Investigating Statistical Models of Music Perception

Guilhem Marion
Supervised by Shihab A. Shamma and Giovanni M. Di Liberto
Laboratoire des Systèmes Perceptifs, ENS (18/02 - 31/07)

Abstract

The neural mechanisms underlying music perception remain poorly understood, this project aims to investigate statistical models of musical surprise and their roles in music perception. In this study, we re-implemented a variant of IDyOM in Python that has been shown more accurate than the original Lisp implementation using both cross-validated likelihood and prediction of EEG recordings of 20 participants passively listening to Bach music. Thinking that the brain behaves similarly as a statistical model of music is part of the statistical learning theory which has been already investigated in the literature. Showing the physiological consistency of such models validates the first hypothesis of this theory (probabilistic prediction hypothesis). The other fundamental hypothesis (statistical learning hypothesis) has been investigated in this study by highlighting enculturation changes after passive exposure to unfamiliar music. The accuracy of EEG decoding of surprise seems not to be related to musical training, we hypothesized that surprise decoding accuracy is driven by audiation capability. We used a pilot experiment to show that mTRF can be used to decode envelope from imagined music and be used as a measure audiation capability. Finally, we presented a new framework for comparing statistical models of surprise that presents several advantages over the currently used methods (negative log-likelihood, behavioral experiments) and can be generalized to any kind of musical modeling, including generative models.

Contents

1 Introduction

Music is, first of all, a social and psychological phenomenon, a shared meaning of sounds. The famous musicologist Jean-Jacques Nattiez [1] takes from Jean Molino the theory of tripartition to explain that the meaning of a musical piece can be given either by the genesis process of the piece, what he calls *poëtic semiosis*, or by the reception process of the listener, what he calls *esthésic semiosis*. According to him, the actual piece in its physical conditions (the score, or even a recording) is empty of sense and does not carry any meaning; only those processes do. So, to understand a piece, one should focus on one of these processes, and the main goal of the musicologist is to understand all of them to, finally, understand the piece.

According to this theory, working only on the actual piece can only lead to descriptive work. This is why, while computational musicology and music generation take more care of the poëtic aspects, this project aims to investigate the esthésic semiosis by studying its physiological correlates: how our brain is processing music to make us feel it?

Recent studies from Marcus Pearce [2] suggested that models based on variable-order Markov chains can reliably mimic the brain processing of melodic expectations, such models are used to predict a surprise signal over time (inversely related to the likelihood) computed note after note from the past. A recent study from Giovanni Di Liberto [3] put forward that the human brain generates a similar signal, implying that stronger cortical responses are generated for more surprising notes.

Because of our clear understanding of their mechanisms, statistical models of surprise give us a great opportunity to test precise hypotheses on the way we perceive music. However, statistical models of surprise have been mostly used by the community of neurosciences of music on indirect physiological indices based on both subjective judgments and behavioral indices [4] [5] [2]. Until now, most of the published papers using direct electrical recordings of the human brain used Event-Related Potential (ERP) from ecologically valid, but non-rhythmic stimuli using only pitches of the same duration [6] [7]. The only study using real music stimuli to correlate neural recordings with surprise signals is under review [3].

Nonetheless, these already-made physiological confirmations of these models already corroborated a first hypothesis: our brain is predicting the next note based on some statistical model while listening to music and one is 'surprised' if the actual next note is far from its expectation. But what about the other notes in the future, do we only have predictions on the next note, or also on the following ones? Do our predictions of notes far in the future affect our expectation of the next one? Music can contain structures at multiple time-scales that may allow for predictions of notes on much longer intervals.

In this thesis, we first aim to implement a statistical model of surprise based on Markov chains in Python. This was previously done with the Lisp programming language [2]. However, the poor accessibility of the Lisp language and the small documentation of the project made it hard to reach and modify, what justify a new implementation. Our goal is to produce a version that is more targeted to the comparison of the computational model with neurophysiological data and to provide us with more modification flexibility in order to embbed physiological hypotheses in the model. In the first project of this report, we investigate whether and how longer-term expectations affect the cortical responses to music. We propose a new model for melodic expectation based on variable-order Markov chains investigating this hypothesis. Then, we use Electroencephalography (EEG) data and the Multivariate Temporal Response Function (mTRF) technique [8] to assess the physiological validity of our two models and compare them to the current Lisp implementation. The presented framework rests on the statistical learning theory [5] who asks several questions related, for example, to the underlying cognitive mechanisms of melodic expectation and enculturation mechanisms. This why, after explaining these ideas in the thesis, we will present some extended pilot projects investigating these questions using our new models.

2 Background

The neural mechanisms underlying music perception remain poorly understood. The literature provides us with two main hypotheses building the theoretical basis of studies using statistical model of surprise: the statistical learning hypothesis and the probabilistic prediction hypothesis [5]. On the one hand, the statistical learning hypothesis says that music enculturation is a process based on an inner model of musical grammar learned over implicit statistical learning. This model can catch structural regularities present in a given musical style the person was exposed to, this model can behave differently with respect to time scales (this idea will be decisive in the design of the model):

short time scale such as a single piece of music (modeled by the short-term model);

long time scale such as an entire lifetime of music listening (modeled by the long-term model).

On the other hand, according to the probabilistic prediction hypothesis, enculturated listeners, in order to organize mental representation of music and to generate culturally appropriate responses, apply their previously trained inner model to generate probabilistic predictions that allow grammatical processing of the piece. In the models we are interested in, the probabilistic predictions are defined as an estimate of the likelihood of notes modeled by an ideal brain of a typical listener. This typical listener is supposed to have only listened to a specific musical corpus (the training set).

Using this paradigm gives the possibility to simulate brain responses to stimuli as the listener was enculturated with a specific kind of music, allowing a new kind of experiments related, for instance, to the effect of musical exposure, implicit learning and cultural distances. This paradigm is so general that is can also apply to different domains such as language [9] allowing, for example, experiments on fundamental differences that language and music can show: the effect of repetition may be one of them.

2.1 IDyOM Model

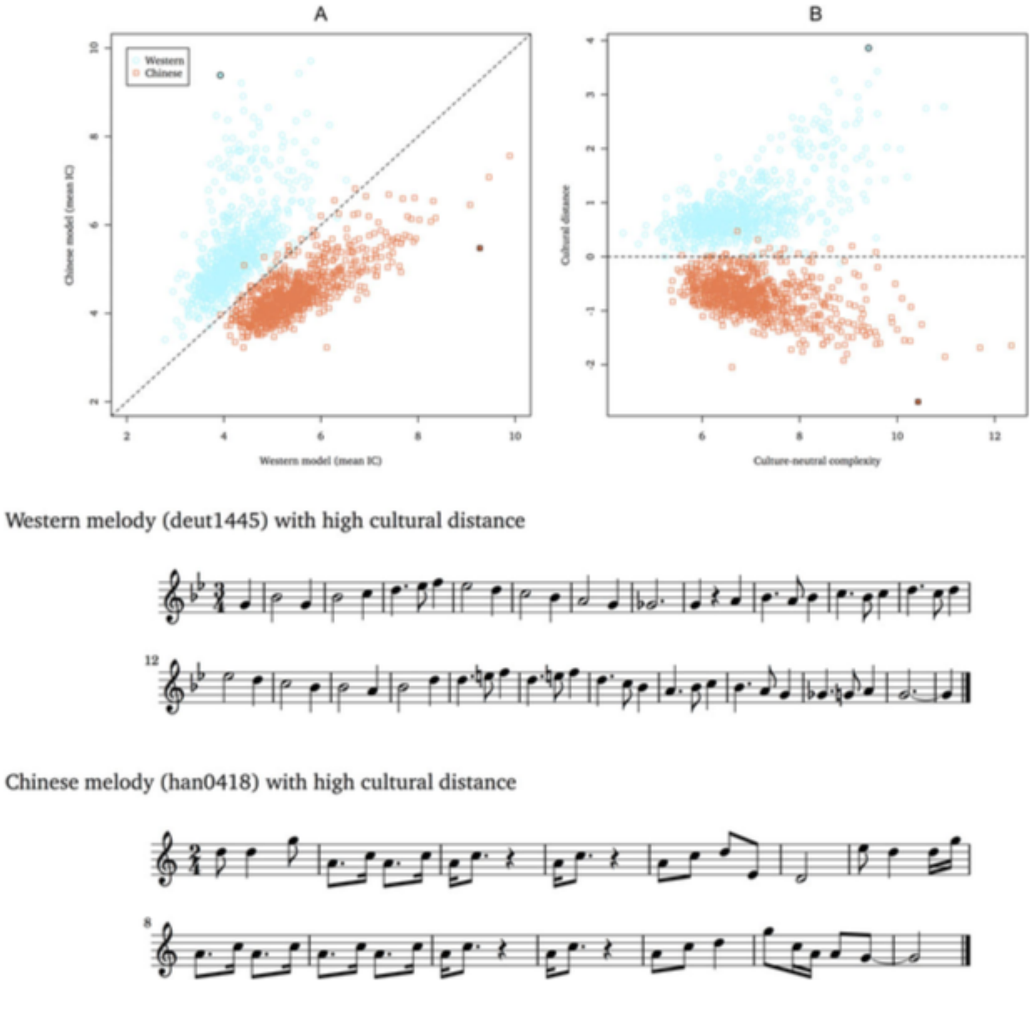
Information Dynamics Of Music (IDyOM) is a statistical model of musical surprise created by Marcus Pearce and published in 2005 in his thesis [2]. It is capable of predicting surprise at the note level, after a training based on variable-order Markov chains. It has provided useful insights to and has been largely used by the community of psychology and neurosciences of music (cited by 248 studies) making it the most used statistical model of surprise. In this section we will present different studies related to the physiological consistency of the model, giving opportunities to a wide range of cognitive studies.

2.2 Surprisal is a Good Tracker for Enculturation

Demorest and Morrison capture the effects of statistical learning in their cultural distance hypothesis: “the degree to which music of any two cultures differ in the statistical patterns of pitch and rhythm will predict how well a person from one of the cultures can process the music of the other.” [10]. In order to catch cultural differences between listener, and therefore to be a good model for enculturation, IDyOM should show good ability to predict cultural distances. This can be verified by predicting how well a person from on culture can process music from other culture. This is what Marcus Pearce [5] showed to prove that IDyOM is capable of simulating enculturation effects through statistical learning.

He trained two instances of IDyOM on two different datasets: one on a corpus of Western folk songs, supposed to model a typical Western listener and one on a corpus of Chinese folk songs, supposed to model a typical Chinese listener. He used both models to predict average surprise (prediction accuracy) both within and between cultures. The Western corpus consists of 769 German folk songs from the Essen Folk Song Collection (data sets *fink* and *erk*). The Chinese corpus consists of 858 Chinese folk songs from the Essen Folk Song Collection (data sets *han* and *natmin*).

A low average surprise means that the data was well predicted by the model, a high average surprise means that the data was badly predicted by the model. We can so have an idea of the accuracy of the model depending on the data. If IDyOM is a good model for enculturation, the Western model will be good for Western music but not for Chinese music, and the Chinese model will be good for Chinese music and not for Western music. Plus, it would be used to classify an unknown piece of music within these two classes.



