Received 23 August 2023; revised 1 December 2023; accepted 5 December 2023. Date of publication 18 March 2024; date of current version 18 June 2024. The review of this article was arranged by Associate Editor H.-Y. Lee.

Digital Object Identifier 10.1109/OJSP.2024.3378594

# Sea-Wave: Speech Envelope Reconstruction From Auditory EEG With an Adapted WaveNet

LIUYIN YANG, BOB VAN DYCK, AND MARC M. VAN HULLE (Fellow, IEEE)

Laboratory for Neuro and Psychophysiology, Department of Neurosciences, KU Leuven, B-3000 Leuven, Belgium

*(Liuyin Yang and Bob Van Dyck contributed equally to this work.)*

CORRESPONDING AUTHORS: LIUYIN YANG; BOB VAN DYCK (e-mail: liuyin.yang@kuleuven.be; bob.vandyck@kuleuven.be).

This work was supported in part by the Belgian Fund for Scientific Research—Flanders under FWO Grant 1S65622N, Grant G0C1522N, Grant G0A4118N, and Grant G0A4321N; in part by European Union's Horizon 2020 Research and Innovation Programme under Grant 857375; in part by the Special Research Fund of the KU Leuven under Grant C24/18/098; and in part by Hercules Foundation under Grant AKUL 043.

**ABSTRACT** Speech envelope reconstruction from EEG is shown to bear clinical potential to assess speech intelligibility. Linear models are commonly used to this end, but they have recently been outperformed in reconstruction scores by non-linear deep neural networks, particularly by dilated convolutional networks. This study presents Sea-Wave, a WaveNet-based architecture for speech envelope reconstruction that outperforms the state-of-the-art model. Our model is an extension of our submission for the Auditory EEG Challenge of the ICASSP Signal Processing Grand Challenge 2023. We improve upon our prior work by evaluating model components and hyperparameters through an ablation study and hyperparameter search, respectively. Our best subject-independent model achieves a Pearson correlation of 22.58% on seen and 11.58% on unseen subjects. After subject-specific fine-tuning, we find an average relative improvement of 30% for the seen subjects and a Pearson correlation of 56.57% for the best seen subject.Finally, we explore several model visualizations to obtain a better understanding of the model, the differences across subjects and the EEG features that relate to auditory perception.

**INDEX TERMS** Computational neuroscience, EEG, speech envelope, WaveNet.

## I. INTRODUCTION

### A. BACKGROUND

Listening to continuous speech elicits a corresponding brain response that can be measured with electroencephalography (EEG). The reconstruction (or decoding) of the presented speech stimulus from EEG is a way to quantify neural tracking, a neural mechanism that supports speech perception and has been demonstrated for both acoustic and higher order representations of speech. Reconstructing the speech stimulus envelope is potentially useful for diagnostic tests, as speech envelope tracking has been successfully linked to speech intelligibility [1], [2], [3], [4]. While linear models have been used extensively to relate EEG to speech (e.g. [2], [5], [6], [7]), they suffer from low reconstruction scores compared to deep learning-based non-linear models [8], [9], [10], especially the subject-independent models. Currently, dilated convolutional non-linear neural networks have been applied successfully to auditory decoding, both for direct regression to a speech feature, such as the speech envelope [10] and fundamental frequency [11], and for classification [12] in a "match-mismatch" paradigm [13]. This development is related to the accessibility of larger datasets, essential for training models with increased capacity. The 2023 ICASSP Signal Processing Grand Challenge provides a substantial dataset that contributes to the further investigation of deep learning-based decoding approaches. The second sub-task focuses on reconstructing speech envelopes from EEG signals to establish correlations between the speech signal and neural activity. This dataset comprises recordings from 85 subjects, who listened to 108 minutes of single-speaker stimuli on average, culminating in approximately 157 hours [14], [15].

### B. PROBLEM STATEMENT

In this paper, we focused on the regression subtask of the Auditory-EEG challenge [16]. The regression task consists of reconstructing the auditory stimulus envelope from the recorded 64-channel EEG signal (from $\mathbb{R}^{64\times t}$ to $\mathbb{R}^{1\times t}$). The challenge organizers proposed the Pearson correlation (Pearson r) for measuring agreement between the predicted and actual envelopes. Reconstruction scores are computed per subject $s$ by averaging across stimuli (Pearson $\mathrm{r}_s$). The final model performance (score) is a weighted sum of the Pearson $\mathrm{r}_s$'s averaged over a set of held-out stories ($S_1$) and a set of held-out subjects ($S_2$), computed as

$$\text{score} = \frac{2}{3}\sum_{s\in S_1}\frac{\text{Pearson r}_s}{|S_1|} + \frac{1}{3}\sum_{s\in S_2}\frac{\text{Pearson r}_s}{|S_2|}. \tag{1}$$

### C. METHODS

Linear models are commonly applied to reconstruct speech envelopes [2], [5], [6], [7], but they suffer from low reconstruction scores, as evidenced by the subject-specific linear model in [10], which achieves a maximal Pearson correlation below 0.2. In 2023, researchers proposed a dilated convolutional network, VLAAI [10], setting the state-of-the-art performance by improving the performance of linear models by 52%. During the competition, only 5 out of the 13 teams that participated managed to improve upon the linear baseline model. Our adapted WaveNet model ranked second in the challenge, while the best model was a transformer-based model [17]. In this study, we introduce a new Sea-Wave model,[1] a novel architecture that advances the performance of our adapted WaveNet model [18]. The improved model architecture is obtained by introducing novel model components and evaluating both existing and novel components through an ablation study. Next, a hyperparameter search is performed to understand the effect of hyperparameter choices and related model attributes, such as receptive field size, on reconstruction performance and generalization capability to unseen speech stimuli and subjects. Additionally, the search enables selecting models with a suitable trade-off between performance and model size for subsequent experiments, evaluation and visualization.

### D. SIGNIFICANCE

This paper proposes a new Sea-Wave model architecture, setting a new benchmark in speech envelope reconstruction when evaluated on both the challenge dataset and the DTU dataset. The model is both compact and computationally efficient, making it well-suited for real-time applications. Subject-finetuning was confirmed to be an effective way of improving model performance. Our best subject-independent model achieves a Pearson r of 22.58% on seen and 11.58% on unseen subjects, an improvement upon our prior work of 30% for the seen subjects. The topographic maps (or scalp plots) depicting model weights and channel importance not only affirm the importance of the auditory cortex but also underscore the involvement of visual and somatosensory areas, suggesting a need for additional exploration into their roles in speech perception.

The paper is organized as follows. We start by detailing our data preprocessing and partitioning. Next, we present the WaveNet-based model submitted to the challenge [18] and the improved model architecture of Sea-Wave. The following sections present the ablation study on the model components, the hyperparameter search and our subject-specific fine-tuning strategy. Finally, a number of model visualizations are explored, yielding insights into the differences between subjects and the EEG features that relate to auditory perception.

## II. DATA PREPROCESSING AND PARTITIONING

We use the preprocessed data provided by the challenge organizer, consisting of speech envelopes and filtered EEG as defined in [2] and [16], respectively. The speech envelope is estimated through a gammatone filter bank [19] comprising 28 subbands, each appropriately spaced by an equivalent bandwidth and centered at frequencies ranging from 50 Hz to 5 kHz. Subsequently, the absolute value of each sample within each filter bank is computed, followed by an exponentiation with a value of 0.6. Finally, all subbands are averaged to generate a unified speech envelope. The EEG is downsampled from its original sampling rate of 8192 Hz to 1024 Hz. Next, artifact removal using a multichannel Wiener filter and common average re-referencing are applied to the EEG. The obtained speech envelopes and EEG are further downsampled to 64 Hz.

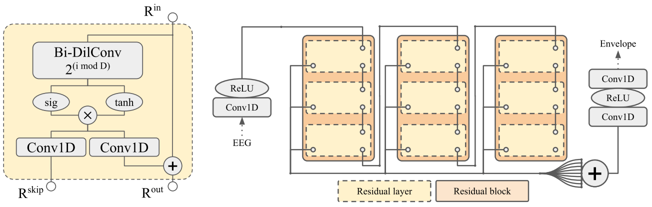
The challenge training and test set defined by the challenge organizer are as follows. The challenge training set comprises EEG responses from 71 subjects listening to various speech stimuli. In total, it encompasses 508 trials from 71 subjects, utilizing 57 different stimuli, resulting in 7216 minutes (120 hours) of data. From the challenge training set, three subsets for model development were defined by a recording-wise stratified split with a ratio of 80/10/10, yielding train, validation and test subsets. The challenge test set comprises two distinct parts: "held-out stories", which contains the recordings from 70 subjects in the training set (seen subjects) for audio stimuli not in the training set, and "held-out subjects", which contains the recordings from 14 subjects not in the training set (unseen subjects) for audio stimuli in the training set [16].

Recognizing the potential benefit of additional training data, we modified the said development subsets as follows. The train and validation subset are combined into a our "training set", while the test subset is repurposed as our "validation set". Our "test set" used for model evaluation is the original challenge test set, containing both held-out subjects and held-out stories.

## III. MODELS

The proposed architecture draws on WaveNet [20] and its subsequent adaptations [21], [22]. While WaveNet was originally

[1] Code will made available in the following GitHub repository: https://github.com/LiuyinYang1101/Sea-Wave.



