Interspeech 2024
1-5 September 2024, Kos, Greece

Towards Self-Attention Understanding for Automatic Articulatory Processes Analysis in Cleft Lip and Palate Speech

Ilja Baumann[1], *Dominik Wagner*[1], *Maria Schuster*[2], *Korbinian Riedhammer*[1], *Elmar Nöth*[3], *Tobias Bocklet*[1,4]

[1]Technische Hochschule Nürnberg, Germany [2]Ludwig-Maximilians University, Germany
[3]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany [4]Intel Labs, Germany

[1]firstname.lastname@th-nuernberg.de

Abstract

Cleft lip and palate (CLP) speech presents unique challenges for automatic phoneme analysis due to its distinct acoustic characteristics and articulatory anomalies. We perform phoneme analysis in CLP speech using a pre-trained wav2vec 2.0 model with a multi-head self-attention classification module to capture long-range dependencies within the speech signal, thereby enabling better contextual understanding of phoneme sequences. We demonstrate the effectiveness of our approach in the classification of various articulatory processes in CLP speech. Furthermore, we investigate the interpretability of self-attention to gain insights into the model's understanding of CLP speech characteristics. Our findings highlight the potential of the self-attention mechanisms for improving automatic phoneme analysis in CLP speech, paving the way for enhanced diagnostics, adding interpretability for therapists and affected patients.

Index Terms: pathologic speech, cleft lip and palate, children's speech, automatic assessment

1. Introduction

Cleft lip and palate (CLP) are congenital conditions arising from an incomplete development of the upper lip and/or roof of the mouth (palate) during fetal development, resulting in a visible split or gap. These can occur independently or simultaneously, unilaterally or bilaterally, and may also involve the jaw. Surgical repair is typically undertaken before the child reaches their first years of age. Beyond challenges in feeding and dental health, CLP significantly impacts speech production, with variations in phonemic alterations among affected individuals. Detailed phonemic analysis facilitates targeted postoperative speech therapies tailored to each patient's specific deficits. While perceptual analyses by experts are standard, they are subjective and time-intensive, taking about 3 hours per child for phoneme-level annotation [1] and thus cannot be performed regularly during the speech therapy.

Our aim is to create an automated system to assess various articulatory processes and examine which audio input features contribute to the final decision, ensuring alignment with expert evaluations by evaluating self-attention of the classification module. Our automated system labels are derived from scores representing distinct articulation processes for each phoneme in an utterance, as perceptually estimated by clinical speech therapists. The selected articulatory processes are usually performed by experts who evaluate CLP speech [2]. We employ a pre-trained wav2vec 2.0 model with a classification module to classify articulatory processes, treating it as a multi-label and multi-class problem based on expert ratings. Although we have phoneme-level annotations available, our system is intentionally trained using only utterance-level labels. We then conduct a detailed analysis at the phoneme level to analyze the degree of alignment between the decisions made by the trained system and the expert labels on phoneme level. Since several studies debate whether attention can lead to explanation of a model or not [3, 4, 5], we conduct a comprehensive analysis of the attention by employing a hierarchical attention relevance computation methodology and employing self-attention head pruning.

1.1. Related work

In the study by Maier et al. [6], a technique was introduced to automatically detect articulation disorders in a cohort of 58 German children with CLP. The approach involved the utilization of multiple classifiers through late fusion and an Automatic Speech Recognition (ASR) system to extract features at both phoneme and word levels while using metrics such as Goodness of Pronunciation (GoP) [7], phoneme posteriors, and word confidence scores. An investigation into the phonological precision of children with CLP was conducted in [8]. Here, phonological precision was automatically estimated across four levels of nasalization: normal, mild, moderate, and severe. Additionally, the authors in [1] analyzed CLP speech at the phoneme level, incorporating speaker-level evaluations. The authors employed a Support Vector Machine (SVM) along with Gaussian Mixture Models (GMM) and fMLLR [9] speaker vectors in their analysis.

