KU LEUVEN

FACULTEIT
INGENIEURSWETENSCHAPPEN

Deep neural networks as a model of speech perception

Soufiane Jhilal

Thesis submitted to obtain the degree of
Masterof Science in Artificial Intelligence,
option Speech and Language Technology

Promotor:
Prof. Dr. Hugo Van hamme

Assessor:
Dr. Jonas Vanthornhout

Supervisors:
Mohammad Jalilpour Monesi & Lies Bollens

Academic year 2021-2022

© Copyright by KU Leuven

Without written permission of the supervisor(s) and the authors it is forbidden to reproduce or adapt in any form or by any means any part of this publication. Requests for obtaining the right to reproduce or utilize parts of this publication should be addressed to KU Leuven, Faculty of Engineering Science - Kasteelpark Arenberg 1, B-3001 Heverlee (Belgium). Telephone +32-16-32 13 50 & Fax. +32-16-32 19 88.

A written permission of the supervisor(s) is also required to use the methods, products, schematics and programs described in this work for industrial or commercial use, and for submitting this publication in scientific contests.

© Copyright by KU Leuven

Zonder voorafgaande schriftelijke toestemming van zowel de promoter(en) als de auteur(s) is overnemen, kopiëren, gebruiken of realiseren van deze uitgave of gedeelten ervan verboden. Voor aanvragen tot of informatie i.v.m. het overnemen en/of gebruik en/of realisatie van gedeelten uit deze publicatie, wend u tot de KU Leuven, Faculteit Ingenieurswetenschappen - Kasteelpark Arenberg 1, B-3001 Heverlee (België). Telefoon +32-16-32 13 50 & Fax. +32-16-32 19 88.

Voorafgaande schriftelijke toestemming van de promoter(en) is eveneens vereist voor het aanwenden van de in dit afstudeerwerk beschreven (originele) methoden, producten, schakelingen en programma's voor industrieel of commercieel nut en voor de inzending van deze publicatie ter deelname aan wetenschappelijke prijzen of wedstrijden.

Foreword

First and foremost, I would like to express my sincere gratitude to my promotor, Pr. Dr. Hugo Van hamme who welcomed me in his research group and gave me the opportunity to work on this interesting topic. I would also like to express my gratitude to my supervisors, Mohammad Jalilpour Monesi & Lies Bollens for their patience, kindness, and invaluable input across every step of my internship. Special thanks also go to my assessor Dr. Jonas Vanthornhout for taking the time to read my work and give his feedback. Finally, I would like to thank my family and friends for their unconditional support of all my endeavours.

Table of Contents

Abstract

Examining the relation between speech stimulus and brain signals can help us better understand how humans process speech and even contribute to improving hearings diagnostics and brain-computer interfaces. In this study we tried to investigate the effects of decision windows, attention and the relevance of silences in the performance of convolutional and long short-term memory models that relate speech to EEG. With the first experiment, we confirmed that longer decision windows lead to improved classification accuracy in a match/mismatch paradigm. We also tried to explore the effect of the participants' attention, as reflected in the order of the stimulus presentation, on the performance of the classification model. No significant effect was found but further investigation is needed. Finally, we tried to reconstuct silences from EEG segments and use the reconstructed speech to achieve good match/mismatch classification performance by exploiting the importance of silences to the decisions of the model. While we did not achieve the best accuracy, the pipeline still reached 60% accuracy with a simple linear match/mismatch decision paradigm.

List of figures and tables

List of figures

List of tables

List of abbreviations and symbols

BCI	brain-computer interface
EcoG	electrocorticography
EEG	electroencephalography
MEG	magnetoencephalography
fMRI	functional magnetic resonance imaging
BOLD	blood-oxygen-level dependent
ERP	event-related potential
AAD	audiory attention detection
DNN	deep neural network
CNN	convoluntional neural network
SVM	support vector machine
LSTM	Long Short-Term Memory
DAE	deep autoencoder
VAD	voice activity detection
ReLU	rectified linear unit

Chapter 1: Introduction

Understanding how the brain processes speech has attracted a lot of attention in the recent years, both in academia and industry. Any interesting findings in this field will not only expedite our progress in modelling human speech perception but more importantly, it can lead to better diagnostics of hearing impairments (Vanthornhout et al., 2018), more efficient architectures for hearing aids (Geravanchizadeh & Zakeri, 2021), improved models for automated speech recognition systems (Baby & Verhulst, 2018) and may even lead to reliable speech brain-computer interfaces (BCIs) (Proix et al., 2022).

A variety of studies attempted to study human speech processing (Ganushchak et al., 2011; Giraud et al., 1997; Hagoort & Brown, 2000; Nourski, 2017; Sanders and Neville, 2003) either by trying to relate auditory stimuli to brain signals (Assaneo & Poeppel, 2018; Bigliassi et al., 2017; Butler & Trainor, 2012; Moumdjian et al., 2018), or by attempting to decode certain properties of speech from those signals (Clayton et al., 2020; Martin et al. 2014; Murphy et al., 2022). Different neuroimaging techniques have been used to record brain signals depending on the paradigm, the subject, the objective of the study and other variables. Several non-invasive recording techniques have been considered for speech analysis (Bocquelet et al., 2016), while other studies rely on invasive neuro-recordings, such as electrocorticography (ECoG), owing to their higher temporal and spatial resolutions (Martin et al., 2018).

1.1 Neuro-recording techniques

Non-invasive neuroimaging techniques (**Figure 1**) including electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI), are the most used techniques in the field

since they do not require any risky brain operations to open the subject's skull in order to record brain signals. By using fMRI recordings per example, it was demonstrated that perceived speech items or types can be successfully decoded with an above chance level accuracy from fMRI blood-oxygen-level dependent (BOLD) activity recorded in subjects' auditory areas (Bonte et al., 2014; Correia et al., 2013; Evans et al., 2013; Formisano et al., 2008). Moreover, speech articulatory features such as place of articulation were also successfully decoded from fMRI recordings (Correia et al., 2015). Nonetheless, despite such promising results, fMRI has a low temporal resolution (Friston et al., 1994) due to its indirect delayed measure of neural activity via the BOLD response (hemodynamic lag) which cannot match real-time speech synthesis.

Figure 1: Illustrative images of different non-invasive neuroimaging techniques used to acquire brain signals for speech decoding. (a) Example of an EEG setup. (b) Example of an MEG setup. (c) Example of an fMRI setup. (Images taken from Ombao et al., 2019)

MEG and EEG are ideal for decoding speech representations thanks to their high temporal resolution. Although MEG provides better spatial resolution of source localization (2-3 mm) than EEG (7-10 mm) (Singh, 2014), MEG equipment are big in size, very expensive, and require electromagnetic shielding, which poses significant challenges to deploying MEG for everyday life use of a BCI system at home.

The current gold standard and most widespread neuro-recording technique for speech and auditory processing studies is EEG, despite its poor spatial resolution and poor signal quality due to the distance of the electrodes from the electrical source in cortex (Michel & Murray, 2012; Zelmann et al., 2013). EEG is relatively cost-effective compared to the other neuro-recording methods, while still giving a good coverage of the brain with excellent temporal resolution which makes it very suitable for auditory studies.

1.2 Brain signals

Brain cells generate electrical activity as a result of millions of action potentials (or nerve pulses) of individual neurons (Schuenke et al., 2020). EEG measures the electrical action potential that occurs when neurons receive and transmit information. When synapses occur on neuron dendrites, a small electric field (dipole) is created along the neuron body due to the difference in charge between the dendrites and the axon (Figure 2). This electric field only lasts for a few milliseconds. The accumulation of these electrical potentials results in brain waves that indicate the activity of the cerebral cortex