(a) Residual layer

(b) Architecture

**FIGURE 1. Original model architecture (Sub-Wave) with detail of a residual layer and the overall architecture, shown for $D = 3$ and $n_b = 3$.**

proposed for speech synthesis, it has since been successfully applied to various time series prediction tasks [23], [24], including the prediction of speech spectrograms from intracranial EEG [25]. Due to its dilated convolution structure, the model can effectively process information on a coarser scale compared to a regular or standard convolution [20]. The WaveNet architecture can be adapted by incorporating a non-causal bi-directional dilated convolution, as proposed by [21], turning Wavenet from a generative and slow model into a discriminative and parallelizable one, as used in DiffWave [22]. This adaptation yields a powerful regression model, that we utilize and adapt for "speech envelope approximation" (SEA) from EEG.

This section is further organized as follows. First, we present our submission to the Auditory-EEG challenge, introducing its WaveNet-based model architecture and discussing its shortcomings. Next, we describe the improved model architecture of Sea-Wave, highlighting the differences with our submission.

### A. MODEL ARCHITECTURE

The model architecture of the model submitted in the challenge, referred to as Sub-Wave in the remainder of this paper, is shown in Fig. 1.

The first layer is a channel-wise 1D convolution (Conv1D) layer and acts as a spatial filterbank that can compress or expand the multichannel EEG into $n_c$ channels. It is followed by a ReLU (Rectified Linear Unit) activation unit and, subsequently, by a stack of $n_l$ residual layers. Within each residual layer, we first encounter a bi-directional, dilated convolution (Bi-DilConv) layer, with a kernel of size $k$ and dilation factor $d$. The dilation factor $d$ increases exponentially with each residual layer, where the maximal dilation factor is controlled with a "dilation cycle" parameter $D$ such that $d_i = 2^{i \bmod D}$ for the $i$-th residual layer. This naturally organizes the residual layers into $n_b$ residual blocks, each consisting of $D$ layers with exponentially growing dilation factors. The dilated convolution is succeeded by a gated activation unit [26], computed as the product of two non-linear activation units (tanh$(\cdot)$ and $sig(\cdot)$). To allow the two activation units to operate independently, the Bi-DilConv layer doubles the number of channels and each unit operates on half of them, yielding the original $n_c$ channels after taking the element-wise product. The gated activation unit is followed by a dropout layer [27]. Next, two separate Conv1D layers are applied: one for the residual connection and another for the skip connection. The output of the residual connection is summed with the original input to the residual block and is propagated to the next block. Finally, all skip connection outputs are summed and scaled and two final Conv1D layers with a ReLU activation unit in between are applied. The final (and deepest) convolution layer combines all channels into a single channel, representing the reconstructed speech envelope. Weight normalization [28] is applied to all convolutional layers, except for the final one.

An important model attribute is the receptive field size, the region of the input space that directly influences the output. It can be computed as

$$rc_{size} = \sum_{i=0}^{n_l - 1} (k-1) \cdot 2^{i \bmod D} + 1, \tag{2}$$

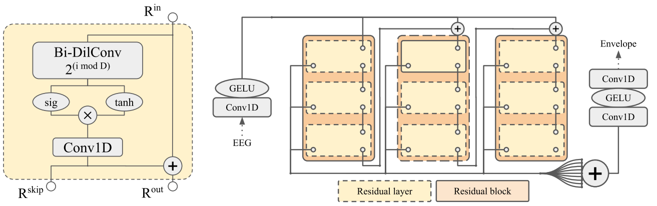
with $rc_{size}$ the receptive field size in time points, $n_l$ the number of residual layers, $D$ the dilation cycle, and $k$ the kernel size.

### B. CHALLENGE SUBMISSION

Sub-Wave, as submitted to the challenge, had 32 residual channels, 40 layers, and a dilation cycle of 7, yielding a receptive field size of 1333 time points (20.83 s). Despite its commendable performance, its architecture and training methodology raise certain concerns. Firstly, the final residual block does not finish a full dilation cycle, since the number of layers is not divisible by the dilation cycle. While mitigated through skip connections from all residual layers, this discrepancy may impact the model's performance and coherence. Secondly, the receptive field size is very large compared to linear model studies, e.g. 0.4 s in [9]. It thus appears implausible that such large receptive field size is necessary for obtaining a good performance. Thirdly, the residual layer components are derived from prior studies, but their significance in envelope reconstruction from EEG is unclear. Finally, the fine-tuning process, did not return significant improvements in the challenge results, prompting for further investigation into its effectiveness. Addressing these issues is critical to refine the model's architecture and optimize its training approach in order to further enhance performance in the Auditory-EEG challenge.

### C. IMPROVED MODEL ARCHITECTURE: SEA-WAVE

The model architecture of Sea-Wave, shown in Fig. 2, differs from Sub-Wave in three aspects. First, each ReLU is replaced by a Gaussian Error Linear Unit (GELU), an activation unit that combines properties from dropout and ReLUs and outperforms ReLUs on numerous datasets [29]. Second, the separate residual and skip Conv1D layers in the residual layers, as used in DiffWave [22], are replaced by a single Conv1D layer, as used in WaveNet [20]. Third, we add a skip connection that sums the spatially filtered EEG to the first layer of each residual block. We refer to this skip connection as an "input-skip" connection.



(a) Residual layer

(b) Architecture

**FIGURE 2. Improved model architecture (Sea-Wave) with detail of a residual layer and the overall architecture, shown for $D = 3$ and $n_b = 3$.**

**TABLE 1. Description of Model Variations Evaluated in the Ablation Study**

| Name | Description |
|---|---|
| Sub-Wave | Baseline model |
| Sub-Wave+ | Baseline model, trained for 20 additional epochs |
| NoGate | Replace gated with sigmoid activation unit |
| SingleConv | Replace separate Conv1D's with a single Conv1D for each residual layer |
| GELU | Replace ReLU with GELU activation unit |
| InSkip | Add input-skip connection for each residual block |
| GELU+InSkip | Combine GELU & InSkip adaptations |
| GELU+InSkip +Concat | Combine C & D and replace addition with concatenation for combining skip connections |

We decided on these model adaptations after our ablation study and hyperparameter tuning experiments.

## IV. EXPERIMENTS

### A. ABLATION STUDY

To attain a better understanding of the model components and their effects on performance, an ablation study is conducted on Sub-Wave, using the subject-independent model submitted during the challenge and described in [18] as the baseline. To this end, we make small adaptations to the baseline model by removing, replacing or introducing model components, while the remaining components are initialized with the weights and biases from the baseline. Each model variation underwent additional training with early stopping on the validation set or until reaching 20 iterations, and was evaluated on the test set. For a fair comparison, the baseline model was also subjected to the same way of using 20 additional iterations of training, referred to as Sub-Wave+. The description of all model variations can be found in Table 1. Note the input-skip connection facilitates the flow of information from the original data to the deep blocks, encouraging the model to refocus on the local information. This ensures that the model can still effectively capture local patterns and contextual information, even in deep layers with a large receptive field size. Meanwhile, since Sub-Wave aggregated all skip connection outputs by addition, we were also interested to see if concatenation could result in better results.

**TABLE 2. Grid Search Hyperparameter Values Per Kernel Size**

| Model Parameter | $k = 3$ | $k = 5$ |
|---|---|---|
| No. Channels | $8, 32, 64, 128$ | 32 |
| Dilation Cycle | 1 to 8 | 1 to 8 |
| No. Residual Blocks | 1 to 5 | 1 to 5 |

### B. HYPERPARAMETER TUNING

Utilizing Sea-Wave, we embark on a comprehensive hyperparameter search experiment through a grid search approach. During this experiment, we assess the effect of the receptive field size, the number of residual layers (model depth), and the number of trainable parameters (model size) on performance, as they are considered the most meaningful model attributes. To this end, we systematically varied the number of channels, dilation cycle, the number of dilation blocks, and kernel size, as seen in Table 2. To reduce the computational cost of the analysis, we evaluated $k = 5$ only for 32 channels.

During this experiment, we employed a modified training and validation split, motivated by two reasons. Firstly, we aimed to evaluate on both seen and unseen subjects without using a test set, as to avoid biasing the optimal hyperparameter choice. Secondly, we wanted to reduce the computational costs of the gird search by considerably reducing the size of the training set. Since our earlier findings [18] indicated that the large variability in performance over subjects can be attributed to the subject rather than to the model architecture, we constructed the modified split as follows. We ranked all subjects with Sub-Wave, grouped every 3 consecutive subjects, randomly selected 2 subjects from each group, and excluded their training data from the training set and included it in the validation set. Note that for seen subjects ($V_1$), the validation set was unchanged, while for the unseen subjects ($V_2$), we evaluated on the combined training and validation data. The final validation performance (score) was computed as the weighted sum of the average Pearson correlation for seen and unseen subjects, similar to (1).

### C. SUBJECT-SPECIFIC FINE-TUNING

In our effort to improve performance on the held-out stories, subject-specific fine-tuning is applied. This approach leverages subject-specific information to optimize the model predictions for each individual subject. During fine-tuning, the subject-independent model is utilized as a starting point, and further training is performed on a subject-by-subject basis. We empirically determined a training strategy that is effective in terms of training time and performance. Our training strategy consists of three steps:

*Subject-independent training:* During this initial phase, examples from the training set were cut using a window size of 5 seconds and a hop size of 0.5 seconds. The mean squared error (MSE) was used as the loss function for training. The dropout ratio was set to 0.2, and a learning rate of 0.0001 was chosen to facilitate rapid convergence. A cyclic learning rate scheduler [30] with a restart period of 10 iterations was applied to achieve better generalization performance. It took approximately 40–50 iterations for the model to reach a validation Pearson correlation above 0.15.

*Subject-specific fine-tuning:* In this step, examples from each subject's training set were cut using a larger window size of 8 seconds and a smaller hop size of 0.1 seconds. The loss function was $0.8 \cdot \text{MSE} + 0.2 \cdot$ Pearson r. A dropout ratio of 0.65 and a small learning rate of 0.00001 were chosen to prevent overfitting. Early stopping, based on the validation set, was used to select the optimal model. Typically, it took approximately 0–10 iterations for the model to reach the highest validation Pearson correlation.