In [10] an interpretation of the classification results was conducted using gradient-based feature attribution methods Integrated Gradients [11] and DeepLIFT [12] based on the same dataset we use in this work. The authors have shown how well the classification model learns meaningful audio segments based on utterance level labels, comparing to phoneme-level annotations of experts. We build upon this work and extend the base model to a large cross-lingual wav2vec 2.0 model with a weighted layer sum of the transformer block outputs and multi-head self-attention classification module. Our main contribution is the analysis of self-attention and how it aligns with the phoneme-level expert labels. Further, we show that different attention heads of the classification module focus on particular phonemes / phonetic patterns and how it affects each articulatory process class by pruning single attention heads.

2. Data

We use recordings of children with CLP, explored in [1]. The recordings were gathered using the picture-naming test PLAKSS (Psycholinguistische Analyse kindlicher Sprechstörungen — Psycholinguistic analysis of children's speech disorders) [13]. The test consists of 99 unique word stimuli, spanning all consonants, vowels, and consonant clusters following German phonotactics. Words with different lengths and stress patterns are included. Children were presented two to four pictograms representing the words to be

10.21437/Interspeech.2024-2134

named, which were recorded sequentially in 33 turns. The dataset comprises recordings from 65 male and 55 female children, aged between 3.8 and 15.7 years (7.9 ± 4.6), the recording length is 4.7 ± 2.2s. Recordings of 120 children were annotated at phoneme-level by one speech therapist with regard to six distinct articulation processes (Hypernasality, Hyponasality, Tension, Elision, Pharyngeal Backing and Interdentality), while five additional speech therapists rated 27 children. Details on the particular processes are provided in the work we build on [1]. We also obtain phoneme-level force alignments in addition to the phoneme-level annotations from the work in [10]. Since the analysis of all six articulatory processes would not allow a detailed explanation on the remaining pages, we decided to focus on two specific processes common in children with CLP: Hypernasality and Tension.

Hypernasality: Excessive air passage through the nose is a frequent occurrence in children with CLP, primarily due to velopharyngeal insufficiency. We will refer to it as nasality for convenience.

Tension: Articulation tension is reduced, primarily leading to a decreased consonant pressure.

We select a single speech turn of the PLAKSS test, consisting of three images for better visualization. The selected turn consists of the three words "Wurst" - [vU6st] (Sausage), "Löwe" - [l2:v@] (Lion) and "Lampe" - [lamp@] (Lamp). The SAMPA (Speech Assessment Methods Phonetic Alphabet) is used for the phonemic presentation throughout the work, since the phoneme annotation was performed using SAMPA. We selected the turn based on the number of utterances and as it is one of the turns with the most annotated phonemes (nasality: 373, tension: 266) to obtain a sufficiently representative evaluation of the two articulatory processes. A fixed train/test split (80%/20%) is used throughout the work, stratified by articulatory process.

3. Methods

We adopt wav2vec 2.0 [14] as the base model and add a classification module for classifying the articulatory processes. To gain an understanding of the classification decisions resulting from the proposed system, we study the attention of the classification model. We analyze the general attention heads of the classification module in order to understand which head focuses on which part of the speech signal. Further, we calculate the hierarchical attention relevance to obtain the attention attribution for each articulatory process, since our classification model is able to predict multiple labels for a single utterance. Lastly, we analyze how attention head pruning affects the results.

3.1. Classification system

We adopt the classification system in [10]: Instead of using the wav2vec 2.0 base model, we employ a cross-lingual wav2vec 2.0 model [15], pre-trained on 53 languages to capture a wider range of acoustic characteristics. In the classification module we use weighted layer sum instead of mean pooling of the transformer block outputs by weighting the outputs of the transformer blocks x^n with trainable parameters w_n. This approach has proven to be effective in training for downstream tasks. We further expand the multi-head self-attention [16] from 8 to 16 heads to comply with the underlying base model. The final classification system consists of a frozen wav2vec 2.0 model and a classification module for multi-label and multi-class prediction. The self-attention module of the classification module takes the $24 \times t \times 1024$-dimensional final hidden representations as inputs, where t refers to 20ms audio frames. Self-attention is succeeded by two blocks comprising dropout [17] with a probability of 0.1, followed by a linear layer with a hidden size of 2048, and GELU [18] activation function. We train the model until convergence using binary cross-entropy (BCE) as loss function with a learning rate of 3×10^{-4} and batch size of 8. The total number of parameters is 340M, number of trainable parameters is 25.2M. Training takes about 3.2 hours on a single A100 GPU.

