# Predicting Imagined Meters in Musical Patterns from MEG Data

Aashna Shroff (aashna), Ben Limonchik (benlimon), and Zoe-Alanah Robert (zrobert7)

## I. Introduction

Musical data is often interpreted within a metrical framework that integrates hierarchical timing information. It has been shown previously that listening to a constant metronome and imagining different meters modulated the resulting auditory evoked response in the temporal lobe and motor-related brain areas such as the motor cortex, basal ganglia, and cerebellum [1]. We aim to use Magnetoencephalography (MEG) brain data to classify the metrical framework (i.e. march, waltz or hemiola) that a person is imagining while listening to a constant metronome. Additionally, we used machine learning techniques to analyze MEG data and identify brain areas and time intervals that are primarily associated with such imagination.

Such an analysis can contribute to music neuroscience research, which can be used to potentially inform music therapy, music education and music performance. Fujioka et. al. [2] mention how such a study can help in stroke rehabilitation to relearn lost motor and sensory skills. The meter of music is its basic rhythmic structure. While this project works to create a basic three-way brainwave-to-meter classification system, a more advanced classifier could ultimately be used by professional musicians to compose music directly from their brain.

Stated clearly, the input to our algorithm is a series of MEG amplitude measurements over a time interval. We then use a number of different classifiers to output a predicted label (march, waltz or hemiola).

## II. Related Work

On a broad level, the two dominant machine learning applications of MEG are 1) electromagnetic brain imaging and 2) brain state classification. Past work related to brain imaging have involved reconstruction and visualizations of neuronal activities based on MEG/EEG measurements, and have been typically used in clinical neuroscience to identify imaging associated with normal brain function and how they might be altered in a disease. For example, machine learning techniques have been used to identify patterns from MEG data for epileptic patients [3], depression [4], and autism in children [5]. Brain state classification algorithms have been typically used to design brain-computer interfaces (BCIs) [6] and making predictions based on MEG data, like movement [7].

Extensive research has been done to use Bayesian Machine Learning for MEG data [8], which we have leveraged in our work. Unfortunately, there have been no attempts to build a system exactly like ours. However, the research mentioned above to make movement predictions shares some similarities; our approaches for preprocessing of MEG data, classifier choice, and evaluation metrics are comparable. That study used regularized linear discriminant analysis and reported accuracy results in the form of percentage of correctly decoded trials. The highest obtained was 67% on a four-way classification system. However, it is also widely known that motor detection is a much easier classification than the imagining tasks that we have attempted to perform.

## III. Dataset and Features

We received our data set from Dr. Takako Fujioka at the Stanford CCRMA Labs, which was collected in a 2014 study [9]. In the experiment, 12 professional musicians were asked to perform listening, tapping, and imagining tasks while under the monitoring of an MEG machine. A constant metronome of 12 unaccented clicks was played in intervals of 390 ms. Participants were asked to imagine one of the following musical rhythms:

$$March: 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2$$
$$Waltz: 1\ 2\ 3\ 1\ 2\ 3\ 1\ 2\ 3\ 1\ 2\ 3$$
$$Hemiola: 1\ 2\ 3\ 1\ 2\ 3\ 1\ 2\ 1\ 2\ 1\ 2$$

Thus, the data looks as follows: 12 participants, 190 trials each, measurements over 94 time intervals at every 0.0128 seconds, and 12,180 amplitude measurements at different spatial locations in the brain. The 12,180 measurements represent the head divided into 12,180 voxels (8 mm cubes). Before it was provided to us, the data was filtered to only include the Beta frequency band (a specific frequency range of human brain activity). The initial MEG readings were divided into 9 different classifications of up, down, and pivot beats for the three rhythms. A data point for a single class thus includes 12,180 measurements for a time segment of 380 milliseconds before and 780 milliseconds after the metronome click. We had in total $190$ trials $\times 12$ participants $\times 9$ classifications $= 20,520$ data points. These were divided in training, validation, and test depending on the classifier being run. It is important to note than for the scope of this report, we were only concerned with 3 of the 9 classes. Given the large amounts of data and the extremely large number of features it would be unreasonable to let any classification model run on the data as a whole. The following preprocessing measures were taken to limit the number of features used:

