Dynamical Hierarchical Self-Organization of Harmonic, Motivic, and Pitch Categories

Ricard Marxer, Piotr Holonowicz, Hendrik Purwins, Amaury Hazan
Music Technology Group, Universitat Pompeu Fabra,Barcelona, Spain
{rmarxer,pholonowicz,hpurwins,ahazan}@iua.upf.edu

Abstract

We introduce a generic model of emergence of musical categories during the listening process. The model is based on a preprocessing and a categorization module. Preprocessing results in a perceptually plausible representation of music events extracted from audio or symbolic input. The categorization module lets a taxonomy of musical entities emerge according to a cognitively plausible online learning paradigm. We show the advantages of using a conceptual clustering method in the musical domain. The system extracts multi-level hierarchies and can be tuned to clustering at various resolutions. The potential of the model is exemplified by exposing it to three different datasets resulting in musical categories of scales, motives, and harmonies consistent with music theory.

1 Introduction

Our cognitively inspired system shows emergence of musical categories during exposure to musical phrases. The simulation environment is based on cognitive grounds in the following respect: a) the cognitive representation is modeled in a self-organizing unsupervised fashion, b) input is processed online and causally, c) categories incrementally evolve and degenerate. The system is general enough to be applied to different musical dimensions. We demonstrate the capacity of the model in three musical domains: scales, melodic motives, and harmony.

Models of unsupervised formation of 'cortical maps' of tonalities have been based on linear and nonlinear methods of dimension reduction [1]. [2] use ART for dynamic clustering without extracting hierarchies. Hierarchical clustering in batch mode has been applied to chords by [3].

With regard to motivic categorization, most of the approaches focus on both segmentation and pattern similarity measures. Language and rule-based models have been introduced [4] in order to extract hierarchical or syntactical representations of the sequences, which are later compared using edit-distance and other domain-specific distances. Another set of models considers multidimensional input and searches for exact repetition along each different dimension [5]. Due to the nature of these methods, the input data must be quantized to obtain a correct behavior.

We present models and experiments that address these problems. The harmonic categorization experiment shows the strength of hierarchical clustering when the data is inherently structured. The motivic clustering experiment acts by continuously generating a hierarchical structure of distance relations between the motives and motivic prototypes that have been segmented. This allows to find motives at different similarity importance and to learn incomplete and fuzzy motivic patterns, and to be directly applicable to non-symbolic data.

2 System overview

The model presented in this paper is part of a larger system called the Music Projector. The system consists of the following units: 1) musical performance on keyboard or music scores parser, 2) preprocessing, 3) categorization, 4) pattern matching and completion, 5) visualization, sonification, and evaluation.

The goal of the system is to create a framework for building computational models of human music listening tasks. The output of each run can be visualized using a series of interactive plots allowing the user to explore the taxonomic hierarchies emerged during the process. This result is complemented by a series of measures on the classification in order to highlight certain behaviors of the models involved.

Our model consists of an implementation of the three first units of the Music Projector. One aim of the model is to demonstrate the emergence of musical hierarchical taxonomies, and the usefulness of the taxonomic information. The second objective of the presented system is to demonstrate how incremental hierarchical clustering can perform online learning with varying resolution. However the model is applicable to any sequential features extracted in an online manner from musical phrases.

3 Unsupervised Hierarchical Online Learning

In this section we present the different datasets and techniques used in our model as well as the methods of evaluation and visualization of the results.

3.1 Datasets

In the first example, we show how the system can be used to extract scales from audio recordings. We analyze *Hoquet*, No. 28 from CD II of *Les Voix du Monde* (Zemp), a piece from the Mbenzele pygmies in the Central African Republic recorded by Arom and Taurelle. The piece is a song described as vocal-instrumental hocketting. A woman's voice regularly alternates with blow sounds from a flute. It consists of a set of 3 main pitches (around 350, 397, 453 Hz), corresponding to the notes f^1, g^1, and a 'high' a^1 one short impulsive exclamation (between 674-720 Hz, around f^2) approximately one octave higher than the lowest main pitch, and one pitch (around 612 Hz, between d^2 and $e\flat^2$) always played by a flute. The flute tone is around a major sixth above the lowest main pitch. The melody is highly repetitive. For the sake of simplicity, this piece is suited to test the algorithm's ability to extract scales from audio.

