LEARNING SUBJECT-INVARIANT REPRESENTATIONS FROM SPEECH-EVOKED EEG USING VARIATIONAL AUTOENCODERS

Lies Bollens\textsuperscript{1,2}, Tom Francart\textsuperscript{2}, Hugo Van Hamme\textsuperscript{1}

\textsuperscript{1}KU Leuven, PSI, Dept. of Electrical engineering (ESAT), Leuven, Belgium
\textsuperscript{2}KU Leuven, ExpORL, Dept. Neurosciences, Leuven, Belgium

1. INTRODUCTION

Recently, much research has gone into modeling how natural running speech is processed in the human brain. A standard approach is to present natural running speech to a subject while EEG signals are recorded. Subsequently, linear models are trained to either reconstruct the speech stimuli, to predict EEG from the stimuli, or to transform both stimulus and neural response to a shared space using a hierarchical model. The archictecture is a variant of Variational Autoencoders (VAE) \cite{Kingma2014} and have found that VAE's may improve subject-invariant modeling of EEG brain neural signals. Though these works yield promising results, little work has been conducted using extremely short windows, necessary for decoding at word-level, as such short windows contain very little information.

Looking at advances in speech processing, a way to accurately model short windows while still keeping track of higher-order features is to use hierarchical models. A factorized hierarchical variational autoencoder (FHVAE) \cite{Bollens2021} is a deep variational inference-based generative hierarchical model, encoding generating factors into two disentangled latent spaces. One space captures high-level slow varying information, whereas the other captures residual fast-changing information. This hierarchical approach allows capturing high-level information while using small time windows.

In this work, inspired by recent progress in both speech recognition and related EEG fields, we adopt the architecture of \cite{Bollens2021} and use FHVAEs to model the neural response to a speech stimuli for extremely short segments of EEG of 500 ms. We propose an extension to the FHVAE architecture, exploiting that the same stimuli can be presented to multiple subjects. This work aims to design a model that can generate subject-invariant representations for short EEG segments, as this is a first step towards decoding at the word level. To this end, we measure how well subject and content are represented in their respective latent spaces and how well these two generating factors are disentangled.

2. METHODOLOGY

2.1. Factorized Hierarchical Variational Autoencoders

Factorized Hierarchical Variational Autoencoders (FHVAE) \cite{Bollens2021} are a variant of Variational Autoencoders \cite{Kingma2014} designed to model sequential data using a hierarchical model. The architecture is shown in figure 1. Each uninterrupted recorded sequence \(i\) of EEG signals \(X_i\) can be decomposed into \(N\) segments of fixed length \(\{x_{i,n}\}_{n=1}^N\). The FHVAE model imposes that the generation of each segment \(x_i\) is conditioned on two latent variables, \(z_1\) and \(z_2\) (omitting superscripts). We refer to \(z_1\) as the latent content variable, as \(z_1\) is generated from a global prior and captures the fast-changing generating factors between segments. On the other hand, we refer to \(z_2\) as the latent subject variable, as \(z_2\) is drawn from a sequence-dependent Gaussian prior, parameterized by an (estimated) sequence-level mean \(\mu_2\) and a fixed variance. The FHVAE learns to encode common factors between segments from the same sequence into \(z_2\). By contrast, \(z_1\) has a zero-mean prior.

In order to avoid collapse (trivial solution for \(z_2\)), a discriminative objective is added: it should be possible to infer the sequence number \(i\) from \(z_2\). An FHVAE introduces an inference model \(q\), since the true posteriors of \(z_1, z_2\) and \(\mu_2\) are intractable. More detailed derivations can be found in \cite{Bollens2021}.

The objective the FHVAE tries to maximize can be summarized as,
3. EXPERIMENTAL SETUP

3.1. Data collection and preprocessing

We present natural running speech to subjects while simultaneously recording the EEG signal. In total, 100 normal-hearing, native Flemish participants contributed to this study. Normal hearing was confirmed using pure tone audiometry and the Flemish MATRIX speech-in-noise test [17]. We present to each participant a minimum of 6 and a maximum of 8 stories, each around 15 minutes in length, chosen from a set of ten different stories and presented in random order. We present the stories binaurally at 62 dBA with Etymotic ER-3A insert phones and ask the subjects a comprehension question after each story. The EEG data is recorded using a 64-channel Biosemi Active-Two EEG at a sampling rate of 8 kHz. Stories are presented using the APEX 4 software platform [18]. Subjects are seated in an electromagnetically shielded and soundproof booth. We remove artifacts from the EEG recording using a multichannel Wiener filter [19]. We bandpass filter the EEG signals between 0.5 Hz and 32 Hz and then downsample to 64 Hz. We use the first and last 40% for training and divide the remaining 20% equally into validation and test set.

