aggregating weighted fusion of the cardinal group representation, $V^k \in \mathbb{R}^{H \times W \times C/K}$.

The feature map channels are produced using a weighted combination over splits. The c^{th} channel being calculated as [24]:

$$V_c^k = \sum_{i=1}^{R} a_i^k(c) U_{R(k-1)+i}, \tag{4}$$

where $a_i^k(c)$ denotes a (soft) assignment weight given by [24]:

$$a_i^k(c) = \begin{cases} \frac{\exp\left(\mathcal{G}_i^c(s^k)\right)}{\sum_{j=0}^{R}\exp\left(\mathcal{G}_j^c(s^k)\right)} & \text{if } R > 1, \\ \frac{1}{1+\exp\left(-\mathcal{G}_i^c(s^k)\right)} & \text{if } R = 1. \end{cases} \tag{5}$$

The mapping $\mathcal{G}_i^c$ determines the weight of each split for the c^{th} channel based on s_k (the global context representation). Finally, the cardinal group representations are concatenated along the channel dimension. The final output Y of the split-attention block is produced using a shortcut connection similar to the one used in the standard residual blocks. There are two cases to be considered when producing the final output, the first being when the input and output feature maps have the same shape, in such a case, the output is $Y = V + X$. In the second case, if the blocks have a stride, the appropriate transformation $\mathcal{T}$ is applied to align the output shapes; in such a case, the output is $Y = V + \mathcal{T}(X)$.

III. Experimental Setup

A. Database and Feature Extraction

We selected two diverse datasets to evaluate the performance of the Transformer and ResNeSt models across different tasks. Firstly, we used the most commonly referred Baby Chillanto database for infant cry classification [28]. Secondly, CREMA-D [29] is used for evaluating both the attention-based models on speech emotion recognition for various duration of audio clips. Baby Chillanto database contains overall 5 different classes of infant cries (i.e., asphyxia, deaf, hunger, normal, and pain). Here, we perform asphyxia *vs.* non-asphyxia classification to investigate the performance of these architectures for binary classification on speech data. We have a total of 340 and 1049 audio samples for each class, respectively. Here, we have only about 23 minutes of data. Furthermore, we choose the CREMA-D dataset, an open-source dataset containing audio samples from 91 different actors (48 males and 43 females). This dataset contains 6 different emotions, namely, anger, disgust, fear, happy, neutral, and sad. Out of these, we use all the classes except neutral for the high-intensity level of emotion. Overall, we have taken about 14.5 minutes of data.

For feature extraction, we computed the *36*-dimensional Mel Frequency Cepstral Coefficients (MFCCs) from the audio clips, including the 0^{th} cepstral coefficient. A window size of 128 ms and a frame shift of 32 ms were used for extracting MFCCs. All audio clips were sampled at a rate of 16 kHz. For the Baby Chillanto database, we used the entire duration of the audio clips, i.e., 1 second. For the CREMA-D database, we considered three different durations: 1, 3, and 5 seconds, corresponding to 32, 94, and 157 feature frames, respectively.

B. Architectural Details

In this subsection, we provide detailed descriptions of the Transformer and ResNeSt architectures used in our experiments, along with the training procedure. Moreover, we have made the codebase publicly available on GitHub [*] to ensure the reproducibility of our experiments.

1) Transformers: Through empirical experiments, we determined that an architecture comprising 6 layers, with 6 attention heads in each sub-layer of the encoder, achieved optimal performance for infant cry classification. Therefore, our architecture consists of a total of 36 (6x6) attentions. Each fully-connected layer contains 1023 neurons with Rectified Linear Unit (ReLU) activation [30]. For the CREMA-D database, when using 1-second audio clips, the best performance was achieved with 6 layers and 6 attention heads in each sub-layer of the encoder. Each fully-connected layer contains 66 neurons. For 3-second audio clips, the optimal configuration included 2 layers with 3 attention heads in each sub-layer, with each fully-connected layer containing 258 neurons and ReLU activation. Similarly, for 5-second audio clips, the best performance was achieved with 2 layers and 12 attention heads in each sub-layer of the encoder, with each fully-connected layer containing 258 neurons and ReLU activation. A single output layer was employed to generate predictions for each input, depending on the specific task. We utilized early stopping criteria with a patience of 5 epochs and a maximum of 500 epochs. The Adam optimizer [31] was used with a learning rate of 0.00005 for both databases.