- Any voxel with a zero-reading over all trials, classes, and participants from the 12,180-set was removed. This resulted in 4,929 remaining cubes, since the other cubes were outside or in the periphery of the brain.
- Wave amplitudes measured on the peripheries of the 94-interval time-frame were found to be unreliable due to moving average preprocessing. Therefore, the first and last 8 MEG measurements on the 94-interval time-frame were removed so only 78 time-intervals remained.
- We augmented our input matrix from 3 dimensions (trial, time, brain voxel) to 2 dimensions (trial, time $\times$ brain voxel). Consequently, each input matrix (see Figure 1) was a (78 $\times$ 4,929 =) 384,462 by 1 matrix, which allowed the feature set to be reduced to 400,000 features from more than 1 million.

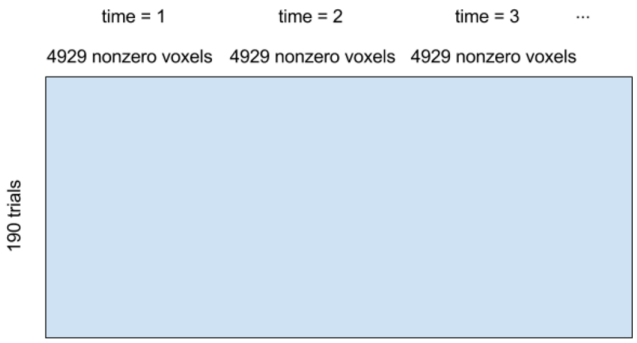


Fig. 1. *Example input matrix after preprocessing*

## IV. Methods

In our research, we narrowed our choice of classifiers to Linear Discriminant, Quadratic Discriminant and Cubic SVM. We chose to start with Discriminative classifiers first because they tend to execute very quickly and their results can easily be visualized (using boundaries) and interpreted. However, Discriminant classifiers tend to perform better when both the predictor $(x_i)$ and the response $(y)$ are numeric since they are good at learning the numerical relationship between successive responses (e.g. when classes are 1-2-3 and there is meaning to the increasing numbers). While our data is numeric (representing amplitudes), our responses are categorical rather than numeric, namely, Waltz, March and Hemiola. Research shows that for these types of problems Support Vector Machines tend to outperform Discriminative classifiers. Furthermore, SVM classifiers are very popular in neuroscience due to their great discriminative power of spectral clustering across different subjects. Therefore, we continued our analysis using SVM classifiers in addition to Discriminant ones.

What follows is a brief description of each of these three classifiers.

### A. *Linear Discriminant and Quadratic Discriminant*

Discriminative classifiers define boundaries to separate different classes of data. In the case of Linear Discriminant analysis, only learn linear boundaries are learnt. Quadratic Discriminant allows more flexibility because it can learn boundaries that are quadratic in shape. This is illustrated in the bottom row of Figure 2.

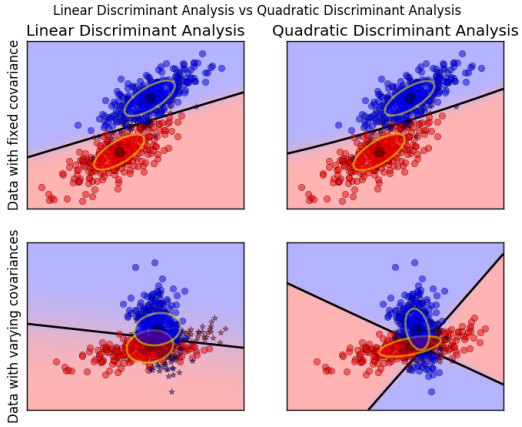


Fig. 2. *The plot shows decision boundaries for Linear Discriminant Analysis and Quadratic Discriminant Analysis.*

Mathematically, for both Linear and Quadratic Discriminants, Bayes rule can be used to estimate the probability of a certain class (i.e. Hemiola, Waltz , March) given the input data set:

$$P(y=k|X) = \frac{P(X|y=k)P(y=k)}{P(X)} = \frac{P(X|y=k)P(y=k)}{\sum_l P(X|y=l)\cdot P(y=l)}$$

where $y$ is the music meter ouput, $X$ is the input MEG data and $l$ goes from 1 to 3 (i.e. the number of classes we have). In addition, to estimate the probability of the data given a target class, a multivariate Gaussian distribution is used:

$$P(X|y=k) = \frac{1}{(2\pi)^n|\sigma_k|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(X-\mu_k)^t\sigma_k^{-1}(X-\mu_k)\right)$$

The difference between Linear discriminant and Quadratic discriminant classifiers is that for the Linear discriminant, it is assumed that target classes share the same covariance matrix (causing a linear decision boundary) while for the Quadratic discriminant, classes may have multiple covariance matrices which leads to the quadratic decision surface.

