Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, July 23-28, 2012, Thessaloniki, Greece
Cambouropoulos E., Tsougras C., Mavromatis P., Pastiadis K. (Editors)

An Exploration of Tonal Expectation Using Single-Trial EEG Classification

Blair Kaneshiro,[*#] Jonathan Berger,[*] Marcos Perreau-Guimaraes,[#] and Patrick Suppes[#]

[*]*Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA, USA*
[#]*Center for the Study of Language and Information, Stanford University, Stanford, CA, USA*

ABSTRACT

We use a machine-learning approach to extend existing averaging-based ERP research on brain representations of tonal expectation, particularly for cadential events. We introduce pertinent vocabulary and methodology, and then demonstrate the use of machine learning in a classification task on single trials of EEG in a tonal expectation paradigm. EEG was recorded while participants listened to two-measure chord progressions that established expectation for resolution to the tonic. Cadential events included the tonic; repeated dominant; ^{b}II; and silence. Progressions were presented in three keys. Classifications were performed on single trials of EEG responses to the cadential events, with the goal of correctly identifying the label of the stimulus that produced the EEG response. Classification of the EEG responses by harmonic function of the cadential endings across keys produced classifier accuracies significantly above chance level. Our results suggest that the harmonic function of the stimulus can be correctly labeled in single trials of the EEG response. We show that single-trial EEG classification can additionally be used to identify task-relevant temporal and spatial components of the brain response. Using only the top performing time ranges or electrodes of the brain response produced classification rates approaching and even exceeding the accuracy obtained from using all time points and electrodes combined.

I. INTRODUCTION

There exists a wide range of research on tonal expectation, which uses, among other methods, averaged event-related brain potentials (ERPs) to describe human brain responses to various degrees of 'appropriateness' of particular chords in harmonic progressions (Leino, 2007). One approach uses short cadential formulae as stimuli. For example, Koelsch et al. (2003) used a five-chord progression ending in a cadence. Replacing the cadential tonic with a major chord built upon the lowered second degree of the scale elicited particular electric responses that enabled comparative differentiation of responses by age and gender.

Inspired by this method's encapsulation of real-world musical contexts to study syntactic irregularities, we designed an expanded stimulus set and applied an alternative method of data analysis, using single-trial classification rather than averaged ERPs to study the relationships between brain responses to stimuli based upon their function in a musical context.

The current study uses the same five-chord cadential progression paradigm with four possible events in the fifth chord: The perfect authentic cadential tonic, and three 'inappropriate' endings replacing the final tonic. These inappropriate endings include a repetition of the penultimate dominant, a major chord rooted on the flatted supertonic (both in root position), and silence.

The aberrant endings represent three types of violated expectation: Two anacolutha, one in which the dominant is repeated, thus delaying the expected resolution by negating the correspondence between metrical placement and harmonic rhythm; the second presenting two non-diatonic elements in the ^{b}II chord, thus interjecting patent surprise. The latter triad appears irregularly in tonal music, typically in first inversion; there it functions as a pre-dominant and, in that context, is referred to as a *Neapolitan*. The third aberrant ending not only denies the expected tonic, but also, by omitting the chord altogether, violates the expectation for a sounding event. We hypothesized that the expected tonic ending and the silent ending would elicit the most distinctive brain responses among the four possible endings.

We present a broad overview of the terminology and methodology of classification, with application toward the analysis of single trials of EEG data. We then demonstrate the use of such techniques by classifying single trials of brain responses to cadential events in short chord progressions. Single-trial classification is used to explore relationships between the responses to stimuli along dimensions of harmonic function and key. We also show that classification can provide a data-driven approach toward identifying relevant spatial and temporal components of the brain responses. For illustrative purposes, we compare responses between two participants.

II. EEG CLASSIFICATION

Machine learning, or *pattern classification*, is a statistical technique in which a model is fit to labeled observations, and then used to predict labels of new observations. This technique, commonly used for large, complex datasets, has many current applications in neuroscience, ranging from fMRI analysis (Norman, 2006) to neural prosthetics (Santhanam, 2006) to noninvasive EEG for motor control (Wolpaw, 2004). Only recently, however, has machine learning been applied to EEG with the goal of investigating music cognition (Schaefer, 2009; Vlek, 2011).