The second dataset used in this paper is the Bach *Inventions* (cf. Figure 3). We choose the *Inventions* by J.S. Bach as an example of extensive motivic musical work. Again the Humdrum Kern encoding is used as representation that preserves the score information. This dataset allows to show the interest of conceptual clustering as a multiresolution clustering technique, capable of clustering segmented musical patterns and to create generalized patterns.

The third dataset processed is the Austrian monophonic folk songs from the Essen folk song collection [6] which will be used to demonstrate the emergence of harmonic representations. This is a collection of 94 scores, containing a total of 1170 instances (bars). The songs are in Humdrum Kern format containing all the information about bar division, tonality and meter. The input to the system has also been completed with metrical weight values of each event extracted using the timing information from the score. This dataset serves us to demonstrate the advantage of conceptual clustering over traditional clustering techniques, by revealing reasonable relations between the clusters that can be used for later processing.

3.2 Preprocessing Unit

This unit performs harmonic, motivic, and pitch preprocessing. The input is triplets of onset time, duration and pitch. In the case of the harmonic processor, incoming bar wise segmented signals are compressed into 12-dimensional vectors of pitch classes. Each component in that vector corresponds to the strength of one pitch class, determined by its frequency of occurrence, overall duration, intensity and metric weight. The key of the piece is extracted from the score and the pitch class profile is

transposed so that the first entry is always the key note. Finally the pitch classes are mapped to scale degrees.

The motivic processor applies a first order difference to deal with intervals of scale degrees. An interval $c - d\sharp$ would be encoded as a scale degree interval of 1.5. Each voice in the Bach *Inventions* is processed separately. The segmentation is performed automatically. Autocorrelation is applied to the 8 first half-bars (considered as an initial cue-finding phase), deriving the length of the motives. The beginning (anacrusis or not) of the motives is determined by the bar position of the first note. Therefore, a musically plausible segmentation can be performed based on the assumption that all motives have same length and that their relative position within a bar (anacrusis or not) is maintained for all motives within the piece. The durations are quantized. Each duration is represented by an adequate number of repeating 32th notes of the same tone (usually the shortest note in the piece).

The pitch preprocessor acts as an audio front end, allowing exposition to audio recordings. The task of the unit is to split the audio sample stream into unique perceptual events, such as notes. The splitting of the stream is done by the use of a custom onset detector based on the complex domain method presented in [7]. It generates inter-onset segments. Finally the events are analyzed and characterized by their duration and pitch which is estimated using the multi comb method in `aubiopitch` [8]. Non-pitched segments are filtered out.

3.3 Categorization Unit

The categorization unit implements clustering methods. The COBWEB [9] serves as a generic framework for hierarchical and incremental clustering. Each node of the tree represents a concept, characterized by the incremental means and standard deviations for each of the dimensions of the incoming feature vector. The edges of the structure represent taxonomic relations. Further works [10, 11] have proposed models to create, in an unsupervised manner, the concept tree based on the sequence of data presented, by the use of a heuristic function to be maximized. The heuristic function used in this paper is the numerical version of the standard category utility function used by Fischer and introduced by Gluck and Corter [12]. Such version of COBWEB was presented in [10] as COBWEB/3 and later extended by [11] as COBWEB/95. The version presented is COBWEB/3 and allows the input of real value attributes and control the specificity:

$$CU_{numeric} = \frac{\sum_k P(C_k) \sum_i \frac{1}{\sigma_{ik}} - \sum_i \frac{1}{\sigma_{iP}}}{4K\sqrt{\pi}} \tag{1}$$

where K is the number of classes, σ_{ik} is the bounded standard deviation for attribute i in class k, and σ_{iP} is the bounded standard deviation for attribute i in the parent node, i.e., the no-class membership case. This formulation adds a new parameter, the threshold at which to bound the standard deviation of the attributes per class σ_{ik} in order to avoid divisions by zero. The acuity parameter of the COBWEB, controls the resolution of discrimination, i.e., the minimum standard deviation taken into account.

