Modelling Human Auditory Time Perception through CNN classifiers of natural sounds

**James Burgess**

**Artificial Intelligence and Adaptive Systems, Msc**

Supervisor: Warrick Roseboom

Candidate Number: 237061

Submitted 31st August 2021

**1.0 Abstract**

**2.0 - Introduction**

Perception is arguably the most crucial aspect of brain function in generating the conscious experience, having to make sense of the external world within a brain hidden from this world. For organisms to behave, interact with and survive in this world perception is vital, with consciousness science thoroughly focused on decoding how noisy perceptual inputs can result in adaptive predictions to directly guide action and conscious thought (Friston, 2005), (Clark, 2013). All organisms that interact with their environment rely on perception in one way or another, detecting light, sound, chemicals, or physical vibrations from humans to the simplest of organisms (Burnett, 2011). Decades of neuroscience research has focused on understanding the transmission and decoding of signals that occur between the sensory organ and the brain. However, the concept of time passing is understood, at least by most complex organisms (Healy et al., 2013) and is undoubtedly a feature of the human conscious experience yet does not rely on a sensory organ to function. With no dedicated ‘input’ to study research on time perception within modern neuroscience had taken a backseat. The ‘internal clock model’ model’ (Treisman, 1963), (Church, 1984) was proposed to explain this conscious experience, yet is flawed in nature of the strong assumption of the presence of a dedicated metronomic like counting system in the brain, to which no significant evidence has been found that such a system is present. Furthermore, time perception is not perfect, we do not perceive ‘clock time’ (aka veridical time) as patterns of behaviour have demonstrated biased trends of overestimations for short durations and underestimations for longer durations (aka Vierordts law) (Lejeune and Wearden, 2009). Most significantly, variations from veridical time have been strongly linked to differing perceptual content and context (Staddon and Higa, 1999), (Zimmermann and Cicchini, 2020). This is such a significant factor in the perception of time that we have developed common phrase directly correlating perceptual content/context with time perception: “A watched pot never boils”, “Like watching paint dry”, “Time flies while you’re having fun”. These content specific variations can be attributed to stimuli properties such as complexity (Block, 1978), rate of change (Herbst, Javadi and Busch, 2012), (Linares and Gorea, 2015) and competing stimuli (Arnold, Tear, Schindel and Roseboom, 2010). Still, the Internal Clock model proved to shows high explanatory power for results generated from constrained scientific experimental conditions such as interval timing, replication, and temporal bisection studies (Church and Deluty, 1977), (Allman, Teki, Griffiths and Meck, 2014), (Gibbon, 1977). Yet this model falls short of explaining the far more naturalistic aspect of time perception in which these clear deviations occur.

To understand and hopefully explain the underlying neural aspects of time perception it is crucial to explore time perception in the domains it is present, i.e., all the time in natural environments for the organism. However, this proves an obvious challenge of controlling for the countless variables that exist in the world we interact with, and an even greater problem of then trying to conclude what aspects affected time perception and importantly how. Recently a new approach has been developed that seeks to solve these challenges. With the growth of machine learning and developments in neural networks (NN) inspired by the activity and learning of neurons (Cox and Dean, 2014), the similarity between convolutional neural networks (CNNs) and cortical perceptual processing in the visual stream has presented the opportunity to model perception processing (Kriegeskorte, 2015). A recent study investigating the dynamics of an image classification CNN, was able to replicate human time perception results, showing clear perceptual content dependent biases (Roseboom et al, 2019). Importantly this was strongly linked to perceptual processing, with model performance most accurately replicating human duration reports when image input to the CNN was confined to the region of visual field currently being processed, observable through gaze tracking of the pupils. These findings were achieved by investigating the dynamics of individual layers as new perceptual information was passed to the network, with greater change in perceptual content, the greater the difference in layer activation between the successive states. Applying a threshold detection mechanism to the network dynamics allowed for detection of salient events, (noticeable changes in perceptual content), where the detection of salient events generated accumulation of ‘fundamental time units’ that are consciously perceived as the flow of time. These fundamental conscious time units can be converted to a quantifiable unit of time, seconds, simply through regression fitted against similar durations presented in seconds as the correlation between the ‘perception’ and the quantifiable is learnt. Thus, this proposes a new model of time perception, one that does not rely on an internal clock or any isolated system, but directly linked to sensory perception.

Thus far, this has only been tested in the visual domain due to the prevalence of image CNN classifiers. To further validate this theory, and eventually reach a stage where time perception is understood as fully intertwined with the distributed nature of consciousness, this theory must be tested in other sensory domains. This project continues from the framework presented in (Roseboom et al, 2019), focusing within the audio domain. Constraining our approach to the auditory domain poses further modelling challenges than with the visual domain. Biologically the CNN and visual stream are highly similar as they can both operate via of a sequence of inputs, frames in a video and sampling frequency of the retina (Eisen-Enosh et al., 2017) followed by feature detection such as edge detection to ‘build’ objects and distinguish between endless object classifications (Canny, 1986). However, audio classification is a continuous ongoing perception, sound waves are not ‘sampled’ by the ear but constantly being obtained and affecting the action of previous and future audio intake though continuous vibration of the cochlear basilar membrane (BM) (Von BÉKÉSY, 1970), (Baby, Van Den Broucke and Verhulst, 2021). Due to this, it did not make sense when audio classification models were being built for speech recognition to replicate the biological process of auditory processing to the cortical level in the same way as speech. Further, significantly little is known about the organization of the cortical stream in which audio is processed when compared to visual processing (Rauschecker and Scott, 2009). Thus, for audio classification, image classifying CNNs are used (Kell et al., 2018), (Shuvaev, Giaffar and Koulakov, 2017), (Eghbali and Hajihosseini, 2019), (Van Meer and Buermann, n.d.).

To evaluate how image representations of audio act both as means for classification of natural environmental audio, and as a means for extracting viable estimation durations; we investigated three different forms of image representation. A basilar membrane displacement image representation, cochleagrams and Mel-Spectrograms. By extracting duration prediction outputs of each model during classification from a hidden layer within the CNN, we could compare these results to estimation reports given by 10 people over the same set of sound files to evaluate model time perception accuracy. Our results from this analysis showed that although the worst of the classifiers, only the Basilar membrane displacement model showed potential for time perception extraction. However, results did not come close to matching human estimation reports. Yet we still believe there is significant room for improvement of this model that can potentially reach a level on-par with human time perception.

**3.0 Methods**

Participants – 3.1

For the online behavioural experiment, 12 adults (age range of 22 to 55, consisting of 3 females and 9 males) were recruited through the lead researcher to voluntarily take part in this pilot research experiment. Two of these participants were later excluded from our final analysis due to failures of completing the attentional component of the experiment, detailed in section 3.2, resulting in a final sample size of n=10. Within the online experiment was a mandatory information sheets and consent forms for participants to register that they understand the presented information and consenting to take part in the experiment. This experiment was approved by the University of Sussex ethics committee.