### B. *Cubic SVM:*

SVMs are used to find a decision boundary that maximizes the margin separating the three different classes (Waltz, March, Hemiola) of our training data points. For simplicity, consider a nonlinear SVM for classifying two-class data. As shown in figure 2, an SVM can be used to change the dimension of the data so that a linear classifier can be applied to it.

Mathematically, for SVMs using the hinge loss function, the following objective function is minimized:

$$\frac{1}{n}\sum_{i=1}^{n}\max(0, 1-y_i(\vec{w}\cdot\vec{x_i}-b))$$

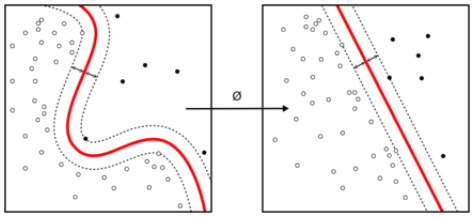


Fig. 3. *Kernel Machine for SVM*

where $x_i$ is our data point $i$, n is the number of data points, $y_i$ is the correct label, b is the intercept and w is the weight vector we want to optimize to maximize the margin on the data. Note that the linear SVM above depends on the dot product between the data point vectors. Thus, to generalize the SVM for nonlinear boundaries, this dot product is replaced with a kernel: $K(w, xi)$.

## V. Results and Analysis

As stated previously, our dataset initially consisted of 78 time intervals $\times$ 4929 non-zero brain voxels = 384,462 features. Just to run any classifiers at all, we had to begin by selecting random $K = 100$ features from this set. This was performed 10 times with no validation and then the average accuracy, precision, and recall was calculated. This is reported in Figure 4 and 5.

Testing was done in two ways (1) using seen subjects which were used for training; and (2) using 2 unseen test subjects that were not included in training. Accuracy for seen subjects was overall very high, even with randomly selected features. This resulted in the conclusion that an individuals brain is consistent trial to trial but brains tend to differ largely person to person.

Results were poor for unseen subjects. As shown in Figure 5, accuracy varied from 29% to 32%. However, in both cases, Cubic SVM outperformed Discriminant classifiers, which is consistent with our analysis in the previous section.

Note: For readability purposes, only the precision and recall results for Waltz is provided instead of all the 3 classes.

| | Linear Discriminant | Quadratic Discriminant | Cubic SVM |
|---|---|---|---|
| **Accuracy** | **70.42** | **44.31** | **97.08** |
| **Precision (Waltz)** | **32.87** | **8.76** | **62.73** |
| **Recall (Waltz)** | **100** | **100** | **100** |

Fig. 4. *Test results for subjects used in training*

Following these results, we decided to leverage scientific domain knowledge to improve results and further reduce the feature set. Brain voxels were spatially averaged according to the Talairach coordinate system, a 3-dimensional coordinate system used to locate brain structures independent from individual differences in the size and overall shape of a brain [10]. Each voxel was mapped to one of 72 well-known regions/cortices of the brain. The set of features mapping to a single cortex was replaced with a single average feature for that cortex over all time intervals. This reduced the set of features to $78 \times 72 = 5616$ features.

| | Linear Discriminant | Quadratic Discriminant | Cubic SVM |
|---|---|---|---|
| **Accuracy** | **29.56** | **30.58** | **31.81** |
| **Precision (Waltz)** | **20.27** | **50.95** | **6.02** |
| **Recall (Waltz)** | **54.81** | **50.03** | **14.66** |

Fig. 5. *Test results for unseen subjects*

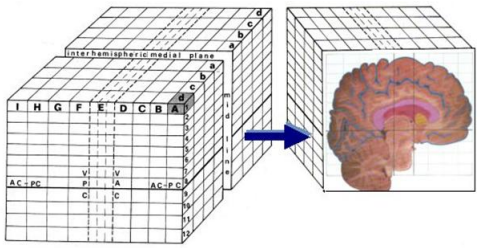


Fig. 6. *Human brain with superimposed Talairach grid*

SVM was run again with no validation. As shown in Figure 7, this feature set performed better but still poorly on unseen people.