3.4 Visualization and Evaluation Unit

Two main representations have been used to assess the quality of the model. A dendrogram allows a visualization of the taxonomic structure. The branch length represents the distance from the centroid of each node to its parent. The radius of the node markers are proportional to the quantity of instances. The darkness represents the distance between the spread vectors of nodes and their parents. Finally, the labels show the name of the node and the class of maximal frequency occurrence in the node when the groundtruth is available. The interactive scatter plot allows sonification and visualization of all the instances that have input the categorization unit. Each instance is represented by a marker identifying the cluster to which it belongs.

A new application of the Strahler number is introduced in this paper. The Strahler number is a measure of rooted tree complexity and has been lately revised to enable its application to non-binary trees [13]. This measure allows us to evaluate the dynamics of the hierarchical clustering, such as

the states of stabilization. The Strahler number calculation used in this paper is the following:

$$\sigma(s) = \begin{cases} 1 & \text{if } s \text{ is leaf} \\ \max\left(\sigma(s_i) + i\right)_{0 \leq i \leq k} & \text{otherwise} \end{cases} \quad (2)$$

where s_i are the children of s in an order such that $\sigma(s_i) \leq \sigma(s_j)$, if $i < j$.

Several approaches have addressed the problem of quantitative evaluation of unsupervised incremental clustering techniques. However in most cases the evaluation is based and compared to non-incremental techniques, diminishing focus on the interest and goal of incremental unsupervised approaches. In these cases, additionally to the content of clusters, the amount of them and their assignment to groundtruth classes must be evaluated.

In [9], three elements of the learning process are considered for the evaluation. The knowledge base, the performance task and the environment. The knowledge base analysis is performed here by comparing the dendrogram results to musicology studies. The performance element quantifies the inference ability of the model, and the environment evaluation tests the incremental nature of the model.

We present a new evaluation method, the incremental clustering F-measure. Differently to the other evaluations of incremental clustering techniques such as prediction accuracy or flexible prediction [9, 14, 11], our evaluation is focused on the incremental clustering performance in the context of transcription for pattern discovery tasks. In this situation the focus goes towards the dynamics of the system during training. Therefore the system is exposed to an annotated set of instances. Although the instances keep the original labels, these are not used during clustering. We then use a new measure of precision and introduce a measure of recall for the resulting cluster configuration, the leaves of the unpruned cluster hierarchy. This allows us to directly apply well known methods of evaluation such as the F-measure. In contrast to the prediction accuracy, we create a stable evaluation measure by considering several hypotheses of class assignment to the clusters, and integrate them by taking an average of their precision and recall measures weighted by each class frequency of occurrence.

In pattern discovery tasks two main situations of the clustering must be penalized: the mixture of instances of different classes into one same cluster and the distribuion of instances of the same class among different clusters. We consider three extreme configurations. The case in which all instances are gathered in a single cluster and the case in which every instance has a different cluster assigned to it, are trivial solutions that output no information and should receive the lowest evaluation possible. This is in contrast to the case where all the instances of each class are gathered in a different and unique cluster, which should be considered as the best solution.

Therefore, in a cluster configuration of a set of groundtruth classes C and a set of clusters K, we define the precision of a given class c in cluster k as the number of occurrences of classes other than c in cluster k, divided by the total number of occurrences of classes other than c. The recall corresponds to the number of occurrences of class c in cluster minus one k divided by the total number of occurrences of class c minus one, we substract one occurrence due to not considering clusters that do not have any instances of a given class:

$$P_{IC}(c,k) = \begin{cases} 1 & \text{if } \sum_{i \neq c}^{C} n_i = 0 \\ 1 - \frac{\sum_{i \neq c}^{C} n_{i,k}}{\sum_{i \neq c}^{C} n_i} & \text{otherwise} \end{cases} \quad (3)$$

