TRANSFORMER-BASED ESTIMATION OF SPOKEN SENTENCES USING ELECTROCORTICOGRAPHY

Shuji Komeiji[1], Kai Shigemi[1], Takumi Mitsuhashi[2], Yasushi Iimura[2], Hiroharu Suzuki[2], Hidenori Sugano[2], Koichi Shinoda[3], and Toshihisa Tanaka[1]

[1] Department of Electronic and Information Engineering, Tokyo University of Agriculture and Technology
[2] Department of Neurosurgery, Juntendo University School of Medicine
[3] Department of Computer Science, Tokyo Institute of Technology

ABSTRACT

Invasive brain–machine interfaces (BMIs) are a promising neurotechnological venture for achieving direct speech communication from a human brain, but it faces many challenges. In this paper, we measured the invasive electrocorticogram (ECoG) signals from seven participating epilepsy patients as they spoke a sentence consisting of multiple phrases. A Transformer encoder was incorporated into a "sequence-to-sequence" model to decode spoken sentences from the ECoG. The decoding test revealed that the use of the Transformer model achieved a minimum phrase error rate (PER) of 16.4%, and the median ($\pm$ standard deviation) across seven participants was 31.3% ($\pm$10.0%). Moreover, the proposed model with the Transformer achieved significantly better decoding accuracy than a conventional long short-term memory model.

Index Terms— Electrocorticogram (ECoG), Brain–machine interface (BMI), Transformer encoder, Sequence to sequence

1. INTRODUCTION

Brain–machine interfacing (BMI), which enables speech to be decoded from human thought, is expected to be used not only by aphasic patients but also as a new communication tool in the future [1]. Several techniques to achieve such BMI, using an invasive electrocorticogram (ECoG) measured by electrodes implanted in the skull, are under development. The ECoG is superior to a surface electroencephalography in terms of spatio and temporal resolution and signal-to-noise ratio; it is particularly suitable for analyzing brain activity related to speech in the high gamma band [2, 3].

A variety of forms of speech decoding using ECoG have been studied, from phoneme-based decoding to sentence-based decoding, and for speaking, listening, and imagining brain activities. To enable isolated-word speech decoding, Pei et al. used a naive Bayes classifier for speaking and imagining tasks [3], and Martin et al. used a support vector machine for speaking, listening, and imagining tasks [4]. To enable sentence-based speech decoding, Viterbi decoding with a hidden Markov model was applied by Herff et al. for a speaking task [5] and by Moses et al. for a listening task [6, 7].

With the advent of deep-learning techniques, recurrent neural networks (RNNs) have been applied to decoding speech from ECoG signals. Sun et al. used a combination of a long short-term memory (LSTM) RNN model [8] and a connectionist temporal classification decoder [9] for speaking and imagining tasks [10]. Makin et al. successfully applied a "sequence-to-sequence" model, composed of an encoder stage and a decoder stage with bidirectional LSTM (BLSTM) for speaking tasks [11]. Moreover, an effective method in training the network with a limited amount of ECoG data has been proposed wherein the network intermediate layer is trained to output speech-latent features with lower dimensions than the input features [11, 12, 13]. Sun et al. [10] and Makin et al. [11] trained LSTM layers in a sequence-to-sequence encoder using mutually synchronized ECoG signals and Mel frequency cepstral coefficients (MFCCs) as inputs and outputs, respectively.

However, it has been generally known that LSTM has drawbacks in learning longer-range dependencies between the input and output sequences. Thus, a so-called Transformer model [14] has been successfully applied in natural language processing (NLP) [15] and automatic speech recognition (ASR) [16]. We therefore hypothesize that the Transformer works efficiently in decoding spoken sentences from the ECoG. This paper is the first to report an invasive BMI that decodes speech using a Transformer embedded in the encoder stage of a sequence-to-sequence model to decode spoken sentences from ECoG signals. The experimental results obtained from seven participants performing the speaking task showed that the proposed model with the Transformer achieved significantly better decoding accuracy than a conventional BLSTM.