Western melody (deut1445) with high cultural distance

Chinese melody (han0418) with high cultural distance

Figure 1: Simulating cultural distance between Western and Chinese listeners. (A) average surprise for all data (Western melodies in blue and Chinese melodies in red), (B) cultural distance over a neutral axis. The dash lines show the equality line ($x = y$) meaning that pieces on that line are equally predicted by the two models and cannot be distinguished by the model. (adapted from [5])

A classification framework based on this technique has been shown to correctly classify 98.52% of the pieces. This shows that IDyOM may be a good model for enculturation (at least for these Western and Chinese folk melodies) and give hope for more complex experiments using the enculturation modeling property of IDyOM.

2.3 Validation by Musicological Analysis

Different techniques have been used to show the physiological consistency of statistical models of surprise. One of them [5] consists in showing the coherence between a musicological analysis of pitch structure and pitch surprisal over time. The author took an analysis of Schubert's *Octet for Strings and Winds* made by Learnard Meyer in *Explaining Music (1973)* [11] and showed that when they found comments related to the notion of surprisal, the same idea was corroborated by IDyOM. Here is an excerpt:

> «When the theme returns in bars 21– 22, Meyer writes that "The triadic implications of the motive are satisfactorily realized ... But instead of the probable G, A follows as part of the dominant of D minor

(V/II)." IDyOM reflects this analysis, estimating a lower probability for the A5 that actually follows than for the G5 that is, again, anticipated (0.013 versus 0.186). [5](p.4) »

There is a number of examples of adequation of IDyOM analyses and the ones from Meyer. Showing that IDyOM, only trained on a corpora of Western music, can generate analyses close to the ones made 20 years ago by a famous musicologist support the idea that IDyOM can be a realiable model of musical enculturation. However, this study is purely qualitative and music analysis is higly subjective. In order to have a better evidence that IDyOM may behave in a similar way than the perception of a Western enculturated person, it would be interesting to replicate this experiment with several subjects.

2.4 Validation by Behavioral Expectedness

This is what Marcu Pearce did by using already collected data from a behavioral experiment designed by Manzara et al. [4] In this experiment, authors asked 15 participants to place bets on the subsequent pitch given an excerpt of Bach choral. Participants continued to bet until the correct note was bet (20 pitches to choose from), then, they moved to the actual next note of the original melody. Incorrect prediction resulted in the loss of the bet b, while a correct prediction was rewarded by incrementing the capital sum in proportion to the amount of the bet: $S_n = 20 \cdot b \cdot S_{n-1}$

The measure of surprise (here called information content) is derived by taking:

$$\text{IC} = log_2(20) - log_2(S), \text{ where S is the average gain for a given note.}$$

The author trained IDyOM only with pitches in order to match with the experimental design. Then, the surprise signals over notes were compared between IDyOM and the data of the experiment. Fig. 2 shows an example on BWV 379 [2].

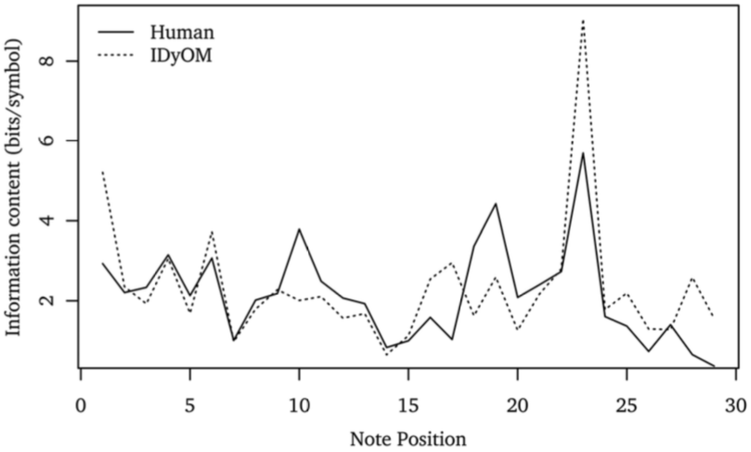


Figure 2: Comparison of surprise signals (information content) between IDyOM simulation and behavioral data from Manzara et al. on BWV 379 Bach choral. (adapted from [2])

IDyOM was able to account for approximately 63% of the variance in the mean uncertainty estimates reported by Manzara et al. This result suggests that human listeners evaluate likelihood of notes in a similar way as IDyOM does. This study was the first (published in Pearce thesis in 2005) to assess an eventual physiological consistency of such models of music.

2.5 Validation by EEG Prediction

Very few studies showed evidence of physiological consistency of IDyOM using recording of neural activity. The first entire study using neural data to prove cortical correlates of a similar process as IDyOM is still under review [3] (but available on bioArxiv). Di Liberto et al. collected 20 participants in Electroencephalography (EEG) and 1 participant in Electrocorticography (ECoG) with electrodes placed in the temporal cortex. Participants were asked to listen to 10 excerpts of Bach music.

IDyOM was trained on a corpus of Western music (excluding the stimuli) in order to simulate the ideal brain of a typical Western participant. This model was used to generate a surprise signal over time for the 10 stimuli, mTRF method was used to predict the EEG data from both the envelope of the stimuli (computed by Hilbert transform) and the surprise signal. The addition of the surprise signal improves, within participants, the prediction accuracy of the EEG data (fig.3).

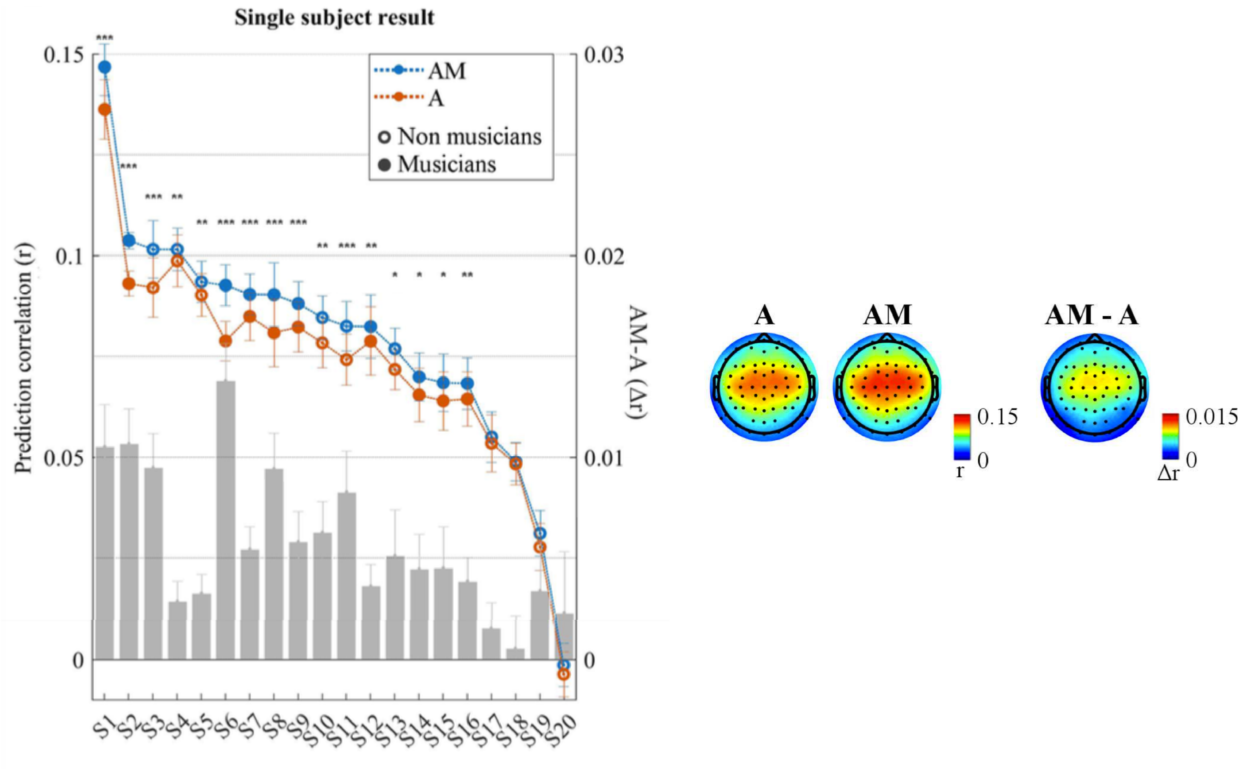


Figure 3: Prediction of unseen EEG data from the envelope of the stimuli (A) and the surprise signal (M) using mTRF method. (left) The enhancement emerged at the individual subject level, the gray bars indicate the improvement. (right) The effect of melodic expectations ($correlation_{AM} - correlation_{A}$) emerged bilaterally on the same scalp areas that showed also envelope tracking. (adapted from [3])

An Event-Related Potential (ERP) analysis has also been made to show that the 20% less surprising notes and the 20% most surprising notes (according to IDyOM prediction) showed significantly different ERPs (fig. 4).

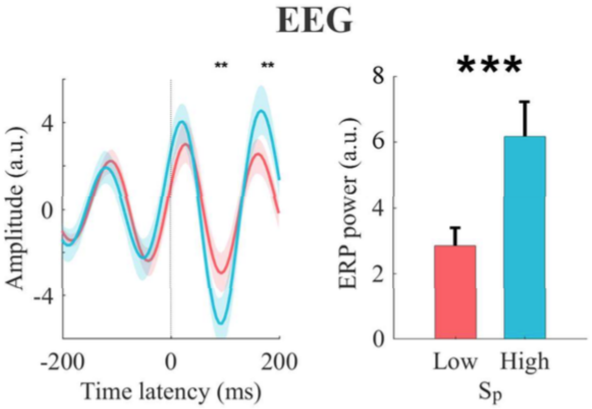


Figure 4: Event-related potentials (ERP) analysis. Shaded areas indicate the 95% confidence interval (across subjects). The right panel shows the total ERP power for the latencies from 0 to 200 ms (* p < 0.001, permutation test). Error-bars indicate the SEM across subjects. (adapted from [3])

This study shows that EEG and ECoG data reflect melodic expectation, and more specifically, that the brain is predicting the next note while listening to music and generate proportionally more powerful brain responses when an expectation is broken. That demonstrates clear evidence of the physiological consistency of the IDyOM model allowing us to imagine more complex EEG studies.

2.6 Electrophysiological Correlates of Melodic Processing in Congenital Amusia

In this study [6] Diana Omigie and colleagues used IDyOM to investigate whether people suffering from congenital amusia[1] process musical expectation in a similar way as others. A known consequence of amusia is bad detection of gross musical violations, however, there has been increasing evidence that amusics show preserved musical ability using implicit methods [13]. Omigie et al. recorded electrophysiological data for both amusics and control participants listening to real melodies. They used IDyOM trained on Western music to model the ideal brain of the participants and generate an ideal surprise signal over time for the stimuli. ERP analysis was used to compare the notes of high surprise and the notes of low surprise to investigate the extent to which the amusic brain differs from that of controls.