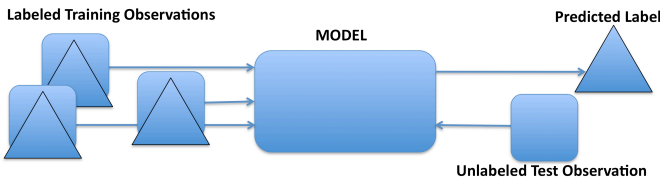

Figure 1. Overview of the classification process. The left portion of the figure shows training observations (squares) combined with their labels (triangles). These data are used to build the model (center). Unlabeled test observations (e.g., right square) are then passed to the model, which outputs a predicted label for each (e.g., right triangle).

Classification is *predictive*, meaning that the goal is to correctly label a future event. This is in contrast to a *descriptive* task, such as averaging-based ERP analysis, where the goal is to summarize or characterize a set of observations whose labels are already known. A simple illustration of the classification process is shown in Figure 1. Bishop (2006) and Hastie (2009) provide thorough coverage of the theory and implementation of machine-learning methods.

For the scope of this paper, we use the term *label* (or *class*, *category*) to refer to the specification of the stimulus, for example "tonic chord" or "dominant chord". We use the term *observation* (or *trial*) to refer to the EEG-recorded brain response to one presentation of a particular stimulus. We use the brain response to predict the label of the stimulus that generated the response. We use the term *channel* (or *electrode*) to refer to one sensor in the EEG montage.

By *single-trial* classification, we mean that the classification task is being performed on single trials of EEG data, with no averaging. We classify the *brain responses* to auditory stimuli, not the auditory stimuli themselves. The methods described here are optimized when the number of observations is balanced across label (i.e., an equal number of trials were collected for each category).

The classifier *training set* is the collection of labeled observations used to build the model, and the *test set* is the collection of unlabeled observations (or observations whose labels are withheld) on which the model is tested. Each observation is described by its *feature vector*, a set of numerical or qualitative descriptors. Feature vectors for EEG classification can be constructed of all samples from all electrodes pertaining to a trial. Many other representations are also possible, including groups of averaged trials; frequency-domain representations; or ICA or other source-space representations. For the current experiment, our initial feature vector is the time course of the data from all electrodes for a given single trial, concatenated into a one-dimensional vector.

A. Dimensionality Reduction

The feature vector for classification can be quite long. In the case of EEG data, concatenating C channels, each contributing N samples of response data, will produce a vector of length CN, a number that can easily enter the thousands. When the number of features is large compared to the number of trials, the data may have very small variance along some dimensions of the space, and the learning model may fail to fit the data in a way that can generalize to new trials. This is sometimes referred to as the "curse of dimensionality" (Tan, 2006, Chapter 2). Additionally, data obtained in channel space are highly correlated and present redundant information in the feature vector.

Dimensionality reduction is often achieved using linear matrix factorization methods. *Principal components analysis* (PCA) is one such method for reducing the dimensionality of a dataset. PCA linearly decomposes a matrix of data into a set of orthogonal vectors sorted in descending order of explained variance, so that the first principal component accounts for the most variance of any single principal component of the dataset (Witten, 2005, Chapter 7). *Singular value decomposition* (SVD) is one method of calculating principal components (Strang, 2003, Chapter 6).

B. Cross-Validation

In order to be able to test a learning model on new trials, a collection of test trials must be set aside. Using all trials to learn a model, and then using those same trials to test its performance, leads to potentially large bias in the estimation. *Cross-validation* is one commonly used technique to attenuate these effects for better estimation of prediction error (Hastie, 2009, Chapter 7).

For an S-fold cross-validation task, data are first randomized and partitioned into S subsets. A total of S classifications are then performed. For fold i, subset S_i will be used as the test (validation) set, while the remaining $S-1$ subsets will be used as the training set to build the model for that fold. The reported rate is the mean rate across the S classifications. Figure 2 illustrates training and test fold partitioning for ten-fold cross-validation.

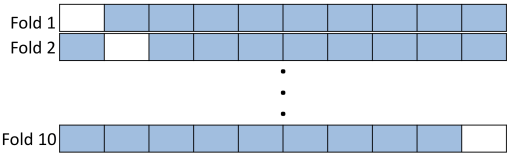

Figure 2. Ten-fold cross-validation. Each row represents one cross-validation fold. The darkened portions of each fold represent the training set – the partition of randomized trials, whose labels are known, that are used to build the model. The white portions represent the test set in each fold. Labels of test trials are withheld and later compared to the classifier's predicted labels to obtain the classifier accuracy.