2. METHODS

2.1. Participants

The seven volunteer participants (four males: js1, js5, js6, and js8; three females: js3, js4, and js7) in this study were undergoing treatment for epilepsy at the Department of Neurosurgery, Juntendo University Hospital. ECoG arrays were surgically implanted on each participant's cortical surface (left hemisphere) to localize their seizure foci. The participants gave written informed consent to participate in this study, which was executed according to a protocol approved by Juntendo University Hospital and the Tokyo University of Agriculture and Technology.

2.2. Experimental Design

ECoGs were recorded for the speaking task, wherein the participants read sentences displayed on a monitor aloud. Each sentence was in Japanese and consisted of three phrases. Each phrase had two candidates to generate one sentence as described in the following: The first phrase was either "w a t a sh i w a" (I) [1] or "k i m i t

This work was supported in part by JSPS KAKENHI 20H00235.

[1] Japanese pronunciation with the corresponding English translation in parentheses.

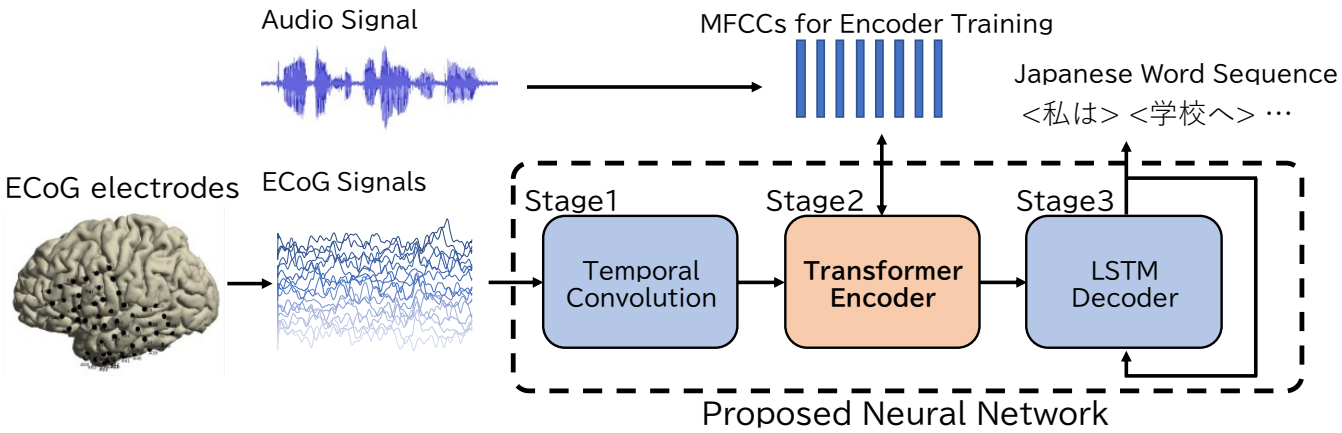


Fig. 1: The decoding pipeline. The details of the Transformer encoder in the proposed neural network, which is the most unique characteristic of this network, are described in Fig. 2.

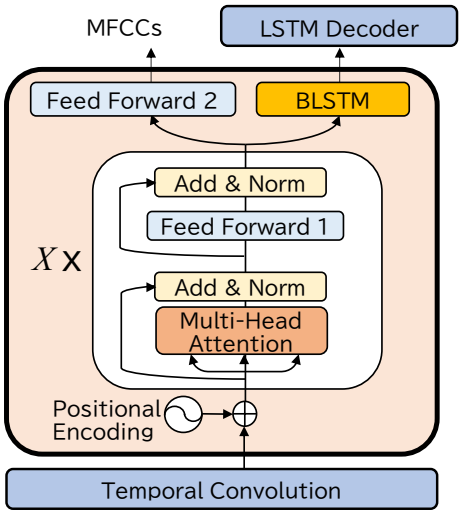


Fig. 2: The proposed structure of the Transformer encoder with a feed-forward layer, which outputs speech-latent variables (MFCCs) and a BLSTM layer connected to the decoder

o" (with you), the second phrase was either "g a q k o: e" (to the school) or "sh o k u b a n i" (to the office), and the third phrase was either "i q t a" (went) or "m u k a u" (go). In consequence, we generated eight ($2 \times 2 \times 2$) patterns of sentences.