Sound Stimuli – 3.2

For this project we focus on natural sound scenes, following from the framework applied to (Roseboom et al, 2019). Being that time perception is a continuous feature of conscious perception, it is thus preferable to investigate continuous and natural sound scenes where time perception will be more prevalent for the participant due to a lesser need for task-based classification such as interpreting speech or music. Considering the importance of stimuli content in time perception (Roseboom et al, 2019), a proposed natural sounds database by (Huang and Elhilali, 2017) was used. Importantly, this database (consisting of 20 audio scenes, however only 15 were available) had labelled each sound scene as either dense or sparse for the number of salient events throughout the scene. 4 scenes were labelled as sparse for salient events, 11 were labelled as dense. Table 1 shows the 15 used sound scenes. In the behavioural experiment we attempted to control for limiting all other external stimuli so that only audio had the maximum contributing factor towards time perception. Participants were instructed to wear headphones for the trials and keep a fixed gaze on a central cross on an otherwise blank screen, excluding for an image of a speaker appearing for 1s to indicate the onset of sound. To control for attention during the trials and to ensure participants kept their fixated gaze, we implemented both visual and auditory attentional response cues, where the central cross would change colour or a tone bleep occurring during the sounds. The timings of these auditory cues were presented at randomly distributed. times within the final half of the chosen sound durations, to bias participants into expecting the possibility of an attention cue up until the end of the sound file. The response times of participants, by pressing space, were recorded for each attentional cue and the mean response time was used as exclusion criteria for participants in final data analysis.

All sound files were presented in stereo format, at 44.1kHz with a 32-bit rate. Durations of sound files were chosen in a range of 5 to 35 seconds, with each sound file repeated four times each at different durations (taken from different sections of the sound scene). We employed a factorial design, such that each participant experienced the same durations for each sound file, with attentional cues on the same sound files, however these were presented in a unique randomised order for each participant. Four duration bands of 7.5s were created to ensure an even distribution of sound files and their durations, 5-12.5s, 12.5-20s, 20-27.5s and 27.5-35s. Each sound file was present in each duration band, and we ensured that each sound was separated from its repeats by a minimum of 4s. Further, each duration band contained two attentional stimuli (one of each modality). To maximise data collection for the smaller sample size of sparse labelled audio, attentional cues were not placed within any sparse labelled scenes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sound name** | **Durations (s)** | **Saliency** **label** | **Rhythmic label** | **Emotional label** |
| Blacksmiths | 12\*, 17, 22, 28 | Dense | Rhythmic | Neutral |
| Blowing Alley | 11, 18, 27, 32 | Sparse | Non-Rhythmic | Neutral |
| Cafeteria | 6, 16, 22, 30\* | Dense | Non-Rhythmic | Neutral |
| Car Chase | 10, 20, 24, 29 | Sparse | Non-Rhythmic | Neutral |
| Carnival | 9, 15, 27, 33 | Dense | Non-Rhythmic | Positive |
| Dog Park | 7, 14, 26\*, 33 | Dense | Non-Rhythmic | Neutral |
| Drum line | 11, 19, 25, 34 | Dense | Rhythmic | Positive |
| Maternity Ward | 5, 13, 21, 28 | Sparse | Non-Rhythmic | Negative |
| Nature sounds | 10\*, 16, 25, 31 | Dense | Rhythmic | Positive |
| Orchestra tuning | 6, 14, 21, 32\* | Dense | Non-Rhythmic | Neutral |
| Piano | 5, 17\*, 24\*, 35 | Dense | Rhythmic | Positive |
| Protests | 8\*, 13, 23, 30 | Dense | Non-Rhythmic | Negative |
| Sailing Barge | 8, 17, 23, 29 | Dense | Rhythmic | Neutral |
| Sports | 9, 15, 26, 31 | Sparse | Non-Rhythmic | Positive |
| Store | 7, 18, 25, 30 | Dense | Non-Rhythmic | Neutral |

*Table 1*

List of sound files and their respective durations, saliency label, rhythmic label and subjective emotional label. Sound files that contained attentional cues are marked with ‘\*’.

Behavioural Experiment Procedure – 3.3

The behavioural experiment was programmed in PsychoPy software (Peirce et al., 2019), and uploaded online to Pavlovia for participants to complete. In the information sheet presented prior to the experiment, it was strongly stressed that participants should not count, refer to a clock or tap along as the sound files were presented so that we could accurately investigate time perception, rather than counting ability. The pilot framework for this experiment comprised of a practice session, where participants received feedback on duration estimations, followed by the main 60 trial section from which we collected data. Practice comprised of 8 trails followed by feedback, presenting the real duration of the sound clip and the difference between the estimation and real duration. Sound files heard during the practice section were taken from the same sound files as the main experiment, thus we ensured the sections of sound were not present in the main experiment to prevent increased familiarity. Each duration band was present twice through practice, with durations of the practice sound files a minimum of 5 seconds difference between the duration of the same sound files in the main experiments. Two attentional cues were also present in the practice session to ensure participants were familiar with the procedure. Feedback for durations was indicated by horizontal mouse movements corresponding to changing the duration estimation score shown on screen, followed by a left click of the mouse to confirm the estimation. This feedback reporting style was chosen to decrease any significant regression to the mean effects that can accompany visual scale reporting (Yu and Chen, 2015). The main experimental component didn’t show participants feedback for all duration reports. Within this section, participants completed 60 trials, 8 of which contained attention cues, such that we collected 52 trials for each participant. Upon receiving feedback from participants, 3 participants reported errors in which sound files unexplainedly became silent after ~1 second for a varying number of trials. As a result of this, we excluded any estimation reports < 1.5 seconds after manually investigating the data where it is clear these errors occurred. This error is still unfortunately unfixed with no correlation between operating software or browser used to run the experiment. Further, we also excluded estimation reports that were identical to the previous trial estimation after half of participants reported that on at least one occasion they accidentally mis-clicked at the beginning of the estimation reporting without moving the mouse. In future experiments we should prevent clicking to confirm until at least 1 second has passed from the onset of estimation reporting.