*Subject-independent fine-tuning:* In order to improve model performance on unseen subjects, a similar fine-tuning approach as in the subject-specific fine-tuning step was employed, but now data from all subjects were used. This step may take up to 30 iterations to achieve the best performance.

In order to prevent overfitting and to ensure the best generalization performance, early stopping was applied during the fine-tuning step.

### D. MODEL PERFORMANCE AND GENERALIZATION

We evaluate the performance of Sea-Wave on the challenge test set, using subject-specific models for held-out stories ($S_1$) and the subject-independent model for the held-out subjects ($S_2$). We compare against the VLAAI model [10], which was provided as a challenge baseline model, and the HappyQuokka system [17], which won the Auditory-EEG challenge and achieved state-of-the-art performance on speech envelope reconstruction.

Additionally, we evaluate Sea-Wave on a subset of the publicly available DTU dataset [31]. This dataset is recorded using a similar EEG system as [14], but in an auditory attention paradigm. Also the stimulus characteristics differ, most notably, the language of the stimuli was Danish, rather than Dutch. We use the single-speaker subset of the recordings, which contains 18 subjects, each with 10 recordings of 50 s. From this subset, we select 2 recordings for subject-specific fine-tuning and keep the remainder as our "DTU test set". Note that, the data is preprocessed as described above.

### E. MODEL INTERPRETATION

Model interpretability is a crucial aspect of this paper, and it adds an essential dimension to the evaluation of the proposed Sea-Wave model. Interpretable models provide insights into how they arrive at their predictions, making it easier for researchers and practitioners to understand the model's decision-making process and trust its outputs. We inspected our subject-independent model in the following three ways.

We visualize the spatial filters learned by the input Conv1D layer, we estimate the filter and channel importance to visualize the most salient ones, and we inspect layer activations during inference on unseen stimuli for each layer.

To estimate filter and channel importance, we performed a zero-one-out and leave-one-in analysis. During the zero-one-out analysis, individual filters or channel inputs were zeroed out, and the resulting performance degradation on the test set was utilized to rank all filters or channels. For the leave-one-in analysis, a single filter or channel was utilized, and the resulting (residual) performance on the test set was utilized.

When inspecting layer activations, we picked one of the highest Pearson correlation test files and one of the worst. We visualize the activations for each layer, yielding insight into how the model transforms EEG into a speech envelope.

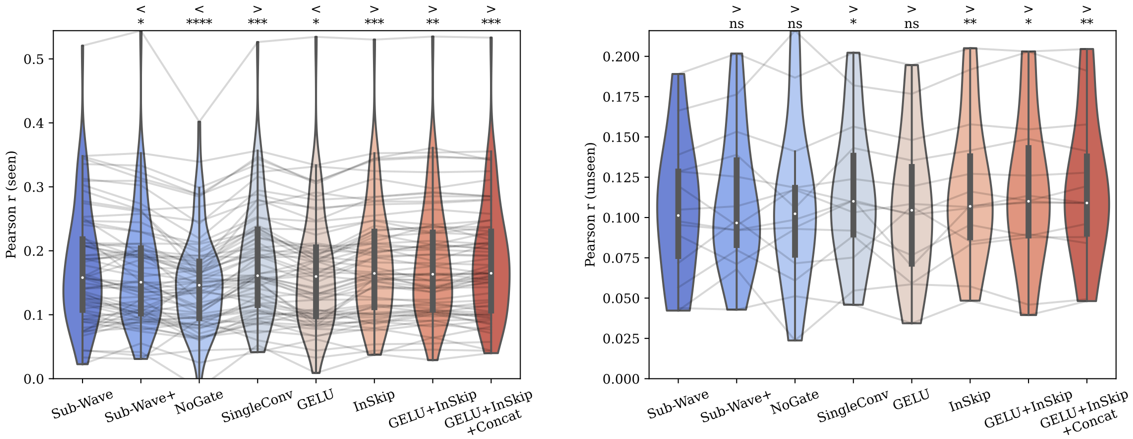
## V. RESULTS

### A. ABLATION STUDY

The results of the ablation study, evaluated on the held-out stories and held-out subjects, are depicted in Fig. 3. A paired t-test with Holm–Bonferroni correction was performed between the baseline and each model variation to assess whether the variant performed better or worse ($p < 0.05$). The asterisk positioned atop the violin plot signifies statistical significance (n.s.: $p \geq 0.05$, $\ast$: $0.01 \leq p \leq 0.05$, $\ast\ast$: $0.05 \leq p \leq 0.01$, $\ast\ast\ast$: $0.01 \leq p \leq 0.001$, $\ast\ast\ast\ast$: $p \leq 0.0001$). The $>$ and $<$ atop the asterisk denote whether the mean performance is above or below the baseline. Each grey line connects performance of the different model variations for a single subject.

Several key observations can be made:

1) The gated activation unit emerges as the most critical component, as its ablation leads to a significant drop in performance, particularly for the seen subjects. This indicates its pivotal role in modeling the envelope and its relation to the EEG. However, it remains unclear whether it captures additional information, or whether its regularizing properties are simply beneficial for modeling acoustic features, as the gated activation unit also works significantly better than the ReLU for modeling audio signals [20] in an auto-regressive setting.
2) The single Conv1D, as used in WaveNet, yields significant improvements compared to two separate Conv1D operations, as in DiffWave. Hence, Sea-Wave employs a single Conv1D.
3) The inclusion of an input-skip connection proves beneficial, particularly for unseen subjects. The input-skip connection enables a better flow of information from the input data to the deep blocks, contributing to an improved generalization of the deep residual layers. Therefore, Sea-Wave employs an input-skip connection.
4) Both addition and concatenation appear effective for information fusion. This is confirmed by an additional paired t-test between the "GELU+InSkip" and "GELU+InSkipConcat" models ($p = 0.3196$). As concatenation requires additional parameters in the last two convolutional layers, Sea-Wave employs addition.
5) Across the various model adaptations, the subject performances remain quite stable relative to each other. This indicates most of the subject variability can be



(a) Held-out stories ($S_1$)

(b) Held-out subjects ($S_2$)

**FIGURE 3. Comparison of performance for model variations evaluated in the ablation study. Performance is evaluated as the Pearson r on the test set for seen ($S_1$) and unseen ($S_2$) subjects. Asterisks show the significance level of a paired two-sided t-test against the baseline (Sub-Wave), symbols > and < indicate whether the mean was higher or lower compared to the baseline, respectively.**

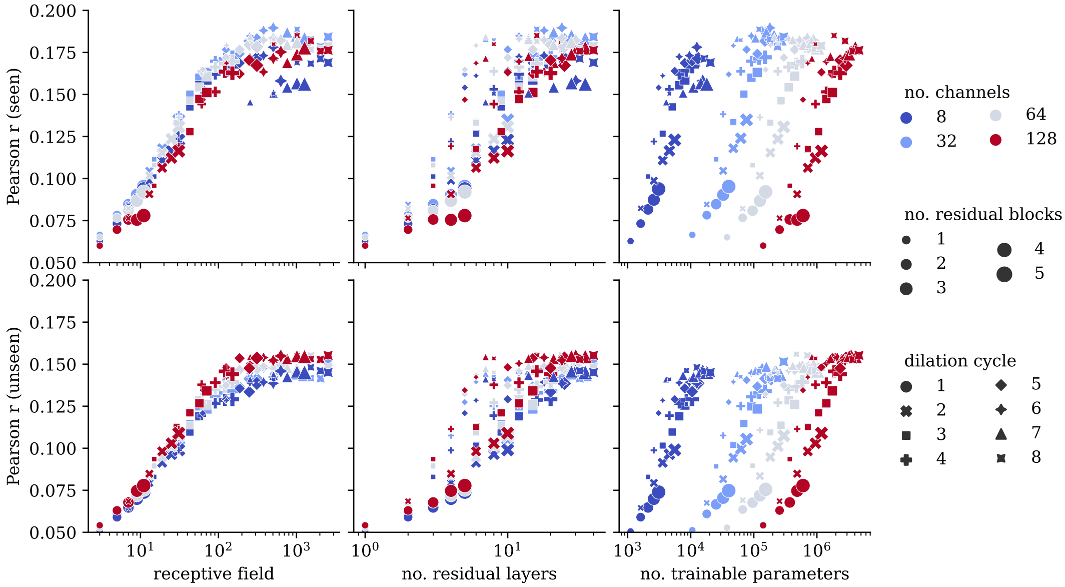


**FIGURE 4. Performance as a function of the receptive field size, the number of residual blocks and the number of trainable parameters, given $k = 3$, during hyperparameter tuning of Sea-Wave. Performance is evaluated as the Pearson r on the validation set for seen ($V_1$) and unseen ($V_2$) subjects.**

attributed to the subject's EEG recordings rather than the model architecture.

### B. HYPERPARAMETER TUNING

Fig. 4 shows the performance in function of the receptive field size, the number of residual layers (model depth), and the number of trainable parameters (model size), for each combination of hyperparameters, given kernel size $k = 3$. A visual inspection indicates the 32 channel models are superior for seen subjects, while 128 channel models generalize slightly better to unseen subjects. It also shows a clear relationship between receptive field size and performance.

