university of groningen / faculty of science and engineering

Dynamic coding in a large-scale, functional, spiking-neuron model

Bachelor's Project Thesis

Loran Knol
Supervisor: Dr J.P. Borst

Abstract: A functional, large-scale, spiking-neuron model of working memory (WM) was adapted to display the patterns often cited in EEG studies as evidence of dynamic coding. The model had a mechanism for temporarily adjusting its own inter-neuron connection strengths following network activation, which served as the main memory mechanism. As for dynamic coding: It is a phenomenon observed in the human brain, in which information is represented in a particular way at time step t, but is represented differently at time step $t+1$, while, crucially, the information itself does not change. In a previous, human EEG study, data were obtained that showed such dynamic coding patterns. Two experiments from that study were conducted with the model, and the model results were compared to the corresponding human data. The comparisons showed that the model and human coding patterns display many similarities, but also some differences. Moreover, the model performed not as well as humans did. Eventually, however, it was concluded that the model did in fact display dynamic coding, which would mean that dynamic coding might simply be a property of any self-modifying network. This calls for a perspective on dynamic coding that is slightly more modest than what is suggested in the existing literature.

1 Introduction

Working memory (WM) is an important part of human cognition (Daneman and Carpenter, 1980). It is involved in maintaining and/or modifying information from the senses or memory in order to complete a specific task (Baddeley, 1992). Although most researchers previously agreed that WM information would be maintained with the help of sustained neuronal firing patterns in the lateral prefrontal cortex (lPFC) (e.g. Curtis and D'Esposito, 2003), more recent studies have noted a decrease of such WM-specific patterns while information maintenance was still required (Sreenivasan, Curtis, and D'Esposito, 2014). Curiously, this decrease would appear even when tasks were executed correctly (Watanabe and Funahashi, 2014), leading to the idea that WM might sometimes function in an effectively *activity-silent* way.

Previous studies (e.g. Wolff, Jochim, Akyürek, and Stokes, 2017) have demonstrated that when WM-specific activity *is* present (prior to decreasing to noise levels), it has the interesting property of not statically representing the information it encodes. Instead, EEG measurements show that information is represented *dynamically*, meaning that while information is encoded in a certain way at time point t, it may be encoded completely differently at time point $t+1$ - but, crucially, the information itself, that what is encoded, does not change.

Stokes (2015) (preceding the study from Wolff et al. (2017), which is just one example of a study incorporating dynamic representations) has attributed these dynamic properties to complex interactions between a network's activity state and its underlying 'hidden' state, where the latter refers to the collection of neurophysiological parameters that determine the network's behaviour (e.g. the amount of calcium at a neuron's presynaptic terminal at a given time). It is called 'hidden' because those parameters, as opposed to network activity, are typically not measured. According to Stokes, a certain stimulus would invoke some activity state

under influence of an already present hidden state. That invoked activity state then modifies the initial hidden state, which causes (through recurrent connections) a new activity state, which again modifies the hidden state, and so on and so forth. This interaction would happen on "the very shortest timescales", according to Stokes.

The present study, however, proposes a model that displays the patterns found in EEG studies cited as proof for dynamic coding, but through a simpler process. More specifically, an existing neural spiking model designed by Pals, Stewart, Akyürek, and Borst (2020) was adapted. It uses the Nengo framework as designed by Eliasmith (2013). While the model does feature a form of hidden states as mentioned above, it does not incorporate the extensive reciprocal interaction between hidden and active states Stokes envisioned. In the following sections, the original Pals et al. model will be discussed in more detail along with the Wolff et al. (2017) experiments with which it was tested. After that, the model adaptations done by this study will be considered, followed by an explanation of certain analyses (employed by Wolff et al.) that test for dynamic coding. Finally, the results of the analyses of the adapted model are compared to the results from Wolff et al., and the capability of the adapted model to display dynamic coding, as well as its implications, are discussed.

1.1 The Pals et al. model

This study takes a model developed by Pals et al. (2020) as its starting point. The model is a large-scale, spiking-neuron model trying to explain activity-silent WM in the context of functional behaviour.

1.1.1 Short-term synaptic plasticity

To achieve a functional model of activity-silent WM, Pals et al. implemented a mechanism known as short-term synaptic plasticity (STSP) (Zucker and Regehr, 2002). STSP refers to a phenomenon, shown by many neurons, in which their firing facilitates (enhances) the connection between pre- and postsynaptic neurons for no more than a few minutes (Zucker and Regehr, 2002). This essentially allows neuronal network activity to alter the network structure and have these alterations persist even after the activity has disappeared. These structural alterations can be considered as a method for holding information, which is an important function for WM. Taking that into account, it is clear how STSP is one of the candidate mechanisms for activity-silent WM.

However, changes in the neuronal network structure are often not measured in neuroimaging studies; it is far more common to measure network activity. For this reason, the variables that determine the network structure (and the exact workings of the STSP-mechanism) are often called *hidden variables*. Two hidden variables in particular are hypothesised to determine the presynaptic neuron's predisposition to fire on a short timescale (Zucker and Regehr, 2002; Mongillo, Barak, and Tsodyks, 2008). One is the build-up of calcium ions at the presynaptic terminal: this build-up happens due to presynaptic firing and facilitates subsequent release of neurotransmitter. The second variable is the amount of neurotransmitter available for release at the presynaptic terminal (resource variable), which depletes due to presynaptic firing. While the calcium build-up declines over time (typically a second), the amount of available neurotransmitter increases. Effectively, the calcium is what enables the facilitation, and the resources variable, in turn, limits the facilitation (i.e. prevents unlimited calcium build-up). Together, they form a calcium-mediated version of the STSP mechanism described above. It is this calcium-mediated STSP mechanism that was implemented in the Pals et al. model.

1.1.2 The Wolff et al. study

This model was designed to use STSP to account for results from an EEG experiment conducted by Wolff et al. (2017). Wolff and colleagues developed a perturbation approach to measure the hidden states of activity-silent WM, i.e. they 'pinged' the brain with a non-specific stimulus so the neural WM networks could 'echo' the information their hidden states contained (not unlike sonar). To demonstrate this approach, they conducted three delayed response experiments, the first two of which are relevant for this study. In both of these two experiments, participants had to maintain randomly oriented gratings in memory (memory items) to be able to compare them to another grating presented at the end of the trial (test item), after some delay.

During this delay, the participants were 'pinged' with a vivid image called an impulse, which would presumably reactivate their activity-silent WM representations. EEG measurements were conducted on the participants during both experiments.

Using information decoding analyses on those EEG data, Wolff and colleagues found that the orientation of the memory item gratings could not only be decoded immediately after initial stimulus presentation (when the items were stored in WM), but also after impulse presentation, consistent with the idea that a 'ping' to WM would elicit an 'echo'. All of this was taken as evidence that WM could save information in an activity-silent way.

In addition, Wolff et al. conducted cross-temporal decoding analyses (CTDAs) on their data recorded just after memory item presentation, one of which is depicted in Figure 1.1. The figure shows, for every time point t_i, how well information from all other time points $t_0, ..., t_n$ could be decoded by a decoder trained on t_i. The plots are symmetrical, so this interpretation is valid both in a row-wise and in a column-wise fashion. High values that lie off-diagonal indicate that decoders can generalise cross-temporally to other time points, i.e. they can decode time points other than those they were trained on. This cross-temporal generalisation is an indication that information is represented statically. In contrast, a CTDA with high decoding values that lie just along its diagonal indicate that a decoder trained on time step t_i could not decode very well at time step t_{i+1}. This absence of cross-temporal decoder generalisation suggests a dynamic way in which information is represented. The CTDAs created by Wolff and colleagues showed such diagonals, and the aim of the present study is to present a large-scale, spiking-neuron model that can show the same CTDA patterns.