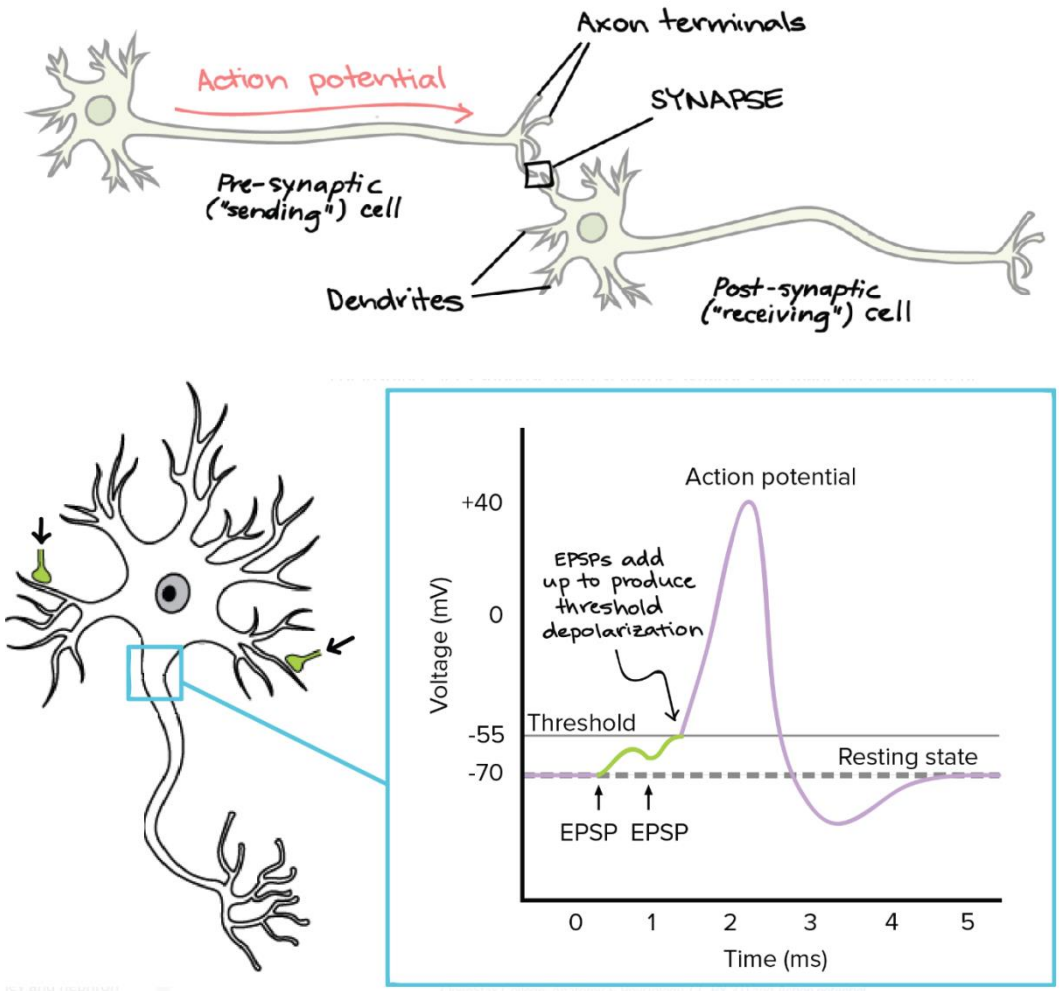


Figure 2: Simplified illustration of neuron anatomy and the occurrence of an action potential. (Image taken from Khan, 2010)

EEG measurements are made by placing electrodes in a standardized way (e.g. Figure 3) over the subject's scalp with a gel in-between to improve conductivity since the brain signal has to travel through the meninges, the thick skull and the skin causing significant deterioration to the signal. EEG signals are captured from the scalp cover multiple frequencies that are usually categorized into specific bands: delta (2 - 4 Hz), theta (4 – 8 Hz), alpha (8 – 12 Hz), beta (15 – 30 Hz), lower gamma (30 – 80 Hz), and upper gamma (80 – 150 Hz). Through the EEG signal measured on the scalp, it is possible, among other possibilities, to analyse multiple stages of auditory speech processing using event-related potentials (ERPs) (Korzyukov et al., 2012) or detect the participant's attention to specific auditory stimuli (Ding & Simon 2012).

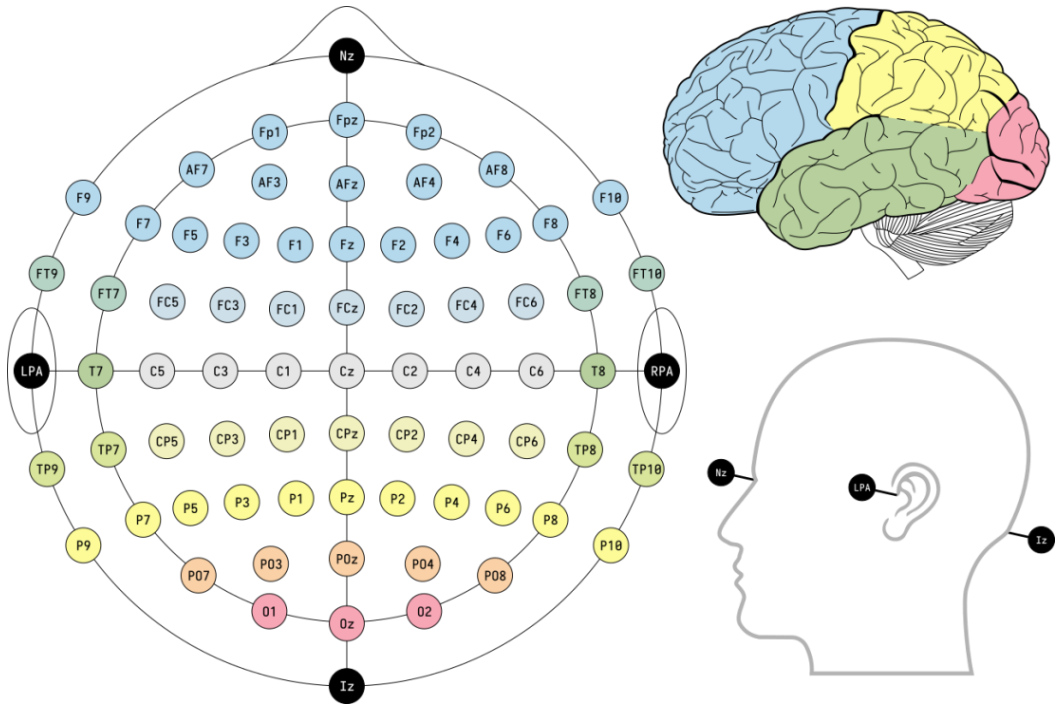


Figure 3: EEG electrode positions in the 10-10 system using modified combinatorial nomenclature, along with the fiducials and associated lobes of the brain. (Image taken from Krol, 2020)

1.3 Auditory attention detection

Using EEG data recorded when subjects were listening to two competing speakers in a binaural experiment; O'Sullivan et al. (2015) were able to identify which speaker was attended to. The reported higher correlations between the audio speech envelopes and the reconstructed EEG-speech envelopes suggest that speech-related signatures are present in EEG, even across experimental protocols and stimuli (O'Sullivan et al., 2015; Crosse et al., 2015). This principle has been later used in many other studies on auditory attention detection (e.g. Biesmans et al., 2017).

The speech envelope has also been considered to play a critical role in grouping acoustic features into auditory objects (Ding & Simon, 2012). Similarly to O'Sullivan et al. (2015), Vanthornhout et al. (2018) extracted the envelope of the EEG signal and correlated it with the auditory stimulus envelope, albeit significant, the correlations were rather small (in the order of 0.1) and a large

variability was observed between trials. Much of this can be attributed to the lower signal to noise ratio of EEG, and the more restricted bandwidth and spatial resolution compared to intracranial recordings, but also to the rather simple linear decoding models that were used. A higher performance was reported when using a relatively simple deep neural network (DNN) (De Taillez et al., 2017).

Deep neural networks have led to significantly improved performance in auditory attention detection experiments, especially when using convolutional neural networks (CNNs) (Kothari et al., 2021). These convolutional models are able to classify the attended speaker directly from EEG data and the speech envelope instead of reconstructing the attended stimulus and comparing the similarity with the actual stimuli. Using a convolutional model, Ciccarelli et al. (2018) managed to decode the attended speaker in around 10 seconds with a median accuracy of 81% which substantially outperforms models that perform classification based on an intermediate measure of similarity/correlation.

1.4 Speech classification

There have also been research advances aiming to classify speech units with EEG. Many studies showed that it is possible to decode isolated speech units such as vowels (Min et al., 2016), syllables (Brigham & Kumar, 2010), phonemes (Sun & Qin, 2016) and words (Rosinová et al., 2017) from EEG recordings. D'Zmura et al. (2009) discerned two imagined syllables (/ba/ and /ku/) from amplitude envelopes in alpha, beta and theta bands of EEG signal with up to 75% accuracy. DaSalla et al. (2009) discerned imagined vowels /a/ and /u/, and a no action state as control with 68% to 78% accuracy using epoched EEG recordings, common spatial pattern filtering, and a non-linear support vector machine (SVM) classifier. Brigham and Kumar (2010) also attempted to discern imagined syllables /ba/ and /ku/; however, they used autoregressive coefficients as features and a k-nearest-neighbor classifier which resulted in accuracy between 42% and 61%. Besides syllables, others attempted binary classification of vocalized and imagined phonemes (Zhao & Rudzicz, 2015) or imagined phonemes (Sun & Qin, 2016). Noteworthy is also the work of Min et al. (2016) on binary imagined vowel classification who achieved an accuracy ranging between 58% and 99% depending on the vowel pair and used classifier.

Rosinová et al. (2017) decoded 50 spoken commands (words) to control a service robot with accuracy up to 5%. However, decoding isolated speech units played to the subjects or imagined by them is still far from decoding continuous conversational speech. In this sense, Sharon et al. (2020) recently achieved a breakthrough by classifying 25 syllable-like units from continuous speech with an accuracy between 34% and 54%, depending on the used dataset and decoded speech mode (imagination, perception, and production). Saha et al. (2019) were also able to demonstrate notable classification results for phonological categories as they achieved accuracy between 57% and 69% using 4-layer CNN with two fully connected hidden layers. They also tried a stand-alone Long Short-Term Memory (LSTM) model which led to a decreased accuracy (below 50%), however, when combining the CNN with the LSTM model, the improved accuracy reached around 60% to 72%. They achieved their best results with an accuracy ranging between 74% and 87% when they combined the CNN and LSTM networks with a deep autoencoder (DAE) that reduced the input dimension before the classification stage (Figure 4).

