Radboud University

Bachelor Thesis

# Transient Generalization Over Different Presentation Frequencies: Lagged Stimuli and Superposition

*Author:*
Pleun Scholten[1]
s4822250

*Supervisors:*
Prof. Peter W.M. Desain[1,2,3]
Dr. Ceci S. Verbaarschot[1,2,3]
Dr. Sara Ahmadi[2]

*A thesis presented for the degree of*
*Bachelor of Science in Artificial Intelligence*

[1]Department of Artificial Intelligence
[2]Donders Institute for Brain, Cognition and Behaviour
[3]Donders Centre for Cognition
Radboud University

July 24, 2020

Radboud University

# Acknowledgments

I would like to thank my supervisors Peter Desain, Ceci Verbaarschot, and Sara Ahmadi, as well as Jordy Thielen for their help. Also, many thanks to Marzieh Borhanazad for allowing me to use her data.

**Abstract**

Attempts to make brain-computer interfaces more generalizable have seen some success. The generative model of Thielen et al. has been able to accurately predict responses to unseen stimuli. However, this model still requires the unseen stimuli to be presented at the same frequency. This is because the transients are assumed to stay the same in the generative model, which is a false assumption when the presentation rate changes. In this thesis, two methods of generalizing transient responses across different presentation rates are investigated. The first method uses lagged stimuli: stimuli where the rate is changed by leaving the length of the stimulus-ON the same, but changing the length of the stimulus-OFF. The second method uses superposition, i.e. linear addition of responses, in order to predict transient responses at different presentation rates. The first method using lagged stimuli showed promise. Classification accuracies were significantly better than chance when predicting responses to stimuli at a rate that was different from the training rate. At the same time, however, the correlations between the transients of the same event at different rates were not significant. The second method using superposition showed no significant results. Some correlations between real transients and transients that were predicted using superposition were significant, but the accuracies were at chance level.

# Contents

# 1 Introduction

Brain-Computer Interfaces (BCI) allow humans to control external devices without the need for muscle control [1]. This can be greatly beneficial for those who cannot communicate via conventional ways, like locked-in patients (LIP) [2, 3]. BCIs can rely on different kinds of brain signals, for instance slow cortical potentials [3], sensorimotor rhythms [4], or visually evoked potentials (VEP) [1, 5, 6]. The latter signals currently allow for the highest information transfer rate, with some VEP-based BCI spellers able to reach a speed of 60 characters per minute [6].

The BCI cycle consists of six distinct stages, as defined in [7]: stimulation, measurement, preprocessing, feature extraction, prediction, and output. This thesis will mostly focus on prediction. The basis of VEP-based BCIs is the fact that foveated and attended flickering targets create a stronger brain response than targets outside the attended area [1]. VEPs arise from the occipital cortex, since visual information is processed there [8, 9]. The stimuli are mapped to computer commands, which determines the output, and by attending a stimulus, this stimulus is more easily recognized from the signal.

Code modulated visually evoked potential (c-VEP) based BCIs are a kind of BCI that use pseudorandom binary codes to determine a flicker pattern [1]. The advantage of using pseudorandom flicker patterns over for instance a steady flicker pattern as used in steady state visually evoked potentials (SSVEP), is that c-VEPs are broad band signals. This makes these signals much more resilient to noise, both externally generated noise, like other electrical devices, or internally generated noise, like muscle movement.

Thielen et al. designed a generative model to predict the brain response to a stimulus [8, 10]. However, the generative model has so far only been used to predict stimulus responses within one presentation rate. If for example a classifier has been trained on data at a presentation rate of 60Hz, the model cannot predict what a response would look like to an event at a presentation rate of 90Hz.

Being able to generalize learned transients across different presentation rates can be of great use, for example in predicting the optimal stimulation frequency for a set of codes. In the code domain, the stimuli with minimal cross- and autocorrelation can be easily calculated. The responses to those stimuli, however, can still be non-optimal, since the responses to the individual events within each stimulus overlap, as illustrated in Figure 1.

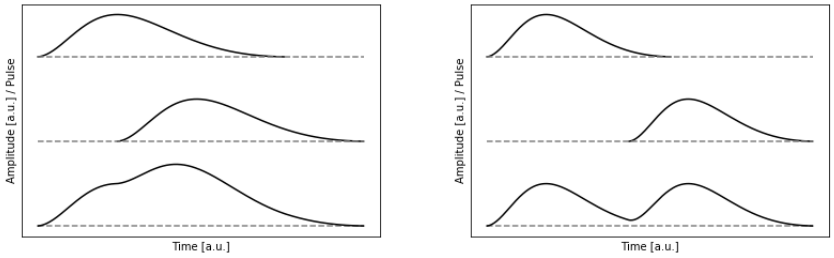



Figure 1: Two imaginary event responses (top and middle row), shifted over time, giving two different overall responses (bottom row)

In this thesis, two different methods to generalize transients across different presentation rates will be investigated:

1) Using lagged stimuli

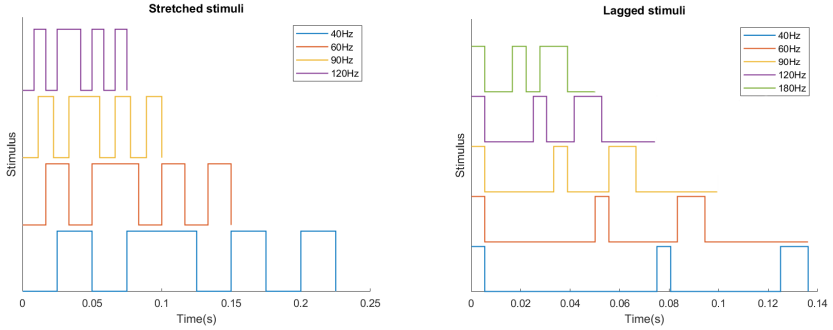
2) Using superposition

Both methods of generalization first require an assumption to be tested. First, the use of lagged stimuli will be explained in Section 1.1, and afterward, in Section 1.2, the use of superposition will be explained.

## 1.1 Generalizing over Lagged Data

The frequency of stimuli can be changed in different ways. The first way is a simple *stretching* of the code, i.e. both the stimulus-ON, where the binary code is 1, as well as the stimulus-OFF, or run-off, where the binary code is 0, are changed by the same ratio, as in Figure 2a. Henceforth, stimuli using this method of changing presentation rate are called *stretched*. The second way is found by introducing a *lag* within the code, i.e. only changing the run-off, leaving the stimulus-ON the same length, as in Figure 2b. Henceforth, stimuli using this method of changing presentation rate are called *lagged*.

When stretching a code, the stimulus-ON length changes, and so we cannot assume that the responses to the events within the stimulus stay the same. For lagged codes however, the size of the stimulus-ON stays the same, so it is possible that the event-responses stay the same when a lagged code is presented at different rates.




(a) Stretched codes

(b) Lagged codes

Figure 2: 'stretched' codes versus 'lagged' codes. Both methods change the frequency of the code, but in a different way.

In theory, the responses to events from lagged stimuli from one rate can be used for stimuli at another rate, since the stimulus-ON events are equal across the rates. This hinges on the assumption that events with the same length stimulus-ON, but a different length stimulus-OFF, are actually equal.

This leads to the first section of this research: providing evidence that events with the same length stimulus-ON, but different length stimulus-OFF have functionally similar transient responses. After this, generalization will be attempted across different rates using lagged stimuli.

## 1.2 Generalizing over Stretched Data

For stretched stimuli however, this event-response equivalence across rates does not hold. Predicting the event responses at one presentation rate, when the event responses at another rate are known, requires a manipulation of the event responses.

The underlying theory which would allow prediction across different presentation rates for stretched stimuli comes from linear systems theory. Linear systems have a key property, called superposition [11], which is defined as

$$T[x1(t) + x2(t)] = T[x1(t)] + T[x2(t)]$$

Here, $xn(t)$ is an input at time $t$ and $T$ is the system transformation, i.e. the output. In terms of VEPs, superposition means that the response to two events is the sum of the responses to the individual events. This is illustrated in Figure 3.

