PHRASE-LEVEL AUDIO SEGMENTATION OF JAZZ IMPROVISATIONS INFORMED BY SYMBOLIC DATA

Jeff Gregorio and Youngmoo E. Kim
Drexel University, Dept. of Electrical and Computer Engineering
{jgregorio, ykim}@drexel.edu

ABSTRACT

Computational music structure analysis encompasses any model attempting to organize music into qualitatively salient structural units, which can include anything in the heirarchy of large scale form, down to individual phrases and notes. While much existing audio-based segmentation work attempts to capture repetition and homogeneity cues useful at the form and thematic level, the time scales involved in phrase-level segmenation and the avoidance of repetition in improvised music necessitate alternate approaches in approaching jazz structure analysis. Recently, the Weimar Jazz Database has provided transcriptions of solos by a variety of eminent jazz performers. Utilizing a subset of these transcriptions aligned to their associated audio sources, we propose a model based on supervised training of a Hidden Markov Model with ground-truth state sequences designed to encode melodic contours appearing frequently in jazz improvisations. Results indicate that representing likely melodic contours in this way allows a low-level audio feature set containing primarily timbral and harmonic information to more accurately predict phrase boundaries.

1. INTRODUCTION

Music structure analysis is an active area of research within the Music Information Retrieval (MIR) community with utility extending to a wide range of MIR applications including song similarity, genre recognition, audio thumbnailing, music indexing systems, among others. Musical structure can be defined in terms of any qualitatively salient unit, from large scale form (e.g. intro, verse, chorus, etc.), to melodic themes and motifs, down to individual phrases and notes.

Paulus [8] categorizes existing approaches to audio-based structural analysis according to perceptual cues assumed to have central importance in determination of structure, namely into those based on *repetition*, *novelty*, and *homogeneity*. A music structure analysis task typically involves a boundary detection step, where individual sections are assumed to be homogeneous, and transitions between sections associated with a high degree of novelty. Novelty is assumed to be indicated by large changes in one or more time-series feature representations that may correspond to perceptually-salient shifts in timbre, rhythm, harmony, or instrumentation. Predicted boundaries can then be used to obtain segments which can be grouped according to similarity in the employed feature space(s). Alternatively, repetition-based methods may be used to identify repeated segments and boundaries directly.

Due to the difficulty in reliabily estimating individual note onsets and pitches, much existing work on music segmentation at the phrase level has been limited to single instruments in the symbolic domain. In a meta-analysis of symbolic phrase segmentation work, Rodríguez López [7] showed that two of the best performing rule-based models in comparative studies include Cambouropoulos's Local Boundary Detection Model (LBDM) [2] and Temperley's *Grouper* [11]. Both relate to Gestalt discontinuity principles, placing phrase boundaries using heuristics derived in part from features of consecutive note onset times, including inter-onset intervals (IOI) and offset-onset intervals (OOI). The LBDM model additionally uses pitch contour information, assuming discontinuity strength is increased by large inter-pitch intervals (IPI). *Grouper* also incorporates knowledge of metrical context and assumes a prior distribution of phrase lengths.

The proposed work focuses on musical structure at the phrase level, specifically identification of phrase boundaries from audio signals. Though we do not directly predict note onsets, durations, or pitches available in symbolic representations, we take advantage of audio-aligned MIDI transcriptions in the supervised training of a Hidden Markov Model (HMM). Using a primarily timbral and harmonic audio feature represenation, we hope to aid in the prediction of phrase boundaries by exploiting correlations between timbral/harmonic cues and common melodic phrase contours represented in the dataset.

2. MOTIVATION

In the audio domain, most existing structural segmentation work attempts to model to large scale form. The self distance matrix (SDM) is a useful representation in this modality, where entries $SDM(i,j) = d(\mathbf{x_i}, \mathbf{x_j})$ represent the distance between all combinations of feature vectors $\mathbf{x_i}$ and $\mathbf{x_j}$ by some distance metric d. This representation lends itself well to identifiable patterns associated with ho-

© Jeff Gregorio and Youngmoo E. Kim. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Jeff Gregorio and Youngmoo E. Kim. "Phrase-level Audio Segmentation of Jazz Improvisations Informed by Symbolic Data", 17th International Society for Music Information Retrieval Conference, 2016.