$$R_{IC}(c,k) = \begin{cases} 1 & \text{if } n_c <= 1 \\ max(\frac{n_{c,k}-1}{n_c-1}, 0) & \text{otherwise} \end{cases} \quad (4)$$

where $n_{c,k}$ is the number of occurrences of class c in cluster k, n_c is the total number of occurrences of class c. The pairs of precision and recall of each cluster are integrated to acheive precision and recall measures per class. This is done by summing all the pairs of a class weighted by the number of occurences of the class in each of the clusters:

$$P_{IC}(c) = \frac{\sum_{k \in K} n_{c,k} P_{IC}(c,k)}{n_c} \quad R_{IC}(c) = \frac{\sum_{k \in K} n_{c,k} R_{IC}(c,k)}{n_c} \quad (5)$$

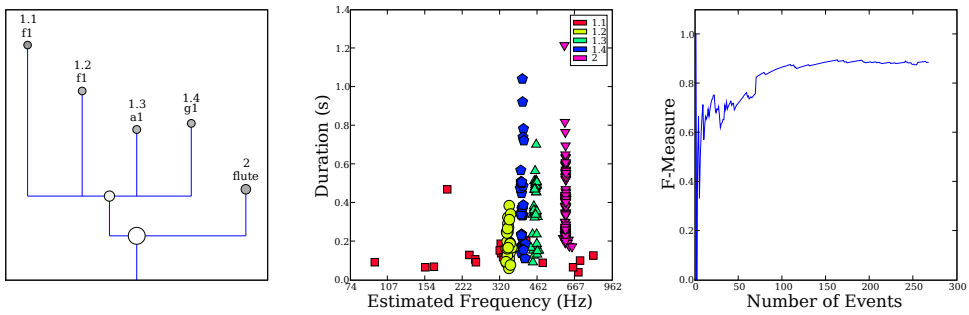

Figure 1: Pitch and duration category dendrogram (left), scatter plot (center) and F-Measure (right) after exposure to *Hoquet*.

The total precision and recall measures are the weighted sum of the per-class measures:

$$P_{IC} = \frac{\sum_{c \in C} n_c P_{IC}(c)}{\sum_{c \in C} n_c} \quad R_{IC} = \frac{\sum_{c \in C} n_c R_{IC}(c)}{\sum_{c \in C} n_c} \tag{6}$$

These new definitions of precision and recall solve some of the problems of applying the traditional measures on incremental clustering configurations, such as the trivial cases of all instances being classified into one single cluster or one cluster created for each instance. This new method allows for a higher range of precision and recall values, permiting more fine grained evaluations.

4 Results

Naturally, the evaluation of unsupervised methods is not straight forward as for classification. Since no hierarchical taxonomy of musical entities is available against which the results can be matched, the results are qualitatively evaluated in the light of music theory.

4.1 Pitch Categories

We present the results of our system exposed to *Hoquet*. In this run the acuity for COBWEB/3 has been set as close as possible to the minimum differences in input data. The parameter is set to 0.02 in order to capture differences of a quarter of a half-tone. The acuity of the duration is set to 0.2. For visualization and analysis purposes the dendrogram has been pruned to a threshold of 2. Node divisions with one of the nodes under the threshold are removed.

From the scatter plot in Figure 1 (center) we can see that the data set is essentially one dimensional (dominated by the frequency dimension). During clustering, categories for the 4 main pitches emerged (vertical accumulation of circles for Node 1.2, pentagons for Node 1.4, upright triangles for Node 1.3 and upside down triangles for Node 2). Node 1.1 consists of one large cluster stretching in horizontal direction with big overlap with the other ones. It contains noise and missclassifications. The fifth pitch of rather impulsive high notes around f^2 is not captured by a separate cluster, but rather mostly classified together with Node 2. In the dendrogram, Figure 1 (left), two nodes appear to represent the note f^1, Node 1.1 however is a mixture of errors from the feature extractor and its label is meaningless, this can be induced from the high distance from its parent node compared to the rest of nodes in the branch.