Audio pre-processing for Neural Network input – 3.4

Prior to input into the neural networks, audio was converted into three visual representations of audio: mel-spectrograms, cochleagrams and basilar membrane (BM) displacement graphs, shown in figure 1. Initially all full-length audio clips were converted to mono and down sampled to 8kHz to improve computational run time of models as this significantly decreased file sizes. To ensure that we had an even sample size of each sound file, so that we could prevent model bias towards classifying any sound class, the initial 75 seconds of every sound clip was used (duration of the shortest file). This allowed us to train models without exclusively using data that would be presented to the models during time perception investigations. However, this was obviously not possible for the shortest duration sound files, but this then allowed us the ability to investigate if overfitting of learning of short duration (raw and unedited) sound files had a contributing factor towards time perception. The sound files were then broken down into 1s long signal representations prior to transformation into visual representations of audio. Mel spectrograms were generated through use of the ‘Librosa’ python module, applying Mel bands to time-frequency signals to logarithmically scale the representation, enhancing representation of optimum human frequency ranges. Cochleagrams were generated using open-source python functions generated by (Kell et al., 2018). Lastly, basilar membrane displacement patterns across 201 central frequency sections of the basilar membrane were predicted using the CoNNear CNN model (Baby, Van Den Broucke and Verhulst, 2021). Images were converted to raw RGB data and rescaled to 100x100 pixels of depth 3 for uniform input shape into the CNN models. To improve classification accuracy of the classifiers and improve generalisation, this data was normalised by dividing all values by 255 such that all values were between 0 and 1. Lastly the ordering of all image data was shuffled batch sizes fed into the models contained a varied distribution of sound classes so that overfitting to specific sound classes didn’t occur.

A picture containing text

Description automatically generated

*Figure 1*

Example images of audio representations used for NN model inputs. All three examples show the same representation of a 1s section taken from the Piano sound file. (A) – Mel Spectrogram representation. (B) – Cochleagram representation. (C) CoNNear basilar membrane displacement predictions.

Classification Models – 3.5

To create our classification model, we recreated a CNN model architecture previously used for spectrogram speech classification (Salamon and Bello, 2017). This model comprises of three convolutional layers with 5x5 receptive fields, two maximum pooling layers with two final fully connected layers. See table 2 for full architecture. Activation functions used for convolutional layers and the non-final dense layer was the ReLu function. Dropout occurs twice in this model, occurring before the final two layers, at a rate of 0.5 in both instances preventing the high potential for overfitting for image classification with high similarities and minimally distinctive features. The final dense layer of the model contained only 15 nodes, activating with a SoftMax activation function outputting the model’s calculated probability of each class, where each node activation corresponds to the probability of a single class. Three different models were generated using this architecture, each specific to the visual representation input type such that we did not have to further account for the models classifying the input type if a single model was used. For each training data set we obtained 75 1s long image representations across 15 classes, generating a whole sample size of n=1125. During model training, this sample size was decreased to hide 20% of the data from the model such that the remaining data could be used to evaluate validation accuracy to observe potential overfitting. For training the model, our data sample labels were one hot encoded allowing use of categorical cross entropy loss function.

Optimisation of network weights and biases was performed using the ‘Adam’ gradient descent optimiser (Kingma and Ba, 2015), allowing for improvement in training compared to traditional SGD as used by (Salamon and Bello, 2017) through utilising adaptive learning rates. Training Epoch number was decided through trial and error until it became apparent that validation loss had plateaued, resulting in epoch iterations of 150, 50 and 100 for BM displacement models, Mel Spectrogram models and Cochleagram models respectively. From judging the results of, testing accuracy, testing loss, validation loss and validation accuracy, seen in figure 11, we were confident that our models all operated to an acceptable level and didn’t show indications of overfitting. However, it was still necessary to perform cross validation studies using a new data set to confirm this. As validation loss and accuracy was only produced by 20% of the shuffled data from the original 75s of audio, it was still likely that hidden data would have been highly similar to the test data, without containing any unseen features within the image representations. To observe if the models were able to generalise to sounds where unseen features were present, the following 50s of audio was collected from all sound files of minimum duration 125s (all bar ‘Sailing Barge’ and ‘Car Chase’). A prime example of the benefit of this is from the ‘Sports’ sound file, where the initial 75s only contains generic crowd noise and cheers while the following 50 seconds contains the outbreak of a football chant of ‘We love you Liverpool we do’. To evaluate the models’ performances across all three input types, and the generalisability of our trained models we generated classification matrices and mean classification accuracy.

Modelling Time Perception – 3.6

To model time perception and compare and contrast results with human performance, we used the same audio data set as used in the behavioural experiment to identify if the model produces similar content dependent biases towards time perception for specific audio.

For reasons further explained in discussion section 5.1, time perception was investigated both with 1s long visual representations, as we have used for training and cross validation, and 0.1s long representations. So that 0.1s representations could be fed into the model, for each 1s long representation 0 padding/replacement was performed to all but a 0.1s section of the representation. The resulting 0.1s data was placed centrally in the 100x100 pixle input to the model as this was essential for consistent Euclidean distance analysis, as if positions of the input are not kept consistent then the resulting Euclidean distance will show significant spikes and peaks (as shown in supplementary image 2). This generated 10 equal dimension representations from a singular representation, increasing our sample size for detecting salient events in time perception analysis by 10fold. An example of this transformation is shown below in figure 2. Model predictions were made for each image representation of a single second from each audio file, such that for the blacksmith sound in duration band 1 (see table 1) that lasted 12 seconds – 12 predictions were made. In order to detect salient changes occurring in network activation, Euclidean distance between single node activation for successive predictions (i.e., second 1-2 vs second 2-3) was calculated for each node in a layer, summing Euclidean distance for each node between successive states such that we could create a single unit distance measure of a single layer. For each sound file, the Euclidean distance was calculated between each successive 1s or 0.1s inputs. Figure 3 below shows the activations mapped to each filter layer between successive inputs into the Blacksmiths sound file.

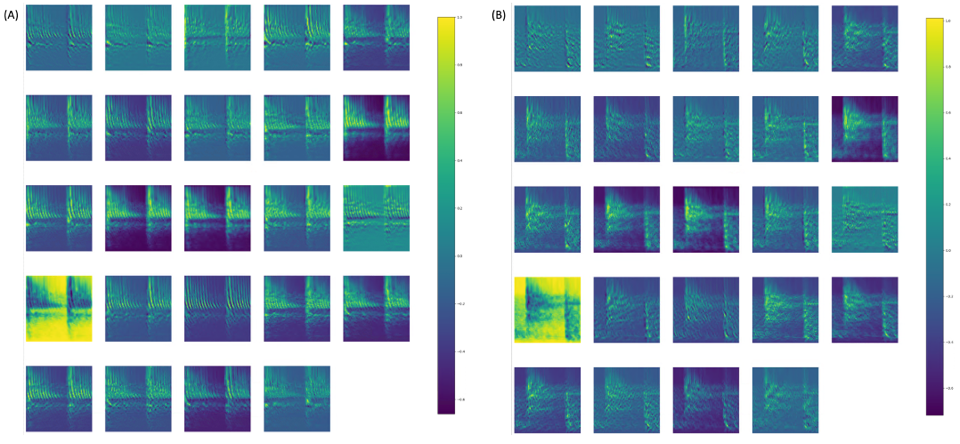
A screen shot of a map

Description automatically generated with low confidence

*Figure 2*

(A) – Full 1s long visual representation of the start of Blacksmith (12s) audio.

(B) – 0 padded 0.1s long visual representation of 0.4s-0.5s from the same Blacksmith (12s) audio



*Figure 3*

Activation comparison between blacksmith BM model input at 3s (A) and 4s (B) for the first convolutional layer with 24 features.