The data revealed a novel effect that was highly comparable in both groups: notes with high IC reliably elicited a delayed P2 component relative to notes with low IC, suggesting that amusic individuals, like controls, found these notes more difficult to evaluate, giving rise to proper processing of pitch expectation in amusics.

3 Method

3.1 IDyOM

IDyOM is a statistical tool used to model melodic expectations at the note level, it is based on variable-order Markov chains and composed of two parts: a long-term model (LTM) that is pre-trained on a musical corpus, and a short-term model (STM) which is trained on the current piece and is used to catch repeating structures during a listening.

3.1.1 Long-Term Model

A Markov chain describes a memorylessness process which means that, for $\forall i, X_i$ sequential random variables,

[1] Amusia is a musical disorder that appears mainly as a defect in processing pitch but also encompasses musical memory and recognition affecting 4% of the general population [12].

$$P(X_k = x|X_{k-1}, X_{k-2}, ..., X_0) = P(X_k = x|X_{k-1})$$

We define S as the set of states and a function $P : S^2 \to [0, 1]$ for the probabilities of jumping from state to state. Such a model can be easily shown as a $n * n$ matrix or a graph $G = (V, E)$ where V (vertexes) are the states of the Markovian process, and E (edges) indicate the transitions probabilities:

Our melodic musical grammar is defined on the alphabet Σ that contains all the musical notes with $P : S^2 \to [0, 1]$ the probability to observe a note knowing the previous one. So one can easily construct the graph $G = (V, E)$ with $V = \Sigma$ and $E = P$ such as G is describing our grammar. This model is known as a first-order Markov Chain.

The fig. 6 shows a simple example of a graph representation of a first-order Markov chain for musical purpose. It is carrying the statistical model representing the beginning of the melody of *Au Clair de la Lune* (fig. 5).

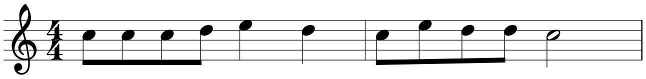


Figure 5: Beginning of *Au Clair de la Lune*.

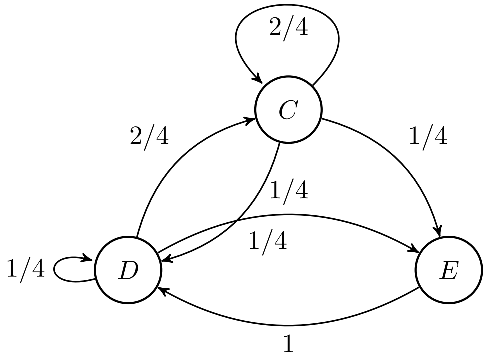


Figure 6: An example of Markov Chain in a graph representation for the melody of *Au Clair de la Lune*. Idiomatic musical gestures are caught, for example, E should always resolve to D and C is the more occurent note.

Because of the highly structured nature of music, it seems not a reasonable assumption to say that the current note depends only on the previous note, musical sentences are constructed over sometimes large numbers of notes and repetitions, for example, show long-term dependencies. By using n-grams, we can still use the Markov model and include long-term dependencies.

An n-gram is an element of the Cartesian product of the original set of states (in our case Σ), 2-gram $\in \Sigma \times \Sigma$, 3-gram $\in \Sigma \times \Sigma \times \Sigma$, ..., and so,

$$n\text{-gram} \in \prod_{k=1}^{n} \Sigma$$

By using n-grams as elements of S the set of states of our Markov Chain, we have the transition probability ω between n-long words, $\forall \omega_k, \forall n$,

$$P(\omega = \omega_k|\omega_{k-1})$$

We can thus get the probability for a given note by,

$$P(Z = z_k|z_{k-1}, ..., z_{k-n}) = \sum_{\{\omega_i\}|\omega_i[0]=z_k} P(\omega = \omega_i|z_{k-1}z_{k-2}...z_{k-n})$$

Variable-order Markov chains have the property to use n-grams of different length and to dynamically adapt the utility of each order. Having a model that can embed n-long temporal dependencies allow us to reasonably use it to model melodic sentences.

As suggested by physiological studies on music perception [5], notes are processed as two independent features: time onset and pitch. Two models are trained on these features, and the final probability is given by their joint likelihood:

$$P(Z = z) = P(Pitch = z_{pitch}) \cdot P(Length = z_{length})$$

Z still follows a probability distribution if $Pitch$ and $Length$ are:

$$\begin{aligned}\sum_{z\in\Sigma} P(Z = z) &= \sum_{i=1}^{n_p}\sum_{j=1}^{n_l} P(Pitch = p_i) \cdot P(Length = l_i)\\ \sum_{z\in\Sigma} P(Z = z) &= \sum_{i=1}^{n_p} P(Pitch = p_i) \cdot \sum_{i=1}^{n_l} P(Length = l_i)\\ \sum_{z\in\Sigma} P(Z = z) &= 1\end{aligned}$$

3.1.2 Training

Transition probabilities are learned over a corpus of musical data. For this purpose, we compute the frequencies over all random variables and use them as probabilities[2].

$$P(Z = z_k|z_{k-1}, ..., z_{k-n}) = \frac{\#z_{k-n}...z_{k-1}z_k}{\#z_{k-n}...z_{k-1}}$$

3.1.3 Merging Information from Different Orders

Once trained, the model provides us with up to n probability distributions (one by order) for a given context. IDyOM uses the following definition of the probability of an event Z depending on the context C:

$$P(Z = z|C) = \alpha(z|C) + \gamma(z|C) \cdot P(Z = z|C_{[1,:]})^3$$

The functions $\alpha()$ and $\gamma()$ are approximated using the Prediction by Partial Matching (PPM) algorithm [2](see [14] for the original method). Note that $P(Z = z|C_{[1,:]})$ refers to the Markov chain of order $n - 1$. By iterating recursively, we visit all orders and assign a weight to every probability distribution. The following method is used:

$$\gamma(z|C) = \frac{t(C)}{\sum_{z'\in\Sigma} \#C \cdot z' + t(C)}, \text{ and,}$$

$$\alpha(z|C) = \frac{\#C \cdot z}{\sum_{z'\in\Sigma} \#C \cdot z' + t(C)}$$

Where $t(z|C)$ denotes the total number of symbol of Σ that have occurred with non-zero frequency in context C. This method allows to acount for the *diversity* of a distribution: a distribution that only saw a couple of n-grams will be less represented than a distribution which saw all the alphabet.

In order to have a numerical definition of surprise for a given note c and a given context C, we use *information content*, usually used in information theory to account for the unpredictability of an event:

$$H(Z = z|C) = -log_2(P(Z = z|C))$$

[2] We use $\#\omega$ as the number of occurrences of the word ω in the corpus.

[3] We note $C_{[x,:]}$ the context C truncated from the x first elements.

3.1.4 Short-Term Model

The short-term model embeds exactly the same computational model than the long-term model, but not trained on a corpus. The short term model is trained during the trial of a given piece, therefore, it only takes into account the very local grammar of the given piece and responds to its structural variations. The distributions of the short-term model and the long-term model are merged using arithmetic mean weighted by the entropies of the models, b is an additional parameters allowing to sharp or smooth the final distribution[4]:

$$P(Z=z) = \frac{E_1^{-b} \cdot P_1(Z=z) + E_2^{-b} \cdot P_2(Z=z))}{E_1^{-b} + E_2^{-b}}$$

Entropy is given by Shannon Entropy as follow:

$$E = -\sum_{z \in \Sigma} P(Z=z) \cdot log_2(P(Z=z))$$

3.2 Evaluation

In order to evaluate implementations and compare models, we need to have an objective measure, we choose cross-validation (k-fold) by mean of likelihood averaged over notes. We call this measure our *theoretical measure*. We divide the data set D into k different subsets D_i. For each subset D_i, we train on $D \setminus D_i$ and evaluate on D_i by the likelihood L, the score is the average over all subsets. The likelihood L for a given subset is the average of likelihoods over all notes:

$$L = \sum_{n \in notes} \frac{P(n|M_i)}{|notes|}$$

With $P(n|M_i)$, the probability returned by the trained model used for the given evaluation.

Using average likelihood over unseen data is based on the idea that a good model should generate a better average likelihood than a worth one. Indeed, notes that appeared should have in mean (because of the law of large numbers) a greater probability than the ones that did not appear; as the probability distribution cannot reach 1 for every event, a more accurate distribution should generate larger likelihoods. Machine Learning studies often use negative log-likelihood to evaluate their models [16], which is basically the same but in the log domain.

3.3 EEG Signal

EEG allows to record non-invasively the brain electrical potentials that reach the surface of the human scalp along with noise from scalp muscles and cerebral motor signals. In the Biosemi 2 system we use, the electrical signal is recorded over 64 electrodes placed on the scalp, the 64 signals are merged to form a matrix of recordings of electrodes over time and represent what we call the EEG signal. There exist other solutions to record brain activity, as fMRI which is based on blood flow and has a way better spatial definition than EEG, but the low temporal resolution of those methods is an issue when studying responses to fast stimuli such as music.

3.4 mTRF

In order to find a given signal into EEG data, an effective method consists in applying a linear regression between a given signal and the EEG one and then compute a correlation between the predicted signal and the target [8]. To this end, we make the assumption that the brain response $r(n,k)$ recorded by an electrode k at index n can be approximated as a linear transformation of the stimulus $s(n)$:

$$\hat{r}(n,k) = \sum_{m=0}^{M} s(n) \cdot w_k(n-m)$$

Where $M = |w_k|$, $N = |s(n)|$ and $M \leq N$

In order to have a definition of the response of the brain, we introduce a residual response $\varepsilon(n,k)$ not predicted by the system such as:

[4] In our implementation we use $b = 1$.

$$r(n,k) = \hat{r}(n,k) + \varepsilon(n,k)$$

We can so write in this form:

$$\forall n, r(n,k) = (s * w_k)(n) + \varepsilon(n,k)$$

By matrix rewriting,

$$R_k = Sw_k + \varepsilon_k$$

With S, Toeplitz matrix of $s(t)$:

$$S = \begin{bmatrix} s(1) & S(N) & \cdots & s(1-N) \\ s(2) & S(1) & \cdots & s(2-N) \\ \vdots & \ddots & \ddots & \vdots \\ s(N) & S(N-1) & \cdots & s(1) \end{bmatrix}$$

And so, stacking all w_k in W,

$$R = SW + \varepsilon$$

The optimization problem is to find the vector W that minimizes this residual response ε, we so define the following optimization problem as an Ordinary Least Squares method over the vector W,

$$\min_W \sum_{k=1}^{K} \sum_{n=1}^{N} [r(n,k) - \hat{r}(n,k)]^2$$

We remark that the minimized error equals the squared residual error,

$$[r(n,k) - \hat{r}(n,k)]^2 = \varepsilon(n,k)^2$$

Thus gives us the following closed formula[5],

$$\hat{W}_k = (S_k^T S_k)^{-1} S_k^T R_k$$

This method is used to predict the EEG signal from a given signal representing the stimuli. The idea is to learn the vector W that applies a linear transformation from $s(t)$ to $r(t,k)$. If this transformation allows to predict $r(t,k)$ with high accuracy it means that the signal $s(t)$ is carrying a lot of information about $r(t,k)$ and so that it is actually being processed in the brain, the process is explained in fig. 7. If the correlation is low, then, one cannot say anything, as the low correlation may be due to the non-linear relation between $s(t)$ and $r(t,k)$ or because $s(t)$ is not carrying relevant information.