The F-Measure curve in Figure 1 shows a convergence of the score with a final value of 88.6 %. The convergence of the curve begins at around 12 events. The instability of the first events of the clustering are due to low initial recall values that are compensated later in the process by the COBWEB merging operations. This means that any operation performed on the output of such clustering algorithm will require tracking such operations in order to make use of such compensations.

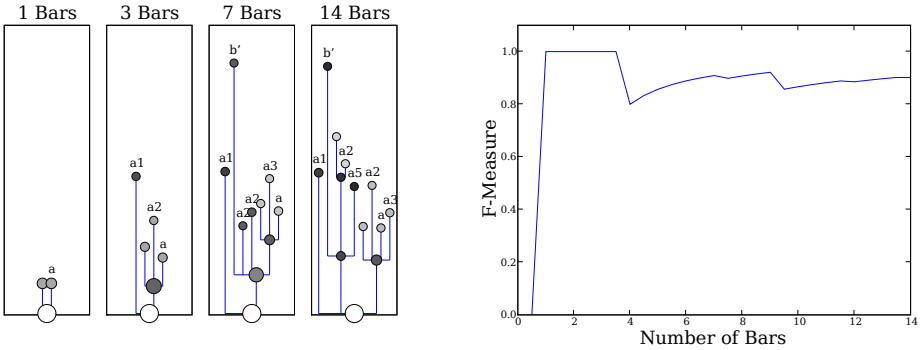

Figure 2: Motivic clustering dendrogram (left) and F-measure plot (right) after exposure to Bach *Inventio no IV*.

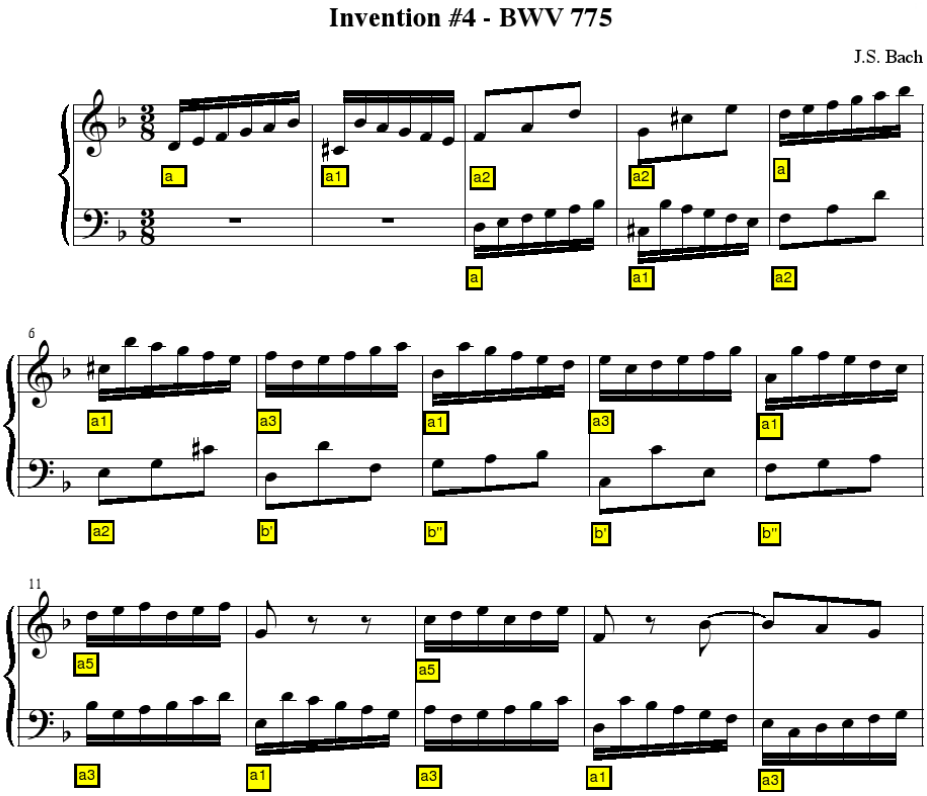

Figure 3: The first 15 bars of Bach's *Inventio no IV*. The bar-long motives from [15]'s analysis are indicated: a, a1, a2, a3, a5, b', b".