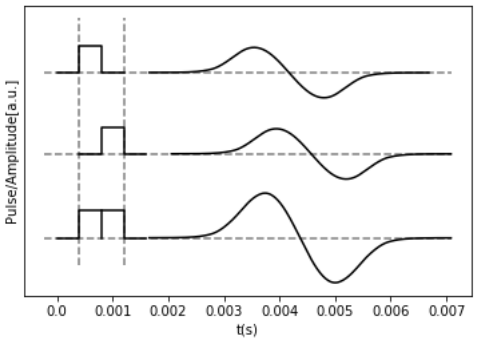
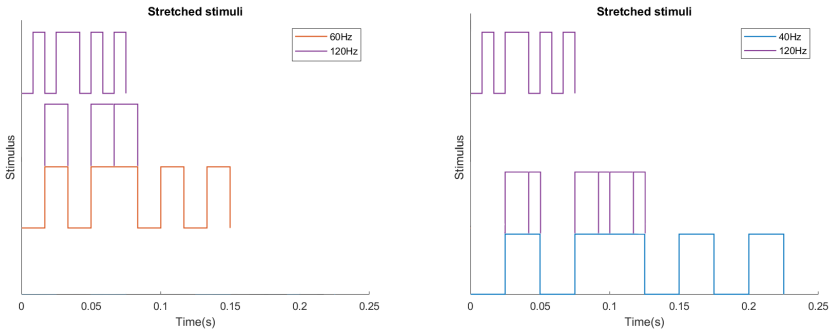



Figure 3: Two simulated events with their response, shifted to align end-to-start, and a longer pulse, exactly the sum of the two individual events, with its response, equal to the sum of the individual responses. This demonstration only holds if VEPs obey to the rules of superposition.

For example, if a classifier is trained on stretched data at 120Hz, but testing data is collected at 60Hz, in theory the transient responses could be predicted by using the long response at 120Hz as a short response at 60Hz, after which superposition of two short 60Hz responses predicts the long response at 60Hz. This is illustrated in Figure 4a. When generalizing from 120Hz to 40Hz, the short event response at 40Hz is theoretically equivalent to the superposition of the short and long event responses at 120Hz, and the long event response at 40Hz is the superposition of two short event responses at 40Hz, as illustrated in Figure 4b.




(a) Superposition to 60Hz

(b) Superposition to 40Hz

Figure 4: Representation of the superposition of stimuli for stretched data needed to generalize from 120Hz to 60Hz and 40Hz. For generalization from 120Hz to 60Hz, no superposition is needed to find the short event response at 60Hz. For the generalization from 120Hz to 40Hz however, the short event response at 40Hz is the superposition of the short- and the long event response at 120Hz.

Superposition is not a new concept in the field of BCI. The generative model assumes that superposition holds for VEPs from event responses to stimulus response, i.e. the stimulus response is calculated as the superposition of all event responses. The results from the generative model seem to be reliable enough to warrant this assumption [5, 8, 10]. There also has been evidence that the SSVEP responses can be explained as the superposition of transient responses [12]. However, no research has been done yet for superposition from event response to event response, and with that the ability to predict responses across rates for stretched data, in VEPs.

This leads to the second section of this research: providing evidence that superposition holds from event response to event response. After this, generalization will be attempted across different rates using the superposition of event responses.

# 2 Methods

## 2.1 Experimental Setup

The data used in this experiment was originally gathered by Borhanazad et al. [5]. Borhanazad et al. used the following experimental setup.

There were ten participants taking part in the experiment with stretched stimuli, and three participants taking part in the experiment with lagged stimuli. The participants were presented with an array of eight RGB LED lights, each covered with a transparent sheet showing a picture of an animal. The

participants sat 35 to 50 centimeters away from the array, chosen by the participant to give them a comfortable view of each LED. The presentation rates that were used differed for stretched and lagged stimuli. The stretched stimuli were presented at 40, 60, 90 and 120Hz, while the lagged stimuli were presented at 40, 60, 90, 120 and 180Hz. There was one run for each stimulation frequency, meaning there were four runs for the stretched stimuli, or five runs for the lagged stimuli. The different rates were presented in random order. Each run consisted of 10 trials, and each trial in turn consisted of the cue of a randomly chosen LED for one second, followed by 4.2 seconds of flashing, followed by one second of rest. Between each run, participants were allowed a short break to prevent fatigue.

The codes that were used in this experiment to determine the flickering pattern of the LEDs were Gold codes. The most frequently used pseudorandom codes used for c-VEP are maximum length sequences (m-sequences), and although m-sequences have excellent cross-correlation, their autocorrelation is often poor [8]. Gold codes are an alternative to m-sequences, and have both good cross-correlation (although worse than m-sequences) but also good autocorrelation [13]. Gold codes are generated from two m-sequences with length $2^{n-1}$, and a maximum cross-correlation of $2^{(n+2)/2}$ where n is the size of the linear feedback shift register.

Borhanazad et al. tested two different kinds of stimuli: stretched and lagged stimuli, as explained in the introduction and shown in Figure 2. Both of these datasets will be used. This means that there was a different number of participants for the two kinds of stimuli.

For the stretched codes, the experiment started with a training session, followed by calibration, followed by a testing session. However, since these testing sessions used early stopping, not every participant has the same amount of data, making for unfair comparison. So instead, 10-fold cross-validation (i.e. leave-one-out cross-validation) was used on the training data for this experiment. For lagged codes, only a training session took place. Hence, for the lagged experiment, 10-fold cross-validation was used on the training data.

Not all classifiers performed equally well. So in order to ensure that the data was both realistic, but that the conditions would still be close to optimal, a subset of the users was made for the stretched data by selecting those who had an average classification rate of at least 80%. Since there were only three participants for the lagged data, the data from all three was used.

## 2.2 General Analysis Pipeline

An analysis pipeline was set up to do the following. The training data was taken in order to calibrate the generative model. In essence, this means that the classifiers now know the generalized event responses to the events within the stimulus.

The generative model can then use the calibrated classifiers and the stimuli matrices to predict the EEG response to those stimuli. It does this by performing convolution: every event evokes an event response, and the stimulus response is

the sum of the event responses shifted over time in accordance with the event onset.

In practice the generative model does as follows: a design matrix $M$ is generated for each stimulus. These design matrices have one axis reflecting the event, one axis reflecting EEG time samples, and one axis reflecting the event response time samples. For each event, the first column then shows a 1 whenever an event of that type is presented. In the next row, the next column shows a 1, until the last column contains a 1.

Calibration has been done by presenting a certain stimulus that contains all possible events. From this, the event responses $R$ are calculated as $R = M^+X$, where $M^+$ is the pseudo-inverse of $M$. The template response to stimulus $j$, $T_j$, is then calculated as $T_j = M_jR$.

During testing, the measured EEG signal is compared to each template, and the template that has the highest correlation with the measured signal is chosen as the attended code.

It should be noted that the transient responses have an arbitrary amplitude. This also means that a transient is completely equivalent if it is flipped over the horizontal axis, and during training, a direction is chosen arbitrarily. In order to ground these transients, the spatial filter is used, and the transients are set to be positive around channel `Oz`. This means that if training results in a spatial filter that is negative at channel `Oz`, both the spatial filter and the temporal filter, i.e. the transients, are negated. This becomes relevant when comparing transients using correlation.

Another important note is regarding the event definition. In the generative model, the transients are determined per event, but the event definitions can be changed. One way of defining the events is short versus long pulses, i.e. 10 or 100 versus 110 or 1100. Here, the run-off is assumed to not matter. Another way would be to also distinguish between the different run-off lengths, i.e. 10 versus 100 versus 110 versus 1100. Here, the run-off is assumed to matter. The former definition will be referred to as *duration*, the latter will be referred to as *components*.

## 2.3 Statistical Tests

All hypotheses, barring the third hypothesis, are tested in two ways.

A) using correlations to test transient similarities.

B) using accuracies to test classifier performance.

Transient similarity is tested using Pearson's correlation. The Pearson's correlation is calculated as

$$\rho(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} \left( \frac{x_i - \bar{x}}{\sigma_x} \right) \left( \frac{y_i - \bar{y}}{\sigma_y} \right)$$

where $x$ and $y$ are two transients, each of $N$ samples, $\bar{x}$ and $\bar{y}$ the means of $x$ and $y$ respectively, and $\sigma_x$ and $\sigma_y$ are the standard deviations of $x$ and $y$ respectively.