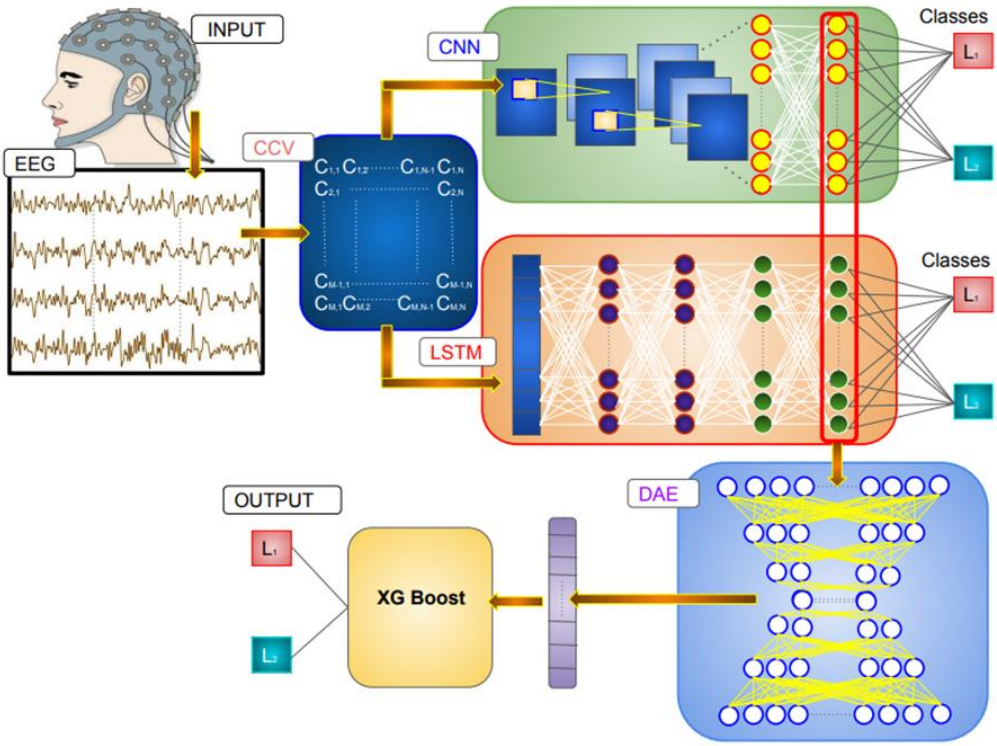


Figure 4: Combined CNN, LSTM and DAE model used by Saha et al. (2019)

1.5 Current study

Examining the relation between the speech stimulus and the EEG response is relevant for several reasons, especially as an objective measure of speech intelligibility (Iotzov & Parra, 2019; Lesenfants et al., 2019; Vanthornhout et al., 2018) serving as a potential useful diagnostics tool in the domain of audiology. Although promising, these methods come with great limitations as the correlations between the true and predicted signal are rather small (in the order of 0.1) with a large variability between trials. These studies relied on simple linear decoding models, with a two-step approach that reconstruct the attended stimulus (e.g. as speech envelope) from the subject's EEG recordings and computes similarity with the true auditory stimuli. Unfortunately, such linear models appear to not be adept for the complex nature of neural signals.

In a similar experiment, but with a simple neural network instead of a linear decoder, higher correlations were obtained (De Taillez et al., 2017). Nonetheless, correlations were still low with large variability between trials which meant long speech segments would be necessary in order to obtain consistent and reliable classification. As such, in order to avoid dealing with the regression problem, Accou et al. (2021) proposed to bypass the traditional two-step approach and use a match/mismatch paradigm. In their study, they used a dilated convolutional network that takes 3 inputs: EEG, a matched speech enveloped of the stimulus and a mismatched envelope in segments of 10s and then tries to identify the stimulus envelope corresponding to the EEG. All recordings (EEG and speech stimuli) were divided into 80%-10%-10% splits for training, validation and test respectively. The best configuration of this model reached over 90% accuracy while all configurations of the model significantly outperformed linear decoders.

The dilated convolutional network (Accou et al., 2021) performed very well on 10s segments. However, it would be interesting to test the model on smaller segments in order to see if we can still achieve good classification with shorter decision windows. The objective is to check if there is a point of diminishing results at 2s or 5s segments per example in which performance does not improve much with longer segments, thus, a segment of 10s may not be necessary.

A second interesting point is to see if the participants'' attention has any effect on the model's performance. In Accou et al.'s (2021) study, subjects actively and continuously listened to multiple stories (each story lasting around 14 to 15 minutes) presented auditory with no specific or consistent rest periods between the presentation of the stories for all subjects. Long experiments can lead to mental fatigue resulting in increased reaction times and a reduction in goal-directed attention reflected by irregular ERPs (Boksem et al., 2005). Listening to stimuli or doing tasks for long hours can cause disinterest and decreased focus of the subjects, preventing them from actively listening and from paying attention to the presented stimuli compared to their active attention at the beginning of a task. It is well known that brain signals show stronger entrainment to the temporal envelope of the attended auditory stimuli (active listening), compared to unattended speech (passive listening) (Ding & Simon, 2012; Horton et al., 2013; Kerlin et al., 2010; Mesgarani & Chang, 2012; O'Sullivan et al., 2015). Furthermore, the latency of neural activity and shape of ERP components can vary drastically depending on the participant's state of arousal and level of attention (Kong et al., 2014). As such, it is worth testing the model according to the order of stimulus presentation both according to story order and according the order of splits in each story.

Finally, in another study, Jalilpour et al. (2021) used an LSTM model with the same match/mismatch paradigm as in Accou et al. (2021). In their study, instead of using the envelope as the speech feature, they tested different levels of speech features in order to determine which level of speech information is being utilized by the model. On segments of 5s, mel spectrogram (a high level speech feature containing information about silences, intensity, and broad phonetic classes) yielded the highest accuracy (84%) among all the features. Furthermore, and maybe most importantly, voice activity detection (VAD) which only contains information about silences, yielded an impressive 75% accuracy. This miss/match global decision is made at the segment (5s) level. An interesting experiment is to try to increase the temporal resolution of those decisions by exploiting the knowledge that silences in the stimulus seem to contribute a lot to the decision (Jalilpour et al. 2021), rather than reducing the segment duration.

Chapter 2: Methodology

This chapter incorporates the explanation of the methodology used in this study. It includes information about the subjects and the dataset used in this study, and details about the models, pipelines and procedures used in the 3 experiments.

2.1 Dataset and subjects

The EEG and speech data used in the following experiment are taken from a larger dataset collected and used for the Accou et al. (2021) and Jalilpour et al. (2021) studies. EEG data was recorded (using a 64-channel system ”ActiveTwo, BioSemi” at a sampling frequency of 8192 Hz) while 63 normal-hearing native Flemish-speaking subjects listened to 10 Flemish stories lasting approximately 14 minutes and 30 seconds each. Subjects were instructed to attentively listen to the presented story and were notified beforehand that content-related questions will be asked at the end of the story in order to motivate them to listen to the story actively. The exact number of stories chosen from the larger datatset for each subject is detailed on Table 1. In total, we have EEG 591 recordings.

Number of subjects	**Stories per subject**
1	4
1	7
5	8
20	9
36	10

Table 1: Number of stories per subject in the current dataset

2.2 Experiment 1

In this first experiment, we wanted to test the dilated convolutional model used by Accou et al. (2021) on smaller segment lengths. The model was used on 10s segment lengths and had median performances ranging between 85% and 91% (a little over 90% for the best configuration). Therefore, we wanted to also test it on smaller segments: 5s, 2s and 1s. The objective here is to see to watch extent does the performance of the model decrease with shorter segments (smaller decision windows). Similar to the Accou et al. (2021) experiment, all recordings (EEG and speech stimuli) were divided into 80%-10%-10% splits for training, validation and test respectively (Figure 5).

