Improved decoding of the speech envelope using a deep neural network

Bernd Accou[1,2], Jonas Vanthornhout[1], Hugo Van hamme[2], Tom Francart[1]

[1] ExpORL, Department of Neurosciences, KU Leuven
[2] PSI, Department of Electrical Engineering (ESAT), KU Leuven

Introduction

Neural tracking

- Finding neural tracking of speech features in the brain
- Can be used as diagnostic tool for speech understanding [Vanthornhout et al., 2018, Accou et al., 2021]

Problems

- Relatively simple linear models are used
- Subject-specific models, which requires collection of training data for each (test-)subject

Proposed solution: A very large augmented auditory inference (VLAAI) network

- Non-linear complex model to decode the envelope from EEG
- Subject-independent model to reduce amount of data needed per subject

Dataset & Preprocessing

Subjects

- 106 normal hearing native Flemish speakers
- Screened with pure-tone audiogram and Flemish MATRIX test

Stimuli

- 2-8 children's stories narrated by a single speaker in Dutch
- $\approx$ 15 minutes in length

This amounts to $\approx$ 195 hours of EEG data (1 hours and 50 minutes of EEG data per subject on average).

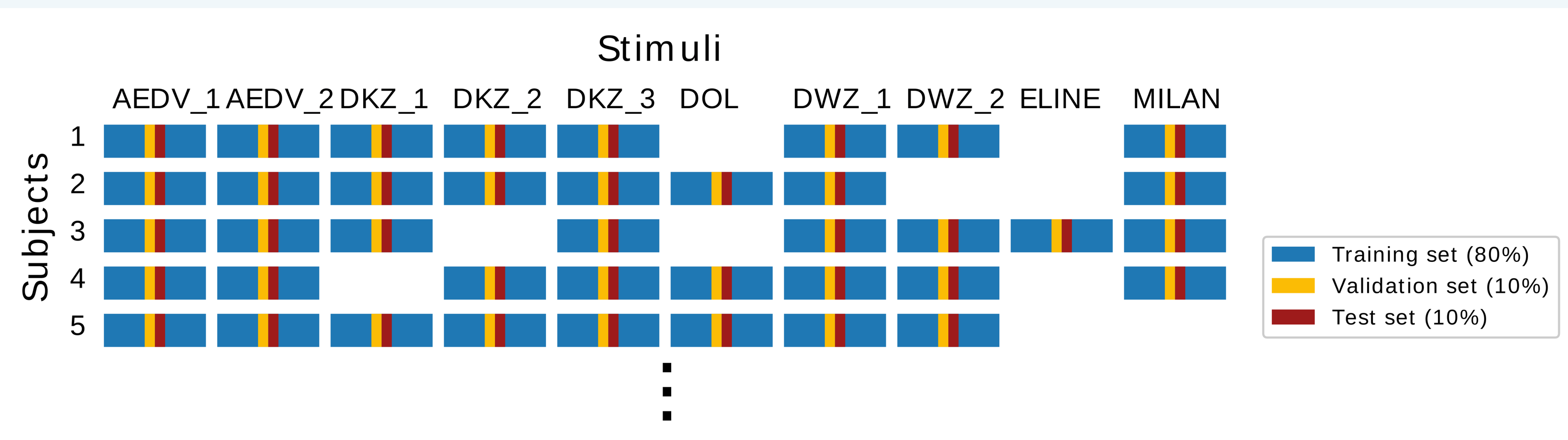

Figure 1: A visualization of the recordings (blue boxes). Test- and validation-set are extracted from the middle of the recording

Preprocessing steps

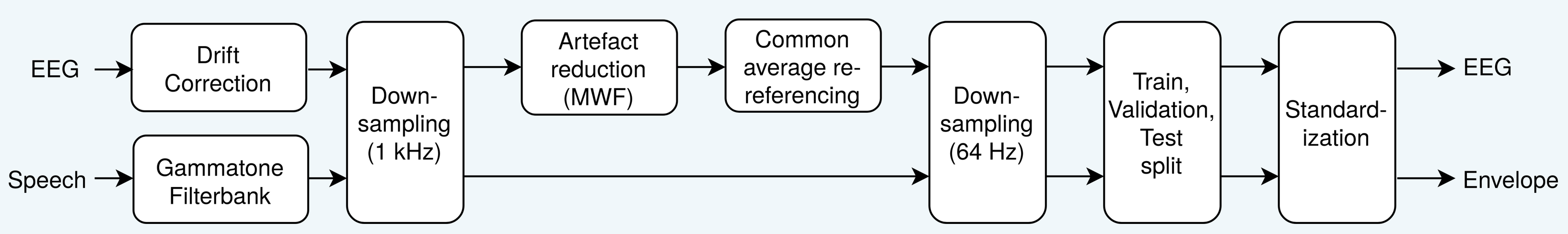

Figure 2: All preprocessing steps performed on the dataset.