4.2 Motivic Categories

In this section, we first present the results of our system exposed to the Bach *Inventio no IV*, composed of a total of 28 instances (14 two-voice bars). The acuity parameter for the COBWEB/3 has been set to minimum possible variation of the input data. Since the input data are scale degree intervals the minimum possible interval is 0.5. We show the resulting dendogram at different instants of the run and the evolution of the F-Measure.

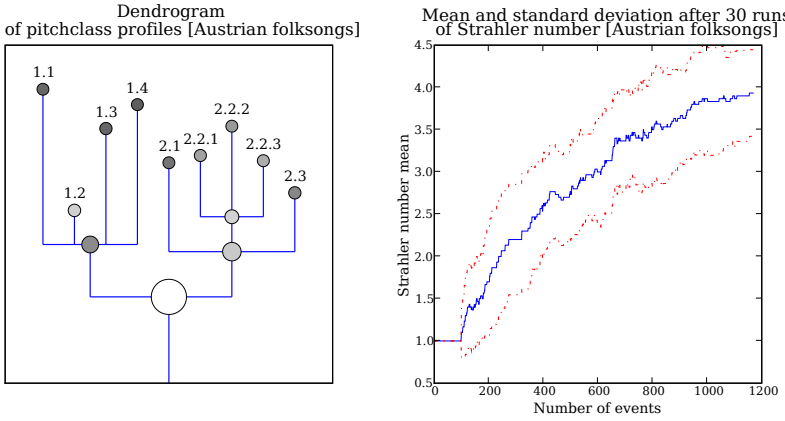

Figure 4: Harmonic clustering dendrogram after one run (left). Strahler number (mean and standard deviation) evolution after 30 exposure to shuffled versions of the Austrian folksong dataset (right).

In Figure 2 we observe that the clustering space evolves incrementally along the arrival of every new bar and the evaluation of it calculated by the use of F-Measure presented in Section 3.4. The dendrogram of Figure 2 (left) shows the cluster configuration at certain interesting points in time, such as the apparition of new annotated patterns, see Figure 3. At the first bar the model has created one cluster for *silence* Node 1 and one for motive *a* Node 2. After 3 bars the cluster configuration consists of two new nodes Node 1 and Node 2.2 representing motives *a1* and *a2* respectevily.

The F-Measure in Figure 2 shows a score of 90.2% according to the groundtruth data in Figure 3. The curve maintains a score between 80% and 100% during the sequence, presenting local minimas at clustering errors, such as in Bars 4, 8, 10 and 12, where instances of classes *b"*, *a3*, *a2* and *a1* are assigned to incorrect clusters respectively.

4.3 Harmonic Hierarchy

We present the results of our system exposed to the Austrian folksongs. In this run the acuity for COBWEB/3 has been set as close as possible to the minimum differences in input data. In this case the smallest difference in an attribute of a pitch profile is the value added by a shortest note in the lowest metrical weight position. By analyzing the scores we find that this would correspond to a 16th note in a position of metrical weight 1/4, assuming a maximum tempo of 120 and a time signature of 4/4, this would result in a smallest possible difference of 0.03125. To avoid over-complexity of the taxonomy we have set the value of the acuity to 0.05. For visualization and analysis purposes the dendrogram has been pruned to a threshold of 50 (approx. 5% of the total). Node divisions with one of the nodes under the threshold are removed.

To evaluate the importance of order of exposure we have repeated the experiment 30 times with different order of presentation of the songs processed. We analyzed the set of resulting hierarchies using the Strahler number complexity measure. We hereby present the entire result of one of the runs and the evolution of the mean bounded by the standard deviation of the Strahler number over all runs.