We analyse the output of the 2nd max pooling layer, 3rd convolutional layer and the 64 node fully connected layer across all 3 model types. We experienced limitations in how many layers we could analyse due to time constraints, which was significantly affected by number of nodes in each layer. This limitation was most notably identified when investigating the first convolutional layer for the BM model, as for the 94x94x24 node layer, across an average of 150 predictions (of 0.1s temporal bin inputs) for each of 60 sound files; ~2,000,000,000 Euclidean distance operations had to be performed resulting in a runtime of over 10 hours to generate Euclidean distance results for each model type. After two iterations of attempting to generate results for this layer, followed by kernel crashes wiping the data, it was deemed an unprofitable use of time. Thus, by limiting analysis to the third convolutional layer (1x18x48), the maximum number of operations to perform was 7.8 million for each model type. For each Euclidean distance analysis, i.e., for each model type new dynamic threshold values needed to be calculated. The equation for updating the dynamic threshold, taken from (Roseboom et al, 2019) is shown below in equation 1. Where Tmax and Tmin are the maximum and minimum threshold values, is the decay time constant, D is the number of individual time steps since threshold surpassed the Euclidean distance (detection of a salient event). To select maximum and minimum values for the thresholds for each input type and model layer we collected: the maximum Euclidean distance across all sound files, average maximum Euclidean distance from each sound file, the minimum Euclidean distance across all sound files, the average minimum Euclidean distance from each sound file and the average Euclidean distance from each sound file. Tmax was set between the global average and the average maximum, and Tmin was set between the global average and average minimum value. From this, were chosen to adjust the decay rate so that only the majority of salient detections were surpassed the threshold. Threshold decay is stochastic through random selection from a gaussian ( distribution centred around 0 with variance of Tmax – Tmin scaled by to adjust the noise effect that either increases or decrease the rate of decay. Upon the threshold surpassing the current Euclidean distance at that time step, threshold resets to the maximum value (Tmax), and is maintained until Euclidean distance decays below threshold. D is increased by 1 for every time step that threshold is below the current Euclidean distance. Tmax, Tmin, and values are shown in table 3 for all model types for 3rd convolutional layer.

|  |  |  |  |
| --- | --- | --- | --- |
| *3rd Convolutional Layer* | Basilar Membrane Model | Cochleagram Model | Mel Spectrogram model |
| Absolute minimum Euclidean distance | 1.34 | 2.32 | 13.86 |
| Average minimum Euclidean distance | 31.64 | 6.87 | 24.12 |
| Absolute maximum Euclidean distance | 303.41 | 177.82 | 255.63 |
| Average maximum Euclidean distance | 185.85 | 86.51 | 162.74 |
| Global average Euclidean distance | 81.89 | 29.12 | 69.47 |
| **Tmax Value** | **133.87** | **57.82** | **116.10** |
| **Tmin Value** | **56.77** | **18.0** | **46.75** |
|  | **50** | **50** | **50** |
|  | **50** | **50** | **50** |

*Table 3*

Dynamic Threshold mechanism parameters choice for 3rd Convolutional layer

Bold values indicate values used in dynamic threshold update equation (equation 1), whilst all other figures are collected from Euclidean distance data for each specific model type in this layer.

Euclidean distance data was collected for every 0.1s within each sound file for each model type, allowing threshold mapping to each sound file. From threshold data we were able to calculate the number of fundamental time units accumulated by the number of times threshold surpasses the Euclidean distance value. From the number of accumulated time units, support vector regression (SVR) fit the accumulation number for each sound file against the true duration of each sound file. Thus, the support vector regression model can predict the duration of a sound file in seconds exclusively through accumulation of time units. The SVR models were used to predict durations for all 3 model inputs for the above mentioned layers. Unfortunately, due to significant time limitations we were unable to evaluate trends of duration predictions for sparse and dense separated sound files. However, this is not that significant considering there is still further work that needs to be done to amend the models to accurately predict human estimation reports.

**4.0 Results**

In this project we aimed to model the newly proposed neural mechanisms of time perception, deviating from the previous consensus of requiring an ‘internal clock’ (Treisman, 1963), (Church, 1984). We aimed to replicate human time perception patterns and biases of natural environmental sound within convolutional neural networks (CNN). Modelling of cortical auditory processing was performed using specially trained CNN natural soundscape classifiers, with specific models generated for multiple visual representations of audio; Basilar membrane displacement/vibrations, Mel Spectrograms and Cochleagrams. The applicability of each audio input representation for use in natural sound classification methods was investigated through training of the CNNs and cross validation across 15 classes of natural sound. Importantly, this also allowed for analysis of the significance of classification capabilities in the role of extracting temporal information from network dynamics. Time perception capabilities of each model were investigated through comparison against 10 human participants completing 60 trials of duration estimations reports. Due to the use of natural sounds database developed for the purpose of classifying auditory salience, we were further able to investigate the well-documented correlation between perceptual content and time perception (Staddon and Higa, 1999), (Roseboom et al, 2019), (Zimmermann and Cicchini, 2020). To demonstrate if the perceptual content of audio presented to both humans and CNN models had a contributing factor to duration estimation reports, direct comparisons between audio containing either sparse or dense salient events were evaluated, along with subjectively labelled rhythmic/non-rhythmic audio and finally also subjectively labelled positive/neutral/negative audio.

Human Duration Estimations for Auditory Stimuli show a consistent underestimation exacerbated with increasing duration – 4.1

Participant estimation reports were compared against the true duration of audio clips for a maximum of 52 trials, with the possibility of single trial exclusions as a result of silent audio files or mis-clicks by participants. Figure 4 below shows a single participant example for these results, showing the estimated vs veridical time and divergence from veridical time as a percentage. This participant shown (participant 1) underestimated duration on all 52 trials, however maintains a relatively linear underestimation pattern for durations 5-15s with a relatively tight distribution compared to durations > 15s. Following this, participant error clearly significantly increases as duration time increases showing a maximum underestimation of 19.2 seconds. Interestingly, no estimated time surpasses 25 seconds, suggesting that this participant was not significantly biased by the upper possible limit of estimation reporting of 48s. However, the participant still shows a clear regression to the mean of the presented audio clips. Figure 5 illustrates the number of estimations this participant made for each 7.5s duration band, showing a significant bias towards duration bands 1 and 2 with no estimations made in band 4. If we were to experience a strong regression to the mean of all sound durations (19.9s) then we would expect band 3 to attribute more estimations than band 1 as the mean value of 19.9s is significantly closer to the lower boundary of band 3 (20s) than the upper boundary of band 1 (12.5s). However, we don’t see this effect with this single participant, rather a general pattern of underestimation that effect increases with duration increase. Overall, this single participant had a mean divergence from veridical time of -7.12s, a variance of 431.68 and an R2 correlation of 0.277 (where 1 would indicate perfect time perception and 0 is fully random selection).