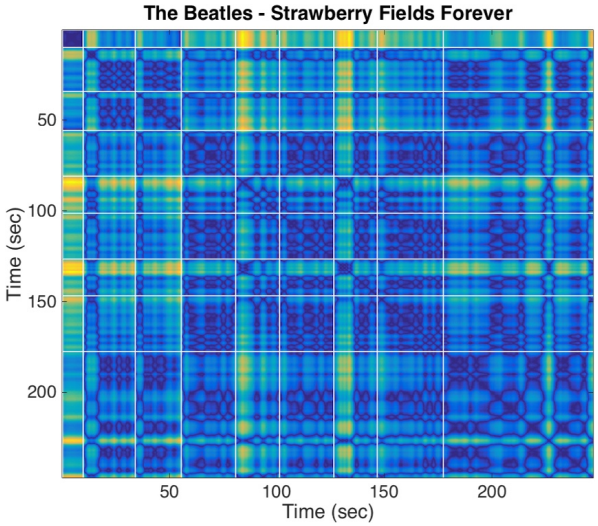


Figure 1. Self-distance matrix with form-level annotations plotted as white lines. Section boundaries often coincide with large blocks of relatively small distance, and some repetitions can be seen as stripes parallel to the main diagonal.

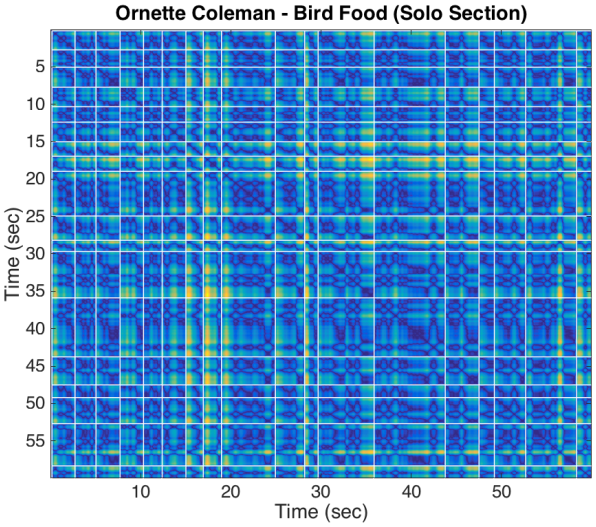


Figure 2. Self-distance matrix with annotated phrase boundaries plotted as white lines. Note the absence of off-diagonal striping patterns indicative of repetition, and the infrequent occurrence of large homogeneous blocks over the duration of entire phrases.

mogeneity, novelty, and repetition principles. Homogeneity within a section is generally associated with low-valued blocks representing small distance, novelty can be seen in the form of transitions between low and high value, and repetition manifests as stripes of low value parallel to the main diagonal. Figure 1 shows an example SDM computed using the timbral and harmonic feature space described in Section 4. Note this matrix is smoothed by averaging distance values from multiple frames around each index, as described in [8], but hasn't been filtered or beat-aligned to enhance repetition patterns, though some are visible.

When attempting audio segmentation at the phrase level, overall feature space homogeneity within single segments may be an unsafe assumption given the shorter time scales involved, in which a performer might employ expressive modulation of timbre. Furthermore, while melodic ideas in a jazz improvisation may be loosely inspired by a theme, extended repetition is usually avoided in favor of maximally unique melodies within a single performance. This context suggests that repetition-based approaches useful for identifying large-scale forms and themes may be inappropriate. Figure 2 shows an example SDM computed at the same resolution as Figure 1 over 60 seconds of a jazz improvisation. Note that while this SDM contains block patterns associated with homogeneity, they don't necessarily align well with entire phrases. This SDM is also almost completely missing any identifiable repetition patterns.

There does exist, however, some degree of predictability in jazz phrase structure that an ideal model should exploit, albeit across a corpus rather than within a single track. We propose a system based on supervised training of a Hidden Markov Model with a low-level audio feature set designed to capture novelty in the form of large timbral and harmonic shifts indicative of phrase boundaries. Existing HMM approaches have included unsupervised training, with a fixed-number of hidden states assumed to correspond to form-level sections [9, 10], instrument mixtures [5], or simply a mid-level feature representation [6]. Our model differs from existing HMM-based approaches in that it attempts to represent common elements of jazz phrase structure directly in the topology of the network, where the ground-truth state sequences are derived from inter-pitch intervals in audio-aligned transcriptions. Likely sequences of predicted states should therefore correspond to melodic contours well-represented in the training data, aiding in the detection of phrase boundaries.

3. DATASET AND PREPROCESSING

In 2014, the first version of the Weimar Jazz Database (*WJazzD*) was released as part of the larger Jazzomat Research Project [1]. The database contains transcriptions of monophonic solos by eminent jazz performers, well representing the evolution of the genre over the 20th century. The database was later expanded to include 299 solo transcriptions from 225 tracks, 70 performers, 11 instruments (soprano/alto/tenor/tenor-c/baritone sax, clarinet, trumpet, cornet, trombone, vibraphone, and guitar) and 7 styles (Traditional, Swing, Bebop, Hardbop, Cool, Postbop, and Free Jazz). The transcriptions were initially generated using state-of-the-art automatic transcription tools and manually corrected by musicology students. In addition to the transcriptions, the database contains a rich collection of metadata and human labels including phrase boundaries, underlying chord changes, form-level sections, and beat locations.