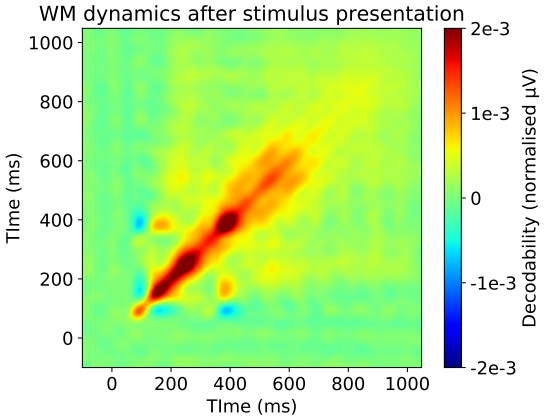

Figure 1.1: A cross-temporal decoding analysis, constructed from EEG data measured when a stimulus was presented and held in WM. The warmer (i.e. more red) the colours, the better the decoding. The clustering of high decoding values on the figure's diagonal and nowhere else indicates little cross-temporal generalisation and thus dynamic coding. Data from Wolff et al. (2017) (Experiment 1).

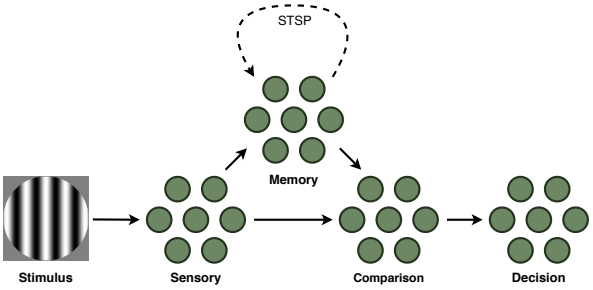

Figure 1.2: The model from Pals et al. (2020). Stimuli are transformed into a vector by the sensory population, which then sends it to the memory and comparison populations. For the first stimulus in a trial, the memory will hold on to its representations via recurrent STSP connections. During second stimulus presentation, memory and sensory will respectively project the first and second stimulus to the comparison population, after which the decision population uses the information from the comparison population to determine how to act on the perceived stimuli difference.

1.1.3 Structure and augmentations

Let us return to the Pals et al. model to consider its structure. The model consisted of two modules - one for each hemisphere - and every module contained a sensory, a memory, a comparison, and a decision population (see Figure 1.2). The grating images linked to the sensory population, which connected to the comparison population in two ways: One connection went directly from sen-

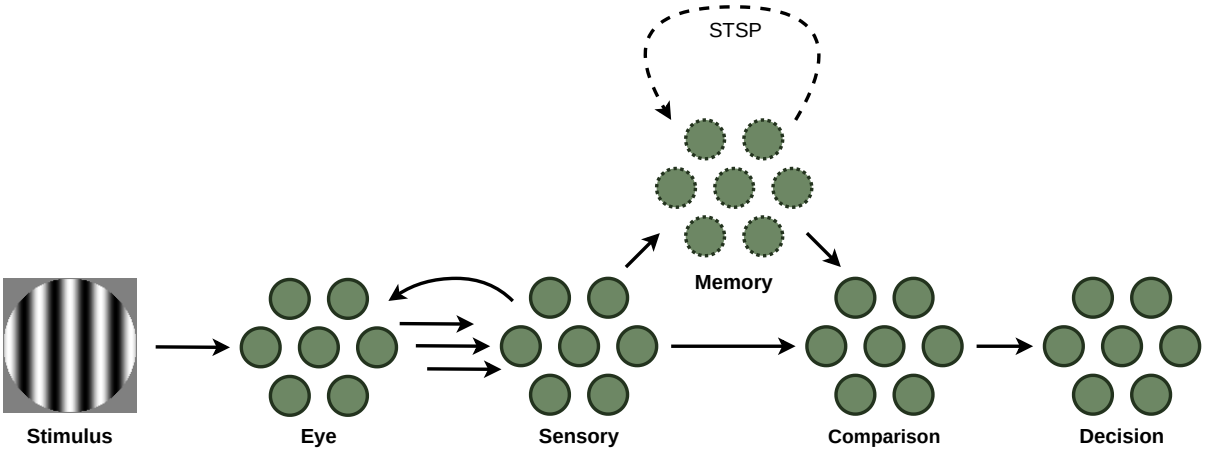

Figure 1.3: The augmented model. An 'Eye' population has been added in between the stimulus and the sensory population. The eye population and the sensory population are connected by both a feed-forward connection with different synaptic delays (the stacked arrows pointing right) and a recurrent connection (the curved arrow pointing left). The neurons of the memory population are subject to noise through intermittent firing (which is indicated by dotted borders).

sory to comparison, while the other one went via the memory population first, and only then to the comparison population. Finally, the comparison population connected to the decision population. All of the aforementioned connections are feed-forward. In addition, a recurrent STSP connection was added to the memory population. This architecture worked well to replicate the behavioural results from Wolff et al. (2017), but it did not suffice to fully replicate the dynamic coding properties found in human EEG data: The CTDA of the original Pals model showed a pattern that lasted just tens of milliseconds, while human data shows patterns that last hundreds of milliseconds (Figure 1.1). In addition, the pattern was that short that it was hard to determine whether it was an example of dynamic coding or simply static coding.

Therefore, in this study, the model has received three augmentations (see Figure 1.3). The expectation was that the length of the original model's CTDA pattern would correlate with the amount of time the stimulus was represented in the model, so the augmentations mainly aimed to increase stimulus representation time. The augmentations are as follows:

- Distributed synaptic delays on a feed-forward connection from an added eye module. This causes the representation of the stimulus to not reach the rest of the model in one piece, but distributed over a prolonged period of time.

- Recurrent connections back from the sensory population towards the eye population, in line with research indicating that lower-level visual areas in the human brain also receive input from higher-level areas (Lamme and Roelfsema, 2000). This was assumed to prolong stimulus representation even further.

- Background noise through intermittent firing in the memory population. Random spikes were expected to reactivate (parts of) the recurrent STSP connections, thereby allowing the saved stimulus representation to be maintained a little longer (albeit in an imperfect way).

For a more in-depth discussion of these augmentations, see Wijs (2020). An additional parameter was the strength of the synaptic delay filter applied to the spiking data of the memory population. This synaptic filter delayed all individual spikes and effectively smeared them out over time, making the data resemble those as measured by EEG devices. The effects of the augmentations and the synaptic filter have been explored, and the resulting model has been fitted to match the Wolff et al. data.

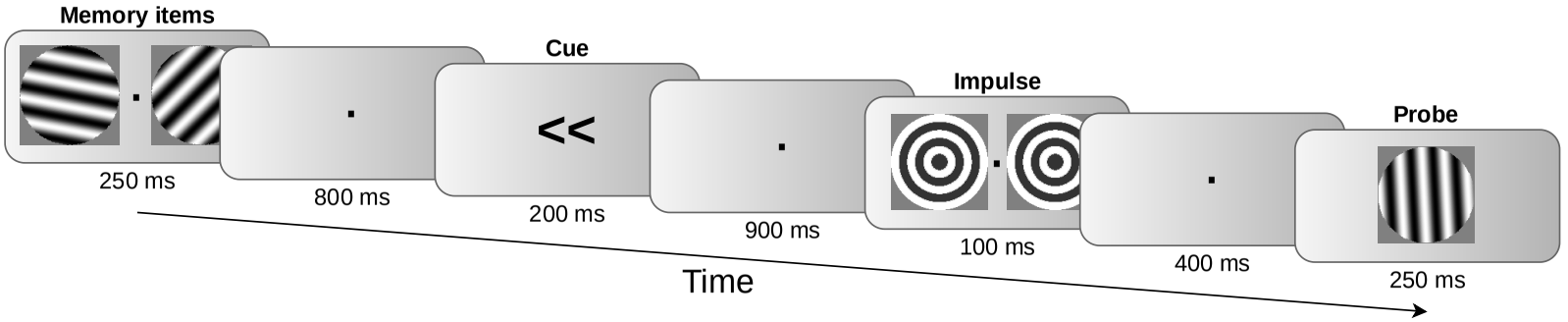

Figure 2.1: Trial schematic for Experiment 1. Two gratings, presented at the start, would have to be kept in memory. One of them would need to be compared to a grating at the end (the probe). Only after memory item presentation would participants receive a cue which item ought to be remembered and which one could be forgotten. Between the cue and the probe, an impulse, a vivid stimulus supposed to function as a 'ping' to WM, appeared. These events were all padded with delays. The duration of every stage is given below the depiction of the stage. The dots in the middle of the stage depictions are fixation dots.