3.2. Classification module attention

Let M be a transformer model equipped with a classification module comprising B blocks. Each block b consists of self-attention in the 24 transformer blocks and an additional classification block. The model is designed to handle an input sequence comprising s tokens, each with a dimensionality of d. The output of M is a classification probability vector y. The self-attention module functions within a reduced subspace d_h of the embedding dimension d. Whereas h is the number of attention-heads, resulting in: $hd_h = d$.

In order to understand to what extent which attention component focuses on relevant regions for classification, we examine the attention weights in the classification module by extracting these for each head N_h of the multi-head attention (MHA) for an input X:

$$\text{MHA}(X) = \sum_{i=1}^{N_h} Att(W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}, W_O^{(i)}, X) \quad (1)$$

where $W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}, W_O^{(i)} \in \mathbb{R}^{d \times d_h}$ denote the query, key, value, and output matrices.

3.3. Hierarchical attention relevance

In every attention layer, we utilize the method outlined in [19] to assess the relevance of attention scores among attention heads, generating a gradient-weighted attention relevance map, denoted as $\bar{A}$. For each attention map, $A^{(b)}$, its gradient is denoted as $\nabla A^{(b)}$. The relevance map is defined as $R^{(n_b)}$ with respect to a target class t, where n_b is the corresponding attention layer. The weighted attention relevance outputs $\mathbf{C} \in \mathbb{R}^{s \times s}$:

$$\bar{A}^{(b)} = I + E_h(\nabla A^{(b)} \odot R^{(n_b)})^+ \quad (2)$$

$$\mathbf{C} = w_{(1)}\bar{A}^{(1)} \cdot w_{(2)}\bar{A}^{(2)} \cdot ... \cdot w_{(n_b)}\bar{A}^{(B)} \quad (3)$$

$\odot$ is the Hadamard product, E_h the mean across the heads. $w_{(n_b)}$ indicate the trainable weights for the weighted layer sum of the classification module. Residual connections in the transformer blocks are handled by adding the identity matrix I to the layer attention matrices. We initialize the relevance propagation with a one-hot vector for the target class t.

3.4. Attention head pruning

That specific attention head's importance varies and individual heads play most important roles, was analyzed in several works [20, 21, 22]. While the model predicts multiple labels for a single utterance, not all attention heads of the classification module are supposed to attend to relevant content for a particular articulatory process. We therefore prune single attention heads to explore the effect on the results. By using head pruning, the output does not contribute to the final output of the attention mechanism, leaving the inputs unchanged for that particular attention head. For this, we introduce variables $z_h^{(i)} \in \{0, 1\}$ to the classification module multi-head attention.

4. Experiments and Results

4.1. Classification results

The overall results across all test turns are comparable to those presented in [10]. However, as our focus lies on a specific recording turn, we solely present the results related to this turn. As shown in Table 1, the results for this individual turn surpass the overall averaged results, potentially due to the larger sample size and number of labeled phonemes per process. The examined articulatory processes in this study are highlighted in bold. Nasality achieves an F1 score of 0.89, while Tension attains an F1 score of 0.83.