C. Interpreting Classification Results

The most commonly reported result from a classification task is the *classification rate* (or *classifier accuracy*) – the percentage of correct classifications. Accuracy is computed as the number of correct classifications divided by the total number of classifications.

Accuracy provides a measure of the overall performance of the classifier, and this number is often useful, especially when high overall accuracy is the goal. However, accuracy does not fully describe the performance of the classifier. Further insight can be gained by analyzing the *confusion matrix* (or *conditional probability matrix*). This matrix plots the percentage or number of actual labels in the rows, against predicted labels in the columns.[1] Therefore, each entry M_{ij} in the confusion matrix M displays the percentage or number of observations belonging to category i that were predicted to be from category j. Rows of the confusion matrix sum to 1 (if plotting percentages) or to the total number of observations actually belonging to the class (if plotting number of observations).[2]

Elements on the diagonal ($i=j$) represent correct classifications, and classifier accuracy can be thus computed as the trace of the confusion matrix divided by the sum of all elements in the matrix. It follows, then, that a particular classification rate could result from a variety of confusion

[1] The transpose of this configuration is also commonly used, with actual labels as columns and predicted labels as rows.

[2] In the transpose case, this will be summation of columns.

matrices. Figure 3 shows an example of two confusion matrices for a three-class problem, which produce identical accuracies but perform quite differently.

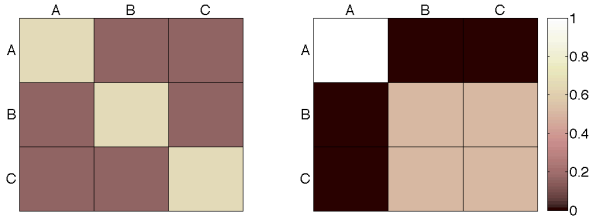

Figure 3. Two confusion matrices with identical accuracy but different performance for a three-class task. Lighter colors correspond to higher values. The monochromatic diagonal and monochromatic off-diagonal in left confusion matrix indicate that the classifier correctly labeled observations from each category with equal accuracy, and additionally mislabeled observations with equal confusion between the other two classes. The right confusion matrix shows that observations from category A were always correctly labeled, while the classifier was unable to distinguish between categories B and C above chance level. Both matrices produce accuracies of 66.7%.

One motivation for the use of single-trial classification in cognitive neuroscience experiments is that classification rates can be interpreted as measures of distance between the brain responses. These distance measures can then be compared to measures of distance between the stimuli that produced them. Brain responses that are more distinct from one another will classify with greater accuracy, whereas brain responses that are similar will have a greater tendency to be confused with one another by the classifier. Therefore, one application of classification is to investigate whether perceptual ratings of difference among stimuli are reflected in classification rates of the associated brain responses. Referring back to the right-hand confusion matrix in Figure 3, if brain responses to stimulus B are never confused with responses to stimulus A, yet are highly confused with responses to stimulus C, we would like to know whether B and C are judged to be highly similar along some perceptual dimension.[3]

The structure of the confusion matrix gives insights into the finer structure of the model, and thus into the structure of the brain responses to the stimuli (Suppes, 2009). Thus, the confusion matrix can be used to determine whether the model captured the structure of the stimuli (e.g., the perceptual distances between A, B, and C, as discussed above).

III. Experiment Methods

A. Stimuli

The stimulus set comprised four chord progressions in each of the three keys of C Major, F Major, and B Major. Each chord progression contained two measures in quadruple meter, with the first measure presenting quarter-note chords of I-vi-ii^6-V^7, setting strong expectation for a resolution to the tonic in the downbeat of the second measure. The event of interest in this experiment was the downbeat of the second measure. This beat contained of one of four endings: Tonic (expected); dominant; $^{\flat}$II; and silence. The remaining three beats of the second measure were always silence. We use the term *cadential endings* to refer to the set of all possible events in the cadence position in the stimulus set, whether or not the event served a proper cadential function within the rules of tonal harmony.