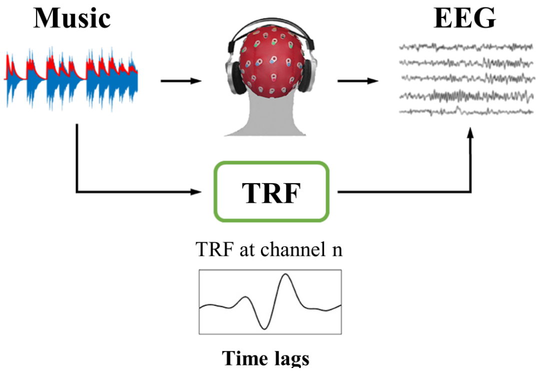


Figure 7: We predict the signal of the EEG from a given representation of the stimuli in order to check its relevance. (adapted from [8])

[5] We note that inverting a matrix can be a hard computational task, so we may want to use another technique like gradient descent in some case.

In order to check hypotheses on the relevance of given musical representations or mechanisms of music processing, we can use different representations of the stimuli as $s(t)$. Recent work showed that envelope and its first derivative allow a correct prediction of the EEG signal [17]. Moreover, studies also suggested that way more information can be present in neural signal [18] and that surprisal can be found in EEG recordings. [3].

3.5 Data

3.5.1 Music Dataset

We used two different datasets which we call:

Training set containing all Bach chorals as midi files made from .kern files written by hands. They are supposed to be rhythmically clean.

Evaluation set containing different Bach concerti:

Bach violin pieces from sonatas and partitas:

- partita num 1 in B minor, allemande
- partita num 2 in D minor, allemande
- partita num 2 in D minor, gigue
- partita num 3 in E major, loure
- partita num 3 in E major, gavotte

Partita for flute in A minor by Bach.

These midi files were gleaned on the internet and were probably recorded with a midi keyboard and are likely to be rythmically imprecise. These files were the stimuli for a previous EEG study whose we used the data.

All surprise signals related to the training set were made using cross-validation. The surprise signals for the evaluation test were generated by training the algorithms with the training set only.

3.5.2 EEG

We used already collected EEG data of people listening to the evaluation test. These data have been recorded for the purpose of a previous study [17].

Twenty healthy subjects (10 male, aged between 23 and 42, M = 29) participated in the EEG experiment. Ten of them were highly trained musicians with a degree in music and at least ten years of experience, while the other participants had no musical background. Each subject reported no history of hearing impairment or neurological disorder, provided written informed consent, and was paid for their participation. The study was undertaken in accordance with the Declaration of Helsinki and was approved by the CERES committee of Paris Descartes University (CERES 2013-11). The experiment was carried out in a single session for each participant. EEG data were recorded from 64 electrode positions, digitized at 512 Hz using a BioSemi Active Two system. Audio stimuli were presented at a sampling rate of 44,100 Hz using Sennheiser HD650 headphones and Presentation software (http://www.neurobs.com). Testing was carried out at Ecole Normale Supérieure, in a dark room, and subjects were instructed to maintain visual fixation on a crosshair centered on the screen and to minimize motor activities while music was presented. EEG signals were digitally filtered between 1 and 8 Hz using a Butterworth zero-phase filter (low- and high-pass filters both with order 2 and implemented with the function filtfilt), and down-sampled to 64 Hz. Results were also reproduced with high-pass filters down to 0.1 Hz and low-pass filters up to 30 Hz. EEG channels with a variance exceeding three times that of the surrounding ones were replaced by an estimate calculated using spherical spline interpolation. All channels were then re-referenced to the average of the two mastoid channels with the goal of maximizing the EEG responses to the auditory stimuli.

4 New Models

4.1 IDyOMpy

IDyOMpy is a variant of IDyOM implemented in Python including the following changes.

4.1.1 Rhythms Quantization

Data gleaned on the internet are often recorded by hand on a midi keyboard. Therefore, rhythm is imprecise, and lead to a Gaussian-like noise added to note duration, for example: [9, 11, 8, 9, 13] in place of [10, 10, 10, 10]. As our algorithm is based on a discrete alphabet, this kind of, apparently small, errors can be catastrophic. Therefore, we chose to over-quantize the vector of note durations in order to avoid that, this technique can be seen as rounding the values to the closest valid value. The quantization parameter can be tuned to any value (possibility to plot cross-validated accuracy over quantization parameter). Fig. 8 shows this plot for the evaluation set.

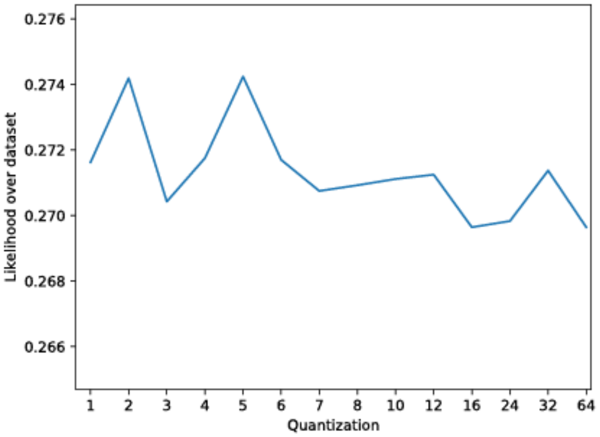


Figure 8: Average likelihood plotted over quantization value for the evaluation set. We used different value for the quantization parameter (avoid rythm imprecision) and computed the likelihood averaged over notes and over pieces.

4.1.2 Data Augmentation

In order to be able to generalize to any keys, we augment the data. For that, we transpose all pitch to 6 semitones up and 5 semitones down in order to reach all 12 tones by minimizing extreme pitches.

4.1.3 Merging Algorithm (smoothing)

Instead of using the PPM algorithm to merge the different orders of the Markov chains, we chose to merge the probability distributions using an arithmetic mean weighted by the entropies of the distributions.

Once trained, the model provides us with up to n probability distributions for a given context, we merge them using arithmetic mean weighted by the entropy of the given generated distribution. We note E_i the entropy of the probability distribution given by the context C_i corresponding to the i-order model.

$$P(Z = z|C) = \frac{\sum_{i=1}^{n} P(Z = z|C_i) \cdot E_i^{-1}}{\sum_{i=1}^{n} E_i^{-1}}$$

Entropy is given by Shannon Entropy as follow:

$$E_i = -\sum_{z \in S_i} P(Z = z|C_i) log_2(P(Z = z|C_i))$$

This method allows better cross-validated predictions over the training set.

4.1.4 Entropy Computation

In our implementation we use entropy to merge the different Markov chains and to merge the the short-term and long-term models. A naive implementation would compute the entropies of the Markov chains in order to get the entropy of the long-term or short-term models. However, because we also used the entropy to merge the Markov chains we already know the entropies of the distributions from wich the long-term and short-term models are drawn.

$$P(Z = z|C) = \frac{\sum_{i=1}^{n} P(Z = z|C_i) \cdot E_i^{-1}}{\sum_{i=1}^{n} E_i^{-1}}$$

Therefore, it was of interest to find a way to compute the entropy of P only from E_i. We showed that using the mean of entropies weighted by themself is a good approximation and allow to drastically reduce the computation time:

$$E = \sum_{i}^{n} E_i \cdot E_i^{-1} / \sum_{i}^{n} E_i^{-1}$$

4.2 JUMP

The large majority of models for music generation and probability distribution estimation relies on predictions of the next note [19]. But, is the brain working this way? Our project aims to investigate if a model based on predictions of notes further than the next one is more physiologically plausible than the current methods. Our model is still based on variable-order Markov chains in order to stay in a controlled theoretical frame. The main idea is that we sometimes have a quite bad expectation (high entropy of the model) for the next note, but for some reasons (repetitions, structure, ...) we may have a very good prediction for one of the following notes. We can think at a musical structure where there is sentence repetition, each sentence ends with a cadence (sub-dominant, dominant, tonic), in this case, the expectation of the end of the sentence can help us predict the next note. We call depth k the size of the jump between the current note and the long-term predicted one.

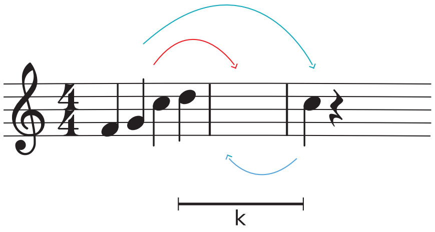


Figure 9: The red line corresponds to the prediction path IDyOM uses, the blue line corresponds to the path JUMP uses.

For a given depth k, and order n,

$$P(X_t = z|X_{t-1}, \cdots, X_{t-n}) = \sum_{x \in \Sigma} P(X_{t+k} = x|X_{t-1}, \cdots X_{t-n}) \cdot P(X_t = z|X_{t+k} = x)$$

The idea is to construct a new IDyOM module having several models with different depth and combine them using arithmetic mean weighted by their entropy.

We so have, for order n and maximum depth K,

$$P(X_t = z|X_{t-1}, \cdot\cdot\cdot, X_{t-n}) = \sum_{k\leq K} E_k \cdot \sum_{x\in\Sigma} P(X_{t+k} = x|X_{t-1}, \cdot\cdot\cdot X_{t-n}) \cdot P(X_t = z|X_{t+k} = x)$$

With E_k, the entropy of the model of depth k.

We hypothesised that the resulting JUMP model would better capture long-term dependencies than IDyOM, thus, allowing us to test whether the low-frequency cortical responses to music reflect such dependencies. We therefore expect JUMP predictions to be more strongly linked to the EEG signal.

5 Results

This section will show the results of the comparison of three models:

IDyOM is the original implementation of IDyOM in Lisp[6] (cf. section 3.1).

IDyOMpy is our Python implementation of IDyOM including some small improvements[7] (cf. section 4.1).

JUMP is a variant of IDyOMpy using conditional probabilities of future events.[7](cf. section 4.2).

In order to compare these models we will use two measures:

Theoretical measure using average likelihood over unseen data based on the idea that a good model should generate a better average likelihood than a worth one.

Physiological measure using the correlation of the predicted EEG. This measure is based on the assumption that a good model should be able to better predict the EEG data as the surprise signal would be more correlated to the brain responses. Because of signal-over-noise ratio, prediction correlations can vary between subjects, this is what we will look at the improvment of the models with respect to the prediction correlation of the envelope only. In case of improvement we will plot the correlations within subjects.

5.1 Theoretical Measure (likelihood)

We computed surprise signals of the training data using cross-validation (k-fold) with $k = 5$. For the evaluation, we trained our model on the training set and computed the surprise signals on the evaluation set.

5.1.1 Training Set (cross-validated)

We use our theoretical measure only on the training data (Bach chorals).

