DEVELOPING A STRUCTURALLY SIGNIFICANT REPRESENTATION OF MUSICAL AUDIO THROUGH DOMAIN KNOWLEDGE

Johanna Devaney and Claire Arthur
School of Music, Ohio State University, USA
{devaney.12, arthur.193}@osu.edu

ABSTRACT

Chroma representations of audio have proven useful for the analysis and classification of audio signals, but they do not distinguish between structurally important and unimportant notes. This paper demonstrates the utility of applying musical domain knowledge in developing a chroma-based representation of musical audio that highlights the harmonically significant parts of the signal. This is achieved by learning a mapping between standard chroma features and a chroma-like representation where only the harmonically-significant notes are retained. The goal of this mapping is to facilitate more robust music similarity and indexing. In order to demonstrate this robustness, this paper describes a music similarity experiment with a new open-source dataset of annotated theme and variations piano pieces.

1. INTRODUCTION

This paper presents a set of similarity-based retrieval experiments that demonstrate the utility of highlighting the harmonically significant parts of the musical surface by combining a proposed chord representation with note information in the symbolic domain and chroma features in the audio domain. The retrieval experiments use an annotated dataset of 6 Beethoven theme and variations, a subset of [1], for the task of predicting of a corresponding theme given a specific variation. In both the audio and symbolic domains, the chord-informed representations outperform the low-level features. This paper also describes a method for learning this chord representation in the audio domain using support vector machines. The learned representation outperforms the chroma features on 5 of 6 of the theme and variation sets, demonstrating the potential of integrating music theoretic domain knowledge in learning a representation of unlabeled data.

2. SIMILARITY EXPERIMENT

6 sets of theme and variations pieces used in the paper were transposed to C major/minor and the opening phrases of the theme and the first 6 variations were manually segmented. Whole phrases were chosen rather than a fixed number of notes because they follow a similar harmonic trajectory over the duration of the phrase from opening tonic to cadence, and thus are considered a unit in music theoretical terms. For the purposes of our experiments the harmonic analysis is represented by a 12-element vector, rather than chord labels. The advantage of this representation is that the relationship between the chords is explicit. For example the typical ordering of diatonic (major) chord labels is I-ii-iii-IV-V-vi-viio, which does not reveal that ii and IV are closer to one another, both in terms of common pitch classes, than IV and V. Once the features were derived, a similarity matrix was generated between each of the variations and each of the 6 themes. The score of the best dynamic programming path, found with Dan Ellis' MATLAB implementation [1], was used as a measure of similarity. The maximum objective theme was selected as the predicted source theme.

The symbolic domain is useful for validating that the structurally significant notes, as inferred from a harmonic analysis, are a more effective representation in the task of similarity-based retrieval than all of the notes in the musical surface, since the notes are more clearly defined in the symbolic domain than they are in the audio domain. Both the note and chord data are represented by 12 element vectors and were sampled at the rate of 24 vectors per quarter note, a division which allows for values as small as a triplet sixteenth note to be represented. In the audio domain, mapping from chroma to significant notes derived from the harmonic analysis approximates the symbolic task. Chroma features are a 12-element vector representing the amount of energy for each pitch-class (chroma) in the audio signal. Assuming the chord and chroma vectors are normalized to the same starting note/chroma, their structure is identical, allowing for direct refinement of the chroma by the chord vectors. The audio recordings were generated through MIDI using the standard piano timbre in the commercial notation software Finale [2], which allowed for the beat locations in the audio signal to be known *a priori*. Chroma vectors were calculated using Dan Ellis' MATLAB implementation [3], where instantaneous frequency estimates from a short-time Fourier transform of the audio are normalized and reduced to 12 bins, corresponding to the 12 chroma. Comparisons were made across all of the frame-wise chroma calculations