When a t-test is performed on correlations, a Fisher transformation is needed first [14]. The Fisher transformation is defined as

$$z = \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right) = \operatorname{arctanh}(\rho)$$

where $\rho$ is the Pearson's correlation, *ln* is the natural logarithm and *arctanh* is the inverse hyperbolic tangent function.

This same principle can be used to get meaningful averages from correlations, which is not the case by default. By first performing a Fisher transformation on a set of correlations, and then averaging, the resulting values can be used meaningfully in a t-test. Also, if a correlation is desired again after averaging, the averaged z-value can be transformed back to a correlation by using the inverse of the function:

$$\rho = \frac{e^{2z} - 1}{e^{2z} + 1} = \tanh(z)$$

## 2.4 Hypothesis 1: Run-off

For the first part of the run-off hypothesis, only the lagged data is examined. First, the average correlation between all transients of the same event are taken over all rates, and compared. If indeed the run-off length does not matter, the event-responses for the same event across all rates should be roughly equivalent. The average correlation will be taken across all rates, of course excluding the correlation with the responses for the same rate. So, for each user, for each event, a correlation table is made containing the correlations between all event-responses. The diagonal, containing the correlations for each rate with itself, is removed, and the tables are averaged to get one average correlation for each user, for each event. This is done for both duration-trained classifiers, with two events, and components-trained classifiers, with four events.

Next, the components-trained classifiers are used to determine the correlations between the two short events (10 and 100) and the two long events (110 and 1100), per user. If the correlations are significantly higher than zero, that means that those two events, are relatively similar.

Cross-validation is also performed, in order to get an estimate of the difference in performance. Whether a longer run-off matters or not is determined by looking at the difference in performance between duration-trained classifiers and components-trained classifiers. If in fact 10 and 100 are perceived as different events, components should perform significantly higher than duration-trained classifiers. If not, duration-trained classifiers and components-trained classifiers should perform roughly equal.

A one-sided t-test with $\alpha = 0.05$ is performed on both the (Fisher transformed) correlations and the accuracies. A t-test on correlations can give insight

into whether the transients are significantly similar, but if the p-value is higher than the significance level, this does not give any proof for the transients being different.

The first null hypothesis (${}^{1A}H_{null}$) is that the correlations between the events with the same number of ones, but different run-off lengths are not significantly different from zero.

The first alternative hypothesis (${}^{1A}H_{a}$) is that the correlations between the events with the same number of ones, but different run-off lengths, are significantly higher than zero.

The second null hypothesis (${}^{1B}H_{null}$) is that the accuracies from components-trained classifiers do not perform significantly different from duration-trained classifiers.

The second alternative hypothesis (${}^{1B}H_{a}$) is that the accuracies from components-trained classifiers do perform significantly higher than duration-trained classifiers.

## 2.5 Hypothesis 2: Superposition

The superposition hypothesis starts with the same general analysis pipeline, run on both duration-trained classifiers and components-trained classifiers, on stretched data. The pipeline is run until the point where the event responses are predicted. Then, one of the event responses can be predicted from the other event responses using superposition, and the real and constructed transients can be compared via both correlation or classifier performance.

For duration events, the long response is predicted as the superposition of two short event responses, one of those shifted by one bit. This is realized by padding two copies of the short event response with zeros, so that one bit worth of zeros (which is the sampling frequency over the presentation rate in terms of samples) is added to the start of one copy, and the end of the other. Next, those two copies are added element-wise, and the result is cut back to the size of the original event response.

For components events, the same is done, but rather than with the response to two short events, the response to 10 and 100 are used. The response to 10 is shifted forward one bit, meaning it gets padded with zeros at the start, and the response to 100 gets padded with zeros at the end. Again, the two padded transients are added element-wise, and the result is cut back to the size of the original event response.

For duration events, the correlation between the short event response and the long event response is measured, as well as between the long event response and the constructed long event response. For components events, three correlations are calculated: between the responses to 10 and 110, the responses to 100 and 110, and finally between the constructed response to 110 and the actual response to 110.

Once these correlations are calculated, a gain in correlation can be calculated. Imagine that a classifier was only trained on short events, but is then tested on codes that contain long events as well. The question is: would it then

be beneficial to perform superposition, in order to estimate the response to the long event? In other words, does superposition have some predictive power? For the transients of duration-trained classifiers, the gain is simply calculated as

$$\text{gain}_{\text{dur}} = \text{corr}(110(0), 110(0)_{\text{con}}) - \text{corr}(110(0),\ 10(0))$$

and for components-trained classifiers, the gain is calculated as

$$\text{gain}_{\text{com}} = \min(\Delta\text{corr}_{10}, \Delta\text{corr}_{100})$$

where

$$\Delta\text{corr}_{10} = \text{corr}(110, 110_{\text{con}}) - \text{corr}(110, 10)$$

and

$$\Delta\text{corr}_{100} = \text{corr}(110, 110_{\text{con}}) - \text{corr}(110, 100)$$

i.e. the minimum gain, compared to the two constituent event responses.

Finally, the classifiers are tested using 10-fold cross-validation. The construction of the transients is inserted into the cross-validation process, meaning that first, the classifier is trained on the training set, then the transients-construction is performed, and then the classifier is applied to the validation set, using those transients containing one constructed event response.

Again, a one-sided t-test with $\alpha = 0.05$ is performed on both the (transformed) correlations and the accuracies, this time between the constructed transients and normal transients.

The first null hypothesis (${}^{2\text{A}}\text{H}_{\text{null}}$) is that both the correlations between the normal transients and constructed transients, as well as the gains, are not significantly different from zero.

The first alternative hypothesis (${}^{2\text{A}}\text{H}_{\text{a}}$) is that the correlations between the normal transients and constructed transients, as well as the gains, are significantly higher than zero.

The second null hypothesis (${}^{2\text{B}}\text{H}_{\text{null}}$) is that the accuracies from the classifiers with constructed transients are not significantly different from chance level.

The second alternative hypothesis (${}^{2\text{B}}\text{H}_{\text{a}}$) is that the accuracies from the classifiers with constructed transients are significantly higher than chance level.

## 2.6 Hypothesis 3: Generalization Using Lagged Data

For the next hypothesis, the transients trained on the lagged data are actually generalized. The classifiers are trained on data from one presentation rate, and tested on the data from another rate.

For each user and event definition (duration and components), for each rate as the training rate, all rates are used as testing rates. These accuracies are then collapsed to an average per user, per event definition. Again, a one-sided t-test with $\alpha = 0.05$ is performed on the classification accuracies.

The null hypothesis (${}^{3}\text{H}_{\text{null}}$) is that the average accuracies are not significant difference from chance level.

The alternative hypothesis (${}^{3}\text{H}_{\text{a}}$) is that the average accuracies are significant higher than chance level.

## 2.7 Hypothesis 4: Generalization Using Superposition

It is important to realize that in order to use superposition, the training rate needs to be an integer multiple of the testing rate. This means that the testing rate is at its largest half the training rate, since then the short event at the testing rate would become the long event at the training rate (see Figure 4).

So, in order to generalize across rates for the available stretched data, the classifiers at 120Hz are taken as the base classifiers. From the rates that were used, 40Hz and 60Hz can express 120Hz as an integer multiple. If the ratio between the training rate (120Hz) and testing rate is even, the short stimulus, and thus the short response according to linear systems theory, can be expressed as the superposition of solely the long stimuli or responses at the training rate (see Figure 4a). If the ratio between the training and testing rate is uneven, the short event response, is set to be the superposition of a number of long event responses plus the superposition of one short event response (see Figure 4b).

First, a t-test with $\alpha = 0.05$ is performed on the correlations, in order to see if the constructed transients are significantly different from the real transients. Second, another t-test is performed on the accuracy, to see whether they perform significantly better than chance level.

The first null hypothesis (${}^{4A}H_{null}$) is that the average correlation between the constructed transients at either 40Hz or 60Hz and the real transients at either 40Hz or 60Hz is not significantly different from zero.

The first alternative hypothesis (${}^{4A}H_{a}$) is that the average correlation between the constructed transients at either 40Hz or 60Hz and the real transients at either 40Hz or 60Hz is significantly higher than zero.