| | SEEN: Cubic SVM | | | UNSEEN: Cubic SVM | | |
|---|---|---|---|---|---|---|
| **Accuracy** | **94.17** | | | **38.98** | | |
| **Confusion Matrix** | | **March** | **Waltz** | **Hemiola** | **March** | **Waltz** | **Hemiola** |
| | **March** | **237** | **2** | **0** | **46** | **72** | **3** |
| | **Waltz** | **105** | **132** | **0** | **28** | **62** | **44** |
| | **Hemiola** | **23** | **221** | **0** | **291** | **243** | **327** |

Fig. 7. *Test results using averaged brain cortex features and Cubic SVM*

It seemed to be overfitting for the 10 people that our classifier was trained on. To test our hypothesis, we applied the $t$-test on each feature to compare the location parameter of each independent data sample:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

where $\bar{x}$ and $\bar{y}$ are the sample means, $s_x$ and $s_y$ are the sample standard deviations, and $n$ and $m$ are the sample sizes. In order to analyze how well-separated the two-groups are by each feature, the empirical cumulative distribution (CDF) of the $p-$values was plotted, as shown in Figure 8.

About 75% of features having p-values smaller than 0.05, meaning there are more than 75% of the original 5616 features that have strong discrimination power. The features

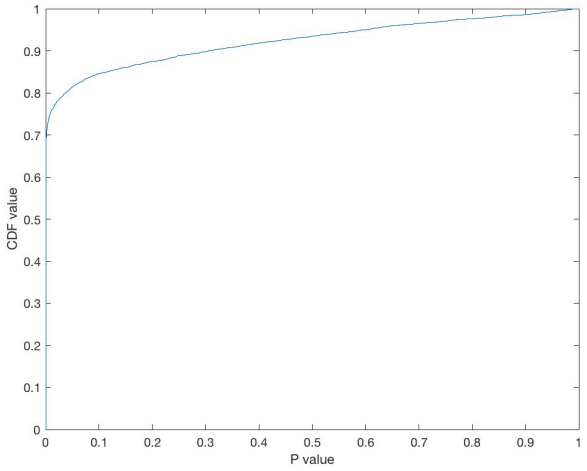


Fig. 8. *Empricial CDF of p-values of different features*

could be sorted according to their p-values and the best ones could be picked, but the number of features to be selected needed to be decided.

To decide this, the MCE (misclassification error, i.e., the number of misclassified observations divided by the number of observations) on the test set was plotted as a function of the number of features. This is shown in Figure 9. To illustrate why resubstitution error is not a good error estimate of the test error, we also show the resubstitution MCE using red triangular marks.

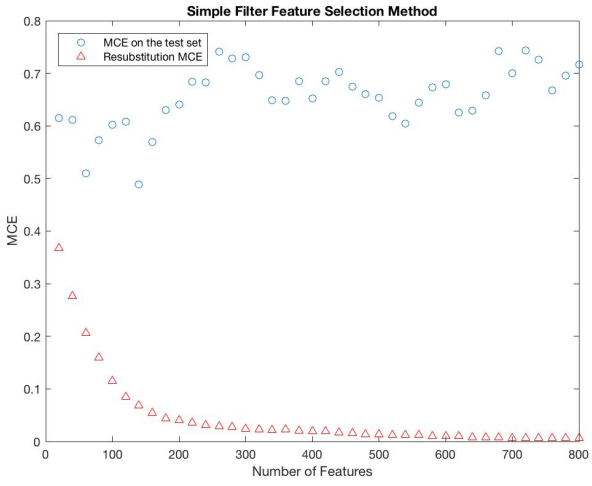


Fig. 9. *Resubstitution and Test MCE for varying feature set sizes*

The resubstitution MCE is over-optimistic. It consistently decreases when more features are used and drops to zero when more than 500 features are used. However, if the test error increases while the resubstitution error still decreases, then overfitting may have occurred. This simple filter feature selection method gets the smallest MCE on the test set when 50 features are used. The plot shows overfitting begins to occur when 75 or more features are used.