3.2. Training and Model Configurations

We present EEG segments of 32 frames, corresponding to 500 ms, as input to the model. We implement the architecture introduced in [16]. \( x \) and \( z_2 \) are predicted by a stacked LSTM network of two layers, each containing 128 cells, followed by two separate single fully connected layers, predicting the conditional mean and variance, respectively. These last layers take as input the output from the last time step from both LSTM layers, summing up to a total of 256 dimensions. The decoder network reconstructing \( z \) consists of a two-layer stacked LSTM decoder network, which takes as input at each time step the concatenation of the sampled \( z_1 \) and \( z_2 \) from the posterior distribution. Subsequently, the predictor network takes the LSTM output of the last layer from each time step and predicts the probability distribution of the corresponding time frame \( p(x_t|z_1, z_2) \). The dimensions of \( z_1 \) and \( z_2 \) are both 32. We set the variances as \( \sigma_{z_1}^2 = \sigma_{z_2}^2 = 0.25 \) and \( \sigma_{a_1}^2 = \sigma_{a_2}^2 = 1 \).

We try the model in two steps. First, we train a default FHVAE, as described in section 2.2, allowing the model to focus on modeling subject-specific generating factors in the latent \( z_2 \) space. We set \( \alpha_{z_2} = 100 \). We will refer to this model as FHVAE. Second, we add the extra regularisation on \( z_1 \) and train our extended FHVAE model, as described in section 2.2, such that the model can now focus on modeling content-dependent, subject-independent generating factors in the latent \( z_1 \) space. We start training with the model weights obtained in the first step. We set \( \alpha_{z_1} = 10000, \alpha_{z_2} = 100 \) and choose a segment batch size \( K = 5000 \). We will refer to this model as Extended FHVAE.

We use ADAM [20] with \( \beta_1 = 0.95 \) and \( \beta_2 = 0.999 \) for both training steps and train over 500 epochs, with early stopping when the variational lower bound on the held-out validation set does not improve for 50 epochs. We use Tensorflow 2.0 for implementation.

3.3. Subject Classification and Disentanglement

We describe the setup used to quantify subject classification performance of the model. The question we answer here is two-fold: does the model succeed at extracting the subject generating factors in the \( z_2 \) latent space and disentangling subject from content information? we group segments of EEG according to subject, which results in a
total of 100 different classes. Train, test, and validation sets are taken as the sets used for the training of the FHVAE, described in section 3.1. For all segments, we infer the latent $z_1$ or $z_2$ and then train a one-layer fully connected network with softmax activation and 100 output classes, using categorical cross-entropy as the loss function and ADAM as optimizer.

To validate our results, we compare with a recently proposed architecture for EEG person authentication [21], which has been compared in a recent survey about methods and challenges in EEG person authentication [22] and yields excellent results. The authors of [21] propose to use an encoder consisting of a convolutional neural network, followed by a single fully connected layer for subject classification. We will refer to this network as CNN. This architecture uses EEG window lengths of 500 ms, making it directly transferable to our paradigm. We use the preprocessing pipeline of our EEG data and sample at 64 Hz, which differs from the 256 Hz proposed by the authors. We use the same train, test, and validation set as above. We jointly train the CNN encoder and output classification network, using categorical cross-entropy as loss function and ADAM as optimizer.

3.4. Content Classification and Disentanglement

We describe the setup used to quantify content classification performance of the model. The question we answer here is two-fold: does the model succeed at extracting the content generating factors in the $z_1$ latent space and disentangling content from subject information?