[6] https://code.soundsoftware.ac.uk/projects/idyom-project

[7] https://guimarion.github.io/IDyOM

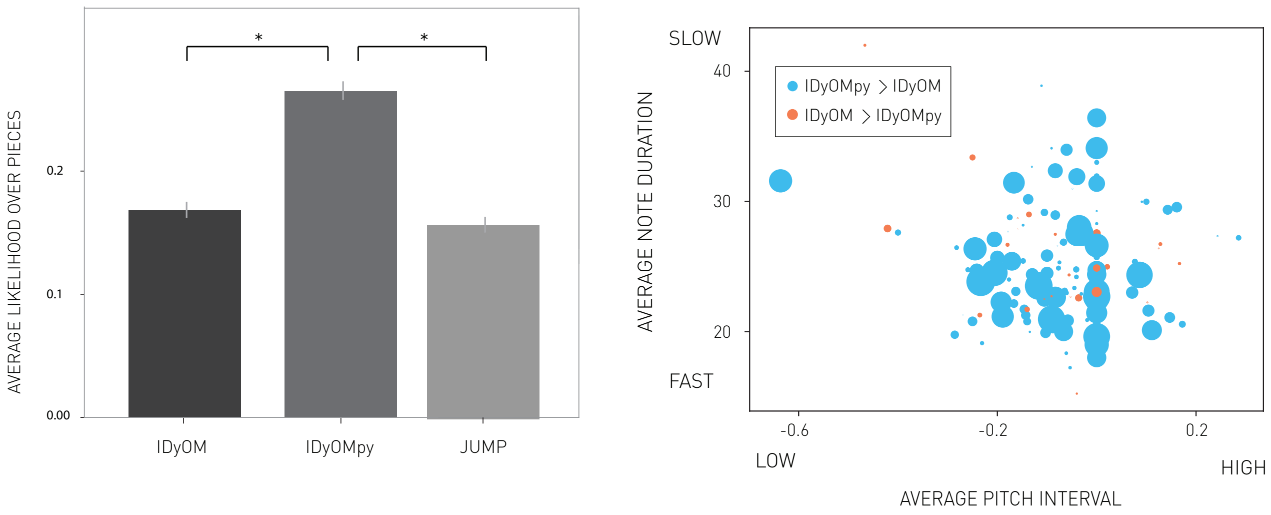


Figure 10: Models comparison on the training set using the theoretical measure ($* : p < 10^{-22}$). (left) Comparison of average likelihood over pieces for the three models using pitch and time onset features. Error bars represent 95% confidence intervals. IDyOMpy was used with maximal order 20 and JUMP with maximal depth 10, both used quantization of 24 notes per beat as the training set was clean and k-fold cross-validation ($k = 5$). (right) Comparison of IDyOM and IDyOMpy within pieces, blue dots resulted in a greater likelihood for IDyOMpy than IDyOM (Lisp), orange dots resulted in a greater likelihood for IDyOM than IDyOMpy, finally the bigger the dots, the greater the difference between the models. Average note duration in expressed in sample, where 1 beat corresponds to 24 samples.

We can see a increasing trend ($p - value = 10^{-22}$ using Wilcoxon test between paired correlations for all pieces) between IDyOM and IDyOMpy and a decreasing trend between IDyOMpy and JUMP ($p - value = 10^{-32}$). When we look at the piece level, we can see that the likelihood increasing trend between IDyOM and IDyOMpy is for 93% of the pieces and that the uniform looking distribution of the likelihood suggests that the algorithm is not overfitting to an obvious sub-space of the data. Between IDyOM and JUMP, the results within pieces are exacerbated and every piece shows a better likelihood for IDyOMpy than JUMP (we don't show the plot).

5.1.2 Evaluation Set (EEG stimuli)

In order to be able to properly interpret the data from the EEG experiment, we use our theoretical measure on the evaluation test (EEG stimuli).

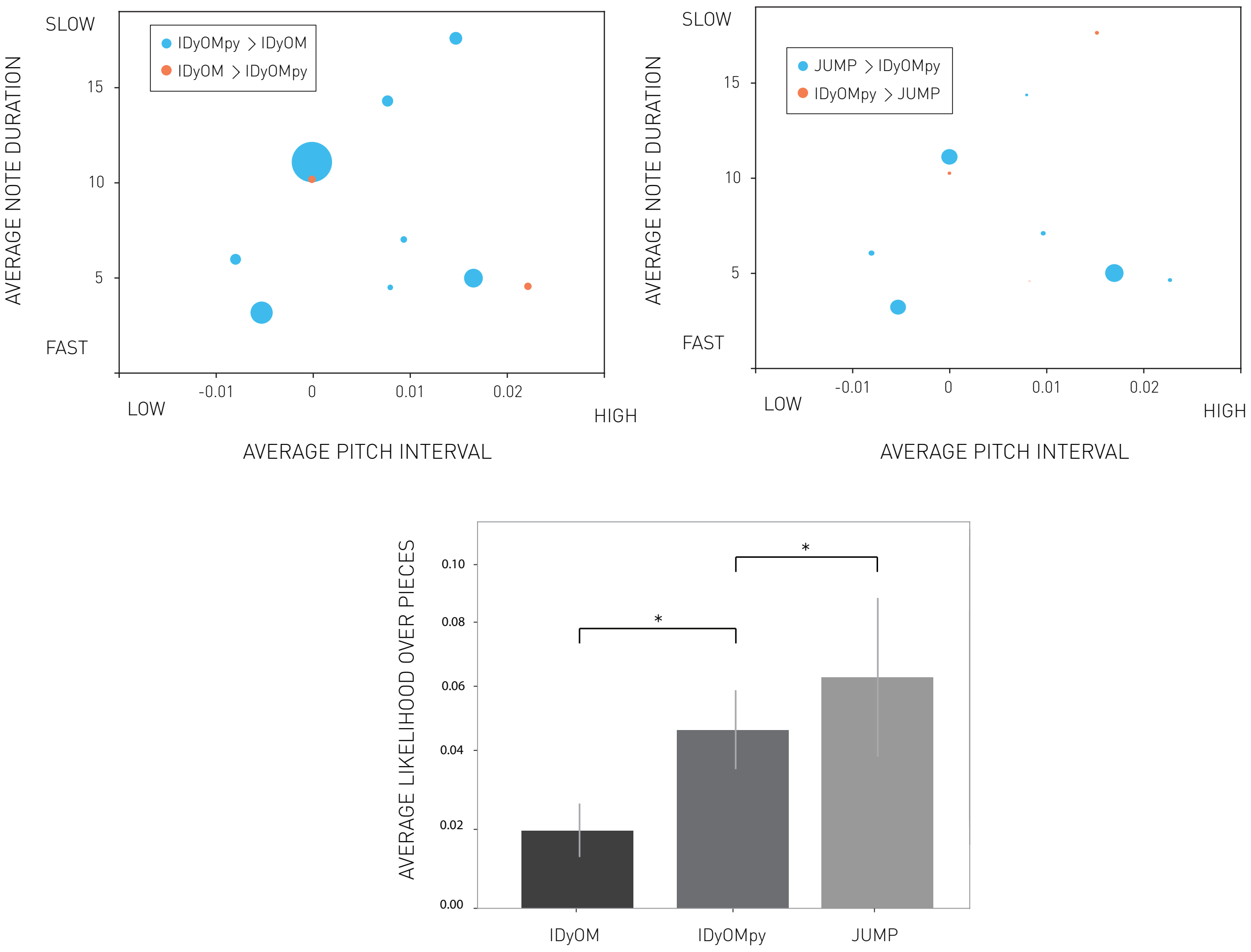


Figure 11: Models comparison on evaluation set using theoretical measure ($* : p < 0.04$). (down) Comparison of the models using average likelihood on the test dataset. (left) Comparison of IDyOM and IDyOMpy piece by piece on the test dataset. Average note duration in expressed in samples, where 1 beat corresponds to 24 samples. (right) Comparison of IDyOMpy and JUMP piece by piece on the test dataset.

We can observe a similar increasing trend between IDyOM and IDyOMpy ($p - value = 0.036$ using Wilcoxon test between paired correlations for all pieces) and a nice distribution over pieces. However, the opposite trend showed in the previous plot between shows IDyOMpy and JUMP ($p - value = 0.024$), which is also quite uniformly distributed over pieces but results in a different ditribution as between IDyOM and IDyOMpy.

5.2 Physiological Measure (eeg)

We compare our models using our physiological measure on EEG data.