Therefore, the features were sorted according to their p-values and the best 50 were selected. This substantially improved the accuracy of the classifier on unseen data, as shown in figure 10

| | SEEN: Cubic SVM | UNSEEN: Cubic SVM |
|---|---|---|
| **Accuracy** | **93** | **43.56** |
| **Confusion Matrix** | **March Waltz Hemiola**<br>**March 192 8 0**<br>**Waltz 118 122 0**<br>**Hemiola 55 225 0** | **March Waltz Hemiola**<br>**March 74 48 13**<br>**Waltz 0 122 8**<br>**Hemiola 291 281 353** |

Fig. 10. *Test results using best 50 features and Cubic SVM*

After seeing an improvement in the classification of unseen people after only including the 50 best features, we analyzed what those features represented temporally and spatially. We mapped each feature index back to a (time interval, brian cortex) pair and found the most common cortices to be the temporal cortex and the premotor cortex. The temporal cortex is responsible for processing of audio input and the premotor cortex is associated with planning and preparing body movements. Looking back to the initial experiment in which participants were imagining themselves tapping to a specific rhythm while listening to a metronome, our feature engineering findings validate the scientific definition of these cortices [11]. When examining the top 50 features temporally, we saw that the most predictive time intervals were immediately following the down beat stimuli. This is visible in figure 11.

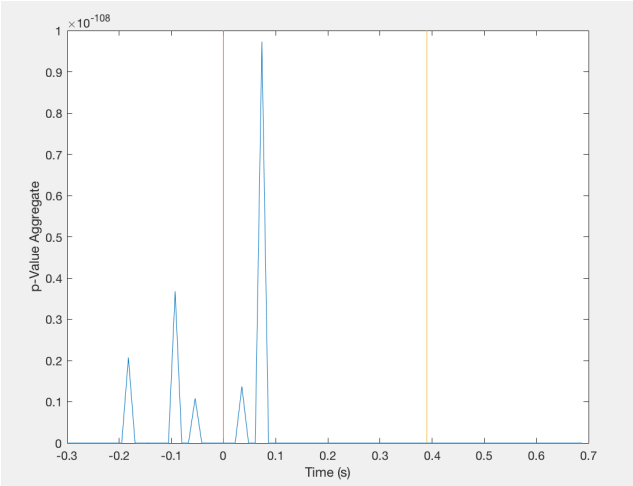


Fig. 11. *Time interval vs. Aggregated p-values*

Throughout our experimental process, we tested on both seen and unseen people. Ultimately, classifying seen people but unseen trials proved to be a trivial machine learning challenge. Classifying entirely unseen brains, on the other hand, was much more difficult. Ultimately, we were not able to get above a 43% accuracy for unseen brains. However, this is largely due to a lack of data. A dataset consisiting of only

12 brains is insufficient to build a classifier that can generalize well.

## VI. Conclusion / Future Work

In conclusion, we found that the best performing classifier for our data was cubic SVM which made sense due to the reasons described in the methods section. As our feature engineering process showed, by removing unnecessary time intervals and irrelevant brain cortexes we were able to improve the generalizability of our model to unseen subjects. More specifically we saw an improvement from 33% precision (almost random) to 43% precision for a three-way classification problem. Given these encouraging results there are additional modifications to our feature set which we believe would have further led to a greater improvement in precision. These include:

- **Reducing our set of 4800 voxels to the few dozens voxels originating from the Temporal Lobe and the Pre-Motor Cortex:** Recall that the MEG readings from 4800 voxels were averaged into 72 physical brain cortexes. While we found that the most indicative averaged features came from the Temporal Lobe and the Pre-Motor Cortex, we suspect that in averaging the MEG readings we may have lost some strong predictors. Therefore, given we now know which cortexes are most relevant to the imagination of music meters, we would have wanted to take a step back and reduce the number of voxels under consideration to the 100 unaveraged voxels originating in the Temporal Lobe and the Pre-Motor Cortex.
- **Applying ICA to our MEG data:** MEG readings are susceptible to noise caused by magnetic fields in the vicinity of the MEG machine. Even in magnetically shielded environments it is almost impossible to avoid incurring magnetic noise from surrounding sources such as the machines power line. Therefore, ICA could be used in order to reduce the noise on our data. We were able to find a few ICA techniques specifically applicable to MEG data however in the absence of an MEG expert during our research it would have been hard to tell if we successfully removed noised from our MEG and thus we decided to focus our efforts on feature engineering since cortices can be more easily detected without prior knowledge of the underlying biology [12].