We keep all EEG segments of the test set, as described in section 3.1 and group them according to their content label $l$, which results in a total of 1538 classes. Each class contains, on average, 80 examples. Instead of directly classifying which of these 1538 classes the latent representations belong to, which would be a high-dimensional problem prone to overfitting given the limited training examples per class, we simplify the classification problem into a binary one. We select a label $l_i$ at timestep $t_i$ and choose a second label $l_{i+1}$ directly after, at timestep $t_i + 500\text{ms}$. We split these two classes into 80% train and 20% test set, using a 5-fold validation scheme. Subsequently, a Linear Support Vector Machine is trained, and we calculate average accuracy over the test set. We conduct this experiment for $k = 1..1537$, effectively utilizing the whole test set.

As a baseline, we take the same approach to selecting classes and training, but instead of inferring the latent variable, we flatten the EEG segments and directly train a Linear Support Vector Machine on the raw EEG.

4. RESULTS

4.1. t-SNE Visualization of Latent Variables

In this section, we report a qualitative visual inspection to evaluate performance of the proposed Extended FHVAE model containing 128 LSTM cells. Therefore, we visualize the latent $z_1$ and $z_2$ space according to different generating factors, which can be either subject or content.

We start by randomly selecting 100 content labels and keeping all segments from all of the 100 different subjects that belong to one of these content labels, resulting in a set of 7453 unique labeled segments. We then infer $z_1$ and $z_2$ for each of these labeled segments and project them onto a two-dimensional space using t-Distributed Stochastic Neighbor Embedding (t-SNE) [23]. For visualization, we color-code the projected $z_1$ and $z_2$ according to their generating factors. A successful model should result in segments of the same sub-

![Fig. 2](image.png)

Fig. 2. t-SNE scatter plots of $z_1$ and $z_2$ latent space. Every point represents one segment. Different colors indicate either content segments or subjects. Different generating factors (subject, content) are shown at the title of each plot.

4.2. Subject Classification and Disentanglement: Experiments

In this section, we report subject classification results, as described in section 3.3. Results are shown in figure 3. We report the mean accuracy over 100 subjects. For clarity, the subject classification accuracy of $z_1$ is left out of the plot. This accuracy ranges between 1.60% and 2.53% for different models, where 1% is the chance level since all 100 subjects are equally present in the data set. During the first stage, while using the original FHVAE architecture, the model should focus on extracting subject information from the EEG segments and disentangling this information from content. Subject classification accuracy on the latent $z_2$ space reaches up to 99.04% for the model using 64 LSTM cells, confirming that during the first stage, the model learns to model subject-specific information in the latent $z_2$ but not in $z_1$, achieving great disentanglement.
During the second stage, while training the Extended FHVAE architecture, the model focuses on extracting content representations in the $z_1$ space, which is reflected in the subject classification accuracy: for the LSTM model containing 32 or 64 cells, there is a significant ($W = 204, p < 0.001$ and $W = 199, p < 0.001$) decrease in accuracy of 1.46% and 3.23% respectively. We report all statistics using a Wilcoxon-signed rank test with normal approximation. This accuracy drop might be indicative of an inherent trade-off in the model between modeling subject and content. However, for the 128 LSTM cells model, the mean accuracy of the FHVAE is 98.94%, which does not decrease in the second step, from which we conclude that, with regard to subject classification accuracy, there is an advantage to using the bigger 128 LSTM cells model.

The CNN model achieves 99.33% subject classification accuracy, which is slightly yet significantly ($W = 750, p < 0.001$) higher than the extended FHVAE model with 128 LSTM cells at 98.96%.

4.3. Content Classification and Disentanglement: Experiments

In this section, we report content classification results, as described in section 3.4. Results are shown in figure 3. We report the average accuracy over stories, over folds. In total, there are nine different stories to which the subjects listen. For the standard FHVAE model, accuracy is about the same for the 32, 64 and 128 LSTM model for $z_1$ (53.65%, 53.55%, and 53.87%) and $z_2$ (55.31%, 55.27%, and 55.03%). The best performing model of these is still significantly lower ($W = 3, p < 0.001$) than the classification accuracy of the naive SVM (57.45%), from which we can conclude that the standard model achieves neither content extraction nor disentanglement.