Table 1: *Classification results of the proposed system with an attention classification module in a multi-class and multi-label scenario for the selected turn consisting of three words. The final column shows the F1 scores from [10] for comparison.*

Articulatory process	Precision	Recall	F1	F1 [10]
Hypernasality	**0.96**	**0.83**	**0.89**	0.71
Hyponasality	0.86	1.00	0.93	0.57
Tension	**0.90**	**0.76**	**0.83**	0.78
Elision	1.00	1.00	1.00	0.27
Pharyngeal Backing	0.80	0.57	0.67	0.58
Interdentality	1.00	0.62	0.76	0.54

4.2. Classification module attention analysis

We first analyze only the attention maps of the classification module for each head. We obtain the attention map $A^{(B-1)}$ for each head, which corresponds to the classification module attention block. Figure 1 shows the attention probabilities of the classification module heads averaged over all utterances of our selected turn. It is evident that the individual heads focus on specific phonemes or phonetic patterns. Head 5 for example focuses on vowels: /U/, /6/ and /2:/. Whereas, head 4 focuses on the specific phonetic pattern /a m p/. We can also observe that some heads have fairly low attention weights, at least for this particular turn, such as head 6 and 15. Yet, this analysis is target label agnostic, it is not evident which head is important for a particular articulatory process. Therefore, we obtain target label specific attention relevance in the next step.

4.3. Hierarchical attention relevance

To obtain the attention attribution for a single target label and to align the attention attribution to the phoneme-level expert ratings, we extract attention relevance maps using the described method in Section 3.3. We assess the relevance for a single articulatory process by initializing a one-hot vector for the target process t. We compute forward for an utterance and sum the logits with the initialized vector and backpropagate the obtained vector. This is performed for each head of the classification module and each utterance of the examined turn. An example of resulting relevance maps for classification module head 13 is shown in Figure 2. The visualized relevance map shows positive attribution to the phoneme /t/, which was also been identified by an expert as nasality.

By using the force alignments, we identify the phoneme boundaries in each utterance and sum the obtained attention relevance scores for each phoneme. The resulting attention attribution is visualized in Figure 3 for two utterances of both examined articulatory processes. Phoneme ranges with an orange background identify the phonemes labeled for the particular articulatory process nasality (a) and tension (b) by the experts.

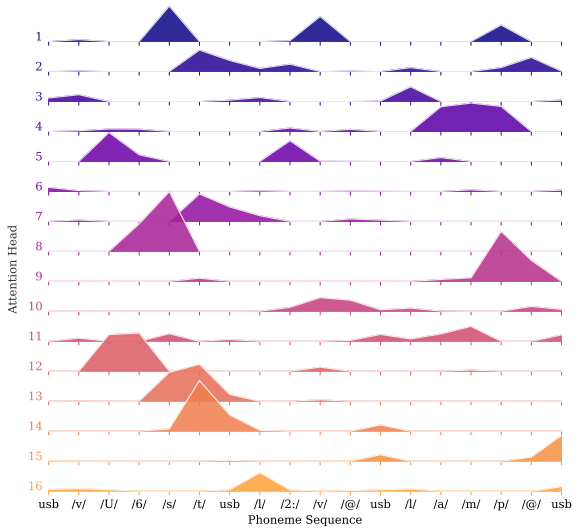


Figure 1: *Attention attribution per head for all utterances of the examined turn. The x-axis shows the spoken phonemes in time-dependent order, y-axis the mean attribution per phoneme segment averaged over all utterances.*

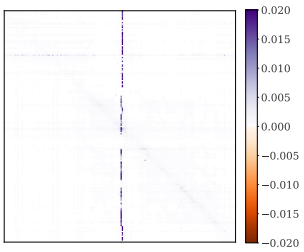


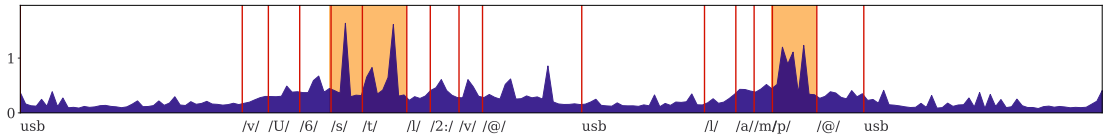
Figure 2: *Gradient-weighted attention relevance map for classification head 13, with positive attention attribution on a particular phoneme /t/ for the articulatory process nasality.*

A holistic overview of the labeled weights per phoneme of the utterance across all utterances of the turn can be seen in Figure 4. It shows the average attention relevance per phoneme for the examined turn. In case of nasality especially phonemes /s/ and /t/ and the pattern /a m p/ are in accordance with the experts ratings, labeled phonemes exhibit the highest weights compared to the unlabeled ones. For tension, the phonemes /U/ and /l/ align well with the expert labels, while /@/ is in contrast to the expert labels. Overall, the tension articulatory process does not align as good as nasality. Often the attention attribution is high even when no expert label was assigned to a phoneme. Notably, this mostly happens when another phoneme was labeled as tension. This could indicate that not all phonemes were labeled accurately.