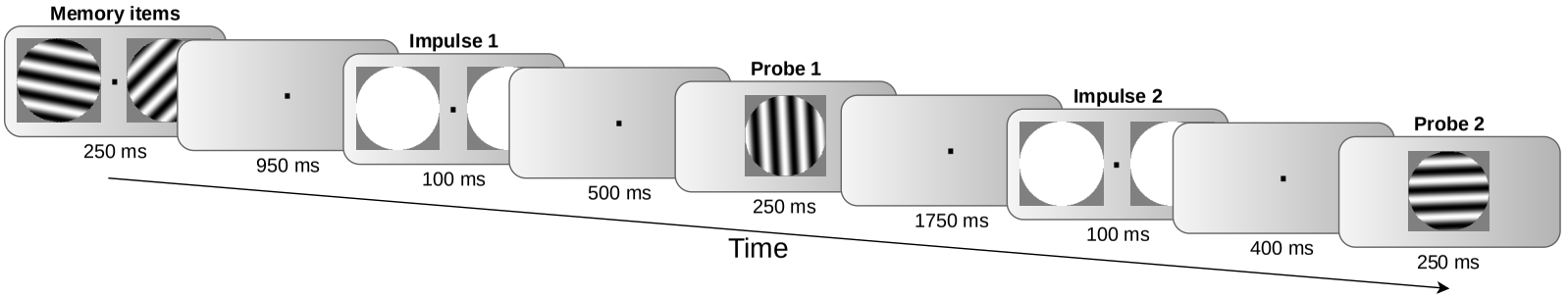

Figure 2.2: Trial schematic for Experiment 2. In this experiment, participants had to maintain two gratings in memory, which would both be tested. Participants knew the order in which the memory items would be tested. After memory item presentation, a first impulse followed, succeeded by the probe for the memory item that would be tested early. Following that, a second impulse would appear with the probe for the late-tested memory item in its wake. All these stages were padded with delays.

2 Methods

To test the augmented model for dynamic coding, it had to perform two tasks which it also performed in Pals et al. (2020). The tasks were adapted experiments from Wolff et al. (2017) and will be explained below. With the data gathered from the task execution by the model, analyses were conducted to be able to construct CTDAs. These analyses will be explained after the experiment descriptions.

Interesting to note is that, before the full experiments were run, several parameter sweeps had been conducted to find the best set of the model parameters mentioned in Section 1.1.3. Discussing those results falls outside the scope of this article, but those who are interested can find some of the sweep results in Appendix A.

2.1 Experiments

Two experiments from Wolff et al. (2017) were conducted with the augmented model, namely Experiments 1 and 2. In Experiment 1, for every trial, subjects had to watch two simultaneously presented, randomly oriented gratings and keep them both in memory. At the end of the trial, a different grating called 'probe', with a different orientation, would appear. One of the two memory gratings would have to be compared to the probe, but participants did not know which one of the two. After the two initial gratings had disappeared and a successive delay, a cue appeared that indicated which of the two items would actually need to be compared to the probe. After a second blank-screen delay period, a short impulse - a vivid image which ought to function as the WM ping - appeared, and after a third delay, the probe (the final test grating) appeared. Subjects had to indicate whether the probe was rotated clockwise or anti-clockwise with respect to the cued memory item. See Figure 2.1.

For Experiment 1, the model was run with 30 different randomisation seeds to simulate 30 different participants. Then, with each 'participant', 1344 trials were conducted, just like Wolff and colleagues did with their human participants.

Experiment 2 was similar, but now both memory items would be tested; participants were told beforehand which item would be tested first (the 'early' item) and which one would be tested second (the 'late' item). After memory item presentation and a delay, an impulse and consequent delay followed. Then probe 1 appeared, which the early item had to be compared to. Another delay followed, with a second impulse. Finally, succeeding a last delay, the final probe appeared which the late item had to be compared to. See Figure 2.2.

For this experiment, 19 'participants' were created which each performed 1728 trials. These numbers are also the same as mentioned by Wolff and colleagues.

2.2 Decodability analysis

The orientation decoding method as outlined in Wolff et al. (2017) was adapted and applied to the output of the memory-module neurons of the model.* Wolff et al. ran separate analyses for the cued and uncued stimuli, so the same was done for the current study. For every experiment trial i, they first calculated all other trials' initial stimulus angle *relative to* trial i's initial stimulus angle. For example, when the angle of trial i is 40°, and the angle of one of the other trials is 39°, the relative angle becomes -1°. They then binned all trials but trial i according to those relative angles. These bins' centres ranged from $-\frac{\pi}{2}$ up to but not including $\frac{\pi}{2}$, where every bin centre was $\frac{\pi}{6}$ apart and the bin width was $\frac{\pi}{6}$, meaning that the bins overlap (see Figure 2.3). Turning a stimulus π rad yields the exact same stimulus, meaning that an angle of $\frac{3\pi}{4}$ rad is treated the same as an angle of $-\frac{\pi}{4}$ rad.

Then, for every time step of the EEG data, the Mahalanobis distances (MDs) between the EEG data of trial i and every bin's EEG data were calculated. The MD is a distance measure for multivariate data that takes into account correlation in those data, and can therefore deal with non-spherical clusters (unlike the Euclidian distance) (De Maesschalk, Jouan-Rimbaud, and Massart, 2000). This is done by incorporating the inverse variance-covariance matrix into the calculation. The more similar the EEG data sets, the smaller the MD. Ideally, the EEG data recorded during the presentation of stimulus angles similar to trial i's stimulus angle should be similar to trial i's EEG data, meaning that the MD should be

*For the code used in these analyses, visit `https://github.com/Valkje/dynamic-coding`.

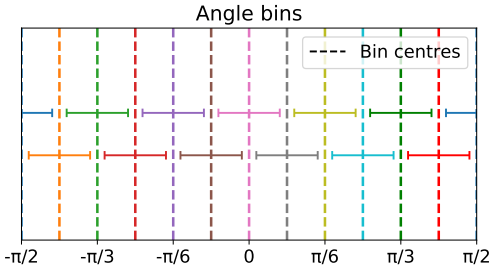

Figure 2.3: An illustration of the stimulus angle bins. The x-axis denotes the angle of the stimuli in radians and the vertical, dashed lines indicate the bin centres. The width of every bin is visualised by a horizontal line on the bin centre that has the same colour as the bin centre line.

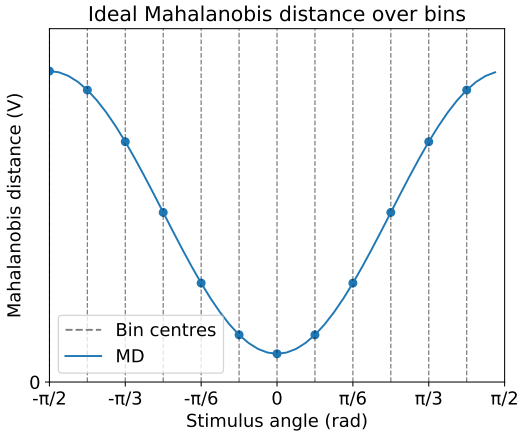

Figure 2.4: Ideal form of Mahalanobis distance across stimulus angle bins from the stimulus angle of trial i. The dashed lines denote the bin centres, while the dots indicate the MD one would expect from the angle bin centres they are placed upon. An infinitesimal number of bins would result in the sinusoidal line drawn through the dots.

smallest for the 0 rad bin and largest for the $\pm\frac{\pi}{2}$ bins, resulting in an MD curve as shown in Figure 2.4.