All chords were presented in root position, except for the third chord, which was presented in first inversion to imply the subdominant. The penultimate (V^7) chord at the end of the first measure contained the leading tone in the top voice, setting melodic expectation for upward resolution to the tonic. The tempo for all stimuli was 96 bpm (625 ms per beat). Figure 4 shows the C-Major portion of the stimulus set.

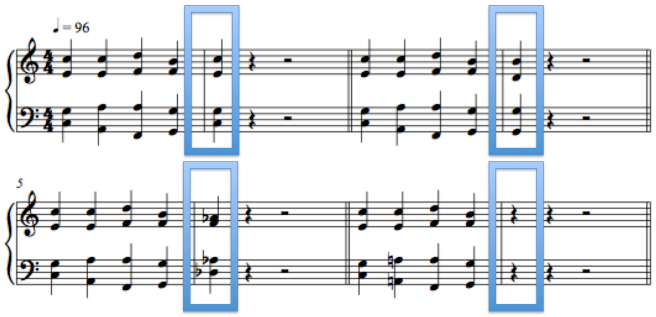

Figure 4. Two-measure chord progressions representing the C-Major portion of the stimulus set. The musical events of interest are the resolutions in the downbeat of each progression's second measure (boxed in rectangles). The four cadential endings are the tonic (top left), dominant (top right), $^{\flat}$II (bottom left), and silence (bottom right).

Stimuli were synthesized in Sibelius using a piano timbre. For each of the twelve progressions, the first 20 ms of the first beat and the last 100 ms of the fifth beat were linearly ramped up and down, respectively, using MATLAB, in order to eliminate clipping at the onset and for a smooth fade out into the remaining silent beats in the second measure.

B. Participants

We present data from two participants. Participant 1 (P1) is a 32-year-old male with ten years of childhood musical instrument training and no formal theory training. Participant 2 (P2) is a 32-year-old female with twenty years of musical instrument training and three years of formal theory training. Participant 2 additionally reports absolute pitch. Both participants report normal hearing and are right handed.

C. Procedure

Stimuli were presented in single-key blocks, with each block containing six progressions ending in the tonic and one of each progression with a deviant ending. Within the block, the two-measure progressions were presented continuously in order to maintain consistent musical meter. In order to establish and reinforce expectation for resolution to the tonic, each block began with at least two progressions resolving to the tonic, and each subsequent deviant-ending progression was preceded by at least one tonic-ending progression. Ordering and arrangement of the progressions within the block were randomized subject to these constraints. Each

[3] Low accuracy can also result from a low-performing classifier. The performance of the classifier should be substantiated prior to drawing conclusions from comparisons that yield low classifier accuracy.

block was followed by a break, the length of which the participant controlled via key press. The keys of the blocks were randomized within each twenty-minute subsession.

Data collection spanned twelve twenty-minute subsessions for each participant, completed in four separate EEG sessions, for a total of 108 blocks presented in each key.

Experiment sessions were conducted at the Suppes Brain Lab at Stanford University. Participants gave informed consent in accordance with Stanford IRB procedures prior to the first experiment session. Stimulus delivery was programmed using Neurobehavioral Systems Presentation software, and stimuli were presented through magnetically shielded Genelec speakers. Participants were seated in a darkened, acoustically shielded booth for experiment sessions, and were instructed only to listen passively to the stimuli while minimizing movement and attending to a visual fix point in the middle of a computer monitor in front of them.

EEG was recorded from 128 monopolar electrodes using the EGI GES 300 system. Data were acquired at a sampling rate of 1 kHz with range of 24 bits, referenced to the vertex.

D. Data Preprocessing

Data preprocessing and classifications were performed using MATLAB. After acquisition, data were highpass filtered using a fourth-order Butterworth filter to eliminate DC offset, then lowpass filtered with an eighth-order Chebyshev Type I filter and downsampled by a factor of 16. The final sampling rate was thus 62.5 Hz, with a Nyquist frequency of 31.25 Hz. Noisy channels were then identified and excluded from further analysis. Eye movements were removed from each subsession dataset using extended Infomax ICA from the EEGLAB Toolbox[4] (Jung, 2000; Perreau-Guimaraes, 2007), and data were converted to average reference. In order to present the classifier with a balanced number of trials from each of the twelve possible cadential endings, we chose one non-initial tonic ending at random, along with all three deviant endings, from each block of nine chord progressions.