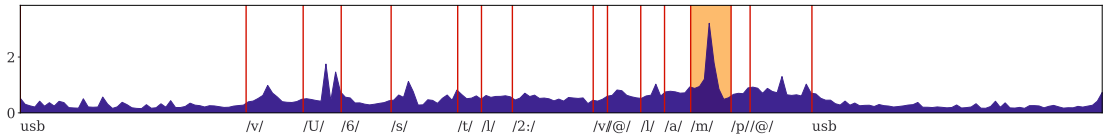
Upon closer examination of the attention relevance maps, we find that some heads offer minimal attributions to a target process. By modifying Equation 2 to include both positive and negative weights, it's evident that certain heads only contribute negatively. This suggests that pruning heads can enhance performance for a single target class.

4.4. Attention head pruning

Through our observation in the attention maps, we investigate the influence of individual attention heads of the classification module and prune individual heads. We only take the classification module attention heads into account, since the wav2vec 2.0

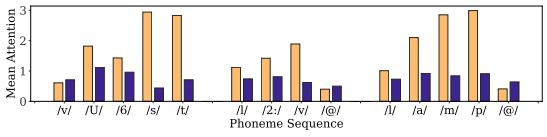


(a) *Nasality example*

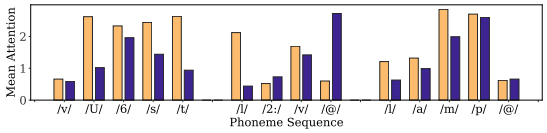


(b) *Tension example*

Figure 3: *Attention attribution for single utterances, x-axis represents phonemes and their borders (red lines) over time, y-axis the mean attribution using the gradient weighted attention attribution method. Phoneme regions with an orange background were labeled with the respective articulatory process by an expert rater.*

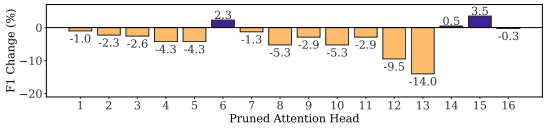


(a) *Nasality*

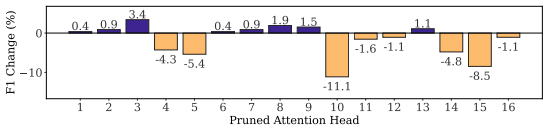


(b) *Tension*

Figure 4: *Attention attribution per turn phoneme showing with a process labeled attention (purple) vs. no label (orange) for all utterances and phonemes of the turn.*



(a) *Nasality results*



(b) *Tension results*

Figure 5: *Attention head pruning results for the articulatory processes (a) Nasality and (b) Tension. The y-axis shows the absolute F1 score change in % for each pruned attention head.*

model was frozen during training and the most relevant information is encoded in the classification module. One individual attention head at a time is pruned for this purpose.

The results in Figure 5 (a) show that pruning head 6, 14 and 15 leads to an absolute F1 improvement of up to 3.5%. In contrast, pruning head 13 leads to a decrease in F1 score of 14%. The analysis of the remaining articulatory processes revealed that an absolute improvement of up to 38% in F1 score can be achieved, e.g., for Pharyngeal Backing process.

5. Discussion

This work tackles the crucial challenge of automatically classifying articulatory processes in children with CLP. The proposed system, utilizing wav2vec 2.0, introduces a method for automated assessment. It categorizes articulatory processes at utterance level, eliminating the necessity for phoneme-level labels.