3.1 MIDI to Audio Alignment

The lack of availability of the original audio tracks used as source material for the database's transcriptions presents some difficulty in taking full advantage of the possibilities for supervised machine learning methods using acoustic features. Toward this end, we were able to obtain 217 of 225 tracks containing the solo(s) as transcribed. Metadata available in *WJazzD* indicate the starting and ending timestamps of the solo sections at a 1-second resolution, which is insufficient for determining ground truth for phrase boundaries associated with note onset times. Additionally, pulling audio files from various sources introduces further uncertainty, as many tracks appear on both original releases and compilations which may differ slightly in duration or other edits.

To obtain ground truth, we trim the original tracks according to provided solo timestamps and employ a tool created by Dan Ellis [4] which uses Viterbi alignment on beat-tracked versions of original audio and resynthesized MIDI to modify and output an aligned MIDI file. Upon inspection, 90 extracted solos produced a suitable alignment that required minimal manual corrections. We parse the database and handle conversion to and from MIDI format in Matlab using the MIDI Toolbox [3], making extensive use of the convenient note matrix format and pianoroll visualizations.

4. AUDIO FEATURES

To represent large timbral shifts, we use spectral flux and centroid features derived from the short-time Fourier transform (STFT), and spectral entropy derived from the power spectral density (PSD) of the audio tracks, sampled at $22050Hz$ with a FFT size of 1024 samples, Hamming windowed, with 25% overlap. Due to our interest in the lead instrument only, features are computed on a normalized portion of the spectrum between $500 - 5000Hz$ to remove the influence of prominent bass lines while preserving harmonic content of the lead instrument.

We also compute two features based on Spectral Contrast, a multi-dimensional feature computed as the decibel difference between the largest and smallest values in seven octave bands of the STFT. Since the resolution of this feature is dependent on the number of frequency bins in each octave, we use a larger FFT of size 4096, which gives meaningful values in four octaves above $800Hz$ without sacrificing much time resolution. The first feature reduces spectral contrast to one dimension by taking the mean difference of all bands between frames. Observing that large positive changes in spectral contrast correlate well with annotated phrase boundaries, we half-wave rectify this feature. The second feature takes the seven-dimensional Euclidian distance between spectral contrast frames.

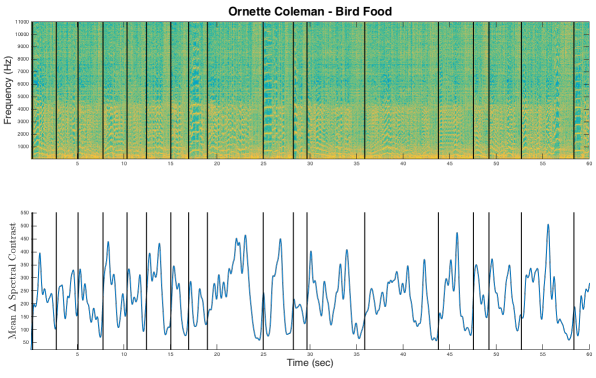


Figure 3. Mean positive difference between spectral contrast frames, plotted against annotated phrase boundaries.

Finally, we include a standard Chromagram feature, which is a 12-dimensional feature representing the contribution in the audio signal to frequencies associated with the twelve semitones in an octave. While the chromagram includes contributions from fundamental frequencies of interest, it also inevitably captures harmonics and un-pitched components. Noting that even precise knowledge of absolute pitch of the lead instrument would be uninformative in determining whether any note were the begining of a phrase, we collapse this feature to a single dimension by taking the Euclidian distance between frames, with the intention of capturing harmonic shifts that may be correlated with phrase boundaries and melodic contours.

All features are temporally smoothed by convolving with a gaussian kernel of 21 samples. All elements in the feature vectors are squared to emphasize peaks. We then double the size of the feature set by taking the first time difference of each feature, which amounts to a second difference for the spectral contrast and chroma features. Later, when evaluating, each feature in the training and testing sets is standardized to zero mean and unit variance using statistics of the training set features.

5. MELODIC CONTOUR VIA HMM

Hidden Markov Models represent the joint distribution of some hidden state sequence $\mathbf{y} = \{y_1, y_2, ..., y_N\}$ and a corresponding sequence of observations $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$, or equivalently the state transition probabilities $P(y_i|y_{i-1})$ and emission probabilities $P(\mathbf{x_i}|y_i)$.

HMMs have been used in various forms for music structure analysis, lending well to the sequential nature of the data, with hidden states often assumed to correspond to some perceptually meaningful structural unit. Unsupervised approaches use feature observations and an assumed number of states as inputs to the Baum-Welch algorithm to estimate the model parameters, which can then be used with the Viterbi algorithm to estimate the most likely sequence of states to have produced the observations.