As can be seen in the figure, the MDs expected in ideal decoding conditions form some sort of sinusoidal - more specifically, a stretched (along the x-axis) and inverted cosine, translated above the line $x = 0$. In other words, re-stretching, mean-centring and inverting the MD curve should yield a cosine when the stimulus is presented and represented in the participant's brain. The stretching is done by multiplying all stimulus angles by 2; the same stretch is applied to the bin width (which changes it from $\frac{\pi}{6}$ to $\frac{\pi}{3}$). The transformations are illustrated in Figure 2.5. When transformed, the MD curve can be convolved with an actual cosine to get a measure of angle decodability: The better the stimulus angle is represented in the brain, the more it will be pronounced in the EEG data, the more the MD curve will look like a transformed cosine and the more the convolution of the transformed MD curve with an actual cosine will result in a high value. This value is called the cosine similarity.

As this cosine similarity has been calculated for every time step for trial i, a cosine similarity curve (also decodability curve) will be created for trial i which shows at which point in the trial the angle could be properly decoded from the EEG data. Such decodability curves are calculated for every trial and used to calculate a mean decodability curve that shows how well an angle could be decoded at which time point for all trials.

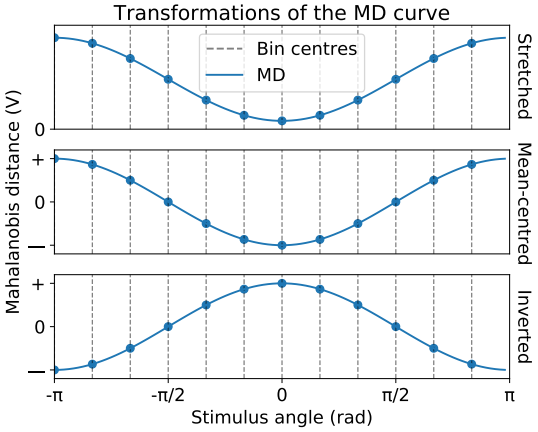

Figure 2.5: The ideal MD curve across bins first stretched along the x-axis, then mean-centred and finally inverted. Note that the domain is now $[-\pi, \pi)$ instead of $[-\frac{\pi}{2}, \frac{\pi}{2})$ due to the stretching. Dashed lines still represent bin centres.

The procedure described above was applied to data from the model. The data consist of several trials of filtered spiking activity of individual neurons from the memory population, for every time step. There were 1500 neurons in the memory population. Experiment 1 lasted 3 seconds and Experiment 2 lasted 4.6 seconds, so with a resolution of 2 milliseconds, that comes down to 1500 time steps for Experiment 1 and 2300 time steps for Experiment 2. To mimic EEG measurements, where the electrical behaviour of groups of cortical neurons is recorded, but also to decrease computational complexity, the K-means algorithm was used to group neurons together into 17 groups (convergence declared when inertia was below 10^{-20}, best of 20 runs). The number 17 was chosen because Wolff et al. used 17 electrodes in their EEG measurements. After clustering, means were calculated for all neurons in a group, resulting in a data structure of [trials] by 17 by [time steps]. Then, noise from a normal distribution ($\mu = 0$, $\sigma = 0.5$) was added to all of the data, as the model neurons are completely silent when no input is presented (i.e. their output is 0). Completely silent neurons would result in both computational errors (division by zero) and a higher than intended decodability (because all neurons look alike when they are silent). Along with the modified neuron data, the angles of the initially presented stimulus were used in the decodability analysis. The bins and their width were kept the same as in Wolff et al..

2.3 Cross-temporal decodability analysis

A cross-temporal analysis (which is an example of multivariate pattern analysis; King and Dehaene, 2014) was conducted to determine the amount of dynamic coding present in the model data. To understand how the analysis works, consider Figure 2.6, which illustrates the process for a single experiment trial i. The main idea is that, for a stimulus angle bin b, and then for every time step t_x, the bin data at t_x are compared to the data of trial i at all other n time steps $t_0, ..., t_n$. 'Compared' in this sense means calculating the MD between the data of b and the data of trial i. This procedure yields n^2 MD values for one bin.

Once the procedure has completed for all bins, there basically are n^2 MD curves, which can undergo the same set of transformations and convolution as explained in Section 2.2. This in turn results in n^2 decodability scores, which can be arranged in

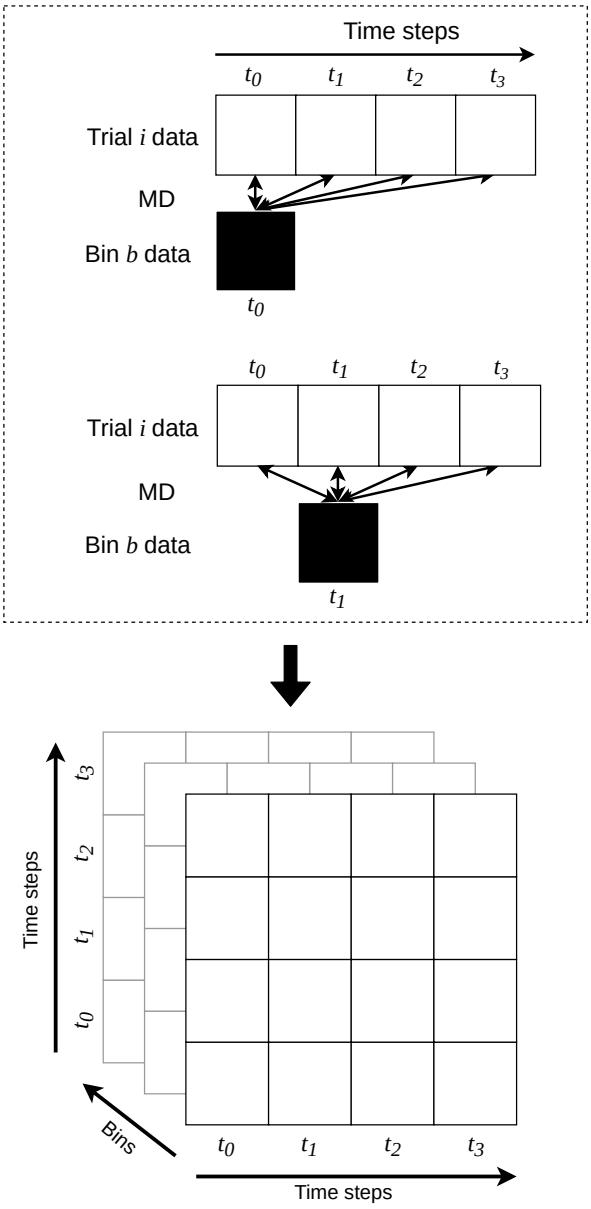

Figure 2.6: Illustration of the cross-temporal decoding process. For every stimulus angle bin b, and then for every time step t_x, the MD is calculated between the EEG state of trial i (17 channel values) at all time steps $t_0, ..., t_n$ on the one hand and the bin mean at t_x on the other hand. This results in an n-by-n grid of MD values for b. Repeating for multiple bins gives multiple grids.

an n-by-n grid. These grids are then averaged together for all trials to get a cross-temporal decoding matrix. A diagonal with high decoding values and low values everywhere else in such matrices is a hallmark of dynamic coding: It indicates that at every time step t_x, there can only be proper decoding in the context of that same time step t_x (i.e. with respect to the bins constructed from the data at that moment t_x), but not in the context of any other time step (e.g. $t_x + 1$).

The preparations of the model data were nearly the same as for the regular decoding analysis: The model neurons were combined into 17 clusters through the K-means clustering algorithm and averaged together for every cluster, after which noise was added. The only adaptation was the splitting of the set of trials of each model participant in two equally sized halves in an effort to further decrease computational complexity; CTDAs were conducted on both halves separately, and the results were averaged together.