The initial feature vector used for classification consisted of the brain response from channels 1 through 124, excluding noisy channels, with 39 samples from each channel (corresponding to 608 ms of the response, after downsampling). The time window for analysis was time-locked to stimulus onset. In total, we used 108 trials of each of the twelve possible cadential endings from each participant, for a total of 1,296 trials per participant.

E. Classifier Implementation

All classifications were performed within-subject using Linear Discriminant Analysis (LDA). PCA was implemented using SVD prior to the randomization for cross-validation. We used ten-fold cross-validation with a nested ten-fold cross-validation performed upon each outer fold's training subset. The number of principal components was optimized to a value *k* between 1 and 200, in the nested folds to avoid biasing the result (Suppes, 2009). The *k*-dimensional subset of the data was then used to classify the trials in the outer test fold. Multi-class classifications were performed using a one-against-all strategy (Bishop, 2006, Chapter 4).

The accuracy for a given classification task is the mean accuracy across the ten outer cross-validation folds. To reduce effects of partitioning in a particular ten-fold cross-validation, we performed each classification three times and report here the mean rate of the three ten-fold cross-validations. *P*-values were calculated under the null hypothesis of a Binomial distribution of the chance-level classification rates with parameters *nTest* (number of trials in a test partition) and 1/*nClasses* (number of categories in the classification).

F. Classification Tasks

The following classifications were performed on the data:

- Grouped by harmonic function (4-class)
- Grouped by harmonic function (pairwise)
- Grouped by tonal key area (3-class)
- No grouping (12-class)
- Grouped by harmonic function, channel-by-channel (4-class)
- Grouped by harmonic function, overlapping 80-ms frames (4-class)
- Grouped by harmonic function, best performing 80-ms frames (4-class)

The channel-by-channel classifications were performed on feature vectors constructed from each electrode's data separately. The overlapping 80-ms frame classifications were performed using frames of six consecutive samples of data from all channels, starting from the first sample and moving forward in three-sample increments (samples 1-6, 4-9, 7-12, etc.). For this configuration, a total of twelve classifications covered the entire range of 39 samples.

IV. Results

We present results from P1 and P2 individually, as well as the mean results. Mean results were calculated by averaging the results across the two participants, as opposed to performing the classifications on the EEG data from the two participants combined.

A. Grouped by Harmonic Function

Grouping trials by harmonic function (tonic, dominant, ^{b}II, and silence) across the three keys reduced the classification task to a four-class problem with chance-level accuracy of 25.0%. P1 responses classified with 41.3% accuracy ($p<10^{-4}$), and P2 with 43.9% accuracy ($p<10^{-5}$). Confusion matrices are shown in Figure 5.

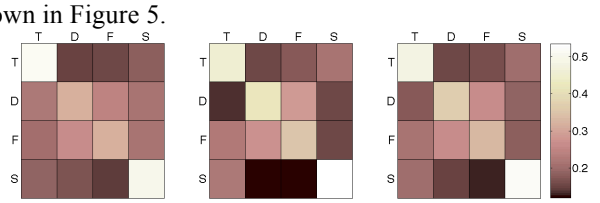

Figure 5. Four-class confusion matrices showing percentage of predicted labels for P1 (left), P2 (center), and mean (right). Brain responses were grouped according to the harmonic function of the stimuli – tonic ("T"), dominant, "D", ^{b}II ("F"), or silence ("S") – across all three keys. The confusion matrix for P1 suggests that the tonic and silent endings classified best, while dominant and ^{b}II endings did not classify well. The confusion matrix for P2 suggests that silence classified best, followed by the tonic and dominant, and then the ^{b}II.

[4] http://sccn.ucsd.edu/eeglab/

B. Pairwise Classification Grouped by Harmonic Function

The four-class classification task above was decomposed into classifications for every pairwise combination of labels, each with chance level of 50.0%. Classification rates and significance levels from this task are shown in Figure 6.