However, after training the extended FHVAE model, classification accuracy of the latent $z_1$ space significantly improves ($W = 0, p < 0.001$) from 54.10% to 62.91% for the model with 128 cells, which indicates that the model succeeds at extracting some relevant content information in the $z_1$ space. The classification accuracy of the smaller LSTM models sees a smaller increase to around 60%, from which we can conclude that, with regard to content classification accuracy, there is an advantage to using the bigger 128 LSTM cells model. For the latent $z_2$ space, performance drops to just barely above the chance level at 51.51%, suggesting better disentanglement for content in the extended FHVAE model.

5. CONCLUSION

This work aims to design a model that can generate subject-invariant representations for short EEG segments of around 500 ms, a necessary temporal resolution for decoding at the word level. To this end, we introduce factorized hierarchical variational autoencoders (FHVAE). FHVAE’s encode the generating factors of the EEG into two disentangled latent spaces. The latent $z_2$ space should encode subject but not content, whereas the latent $z_1$ space should encode content but not subject. In order to exploit that it is possible to generate multiple EEG recordings of one or more subjects listening to the same stimulus, we propose an adaption to the original FHVAE architecture: Extended factorized hierarchical variational autoencoders.

We train both models using EEG recordings from 100 subjects who listen on average to 8 stimuli, each around 15 minutes. In a first step, we train the default FHVAE architecture, allowing the model to focus on modeling subject-specific generating factors in the latent $z_1$ space. In a second step, we use the learned model weights and train the Extended FHVAE architecture, allowing the model to focus on modeling content-specific generating factors in the latent $z_2$ space.

Subject classification accuracy between all 100 subjects reaches 98.94% and 98.96% for the latent $z_1$ representations of respectively the FHVAE and the Extended FHVAE architecture, showing that the default FHVAE model succeeds at extracting subject information and that this information is kept in the second stage. Subject classification accuracy for the latent $z_1$ is for both models at around 2%, suggesting great disentanglement with regard to subject. At 99.33%, subject classification for the state-of-the-art CNN model is slightly superior, but its sole goal is to represent subject, thereby discarding content information. The Extended FHVAE however provides disentangled representations of both content and subject.

Binary content classification accuracy for the default FHVAE model reaches 53.87% and 55.27% for respectively the latent $z_1$ and $z_2$ representations. Content classification accuracy for the Extended FHVAE model improves to 62.91% and 51.51% for respectively the latent $z_1$ and $z_2$ representations, confirming that the Extended model succeeds at modeling content generating factors in $z_1$ but not in $z_2$, something which the default FHVAE fails to achieve.

We conclude that the Extended FHVAE model successfully models both subject and content generating factors and disentangles these two while using small EEG segments of around 500 ms. Discrimination between such small-scale segments might be a first step towards decoding at the word level.
6. REFERENCES

[1] Giovanni M. Di Liberto, James A. O’Sullivan, and Edmund C. Lalor, “Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing.” Current Biology, vol. 25, no. 19, pp. 2457–2465, Oct. 2015.

[2] Alain de Cheveigné, Daniel D. E. Wong, Giovanni M. Di Liberto, Jens Hjortkjaer, Malcolm Slaney, and Edmund Lalor, “Decoding the auditory brain with canonical component analysis,” NeuroImage, vol. 172, pp. 206–216, May 2018.

[3] Jonas Vanthornhout, Lien Decruy, E. Verschueren, L. Decruy, and Jan Wouters, “Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope,” Journal of the Association for Research in Otolaryngology, vol. 19, no. 2, pp. 181–191, Apr. 2018.

[4] D. Lesenfants, J. Vanthornhout, E. Verschuiren, L. Decruy, and T. Francart, “Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations,” vol. 380, pp. 1–9.

[5] Giovanni Di Liberto, Jingping Nie, Jeremy Yeaton, Bahar Khaliqinejad, Shihab Shamma, and Nima Mesgarani, “Neural representation of linguistic feature hierarchy reflects second-language proficiency.”

[6] Mohammad Jalilpour Monesi, Bernd Accou, Jair Montoya-Martínez, Tom Francart, and Hugo Van Hamme, “An LSTM based architecture to relate speech stimulus to eeg,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 941–945, ISSN: 2379-190X.