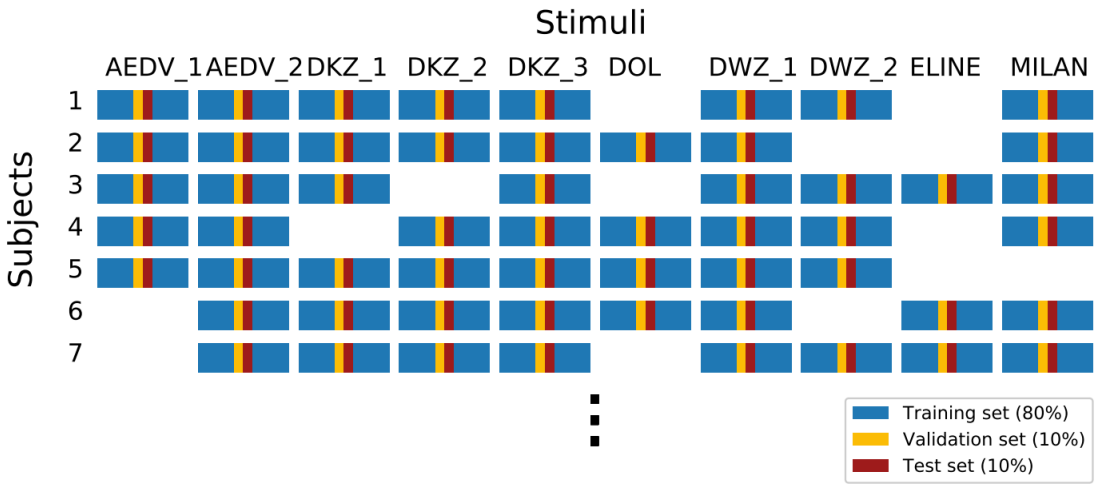


Figure 5: Visualization of data of 7 first subjects within the dataset. Each rectangle represents a recording. (Image taken from Accou et al., 2021)

In this match/mistach paradigm, the model takes in 3 inputs, one for EEG data segments and two for speech envelope segments. Data was presented to the models 4 times, each time according to a different segment length (10s, 5s, 2s and 1s) with an overlap of 90%. The imposter speech envelope segment starts one second after the end of the matched speech envelope segment (Figure 6).

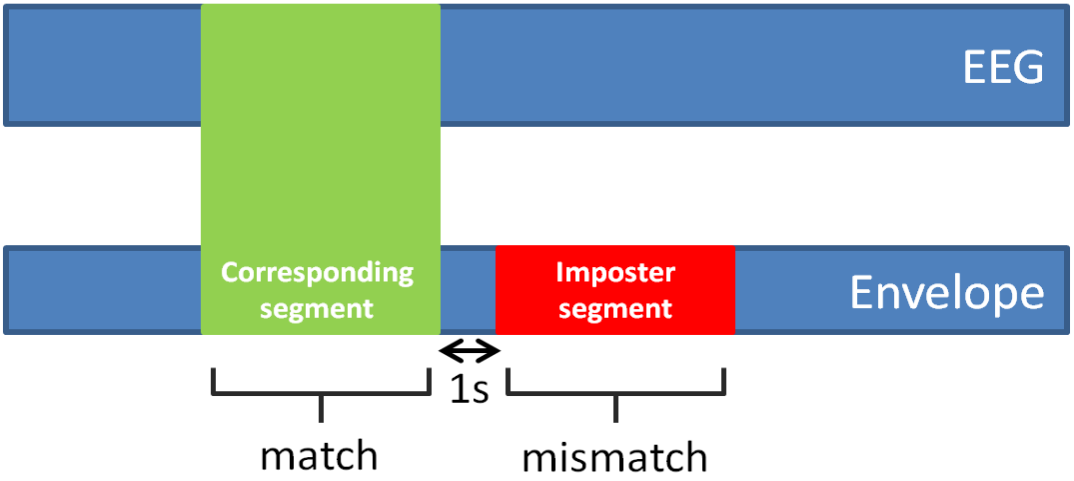


Figure 6: Extracting match and mismatch envelope spectrograms. Segment length set at 10s, 5s, 2s and 1s according to each trial.

The dilated convolutional network is shown in Figure 7. To compensate for the brain delay following the presentation of an auditory stimulus, the model applies an integration window of 250ms as a convolution that slides over the EEG segment and linearly combines all channels over the next 250ms into a reconstructed speech envelope sample. A single sigmoid neuron is fed the cosine similarity between both input envelopes and the reconstructed envelope. Then, the classification of which envelope input matches with the EEG is computed by this sigmoid neuron.

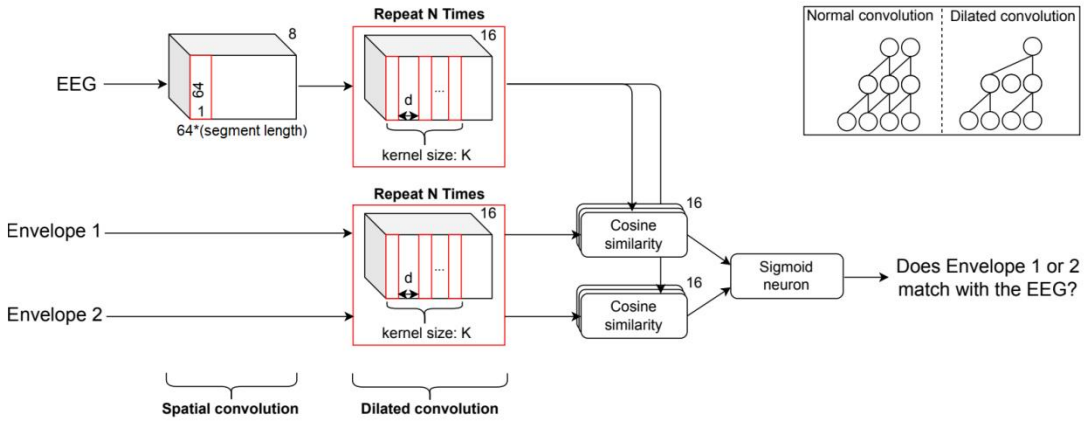


Figure 7: The structure of the dilated convolutional network. (Image adapted from Accou et al., 2021)

As described by Accou et al. (2021), the model starts first with a convolutional layer with 8 filters that combine all EEG channels, on a spatial and linear level. Then, N dilated convolutions with kernel size K are applied to the spatially filtered EEG and the two stimulus envelopes. In this experiment we used N=3 dilated convolutions with kernel size K=3 as this was the configuration with the best performance. The dilation factor d for layer Ln is chosen to be Kn−1 (as in van den Oord et al., 2016) in order to minimize the amount of parameters per receptive field size. After each dilated convolution, a rectified linear unit (ReLU) nonlinearity is used. Cosine similarity is applied in order to compare each EEG representation to each stimulus representation after dilated convolutions. Then, based on the cosine similarity scores, the sigmoid layer classifies match/mismatch.

2.3 Experiment 2

In the second experiment, we wanted to check if participants' attention, as reflected by story order and segment order within stories, had any effect on the model's performance. While we don't have exact information about how many times and when each subject took a break, nor any objective measure of the subjects' focus and attention while listening to the stories, we have the presentation order of the stories. Hence, we could assume that as the experiment extends, the participants are more likely to get tired or bored, and thus, be more likely to be inattentive to the auditory stimuli. Aside from story order, we can also investigate this effect according to segments within the stories themselves. While a participant is more likely to be more attentive during the first story than during the presentation of the seventh story per example, it is also conceivable that a participant might be more attentive at the beginning of each story than at the middle or end of the story. Accordingly, we can train the model and then test it according to story order as well as according to segment order with the stories.

The same model used in experiment 1 (as displayed in Figure 7) is used in this experiment with 10s segments. However, the organization of the data the model is trained and tested on is different. Out of the 63 subjects, we randomly picked 20 participants for which we have 10 EEG recordings. Those 20 participants'

brain recordings were kept apart as hold out data for testing and were not included in training. The EEG recordings of the remaining 43 participants were used for training and validation. Each recording was split 90%-10% for training and validation respectively.

The model was trained once and then tested multiple times according to our paradigm. The model was tested 10 times on the holdout data according to story order such as the first time it was only tested on the first story each subject listened to. Then it was tested on the second story, and then the third, and so forth up to the tenth story (see Figure 8).