Chart, scatter chart

Description automatically generated

*Figure 4*

Single participant example of results generated from online behavioural time perception experiment over 52 trials.

In both images, the red line indicates that any data point below the line is an underestimation of duration, and anything above as an overestimation.

(A) – Estimated time vs Veridical time of each audio clip.

(B) – Divergence of the estimation from veridical time

Chart, histogram

Description automatically generated

*Figure 5*

Distribution of participant 1 estimations into the four duration bands used to evenly distribute sound file durations. Each duration band had 15 sound files each, bars that surpass the red line have attributed more estimations within that duration band than were actually present

To ensure data analysis was only using reliable data from participants who actively paid attention to sound audio and were not distracted/on their phones, the mean response time across all 8 attentional cues was used as an exclusion measure. These results can be seen below in figure 6. Participant 9 did not engage with any behavioural cues indicating instructions were not read (thus importantly we are unsure if they read instructions not to count), leading to automatic exclusion. Although a maximum response time that would result in exclusion had not been established prior to this analysis, the clear outlier from participant 10 from the rest of the participants warrants exclusion as well. For future experiments using this framework, we would suggest a maximum mean response time of 1s to allow inclusion.

Chart, box and whisker chart

Description automatically generated

*Figure 6*

Mean response times taken from 8 attentional cues (audio and visual are averaged together as the same response was needed for both). Error bars represent the standard deviation across all 8 trials.

Prior to increasing the number of participants to a level that allows for reliable statistical testing, we needed to first evaluate the framework of the behavioural experiment. The efficacy of the experiment was quantified using two methods, firstly we tracked the mean divergence from veridical time across each participant for the order they received trials in, i.e., the mean divergence from the 1st trial of all participants regardless of sound scene or duration etc. Secondly, we investigated the attentional response time for each attention cue the participants responded to in order they were presented. The results from these investigations can be seen in figure 7a and 7b respectively. Figure 7a unsurprisingly shows a large amount of variation in estimations, however there is an apparent trend of diverging further from veridical time as trial number increases suggesting participants become less accurate due to repetition fatigue. On the other hand, figure 7b shows a trend supporting that participants’ average response time didn’t decrease with inter-participant variation decreasing with increased attention cues. This suggests participants maintained a consistent attention level given cue appearance were randomly selected with no means to predict the onset. It is worth noting many attention response times fall below 0s, this is due to errors in pavlovia latencies and oversight in our means of data collection. Key response time was saved and directly compared to the planned cue time onset, however due to the apparent variation in onset of cues we are not perfectly mapping response time. This is a fixable issue in future investigations using this framework by forcing saving of the onset and offset times of attention cues.

Chart, scatter chart

Description automatically generated

*Figure 7*

Investigating behavioural framework though response patterns of participants over time.

(A) – Mean divergence from veridical time as a function of the trial number presented to the participants chronologically.

(B) – Attentional response time for chronological presentation of attention cues 1-8 for all 10 participants

Analysis across all participants provides the average datasets we aim to replicate through CNN modelling, demonstrating the general pattern of human time perception complete with any prevalent content specific biases. These averaged results are shown in figure 8 below, illustrating both mean estimated time for each duration across all participants that gave valid estimations, along with the mean divergence from veridical time for each presented duration. The mean was taken for each duration presented rather than each sound file as multiple sound files were attributed for the same duration, resulting in visualising 29 consecutive data points rather than 52. While it can be argued this could possibly omit content specific biases in estimations, figures 9,10 and 11 are dedicated to representing the comparison between different classes of sound files. Figure 8a shows a highly similar trend to figure 4a, besides the unexpected low duration estimation at 10s, the estimations are near linear up until the 15s period and show minimum variation across each participant. However, as seen for the individual participant, the variance of estimations significantly increases beyond 15 seconds, with increasing underestimations as duration increases. The standard deviation error bars illustrate well the clear pattern of underestimation, with the upper boundary of error bars only surpassing the diagonal on four occasions. Averaged data shows a far more consistently linear pattern of estimations, demonstrated by the linear regression line of best fit (LOBF) in figure 8b. The average divergence from veridical time was - 4.0s, averaged across all participants.

Chart, line chart

Description automatically generated

*Figure 8*

Average human time perception for 52 natural soundscapes (across 15 classes) for 10 participants.

(A) – Mean of duration estimations for each audio true duration

(B) – Mean divergence from veridical time for each audio true duration

Blue lines represent standard deviation error bars. Red line indicates the perfect correlation between estimation and true duration, anything above is an overestimation, anything below is an underestimation.

Sound files were categorised in three different approaches to investigate if patterns of time perception were dependent on audio content present in either: sparsity of salient events, rhythmic quality, and emotional content. Sounds were subjectively labelled by the researcher as either positive/neutral/negative or as rhythmic or non-rhythmic whereas previous use of the natural sounds saliency database had supported the classification of sounds as either sparse or dense for salient events (Huang and Elhilali, 2017). Figure 9 groups together both rhythmic content and emotional content separated results as these both lack any distinguishable differences from the separated sound classes. Linear regression functions were fitted to each class within the categories however these clearly demonstrate insignificant differences and thus are not investigated further. Separations of sounds into sparse vs dense for the content of salient events within a sound file is demonstrated below in figure 10, as we further expand the analysis by computing the mean deviation from veridical time relative to the mean. In figure 10a, it is only marginally noticeable there is a difference in trends between the two categories, interestingly the regression fitted LOBFs show highly similar gradients (0.71 vs 0.74 for sparse and dense/salient respectively), however, there is a clear downshift in the sparse LOBF, showing a greater extent of underestimations. This effect is further seen in figure 10b, with sparse sound files showing a mean deviation % of -33.75 vs -21.34 for dense/salient sound files. We are confident this result does not come from a difference in duration distribution of the two classes simply representing the increase in underestimation with increased time due to model design ensuring an even distribution of sparse and dense in each duration band (11 and 4 in each band respectively). Further if this was the case, we would expect a more significant difference between gradients of LOBFs.

Chart, line chart, scatter chart

Description automatically generated

*Figure 9*

Observations of estimation pattern results from sound files subjectively classed into (A) ‘Rhythmic (*purple*) vs Non-rhythmic (*green*)’ and (B) ‘Positive (*green*) vs Neutral (*yellow*) vs Negative (*red*)’

Dashed lines indicated calculated linear regression lines of best fit

Chart, scatter chart

Description automatically generated

*Figure 10*

Variation in duration estimation patterns comparing sparse and dense salient event containing sound files, categorised by (Huang and Elhilali, 2017).

(A) – Estimated time vs Real duration separating sparse sound files (*green*) and dense sound files (*pink*). Dashed lines of equivalent colour represent linear regression lines of best fit.

(B) – Mean deviation as a percentage relative to the mean estimate of the category type. Bar graphs show mean deviation of all data points, overlayed against deviation of mean estimates from each sound file (16 sparse vs 36 Dense).