Figure 4 (left) shows the dendrogram generated after being exposed to 94 folk songs. Figure 5 entails a detailed representation of the centroids and spreads of the nodes. In the root node on top we see the profile of the diatonic major scale. This profile can be seen as the tonal essence of the folk song collection. The root node splits into two subsequent nodes: Node 1 (542 instances) characterizes the major triad which is dominant in Austrian folk song serving as a foil for the Viennese Classic extensively featuring the triad (especially the tonic one). Node 2 (628 instances) captures the diatonic major scale with deemphasized tonic and third note. Node 1 further splits into four nodes (third row of Figure 5): Each of the Nodes 1.1, 1.3, 1.4 represents one tone of the triad. This corresponds to the fact that long notes in the Austrian folk songs are mostly the triad notes, especially the tonic note

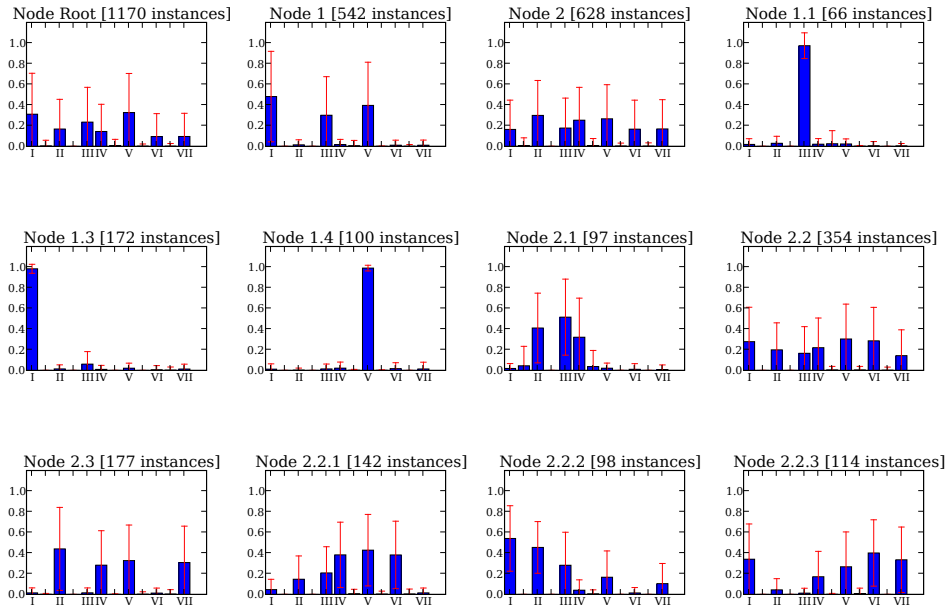

Figure 5: Harmonic clustering main hierarchy (centroid and spread of each cluster) after one exposure to the Austrian folksong dataset.

(e.g. in the end of a piece) reflected by the relatively large number of instances (172). Node 2 splits into three nodes. Node 2.1 reveals a feature in the Austrian folk songs: The embellishments around the third note including chromatic shades, i.e. some non-diatonic notes between I-II, and IV-V appear. Node 2.3 (177) clearly depicts the well-known dominant seventh chord. Node 2.2 contains various scale runs on the diatonic scale. The latter further splits into three subnodes: Subnode 2.2.1 (142 instances) originates in a typical pattern consisting of a scale fragment (or a third-fifth jump) mounting up to the sixth and then returning. Subnode 2.2.2 encodes a scale fragment I-II-III, however the presence of the V and VII is not clear. Similarly, Subnode 2.2.3 seems to describe a scale run from IV to I or serves a container of various substructures that reveal on the next-lower level. The Strahler number (Figure 4) describing the complexity of the tree shows a stabilization during exposure.

5 Conclusions and Future Work

We have presented a new approach to music listening modeling, by using a concept clustering technique for the incremental unsupervised online classification of musical events. This approach shows two advantages in face of other traditional clustering methods. The first advantage is the direct gain in information, given the taxonomy structure created by the hierarchical clustering process. This has been showed by exposing the model to harmonic data which is inherently hierarchical. Secondly the conceptual clustering allows classification at different resolution levels revealing clusters that otherwise would be harder to find, such as in the results described when exposing the system to motivic data. We have represented motivic patterns as incremental means and variances of each of the events. This allows to consider means highly prototypical when the variance is small. On the other hand, components with high variances indicate features that are less significant for that category.