3 Results

After conducting both experiments and collecting the corresponding model data, the analyses as outlined in Sections 2.2 and 2.3 were applied. This section will present the results of those analyses, while also considering the results from the model and the human data, gathered by Wolff and colleagues, together, as to ascertain how well the model approaches the human results. In all of the figures that follow, the term 'Data' is used to refer to the human data and results, while the term 'Model' naturally refers to the results of the augmented model.

In order to be able to compare the Data and Model results, however, some conversion has to be done first. The Model decodability scores are namely a factor of about 10 higher than the Data decodability scores. To bring all scores to the same scale, z-scores have been calculated for every individual analysis. (Means and standard deviations are recalculated for every analysis instance, unless stated otherwise.)

3.1 Experiment 1

As for Experiment 1, Figure 3.1 shows the Data and Model z-scores for both the regular (A and C, respectively) and cross-temporal (B and D, respectively) decodability of both memory items combined, during and after their presentation. The curves in A and C are relatively similar, with a steep ascent and a (relatively) smooth decay. However, it has to be mentioned that while the Model curve remains high for roughly the stimulus duration (250 ms) before decaying, the Data curve starts decaying rather quickly after reaching its peak. Moreover, the Data curve shows a second peak, which the Model does not.

Considering the CTDAs in B and D, some more differences can be seen. The cross-temporal Data results show a very strong diagonal that almost continues up to the 750^{th} millisecond. The Model CTDA, however, shows a diagonal that seems slightly thicker, shorter and less strong compared to its own surrounding decodabilities. More notably, the Model CTDA shows vertical and horizontal strokes of decodability (referred to as 'arms'), which start around 200 ms and continue for as long as the diagonal continues. The Data CTDA, however, shows no arms. Another difference is the appearance of periodic spots of slightly higher decodability in the Data CTDA, which are absent in the Model CTDA. Having said that, the Model CTDA definitely seems to show a diagonal that has at least roughly the same length as the one in the Data CTDA.

To show that the augmented model can still maintain its original function (being a model of activity-silent WM), Figure 3.2 considers the decodability of both the cued and uncued memory items after impulse presentation. Recall that the impulse is a vivid stimulus which was supposed to work as a 'ping' to WM and re-elicit silent WM item representations. As can be seen in the Model part of the figure (right pane), after impulse presentation (grey bar), both cued and uncued memory items could be decoded from the model data, although the cued item could be decoded slightly better than the uncued one. In contrast, the left pane of the figure shows that only the cued item could be decoded from the Data; according to calculations by Wolff et al. (2017), the uncued decodability was not significantly different from 0. This reveals an error in the augmented model: The uncued item can be decoded from its data, while that should not be the case.

Since the (dynamic) coding patterns have been

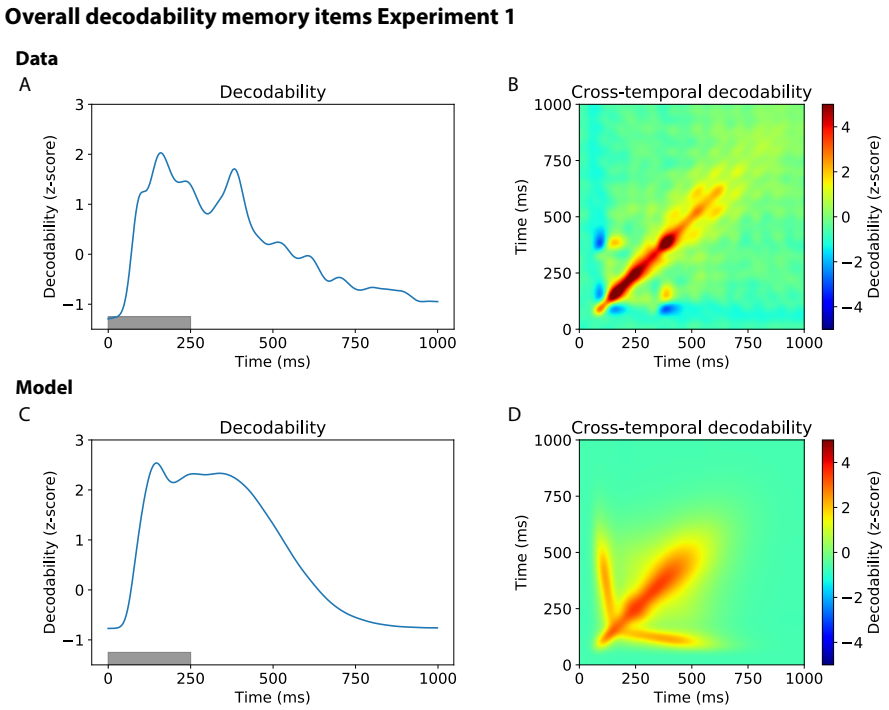

Figure 3.1: The overall decodability of the combined (cued and uncued) memory items, just after their presentation. The grey bars indicate memory item presentation. Data: The results of human data from Wolff et al. (2017). Model: The results of the data from the augmented model. All decodability scores are z-scores. ***A*** Decodability curve of the memory items from the Data. ***B*** The CTDA of the memory items from the Data. ***C*** Decodability curve of the memory items from the Model. ***D*** The CTDA of the memory items from the Model.

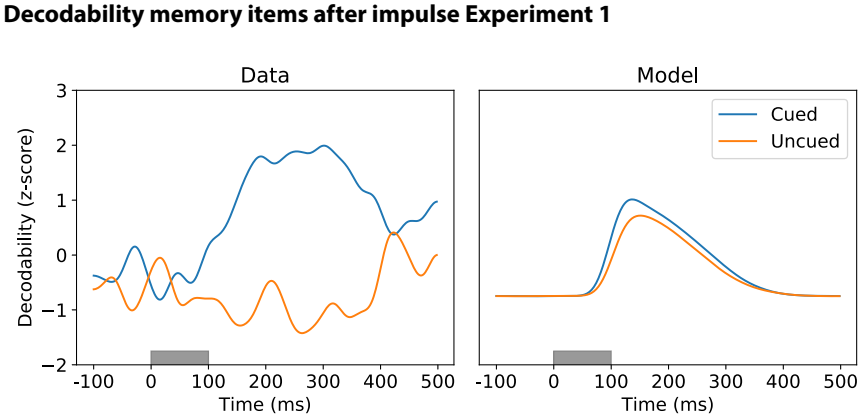

Figure 3.2: The separate decodability curves of the cued and uncued memory items after the impulse in Experiment 1. Time is shown relative to the impulse onset, and grey bars indicate impulse presentation. The left pane shows the curves for the Data (whose z-scores make use of the same mean and standard deviation), while the right pane shows the curves for the Model (which also share their mean and standard deviation).

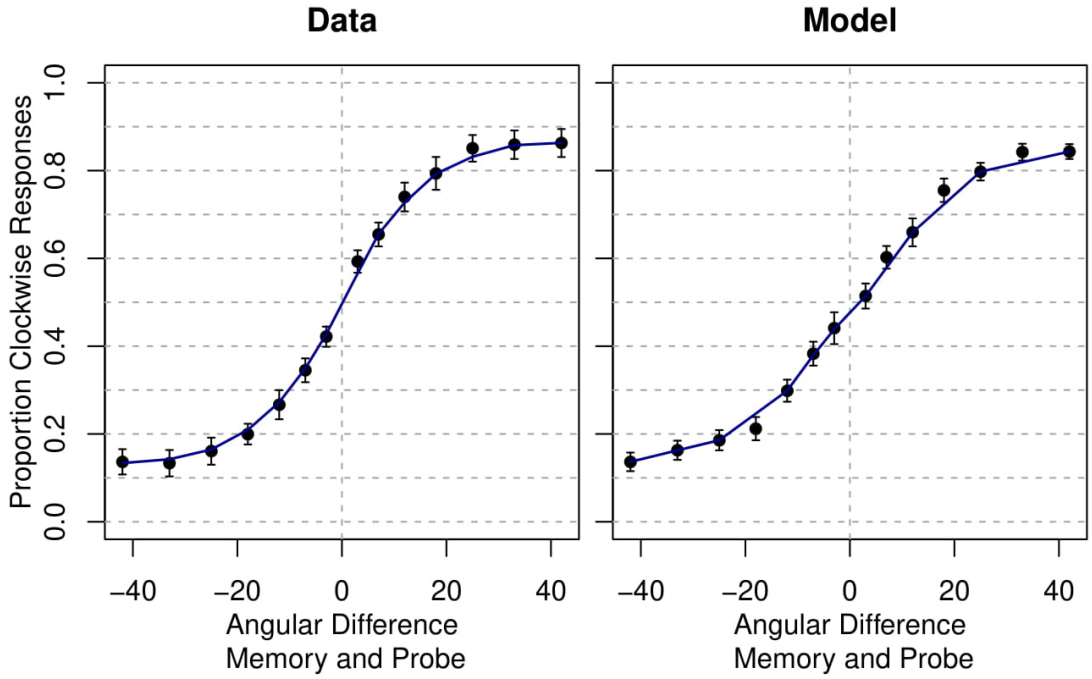

Figure 3.3: The Data (left) and Model (right) performance. The x-axis indicates how much the cued memory item and probe differ in degrees, while the y-axis reports the proportion of clockwise responses given for the angular difference on the x-axis.