The second null hypothesis (${}^{4B}H_{null}$) is that the average accuracy using the constructed transients at either 40Hz or 60Hz is not significantly different from chance level.

The second alternative hypothesis (${}^{4B}H_{a}$) is that the average accuracy using the constructed transients at either 40Hz or 60Hz is significantly higher from chance level.

## 2.8 Summary of the Hypotheses

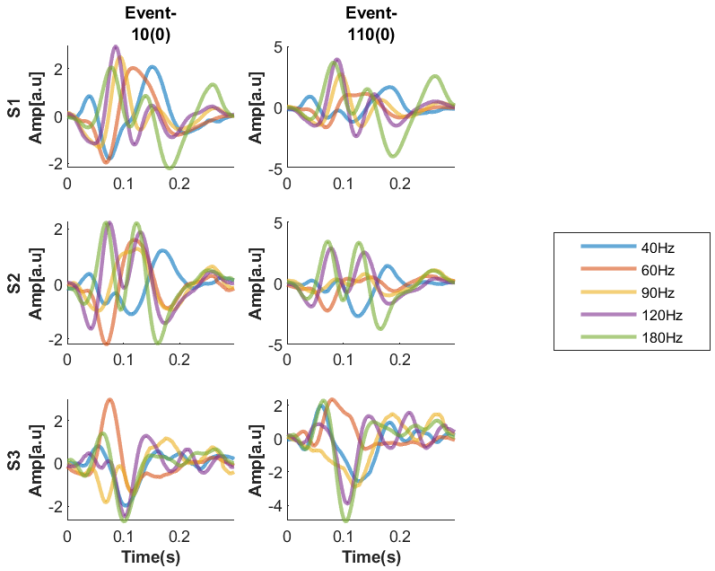
To clarify and summarize the four hypotheses and their interaction, see the table below.

| H | Name | Stimulus | Research Question | Example |
|---|---|---|---|---|
| $^{1}$H | Run-off test | Lagged | Does the run-off length matter for the transient response in lagged data? | Is the response to 10 the same as to 100, the same as to any event of the form $10^{+}$? |
| $^{2}$H | Superposition test | Stretched | Can event responses be predicted accurately using superposition in stretched data? | Can for duration-trained classifiers, the response to 110(0) be predicted as the superposition of two responses to 10(0)? |
| $^{3}$H | Generalization on lagged data | Lagged | Can lagged data be used to generalize transients over presentation rates? | Can transients trained on lagged data at 90Hz be used to classify data at 60Hz? |
| $^{4}$H | Generalization using superposition | Stretched | Can superposition be used to generalize transients over presentation rates for stretched data? | Can the transient responses at 120Hz be manipulated, using superposition, in order to classify data at 60Hz? |

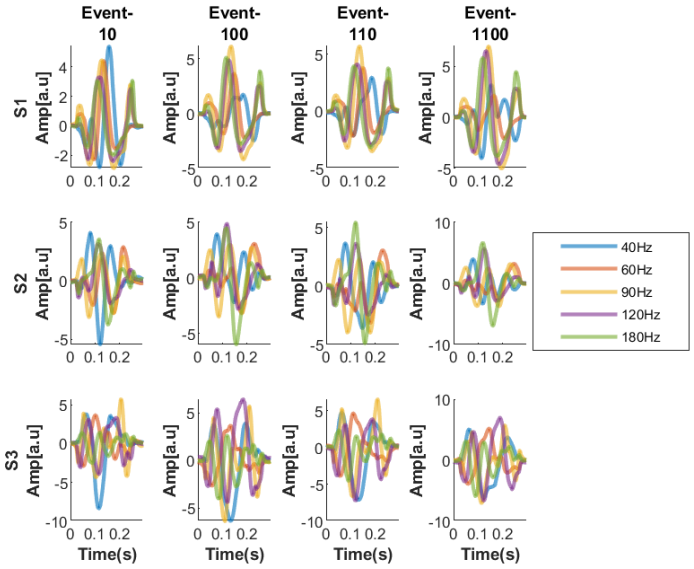
# 3 Results

## 3.1 Hypothesis 1: Run-off

The event responses were plotted per user, per event, over all rates. The responses for event type duration are plotted in Figure 5a, and the responses for event type components are plotted in Figure 5b.




(a) Event responses for event type duration




(b) Event responses for event type components

Figure 5: Event responses per user, per event for all rates, including the average response for lagged data.

For the first part of the run-off hypothesis, the transients for the lagged data were examined. The correlations for each user, per event, averaged over all rates

were:

| User | e10(0) | e110(0) |
|---|---|---|
| S1 | 0.1854 | 0.1313 |
| S2 | 0.0327 | 0.0197 |
| S3 | 0.3838 | 0.2399 |

for duration, and

| User | e10 | e100 | e110 | e1100 |
|---|---|---|---|---|
| S1 | 0.4586 | 0.4530 | 0.2791 | 0.2390 |
| S2 | 0.0644 | 0.2325 | 0.1511 | 0.1698 |
| S3 | -0.0934 | -0.1575 | -0.0874 | -0.1464 |

for components.

A t-test was performed to see if those correlations between the responses to events with the same number of ones were significantly different from zero, i.e. if the transients were significantly similar.

| Event | p-value | reject ${}^{1A}H_{null}$ |
|---|---|---|
| Duration | 0.1873 | no |
| Components | 0.4421 | no |

The correlations between the responses to the short and long events of the components-trained classifiers are averaged per user, and shown per rate in Figure 6.

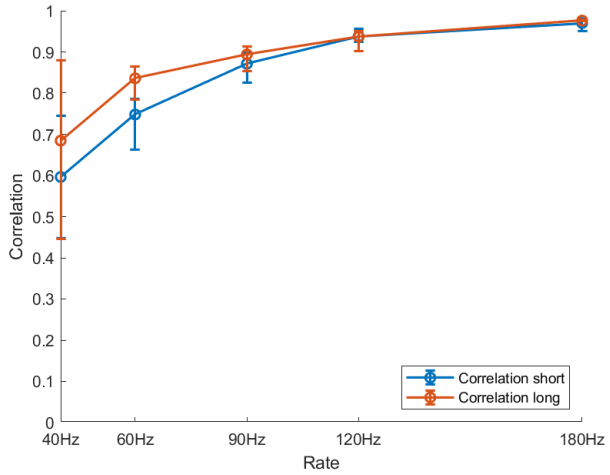



Figure 6: The correlations between either the short events (10 or 100) or the long events (110 or 1100) for components per rate, for lagged data. The error bars indicate the highest and lowest correlation.

Both the correlations between the long and short events for components, as shown in Figure 6, were significantly similar.

| Event | p-value | reject ${}^{1A}H_{null}$ |
|---|---|---|
| Short | 9.437e-04 | yes |
| Long | 3.620e-04 | yes |

The accuracies achieved on lagged stimuli, using either duration-trained classifiers or components-trained classifiers, were averaged over users, and are shown per event type in Figure 6 and Figure 7.

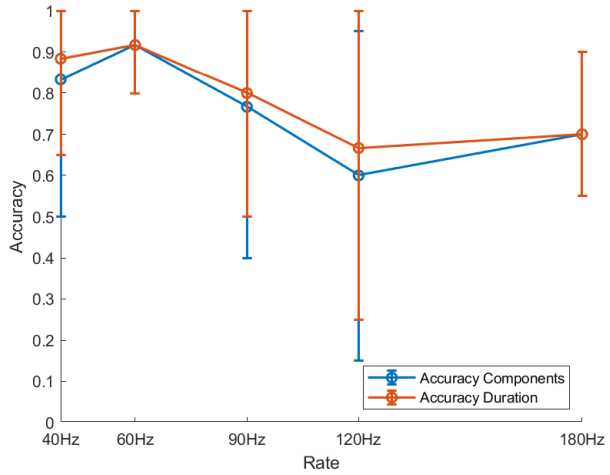



Figure 7: The classification accuracy per rate, for duration- and components-trained classifiers, for lagged data. The error bars indicate the highest and lowest performance.