To quantify the relationship between the performance and these three model attributes, we fitted a sigmoidal curve for each model parameter, as described in (A1). We estimate a threshold $x^{th}$ above which performance is stable by calculating the intercept of the slope at the inflection point of the sigmoidal curve and its upper asymptote, as described in (A2). Fig. 5 shows the fitted sigmoidal curves and Table 3 presents the threshold and the coëfficient of determination ($R2$) for each fitted curve. Only for receptive field size, we obtained a good fit ($R^2 = 99.81$) and thus a reliable threshold estimate. The threshold is around 100 samples (±1.56 sec), and thus models with smaller receptive field sizes can be considered

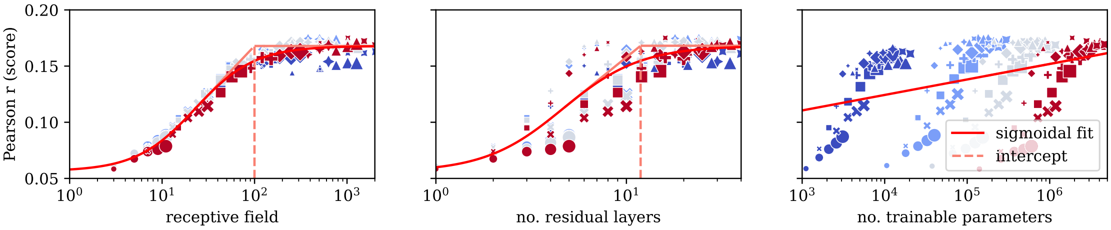


**FIGURE 5.** Sigmoidal curve fit for performance (score) with receptive field size, the number of residual blocks and the number of trainable parameters, given $k = 3$.

**TABLE 3.** Threshold and Coëfficient of Determination ($R^2$) for Each Sigmoidal Curve Fit

| Model Attribute | $x^{th}$ | $R^2$ |
|---|---|---|
| Receptive Field | 99.81 | 0.97 |
| No. Residual Layers | 11.87 | 0.79 |
| No. Trainable Parameters | NaN | 0.14 |

**TABLE 4.** Comparison of Kernel Size 3 and 5, Given $n_c = 32$ and $rc_{size} \geq 100$

| Paired on | Sample size | $p$ |
|---|---|---|
| Dilation Cycle and No. Residual Blocks | 21 | 0.4288 |
| Receptive Field Size | 9 | 0.2779 |
| Approx. # Trainable Parameters | 7 | 0.6364 |

sub-optimal, while those with large receptive fields tend to be overparameterized.

To assess whether a larger kernel is beneficial, we compare the performance (score) for models with a kernel size 3 and 5 with a paired t-test, Holm-Bonferroni corrected, given 32 channels and a receptive field size larger than 100. The models are paired on the remaining hyperparameters, dilation cycle and number of residual blocks. As this comparison does not factor in the increased receptive field and model size associated with a kernel size of 5, we proceed with two additional comparisons. The first compares only models with equal receptive field sizes, the second compares only models with approximately equal model sizes (rounded to 2 significant numbers). Note that both additional comparisons have a lower sample size, as unpaired models are discarded. The results of the paired t-tests, shown in Table 4, indicate a larger kernel size is not significantly better regardless of the scenario.

However, upon repeating this analysis for seen and unseen subjects separately, shown in Table 5, we find a kernel size of 5 to be significantly better for unseen subjects. This also holds for models with equal receptive field sizes. This indicates that a faster growth of the receptive field, a consequence of the larger kernel, may be beneficial for generalization.

In the above comparisons we ensured a receptive field size larger than 100, excluding many sub-optimal models. Now, we compare the 3 best models per hyperparameter value, ensuring only the best models for a given hyperparameter value are included. We perform a paired t-test for all combinations of values per hyperparameter, presented in Table 6. For the number of channels, we find that $n_c = 8$ is significantly worse than all other choices. Interestingly, increasing the number of channels from 32 to 64 significantly worsens performance, indicating that either 32 or 128 is optimal. For the number of residual blocks, there is no optimal choice, as differences are insignificant. For the dilation cycle, we find that $D = 4$ and $D = 5$ are significantly worse than all other choices.

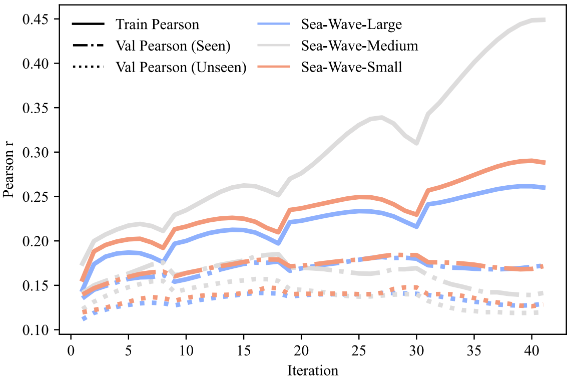


**FIGURE 6.** Evolution of Pearson r during training iterations.

Based on these results, we pick 3 Sea-Wave models for further experimentation and evaluation, see Table 7. Since in the challenge evaluation the unseen subjects are weighted less, we stick to a kernel size of 3 for all three models. To examine differences in model convergence, we shown the evolution of the Pearson correlation during training for these three models in Fig. 6.

### C. SUBJECT-SPECIFIC FINE-TUNING

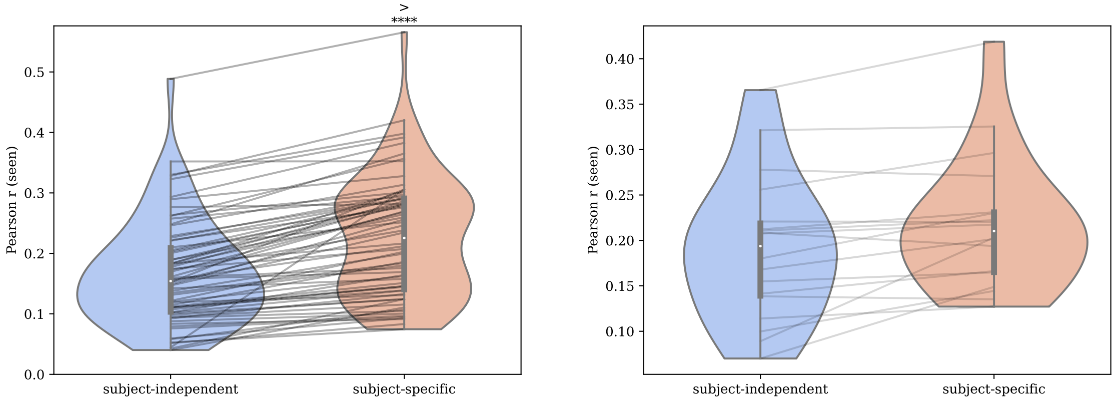
Fig. 7(a) shows the effect of subject-specific fine-tuning on performance. It is clear the approach yields large performance gains, as it raises the mean Pearson r on seen subjects ($S_1$) from 16.77% to 22.84%, confirming previous studies on subject-specific fine-tuning, e.g. [10].

### D. MODEL PERFORMANCE AND GENERALIZATION

Table 8 presents the performance on the challenge test set of the VLAAI model [10], the challenge winning model (HappyQuokka) [17], our runner-up Sub-Wave model, and the

**TABLE 5.** Comparison of Performance (score) for Kernel Size 3 and 5, Given $rc_{size} \geq 100$

| | | $p$ | | mean Pearson r | | | |
|---|---|---|---|---|---|---|---|
| | | | | seen | | unseen | |
| Paired on | Sample size | seen | unseen | $k = 3$ | $k = 5$ | $k = 3$ | $k = 5$ |
| Dilation Cycle and No. Residual Blocks | 21 | 1 | 0.02161 | 0.1819 | 0.1808 | 0.1427 | **0.1457** |
| Receptive Field Size | 9 | 1 | 0.01548 | 0.1822 | 0.1818 | 0.1422 | **0.1458** |
| Approx. # Trainable Parameters | 7 | 1 | 1 | 0.1816 | 0.179 | 0.1429 | 0.1444 |



(a) Held-out stories ($S_1$)

(b) Held-out stimuli (DTU test set)

**FIGURE 7.** Comparison of performance for Sea-Wave-Small before and after subject-specific fine-tuning. Performance is evaluated as Pearson r on held-out stories ($S_1$) and held-out stimuli (DTU test set).

**TABLE 6.** Comparison of Performance (score) for the 3 Best Performing Models Per Hyperparameter Value

| Hyperparameter | | $p$ |
|---|---|---|
| No. Channels | 8 vs. 32 | 0.001622 |
| | 8 vs. 64 | 0.004238 |
| | 8 vs. 128 | 0.01189 |
| | 64 vs. 32 | 0.0242 |
| No. Residual Blocks | x | x |
| Dilation Cycle | 4 vs. 5 | 0.0001872 |
| | 4 vs. 6 | 0.0008465 |
| | 4 vs. 7 | 0.0007596 |
| | 4 vs. 8 | 0.0001546 |
| | 5 vs. 6 | 0.03871 |
| | 5 vs. 7 | 0.03871 |
| | 5 vs. 8 | 0.003543 |

Paired T-Test with Holm-Bonferroni correction, only P-Values below 0.05 are shown.