Since each sentence pattern was displayed ten times, the total number of sentences read by each participant was 80. The 80 sentences were displayed to each participant in a random order.

2.3. Data Acquisition

ECoG signals were recorded using biosignal amplifiers (g.HIAMP, g.tec Medical Engineering GmbH). In addition to ECoG signals, acoustic signals, such as participants' speech signals, were recorded with a microphone (ECM360, SONY) and digitized in sync with the ECoG signals. The ECoG and acoustic signals were then digitized at sampling rates of 1,200 Hz for participants js1 and js3 (Group 1) and 9,600 Hz for participants js4, js5, js6, js7, and js8 (Group 2)[2].

2.4. ECoG Preprocessing

The ECoG signals from electrodes with visible artifacts or excessive noise were removed. As a result, the numbers of electrodes used for each participant in the analysis were as follows: 61 for js1, 38 for js3, 54 for js4, 49 for js5, 51 for js6, 29 for js7, and 48 for js8 through the experiments.

Next, we trimmed the ECoG signals and the audio signals from the starting point of each spoken sentence. The length of each trimmed signal was the maximum duration of the speech signals given by each participant. This operation prevented the speech decoder from predicting sentences based on their trimmed length. These trimmed ECoG signals were anti-aliased (low-pass filtered at 200 Hz) and downsampled to 400 Hz. IIR notch filters of 50 and 100 Hz were applied to suppress the power supply noise and the harmonic. Finally, the analytic amplitudes at each electrode were extracted in each of eight adjacent frequency bands, between 70 and 150 Hz, filtered by eight finite impulse response (FIR) bandpass filters, averaged across bands, and downsampled to 200 Hz [6, 7, 11, 17]. The eight FIR bandpass filters were designed with passbands of 68–78, 74–84, 82–92, 91–102, 101–112, 112–124, 124–136, and 137–150 Hz. The amplitudes of the analytic signal were then z-scored, and ECoG preprocessed signals were obtained.

2.5. Network Architecture

Figure 1 shows the proposed architecture of the sequence-to-sequence style artificial neural network [18]. The network processes the sequences in three stages: a temporal convolution network, an encoder network, and a decoder network. As depicted in Fig. 2, the proposed architecture is characterized by the use of a Transformer encoder in the encoder stage, whose outputs are trained to be MFCCs. If only BLSTM layers are used in the encoder stage, the proposed model corresponds to the model proposed by Makin et al. [11]. Although it may be possible to use a Transformer decoder in the decoder stage, the LSTM decoder was adopted for reasons of sufficiency for decoding simple sentences.

[2]The use of the low sampling rate of 1,200 Hz for Group 1 (js1 and js3) was an accident during the experiment. However, we did not discard Group 1 from the analysis.

The temporal convolution (convolutional neural network; CNN) layer effectively downsamples the ECoG preprocessed signals. There are K of these signals (where K is the number of electrodes), each with a length of L; the signals are processed by C convolutional filters. Each convolutional filter with a kernel size of $W \times K$ is applied to the K signals with a stride of W; as a result, the CNN layer outputs a sequence of N (= integer$(L/W) + 1$) vectors of C with dimmension of C.

In the second stage, the Transformer encoder receives the outputs of the temporal convolution layer and is composed of a stack of X identical layers, each with two sub-layers. The first is the multi-head self-attention mechanism and the second is a fully connected feed-forward network. A residual connection around each of the two sub-layers, followed by layer normalization, is employed. The outputs of the Transformer encoder are the sequence of feature vectors of dimension C with a length of N, the same size as the inputs. The outputs of the stack of X identical layers are input to two different layers; one is BLSTM, whose last hidden and cell states are input to the next stage, and the other is the feed-forward network called "feed-forward 2" to represent the sequence of MFCCs. The "feed-forward 2" layer plays an important role in regularizing the outputs of the Transformer encoder. These outputs span a latent feature space with lower dimensions and mitigate the problem of limited training data [11, 12, 13]. In training, the feed-forward outputs are made to approximate to the sequence of MFCCs whose length is the same as that of the outputs of the temporal convolution layer N. Note that in the test phase, the MFCCs are not applied to the output of the encoder.