[7] Bernd Accou, Mohammad Jalilpour Monesi, Hugo Van hamme, and Tom Francart, “Predicting speech intelligibility from EEG using a dilated convolutional network,” arXiv:2105.06844 [cs, eess, q-bio], May 2021, arXiv: 2105.06844.

[8] Alain de Cheveigné, Malcolm Slaney, Søren A. Fuglsang, and Jens Hjortkjaer, “Auditory stimulus-response modeling with a match-mismatch task,” Journal of Neural Engineering, vol. 18, no. 4, pp. 046040, May 2021.

[9] Ozan Özdenizci, Ye Wang, Toshiaki Koike-Akino, and Deniz Erdoˇgmu¸s, “Learning Invariant Representations From EEG via Adversarial Inference,” IEEE Access, vol. 8, pp. 27074–27085, 2020.

[10] Jinpeng Li, Shuang Qiu, Changde Du, Yixin Wang, and Huiguang He, “Domain Adaptation for EEG Emotion Recognition Based on Latent Representation Similarity,” IEEE Transactions on Cognitive and Developmental Systems, vol. 12, no. 2, pp. 344–353, June 2020.

[11] Soheil Rayatdoost, Yufeng Yin, David Rudrauf, and Mohammad Soleymani, “Subject-Invariant Eeg Representation Learning For Emotion Recognition,” in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), June 2021, pp. 3955–3959, ISSN: 2379-190X.

[12] Xiang Li, Zhigang Zhao, Dawei Song, Yazhou Zhang, Jingshan Pan, Lu Wu, Jidong Huo, Chunyang Niu, and Di Wang, “Latent Factor Decoding of Multi-Channel EEG for Emotion Recognition Through Autoencoder-Like Neural Networks,” Frontiers in Neuroscience, vol. 14, pp. 87, Mar. 2020.

[13] Juan Lorenzo Hagad, Tsukasa Kimura, Ken-ichi Fukui, and Masayuki Numao, “Learning Subject-Generalized Topographical EEG Embeddings Using Deep Variational Autoencoders and Domain-Adversarial Regularization,” Sensors, vol. 21, no. 5, pp. 1792, Jan. 2021.

[14] Diederik P. Kingma and Max Welling, “Auto-Encoding Variational Bayes,” arXiv:1312.6114 [cs, stat], May 2014, arXiv: 1312.6114.

[15] Wei-Ning Hsu, Yu Zhang, and James Glass, “Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data,” arXiv:1709.07902 [cs, eess, stat], Sept. 2017, arXiv: 1709.07902.

[16] Wei-Ning Hsu and James Glass, “Scalable Factorized Hierarchical Variational Autoencoder Training,” arXiv:1804.03201 [cs, eess, stat], June 2018, arXiv: 1804.03201.

[17] Heleen Luts, Sofie Jansen, Wouter Dreschler, and Jan Wouters, “Development and normative data for the Flemish/Dutch Matrix test,” 2014.

[18] Tom Francart, Astrid van Wieringen, and Jan Wouters, “APEX 3: a multi-purpose test platform for auditory psychophysical experiments,” Journal of Neuroscience Methods, vol. 172, no. 2, pp. 283–293, July 2008.

[19] Ben Somers, Tom Francart, and Alexander Bertrand, “A generic EEG artifact removal algorithm based on the multi-channel Wiener filter,” Journal of Neural Engineering, vol. 15, no. 3, pp. 036007, Feb. 2018.

[20] Diederik P. Kingma and Jimmy Ba, “Adam: A Method for Stochastic Optimization,” arXiv:1412.6980 [cs], Jan. 2017, arXiv: 1412.6980.

[21] Ozan Özdenizci, Ye Wang, Toshiaki Koike-Akino, and Deniz Erdoˇgmu¸s, “Adversarial Deep Learning in EEG Biometrics,” IEEE signal processing letters, vol. 26, no. 5, pp. 710–714, May 2019.

[22] Amir Jalaly Bidgoly, Hamed Jalaly Bidgoly, and Zeynab Arezoumand, “A survey on methods and challenges in EEG based authentication,” Computers & Security, vol. 93, pp. 101788, June 2020.

[23] Laurens van der Maaten and Geoffrey Hinton, “Visualizing Data using t-SNE,” Journal of Machine Learning Research, vol. 9, no. 86, pp. 2579–2605, 2008.