To train and evaluate the models in the experiment section, windows of 5 seconds with an overlap of 80% were used

Model architecture

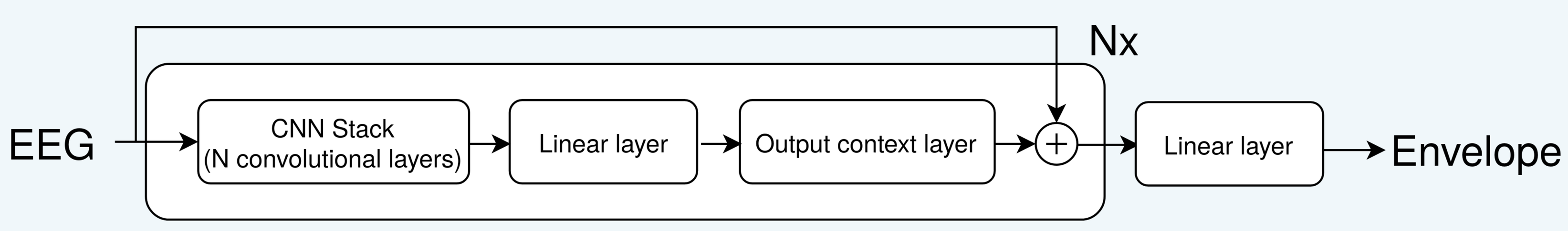

Figure 3: The architecture of the new proposed VLAAI network.

The VLAAI network consists of 3 modules which can be repeated $Nx(=4)$ times:

1. **CNN stack**
   This module consists of 5 convolutional layers with a kernel size of 8. The first 3 and last 2 layers have 256 and 128 filters respectively. After each convolutional layer, layer normalization, a LeakyReLU activation and zero-padding (7 samples at the end of the sequence) were applied
2. **Linear layer**
   A linear combination of the output filters of the CNN stack.
3. **Output context layer**
   Predicting a better sample based on 31 previous samples + the current sample.

VLAAI is trained with negative Pearson coefficient as a loss function

Comparison to baseline models

All models utilized in this section were trained with EEG data of all participants (i.e. **subject-independent** models).

The proposed VLAAI network is compared to 3 baseline models:

1. **Linear decoder**
   A linear decoder with an integration window of 500 ms, trained with negative pearson coefficient as loss.
2. **CNN**
   The CNN model of Thornton et al. [2022], which was based on the EEGNET architecture [Lawhern et al., 2018]
3. **FCNN**
   The FCNN model of Thornton et al. [2022], which was based on the architecture of de Taillez et al. [2017].