Newly trained CNN models favour Mel-Spectrograms over Cochleagrams and Basilar Membrane Displacement as audio image representation classifiers – 4.2

Three newly made CNN models, replicating model architecture from (Salamon and Bello, 2017) with details outlined in section 3.5, were investigated for training/testing accuracy, training/testing loss, cross validation accuracy and computational run times. Figure 11 below shows training history of each model type using training data from the initial 75s of each unedited sound file. These initial accuracy results show a clear advantage for the Mel Spectrogram input specific model, in terms of maximum validation accuracy reached for both train and test data and in computational efficiency for training, requiring half as many epochs as the cochleagram specific model and a third as many as the BM model needed. The only metric that the Mel spectrogram model was outperformed in was final test and train loss score, which the Cochleagram showed a minor improvement in. This indicates a possibility that the Cochleagram specific model begins to overfit for the training data, shown by the training loss continuing to decrease past the test loss. Although it is common for training loss to be lower than test loss, and vice versa for classification accuracy, this outcome is the effect of two dropout functions applied to both fully connected layers in the model at a high rate of 0.5 for each.

For further validation of model performance, cross validation was performed on the following 50s of audio of all sound files bar ‘Sailing Barge’ and ‘Car Chase’ as these durations were too short and only data unseen to the model during training can be used for cross validation. To evaluate this, each model was made to predict each shuffled 1s temporal bin for 50 samples each from the presented 13 classes. As model predictions were outputted as probabilities of each 15 possible classes, the node with maximum activation (highest probability) was taken as the prediction label for one hot classification. From these results we generated confusion matrices for each model type, shown in figure 12. Overall classification accuracy was calculated from confusion matrices, shown below in bar graph figure 13. This clearly shows that Mel spectrogram specific models maintained superiority in classification accuracy when tested on unseen data, with a 20% better classification accuracy than the next best model of the Cochleagram specific model. Interestingly, from the confusion matrices it becomes apparent that different models respond differently to each sound. Mentioned in the methods section 3.5, we identified the Sports sound file as a candidate to observe if the models can generalise due to previously unseen features. Interestingly the Cochleagram model performs best at 90% accuracy (vs 82% and 62% for Mel spectrogram and Basilar membrane model respectively), suggesting our earlier hypothesis that the cochleagram overfits is not the case. Yet however we do still see a classification accuracy drop from ~90% to 70% for the cochleagram model. Another surprising factor in the confusion matrices can be seen in sounds DogPark and Carnival, which do contain highly distinct features such as dog barking and whistles for DogPark and cheers and fairground ride sounds in carnival, however these features are relatively distributed within the model and will not occur in every 1 second bin. Yet still, the Mel-spectrogram model is able to classify both of these classes with 100% accuracy, compared to relatively average performances by the other models.

Graphical user interface, chart

Description automatically generated

*Figure 11*

Training history results of each input specific model showing classification accuracy and loss calculations at each epoch for training (*purple*) and test (*gold*) data sets. Test data was generated from 20% of the shuffled 1s long inputs from the 75s sample of 15 sound files. Batch size was 64 so that the majority of class types were inputted at each batch multiple times to ensure sudden overfitting to a class type didn’t occur. Epoch number was decided through trial and error of training until a plateau of validation loss was reached, as continuing epochs after this point can lead to overfitting of the training data. Loss function was calculated via categorical cross entropy and optimised with the Adam optimiser (Kingma and Ba, 2015).

(A-B) Basilar membrane model classification accuracy and loss function over 150 epochs

(C-D) Mel spectrogram model classification accuracy and loss function over 50 epochs

(E-F) Cochleagram model classification accuracy and loss function over 100 epochs

\*Unfortunately, these graphs do not display a uniform Y axis, as training history data was eventually lost, it was not possible to rescale these graphs without retraining the models.

Chart, scatter chart

Description automatically generated

*Figure 12*

Confusion matrices for one hot prediction output of each input specific model. All confusion matrices are mapped to the same heatmap scale, where darker purple indicates more occurrences of predictions made against the true label sound classes. As only 50 inputs were added to the model, confusion matrix results are doubled to represent classifications as percentages. A fully perfect classifier will only show active (>0) cells along the diagonal where true label matches predicted label. Rows along the matrix showing the number (x2) of true label inputs are expected to be empty for CarChase and SailingBarge as no sound classes with these labels were entered to the model. The sum of each cell for a whole row equals 100 in every other condition.

(A) – Basilar membrane input specific model. (B) – Cochleagram confusion matrix. (C) – Mel Spectrogram confusion matrix

Chart, bar chart

Description automatically generated

*Figure 13*

Cross validation accuracy results calculated from confusion matrix as the mean of each cell along the diagonal (correct predictions).

Only the Basilar Membrane displacement input specific model shows capabilities to predict duration estimates, with significant capability for improvement – 4.3

Figure 14 demonstrates an example of the dynamic threshold application with accumulating for the Blacksmiths 12s sound file. Results for the Mel spectrogram model and the BM model are as expected, threshold levels only detect significantly salient changes for successive activation states, which result in a relatively linear increase of accumulated time units for this sound file. However, the results taken from the cochleagram model clearly show a threshold level that is too low, effectively always accumulating time steps as almost all increase in Euclidean distance results in accumulating a time unit. This is a concerning result given that threshold levels were calculated via average Euclidean data over all sound files, suggesting high variance of Euclidean distance for each sound file.

Chart

Description automatically generated

*Figure 14*

Dynamic threshold applied to Euclidean distance data for successive activation states of the third convolutional layer of each model architecture. Euclidean distance surpassing threshold levels results in the accumulation of a time step, shown increasing over the successive inputs with red dashed vertical bars correlating to the time at which a time unit was accumulated.

(A, B) – Basilar membrane model Euclidean distance and resulting time step accumulation

(C, D) – Mel Spectrogram model Euclidean distance and resulting time step accumulation

(E, F) – Cochleagram model Euclidean distance and resulting time step accumulation

Analysing the accumulated time unit’s vs true duration of each sound file shows the general trend collected by the dynamic threshold. The basilar membrane model (15a) shows the strongest trend, even demonstrating a stronger linear trend and lesser variation for sound files <15s in duration as seen in human duration estimations. The cochleagram model (15b) on the other hand shows what appears to be significantly minimal correlation between accumulated time units and true duration, a trend that we should expect to see if this model could accurately predict time perception. The Mel spectrogram model (15c) shows an interesting trend, with evident increase in accumulation with increasing duration but at a very minor rate.

The final predictions generated by each SVR model are shown in figure 16. It is immediately clear that neither the cochleagram or Mel spectrogram have been able to predict duration estimations, showing no trend with increasing true duration and all presenting incredibly low duration estimates. On the other hand, although the basilar membrane model is clearly not perfect by not replicating the trend shown for human participants in figure 8a, there is still an improved trend in comparison to the other two models and indicating potential for improvement.