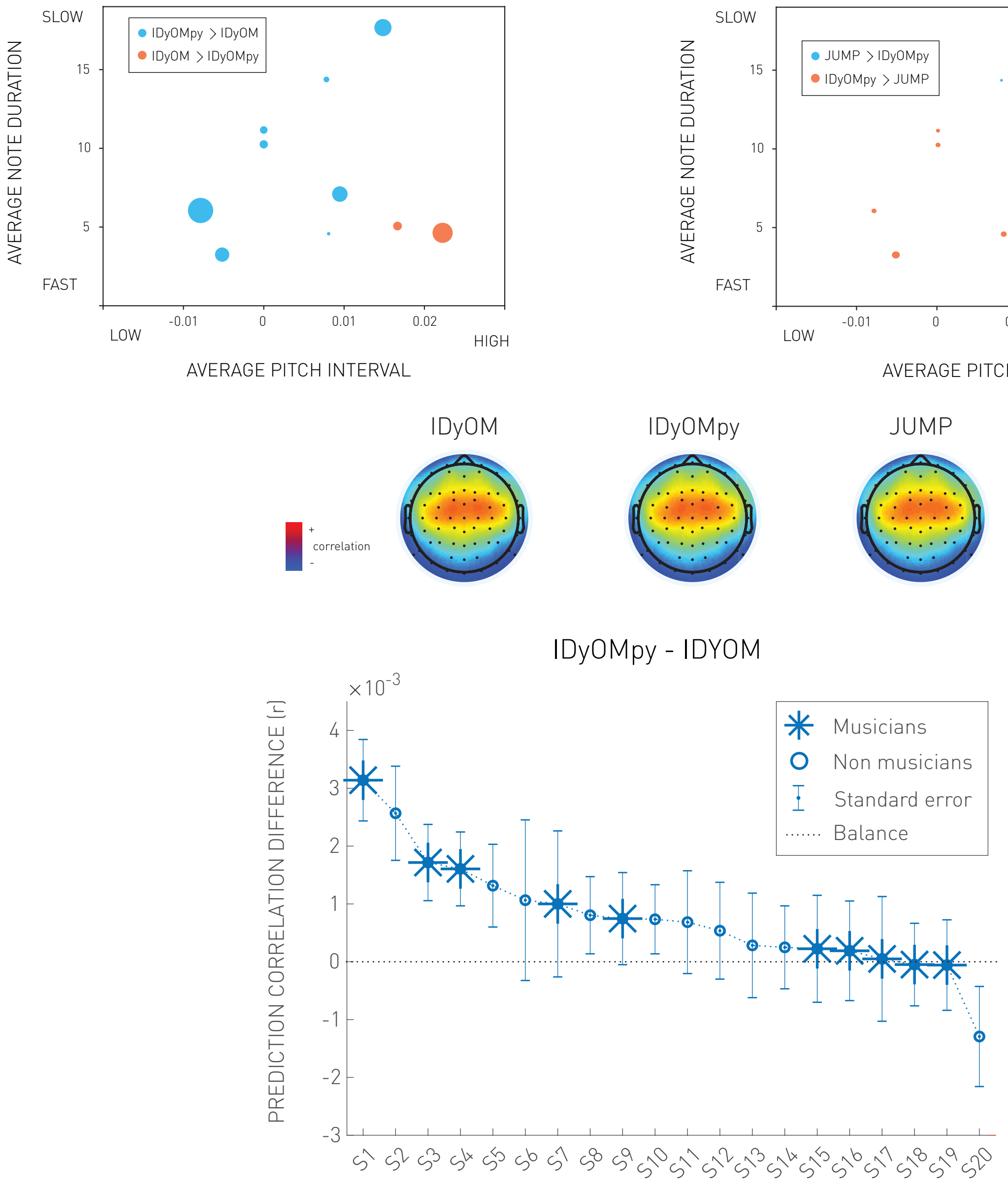


Figure 12: Models comparison using physiological measure (EEG prediction). EEG was predicted (leave-one-out cross-validation) using mTRF from the envelope of the stimuli, the first derivative and the envelope and the surprise signal generated by each model, we use the correlation between the predicted EEG and the true EEG as what we call physiological measure. (left) Comparison of IDyOMpy and IDyOM piece by piece. (right) Comparison of IDyOMpy and JUMP piece by piece. Average note duration is expressed in sample, where 1 beat corresponds to 24 samples. (center) Comparison of the topoplot (spatial correlation). mTRF is used to predict all electrodes, the average prediction correlation within electrodes is plotted over the scalp. (down) Comparison of IDyOM and IDyOMpy by participants ($p = 10^{-3}$). We plotted the correlations from IDyOMpy subtracted by the correlation from IDyOM ($r_{idyompy} - r_{idyom}$). The dash line is the equality line where IDyOM and IDyOMpy get same prediction correlations. The error bars correspond to the standard error computed across pieces. Note that subject S20 (only subject not showing improvement from IDyOMpy) did not show any improvement by adding surprise signal from IDyOM to the envelope.

Between IDyOM and IDyOMpy, we observe a similar increasing trend (stats between subjects below) as shown with the ohter measures and still a uniform distribution over pieces. Note that the distribution over pieces is quite similar as the one showed using the theoretical measure on the same pices.

Between IDyOMpy and JUMP, we observe a similar trend as in the first analysis (theoretical measure on the training set) but the opposite trend as shown using the theoretical measure on evaluation test (same data). Also, note that the distribution over pieces is drastically different from the one using the theoretical measure.

When we look at the comparison between IDyOM and IDyOMpy within-subjects, we observe a clear trend ($p = 10^{-3}$ using Wilcoxon test between paired correlations for all subjects) showing an improvement for every subject but one. It seems that there are no correlations between the quality of the improvement and the musical training of the participants.

Finally, the topoplots show that the spatial correlations are the same for all models, which is, by the way, the same as the envelope and its first derivative (cf. section 2.4).

6 Discussion

We hypothesized that we could replicate results from [3] in order to prove the consistency of the statistical learning theory and, more importantly, that our new implementation of IDyOM may be a better model for studies related to the statistical learning theory of music. Likewise, we also hypothesized that a model based on the prediction of notes further than the subsequent one would allow better long-term dependencies and a better prediction of the EEG signal suggesting similar cortical mechanisms.

First, showing that theoretically valid statistical models of surprise can reliably predict EEG data lighted on the probabilistic prediction hypothesis [5] saying that, in order to structure the musical discourse, the listener uses an inner probabilist model of musical expectation. However, the probabilistic prediction hypothesis is only one of the two bases of the statistical learning theory, the other one being the statistical learning hypothesis. This hypothesis says that this inner model of musical expectation is learned throughout a lifetime of music listening, this mechanism is called enculturation. This means that every listener is supposed to be enculturated differently depending on his/her own experience of music and that people who listened to music from drastically different cultures are supposed to be more differently enculturated than people from the same musical world. This hypothesis (statistical learning hypothesis) has never been proved physiologically [5]. One possible continuation of this study would be to design an experiment to highlight this mechanism in human cortical recordings. As an example, showing that enculturation of people change after intensive passive listening of unfamiliar music would be of interest.

We also showed that our implementation generalizes better in terms of cross-validated likelihood but also that the surprisal from our implementation allows better predictions of the EEG signal for 19 participants over 20. This interesting result can be a step for studies of melodic expectation as this model can enable us to look at more subtle expectation change, and therefore looking at, for example, surprise decoding accuracy over time to look for mechanisms as attention. Also, having a more accurate model can allow us to look at the subtle changes in people enculturation after passive listening of unfamiliar music. Finally, this model can be of interest in the community of music psychology and neurosciences for its usability and modularity.

In order to show the accuracy differences between our implementation and the original one, we presented a framework for musical statistical model benchmarking. Indeed, statistical models of music are usually evaluated using negative log likelihood [16] [2] or through behavioral experiments, often processed online [20] [21]. This framework allows having an objective measure based on cortical correlates of music listening. The literature abounds with statistical models of music: from Markov chains [21] [2] and Bayesian models [22] [23] to Recurrent Neural Network-based ones [19] [20] [16]. Using our framework to evaluate these models would allow us to understand deeper differences (surprisal decoding over time for example) they can have. One question would be to know whether this evaluation framework is sensibly close to the usual behavioral tests supposed to catch the convincibility of the models. Showing that our framework can say a lot on the perceived accuracy of models can lead to fast improvement in musical modeling as we don't have to collect new data for each model.

Finally, (Di Liberto 2019) [3] showed that surprise decoding accuracy differences in musicians and non-musicians were not clear. We showed a similar result in the improvement of our implementation over the original one. This allows us to hypothesize that surprise decoding accuracy is not driven by the property musician/non-musician but maybe by another one. Precisely, Edwin Gordon, American music learning theorist, proposed in 1975 [24] the term audiation to designate the understanding and internal realization of music, in other words, the sensation of hearing or feeling an individual sound when it is not physically present. He suggests that "audiation is to music what thought is to language". Gordon points out that audiation is a potential element of musical ability, arguing that to demonstrate musical ability, audiation is necessary. He introduced several tests of musical ability based on audiation. In his theory, Edwin Gordon defines several stages of audiation whose highest is:

> « Stage 6: Anticipating and predicting tonal patterns and rhythm processes [25]. »

Meaning that audiation abilities are necessary for musical prediction and anticipation, which is exactly what we try to catch in surprise tracking. We can, therefore, hypothesize that the adequacy of the surprise signal produced by our statistical model and the listener EEG recording would depend on this ability in the listener. This question opens a wide range of studies of music and the brain: first, it would interesting to see whether audiation is a musical metaphor and if there is any physiological meaning of it. For example, it would of interest to see whether or not it is possible to decode features of stimuli from EEG of people imagining the stimuli. This would be physiological evidence of the idea of audiation. Then, showing that audiation capabilities (Edwin Gordon defined behavioral tests for audiation capabilities) are related to the accuracy of surprise decoding can highlight the stage 6 of Gordon theory and open doors to underlying mechanisms of melodic expectation.

Unfortunately, our hypothesis on predictions of notes further than the subsequent one (JUMP) cannot be answered because JUMP is not a better predictor of the EEG signal. This can be explained by the parameters (i.e., depth of the predictions and depth merging algorithm) and the implementation. More attention can be driven into the implementation and can eventually lead to positive results. Also, other methods can be used to answer this question. As an example, we can hide notes to the listener while recording EEG so the listener has to predict the following notes with a depth k corresponding to the length of the blank. Our model allows getting the *true* surprise value for these notes. Regressing the surprise signal after deleting the hidden notes can allow us to investigate the question of the jumps in the predictions, and to see if musically trained listeners can predict at bigger depth than others.

This study showed a technical improvement in the domain of statistical learning theory of music and neurosciences and psychology of music. This new implementation will be released as a tool for the community. Besides, this study asks several questions (audiation, enculturation, models comparison) that have been partially investigated in pilot studies. The next section will show preliminary results of these studies.

7 Future Projects

7.1 Decoding Imagined Music

7.1.1 Scientific Question

As shown in this study (cf. section 6), audiation can be an explanation of differences in surprise decoding accuracy. The first related question is to understand whether audiation corresponds to a physiological mechanisms: is imaging music generates auditory cortical responses?

EEG data have been collected as two sets: one participant in Paris where three conditions are investigated: playing, silent playing and imagining using familiar melodies for the participant (professional theremin player) and two participants in Montreal investigating two conditions: listening and imagining using Bach choral melodies.

We use envelope tracking and surprise signal from IDyOMpy to predict the EEG data, we hypothesize that these signals are present in the imagined EEG.

7.1.2 Using Envelope (first dataset)

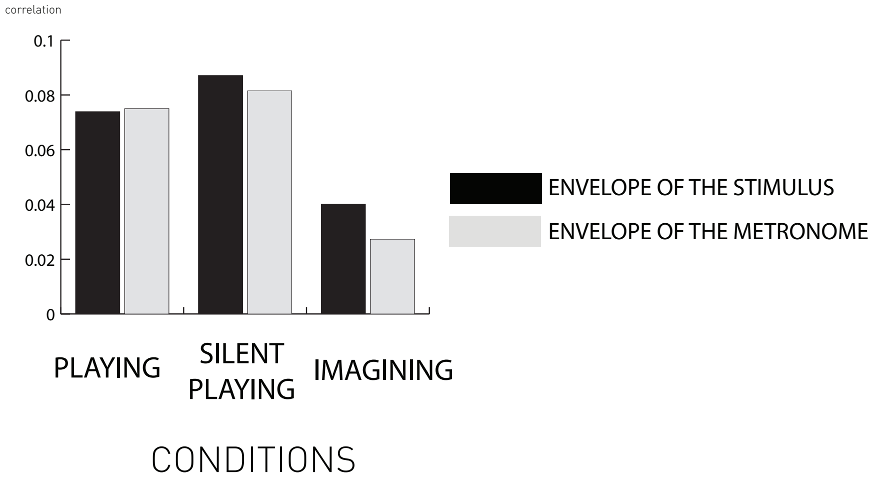


Figure 13: Comparison of the prediction accuracy from the envelope of the metronome (heard sound) and the stimulus (played or imagined sound) with lags window of [-50, 350] ms.

We observe that adding the information of the envelope of the real stimuli (not the metronome) helps to predict the EEG signal, meaning that there is correlates of the stimuli in the EEG data. Shuffling the pieces kills the improvement.

7.1.3 Using CCA between conditions (first dataset)

EEG is available for 3 conditions: playing, silent playing, imaging. CCA was applied to each pair of conditions to try to identify brain responses common across conditions.