[1] http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw/

[2] http://www.finalemusic.com

[3] http://www.ee.columbia.edu/ dpwe/resources/matlab/chroma-ansyn/

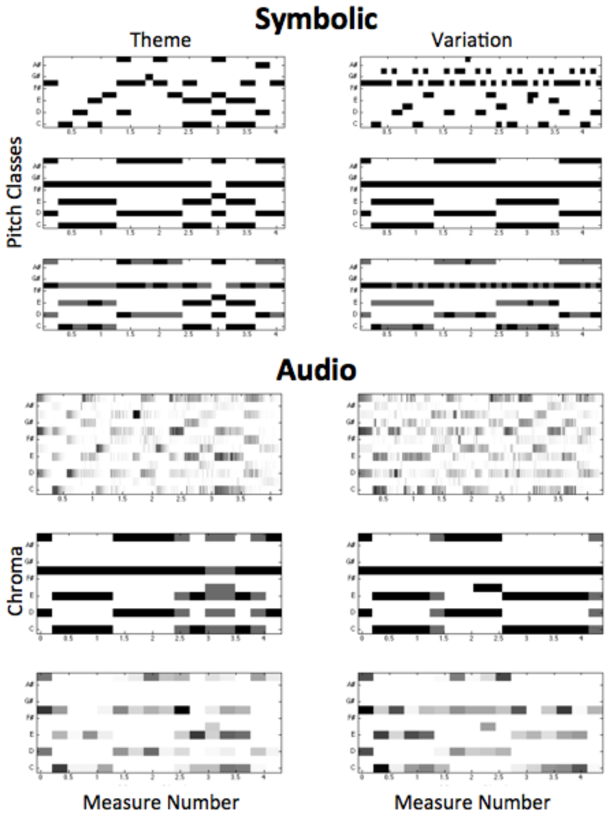

Figure 1. The plot on the left shows an example of the symbolic encoding and the plot on the right shows an example of the audio encoding. Within each plot, the top row shows the representation of the notes, the middle row shows the representation of the notes derived from the harmonic analysis, and the bottom shows the two combined.

(where the frame-size was set to approximate the size of the symbolic frames). For the harmonically significant chroma features, the chroma and harmonic analyses were averaged across each beat. In the symbolic domain, it is possible to work on a note-by-note basis, but since that information is not always available in the audio domain, a more generalized segmentation of the signal was used. Examples of the symbolic and audio representations of the data are shown in the top and bottom plots in Fig. 1.

In order to assess the performance of a learned representation on the same task, a set of 12 support vector machines (SVM) were trained for multi-label classification using the MATLAB based LIBLINEAR library [4] . For each beat, the input to the SVM was one element of the corresponding chroma vector (averaged over the duration of the beat from the frame-wise estimates). The output is the predicted value of whether that chroma is harmonically significant or not. The SVM was trained on the harmonically-significant chroma representation, using a leave-one-out paradigm across the 6 theme and variations sets. The output predictions were recombined into 12-element vectors for each frame of the variations and the same dynamic programming-based similarity procedure used above was run on the prediction for each variation.

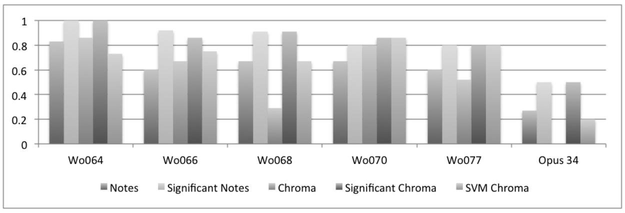

Figure 2. F1 scores on theme retrieval task for the first 6 variations of each piece for the symbolic (Notes and Significant Notes) and audio (Chroma, Significant Chroma, and SVM Chroma) tasks.

The results of the experiments are shown in Fig. 2. The F1 scores for the symbolic experiment show that there is a marked improvement in performance on the similarity retrieval task between the note and harmonically significant note representations. Likewise, the trend in the F1 scores for the audio experiment mirrors the trend in the symbolic task, with the harmonically-informed features consistently outperforming the naive ones. Overall, the SVM-predicted representation does not perform as well as the hand-coded harmonically significant chroma, but it does out perform the standard chroma representation for all but one set.

3. CONCLUSIONS

This paper evaluated the utility of using music theoretic domain knowledge, as expressed by a vector-based chord representation, on a similarity-based retrieval task in both the symbolic and audio domains. In these two domains, the vector-based chord representation outperformed the symbolic notes and standard chroma, respectively, on the task of identifying the corresponding theme for a given variation. This paper also provided a proof of concept use of this chord representation for learning a mapping of audio chroma features to structurally significant notes (chord-tones) using support vector machines. The learned representation was shown to be more effective than chroma features on the retrieval task for 5 of 6 theme and variations sets. This work addresses one of the main challenges for music similarity, namely, finding a way to emphasize the more structurally significant parts of the musical surface and/or de-emphasize the less significant parts.

4. ACKNOWLEDGEMENTS

This work was supported by a Google Faculty Research Award.

5. REFERENCES

[1] J. Devaney, C. Arthur, N. Condit-Schultz, and K. Nisula. Theme and variation encodings with roman numerals (TAVERN): A new data set for symbolic music analysis. In *Proceedings of the International Society of Music Information Retrieval conference*, pages 728–34, 2015.

[4] http://www.csie.ntu.edu.tw/ cjlin/liblinear/