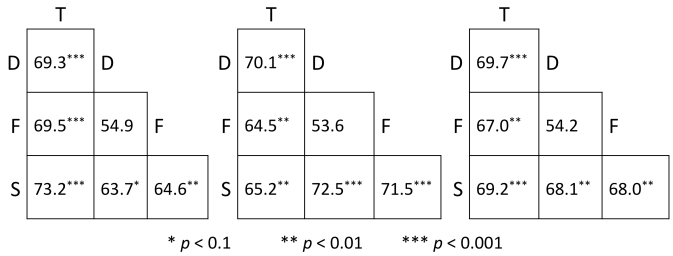

Figure 6. Pairwise classification rates and significance levels for P1 (left), P2 (center), and mean (right) of brain responses grouped by harmonic function of tonic ("T"), dominant ("D"), ^{b}II ("F"), and silence ("S") across all three keys. For P1, the best classifying pair was the tonic versus silent endings. For P2, the best classifying pair was the dominant versus silent; ^{b}II versus silent and tonic versus dominant show similar performance. For both participants, dominant versus ^{b}II responses did not classify significantly above chance level.

C. Grouped by Key

For comparison against the grouping by harmonic function, we grouped brain responses along the other dimension of key area of the progressions (C Major, F Major, and B Major). Classification rates for stimuli grouped by key were not significantly above chance level of 33.3%. P1 data classified at 38.3% (p=0.11), and P2 at 36.4% (p=0.25). Confusion matrices are shown in Figure 7.

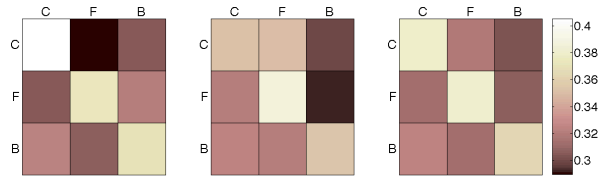

Figure 7. Three-class confusion matrices showing percentage of predicted labels for P1 (left), P2 (center), and mean (right). Stimuli were grouped by key area (C, F, or B Major). While the confusion matrices suggest differentiation of C Major for P1 and F Major for P2, classification results were not statistically significant for brain responses grouped along this dimension.

D. Twelve-Class Results

Responses to all twelve possible cadential endings (four endings per key, three keys) were classified, with chance accuracy of 8.3%. The classification rate for P1 was 15.1% (p<0.01) and for P2 was 13.6% (p<0.05). The twelve-class confusion matrices are shown in Figure 8.

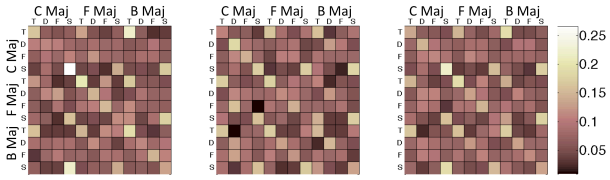

Figure 8. Twelve-class confusion matrices showing percentage of predicted labels for P1 (left), P2 (center), and mean (right). Since four cadential endings were presented per key, elements at positions incremented four units to the left or right of the main diagonal in a given row indicate stimuli of the same harmonic function in a different key.

E. Increased Spatial Resolution

Classification was performed for the four-class condition (grouped by harmonic function) meaning that a separate classification was performed on each electrode's data. Each electrode's classification rate is mapped to its location in Figure 9. The best performing electrode for P1 produced a rate of 42.8% (p<10^{-5}), and the best electrode for P2 produced a rate of 39.9% (p<10^{-4}).

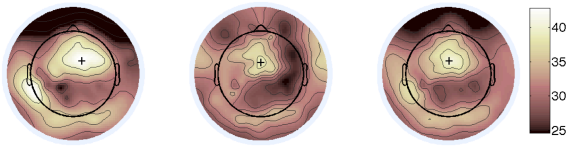

Figure 9. Map of rates for classifications performed on each electrode's data separately for P1 (left), P2 (center), and mean (right). The location of the highest performing electrode for each plot is marked with a cross ('+').

F. Increased Temporal Resolution

Classifying overlapping six-sample, 80-ms frames of the brain response using all electrodes produced the rates shown in Figure 10. Horizontal lines spanning the plots represent significance thresholds from p=0.1 to p=10^{-4}. The maximal rate for P1 was 41.0% (p<10^{-4}) in the seventh frame, 288-368 ms after stimulus onset. For P2, the maximal rate was 39.9% (p<10^{-4}) in the eighth frame, 336-416 ms after stimulus onset.