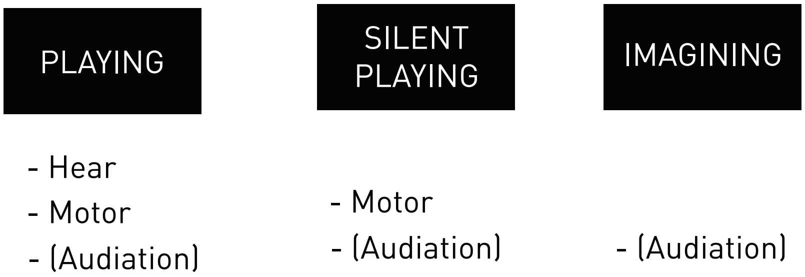


Figure 14: Conditions of the first experiment, we hypothesized that audiation is present in all the conditions, especially in silent playing and imaging.

Data from each condition were centered (mean removed), normalized and submitted to PCA, and the PC series was truncated to 10 PCs to limit overfitting. The following plot shows the correlation between CCs for each CC pair (no cross-validation), as a function of overall shift between data of conditions being compared:

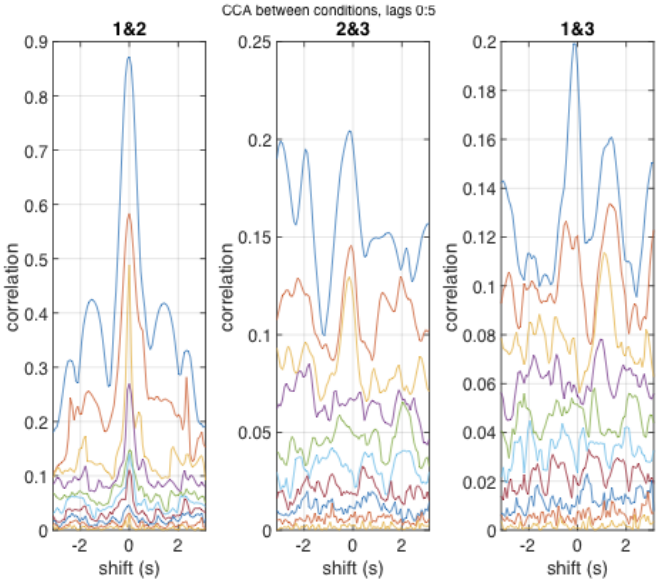


Figure 15: Correlations between conditions from CCA using a time shift.

There seems to be a strong correlation between 1&2 for several CC pairs (compare the height of the peak at zero shift to background). There seems to be a much weaker correlation between 1&3 for one CC, and none (or very weak) between 2&3.

We took the first CCA analysis to transform silent playing condition and imagined condition into a maximally correlated space. Then, we tried to predict the embedded imagined EEG from the envelope of the stimuli using TRF with lags of [-50, 350].

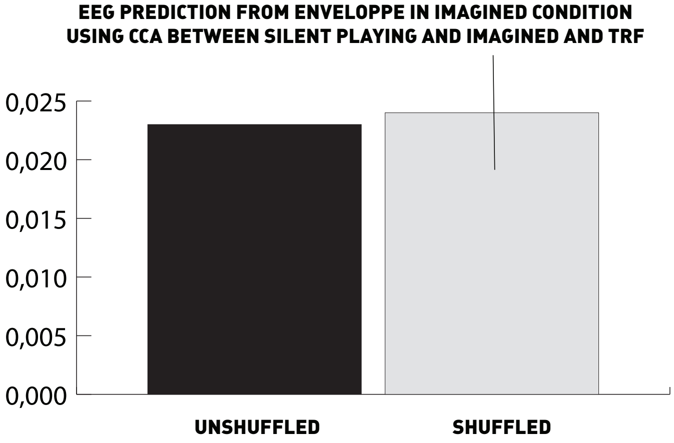


Figure 16: Correlation of imagined and silent playing conditions without and with stimuli shuffling.

We can see that the weak correlation we found is completely due to chance. We then conclude that CCA is not a good analysis for the project and does not allow us to replicate results we found using TRF from the envelope to raw EEG. We think that doing CCA between these conditions result in a signal which is not indicative of the envelope of the imagined sound, the metronome beat is, for instance, much more correlated between these two conditions such that the PCA destroys the audiation correlates.

7.1.4 Using Surprise (second dataset)

We tried to replicate the result found with the first dataset using mTRF to predict imagined EEG from the envelope of the stimuli. For that, we used the envelope of the beat (metronome), the envelope of the stimuli, and the surprise signal from IDyOMpy as regressors for the EEG prediction. As a baseline, we constructed a null-model by training mTRF on shuffled data, of which the prediction should be interpreted as being merely by chance.

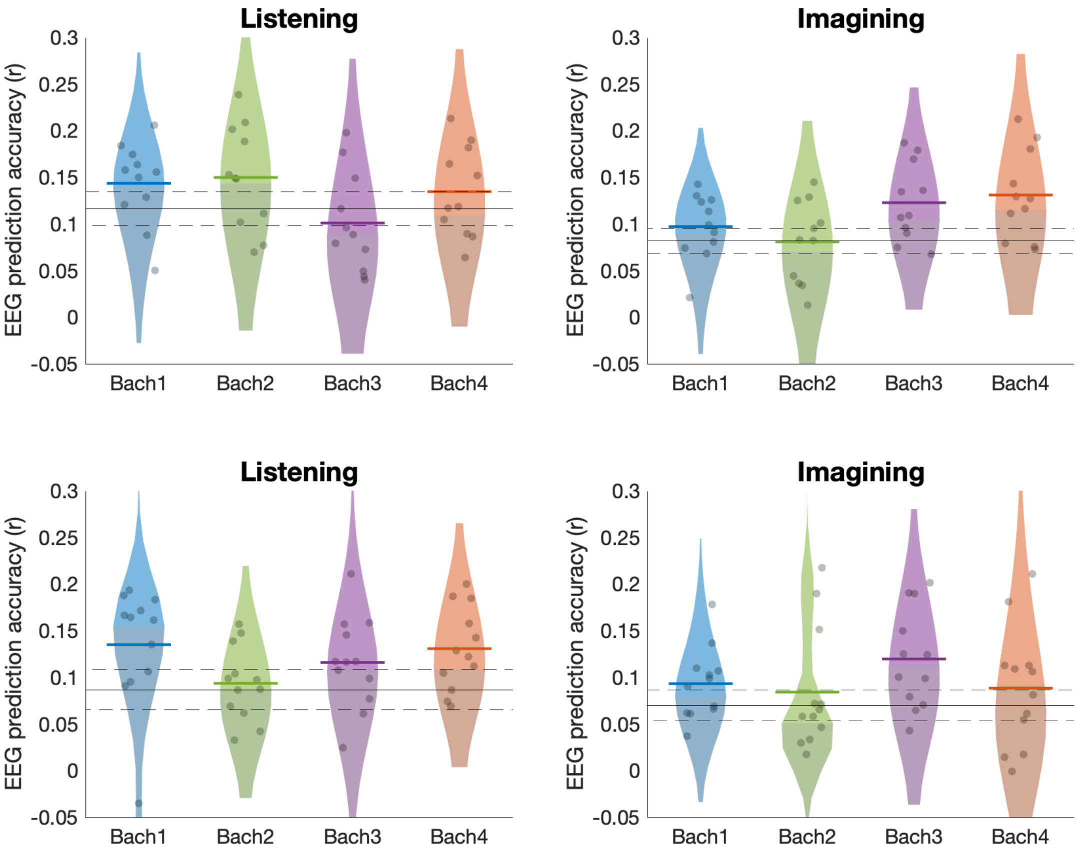


Figure 17: Encoding prediction accuracy. Violin plots show smoothed histograms of EEG prediction correlation for each Bach choral. Gray dots mark individual trials (i.e., leave-one-trial-out cross-validation fold) and colored thick horizontal lines mark the mean of each distribution. Black horizontal lines mark the mean (solid lines) and 95% confidence intervals (dashed lines) of the null-models (chance level). We computed Wilcoxon test using the mean over null-models (10) correlations and the real correlations paired by trials. In imaging condition: $p = 0.0086$ (subject 1) and $p = 0.0250$ (subject 2). In listening: $p = 0.2249$ (subject 1) and $p = 0.0086$ (subject 2).

We can see that the decoding of the imagination condition is significant for both subjects, which means that we found cortical correlates of the stimuli. However, the listening condition is not significant for subject 1, which can be explained by the hardness to stay focus in passive conditions.

7.1.5 Conclusion

This project gives good results for EEG decoding of imagined music using mTRF. However, we need to replicate the experiment with more subjects. Also, it would be very interesting to have a wide range of subjects in order to find an eventual correlation between results of the audiation test from Gordon and the accuracy of the EEG decoding of imagined music. Finally, it will be necessary to look at the spatial correlations as well as the weights of the mTRF in order to compare them with the listening condition.

7.2 Implicit Learning of Novel Musical Genre through Exposure

7.2.1 Scientific Question

In this thesis, we highlighted the consistency of the probabilistic prediction hypothesis (cf. section 6). However, the probabilistic prediction hypothesis is only one of the two bases of the statistical learning theory, the other one being the statistical learning hypothesis. This hypothesis says that this inner model of musical expectation is learned throughout music listening. This enculturation mechanism can, therefore, be tracked by enculturation modeling. We showed (cf. section 2.2) that IDyOM is a good model for enculturation. We hypothesize that, after relatively intensive listening of unfamiliar music, our inner enculturation model will change. This project aims to investigate whether enculturation changes can be tracked using our statistical model of musical surprise (IDyOMpy).

7.2.2 Experiment

We use a 2h dataset of Chinese music (dataset *shanxi* from *Essen Folk Song Collection*) rendered using *Guzheng Kontakt library* from *WavesFactory*. We splitted the dataset in three: training (1h30), test1 (15min) and test2 (15min). The experiment consists in three steps:

Pre-Exposure: participants are asked to listen the test1 set while we record EEG.

Exposure: participants are asked to listen to the whole training set (1h30) every day for two days.

Post-Exposure: participants are asked to listen to test1 and test2 while we record the EEG.

We recorded 3 participants, 2 naive listeners (supposed to be mildly exposed to this music) and 1 for who this music was familiar.

7.2.3 Analysis

We hypothesized that before the exposure the inner model of surprise is closer to a common Western model of surprise than to a proper ‘Chinese‘ model of surprise. And that after exposure, the inner model of the listener should be closer to the Chinese model due to the statistical learning hypothesis. We so simulate these two models by training IDyOMpy with the training set (Chinese model) and with western music (Western model). A recent study [5] showed that IDyOM model behaves strictly differently for these two datasets, and therefore make a comparison meaningful.

Therefore, we expect the Western model to be more predictive before exposure and the Chinese model to be more predictive after exposure, in other words, to see the relative utility of the Chinese model to be increased after exposure.