Finally, our project focused on differentiating the downbeats of three different music meters: Hemiola, Waltz and March. Nevertheless, in our disposal we had data for up beats, middle beats and pivotal beats for the Hemiola readings. Given greater computational resources we would have attempted to do a 3 way classification of the different beats within the Hemiola meter (i.e. D3, M3, U3) in order to see if we can clearly distinguish beats within the same meter. Combining this research with the 3-way classification of different meter types would have enabled us to better estimate the feasibility of composing music purely based on MEG waves, since both a meter and a certain beat within a given meter has to be chosen by a subject. While we intended to pursue this classification problem, due to lack of precision for the 3-way classification problem, we decided to focus our efforts on feature engineering.

## Acknowledgment

The authors would like to thank Dr. Takako Fujioka of the Stanford CCRMA Lab, who provided us with access to an 80 gigabyte dataset from her 2014 experiment with the Rotman Research Institute in Toronto, ON. Dr. Fujioka has dedicated her career to studying the intersection of music and neuroscience and was incredibly helpful in understanding the data set and providing us with external resources.

## References

[1] Fujioka, T., Ross B., Trainor L. J., Beta-Band Oscillations Represent Auditory Beat and Its Metrical Hierarchy in Perception and Imagery. Journal of Neuroscience 11 November 2015, 35 (45) 15187-15198; DOI: http://dx.doi.org/10.1523/JNEUROSCI.2397-15.2015.

[2] Altenmller, E., Demorest, S.M., Fujioka, T.,et. al. (2012), Introduction to The Neurosciences and Music IV: Learning and Memory, Annals of the New York Academy of Sciences, 1252: 116. doi:10.1111/j.1749-6632.2012.06474.x.

[3] J. Jung, R. Bouet, C. Delpuech, et. al., The value of magnetoencephalography for seizure-onset zone localization in magnetic resonance imaging-negative partial epilepsy, Brain Oct 2013, 136 (10) 3176-3186; DOI: 10.1093/brain/awt213.

[4] B. Hosseinifard, M. Hassan Moradi, R. Rostami, Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal, Computer Methods and Programs in Biomedicine, Volume 109, Issue 3, March 2013, Pages 339-345, ISSN 0169-2607.

[5] J.E. Cardy, E.J. Flagg, W. Roberts, T.P.L. Roberts, Auditory evoked fields predict language ability and impairment in children, International Journal of Psychophysiology, Volume 68, Issue 2, May 2008, Pages 170-175, ISSN 0167-8760.

[6] J. Mellinger, G. Schalk, C. Braun, H. Preissl, et. al., An MEG-based braincomputer interface (BCI), NeuroImage, Volume 36, Issue 3, 1 July 2007, Pages 581-593, ISSN 1053-8119,

[7] M. Hallett, O. Bai and C. Bonin, "Predicting Movement: When, Which and Where," 2007 IEEE/ICME International Conference on Complex Medical Engineering, Beijing, 2007, pp. 5-7. doi: 10.1109/ICCME.2007.4381681/

[8] W. Wu, S. Nagarajan, Z. Chen, "Bayesian Machine Learning: EEG/MEG signal processing measurements," in IEEE Signal Processing Magazine, vol. 33, no. 1, pp. 14-36, Jan. 2016. doi: 10.1109/MSP.2015.2481559.

[9] Fujioka, T., Fidali B. C., Ross B. Neural correlates of intentional switching from ternary to binary meter in a musical hemiola pattern. Front. Psychol., 12 November 2014, http://dx.doi.org/10.3389/fpsyg.2014.01257.

[10] Talairach J, Tournoux P, Co-planar stereotaxic atlas of the human brain. Thieme, New York, 1998.

[11] "Temporal Lobe - The Brain Made Simple." Temporal Lobe - The Brain Made Simple. The Brain Made Simple, n.d. Web. 16 Dec. 2016.

[12] Kawakatsu, Masaki. "Application of ICA to MEG Noise Reduction." (n.d.): n. pag. Tokyo Denki University.