2) ResNeSt: For our ResNeSt experiments, we employed the ResNeSt-50 model as proposed in the original paper. We used the radix-major implementation, which leverages standard CNN operators and group convolution, resulting in faster computation compared to the cardinality-major implementation. In our experiments, we set the radix value to 2. For both the Baby Chillanto and CREMA-D datasets, we found that a learning rate of 0.001 and a dropout rate of 0.2 yielded the best results across all durations of audio clips. We utilized the Adam optimizer to optimize the model parameters. Early stopping criteria with a patience of 5 epochs and a maximum of 500 epochs were applied. In the case of infant cry classification, the fully-connected layer at the end of the ResNeSt model employed the sigmoid activation function, suitable for the binary classification task. On the other hand, for the multi-class classification task in the CREMA-D database, we utilized the softmax activation function.

IV. Experimental Results

To evaluate the performance and effectiveness of both models, we employed standard performance metrics such as Area Under Curve (AUC), F1 score, Recall, and Accuracy [25]. Additionally, we conducted experiments using various durations of audio clips to understand the models' behavior under different conditions. In order to simulate real-life scenarios and assess the impact of different attention mechanisms,

[*] https://github.com/hastinmodi/Attentions_for_speech_classification/

TABLE I
PERFORMANCE (MEAN AND STANDARD DEVIATION) OF DIFFERENT MODELS ON INFANT CRY CLASSIFICATION TASK AFTER REMOVING INPUT FRAMES ON TEST SET

Method	% Frames Removed	AUC	Recall	F1	Accuracy
ResNet - without attention	0%	0.9765	0.8676	0.8676	93.5%
Transformer	0%	**0.9918**	**0.9706**	**0.9296**	**96.4%**
	15%	0.9896 (+/- 0.001)	0.9294 (+/- 0.012)	0.9041 (+/- 0.0084)	95.2% (+/- 0.4%)
	20%	0.9892 (+/- 0.001)	0.9882 (+/- 0.007)	0.9143 (+/- 0.0075)	95.4% (+/- 0.4%)
ResNeSt	0%	**0.9821**	**0.8971**	**0.9037**	**95.32%**
	15%	0.981 (+/- 0.002)	0.8676 (+/- 0.01)	0.8846 (+/- 0.012)	94.5% (+/- 0.65%)
	20%	0.987 (+/- 0.001)	0.8736 (+/- 0.04)	0.885 (+/- 0.027)	94.5% (+/- 1.2%)
Method	**Length of Audio**	**AUC**	**Recall**	**F1**	**Accuracy**
Transformer	0.5 sec	0.9847	0.9559	0.8966	94.6%
	0.6 sec	0.0597	0.7647	0.3161	19.0%
	0.7 sec	0.9905	0.9853	0.9116	95.3%
	0.8 sec	0.8164	1	0.4172	31.7%
	0.9 sec	0.9914	0.9706	0.9231	96.0%
ResNeSt	0.5 sec	0.9919	0.8971	0.9037	95.3%
	0.6 sec	0.9920	0.8971	0.8971	95.0%
	0.7 sec	0.9929	0.8824	0.9023	95.3%
	0.8 sec	0.9909	0.8971	0.8971	95.0%
	0.9 sec	0.9907	0.9265	0.9131	95.7%

we randomly removed 15-20% of frames during the test phase. For infant cry classification, we also examined audio clips ranging from 0.5 to 0.9 seconds in length. In this Section, we first analyze both approaches for infant cry classification, and later, we conduct evaluations on emotion recognition.

A. Infant Cry Classification

Table I presents the performance comparison of the two attention-based methodologies. The results indicate that both models perform similarly in terms of accuracy, AUC, and F1 score when no frames are removed. However, when considering recall, which is particularly crucial in healthcare applications, the Transformer model exhibits superior performance over ResNeSt. To evaluate the models' robustness, we conducted experiments by randomly removing 15% and 20% of input frames, as well as decreasing the audio clip length, while evaluating the pre-trained models. Each experiment was repeated five times for more accurate measurements of robustness. From the results in Table I, we observe that the Transformer model displays greater robustness than ResNeSt when frames are randomly removed. However, when decreasing the audio length, ResNeSt appears to be more resilient than the Transformer. Moreover, both the Transformer and ResNeSt models outperform the baseline approach (sc-transfer) by 10.9% and 4.3%, respectively. Notably, when comparing the attention-based models with the non-attention-based ResNet model, we observe that the attention-based methods yield superior performance.