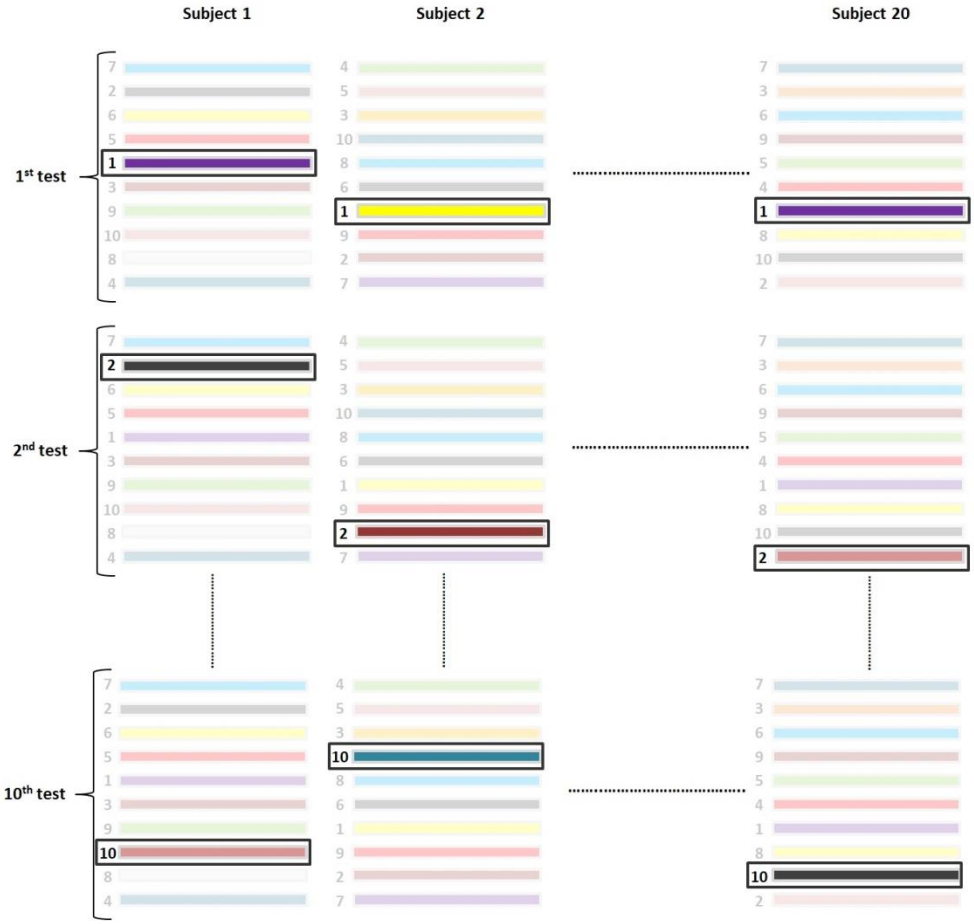


Figure 8: Organization of testing data according to story presentation order. Each colored rectangle represents a recording while the numbers represent the order of the presentation of the story associated with that recording. Each bold highlighted rectangle reprents the data used in that testing phase.

Furthermore, the model was also tested 10 times according to splits within stories. Each EEG recording and speech envelope were split into 10 equal partitions. For the first test, the model was tested on the first partition of each recording, which is the first 10% of each recording. Subsequently, the model was tested on the second 10% split of each recording, then the third, and so forth up to the 10th and final 10% split (see Figure 9).

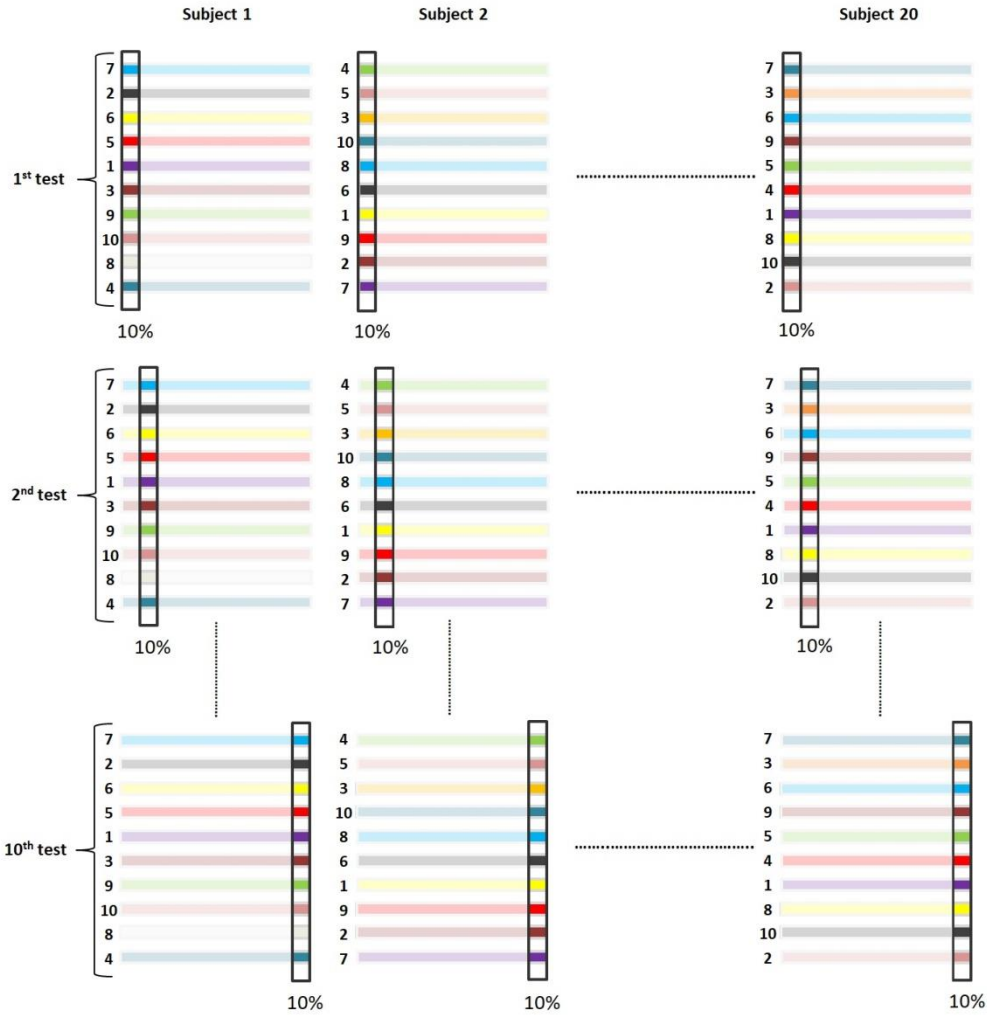


Figure 9: Organization of testing data according to split/segment order within stories. Each colored rectangle represents a recording and the numbers represent the order of the presentation of the story associated with that recording. Each bold highlighted rectangle reprents the data used in that testing phase.

2.4 Experiment 3

In the third and final experiment, we want to exploit the knowledge that silences in the stimulus seem to contribute a lot to the decision (Jalilpour et al., 2021), rather than reducing the segment duration to increase the temporal resolution of our decisions. We will use the same LSTM model (Figure 10) used by Jalilpour et al. (2021) with the same match/mismatch paradigm on 5s segments. As speech feature, we will use AnyPhoneme instead of VAD. AnyPhoneme is a one-hot vector which determines at each time step whether it is silence (0) or any phoneme (1). We will first run this model to set a reference for the accuracy of the model. We expect to get around 72% accuracy wich is the accuracy yielded by the model in Jalilpour et al.'s (2021) experiment. The data is divided into 80%-10%-10% splits for training, validation and test respectively as in Figure 5.

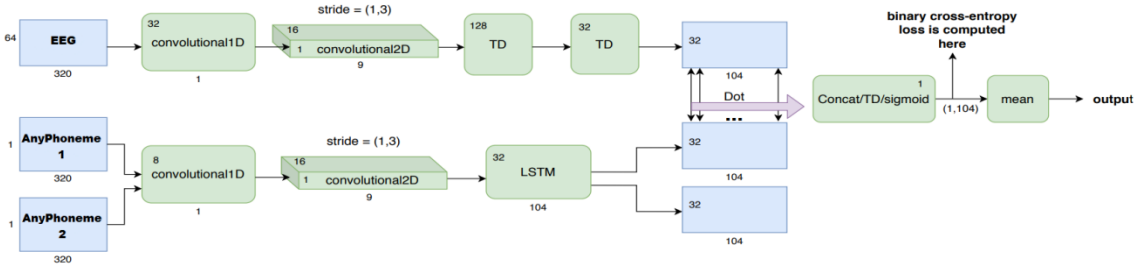


Figure 10: The LSTM-based model for match/mismatch classification. TD refers to time distributed which applies a dense layer to every temporal slice of the input. Dot is a layer that applies dot product (cosine similarity) between EEG representation and speech representation for each time step. (Image taken from Jalilpour et al., 2021)

We will then use a simple bidirectional LSTM classifier that takes in 5s segments of EEG and predicts silences in the speech. For that, the model will classify each datapoint in the EEG segment as 0 for silences and 1 for speech (zeros and ones vector similar to AnyPhoneme). The model will be trained on the same data used for the LSTM match/mistach model. Since the correlations in predictions are quite low, we do not expect very accurate point-wise decisions. We expect an accuracy of around 52%. Nevertheless, aligning speech and EEG (i.e. estimating brain response times) could be possible. As such, we would like to know if the 52% accuracy will result in around 72% accuracy in the match/mismatch task. To do

that we will use the output of the bidirectional LSTM classifier, the zeros and ones vector output from each 5s EEG segment input. We will then try to pick the correct/matched AnyPhoneme segment by comparing our output vector with the 2 AnyPhoneme segments and picking the segment with the most matching datapoints (see Figure 11).

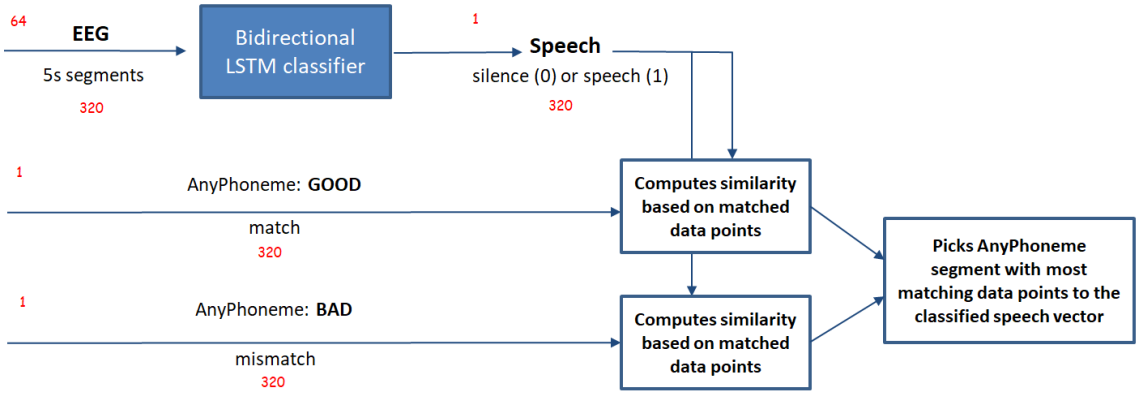


Figure 11: Structure of the proposed match/mismatch paradigm using a bidirectional LSTM classifier.