**TABLE 7.** Hyperparameter Values for Evaluated Models and Corresponding Receptive Field Size, Number of Trainable Parameters and Computational Complexity (Expressed in FLOPs)

| Model | $n_c$ | $n_l$ | $D$ | $rc_{size}$ | no. params | no. FLOPs |
|---|---|---|---|---|---|---|
| Sub-Wave | 32 | 40 | 7 | 1333 | 341k | 54M |
| Sea-Wave-Large | 128 | 16 | 8 | 1021 | 1873k | 299M |
| Sea-Wave-Medium | 32 | 40 | 5 | 497 | 298k | 47M |
| Sea-Wave-Small | 32 | 20 | 5 | 249 | 150k | 24M |

**TABLE 8.** Comparison of Performance for Different Models on the Challenge Test Set

| Model | Held-out Stories | Held-out Subjects | Score |
|---|---|---|---|
| VLAAI | On average: 0.1614 | | |
| HappyQuokka | 0.1895±0.0869 | 0.0976±0.0444 | 0.1589 |
| Sub-Wave | 0.1741±0.0913 | 0.1123±0.0447 | 0.1535 |
| Sea-Wave-Large | 0.1916±0.0968 | 0.1086±0.0454 | 0.1639 |
| Sea-Wave-Medium | 0.2112±0.1039 | 0.1046±0.0489 | 0.1754 |
| Sea-Wave-Small | 0.2258±0.0985 | 0.1158±0.0410 | 0.1891 |

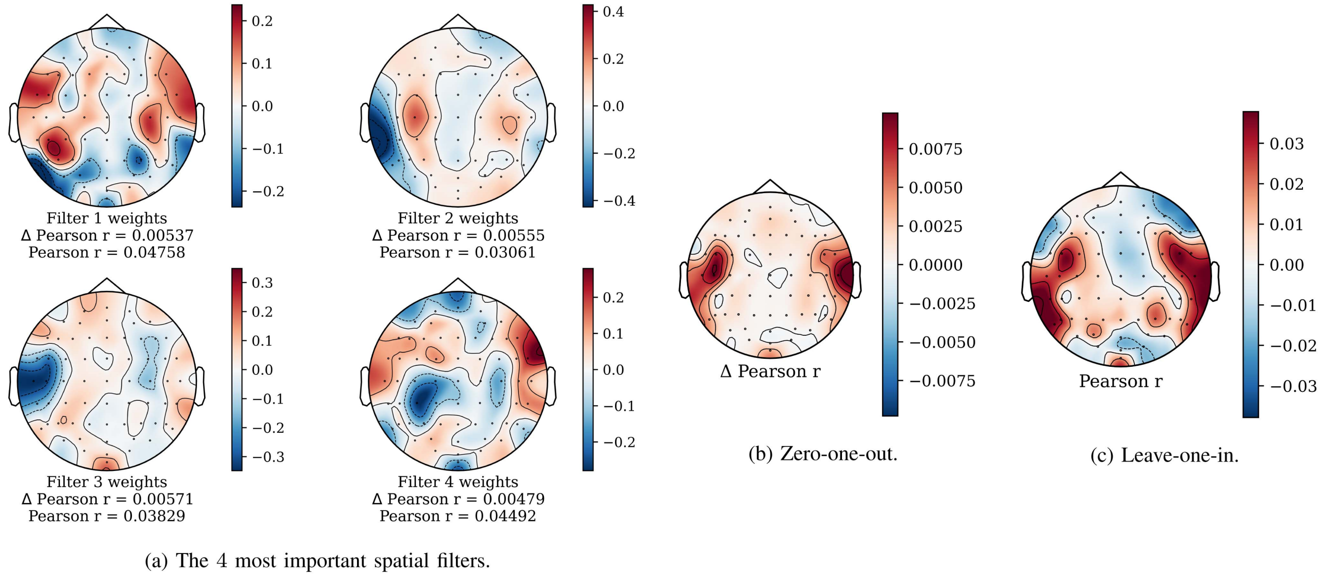
three selected Sea-Wave models (see Table 7). The performance for held-out stories ($S_1$) and held-out subjects ($S_2$) is also included. In Table 10, we report the Pearson correlations per subject using the subject-independent, subject-specific, and randomly initialized Sea-Wave-Small models.

We also evaluated the subject-independent Sea-Wave-Small model on the DTU test set, to test how well the model generalizes to another dataset. A comparison before and after subject-specific fine-tuning is shown in Fig. 7(b).

## E. MODEL INTERPRETATION

We use a subject-independent Sea-Wave-Small for all visualizations, as it is the smallest and best performing model.

*Topographic map of spatial patterns:* As this model has 32 channels, we can extract 32 spatial filters from the Conv1D input layer. To find the most important filters, we rank these



(a) The 4 most important spatial filters. (b) Zero-one-out. (c) Leave-one-in.

**FIGURE 8.** Topographic map of the four most important spatial filters and the channel importance estimated with zero-one-out, and leave-one-in analysis.

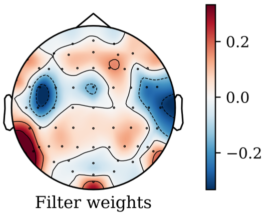


**FIGURE 9.** Topographic map of the spatial filter of a subject-independent single-channel Sea-Wave-Small.

filters based on a zero-one-out and leave-one-in analysis. Filters that appear in the top-8 for both analyses are selected for visualization. Fig. 8(a) shows the resulting 4 filters and displays their filter weights on a topographic map. A complete overview of the ranked filters for zero-one-out and leave-one-in analysis is given in Fig. 12(a) and (b), respectively.

*Topographic map of channel importance:* Fig. 8 shows the channel importance, estimated through zero-one-out and leave-one-in, on a topographic map. In Fig. 8(b), the performance drop (Δ Pearson r) contributed to the zeroed channel is shown, while Fig. 8(c) shows the residual performance (Pearson r) when all other channels are zeroed. This analysis provides insights into the relevance of each input channel and its contribution to the model's overall performance.

*Layer activation patterns:* In Fig. 10(a), the first and last layer activations for the best performing subject and stimulus pair (subject 44, test file 02, Pearson r of 0.532) are shown. Conversely, Fig. 10(b) shows these for the worst performing subject and stimulus pair (subject 63, test file 02, −0.026). Next, we evaluate the contributions of different skip connections by passing them individually through the subsequent channel-wise convolution layers, instead of passing their scaled sum, using the same subjects and test file. Fig. 11(a) and (b) show the reconstructed envelope for each skip connection, with layer 0 and layer 19 representing the first and the last skip connection. The black line represents the actual speech envelope in both figures.

## VI. DISCUSSION

### A. ABLATION STUDY

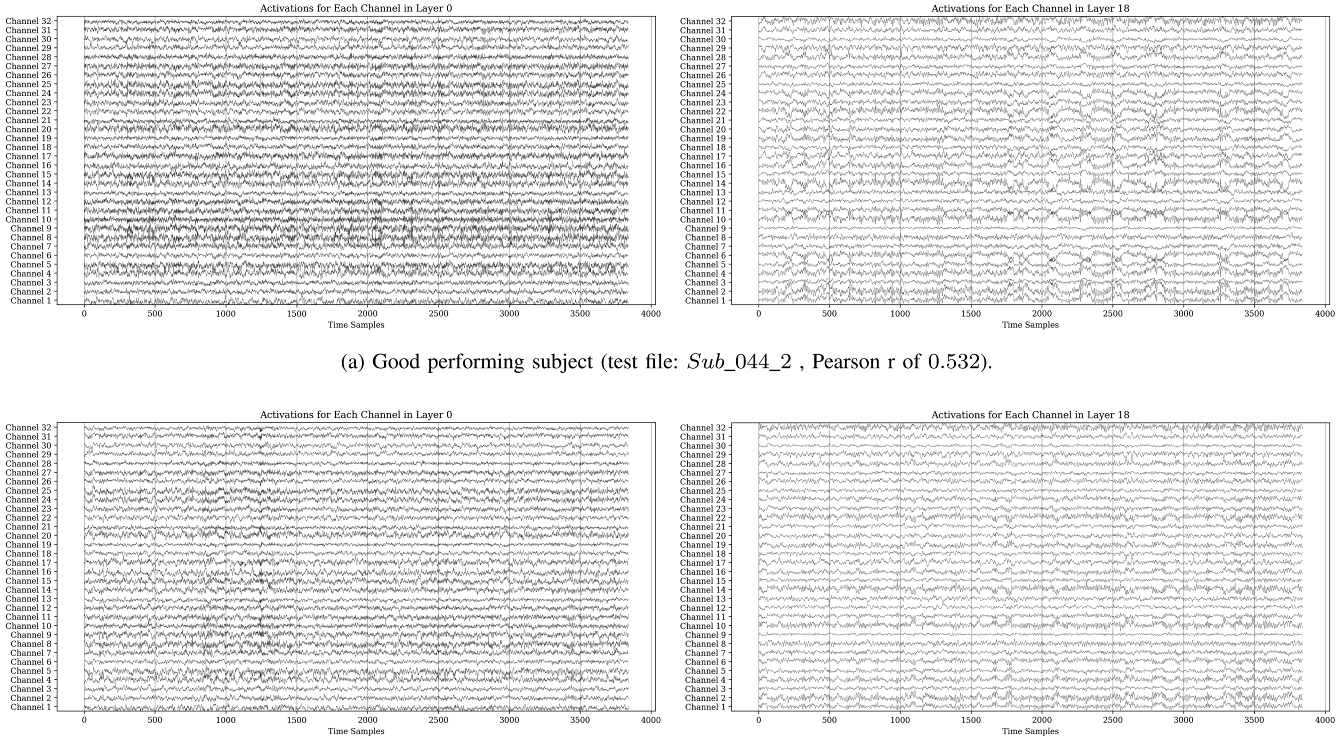
The insights gained from the ablation study have played a crucial role in the development of the Sea-Wave model. Leveraging these findings, specific architectural choices have been made to optimize the model's performance in speech envelope reconstruction from auditory EEG recordings.

In particular, the Sea-Wave model excludes the separate skip Conv1D operation within the residual layer, as this component was found to have a limited impact on model performance. Instead, the model incorporates the new skip connection with GELU activation units, which were identified as critical components in achieving better results.