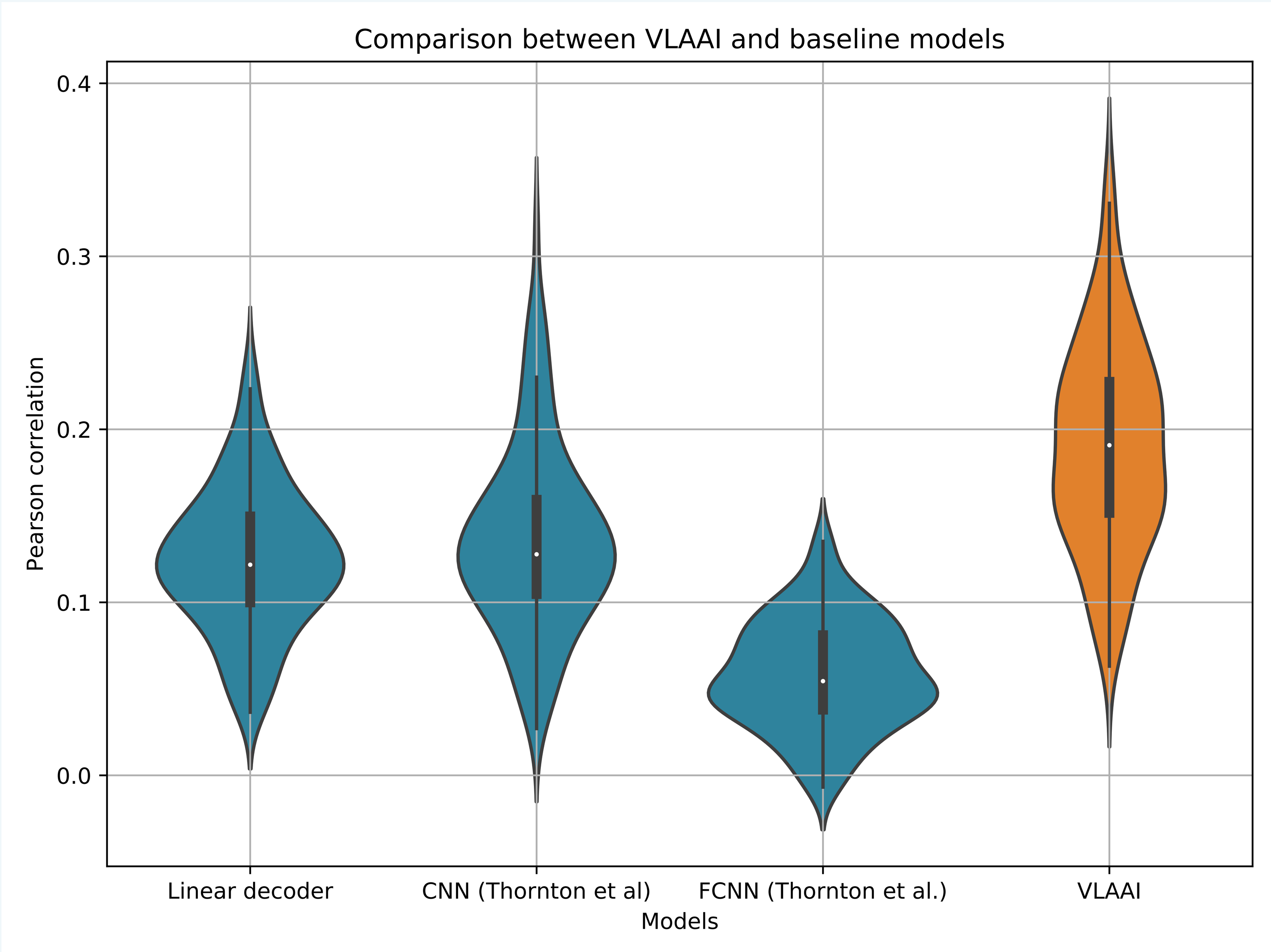

Figure 4: Comparison between the VLAAI network with 3 baseline models for 5 second segments. Each point in the boxplot represents the reconstruction score for a subject, averaged across stimuli.

Models were compared using a Wilcoxon signed-rank test using Holm-Bonferroni correction. All models performed significantly different ($p \leq 0.05$). The VLAAI network significantly outperforms all baseline models

Discussion

- While Thornton et al. [2022] could be replicated using their own provided data, the FCNN model performed worse when trained/applied on the datasets presented here (see Figure 4). A possible explanation is that the hyperparameters for regularizing the population models in Thornton et al. [2022] are not optimal for our dataset
- The VLAAI network outperforms the baseline models substantially (a relative improvement in median accuracy of 56.74% compared to the linear decoder baseline). A better decoder model, such as VLAAI, might reveal neural tracking and/or effects that were hidden previously

Acknowledgments & References

The work is funded by KU Leuven Special Research Fund C24/18/099 (C2 project to Tom Francart and Hugo Van hamme).Research funded by a PhD grant (1S89620N) of the Research Foundation Flanders (FWO).This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 637424, ERC starting Grant to Tom Francart).

B. Accou, M. J. Monesi, H. V. hamme, and T. Francart. Predicting speech intelligibility from EEG in a non-linear classification paradigm $\less$sup$\greater$\ast$\less$/sup$\greater$. *Journal of Neural Engineering*, 18(6):066008, Nov. 2021. ISSN 1741-2552. doi: 10.1088/1741-2552/ac33e9. URL https://doi.org/10.1088/1741-2552/ac33e9. Publisher: IOP Publishing.

T. de Taillez, B. Kollmeier, and B. T. Meyer. Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. *European Journal of Neuroscience*, 51(5):1234–1241, 2017. ISSN 1460-9568. doi: 10.1111/ejn.13790. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.13790. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejn.13790.

V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, July 2018. ISSN 1741-2552. doi: 10.1088/1741-2552/aace8c. URL https://doi.org/10.1088/1741-2552/aace8c. Publisher: IOP Publishing.

M. Thornton, D. Mandic, and T. Reichenbach. Robust decoding of the speech envelope from EEG recordings through deep neural networks. *Journal of Neural Engineering*, 19(4):046007, July 2022. ISSN 1741-2552. doi: 10.1088/1741-2552/ac7976. URL https://doi.org/10.1088/1741-2552/ac7976. Publisher: IOP Publishing.

J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart. Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. *Journal of the Association for Research in Otolaryngology*, 19(2):181–191, Apr. 2018. ISSN 1438-7573. doi: 10.1007/s10162-018-0654-z. URL https://doi.org/10.1007/s10162-018-0654-z.