Chapter 3: Results

This chapter is reserved to presenting, analysing, and discussing the findings of the study.

3.1 Experiment 1

In this first experiment, we wanted to test the dilated convolutional model used by Accou et al. (2021) on smaller segment lengths. We trained and tested the model on 4 different segment lengths: 10s, 5s, 2s and 1s. Our tests revealed decreasing classification accuracy as the segment length got smaller. The median accuracy for the 10s segments was around 88%, around 81% for 5s segments, 70% for 2s segments and 64% for 1s segments (Figure 12).

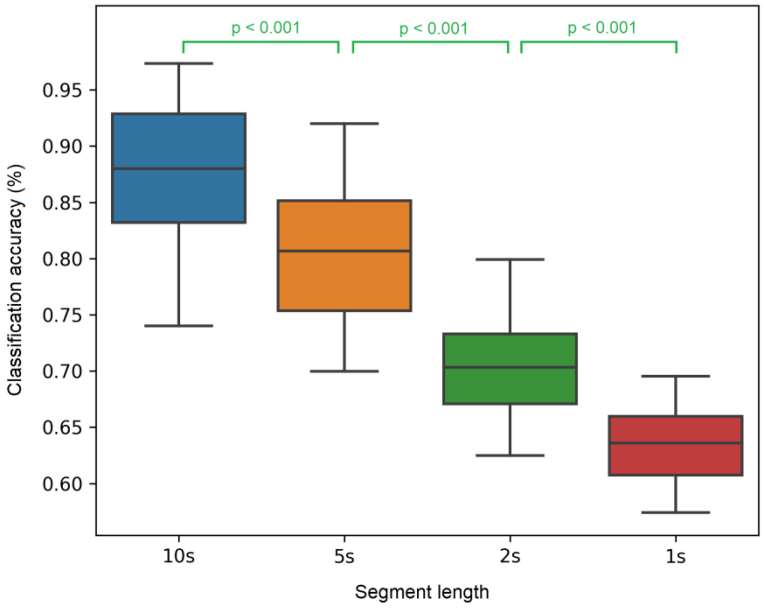
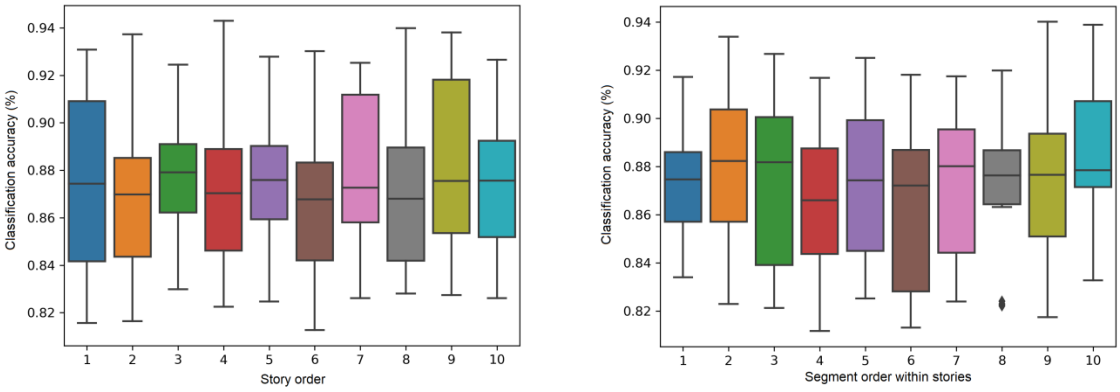


Figure 12: Classification accuracy per segment length. Boxplots represent 63 subjects.

By using Wilcoxon signed rank test with Holm-Bonferroni correction for multiple comparisons, we found that the classification accuracy according to each segment length outperformed the classification accuracy of the other smaller segments significantly ($p < 0.001$). These results were expected as longer segments mean longer decision windows and more data points allowing the model to make better decisions. If segments get longer, the improvement of the model will probably not increase drastically since it is already at almost 90% accuracy for 10s segments.

3.2 Experiment 2

In the second experiment, we trained the model once (on segments of 10 seconds) and then tested it multiple times on holdout data according to story order and according to segment order within stories. Classification accuracies according to story order and according to segment order within stories are presented on Figure 13.



(a) Classifictaion accuracy according to story order

(b) Classifictaion accuracy according to segment order within stories

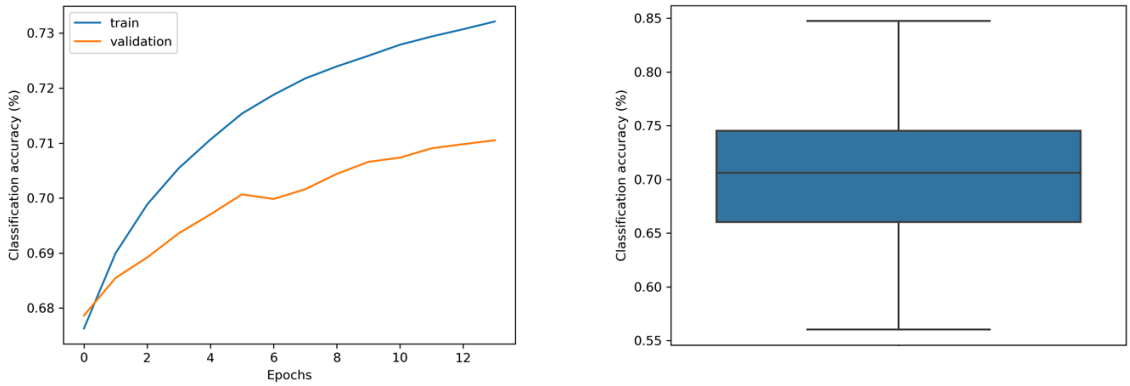
Figure 13: Classification accuracy according to (a) story order and (b) segment order with stories. Boxplots represent 20 subjects.

All median classification accuracies ranged between 86% and 88%. All statistical comparisons were insignificant demonstrating that the performance of the model is not affected by story order or within story segment order. However, this result does not necessarily imply that the performance of the model is not affected by the participants' active attention and focus on the presented auditory stimuli.

While we assumed that participants may have been bored, fatigued and less attentive while listening to stories at the end of the experiment compared to the beginning, we cannot assert that with certainty. We also do not know exactly how often each participant took a break during the EEG acquisition session. If participants took breaks often, then looking at story order would be pointless. Comparing accuracy performance according to within story segment order might also be limited due to the length of each story being around 14 minutes which may not be long enough to tire or bore a participant as to to render him less attentive to the stimuli. It is also possible that those effects are existent but not strong enough to affect the model. The model can also be basing its classification decisions on robust speech features that are not easily affected by the participant attention. In all cases, in order to make a correct conclusion on whether such a classification model is affected by the participant' active attention to the stimuli, the experimental design and data acquisition should be carried out with this objective in mind in order to acquire data in two different conditions: active and passive listening. Then, a classification model can be tested on this data in order to reach a more comprehensive conclusion.

3.3 Experiment 3

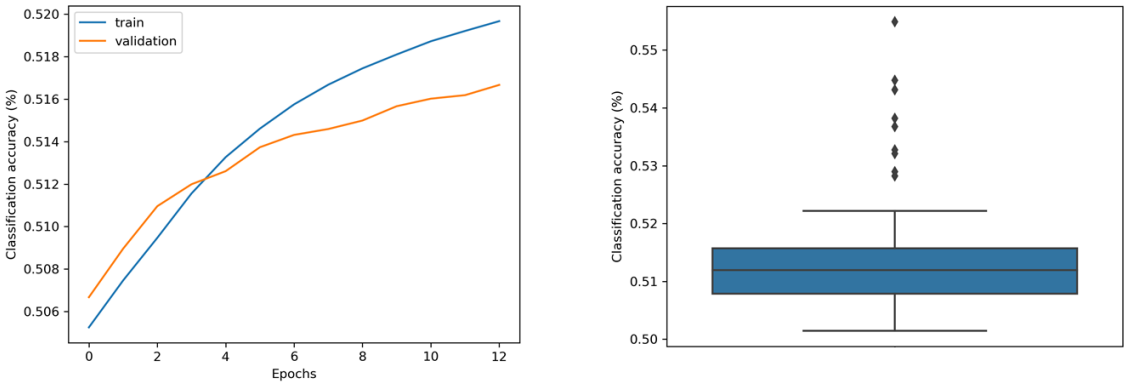
In this last experiment, we first tested the LSTM match/mismatch model on 5s segments of EEG and AnyPhoneme. As expected, we got a median classification accuracy of around 71% (Figure 14). This result is similar to the classification accuracy achieved by Jalilpour et al. (2021). This classification accuracy will serve as a baseline that we will compare to the accuracy we get with our linear classification after we reconstruct AnyPhoneme from EEG segments.