The third stage is the decoder network layer with an LSTM layer whose initial hidden and cell states are equal to the last ones of the BLSTM in the second stage. The inputs and outputs of the decoder are the sequences of one-hot encoding vectors that represent phrases, the start of the sentence, or the end of the sentence. To adjust the dimensions of the inputs and outputs of the LSTM based on those of the one-hot vectors, the feed-forward layers named "feed-forward 3" and "feed-forward 4" are put before the input and after the output of the LSTM.

2.6. MFCCs Used in Training

The trimmed audio signals were used to calculate MFCCs, which were calculated with the python_speech_features package using 20-ms sliding frames with a slide of $1/200$. For two participant groups, Groups 1 and 2, we considered two scenarios regarding the MFCCs used in the encoder.

For Group 1, due to the low sampling rate for speech (1,200 Hz), we did not use the speech signals recorded with each participant's ECoG in a synchronized way. Instead, the MFCCs used in the training phase were the MFCCs obtained by the speech of js4[3]. For Group 2, we used the MFCCs obtained from their speech signals recorded at the same time as the ECoG signals.

2.7. Network Implementation and Parameters

The network parameters used in the experiment are shown in Table 1. For the layers shared by the model proposed by Makin et al. [11], the same values were taken as parameters to simplify the comparison with the proposed model.

As can be seen in Table 1, the number of convolutional filters (C) and kernel size (W) were set to 100 and 12, respectively. The other kernel size K remains a variable number because it depends on the participants. The output dimension of "feed-forward 2" was set to 13 because we set the dimension of MFCCs to 13. The output dimension of "feed-forward 4" was set to eight for a vocabulary size of six plus two (the start and end tokens). The variables B and N in Table 1 represent the batch size, which is a training parameter, and the sequence length, which depends on data.

[3]The reason js4 was chosen as the MFCCs for Group 1 is that js4 achieved the best performance in decoding speech among Group 2.

Table 1: Network parameters

Layer name	Parameter	Output shape
Temporal convolution	filters = 100	(B, N, 100)
	kernel size = (12, K)	
	strides = (12,1)	
	dropout = 0.1	
Transformer Encoder	$X = 2$	(B, N, 100)
	attention head = 10	
	hidden units = 100	
	dropout = 0.5	
BLSTM	forward units = 400	(B, N, 400)
	backward units = 400	hidden = (800)
	dropout = 0.5	cell = (800)
Feed-forward 2	units = 13	(B, N, 13)
Feed-forward 3	units = 800	(B, 800)
LSTM	units = 800	(B, 800)
	dropout = 0.1	
Feed-forward 4	units = 8	(B, 8)

Table 2: Training parameters

Parameter name	Value
Learning rate	0.0005
MFCC-penalty weight, λ	0.1
Batch size	16
# training epochs	800

2.8. Network Training

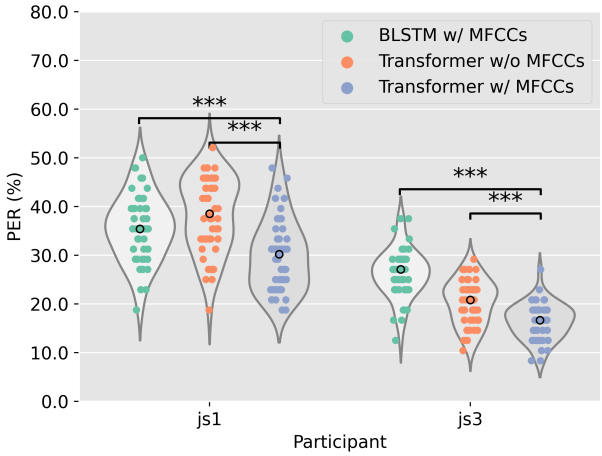
The proposed model is trained by reducing both the encoder and decoder losses, which are the differences between the encoder outputs and MFCCs, and between the decoder outputs and the phrase sequence, respectively. The losses are summed with an MFCC-penalty weight λ. The training parameters are shown in Table 2. Some are the same parameters used by Makin et al. [11].