Chart, scatter chart

Description automatically generated

*Figure 15*

Accumulated time unit’s vs true durations for each of the 60 sound files present. As no attentional cues were present in model inputs there was no need to exclude these from final analysis.

(A) - Basilar membrane specific model.

(B) – Cochleagram specific model

(C) – Mel spectrogram specific model

Chart, line chart

Description automatically generated

*Figure 16*

Predictions of duration in seconds calculated by the SVR models trained on each individual dataset, plotted against the true duration of each sound file

(A) – Basilar membrane specific model.

(B) – Cochleagram specific model

(C) – Mel spectrogram specific model

**5.0 – Discussion**

Auditory classification as a means to generate time perception – 5.1

From the offset, modelling human auditory perception using image representations of audio is a flawed approach. Although the strong correlation between CNN image classification and the visual processing cortical pathway is well documented (Kriegeskorte, 2015). The limited understanding, we have of the auditory cortical processing stream prevents any such approach, in isolation, from concluding accurate modelling of the neural mechanisms involved. However, this strong correlation between visual processing methods and CNN processing strongly supports the parent paper (Roseboom et al, 2019) that we are extending as strong evidence for time perception mechanism present within perceptual processing. Thus, in collaboration with (Roseboom et al, 2019), a successful modelling approach of auditory time perception of faithfully replicated human reports would strongly support the generalised theory that time perception can be processed independently of dedicated systems such as the internal clock (Treisman, 1963), (Church, 1984). This can be viewed as an Occom’s razor argument, if perceptual content can be extracted from a highly simplistic system such as the classification models we generated, then there should be no need for dedicated systems. A dedicated time perception system within the brain would require complex inter-modality interactions as opposed to a global embedded process.

The issues of modelling auditory perception were continuous throughout this project. Most significantly were the issues of temporal bins as successive inputs rather than continuous input. This project could not have been justified if classification of natural auditory scenes failed, as this would indicate we are not modelling auditory processing at all and only observing random dynamics of networks rather than dedicated patterns of activity representing audio. This was the initial justification of choosing second long inputs as training data for our models, we wanted to ensure that each input contained enough distinct features that could be identified by the models to undergo classification. However, as we quickly observed when analysing Euclidean distance of second long inputs only 12 data points, the dynamic threshold would have been minimally impactful. With only 12 time steps for a single sound file (in the case of Blacksmiths (12s), this is therefore the maximum number of accumulations if the threshold was constantly active (and thus not detecting salient events). Further, regression of a smaller number of accumulated time steps will be significantly negatively impactful for extracting time perception predictions, which was made very clear from the results of Mel spectrogram support vector regression outputs (figure 15C and 16C). Temporal bins must be short enough that they are able to accurately indicate when a salient event occurs and must have a high enough number of total bins to identify all significant salient events. All of these issues present were only increased by classifying natural environmental sound scenes as opposed to music or speech simply due to the lack of distinct features to classify. This can be seen by comparison of our cross validation accuracy results shown in figure 13 when compared to the study investigating spectrograms vs cochleagrams for speech recognition of 10 classes of numbers (Van Meer and Buermann, n.d.). Even though the classification model architecture comes directly from this paper, they demonstrated cochleagrams as the superior classification model type. The high dropout rate used in this architecture did significantly benefit our classification problem though prevention of over generalisation. Given that very high classification accuracy percentage did not seem to be a significant factor in generating the more accurate duration predictions, this approach for large generalisation at the cost of classification accuracy is certainty one that we wish to include to further iterations of these models.

Proposed model and experimental framework redesign – 5.2

For future extensions of this model, we believe in this project that clear paths for improvement have been generated. Given the significant improvement in general duration estimations of the Basilar membrane model compared to both other models, this is definitely where future attention should be focused. This is a very promising direction given the significantly more biologically plausible representation of audio that the basilar membrane input generates as this is the lowest level of input entered to the auditory processing pathway. However, we have not given up faith with the Mel spectrogram model as showing capabilities to model time perception. The general increasing trend of accumulated time units with increasing true duration seen in figure 15C suggests that with improved regression classification for predicting duration is possible. This is likely achievable via increasing sample size or adjusting the dynamic threshold variables so that more time units are accumulated would significantly improve time perception beyond consistent failures in prediction. We also believe the lack of sample size of presented sound files prevented improved estimation predictions for the BM model, limiting the discernment capabilities of the SVR model evidenced by the relatively strong trend between accumulated time units and increasing true duration of sound files shown in figure 15A. Although we believe the current data set, we could have easily shown an improved duration prediction pattern for the Mel Spectrogram model via data up scaling so that there are greater differences between data points though easier classification, this goes against our aim of biologically modelling the neural mechanisms. The significantly larger variance in Euclidean distance distributions across different sound files for the cochleagram input model is the main factor for not focusing on this in future extensions. This will be a significantly harder issue to fix though standardising Euclidean distances, which is effectively downscaling the classification capabilities of the model. This will likely result in decrease abilities of feature detection within each sound file resulting in highly variable duration estimations.

We further want to focus on the basilar membrane model in future applications as in hindsight we feel this input type was most negatively hindered by our model design. By visual comparison between figure 1C and 2A, showing the basilar membrane displacement image representation of audio (albeit of different sound files) pre and post (respectively) down-scaling the image shows that there is a significant amount of information loss. 100x100 pixels was chosen as a uniform input size so that model architecture was consistent clearly benefited the Mel spectrogram most which would have undergone lesser information loss. Decreasing the information loss of images before model input can be performed by increasing input dimensions and decreasing the temporal bin size of inputs. However, both of these will come at significant increases in computational complexity for calculating Euclidean distances within (larger) layers and training respectively. To combat this, we suggest importing images into the models in grey scale format rather than colour, reducing the size of the input dimensions 3 fold. If this adjustment were to be done and still compare against other audio image representations, to evaluate each input type on an even playing field I would suggest quantifying and keeping consisten the extent of information loss that occurs in the image pre and post downscaling though an information complexity quantifier such as LZ complexity (Zhang, Wei, Di Maria and Liu, 2016). This isn’t a suggesting that this will significantly improve classification accuracy as this is not what we are interested in, but through a balanced ability to detect features within the audio, detection of time perception dependent salient features will be more evenly identified.