(a) Training and validation (b) Testing

Figure 14: (a) Training, validation and (b) testing of the LSTM match/mismatch model. Boxplot represents 63 subjects.

We then used the bidirectional LSTM classifier to reconstruct AnyPhoneme vector (0 for silences, and 1 for speech). As expected, the model did not produce very accurate point-wise decisions as the classification accuracy was around 51% (Figure 15). The labels are balanced so the chance level is 50%.



(a) Training and validation (b) Testing

Figure 15: (a) Training, validation and (b) testing of the bidirectional LSTM classifier. Boxplot represents 63 subjects.

The reconstructed speech (AnyPhoneme) vector from EEG segments was then used to pick the matching AnyPhoneme segment by linearly picking the speech segment with the most matching data points. This pipeline produced an accuracy of around 60% which is significantly low ($p<0.001$) compared the LSTM match/mismatch pipeline (Figure 16). While we did not achieve the targeted accuracy, this pipeline still produced a classification accuracy above change level.

Despite the accuracy of predicting silences at each data point being only 51%, we managed to achieve 60% match/mismatch classification using the reconstructed speech vector and a simple linear process. These results further confirm that information about silences is very useful in this classification task. It also appears that the model does not really rely on whether or not there is silence at each data point, but rather on the distrubution and concentration of silences in a given part of a segment.

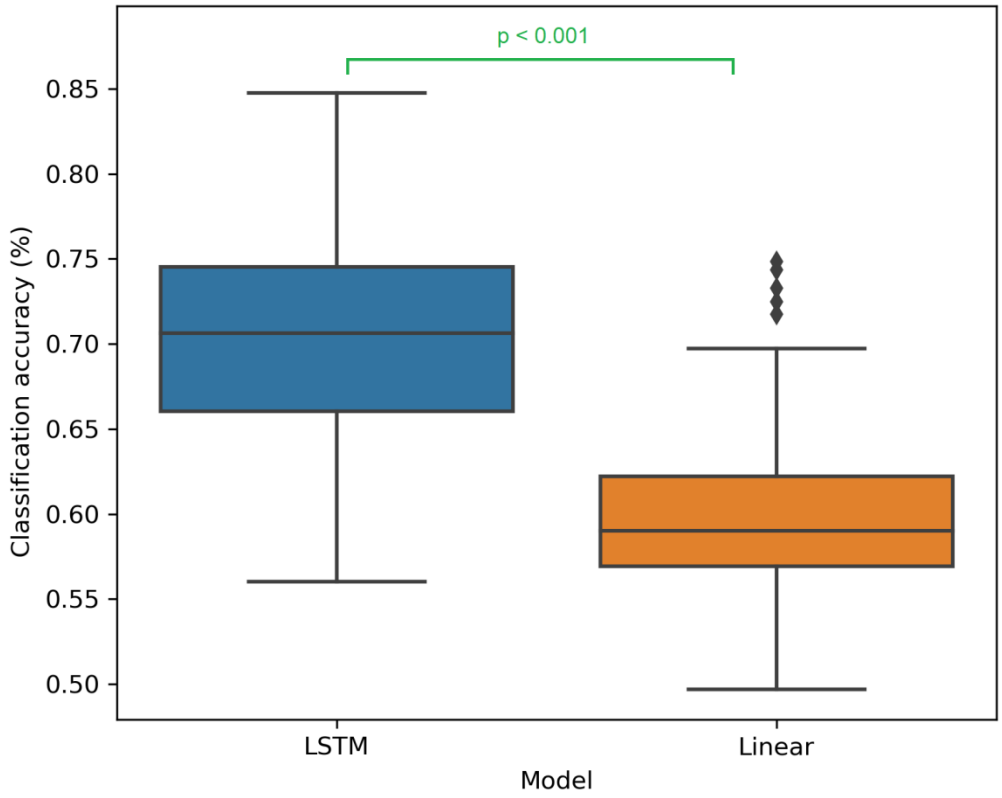


Figure 16: Classification accuracy of the LSTM match/mismatch model and the Linear match/mismatch classification based on the reconstructed AnyPhoneme speech vector.

Chapter 4: Conclusion

The main objectives of this study were to investigate the effects of attention according to the order of stimulus presentation on the performance of the model and to explore the relevance of silences to the match/mismatch classification by trying to increase the temporal resolution of our decisions using reconstructed speech to align EEG and speech stimuli.

In the first experiment, we wanted to test to what extent the classification accuracy will be affected when we decrease the decision windows. We trained and tested the model on 4 different segment lengths: 10s, 5s, 2s and 1s. The performance of the model was best at 10s segments and decreased significantly with smaller decision windows. While, it is true that classification performance will not keep increasing with longer segments as it will reach a point of diminishing returns at a certain threshhold, for segments between 1s and 10s, every increase in decision windows leads to substantial imporvements to the model's classification accuracy.

In the second experiment, we wanted to verify if there are any effects of participants' attention on the performance of the model. No significant effect was found neither according to story order nor according to segment order within stories. This result is not enough to assert that no such effect exists on a similar classification model because the data collection and the experimental paradigm was not designed with this experiment in mind. An effect may exist and could affect the classifictaion of a similar model, however, we can for sure say that the order of presentation of the auditory stimuli (either between or with stories) in the context of this dataset, has no effect on the classification accuracy of the model.

In the final experiment, we aimed to exploit the importance of silences for the decisions of the model in order to investigate the possibility of achieving equivalent performance by reconstructing speech and silence information from EEG, and then matching it to the corresponding speech segment. While the classification accuracy of the reconstructed speech was only 51%, with the reconstructed speech vector and a linear matching paradigm we managed to achieve 60% accuracy further confirming the relevance of silences to the decisions.

Bibliography

Accou, B., Jalilpour Monesi, M., Montoya, J., Van hamme, H., & Francart, T. (2021). Modeling the relationship between acoustic stimulus and EEG with a dilated Convolutional Neural Network. *2020 28th European Signal Processing Conference (EUSIPCO)*. https://doi.org/10.23919/eusipco47968.2020.9287417

Assaneo, M. F., & Poeppel, D. (2018). The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science Advances*, *4*(2). https://doi.org/10.1126/sciadv.aao3842

Baby, D., & Verhulst, S. (2018). Biophysically-inspired features improve the generalizability of neural network-based Speech Enhancement Systems. *Interspeech 2018*. https://doi.org/10.21437/interspeech.2018-1237

Biesmans, W., Das, N., Francart, T., & Bertrand, A. (2017). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(5), 402–412. https://doi.org/10.1109/tnsre.2016.2571900

Bigliassi, M., Karageorghis, C. I., Wright, M. J., Orgs, G., & Nowicky, A. V. (2017). Effects of auditory stimuli on electrical activity in the brain during cycle ergometry. *Physiology & Behavior*, *177*, 135–147. https://doi.org/10.1016/j.physbeh.2017.04.023

Bocquelet, F., Hueber, T., Girin, L., Chabardès, S., & Yvert, B. (2016). Key considerations in designing a speech brain-computer interface. *Journal of Physiology-Paris*, *110*(4), 392–401. https://doi.org/10.1016/j.jphysparis.2017.07.002

Boksem, M. A. S., Meijman, T. F., & Lorist, M. M. (2005). Effects of mental fatigue on attention: An ERP study. *Cognitive Brain Research*, *25*(1), 107–116. https://doi.org/10.1016/j.cogbrainres.2005.04.011

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., & Formisano, E. (2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *Journal of Neuroscience*, *34*(13), 4548–4557. https://doi.org/10.1523/jneurosci.4339-13.2014

Brigham, K., & Kumar, B. V. (2010). Imagined speech classification with EEG signals for Silent Communication: A preliminary investigation into synthetic telepathy. *2010 4th International Conference on Bioinformatics and Biomedical Engineering*. https://doi.org/10.1109/icbbe.2010.5515807

Butler, B. E., & Trainor, L. J. (2012). Sequencing the cortical processing of pitch-evoking stimuli using EEG analysis and source estimation. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00180

Ciccarelli, G., Nolan, M., Perricone, J., Calamia, P. T., Haro, S., O'Sullivan, J., Mesgarani, N., Quatieri, T. F., & Smalt, C. J. (2019). Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-47795-0

Clayton, J., Wellington, S., Valentini-Botinhao, C., & Watts, O. (2020). Decoding imagined, heard, and spoken speech: Classification and regression of EEG using a 14-channel dry-contact mobile headset. *Interspeech 2020*. https://doi.org/10.21437/interspeech.2020-2745

Correia, J. M., Jansma, B. M., & Bonte, M. (2015). Decoding articulatory features from fmri responses in dorsal speech regions. *Journal of Neuroscience*, *35*(45), 15015–15025. https://doi.org/10.1523/jneurosci.0977-15.2015