Another t-test was performed to see if the accuracies between the duration-trained classifiers and components-trained classifiers were significantly different.

| Experiment | p-value | reject ${}^{1B}H_{null}$ |
|---|---|---|
| Accuracy | 0.4830 | no |

## 3.2 Hypothesis 2: Superposition

To get an understanding of what the constructed transients look like, compared to both the real transient and its constituent parts, Figures 8, 9, 10 and 11 are shown below. Figure 8 (duration) and Figure 10 (components) show instances where there is a gain in correlation, i.e. the superposition is beneficial, while Figure 9 (duration) and Figure 11 (components) show instances where there is a loss in correlation, i.e. the superposition is not beneficial.

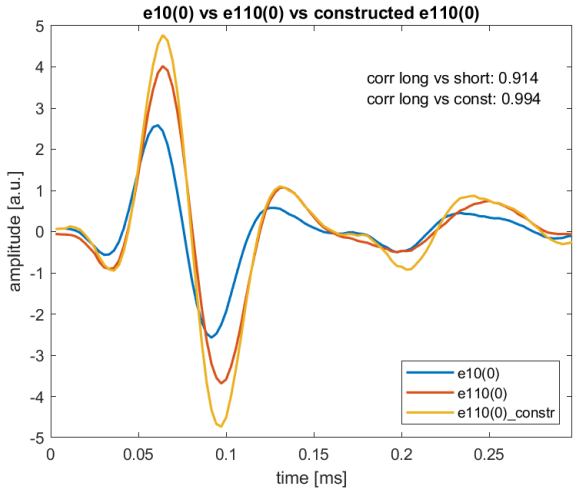



Figure 8: The event responses for type duration, including the constructed long response, for subject 2, at 120Hz, for stretched data. Superposition results in a gain in correlation in this case.

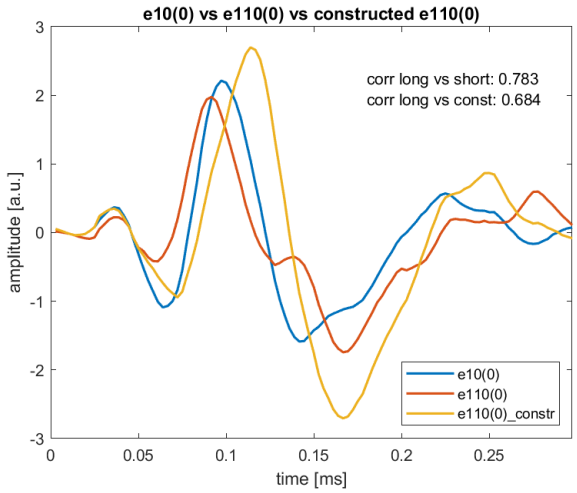



Figure 9: The event responses for type duration, including the constructed long response, for subject 7, at 40Hz, for stretched data. Superposition results in a loss in correlation in this case.

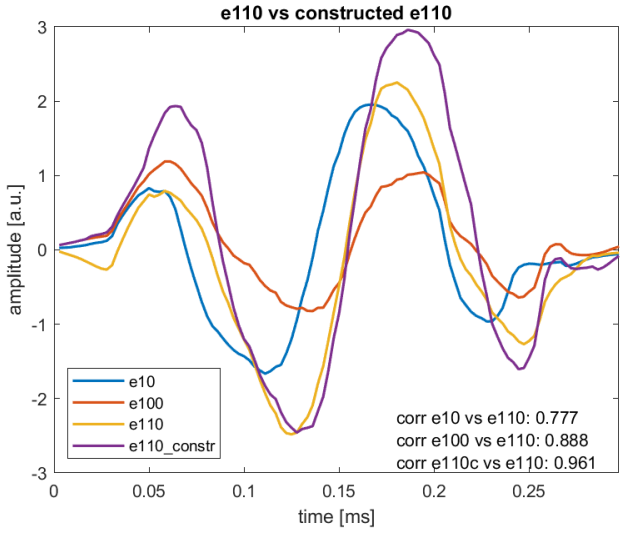



Figure 10: The event responses for type components, including constructed the 110 response, for subject 2, at 60Hz, for stretched data. Superposition results in a gain in correlation in this case.

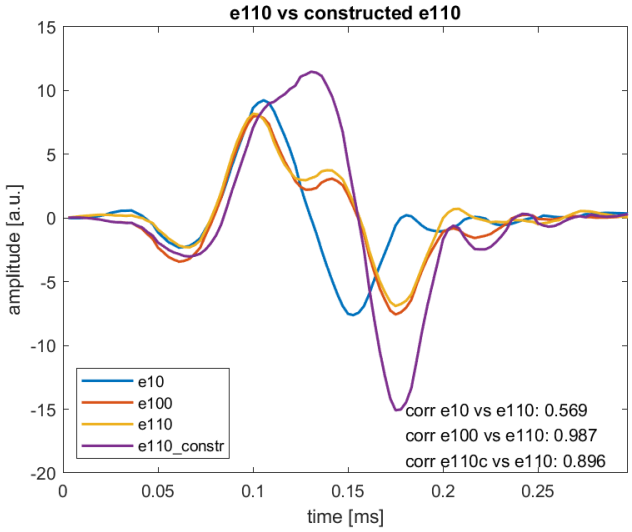



Figure 11: The event responses for type components, including the constructed 110 response, for subject 5, at 40Hz, for stretched data. Superposition results in a loss in correlation in this case.

First, the correlations between the true long response and the constructed

long response were calculated for duration, and the correlations between the true response to event 110 and the constructed response to event 110 were calculated for components. These are plotted in Figure 12.

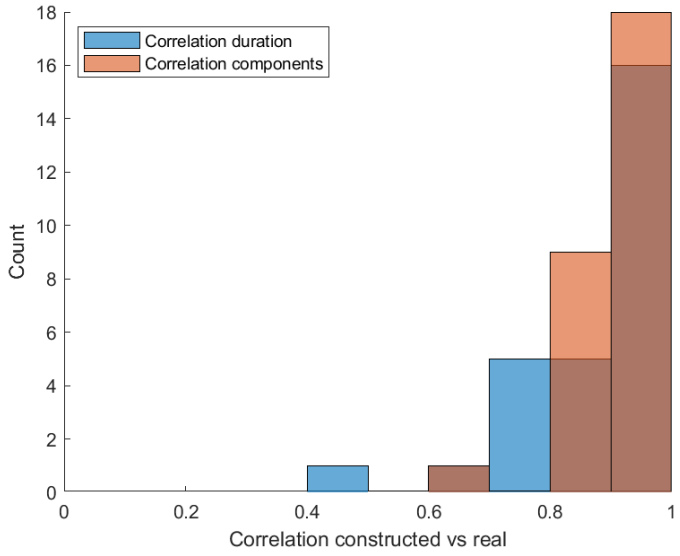



Figure 12: The average correlations between the real long or 110 response and the constructed long or 110 response

Next, a t-test with was performed on those correlations.

| Event | p-value | reject ${}^{2A}H_{null}$ |
|---|---|---|
| Duration | 2.344e-3 | yes |
| Components | 1.607e-3 | yes |

After that, the correlation gains were calculated with regards to the constituent parts, which are shown in Figure 13.

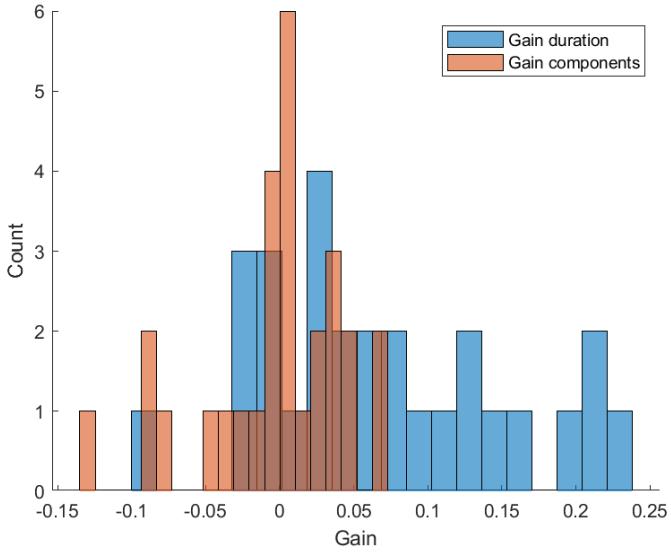



Figure 13: The average correlations gains when comparing the constructed long or 110 response to their constituent parts.