Regarding information fusion, the model uses addition as the preferred method due to its ability to yield a smaller number of trainable parameters while still maintaining competitive performance.

### B. HYPERPARAMETER TUNING

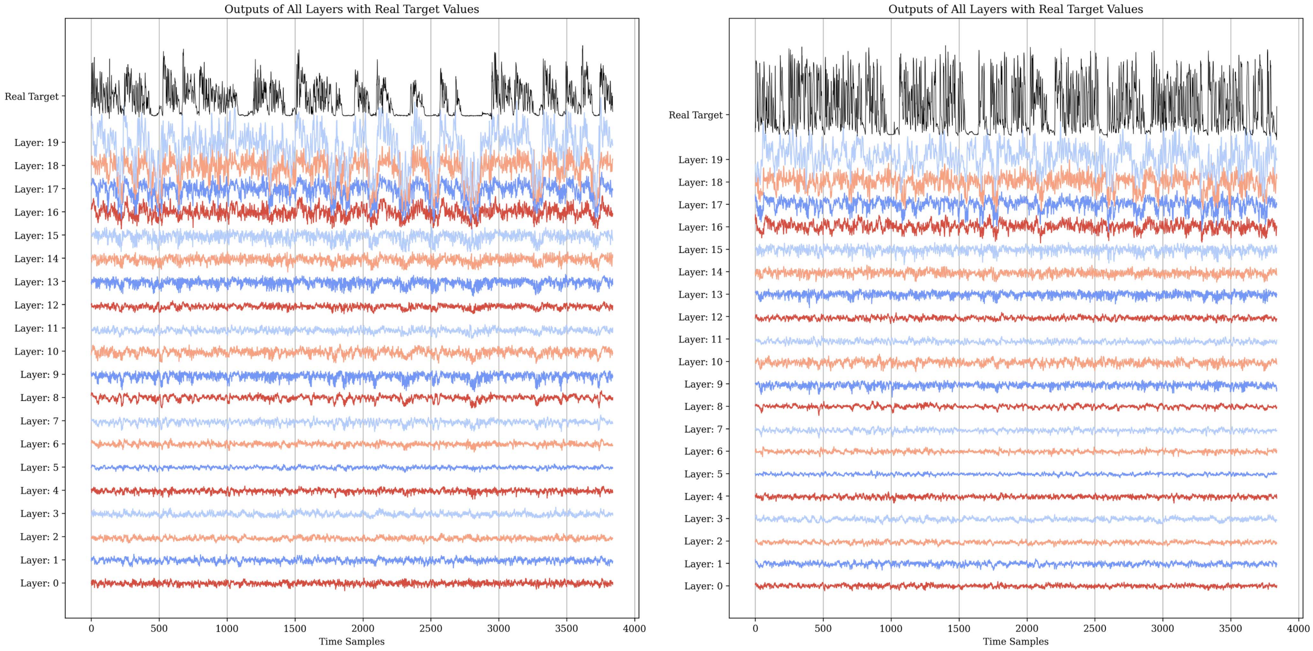
Our results suggest that receptive field size is the most important model attribute to consider. This is indicated by the high coëfficient of determination of the sigmoidal curve fit. Surprisingly, good performance can be reached for nearly all model depths and sizes, given an appropriate receptive field size. For the other model attributes, the inferior sigmoid curve fit prevents drawing strong conclusions. Note that the apparent



(a) Good performing subject (test file: $Sub\_044\_2$ , Pearson r of 0.532).

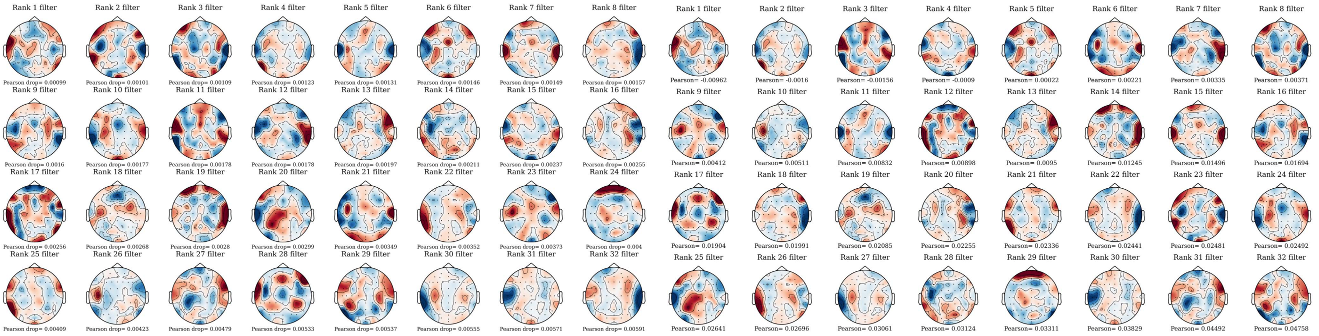
(b) Bad performing subject (test file: $Sub\_063\_2$, Pearson r of $-0.026$).

**FIGURE 10.** **Activation patterns from the first and last layer for a good and bad performing subject.**



(a) Good performing subject (test file: $Sub\_044\_2$ , Pearson r of 0.532). (b) Bad performing subject (test file: $Sub\_063\_2$, Pearson r of $-0.026$).

**FIGURE 11.** **Outputs using a single layer for a good and bad performing subject. The color scheme completes one cycle for each residual block.**



(a) Zero-one-out (Δ Pearson r).

(b) Leave-one-in (Pearson r).

**FIGURE 12.** Ranked 32 spatial patterns using zero-one-out and and leave-one-in analysis.

threshold of 12 residual layers may be better explained in terms of receptive field size, as shallower models have smaller receptive field sizes. With regard to kernel sizes 3 and 5, they appear equivalent for seen subjects, but the latter can be beneficial for generalizing to unseen subjects.

For all models, a single iteration yields a model performance on par with that of linear models, with subsequent iterations slowly improving it further. However, some architectures would benefit from tuning the learning rate and number of training iterations. This is evidenced by Sea-Wave-Medium, which overfitted faster than the other two models, arguably because of its depth (40 layers), as shown in Fig. 6. Furthermore, we also found that the models with 8 channels had not converged fully, potentially contributing to the low performance compared to models with more channels.

### C. SUBJECT-SPECIFIC FINE-TUNING

For a few subjects, performance did not improve. This phenomenon may be attributed to overfitting, where the model becomes too tailored to the specific characteristics of listening to certain training speech audio files during the fine-tuning process. Consequently, when evaluated on unseen data or held-out stories, these subjects' performance may not generalize as effectively.

### D. MODEL PERFORMANCE AND GENERALIZATION

From Table 8, we observe all three Sea-Wave models outperformed the state-of-art HappyQuokka model, especially in terms of held-out stories. The effectiveness of the subject-specific fine-tuning strategy is evident in the improved performance of all three Sea-Wave models. Interestingly, the small model with fewer parameters, achieved the best performance among the Sea-Wave variants. This observation aligns with the idea that wider (128-channel) and deeper models may be more prone to overfitting. The smaller model strikes a better balance between model complexity and the available data, allowing it to generalize more effectively and avoid overfitting issues that may be encountered with larger and more complex architectures. Furthermore, it proves that having a more focused view of local patterns is more helpful, instead of having a large receptive field, when models attempt to incorporate information from distant points.

**TABLE 9.** Sigmoidal Curve Fit: Parameter Values

| Model Attribute | k | L | x0 | b |
|---|---|---|---|---|
| Receptive Field | 0.11 | 3.13 | 1.36 | 0.06 |
| No. Residual Layers | 0.11 | 1.53 | 2.12 | 0.06 |
| No. Trainable Parameters | 8.69 | 12.03 | 0.00 | −4.21 |

The model also performs well on the DTU test set and outperforms the performance reported in [10]. Through subject-specific fine-tuning, significant improvement could be achieved by using very few training examples (2 files corresponding to 100 s of speech).

### E. MODEL INTERPRETATION

The foremost four spatial patterns along with the single-channel pattern exhibit notable similarities. Specifically, post-training, these patterns display clustered positive and negative weights in regions corresponding to the auditory cortex, the vision cortex (Oz), and somatic sensory associated areas (P3 and P5). Filter 1 manifests a symmetric pattern, while Filters 2 and 3 demonstrate prominent negativity in the left auditory cortex. Conversely, Filter 4 exhibits positivity in the left auditory cortex. Consequently, we postulate that these distinct filters capture and model diverse interactions among areas, collectively contributing to the final predictions.