The self-attention analysis provides insight into the attribution of attention to specific articulatory processes in CLP speech, allowing for a more nuanced understanding of how the model makes its decisions. By quantifying the relevance of attention scores per class, phonemes or phonetic patterns can be identified which are most influential in determining particular articulatory abnormalities, thus enhancing the interpretability of the model's predictions. The identification of negative influences of single attention heads for a single class highlights the complexity of the multi-label setting. This finding underscores the importance of evaluating individual attention mechanisms within the model to understand their contribution to classification performance accurately. The exploration of attention head pruning represents a significant advancement in the optimization of the classification system. By selectively removing attention heads that do not contribute meaningfully to the classification task, the model can be optimized and potentially improve its efficiency and reliability, while providing interpretable decisions. This approach demonstrates a proactive strategy for model refinement, emphasizing the importance of careful attention to the design and implementation of attention mechanisms in multi-class and multi-label models. Finally, attention analysis offers a less resource-intensive alternative to methods like Integrated Gradients and DeepLIFT, making it more accessible for practical application in clinical settings.

6. Conclusion

Our study presents a novel approach for automated assessment of articulatory processes in children with CLP speech and the role of attention for improved classification systems. Attention analysis and pruning techniques enhance the interpretability and efficiency of the model, paving the way for more accurate and reliable clinical assessments. We have shown, using a hierarchical attention attribution method, that attention attribution aligns with expert labels. Further, pruning single attention heads, enhances absolute classification results of up to 38% F1 score.

Future work will focus on refining attention mechanisms to prevent the negative influence of single attention heads on classification performance. One potential avenue is the implementation of gating or re-scoring mechanisms within the attention module to mitigate the impact of disruptive attention heads. Since gating experiments are beyond the scope of this paper, it will be part of future investigation.

7. Acknowledgements

This work is supported by the Bavarian Ministry of Health, Care and Prevention, with funding derived entirely from the European Union's NextGenerationEU program.

8. References

[1] T. Bocklet, K. Riedhammer, U. Eysholdt, and E. Nöth, "Automatic phoneme analysis in children with cleft lip and palate," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7572–7576.

[2] A. Harding and P. Grunwell, "Active versus passive cleft-type speech characteristics," *International journal of language & communication disorders*, vol. 33, no. 3, pp. 329–352, 1998.

[3] S. Jain and B. C. Wallace, "Attention is not Explanation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3543–3556.

[4] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 11–20.

[5] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? an analysis of BERT's attention," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 276–286.

[6] A. Maier, F. Hoenig, T. Bocklet, E. Noeth, F. Stelzle, E. Nkenke, and M. Schuster, "Automatic detection of articulation disorders in children with cleft lip and palate," *The Journal of the Acoustical Society of America*, vol. 126, pp. 2589–602, 11 2009.

[7] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2, pp. 95–108, 2000.

[8] T. Arias-Vergara, E. Londoño-Mora, P. A. Pérez-Toro, M. Schuster, E. Nöth, J. R. Orozco-Arroyave, and A. Maier, "Measuring Phonological Precision in Children with Cleft Lip and Palate," in *Proc. INTERSPEECH 2023*, 2023, pp. 4638–4642.

[9] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[10] I. Baumann, D. Wagner, M. Schuster, E. Nöth, and T. Bocklet, "Towards interpretability of automatic phoneme analysis in cleft lip and palate speech," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[11] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[12] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17, 2017, p. 3145–3153.

[13] A. Fox-Boyer, *PLAKSS-II: Psycholinguistische Analyse kindlicher Aussprachestörungen*. Pearson, 2014.

[14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[15] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. INTERSPEECH 2023*, 08 2021, pp. 2426–2430.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[18] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[19] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 782–791.

[20] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5797–5808.

[21] M. Xia, Z. Zhong, and D. Chen, "Structured pruning learns compact and accurate models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1513–1528.

[22] M. Behnke and K. Heafield, "Losing heads in the lottery: Pruning transformer attention in neural machine translation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, Nov. 2020, pp. 2664–2674.