To further improve this report and generate supporting evidence of (Roseboom et al, 2019), improvements to the behavioural framework can also be undertaken. As we demonstrated evidence of a perceptual content specific contribution to time perception estimation in sparse vs dense salient events within audio, we can further extend this investigation by observing other metrics of saliency such as stationarity of the sounds. Although we have only demonstrated a very small sample size of participants in our behavioural results, we believe the effect observed comparing the sparse and dense audio scenes was perceptually relevant due to the comparison between figure 10B and supplementary figure 1. Supplementary figure 1 shows the results of sparse vs dense sound files prior to participant exclusion as a result of failing the attentional tasks, in which results showing opposite trends with significant outliers. The removal of 2 out of 12 participants flipped this trend, suggesting the participates not accurately completing the attentional component of the experiment were less consciously perceptive of the audio, thus resulting in estimations that didn’t represent this perceptual content. However, it can also be argued that the removal of only 2 participants changing the data trends so significantly shows the lack of statistical power with our current experimental paradigm and sample size. With the rest of the behavioural experiment, we were happy with results indicating that participants continued the attentional response task at a consistent level showing that attention was kept throughout all 60 trials. However, the trend we observe with duration estimation accuracy decreasing as trial number increase suggests that participants care less about accurately estimating the duration. We suggest this is easily fixable with adding an optional break within the 60 trials and rewarding participants with a simple ‘well done’ to maintain willingness to accurately estimate duration.

Although currently out of the scope of this project, an end goal would be to demonstrate a similar framework using model architectures inspired by (Kell et al., 2018). This CNN model architecture for speech and music classification is currently the most biologically plausible approach to modelling the auditory pathway supporting fMRI results of divergence pathways for both speech and music. However due to the omission of testing natural sound files in this architecture, this possible extension could further explain cortical organisation.

**Bibliography**

Allman, M., Teki, S., Griffiths, T. and Meck, W., 2014. Properties of the Internal Clock: First- and Second-Order Principles of Subjective Time. *Annual Review of Psychology*, 65(1), pp.743-771.

Arnold, D., Tear, M., Schindel, R. and Roseboom, W., 2010. Audio-Visual Speech Cue Combination. *PLoS ONE*, 5(4), p.e10217.

Baby, D., Van Den Broucke, A. and Verhulst, S., 2021. A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications. *Nature Machine Intelligence*, 3(2), pp.134-143.

Block, R., 1978. Remembered duration: Effects of event and sequence complexity. *Memory & Cognition*, 6(3), pp.320-326.

Block, R., Hancock, P. and Zakay, D., 2010. How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica*, 134(3), pp.330-343.

Burnett, S., 2011. Perceptual Worlds and Sensory Ecology. *Nature Education Knowledge*, 3.

Canny, J., 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6), pp.679-698.

Church, R. and Deluty, M., 1977. Bisection of temporal intervals. *Journal of Experimental Psychology: Animal Behavior Processes*, 3(3), pp.216-228.

CHURCH, R., 1984. Properties of the Internal Clock. *Annals of the New York Academy of Sciences*, 423(1 Timing and Ti), pp.566-582.

Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), pp.181-204.

Cox, D. and Dean, T., 2014. Neural Networks and Neuroscience-Inspired Computer Vision. *Current Biology*, 24(18), pp.R921-R929.

Eghbali, H. and Hajihosseini, N., 2019. Deep Convolutional Neural Network (CNN) for Large-Scale Images Classification. *SSRN Electronic Journal*,.

Eisen-Enosh, A., Farah, N., Burgansky-Eliash, Z., Polat, U. and Mandel, Y., 2017. Evaluation of Critical Flicker-Fusion Frequency Measurement Methods for the Investigation of Visual Temporal Resolution. *Scientific Reports*, 7(1).

Friston, K., 2005. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), pp.815-836.

Gibbon, J., 1977. Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, 84(3), pp.279-325.

Healy, K., McNally, L., Ruxton, G., Cooper, N. and Jackson, A., 2013. Metabolic rate and body size are linked with perception of temporal information. *Animal Behaviour*, 86(4), pp.685-696.

Herbst, S., Javadi, A. and Busch, N., 2012. How long depends on how fast - perceived flicker frequencies dilate subjective duration. *Journal of Vision*, 12(9), pp.141-141.

Huang, N. and Elhilali, M., 2017. Auditory salience using natural soundscapes. *The Journal of the Acoustical Society of America*, 141(3), pp.2163-2176.

Joseph, S., 2015. Jos J. Eggermont. Auditory Temporal Processing and Its Disorders. *Perception*, 44(12), pp.1431-1433.

Kell, A., Yamins, D., Shook, E., Norman-Haignere, S. and McDermott, J., 2018. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3), pp.630-644.e16.

Kingma, D. and Ba, J., 2015. Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations*.

Kriegeskorte, N., 2015. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1(1), pp.417-446.

Lejeune, H. and Wearden, J., 2009. Vierordt'sThe Experimental Study of the Time Sense (1868) and its legacy. *European Journal of Cognitive Psychology*, 21(6), pp.941-960.

Linares, D. and Gorea, A., 2015. Temporal frequency of events rather than speed dilates perceived duration of moving objects. *Scientific Reports*, 5(1).

Peirce, J., Gray, J., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E. and Lindeløv, J., 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), pp.195-203.

Rauschecker, J. and Scott, S., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, 12(6), pp.718-724.

Roseboom, W., Fountas, Z., Nikiforou, K., Bhowmik, D., Shanahan, M. and Seth, A., 2019. Activity in perceptual classification networks as a basis for human subjective time perception. *Nature Communications*, 10(1).

Salamon, J. and Bello, J., 2017. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters*, 24(3), pp.279-283.

Shuvaev, S., Giaffar, H. and Koulakov, A., 2017. Representations of Sound in Deep Learning of Audio Features from Music.

Staddon, J. and Higa, J., 1999. TIME AND MEMORY: TOWARDS A PACEMAKER-FREE THEORY OF INTERVAL TIMING. *Journal of the Experimental Analysis of Behavior*, 71(2), pp.215-251.

Treisman, M., 1963. Temporal discrimination and the indifference interval: Implications for a model of the "internal clock". *Psychological Monographs: General and Applied*, 77(13), pp.1-31.

Urban Sound Datasets. n.d. *UrbanSound8K*. [online] Available at: <https://urbansounddataset.weebly.com/urbansound8k.html> [Accessed 28 August 2021].

Van Meer, T. and Buermann, M., n.d. Speech recognition using very deep neural networks: Spectrograms vs Cochleagrams.

VON BÉKÉSY, G., 1970. Travelling Waves as Frequency Analysers in the Cochlea. *Nature*, 225(5239), pp.1207-1209.

Yu, R. and Chen, L., 2015. The need to control for regression to the mean in social psychology studies. *Frontiers in Psychology*, 5.

Zhang, Y., Wei, S., Di Maria, C. and Liu, C., 2016. Using Lempel–Ziv Complexity to Assess ECG Signal Quality. *Journal of Medical and Biological Engineering*, 36(5), pp.625-634.

Zimmermann, E. and Cicchini, G., 2020. Temporal Context affects interval timing at the perceptual level. *Scientific Reports*, 10(1).

**Supplementary Materials**

Chart, line chart

Description automatically generated

*Supplementary figure 1*

a

A picture containing text, antenna, tool

Description automatically generateda

*Supplementary figure 2*

Failed attempt at calculating Euclidean distance due to 0 padded 0.1s temporal bins not localised in the same position of the input dimensions