Overall, the model demonstrates interesting advances over traditional techniques, however further work will be needed to turn the system into a totally unsupervised process. Bar divisions and metrical weights in harmonic categorization could be determined automatically to better process expressive timing. Also motive processing should be extended from equal length to variable length motives. Another limitation of the system is the use of a highly order-dependent categorization algorithm such

as COBWEB/3, but the choice of such algorithm was based on its generality, not on its performance, and now the system is prepared to accept any conceptual clustering technique.

Future research will involve studying further possibilities for incremental conceptual clustering techniques more robust to the order of exposure. Another goal is the automatic tuning of the acuity measure or other resolution control parameters by higher order cognition tasks such as pattern recognition and expectation [16]. Finally other timbral features will be included for better discrimination and classification of events.

Acknowledgments

This work is funded by EU Open FET IST-FP6-013123 (EmCAP) and the Spanish TIC project ProSeMus (TIN2006-14932-C02-01). Thanks to Hans Peter Reutter for advice in music theory [17] and to Ines Salselas for preparing musical data. Thanks are due to Perfecto Herrera for his comments.

References

[1] Hendrik Purwins. *Profiles of Pitch Classes - Circularity of Relative Pitch and Key: Experiments, Models, Computational Music Analysis, and Perspectives*. PhD thesis, Berlin University of Technology, 2005.

[2] I. Taylor and M. Greenhough. Modeling pitch perception with adaptive resonance theory artificial neural networks. *Connection Science*, 6(2-3):135–154, 1994. Journals Oxford Ltd.

[3] M. C. Mozer. Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multi-scale processing. *Connection Science*, 6(2 & 3):247–280, 1994.

[4] Rens Bod. A unified model of structural organization in language and music. *Journal of Artificial Intelligence Research*, 17:289–308, 2002.

[5] D. Meredith, K. Lemstrom, and G. Wiggins. Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4):321–345, 2002.

[6] H. Schaffrath. The essen folksong collection in the humdrum kern format, 1995.

[7] C. Duxbury, J. Bello, M. Davies, and M. Sandler. Complex domain onset detection for musical signals. *Proceedings Digital Audio Effects Workshop (DAFx)*, 2003.

[8] Paul Brossier. *Automatic Annotation of Musical Audio for Interactive Applications*. PhD thesis, Queen Mary University of London, UK, August 2006.

[9] Douglas H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Mach. Learn.*, 2(2):139–172, 1987.

[10] K. McKusick and K. Thompson. Cobweb 3: A portable implementation. *Technical Report No. FIA-90-6-18-2*, 1990.

[11] Jungsoon Yoo and Sung Yoo. Concept formation in numeric domains. In *CSC '95: Proceedings of the 1995 ACM 23rd annual conference on Computer science*, pages 36–41, New York, NY, USA, 1995. ACM Press.

[12] M. Gluck and J. Corter. Information, uncertainty, and the utility of categories. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, pages 283–287, 1985.

[13] David Auber, Maylis Delest, Jean-Philippe Domenger, Ph. Duchon, and Jean-Marc Fdou. New Strahler numbers for rooted plane trees. *Proceedings of the Third Colloquium on Mathematics and Computer Science*, pages 203–214, 2004.

[14] Kathleen B. McKusick and Pat Langley. Constraints on tree structure in concept formation. In *IJCAI*, pages 810–816, 1991.

[15] O. Lartillot and P. Toiviainen. Motivic matching strategies for automated pattern extraction. *Musicae Scientiae*, Disco.4A/RR:281–314, 2007.

[16] A. Hazan, P. Brossier, P. Holonowicz, P. Herrera, and H. Purwins. Expectation along the beat: A use case for music expectation models. In *Proceedings of International Computer Music Conference 2007*, Copenhagen, Denmark, 2007.

[17] Hans Peter Reutter. Approach to a melodic analysis of several Austrian folk songs seen from a music theoretical point of view. http://www.satzlehre.de, 2006.