2.9. Evaluation Items

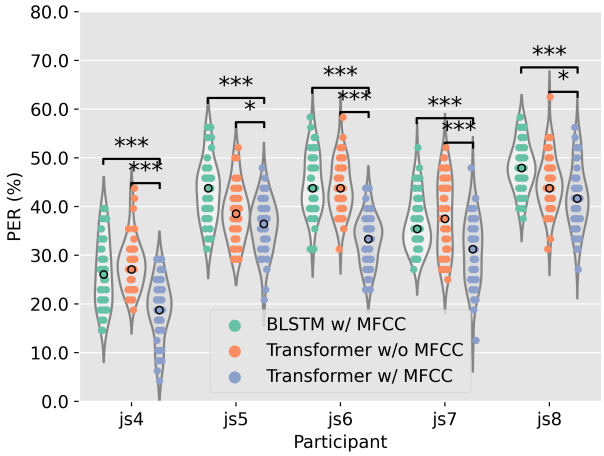
Three models below are evaluated in this paper:

- BLSTM with MFCCs
- Transformer without MFCCs
- Transformer with MFCCs

The BLSTM-with-MFCCs model has two BLSTM layers in the encoder instead of positional encoding and a stack of $X = 2$ identical layers in Fig. 2. It is equivalent to the model proposed by Makin et al. [11]. The parameters of the two BLSTM layers were the same as those of the last BLSTM layer in the encoder in Fig. 2. The Transformer-without-MFCCs and Transformer-with-MFCCs models are the proposed model, trained without or with the MFCCs; i.e., the MFCC-penalty weight λ has been set to 0.0 or 0.1. A comparison of the BLSTM-with-MFCCs and Transformer-with-MFCCs models



(a) Group 1: Using the MFCCs of participant js4.



(b) Group 2: Using MFCCs synchronized with the ECoG of each participant.

Fig. 3: PERs of the decoded sentences. The vertical axis shows the phrase error rate (PER); the horizontal axis shows the participants. The "o" marks for participants and models indicate medians. The asterisks indicate the significance between the two models. The p-values were computed with a Wilcoxon signed-rank test (*: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$). (a) Group 1 shows the models using MFCCs derived from the speech of participant js4. (b) Group 2 shows the models with MFCCs derived from the participant's speech, as synchronized with ECoG.

shows the effectiveness of using a Transformer. The comparison of the Transformer models with and without MFCCs shows the contribution of the MFCCs to the training of the encoder.

2.10. Evaluation Methods

Five-fold cross-validation was used for the evaluation. All 80 trials for each participant were divided into a training set of 64 trials and a test set of 16 trials. Each fold of the inferred 16 sentences from the test set was evaluated in terms of a phrase error rate (PER). Thus, five PERs were obtained across the cross-validation. To mitigate the problem of non-repeatability of the model training due to the randomness of the initial weights of the model and calculation by GPUs, the series of processes from training to evaluation was repeated ten times. As a result, 50 PERs were obtained for each participant. The significant-difference test used to compare the two models was the Wilcoxon signed-rank test using 50 PERs.

3. RESULTS

Figure 3 shows the PERs by participants for each of the three models for (a) Group 1 and (b) Group 2. According to Fig. 3 (a) and (b), the Transformer-with-MFCCs model was found to be significantly the most effective for all participants. The p-values were less than 0.001, with the exception of the comparison of Transformer models with or without MFCCs for participants js5 and js8.

The average PERs for each subject according to the Transformer-with-MFCCs model with standard deviations were $29.9 \pm 7.3\%$ for js1, and $16.4 \pm 3.7\%$ for js3 in Group 1, and $18.7 \pm 6.3\%$ for js4, $35.8 \pm 5.7\%$ for js5, $33.1 \pm 5.0\%$ for js6, $30.9 \pm 7.0\%$ for js7, and $41.9 \pm 6.5\%$ for js8 in Group 2. The Transformer-with-MFCCs model achieved an average PER below 20% for js4 and js3.