Another t-test was performed on the gains, so see if the construction is beneficial when compared to the constituent parts.

| Event | p-value | reject ${}^{2A}H_{null}$ |
|---|---|---|
| Duration | 3.520e-3 | yes |
| Components | 0.4159 | no |

For the second part of the superposition hypothesis, the accuracies were measured. Either the real transients were used, or the transients where one of the event responses was constructed, were used. To illustrate, three users were selected such that with the non-constructed transients, one user had high accuracy, one user had mediocre accuracy and one user had low accuracy, both for duration and components. The accuracies for these three users are shown in Figure 14.

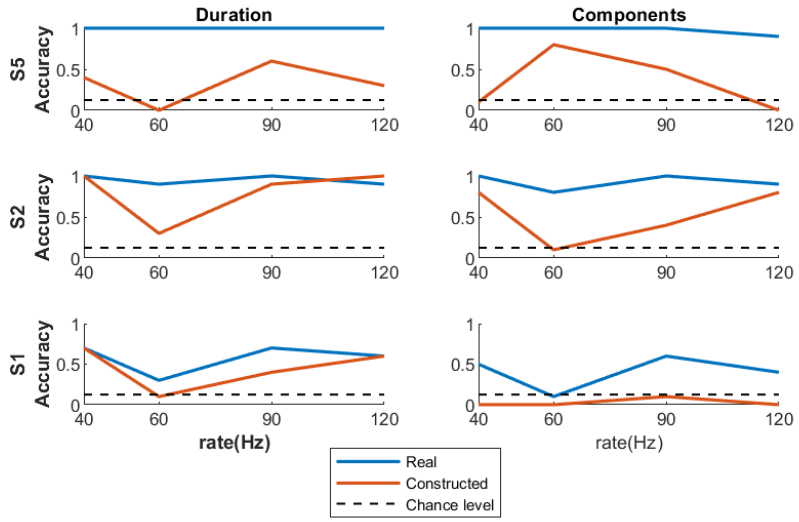



Figure 14: The accuracies for three users on stretched data, for all rates, for both duration and components, for either the real transients, and the transients where one of the event responses was replaced by a version constructed using superposition.

A third t-test was performed on the accuracies, using the constructed long response or 110 response, with chance level accounted for.

| Event | p-value | reject ${}^{2B}H_{null}$ |
|---|---|---|
| Duration | 0.053567 | no |
| Components | 0.030639 | yes |

## 3.3 Hypothesis 3: Generalization Using Lagged Data

The lagged generalization experiment was the only experiment where just the accuracies were tested, and not the correlations, since those were already tested in the run-off experiment (Section 3.1).

The accuracies per user, per event type, for all training rates and all testing rates are shown in Figure 15.

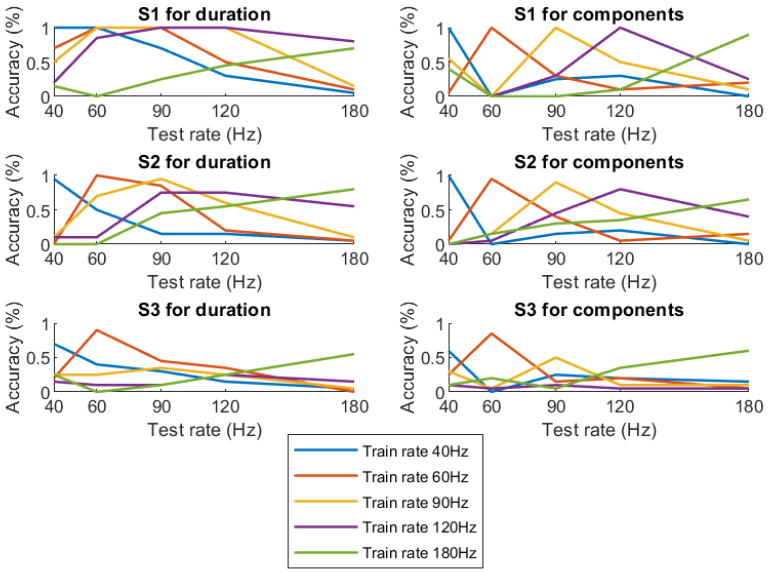



Figure 15: The accuracies per user, per event type, for all training rates and all testing rates on lagged data.

A t-test was performed on the average accuracies of the classifiers per user for lagged data, training on one rate, and testing on all rates. Chance level was accounted for, and the instances where the train- and test rate were equal were removed, in order to test only the generalized rates.

| Event | p-value | reject ${}^{3}\mathrm{H}_{\mathrm{null}}$ |
|---|---|---|
| Duration | 2.652e-04 | yes |
| Components | 7.886e-2 | no |

## 3.4 Hypothesis 4: Generalization Using Superposition

To get an understanding of what the predicted transients look like, Figure 16 and 17 are two examples that show the real transients at 40 and 60Hz against the transients predicted from 120Hz using superposition, as well as the responses at 120Hz. Figure 16 is from a subject that had relatively high correlation, while Figure 17 is from a subject with very low correlation.

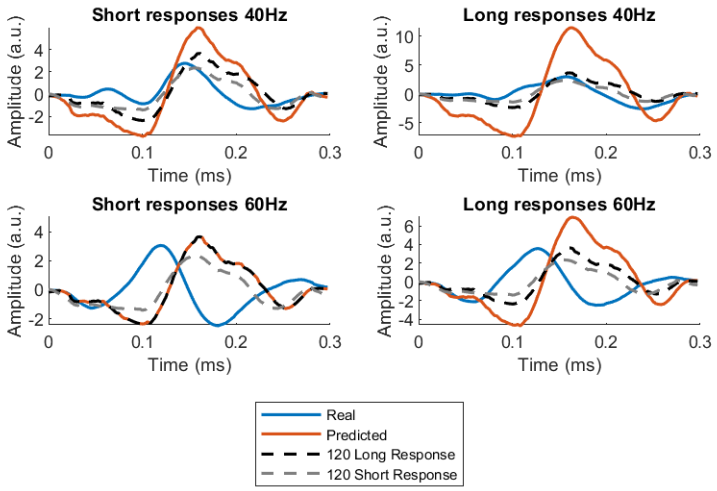



Figure 16: The transients at 40Hz and 60Hz, both real and predicted using superposition from 120Hz for stretched data, for subject 9. The predicted transients look somewhat like the real transients.

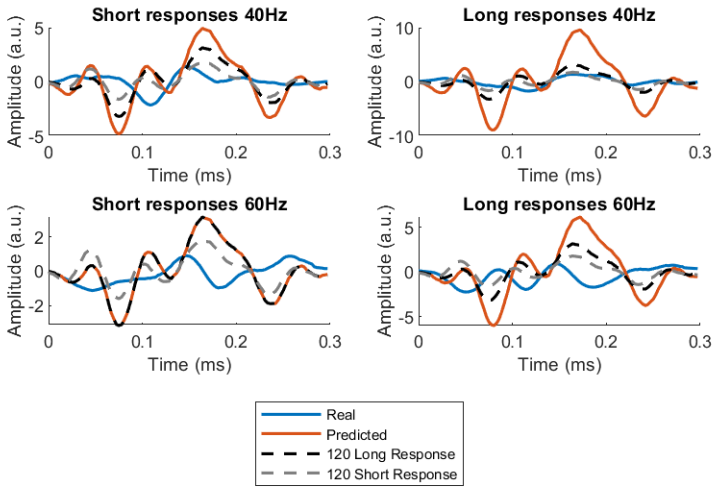



Figure 17: The transients at 40Hz and 60Hz, both real and predicted using superposition from 120Hz for stretched data, for subject 1. The predicted transients look very little like the real transients.

The correlations between the real event responses at 40 and 60Hz and the event responses constructed from 120Hz transients, per user, are shown below in Figure 18.