7.2.4 Results

We track the evolution of the improvement of the EEG prediction by the surprise generated by the Chinese model. The plot shows:

$$correlationChineseSurprise/correlationWesternSurprise$$

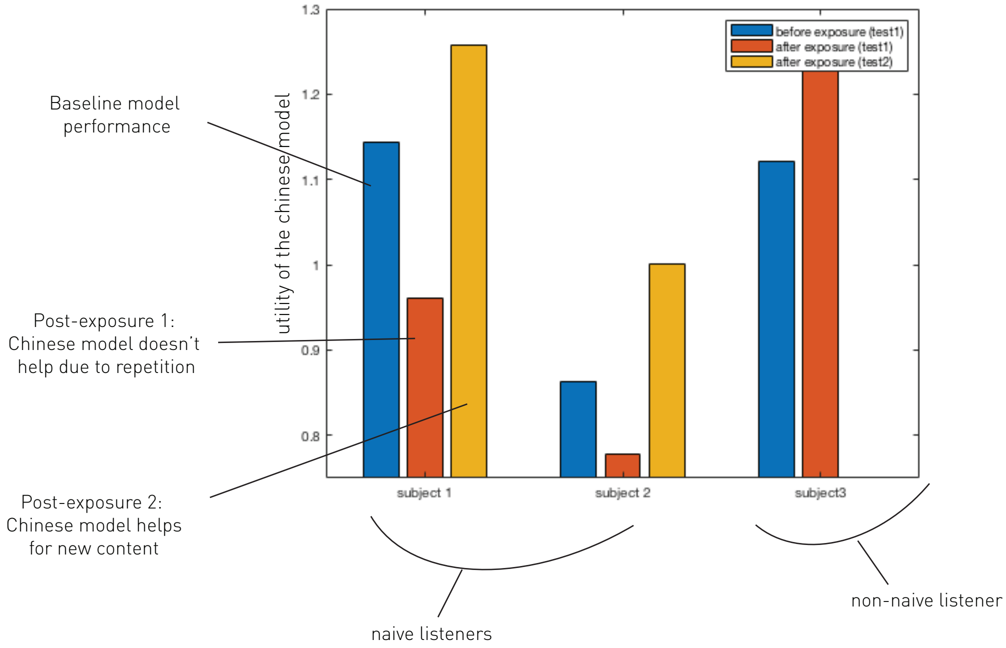


Figure 18: Comparison of the utility of the Chinese model over the Western model after and before exposure.

We observe a decreasing trend of the utility of the Chinese model after exposure for the data of test1, these data were the same as before exposure. We also observe an increasing trend after exposure for the data of test2 with respect to data of test1 before exposure. The same pattern is observed for both naive subjects.

7.2.5 Discussion

Data from test1 was already known by the subjects from the first experiment (before exposure), therefore, we can interpret the increasing utility of the Western model in the second listening (after exposure) as a marker of the remembering of the pieces. As the pieces were already known through the Western model, we can hypothesize that in the second listening, it is more likely to process the same piece through the same model. However, it is also likely that the inner Chinese model and the inner Western model of the listeners are more separated after exposure, and so the utility of what is specific to the new Chinese model (after exposure) decreases in the second listening.

On the other hand, the trend before exposure (test1) and after exposure (test2) is very clear and fit exactly to our expectations. New musical contents are not associated with any specific model and will, therefore, be likely to be processed through the Chinese model newly trained during exposure. This model being more accurate, it results in better prediction of the EEG data.

Even if results are hard to interpret, the same trend is shown for both naive listeners and comfort us in our interpretation that this experiment may be evidence of the statistical learning hypothesis. However, more subjects need to be collected to be confident in the data and, therefore, in the interpretation.

8 Conclusions and Future Directions

In this project, we showed that IDyOM is a good model for musical enculturation but that this model was too difficult to modify and access. Therefore, we implemented a new version in Python showing better results for both theoretical (average likelihood comparison) and physiological (EEG prediction correlation) measures on Bach chorals and concerti. This new implementation will be of interest of the community of neurosciences and psychology of music, and therefore will be released freely on the internet. We also showed that the modularity of our new implementation allows implementing physiological hypotheses into variants of our model. Using our framework to assess the validity of such models can lead to evidence of similar cortical mechanisms. Replicating such clear evidence of the consistency of statistical models of surprise validates the probabilistic prediction hypothesis. However, this hypothesis is only one of the two bases of the statistical learning theory. The other hypothesis (statistical learning hypothesis) has been investigated in an additional pilot project showing great preliminary results of enculturation change after passive exposure to unfamiliar music.

The new framework we used to compare models of music surprise can be used as a benchmark for statistical models of surprise, replacing the usually used negative log-likelihood. Also, this framework can be generalized to all statistical models of music using the notion of likelihood and allows several advantages over the current benchmark methods such as: (1) it is a physiological and objective measure of statistical models, (2) it does not need additional data when a model is changed, (3) it does not require any part of the training data to evaluate. Therefore, this framework will also be shared with the community and is already used by researchers at Maryland University to evaluate a new model of surprise [22].

Finally, we hypothesized that the accuracy of EEG decoding of surprise is not directly driven by musical training but by audiation capability. We showed for the first time, in a pilot project, clear evidence that audiation is not a musical metaphor but has a cortical meaning as it is possible to decode envelope from EEG of imagined music. It is planned to keep this project going in order to show whether audiation capabilities are correlated to surprise decoding accuracy.

This research will be continued in a Ph.D. at Laboratoire des Systèmes Perceptifs supervised by Prof. Shihab Shamma, the first project will be to reproduce the imagined music decoding with a reasonable amount of subjects (20) and a very first step will be to see if using tactile metronome instead of regular metronome may improve the decoding.

A part of the pilot experiments for future projects was made at the 2019 Telluride Neuromorphic Cognition Engineering Workshop. I would like to thank all the participants and faculty of the workshop for their very warm welcome and their help and a special thanks to the people who helped me for the projects that I used in this report, by projects:

Decoding imagined music: Giovanni Di Liberto (ENS), Seung-Goo Kim (Duke), Aditya G. Nair (UW), Lauren Fink (UC, Davis), Alain de Cheveigné (ENS), Shihab Shamma (ENS), Lisa Margulis (Princeton).

Implicit learning trough passive exposure: Pablo Ripollés (NYU), Claire Pelofi (NYU), Sandeep Kothinti (JHU), Shihab Shamma (ENS), Mounya Elhilali (JHU)

Physiological measure of statistical models of music: Sandeep Kothinti (J.Hopkins), Shihab Shamma (ENS), Mounya Elhilali (J.Hopkins)

I would also like to thank Daniel Pressnitzer for introducing me to the Laboratoire des Systèmes Perceptifs, Yves Boubenec for introducing to my current and future supervisor, Shihab Shamma and all the personnel of the LSP and IRCAM.

Finally, I want to tremendously thank (words are weak) **Giovanni Di Liberto** and **Shihab Shamma** for their extremely warm welcome to the laboratory, their daily happiness, their extraordinary supervising and for sharing their precious knowledge with me during these scientifically very exciting 6 months. They truly allowed me to build all these projects and collaborations that will be following me during all my Ph.D.

References

[1] J.-J. Nattiez. *Musicologie générale et sémiologie.* Christian Bourgois Editeur, 1987.

[2] M. T. Pearce. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition.* PhD thesis, 2005.

[3] Giovanni M. Di Liberto, Claire Pelofi, Roberta Bianco, Prachi Patel, Ashesh D. Menhta, Jose L. Herrero, Nima Mesgarani, Alain de Cheveigné, and Shihab Shamma. Cortical encoding of melodic expectations in human temporal cortex. *in preparation, url:https://www.biorxiv.org/content/10.1101/714634v2.*

[4] Leonard C. Manzara, Ian H. Witten, and Mark James. On the entropy of music: An experiment with bach chorale melodies. *Leonardo Music Journal*, 2(1):81–88, 1992.

[5] M. T. Pearce. Statistical learning and probabilistic prediction in music cognition: mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, (1423), 2018.

[6] Diana Omigie, Marcus T. Pearce, Victoria J. Williamson, and Lauren Stewart. Electrophysiological correlates of melodic processing in congenital amusia. *Neuropsychologia*, 51(9):1749 – 1762, 2013.

[7] Diana Omigie, Marcus T. Pearce, Katia Lehongre, Dominic Hasboun, Vincent Navarro, Claude Adam, and Severine Samson. Intracranial recordings and computational modelling of music reveal the time-course of prediction error signaling in frontal and temporal cortices. *Journal of Cognitive Neuroscience*, 31(6):855–873, April 2019.

[8] Michael J. Crosse, Giovanni M. Di Liberto, and Edmund C. Lalor. The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli. *Frontiers of Human Neuroscience*, 10(604), 2016.

[9] Sascha Griffiths, Matthew Purver, and Geraint Wiggins. From phoneme to morpheme: A computational model. 11 2015.

[10] Steven Morrison, Steven Demorest, and Laura Stambaugh. Enculturation effects in music cognition: The role of age and music complexity. *Journal of Research in Music Education*, 56, 01 2008.

[11] L.B. Meyer. *Explaining Music: Essays and Explorations.* Berkeley, CA: University of California Press., 1973.

[12] H. KALMUS and D. B. FRY. On tune deafness (dysmelodia): frequency, development, genetics and musical background. *Annals of Human Genetics*, 43(4):369–382, 1980.

[13] Krista L. Hyde, Robert J. Zatorre, Timothy D. Griffiths, Jason P. Lerch, and Isabelle Peretz. Morphometry of the amusic brain: a two-site study. *Brain*, 129(10):2562–2570, 08 2006.

[14] A. Moffat. Implementing the ppm data compression scheme. *IEEE Transactions on Communications*, 38(11):1917–1921, Nov 1990.

[15] K. Agres, S. Abdallah, and M. T. Pearce. Information-theoretic properties of auditory sequences dynamically influence expectation and memory. *Cognitive Science*, (42), 2018.

[16] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, and Douglas Eck. An improved relative self-attention mechanism for transformer with application to music generation. *CoRR*, abs/1809.04281, 2018.

[17] Giovanni M. Di Liberto, Claire Pelofi, Shihab Shamma, and Alain de Cheveigné. Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening. *Acoustical Science and Technology (AST)*, in press.

[18] Jonas Obleser, Björn Herrmann, and Molly J Henry. Neural oscillations in speech: Don't be enslaved by the envelope. *Frontiers in human neuroscience*, 6:250, 08 2012.

[19] Jean-Pierre Briot, Gaëtan Hadjeres, and François Pachet. Deep learning techniques for music generation - A survey. *CoRR*, abs/1709.01620, 2017.

[20] Gaëtan Hadjeres, François Pachet, and Frank Nielsen. DeepBach: a steerable model for Bach chorales generation. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1362–1371, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[21] Jon Gillick, Kevin Tang, Robert M. Keller, and Harvey Mudd Collge. Learning jazz grammars. In *in Proceedings of the SMC 2009 - 6th Sound and Music Computing Conference*, 2009.

[22] Nicholas Huang and Mounya Elhilali. Auditory salience using natural soundscapes. *The Journal of the Acoustical Society of America*, 141:2163–2176, 03 2017.

[23] Ryan Prescott Adams and David J. C. MacKay. Bayesian Online Changepoint Detection. *arXiv e-prints*, page arXiv:0710.3742, Oct 2007.

[24] R. C. Gerhardstein. The historical roots and development of audiation: A process for musical understanding. *In Hanley, B. Goolsby, T.W. (Eds.) Musical understanding: Perspectives in theory and practice*, 2002.

[25] Edwin Gordon. *Learning Sequences in Music: A Contemporary Learning Theory.* Chicago: GIA Publications, 2007.