G. Grouping by Harmonic Function for Best Times

Using the results from the previous section, the four-class classification was performed once more, using only the samples from the three top-performing frames for each participant. For P1, these were frames 5-7, covering 192-368 ms after stimulus onset. The classification rate for P1 was 43.2% (p<10^{-5}) compared to chance level of 25.0%. The classification for P2 used frames 6-8, covering 240-416 ms after stimulus onset. The classification rate for P2 was 45.1% (p<10^{-6}). Confusion matrices for both participants are shown in Figure 11. As time frames do not align exactly between participants, we do not present mean results for this condition.

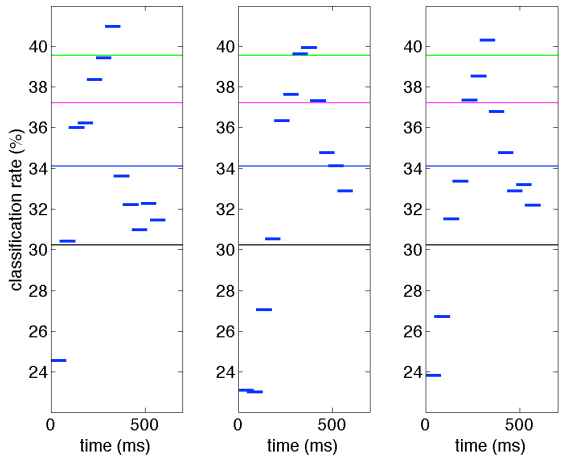

Figure 10. Six-sample classification rates plotted as a function of time for P1 (left), P2 (center), and mean (right). Short horizontal lines indicate the range of time of the brain response used for each classification. Horizontal lines spanning the entire plot indicate *p*-value thresholds for significance levels of 0.1 (black), 0.01 (blue), 0.001 (pink), and 10^{-4} (green). For both subjects, the first time frame does not classify above chance; the classification rate then increases quickly up to its peak between 300-400 ms, before falling to lower significance levels.

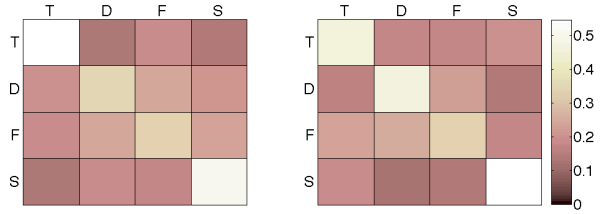

Figure 11. Four-class confusion matrices for classification grouped by harmonic function for the three best performing time frames for P1 (left) and P2 (right). The classification for P1 used data 192-368 ms after stimulus onset and achieved a rate of 43.2%. The classification for P2 used data 240-416 ms after stimulus onset and achieved a rate of 45.1%.

V. DISCUSSION

The classifier performed significantly above chance in the four-class tasks, suggesting that the differences between the four cadential endings in the EEG response are detectable at even the single-trial level, and that the response includes features that are invariant across key area. The relatively high values for the tonic and silent endings on the diagonals of the confusion matrices may indicate that these two stimuli are the most distinguishable in the brain response, while responses to the dominant and ^{b}II endings show high confusion with one another. The pairwise classification rates further support these results, in that the dominant versus ^{b}II condition was by far the lowest performing pair.

The pair rates point to another potential difference in processing between the two participants. For P1, the silent ending is best differentiated from the tonic, and less so from the ^{b}II and dominant. In contrast, for P2 the silent ending is the *least* differentiated from the tonic, classifying better against the other two endings.

The three-class rates for stimuli grouped by key did not produce statistically significant rates. This suggests that the cadential events were processed more by their function in the progression than by the surrounding key area, even for P2, who reported absolute pitch.

The twelve-class accuracy is less statistically significant than the four-class. However, the confusion matrix may suggest that misclassifications tended to occur among stimuli serving the same harmonic function in other keys (tonic endings in C, for example, being confused mainly with tonic endings in F and B). This is shown by the increased percentage of predicted labels occurring in increments of four units to the left or right of the main diagonal of the matrices.

While not a replacement for source analysis, classifications on individual electrodes did help us to identify the most task-relevant electrodes for the four-class classification. Additionally, the performance of the single best electrode for each participant was shown to approach the performance of the full set of electrodes, with the rate for P1 actually exceeding slightly the rate of the full set (42.8% for one electrode compared to 41.3% for the full set).