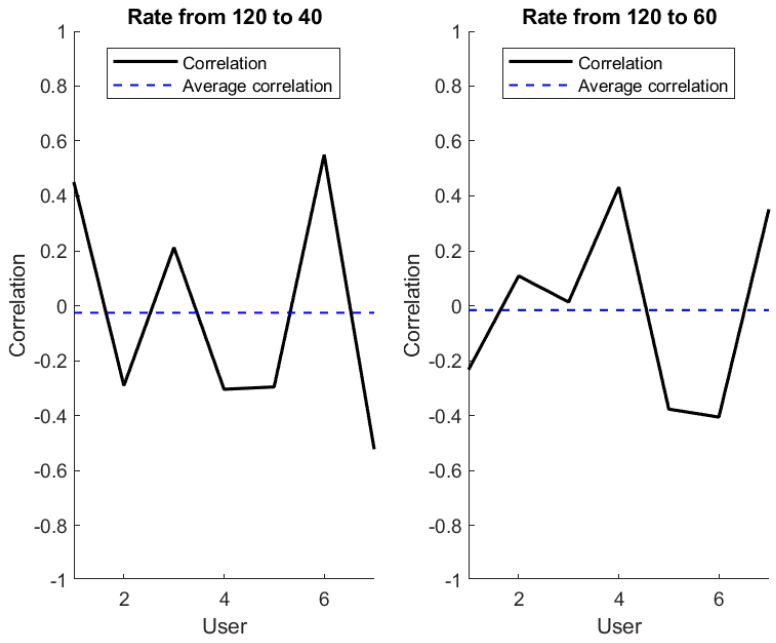



Figure 18: The correlations for all users between the real transients at 40Hz and 60Hz and the transients constructed from 120Hz for stretched data.

A t-test was performed on those correlations, to see if the two transients were significantly similar.

| Testing rate | p-value | reject ${}^4\mathrm{H}_{\mathrm{null}}$ |
|---|---|---|
| 40Hz | 0.9476 | no |
| 60Hz | 0.4727 | no |

The accuracies that were achieved using the constructed transients are shown in Figure 19 for a few subjects. These subjects were chosen so that one had relatively high accuracies, one had relatively mediocre accuracies, and one had poor accuracies. Note that there is no data point at 90Hz, since it is impossible to use superposition generalize to 90Hz from 120Hz. This is because $\frac{120}{90}$ is not a whole number, as explained in Section 2.7.

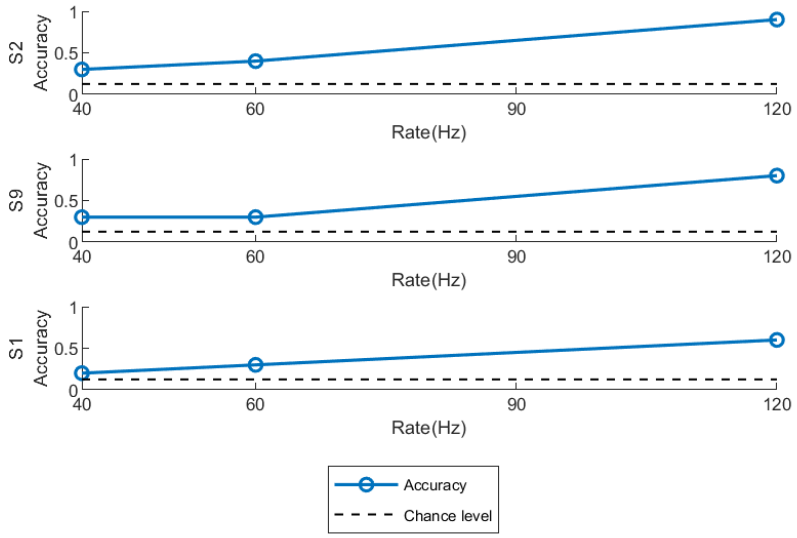



Figure 19: The classification accuracies for a subset of users, at 40Hz and 60Hz, using the transients constructed from 120Hz.

The second t-test was performed on the accuracies, corrected for chance.

| Testing rate | p-value | reject $^4H_{null}$ |
|---|---|---|
| 40Hz | 0.9476 | no |
| 60Hz | 0.4727 | no |

# 4 Discussion

For the run-off hypothesis, the similarity between the transient responses to events with the same number of ones, but a different number of zeros was investigated. The correlations between the responses to events with the same number of ones were not significantly different from zero, meaning that there was no statistical similarity between the transients of the same event at different rates for lagged data. However, at the same time, classifiers using the event definition duration performed roughly equal to classifiers using the event definition components, since there was no significant difference between those two sets of classification rates. This makes it ambiguous whether the run-off length is important to the classifier or not.

What is odd, is that it seems as if the trend lines in Figure 6 and Figure 7 are opposing: the higher the rate, the higher the correlation seems to be, but the higher the rate, the lower the accuracy becomes. This holds for both components and duration, but stronger for components. This could indicate that when a classifier trains on components, it relies on those four components

being different, and so the more similar those event responses are, the worse the classifier performs.

Looking at the responses in Figure 5 per event, per user over all rates, it seems like specifically subject 1 for components has quite similar responses, with the notable exception at 40Hz. This is also confirmed by the correlation table. As discussed in Section 2.2, the amplitudes of the transients can flip over the horizontal and still be equivalent to the classifier. This is visible for certain events for certain users, e.g. in the plot of short event response of duration event types for subject 3 (Figure 5a, left bottom plot), just before 0.1s, the yellow transient goes against the majority of the other transients.

In general, it seems like the correlations between the different events are higher for components than duration. This gives the expectation that in the generalization using lagged data, components will generalize better, i.e. the accuracies will be higher for components-trained classifiers when the training rate and testing rate are not equal.

In the superposition hypothesis, the principle of superposition was investigated, by predicting one of the event responses within the same presentation rate. For duration, the long event response was predicted as the superposition of two short event responses, and for components the response to 110 was predicted as the superposition of the responses to 100 and 10.

As could be seen in the results of the superposition hypothesis (Section 3.2), both the constructed long response for duration and the constructed response to 110 for components had significant similarities to their real counterpart, i.e. the correlations were strongly skewed towards 1. However, the correlations were also relatively high between the constituent transients and the to-be-constructed transient, meaning for duration the correlation between the short event response and the long event response was high, and for components the correlations between the responses to 10 or 100 and 110 were high. That means that the gains show a more fair comparison of how useful the constructions are. And, as shown with both the t-test and in Figure 13, the correlation gain is not statistically significant for components, but was significant for duration. In fact, for components, quite a few constituent transients resemble the target transient more closely than the constructed transient (see Figures 9 and 11). However, it seems like this is a viable method to use for duration-trained classifiers.

Just these results regarding the correlation indicate that in the generalization over rates using superposition, the constructed transients could give high accuracies when the duration event type is used. However, the accuracies were not significantly higher than chance level for duration-trained classifiers, while they were for components-trained classifiers. One explanation could be that the components-trained classifiers still contained three out of four original transients, which leads to most of the stimulus response still staying intact. At the same time, even if the correlations for the transients of duration-trained classifiers are significantly different from zero, that does not mean they are necessarily high. This could lead to the effect seen with the accuracies in Section 3.2.

In the experiment generalizing over rates using lagged stimuli, the transients were used for rates other than the training rate, in order to see if classification accuracy would be significantly higher than chance level. If the instances where the training- and testing rate are the equal are excluded, the components do not show any significant performance. For event type duration however, there is significantly better performance.

This can also be seen from Figure 15: for at least some subjects, the lagged duration transients seem to generalize relatively well to rates that are not too different. For example, for subject one, when training at 120Hz, testing at 90Hz, 120Hz and 180Hz all gave decent accuracies, and arguably at 60Hz as well. It also seems that the more extreme the difference is between the training rate and testing rate, the less well the classifier performs.

For components, there was much worse generalization over different rates. This goes against the prediction made from the results in the run-off hypothesis, where the correlations were higher between the event responses for components. One possible reason could be that it is inconsistent to use component event types over different rates: by using component events, the assumption is made that the run-off length matters. But at the same time, it is also assumed that the run-off length does not matter, since a generalization is made over different lagged rates, i.e. by increasing or decreasing the run-off length. This inconsistency could make it difficult for the classifier to decide whether a signal should be seen as the response to the lagged equivalent of 10 or 100.

Another problem could be that duration-trained classifiers have double the data to learn the event responses compared to components-trained classifiers, since two events for components,e.g. 10 and 100, are counted as the same event for duration, e.g. 10(0).

Subject 3 performed quite poorly on lagged data, as can be seen in the bottom row of Figure 15, even when the training and testing rate was equal. However, removing subject three from the analysis did not make a difference in the result.