Correia, J., Formisano, E., Valente, G., Hausfeld, L., Jansma, B., & Bonte, M. (2013). Brain-based translation: Fmri decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *Journal of Neuroscience*, *34*(1), 332–338. https://doi.org/10.1523/jneurosci.1302-13.2014

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, *35*(42), 14195–14204. https://doi.org/10.1523/jneurosci.1829-15.2015

DaSalla, C. S., Kambara, H., Sato, M., & Koike, Y. (2009). Single-trial classification of vowel speech imagery using common spatial patterns. *Neural Networks*, *22*(9), 1334–1339. https://doi.org/10.1016/j.neunet.2009.05.008

de Taillez, T., Kollmeier, B., & Meyer, B. T. (2017). Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. *European Journal of Neuroscience*, *51*(5), 1234–1241. https://doi.org/10.1111/ejn.13790

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. https://doi.org/10.1073/pnas.1205381109

D'Zmura, M., Deng, S., Lappas, T., Thorpe, S., & Srinivasan, R. (2009). Toward EEG sensing of imagined speech. *Human-Computer Interaction. New Trends*, 40–48. https://doi.org/10.1007/978-3-642-02574-7_5

Evans, S., Kyong, J. S., Rosen, S., Golestani, N., Warren, J. E., McGettigan, C., Mourao-Miranda, J., Wise, R. J., & Scott, S. K. (2013). The pathways for intelligible speech: Multivariate and univariate perspectives. *Cerebral Cortex*, *24*(9), 2350–2361. https://doi.org/10.1093/cercor/bht083

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "who" is saying "what"? brain-based decoding of human voice and speech. *Science*, *322*(5903), 970–973. https://doi.org/10.1126/science.1164318

Friston, K. J., Jezzard, P., & Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping*, *1*(2), 153–171. https://doi.org/10.1002/hbm.460010207

Ganushchak, L. Y., Christoffels, I. K., & Schiller, N. O. (2011). The use of Electroencephalography in language production research: A Review. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00208

Geravanchizadeh, M., & Zakeri, S. (2021). Ear-EEG-based Binaural Speech Enhancement (EE-BSE) using auditory attention detection and audiometric characteristics of hearing-impaired subjects. *Journal of Neural Engineering*, *18*(4), 0460d6. https://doi.org/10.1088/1741-2552/ac16b4

Giraud, A. L., Garnier, S., Micheyl, C., Lina, G., Chays, A., & Chéry-Croze, S. (1997). Auditory efferents involved in speech-in-noise intelligibility. *NeuroReport*, *8*(7), 1779–1783. https://doi.org/10.1097/00001756-199705060-00042

Hagoort, P., & Brown, C. M. (2000). Erp effects of listening to speech: Semantic erp effects. *Neuropsychologia*, *38*(11), 1518–1530. https://doi.org/10.1016/s0028-3932(00)00052-x

Horton, C., D'Zmura, M., & Srinivasan, R. (2013). Suppression of competing speech through entrainment of cortical oscillations. *Journal of Neurophysiology*, *109*(12), 3082–3093. https://doi.org/10.1152/jn.01026.2012

Iotzov, I., & Parra, L. C. (2019). EEG can predict speech intelligibility. *Journal of Neural Engineering*, *16*(3), 036008. https://doi.org/10.1088/1741-2552/ab07fe

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *Journal of Neuroscience*, *30*(2), 620–628. https://doi.org/10.1523/jneurosci.3631-09.2010

Khan, S. (2010). *The neuron and nervous system: The synapse*. Khan Academy. Retrieved from https://www.khanacademy.org/science/biology/human-biology/neuron-nervous-system/a/the-synapse.

Kong, Y.-Y., Mullangi, A., & Ding, N. (2014). Differential modulation of auditory responses to attended and unattended speech in different listening conditions. *Hearing Research*, *316*, 73–81. https://doi.org/10.1016/j.heares.2014.07.009

Korzyukov, O., Karvelis, L., Behroozmand, R., & Larson, C. R. (2012). ERP correlates of auditory processing during automatic correction of unexpected perturbations in voice auditory feedback. *International Journal of Psychophysiology*, *83*(1), 71–78. https://doi.org/10.1016/j.ijpsycho.2011.10.006

Kothari, M., Joshi, S., Nandanwar, A., Jaiswal, A., & Baths, V. (2021). Deep Neural Networks on EEG Signals to Predict Auditory Attention Score Using Gramian Angular Difference Field. https://doi.org/https://doi.org/https://doi.org/10.48550/arXiv.2110.12503

Krol, L. R. (2020). *Eeg electrode positions in the 10-10 system using modified combinatorial nomenclature, along with the fiducials and associated lobes of the brain.* Wikipedia. Retrieved from https://en.wikipedia.org/wiki/10%E2%80%9320_system_(EEG).

Lesenfants, D., Vanthornhout, J., Verschueren, E., & Francart, T. (2019). Data-driven spatial filtering for improved measurement of cortical tracking of multiple representations of speech. *Journal of Neural Engineering*, *16*(6), 066017. https://doi.org/10.1088/1741-2552/ab3c92

Martin, S. Ã., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N. E., Rieger, J., Schalk, G., Knight, R. T., & Pasley, B. N. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in Neuroengineering*, *7*. https://doi.org/10.3389/fneng.2014.00014

Martin, S., Iturrate, I., Millán, J. del, Knight, R. T., & Pasley, B. N. (2018). Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00422

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236. https://doi.org/10.1038/nature11020

Michel, C. M., & Murray, M. M. (2012). Towards the utilization of EEG as a brain imaging tool. *NeuroImage*, *61*(2), 371–385. https://doi.org/10.1016/j.neuroimage.2011.12.039

Min, B., Kim, J., Park, H.-jun, & Lee, B. (2016). Vowel imagery decoding toward silent speech BCI using extreme learning machine with Electroencephalogram. *BioMed Research International*, *2016*, 1–11. https://doi.org/10.1155/2016/2618265

Moumdjian, L., Buhmann, J., Willems, I., Feys, P., & Leman, M. (2018). Entrainment and synchronization to auditory stimuli during walking in healthy and neurological populations: A methodological systematic review. *Frontiers in Human Neuroscience*, *12*. https://doi.org/10.3389/fnhum.2018.00263

Murphy, A., Bohnet, B., McDonald, R., & Noppeney, U. (2022). Decoding part-of-speech from human EEG signals. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, *1*. https://doi.org/10.18653/v1/2022.acl-long.156

Nourski, K. V. (2017). Auditory processing in the human cortex: An intracranial electrophysiology perspective. *Laryngoscope Investigative Otolaryngology*, *2*(4), 147–156. https://doi.org/10.1002/lio2.73

Ombao, H., Lindquist, M., Thompson, W., & Aston, J. (2019). *Handbook of Neuroimaging Data Analysis*. CRC press.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. https://doi.org/10.1093/cercor/bht355

Proix, T., Delgado Saa, J., Christen, A., Martin, S., Pasley, B. N., Knight, R. T., Tian, X., Poeppel, D., Doyle, W. K., Devinsky, O., Arnal, L. H., Mégevand, P., & Giraud, A.-L. (2022). Imagined speech can be decoded from low- and

cross-frequency intracranial EEG features. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-021-27725-3

Rosinova, M., Lojka, M., Stas, J., & Juhar, J. (2017). Voice command recognition using EEG signals. *2017 International Symposium ELMAR*. https://doi.org/10.23919/elmar.2017.8124457

Saha, P., Fels, S., & Abdul-Mageed, M. (2019). Deep learning the EEG manifold for phonological categorization from active thoughts. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp.2019.8682330

Sanders, L. D., & Neville, H. J. (2003). An ERP study of continuous speech processing. *Cognitive Brain Research*, *15*(3), 228–240. https://doi.org/10.1016/s0926-6410(02)00195-7

Schuenke, M., Schulte, E., Schumacher, U., MacPherson, B., Stefan, C., Wesker, K. H., & Voll, M. (2020). *Head, neck, and neuroanatomy (thieme atlas of anatomy)*. Thieme Medical Publishers.

Singh, S. P. (2014). Magnetoencephalography: Basic principles. *Annals of Indian Academy of Neurology*, *17*(5), 107. https://doi.org/10.4103/0972-2327.128676

Sun, P., & Qin, J. (2016). Neural networks based EEG-speech models. https://doi.org/http://arxiv.org/abs/1612.05369

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. https://doi.org/https://doi.org/10.48550/arXiv.1609.03499

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., & Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *Journal of the Association for Research in Otolaryngology*, *19*(2), 181–191. https://doi.org/10.1007/s10162-018-0654-z

Zelmann, R., Lina, J. M., Schulze-Bonhage, A., Gotman, J., & Jacobs, J. (2013). SCALP EEG is not a blur: It can see high frequency oscillations although their generators are small. *Brain Topography*, *27*(5), 683–704. https://doi.org/10.1007/s10548-013-0321-y

Zhao, S., & Rudzicz, F. (2015). Classifying phonological categories in imagined and articulated speech. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp.2015.7178118