B. Speech Emotion Recognition

We adopted a similar evaluation strategy to assess the robustness of both architectures. Additionally, we conducted training and evaluations on audio clips of different durations, namely 1, 3, and 5 seconds. Similar to infant cry classification, we randomly removed 15% and 20% of frames during evaluation. The results presented in Table II and Table III demonstrate that the Transformer model outperforms ResNeSt when no feature frames are removed. However, when dealing with longer audio clips, ResNeSt exhibits greater robustness

TABLE II
PERFORMANCE (MEAN AND STANDARD DEVIATION) OF TRANSFORMER ON EMOTION RECOGNITION TASK AFTER REMOVING 0%, 15%, AND 20% OF INPUT FRAMES ON THE TEST SET

Duration	% Frames Removed	AUC	F1	Accuracy
1 second	0%	**0.8247**	**0.4654**	**52.2%**
	15%	0.7982 (+/- 0.011)	0.4174 (+/- 0.008)	48.89% (+/- 0.78%)
	20%	0.7948 (+/- 0.012)	0.2319 (+/- 0.022)	38.45% (+/- 1.7%)
3 seconds	0%	**0.8444**	**0.5961**	**61%**
	15%	0.8195 (+/- 0.023)	0.4723 (+/- 0.022)	50.4% (+/- 2.3%)
	20%	0.7932 (+/- 0.025)	0.4146 (+/- 0.026)	46.9% (+/- 2.6%)
5 seconds	0%	**0.8569**	**0.5862**	**56.7%**
	15%	0.3458 (+/- 0.028)	0.039 (+/- 0.008)	6.2% (+/- 1.3%)
	20%	0.348 (+/- 0.022)	0.041 (+/- 0.017)	6.9% (+/- 2.2%)

TABLE III
PERFORMANCE (MEAN AND STANDARD DEVIATION) OF RESNEST ON EMOTION RECOGNITION TASK AFTER REMOVING 0%, 15%, AND 20% OF INPUT FRAMES ON TEST SET

Duration	% Frames Removed	AUC	F1	Accuracy
1 second	0%	0.7346	0.4217	43.3%
	15%	0.734 (+/- 0.014)	0.4133 (+/- 0.016)	43.11% (+/- 1.45%)
	20%	**0.7524** (+/- 0.020)	**0.4348** (+/- 0.018)	**45.11%** (+/- 2.4%)
3 seconds	0%	**0.8301**	**0.4622**	**50%**
	15%	0.7971 (+/- 0.016)	0.4358 (+/- 0.039)	47.22% (+/- 3.2%)
	20%	0.7741 (+/- 0.016)	0.4008 (+/- 0.033)	44.2% (+/- 2.4%)
5 seconds	0%	**0.8443**	**0.4739**	**51.1%**
	15%	0.8115 (+/- 0.009)	0.4471 (+/- 0.031)	49.78% (+/- 2.9%)
	20%	0.7596 (+/- 0.011)	0.4323 (+/- 0.049)	47.78% (+/- 4.1%)

compared to the Transformer. Interestingly, when evaluating both approaches on the full duration of audio clips (i.e., 5 seconds), we observe that the Transformer performs poorly when feature frames are randomly removed.

V. SUMMARY AND CONCLUSIONS

This paper introduced two state-of-the-art attention-based methodologies, namely Transformer and ResNeSt, which address classification problems from the perspectives of sequential and spatial processing, respectively. Specifically, we conducted experiments on infant cry and speech emotion classification tasks to evaluate the effectiveness and robustness of both architectures. Moreover, we examined the performance of these models across different audio durations. Through a series of comprehensive experiments, including the evaluation of robustness by randomly removing frames, we made several key observations.

Overall, the results demonstrate that the Transformer model outperforms ResNeSt in terms of all performance metrics considered. Particularly, the Transformer exhibits superior performance in accuracy, AUC, F1 score, and recall, making it a compelling choice for both infant cry and speech emotion classification tasks. However, when examining robustness,

ResNeSt emerges as the more resilient model, especially for longer-duration audio clips. This finding highlights the importance of considering both performance and robustness aspects when selecting an appropriate model for specific applications.