In the last generalization for stretched data, superposition was used to generalize from one rate to another, as explained in Section 2.7, and illustrated in Figure 4. The first t-test showed that the correlations were not significantly different from zero, as is also shown in Figure 18. The second t-test showed that the accuracies were not significantly different from chance level, as illustrated in Figure 19. So it seems that when all the transients are constructed, unlike in the superposition hypothesis, where only one of the two or four transients was constructed, both the correlations and the accuracies go down dramatically.

## 4.1 Limitations

The statistical tool used to show whether two transients were different was a simple t-test performed on either the correlation or the accuracy. However, t-tests are not very appropriate for BCI data. Specifically, using t-tests on correlations is a relatively indirect measure of comparison. It would be best to

directly compare two transient responses for equality, but due to the high auto-correlation inherent to time-series data, this is not possible with t-tests.

Another issue this research has, is the indirect measurement that is classifier performance. Similarly, it is important to keep in mind that the transient responses that the classifier has learned are not the real brain responses, but rather a temporal filter. Therefore, these results should be reconsidered when taken outside of the field of BCI.

As can be seen in Figures 8, 9, 10, and 11 in Section 3.2, the event-to-event superposition is not quite a simple linear addition. One of the most notable differences is the larger amplitude in many of the predicted responses. This can indicate that some normalization factor might be necessary. And, as can be seen in Figure 9 and Figure 11, many times the general shape of the transients are also wrong, leading to a loss of correlation. The better performing constructed transients from the generalization using superposition do look promising, e.g. in Figure 16 (remember that the transients can be flipped horizontally, as explained in Section 2.2). However, specifically for 40Hz, but to a lesser extent also for 60Hz, the amplitudes of the constructed transients are much higher than the real transients.

# 5 Conclusion

In this thesis, two different methods of generalizing transients across presentation rates were explored. The first method used lagged stimuli, where all events of the same type have the same number of ones, regardless of the presentation rate. The second method used superposition in the more commonly used stretched stimuli.

First, some evidence was gathered to find out whether the run-off lengths mattered to the classifiers. The results showed that events with the same number of ones, but a different number of zeros, were not similar in terms of correlation. At the same time, the accuracies did not change significantly, whether events with the same number of ones, but a different number of zeros were seen as the same or as different events.

Next, the concept of superposition was explored, still within one presentation rate. The results showed that the constructed long event response was relatively similar to the real long event response. When the similarity of the constituent transients was taken into account, the similarity was still significant for duration events, but not for components events. However, the accuracies showed results significantly above chance level for components events, but not for duration events.

After that, the first generalization across rates was done on lagged data. The results showed that classifiers with event type duration had significant results on testing data that had a different presentation rate from the training data. Classifiers trained on components did not, however.

Finally, stretched data was generalized across rates, using superposition. The results showed both insignificant correlation between the predicted transients and the real transients, as well as classification accuracies around chance level.

So in conclusion, lagged data shows promise in being generalizable across different presentation rates, specifically when the training- and testing rate are not too far apart. It also provides evidence that stretched data cannot be generalized over different presentation rates via a simple linear superposition.

## 5.1 Future work

As discussed in Section 4.1, the statistics used in this experiment were not very suitable for EEG and BCI data. One large problem is inherent to the field of frequentist statistics, which is the inability to prove functional equivalence. The field of Bayesian statistics however, has developed some promising techniques, specifically a technique called *Bayesian estimation supersedes the t-test* (BEST) [15]. One potential pitfall could be that the BEST algorithm requires the definition of a *region of practical equivalence*. This might be difficult to define for BCI data, since it still not fully know which specific features of transients are important for functional similarity (see the contradictory results of the superposition hypothesis in Section 3.2). So as of now, it might be difficult to define precisely what transients should, or should not be seen as practically equivalent.

One more event type that could be explored would be the rising and/or falling edge, where the events 01 and 10 could either be defined as different events, or as the same single event. This event definition assumes that the important feature is a change in the stimulus, rather than a stimulus being on. Similar to lagged data, if the events are defined to be rising/falling edges, the transients from one rate could theoretically be used for another rate.

As discussed in Section 1.2, superposition is an idea from linear systems theory. If c-VEPs are in fact true linear systems, other features from linear systems theory could be exploited. This entails approximating a Dirac delta function, which would be a true impulse, defined as $\delta(x) = 0$ for $x \neq 0$ and $\delta(x) \to +\infty$ for $x = 0$. This can be approximated in VEPs by learning the response to e.g. the short event at 360Hz. Theoretically, one could estimate the response to complex stimuli while only having measured the response to this one, very short impulse, by exploiting the property of superposition. However, before any further steps towards further generalization are taken, the generalization of stretched data needs to be vastly improved.

One last, but major drawback of this experiment was the lack of data, specifically for the lagged experiments, where there were only three subjects, of which one performed consistently very poorly. Therefore, it is recommended to re-do the experiment on lagged stimuli, with more subjects.

# References


[1] G. Bin, X. Gao, Y. Wang, B. Hong, and S. Gao, "Vep-based brain-computer interfaces: time, frequency, and code modulations [research frontier]," *IEEE Computational Intelligence Magazine*, vol. 4, no. 4, pp. 22–26, 2009.

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

[3] N. Birbaumer, "slow cortical potentials: Plasticity, operant control, and behavioral effects," *The Neuroscientist*, vol. 5, no. 2, pp. 74–78, 1999.

[4] B. Blankertz, C. Sannelli, S. Halder, E. M. Hammer, A. Kübler, K.-R. Müller, G. Curio, and T. Dickhaus, "Neurophysiological predictor of smr-based bci performance," *NeuroImage*, vol. 51, no. 4, pp. 1303 – 1309, 2010.

[5] M. Borhanazad, J. Thielen, J. Farquhar, and P. Desain, "The effect of high and low frequencies in c-vep bci," in *Proceedings of the 8th Graz Brain-Computer Interface Conference 2019*, 2019.

[6] M. Spüler, W. Rosenstiel, and M. Bogdan, "Online adaptation of a c-vep brain-computer interface(bci) based on error-related potentials and unsupervised learning," *PLOS ONE*, vol. 7, pp. 1–11, 12 2012.

[7] M. van Gerven, J. Farquhar, R. Schaefer, R. Vlek, J. Geuze, A. Nijholt, N. Ramsey, P. Haselager, L. Vuurpijl, S. Gielen, and P. Desain, "The brain–computer interface cycle," *Journal of Neural Engineering*, vol. 6, p. 041001, 07 2009.

[8] J. Thielen, P. van den Broek, J. Farquhar, and P. Desain, "Broad-band visually evoked potentials: Re(con)volution in brain-computer interfacing," *PLOS ONE*, vol. 10, pp. 1–22, 07 2015.

[9] S. Ahmadi, M. Borhanazad, D. Tump, J. Farquhar, and P. Desain, "Low channel count montages using sensor tying for VEP-based BCI," *Journal of Neural Engineering*, vol. 16, p. 066038, 11 2019.

[10] J. Thielen, P. Marsman, J. Farquhar, and P. Desain, *Re(con)volution: Accurate Response Prediction for Broad-Band Evoked Potentials-Based Brain Computer Interfaces*, pp. 35–42. 08 2017.

[11] A. V. Oppenheim, "Generalized superposition," *Information and Control*, vol. 11, no. 5, pp. 528 – 536, 1967.

[12] A. Capilla, P. Pazo-Alvarez, A. Darriba, P. Campo, and J. Gross, "Steady-state visual evoked potentials can be explained by temporal superposition of transient event-related responses," *PLOS ONE*, vol. 6, pp. 1–15, 01 2011.

[13] R. Gold, "Optimal binary sequences for spread spectrum multiplexing (corresp.)," *IEEE Transactions on Information Theory*, vol. 13, no. 4, pp. 619–621, 1967.

[14] D. M. Corey, W. P.Dunlap, and M. J. Burke, "Averaging correlations: Expected values and bias in combined pearson rs and fisher's z transformations," *The Journal of General Psychology*, vol. 125, no. 3, pp. 245–261, 1998.

[15] J. K. Kruschke, "Bayesian estimation supersedes the t test," *Journal of Experimental Psychology: General*, vol. 142, no. 2, pp. 573–603, 2013.