4. DISCUSSION

First, the proposed model achieved an average PER below 20% for js3 and js4. A recent study [11] suggested that the use of more dense electrodes may improve the PER in spite of the larger size of the vocabulary. Makin et al. [11] achieved a word-error rate as low as 3% for one specific participant using the BLSTM-with-MFCCs model for a vocabulary size of 250 with the ECoG using 250 electrodes. This vocabulary size is 41 times larger than the size used in this paper. Also, we used a less dense electrodes array than Makin et al. used [11]. It should be noted that the electrode specifications were determined to treat epilepsy and cannot be easily changed. However, it is worth noting that our study showed that PER could be improved by using the Transformer model, even with a lower density electrode array.

Second, the Transformer encoder achieved superior accuracy than the BLSTM for all seven participants; this implies that the longer-range dependencies [14], which are said to be learned by the Transformer encoder, may be useful for decoding speech from ECoG signals. Although most studies have paid attention to the local characteristics of the ECoG [3, 4, 5, 6, 7], this paper found the importance of the ECoG's longer-range dependencies. Future work will include a detailed analysis based on looking at the output of the intermediate layers, such as attention layers, and clarifying which part of the cortex and which time points are important for speech decoding.

Finally, using MFCCs contributed to the training of the Transformer encoder. In particular, from the results of Group 1, the models for the participants js1 and js3 with the MFCCs from js4 achieved better accuracy than those without MFCCs. This implies that MFCCs used in the training phase is not necessarily calculated from the speech of the same person. This finding suggests that it is possible to train imagining tasks where we cannot use MFCCs calculated from synchronized speech signals.

5. REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[2] N. E. Crone, D. L. Miglioretti, B. Gordon, and R. P. Lesser, "Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band," *Brain: A Journal of Neurology*, vol. 121, no. 12, pp. 2301–2315, 1998.

[3] X. Pei, E. C. Leuthardt, C. M. Gaona, P. Brunner, J. R. Wolpaw, and G. Schalk, "Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition," *Neuroimage*, vol. 54, no. 4, pp. 2960–2972, 2011.

[4] S. Martin, P. Brunner, I. Iturrate, J. d. R. Millán, G. Schalk, R. T. Knight, and B. N. Pasley, "Word pair classification during imagined speech using direct brain recordings," *Scientific Reports*, vol. 6, no. 1, pp. 1–12, 2016.

[5] C. Herff, D. Heger, A. De Pesters, D. Telaar, P. Brunner, G. Schalk, and T Schultz, "Brain-to-text: decoding spoken phrases from phone representations in the brain," *Frontiers in Neuroscience*, vol. 9, pp. 217, 2015.

[6] D. A. Moses, N. Mesgarani, M. K. Leonard, and E. F. Chang, "Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity," *Journal of Neural Engineering*, vol. 13, no. 5, pp. 056004, 2016.

[7] D. A. Moses, M. K. Leonard, J. G. Makin, and E. F. Chang, "Real-time decoding of question-and-answer speech dialogue using human cortical activity," *Nature Communications*, vol. 10, no. 1, pp. 1–14, 2019.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine learning*, 2006, pp. 369–376.

[10] P. Sun, G. K. Anumanchipalli, and E. F. Chang, "Brain2char: a deep architecture for decoding text from brain recordings," *Journal of Neural Engineering*, vol. 17, no. 6, pp. 066015, 2020.

[11] J. G. Makin, D. A. Moses, and E. F. Chang, "Machine translation of cortical activity to text with an encoder–decoder framework," *Nature Neuroscience*, vol. 23, no. 4, pp. 575–582, 2020.

[12] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific Reports*, vol. 9, no. 1, pp. 1–12, 2019.

[13] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[16] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K.l Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *arXiv:2106.07447*, 2021.

[17] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation," *Nature*, vol. 495, no. 7441, pp. 327–332, 2013.

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.