discussed, the focus can now be shifted towards the performance of the model; after all, the augmented model should be functional. More specifically, as it models parts of human cognition, it should ideally be as good as (and not worse or better than) humans. Figure 3.3 shows the Data performance next to the Model performance. Absolutely perfect task performance (which would not be human) would give 'curves' that are a flat 0 (no clockwise responses given) for negative angular differences, rise straight up at an angular difference of 0 degrees and are a flat 1 (only clockwise responses given) for positive angular differences. The Data and Model curves are rather similar, with both subfigures showing an S-shaped curve. This indicates the augmented model is still a good approximation for human behaviour, just like the original Pals et al. (2020) model.

3.2 Experiment 2

Considering Experiment 2, Figure 3.4 sheds some more light on the dynamics of the Model with respect to the Data. The figure has the following build-up: A and C denote, respectively, the Data and Model decodability curves for the early- and late-tested items. Meanwhile, B and D, associated with the Data and Model results, respectively, each show two separate CTDAs; the left one in each subfigure depicts the dynamics of the early-tested item, while the right one gives the CTDA of the late-tested item.

Comparing to the results from Experiment 1, a number of differences, but also similarities, can be seen. For example, the Data curves in A do not show a clear second peak such as the one seen in Figure 3.1A, but are otherwise rather similar to it, with a steep ascent and a slow decay. However, the CTDAs in B do differ from the CTDA in Figure 3.1B, with the tested-early CTDA (left pane) showing a thicker diagonal, while the tested-late CTDA (right pane) shows a thinner one. The Model curves in C, on the other hand, are nearly identical to the curve in Figure 3.1C. The same can be said for the Model item representation dynamics in D, which both look very similar to the dynamics showed in

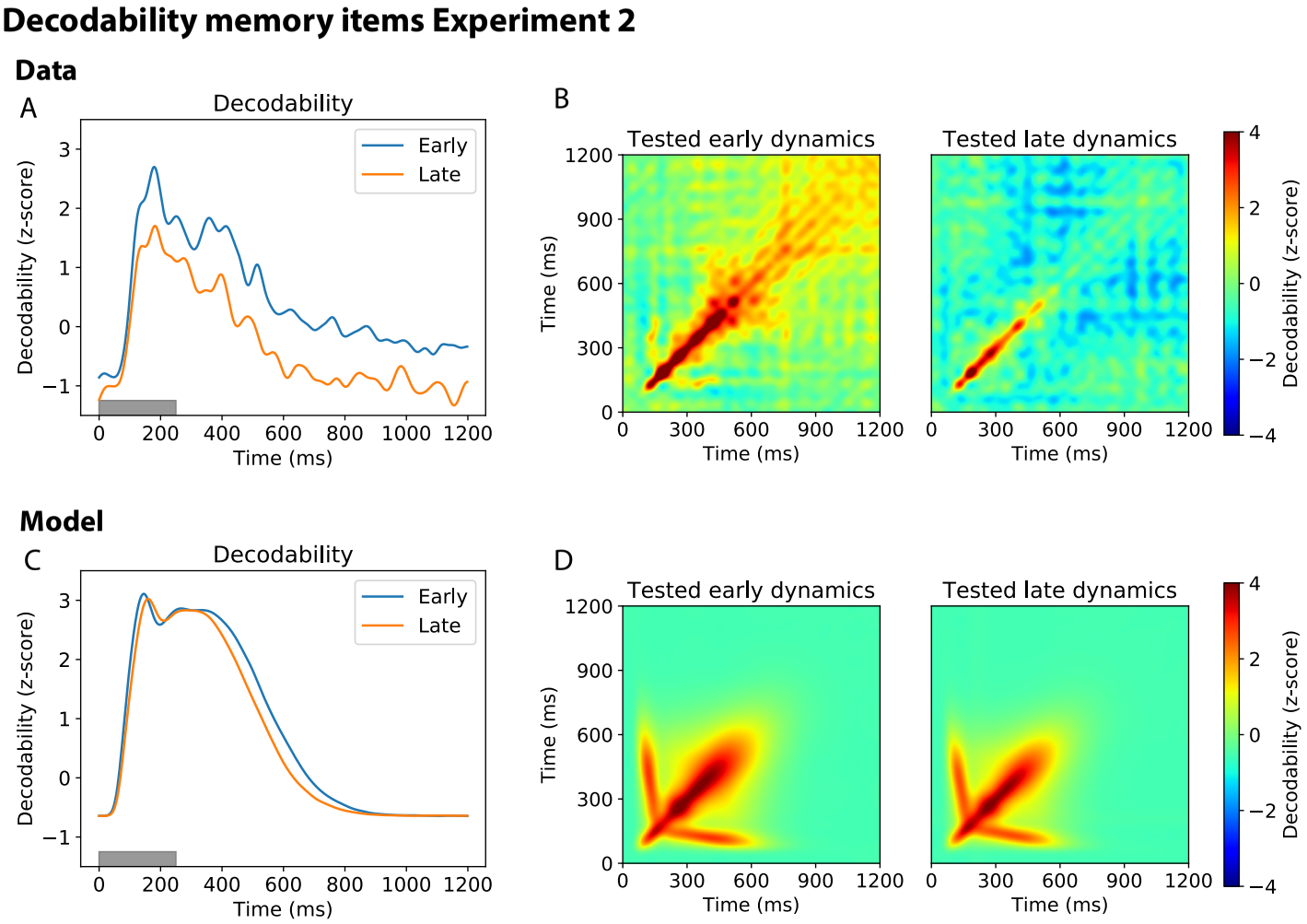

Figure 3.4: The decodability of the early- and late-tested item, just after their presentation. The top row shows the Data results, the bottom row shows the Model results (after Wolff et al., 2017). Each sub-figure (A, B, C and D) has a separate shared mean and standard deviation used in the calculation of the z-scores. ***A*** The decodability curves for the early- and late-tested Data memory items after their presentation. ***B*** The CTDAs corresponding to *A*. ***C*** The decodability curves for the early- and late-tested Model memory items after their presentation. ***D*** The CTDAs corresponding to *C*.