A similar pattern emerges for both zero-one-out and leave-one-out channel-importance analysis, where the auditory cortex area is most important. This is consistent with prior studies [32], as well as the spatial patterns, conforming the importance of the audio cortex in auditory EEG decoding. Meanwhile, we also found some positively correlated channels around the vision cortex and somatic sensory-associated areas. The presence of positively correlated channels in the vision cortex and somatic sensory associated areas indicates a potential cross-modal interaction between auditory, somatosensory processing, and visual processing. In previous studies, cross-modal interactions between different sensory modalities have been observed. For instance, researchers

**TABLE 10.** Performance (score) on the Test Set for Each Subject, Using a Subject-Independent, a Subject-Specific, and a Randomly Initialized Sea-Wave-Small

| Subject | Subject-independent | Subject-specific | Random | Subject | Subject-independent | Subject-specific | Random |
|---|---|---|---|---|---|---|---|
| sub-002 | 0.1896±0.0592 | 0.2596±0.0752 | –0.006±0.0369 | sub-003 | 0.2932±0.0819 | 0.3828±0.0651 | –0.0045±0.0387 |
| sub-004 | 0.1577±0.0561 | 0.1691±0.0288 | –0.0048±0.0236 | sub-005 | 0.2287±0.0484 | 0.291±0.0327 | 0.0044±0.0295 |
| sub-006 | 0.2634±0.0451 | 0.301±0.0622 | 0.0005±0.0382 | sub-007 | 0.1806±0.0645 | 0.2649±0.0404 | –0.0033±0.039 |
| sub-008 | 0.248±0.0834 | 0.3651±0.0588 | –0.0009±0.026 | sub-009 | 0.3299±0.0565 | 0.3981±0.0343 | 0.0012±0.0447 |
| sub-010 | 0.0938±0.0607 | 0.1047±0.0626 | 0.0028±0.0287 | sub-011 | 0.1342±0.0318 | 0.1841±0.0642 | –0.0022±0.0294 |
| sub-012 | 0.0402±0.0558 | 0.1106±0.0572 | –0.0028±0.0195 | sub-013 | 0.2462±0.06 | 0.3136±0.0693 | 0.0022±0.0247 |
| sub-014 | 0.2219±0.0399 | 0.2761±0.0519 | 0.0017±0.0243 | sub-015 | 0.1107±0.0482 | 0.146±0.0464 | –0.0043±0.0269 |
| sub-016 | 0.1523±0.1242 | 0.2466±0.0885 | 0.0004±0.0337 | sub-017 | 0.1561±0.0349 | 0.2121±0.0613 | 0.0007±0.0348 |
| sub-018 | 0.0516±0.0625 | 0.0832±0.0751 | 0.0023±0.029 | sub-019 | 0.1215±0.048 | 0.2028±0.0893 | 0.0043±0.0255 |
| sub-020 | 0.2003±0.0631 | 0.2825±0.0807 | –0.0037±0.0495 | sub-021 | 0.1199±0.0614 | 0.1625±0.0456 | –0.001±0.0235 |
| sub-022 | 0.1597±0.0322 | 0.1759±0.0721 | –0.0013±0.0434 | sub-023 | 0.1612±0.0649 | 0.2696±0.0752 | 0.0106±0.0296 |
| sub-024 | 0.1119±0.0519 | 0.1512±0.0656 | –0.0041±0.0304 | sub-025 | 0.203±0.0441 | 0.2806±0.0362 | –0.0029±0.0371 |
| sub-026 | 0.1377±0.0294 | 0.1997±0.0532 | 0.0047±0.0303 | sub-027 | 0.1312±0.0473 | 0.1803±0.037 | –0.0032±0.0292 |
| sub-028 | 0.1074±0.0455 | 0.1459±0.0366 | –0.0025±0.0298 | sub-029 | 0.1101±0.0677 | 0.1445±0.0302 | 0.0034±0.0287 |
| sub-030 | 0.0878±0.0487 | 0.1004±0.0541 | –0.0004±0.0295 | sub-031 | 0.1742±0.0552 | 0.2073±0.0444 | 0.0009±0.0365 |
| sub-032 | 0.352±0.0565 | 0.3523±0.0556 | 0.0041±0.0335 | sub-033 | 0.2208±0.0424 | 0.29±0.0399 | –0.0011±0.0279 |
| sub-034 | 0.1773±0.0331 | 0.2811±0.0464 | –0.0059±0.0278 | sub-035 | 0.1469±0.0598 | 0.2345±0.0579 | –0.0059±0.0353 |
| sub-036 | 0.2765±0.0629 | 0.2819±0.0463 | –0.0006±0.0409 | sub-037 | 0.1831±0.0507 | 0.2164±0.0505 | –0.003±0.0341 |
| sub-038 | 0.21±0.0495 | 0.2517±0.0518 | –0.0064±0.0288 | sub-039 | 0.1825±0.0743 | 0.3034±0.0727 | –0.0033±0.0258 |
| sub-040 | 0.2048±0.1062 | 0.292±0.0714 | –0.0059±0.0357 | sub-041 | 0.1831±0.0541 | 0.2863±0.0703 | 0.0085±0.0383 |
| sub-042 | 0.3227±0.0554 | 0.3918±0.0545 | –0.0017±0.0376 | sub-043 | 0.3282±0.0323 | 0.4201±0.0379 | –0.0159±0.0533 |
| sub-044 | 0.4885±0.0377 | 0.5657±0.0328 | 0.0047±0.0644 | sub-045 | 0.2574±0.0401 | 0.2942±0.0627 | 0.0061±0.0304 |
| sub-046 | 0.152±0.0343 | 0.3057±0.0642 | 0.0044±0.0393 | sub-047 | 0.2618±0.0551 | 0.3566±0.0809 | –0.0009±0.0401 |
| sub-048 | 0.2895±0.0457 | 0.3279±0.016 | –0.0031±0.0476 | sub-049 | 0.1144±0.0856 | 0.1861±0.0959 | 0.003±0.0328 |
| sub-050 | 0.0445±0.013 | 0.3059±0.0543 | –0.0034±0.0377 | sub-051 | 0.0581±0.0527 | 0.1072±0.1007 | 0.0009±0.0438 |
| sub-052 | 0.1696±0.0701 | 0.2686±0.0527 | –0.0014±0.0269 | sub-053 | 0.0981±0.0617 | 0.1249±0.0316 | 0.0031±0.0282 |
| sub-054 | 0.0426±0.073 | 0.1244±0.0866 | 0.0029±0.0317 | sub-055 | 0.1422±0.1052 | 0.1578±0.1056 | –0.0041±0.0245 |
| sub-056 | 0.1484±0.0595 | 0.2562±0.0344 | 0.0006±0.0328 | sub-057 | 0.2259±0.0713 | 0.2975±0.051 | –0.0053±0.0367 |
| sub-058 | 0.1889±0.0403 | 0.2788±0.022 | 0.0001±0.0209 | sub-059 | 0.1728±0.0462 | 0.2382±0.0375 | 0.0004±0.038 |
| sub-060 | 0.102±0.0389 | 0.1386±0.0335 | –0.0011±0.0262 | sub-061 | 0.0925±0.0654 | 0.1328±0.0636 | –0.0008±0.0275 |
| sub-062 | 0.1023±0.0845 | 0.1643±0.0366 | –0.0026±0.0362 | sub-063 | 0.0527±0.0549 | 0.0746±0.0902 | –0.0018±0.0299 |
| sub-064 | 0.0791±0.0649 | 0.0967±0.0801 | 0.0027±0.0286 | sub-065 | 0.1006±0.0683 | 0.1241±0.1169 | –0.0005±0.0288 |
| sub-066 | 0.0949±0.0458 | 0.1155±0.0321 | 0.0035±0.0288 | sub-067 | 0.0595±0.0682 | 0.093±0.0528 | –0.0006±0.0333 |
| sub-068 | 0.0759±0.0585 | 0.0947±0.0347 | –0.002±0.0314 | sub-069 | 0.0836±0.0579 | 0.0914±0.0435 | 0.0014±0.0297 |
| sub-070 | 0.1191±0.0205 | 0.131±0.0416 | –0.0011±0.0281 | sub-071 | 0.1105±0.06 | 0.139±0.0623 | 0.0029±0.0398 |
| sub-072 | 0.1795±0.0684 | NA | –0.0041±0.0317 | sub-073 | 0.0929±0.0598 | NA | 0.0025±0.0362 |
| sub-074 | 0.1457±0.0651 | NA | –0.0045±0.0298 | sub-075 | 0.1935±0.071 | NA | –0.0022±0.0333 |
| sub-076 | 0.0537±0.0385 | NA | 0.0048±0.0332 | sub-077 | 0.0767±0.0661 | NA | 0.0041±0.0392 |
| sub-078 | 0.1085±0.0737 | NA | 0.0017±0.0305 | sub-079 | 0.0935±0.0511 | NA | 0.0026±0.0347 |
| sub-080 | 0.1251±0.0827 | NA | 0.0027±0.0318 | sub-081 | 0.0509±0.0536 | NA | 0.0007±0.0343 |
| sub-082 | 0.105±0.0796 | NA | 0.0022±0.0341 | sub-083 | 0.089±0.0642 | NA | –0.001±0.0256 |
| sub-084 | 0.1376±0.0891 | NA | 0.0012±0.0392 | sub-085 | 0.1119±0.0673 | NA | 0.0031±0.0349 |

Note subject-specific models are only reported for seen subjects ($S_1$).

in [33] discovered that listening to sounds can influence the visual dorsal pathway, enhancing attention and memory for objects' locations. Similarly, in another study [34], an interaction between auditory and somatosensory processing was reported. This audio-tactile cross-modal interaction was found to expedite initial cortical activity in the human somatosensory cortex, indicating that such interactions occur during the early stages of cortical processing. This faster processing in the sensory cortex could potentially contribute to shorter reaction times under multisensory integration. These findings highlight the intricate connections between different sensory systems and how they mutually influence each other to enhance cognitive processing.

Comparing the activations of the first and last layers for both subject 44 and subject 63, several observations can be made:

1) The activations in the first layer displayed characteristics resembling raw EEG signals, with prominent high-frequency components. However, as the signal propagated through the subsequent layers, the activations appeared much smoother. This suggests that the model acts as a low-pass filter, similar to findings in linear model studies.
2) The observation that many residual channels share similar activations or exhibit counter-phase information suggests that redundant information is present across

multiple channels, and the use of a small number of channels may suffice to capture the essential features for auditory EEG decoding.