As we look ahead, our future work will focus on further enhancing the performance of these models. This may involve exploring novel architectural modifications, leveraging additional contextual information, or incorporating domain-specific knowledge. By continually refining and advancing these attention-based methodologies, we aim to improve their applicability and impact in various real-world scenarios.

REFERENCES

[1] H. Malaviya, J. Shah, M. Patel, J. Munshi, and H. A. Patil, "Mspec-net: Multi-domain speech conversion network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7764–7768.

[2] M. Patel, M. Parmar, S. Doshi, N. Shah, and H. Patil, "Novel inception-gan for whispered-to-normal speech conversion," in *Proc. 10th ISCA Speech Synthesis Workshop*, pp. 87–92.

[3] M. Parmar, S. Doshi, N. J. Shah, M. Patel, and H. A. Patil, "Effectiveness of cross-domain architectures for whisper-to-normal speech conversion," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[4] M. Purohit, M. Patel, H. Malaviya, A. Patil, M. Parmar, N. Shah, S. Doshi, and H. A. Patil, "Intelligibility improvement of dysarthric speech using mmse discogan," in *2020 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2020, pp. 1–5.

[5] M. Patel, M. Purohit, J. Shah, and H. A. Patil, "Cinc-gan for effective f 0 prediction for whisper-to-normal speech conversion," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 411–415.

[6] M. Patel, M. Parmar, S. Doshi, N. J. Shah, and H. A. Patil, "Novel adaptive generative adversarial network for voice conversion," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1273–1281.

[7] M. Patel, M. Purohit, M. Parmar, N. J. Shah, and H. A. Patil, "Adagan: Adaptive gan for many-to-many non-parallel voice conversion," 2019.

[8] D. G. Rajpura, J. Shah, M. Patel, H. Malaviya, K. Phatnani, and H. A. Patil, "Effectiveness of transfer learning on singing voice conversion in the presence of background music," in *2020 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2020, pp. 1–5.

[9] S. Gupta, A. T. Patil, M. Purohit, M. Parmar, M. Patel, H. A. Patil, and R. C. Guido, "Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments," *Neural Networks*, vol. 139, pp. 105–117, 2021.

[10] M. Purohit, M. Parmar, M. Patel, H. Malaviya, and H. A. Patii, "Weak speech supervision: A case study of dysarthria severity classification," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 101–105.

[11] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2018, pp. 1564–1574.

[12] S. Chaudhari, G. Polatkan, R. Ramanath, and V. Mithal, "An attentive survey of attention models," *Transactions on Intelligent Systems and Technology (TIST)*, {Last Accessed on: Oct, 2021}.

[13] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2015, pp. 577–585.

[14] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, United States, 2017, pp. 3156–3164.

[15] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, {Last Accessed on: Oct, 2021}.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, {Last Accessed on: Oct, 2021}.

[17] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015*, {Last Accessed on: Oct, 2021}.

[18] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," *arXiv preprint arXiv:1601.06733*, {Last Accessed on: Oct, 2021}.

[19] 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 (NIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning (ICML)*, Lille, France, 2015, pp. 2048–2057.

[21] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888.

[22] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *INTERSPEECH*, Graz, Austria, 2019, pp. 2578–2582.

[23] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Hawaii, USA, 2018, pp. 1771–1775.

[24] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[25] C. C. Onu, J. Lebensold, W. L. Hamilton, and D. Precup, "Neural transfer learning for cry-based diagnosis of perinatal asphyxia," *arXiv preprint arXiv:1906.10199*, {Last Accessed on: Oct, 2021}.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, {Last Accessed on: Oct, 2021}.

[27] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Müller, S. Stüker, and A. Waibel, "Very deep self-attention networks for end-to-end speech recognition," *arXiv preprint arXiv:1904.13377*, {Last Accessed on: Oct, 2021}.

[28] C. Ji, S. Basodi, X. Xiao, and Y. Pan, "Infant sound classification on multi-stage CNNs with hybrid features and prior knowledge," in *International Conference on AI and Mobile Services*. Springer, 2020, pp. 3–16.

[29] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[30] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, {Last Accessed on: Oct, 2021}.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, {Last Accessed on: Oct, 2021}.