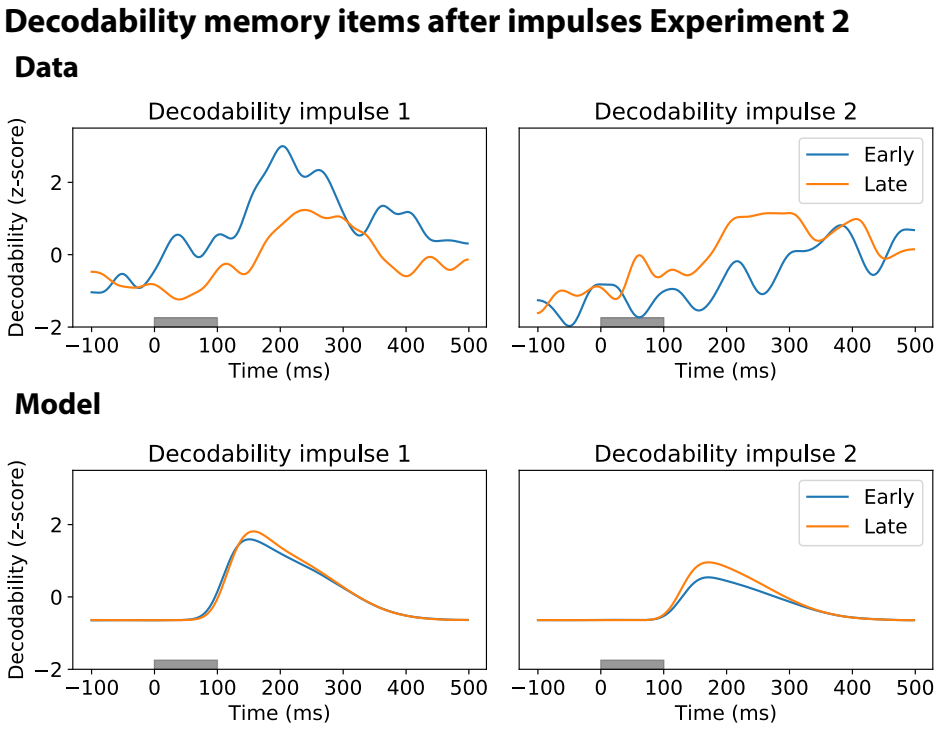

Figure 3.5: The decodability curves of the early- and late-tested items after the impulses of Experiment 2. The top row shows the Data curves (after Wolff et al. (2017)), the bottom row shows the Model curves. The left column shows the curves around the first impulse, the right column shows them around the second impulse. Time is relative to the corresponding impulse, and the grey bars indicate either the first impulse (left column) or the second one (right column). The Data z-scores share the mean and standard deviation used in their z-score calculation, as do the Model z-scores.

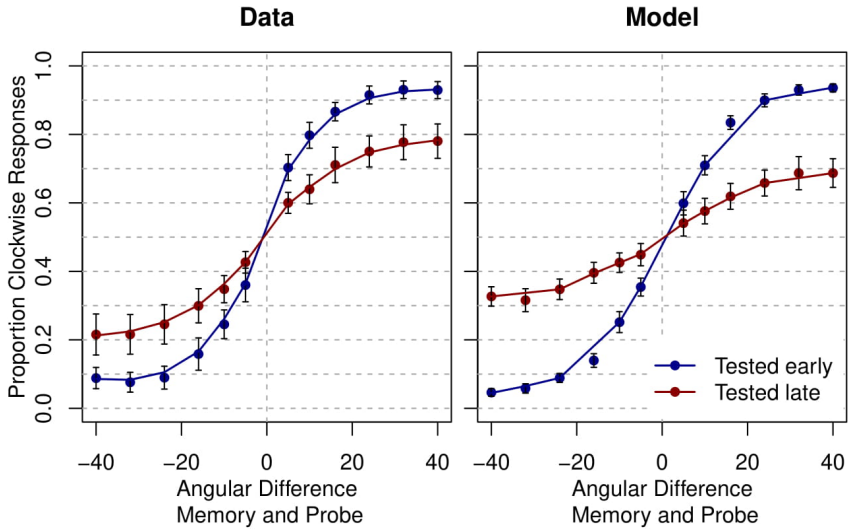

Figure 3.6: Data and Model performance for the early- and late-tested items.

Figure 3.1D.

Comparing the different results within Figure 3.4 to each other, more insights can be gained. For instance, there is not much of a difference between the decodability curves (C) and between the dynamics (D) of the early- and late-tested items for the Model. For the Data, however, there seem to be substantial differences. A suggests that the early-tested item was decoded much better than the late-tested item. In B, the differences seem even more extreme: While the tested late dynamics show a diagonal that resembles the diagonal shown in Figure 3.1B (albeit a bit shorter), the tested early dynamics show a diagonal that is stronger, continues for a longer amount of time and seems to 'fan out' as time progresses. When looking closely at the start of the diagonal (around the 200^{th} millisecond), two little arms can be seen, which are very clearly present in the Model dynamics.

Taking this into account, in some respect both the early- and late-tested Model dynamics are more similar to the late-tested Data dynamics, as neither shows a fanning out of the diagonals and the diagonals are of roughly the same length. However, in a different sense the Model dynamics are more similar to the early-tested Data dynamics, as they all show some sort of arms. Additionally, the thickness of the Model diagonals seem to match the early-tested Data diagonal better.

Looking further, especially at the decodability of the items after the impulses (Figure 3.5), a similar picture arises. The Data curves show a clear difference, with the early-tested item being more strongly represented after the first impulse, while after the second impulse, the late-tested item is more prominent. The Model curves show markedly less difference after each impulse.

Finally, the performance on Experiment 2 has to be considered. Considering Figure 3.6, it seems that the Model and Data performance are very similar to each other for the early-tested item, which would mean the Model is still a good behavioural approximation for the Data. However, concerning the performance for the late-tested item, the Model in its augmented form seems to perform quite worse than the Data, in contrast with the original model. Evidently, the late-tested item is not as well-preserved in the Model as it should be.

4 Discussion

A model created by Pals et al. (2020), originally designed to be a functional implementation of activity-silent WM and able to execute tasks from Wolff et al. (2017), was augmented to display dynamic coding patterns visible in human EEG data. The experiments from Wolff and colleagues were conducted again with the augmented model, and the resulting data were fed into decodability analyses. The results were compared to the human ones. This section will further discuss the similarities and differences between the model and human data, suggest points of improvement and further research, and finish by drawing conclusions on the overall fitness of the model, as well as on its implications for dynamic coding.

4.1 Decodability

In Section 3.1 and 3.2, all Model decodability curves are very similar with a steep ascent and a graceful decay, while the Data show some more variability in its decodability curves. One example of this is the clear second peak seen in the Data curve after memory item presentation in Experiment 1, which is not present in the Model curve. This might reflect some other processes being active at the same time in every trial in the participants' brains that are absent in the model (e.g. attentional shifts, feedback from higher-order stimulus-processing areas to lower-order areas et cetera), which in some way reinstate the stimulus representation.

As for Experiment 2, there seems to be a clear difference between the early and late Data decodability curves, while the Model curves are almost completely identical. A possible explanation for this phenomenon might be attention: Participants knew in advance which item would be tested first, so it is not unreasonable to assume that they might pay more attention to that item. Attention is known for strengthening neural responses to attended stimuli (Kastner and Ungerleider, 2000), which would suggest that it might also strengthen the WM representation of the early-tested item, explaining the difference in the maxima of the Data curves. The model does not implement attentional control. Instead, attentional effects were simulated by reducing the late-tested item strength to 90% of the

early-tested input, which might not be enough to create a differentiation between the early and late Model decodability curves.

4.2 CTDAs

The CTDAs show a picture that match the one shown by all the decodability curves. All Model CTDAs look alike, while the Data CTDAs show some interesting differences; not only among themselves, but also when compared to the Model CTDAs. Differences that already become clear from Experiment 1 are the periodic spots that appear in the Data CTDA but not in the Model one. They could be the result of some other, consistently recurring brain processes, just like such processes could cause the second peak seen in the Data decodability curve of Experiment 1. Another difference between the Data and Model CTDA is the absence of the arms in the Data CTDA. Perhaps this is related to the clean and stronger representation of memory items in the Model, which are undisturbed by any other processes, while in the human brain, these representations might have to contend with others. This view would fit in with the fact that the decodability curve of the Model has a higher maximum than the Data curve.