3) From the model predictions using single-layer outputs, it is clear that the deep layer's activations contribute more to the final envelope prediction. As shallow layers feature smaller amplitudes and shared little envelope-like features, while the deep layers, especially the last layer, captured the dynamics of the envelopes.
4) Comparing the two single-layer output plots, we observe that the model seems to be good at modeling the silence periods. This is evident from the clear drops in outputs observed in the deep layers during these periods. The model's ability to accurately capture and monitor the periods of silence or low activity reflects its proficiency in detecting and representing calm or non-speech segments. On the other hand, in Fig. 11(b), the subject for which the Pearson correlation was the worst. The model struggles to accurately capture rapid changes in the envelope, leading to discrepancies between its predictions and the ground truth. This finding indicates that the model might encounter challenges when dealing with complex and rapidly fluctuating speech envelopes.

### F. SIGNIFICANCE AND FUTURE WORK

Our study yielded several key findings. Firstly, we determined the optimal receptive field size for the reconstruction task to be around 1.56 seconds, offering valuable insights for future application design and research. Secondly, we verified the effectiveness of the fine-tuning strategy to leverage the performance. Thirdly, the proposed Sea-Wave model achieved a new state-of-the-art performance on both the competition dataset and the DTU dataset. In this study, the top-performing Sea-Wave-Small is both compact and computationally efficient, paving the way for real-time applications. Lastly, the topoplots not only highlight the importance of the auditory cortex but also the contribution of visual and somatosensory areas, suggesting further investigation of their roles in speech intelligence. However, certain challenges persist. Firstly, notable subject variability was observed, with the best subject achieving a Pearson score above 0.5 but the worst subject below 0.1. Determining whether this variability originates from neurological differences among subjects or model-related factors remains uncertain. Secondly, enhancing the population model's performance would be a significant stride towards real-world applications, as developing subject-specific models demands data and time resources. Lastly, there is room for further enhancements in terms of model architecture. Given the constraints of time and computation resources, the grid search conducted in this study may not have uncovered the most optimal hyperparameters, suggesting potential for improvement. Additionally, exploring the integration of conditioner modules, as demonstrated in Wavenet [20] and HappyQuokka [17], to address subject-specific features appears to be a promising avenue for future research. Instead of training individual models for each subject, training a compact conditioner module that captures inter-subject variability could offer a more practical solution.

## VII. CONCLUSION

This study presents Sea-Wave, a novel WaveNet-based model for speech envelope decoding from EEG. We show that it outperforms state-of-the-art models and boasts increased interpretability. On average, the best model achieves a Pearson correlation of 0.2258 on the held-out stories and 0.1158 on the held-out subjects. Using an ablation study, we identify the gated activation unit and input-skip connections as critical model components. Furthermore, our model visualizations show evidence for cross-modal interactions between auditory, visual, and somatosensory processing. In addition, they show the model excels in predicting the silences. Improving the generalization to unseen subjects and stimuli remains challenging. Examining the subject-specific models and their differences, could provide a better insight into individuals differences in auditory EEG responses and inform personalized approaches for decoding auditory processes. Lastly, improving reconstruction of the speech envelope increases the potential utility in diagnostic tests assessing speech intelligibility.

## APPENDIX A

### A. SIGMOIDAL CURVE FIT

A sigmoidal curve parameterized as

$$f(x_{\log}) = \frac{L}{1 + \exp\left(-k \cdot (x_{\log} - x_0)\right)} + b, \qquad \text{(A1)}$$

is fitted per model attribute $x_{\log}$ on a logarithmic scale. The threshold $x_{\log}^{th}$ is calculated as the intercept between the slope at the inflection point of the sigmoidal curve and its upper asymptote as follows:

$$\begin{aligned} x_{\log}^{th} &= \frac{df}{dx}(x_0)^{-1} * (y_{\max} - f(x_0)) + x_0, \\ &= \frac{k \cdot L}{4} \cdot \left(\frac{L}{2} + b\right) + x_0, \end{aligned} \qquad \text{(A2)}$$

since

$$f(x_0) = \frac{L}{2} + b, \quad \frac{df}{dx}(x_0) = \frac{k \cdot L}{4}, \text{ and } y_{\max} = L + b.$$

## REFERENCES

[1] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: Functional roles and interpretations," *Front. Hum. Neurosci.*, vol. 8, May 2014, Art. no. 311.

[2] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart, "Speech intelligibility predicted from neural entrainment of the speech envelope," *J. Assoc. Res. Otolaryngol.*, vol. 19, pp. 181–191, Apr. 2018.

[3] I. Iotzov and L. C. Parra, "EEG can predict speech intelligibility," *J. Neural Eng.*, vol. 16, Jun. 2019, Art. no. 036008.

[4] B. Accou, M. J. Monesi, H. V. Hamme, and T. Francart, "Predicting speech intelligibility from EEG in a non-linear classification paradigm," *J. Neural Eng.*, vol. 18, Dec. 2021, Art. no. 066008.

[5] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *Eur. J. Neurosci.*, vol. 31, pp. 189–193, Jan. 2010.

[6] G. Di Liberto, J. O'Sullivan, and E. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Curr. Biol.*, vol. 25, pp. 2457–2465, Oct. 2015.

[7] M. J. Crosse, G. M. D. Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli," *Front. Hum. Neurosci.*, vol. 10, Nov. 2016, Art. no. 604.

[8] T. Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *Eur. J. Neurosci.*, vol. 51, pp. 1234–1241, Mar. 2020.

[9] M. Thornton, D. Mandic, and T. Reichenbach, "Robust decoding of the speech envelope from EEG recordings through deep neural networks," *J. Neural Eng.*, vol. 19, Aug. 2022, Art. no. 046007.

[10] B. Accou, J. Vanthornhout, H. V. Hamme, and T. Francart, "Decoding of the speech envelope from EEG using the VLAAI deep neural network," *Sci. Rep.*, vol. 13, Jan. 2023, Art. no. 812.

[11] C. Puffay, J. V. Canneyt, J. Vanthornhout, H. V. Hamme, and T. Francart, "Relating the fundamental frequency of speech with EEG using a dilated convolutional network," Jul. 2022. *arXiv:2207.01963*.

[12] B. Accou, M. J. Monesi, J. Montoya, H. V. Hamme, and T. Francart, "Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network," in *Proc. 28th Eur. Signal Process. Conf.*, 2021, pp. 1175–1179.

[13] A. De Cheveigné, M. Slaney, S. A. Fuglsang, and J. Hjortkjaer, "Auditory stimulus-response modeling with a match-mismatch task," *J. Neural Eng.*, vol. 18, Aug. 2021, Art. no. 046040.

[14] B. Accou, L. Bollens, M. Gillis, W. Verheijen, H. V. Hamme, and T. Francart, "SparrKULee: A speech-evoked auditory response repository of the KU leuven, containing EEG of 85 participants," *Neuroscience*, Jul. 2023.

[15] L. Bollens, B. Accou, H. V. Hamme, and T. Francart, "SparrKULee: A speech-evoked auditory response repository of the KU Leuven, containing EEG of 85 participants," 2023.

[16] L. Bollens, B. Accou, H. V. Hamme, and T. Francart, "ICASSP 2023 auditory EEG decoding challenge," KU Leuven RDR, 2023.

[17] Z. Piao, M. Kim, H. Yoon, and H.-G. Kang, "HappyQuokka system for ICASSP 2023 auditory EEG challenge," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–2.

[18] B. V. Dyck, L. Yang, and M. M. V. Hulle, "Decoding auditory EEG responses using an adapted WaveNet," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–2.

[19] P. L. Søndergaard and P. Majdak, "The auditory modeling toolbox," in *The Technology of Binaural Listening*, [Modern Acoustics and Signal Processing Series], (J. Blauert). Berlin, Heidelberg: Springer, 2013, pp. 33–56.

[20] A. V. D. Oord et al., "WaveNet: A generative model for raw audio," Sep. 2016. *arXiv:1609.03499*.

[21] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 5069–5073.

[22] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," Mar. 2021. *arXiv:2009.09761*.

[23] J. Boilard, P. Gournay, and R. Lefebvre, "A literature review of WaveNet: Theory, application, and optimization," in *Audio Engineering Society Convention*, vol. 146. New York, NY, USA: Audio Engineering Society, Mar. 2019.

[24] M. Guven and F. Uysal, "Time series forecasting performance of the novel deep learning algorithms on stack overflow website data," *Appl. Sci.*, vol. 13, Apr. 2023, Art. no. 4781.

[25] R. Wang, Y. Wang, and A. Flinker, "Reconstructing speech stimuli from human auditory cortex activity using a WaveNet approach," Nov. 2018. *arXiv:1811.02694*.

[26] A. V. D. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," Jun. 2016. *arXiv:1606.05328*.

[27] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," Jul. 2012. *arXiv:1207.0580*.

[28] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," Jun. 2016. *arXiv:1602.07868*.

[29] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," Jun. 2023. *arXiv:1606.08415*.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," Jan. 2019. *arXiv:1711.05101*.

[31] S. A. Fuglsang, D. D. Wong, and J. Hjortkjær, "EEG and audio dataset for auditory attention decoding," Mar. 2018.

[32] K. Ignatiadis, R. Barumerli, B. Tóth, and R. Baumgartner, "Effects of individualized brain anatomies and EEG electrode positions on inferred activity of the primary auditory cortex," *Front. Neuroinform.*, vol. 16, Oct. 2022, Art. no. 970372.

[33] V. Marian, S. Hayakawa, and S. R. Schroeder, "Cross-modal interaction between auditory and visual input impacts memory retrieval," *Front. Neurosci.*, vol. 15, Jul. 2021, Art. no. 661477.

[34] S. Sugiyama, N. Takeuchi, K. Inui, M. Nishihara, and T. Shioiri, "Effect of acceleration of auditory inputs on the primary somatosensory cortex in humans," *Sci. Rep.*, vol. 8, Aug. 2018, Art. no. 12883.