Performing the classification on smaller temporal frames of the full response identified useful temporal components. We saw that for both participants, the very early stages of the response did not contribute significantly to successful classification, while by 400 ms into the response, classification rates had reached high levels of statistical significance. Here, too, the best performing six-sample time frame for each participant was found to perform nearly as well as the entire 39-sample response. We also note that the time course of classification differed slightly between participants, with the rates for P1 reaching peak accuracy sooner after stimulus onset than the rates for P2.

The final four-class classification was performed using data from only the top three overlapping frames (12 samples) that were identified for each participant. Despite the lower dimensionality of this initial feature vector, the classifier obtained rates slightly higher than the rates from the full feature vector for both participants: 43.2% compared to 41.3% for the full set for P1; and 45.1% compared to 43.9% for P2.

The high accuracy resulting from classifying task-relevant spatial and temporal components of brain responses suggests that the full feature vector may not be necessary to achieve significant classifier performance. In fact, higher accuracy may be achieved by identifying the best performing electrodes and time points for the classification task at hand, and then performing the classifications using a reduced feature vector composed of those elements.

VI. CONCLUSION

We have shown that single-trial classification of EEG data can be used as an alternative to traditional ERP analysis as a means of exploring relationships between brain responses to musical stimuli. Using the classification approach, we showed that single trials of brain responses to cadential endings could be correctly labeled with the harmonic function of the stimulus with high statistical significance. Features of the brain responses to the different cadential endings generalize across multiple musical keys. The confusion matrix of classifier output was utilized to assess relationships between the brain responses. We have additionally shown that classification can be used to identify robust spatial and temporal features of the brain response, and that these features,

despite using only a subset of the electrodes or time points of the brain response data collected for a given trial, will classify with accuracy that approaches, and in some cases even exceeds, accuracy using the entire feature vector.

There exist a number of other possible applications of single-trial EEG classification. For instance, classifier output could be used as the basis for multidimensional scaling. Another possibility involves across-participants classification, either by aggregating data across participants (Schaefer, 2011) or by testing learned models on trials from new participants. The success we have seen thus far using single-trial EEG classification to investigate processing of harmonic function and expectation in brain responses encourages us to explore this topic further in larger-scale experiments.

ACKNOWLEDGMENT

The authors wish to thank Duc T. Nguyen for collecting the EEG data.

REFERENCES

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology, 37*, 163-178.

Koelsch, S., Grossmann, T., Gunter, T. C., Hahne, A., Schröger, E., & Friederici, A. D. (2003). Children processing music: Electric brain responses reveal competence and gender differences. *Journal of Cognitive Neuroscience, 15*, 683-693.

Leino, S., Brattico, E., Tervaniemi, M., & Vuust, P. (2007). Representation of harmony rules in the human brain: Further evidence from event-related potentials. *Brain research, 1142*, 169-177.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *TRENDS in Cognitive Science, 10*, 424-430.

Perreau-Guimaraes, M., Wong, D. K., Uy, E. T., Grosenick, L., & Suppes, P. (2007). Single-trial classification of MEG recordings. *IEEE Transactions on Biomedical Engineering, 54*, 436-443.

Santhanam, G., Ryu, S. I., Yu, B. M., Afshar, A., & Shenoy, K. V. (2006). A high-performance brain-computer interface. *Nature, 442*, 195-198.

Schaefer, R. S., Desain, P., & Suppes, P. (2009). Structural decomposition of EEG signatures of melodic processing. *Biological Psychology, 82,* 253-259.

Schaefer, R. S., Farquhar, J., Blokland, Y., Sadakata, M., & Desain, P. (2011). Name that tune: Decoding music from the listening brain. *Neuroimage, 56*, 843-849.

Strang, G. (2003). *Introduction to linear algebra* (3rd ed.). Wellesley, MA: Wellesley-Cambridge Press.

Suppes, P., Perreau-Guimaraes, M., & Wong, D. K. (2009). Partial orders of similarity differences invariant between EEG-recorded brain and perceptual representations of language. *Neural Computation, 21*, 3228-3269.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston, MA: Pearson Education.

Vlek, R. J., Schaefer, R. S., Gielen, C. C. A. M., Farquhar, J. D. R., & Desain, P. (2011). Shared mechanisms in perception and imagery of auditory accents. *Clinical Neurophysiology, 122*, 1526-1532.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

Wolpaw, J. R., & McFarland, D. J. (2004). Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proceedings of the National Academy of Sciences, 101*, 17849-17854.