Considering Experiment 2, the differences between the dynamics of the early- and late-tested memory items are more pronounced for the Data than for the Model, corresponding to characteristics of the respective decodability curves. More specifically, for the Data, the early-tested diagonal is thicker than the late-tested one, and also 'fans' out into some semi-static pattern as time progresses. In addition, it has two small arms at roughly the same location the Model CTDAs have them, although the Data arms do not last quite as long as the Model ones. The variability in both the arms and the thickness of the diagonal can again be explained with the decodability curves: The higher the curve's maximum, the more pronounced the arms and the thicker the diagonal. This suggests that the strength of the item representation in WM has an effect on its dynamics as displayed in a CTDA, which in turn relates to the presumed effects of attention.

However, stimulus strength alone does not explain the high-decodability fan visible in the early-tested Data dynamics. An interesting note is that such fans also appear in the model parameter sweep, specifically for model versions with strong recurrent connections (Appendix A). A possible reason for the fan might then be the rehearsive aspects often associated with attention: Reactivation of the neurons responsible for the item representation would lead to an extended diagonal that eventually becomes static.

4.3 Impulse response

The responses to the impulse show a rather critical point of improvement for the model, especially in Experiment 1. Whereas the uncued Data decodability curve remains close to its minimum, the uncued Model decodability curve becomes nearly as high as the cued one. Clearly the model holds on too well to the uncued item, perhaps due to all the augmentations that were intended to prolong the item activation in the first place. The response to the second impulse from Experiment 2 suggests such a deficit as well: Although the early-tested item is no longer needed, its decodability is not much lower than that of the late-tested item. Combined with the fact that Wolff et al. (2017) showed that the late-tested item curve was significantly different from 0 and the early-tested item was not, this also indicates an item maintenance that works too well. Adding more noise to the model, in order to distort item representations more quickly, seems like a potential solution.

The fact that the Model impulse responses to the first impulse of Experiment 2 are nearly identical, is to be expected: Neither item has been reactivated more than the other at the time the first impulse arrives. Both items also start out with similarly strong representations after stimulus presentation (Figure 3.4C), presumably due to the lack of attentional control.

4.4 Performance

The Model performance for Experiment 1 and the early-tested item of Experiment 2 is roughly the same as the Data performance, which is good. However, the Model seems to perform quite worse on the second probe of Experiment 2, for the late-tested item. So while the uncued and early-tested item seem to be preserved too well in the model, the late-tested item seems to be preserved too poorly, which

would appear rather contradictory at first sight. However, attention might explain this contrast as well: Wolff and colleagues observed a strong lateralisation in their EEG data after the early-tested item was probed, and suspected that this lateralisation might reflect a shift in attention towards the initially deprioritised, late-tested item. Adding more noise to the Model to allow it to quickly 'forget' unnecessary items, while also adding some form of attentional control that would allow it to reinstate WM items that become important only later on, would then seem the right approach for a more human-like model.

4.5 Dynamic coding

Overall, the model data show many similarities to the human data, while also showing some important differences. To combat these differences, it would seem that the addition of attentional control paired with a bit more noise added to the memory population might suffice. Most crucially, however, the model displays clear diagonals in its CTDAs, which are hallmark patterns often cited in the literature as evidence for dynamic coding. Namely, as mentioned before, such diagonals indicate that a decoder trained on data at some time step t can decode data at that same time step t very well, but data at a more distant time step, say $t + 1$, rather poorly, which suggests a dynamic item representation. The fact that the model displays such patterns has a number of implications.

One of the first is that the calcium-mediated STSP mechanism as proposed by Mongillo et al. (2008) and implemented here by Pals et al. (2020) is a good basis mechanism for displaying dynamic coding. The original model by Pals and colleagues also needed some extensions, however: The pattern it displayed in a CTDA was simply too short to be classified as either dynamic or static. As the present study has shown, prolonging the WM item representations was the key to seeing dynamic patterns appear in the model data. Having said that, prolonging the representations would seem more of a practical necessity for being able to properly see the patterns, rather than a fundamental property that a dynamically coding network should have. In contrast, the self-modifying nature of networks that implement an STSP mechanism might be what enables dynamic coding. After all, if a network has changed its structure due to previous activation, it is unlikely it will represent the some input it received before in exactly the same way as it did then.

Then what does this mean for the dynamic coding framework as proposed by Stokes (2015)? Stokes sees dynamic coding as the result of a complex reciprocal interaction between some network's activity states and its hidden states, which would result in a complex trajectory through the activity state space of a neuron or a complete neuronal population. In some sense, this view seems rather fitting: The augmented model uses a hidden state, namely the calcium and resource variables, and that state is modified through network activity. This in turn allows for dynamic coding, and, if so desired, the accompanying trajectories through activity state space by the memory population, or even its individual neurons, can very well be called complex.

However, it is questionable whether these terms are appropriate for what dynamic coding might actually be: A property of any self-modifying activation network. Dynamic coding might be simpler than the 'big terms' that Stokes uses would make the reader suspect. For instance, "temporal variability at the very shortest timescales", as Stokes mentions, turned out to be hardly necessary for dynamic coding: Although rapidly changing neurophysiological parameters might have some influence on WM dynamics in the human brain, a resolution of two milliseconds proved enough for the model to mimic those dynamics. The underlying conceptual framework that Stokes proposes seems appropriate, but it might be best to appraise it in a more modest way, foregoing a focus on terms like "a complex spatiotemporal trajectory through state space".

In contrast, a more interesting question that still remains unanswered is how a network can maintain the same information through time even though the information *representation* is dynamic. That question has not been addressed by this study, but judging by the fact that the augmented model can still perform its task while also displaying dynamic coding, the model is capable of this information maintenance even though the information representation is dynamic. This makes the current model a promising tool for future research on dynamic coding and its importance for human WM and broader brain functioning.

References

A. Baddeley. Working memory. *Science*, 255(5044): 556–559, 1992.

C.E. Curtis and M. D'Esposito. Persistent activity in the prefrontalcortex during working memory. *Trends in Cognitive Sciences*, 7(9):415–423, 2003.

M. Daneman and P.A. Carpenter. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4): 450–466, 1980.

R. De Maesschalk, D. Jouan-Rimbaud, and D.L. Massart. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

C. Eliasmith. *How to build a brain: A neural architecture for biological cognition.* New York, NY: Oxford University Press, 2013.

S. Kastner and L.G. Ungerleider. Mechanisms of visual attention in the human cortex. *Annual Review Neuroscience*, 23(1):315–341, 2000.

J.R. King and S. Dehaene. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210, 2014.

V.A.F. Lamme and P.R. Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neuroscience*, 23(11): 571–579, 2000.

G. Mongillo, O. Barak, and M. Tsodyks. Synaptic theory of working memory. *Science*, 319(5869): 1543–1546, 2008.

M. Pals, T.C. Stewart, E.G. Akyürek, and J.P. Borst. A functional spiking-neuron model of activity-silent working memory in humans based on calcium-mediated short-term synaptic plasticity. *PLoS Computational Biology*, 16(6): e1007936, 2020.

K.K. Sreenivasan, C.E. Curtis, and M. D'Esposito. Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, 18(2):82–89, 2014.

M.G. Stokes. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences*, 19(7):394–405, 2015.

K. Watanabe and S. Funahashi. Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nature Neuroscience*, 17(4):601–611, 2014.

C. Wijs. Dynamic coding in a neural model of activity-silent working memory. Unpublished manuscript, 2020.

M.J. Wolff, J. Jochim, E.G. Akyürek, and M.G. Stokes. Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, 20(6):864–871, 2017.

R.S. Zucker and W.G. Regehr. Short-term synaptic plasticity. *Annual Review of Physiology*, 64(1): 355–405, 2002.

A Sweep results

The resulting dynamics of a parameter sweep over three parameters, namely *(1)* the strength of the feed-forward connection with distributed latency from the eye to the sensory population, *(2)* the strength of the recurrent connection from the sensory to the eye population and *(3)* the standard deviation of the noise added to the memory population. Every column indicates a particular combination of feed-forward connection strength and recurrent connection strength. Each row indicates a specific standard deviation of the noise. A wide range of patterns becomes available, including the fanning out of the diagonal seen in the human data of Experiment 2 from Wolff et al. (2017).