Paulus notes that unsupervised HMM-based segmentation tends to produce unsatisfactory results on form-level segmentation tasks due to observations relating to individual sound events [8], a shortcoming which has led to observed state sequences being treated as a mid-level representations in subsequent work. We revisit HMMs as a segmentation method specifically for phrase-level analysis due to the particular importance of parameters of individual sound events rather than longer sections. Specifi-

cally, we postulate that phrases drawn from the jazz vocabulary follow predictable melodic contours, and incorporating ground-truth knowledge of the distribution and temporal evolution of these contours as observed in a large dataset of phrases through supervised training may help in identification of phrase boundaries.

6. EXPERIMENTS

We first evaluate a 2-state HMM, with states $y_i \in \{0, 1\}$ corresponding to the absence or presence (respectively) of the lead instrument during the i^{th} audio frame. Though a 2-state graphical model is trivial and offers no advantages over any other supervised classification method, we include it here simply as a basis for comparison with the multi-state models to evaluate the efficacy of adding states based on ground-truth pitch contour.

To estimate an upper bound on expected performance of our audio-based models, we evaluate two symbolic segmentation models using features of precisely known note pitches and onset times. First, we evaluate the Local Boundary Detection Model (LBDM) implementation offered by the MIDI Toolbox [3]. The LBDM outputs a continuous boundary strength, which we use to tune a boundary prediction threshold for maximum f-score via cross validation. Second, we train a 2-state HMM, where the model state y_i then corresponds to the i^{th} note rather than the i^{th} audio frame, and takes the value 1 if the note is the first in a phrase, and 0 otherwise. Observations $\mathbf{x_i}$ similarly correspond to features of individual note events including the IOI, OOI, and IPI. Results of the symbolic segmentation models are shown in Table 1(a).

To directly encode melodic contour in the network topology for the multi-state, audio-based HMMs, we extract ground-truth state sequences based on quantization levels of the observed inter-pitch interval (IPI) in the transcription. The following indicate the state of the network following each IPI, where the state remains for the duration of the note, rounded to the nearest audio frame:

$$\textbf{5-State} \quad y_i = \begin{cases} 0 & \text{lead instrument absent} \\ 1 & \text{first phrase note} \\ 2 & IPI < 0 \\ 3 & IPI = 0 \\ 4 & IPI > 0 \end{cases}$$

$$\textbf{7-State} \quad y_i = \begin{cases} 0 & \text{lead instrument absent} \\ 1 & \text{first phrase note} \\ 2 & IPI < -5 \\ 3 & -5 \leq IPI < 0 \\ 4 & IPI = 0 \\ 5 & 0 < IPI \leq 5 \\ 6 & IPI > 5 \end{cases}$$

The 5-state model simply encodes increasing/decreasing/unison pitch in the state sequence. The 7-state model further quantizes increasing and decreasing pitch into intervals greater than and less than a perfect fourth. Each HMM requires a discrete observation sequence, so the 10-dimensional audio feature set described in Section 4 is discretized via clustering using a Gaussian Mixture Model (GMM) with parameters estimated via Expectation-Maximization (EM).

We note that in the solo transcriptions, there are many examples of phrase boundaries that occur between two notes played legato (i.e. the offset-onset interval is zero or less than the time duration of a single audio frame). When parsing the MIDI data and associated boundary note annotations to determine the state sequence for each audio frame, in any such instance where state 1 is not preceded by state 0, we force a 0-1 transition to allow the model to account for phrase boundaries that aren't based primarily on temporal discontinuity cues.

7. RESULTS

Evaluation of each network is performed via six fold cross-validation, where each fold trains the model on five styles as provided by the *WJazzD* metadata, and predicts on the remaining style. We note that *WJazzD* encompasses seven styles, but the 90 examples successfully aligned to corresponding audio tracks did not include any traditional jazz. Though the sequence of states predicted by the model include the contour-based states, our reported results only consider accuracy in predicting a transition to state 1 in all cases.

Precision, recall, and f-score metrics reported in form-level segmentation experiments typically consider a true positive to be a boundary identified within 0.5 and 3 seconds of the ground truth. Considering the short time scales involved with phrase-level segmentation, we report metrics considering a true positive to be within one beat and one half beat, as determined using each solo's average tempo

Model	P_n	R_n	F_n
LBDM	0.7622	0.7720	0.7670
HMM, 2-State	0.8225	0.8252	0.8239

(a) Symbolic models

Model	P_{1B}	R_{1B}	F_{1B}
HMM, 2-State	0.6114	0.5584	0.5837
HMM, 5-State	0.5949	0.6586	0.6251
HMM, 7-State	0.6116	0.6565	0.6333

(b) Audio models, true positive within one beat of annotation

Model	$P_{0.5B}$	$R_{0.5B}$	$F_{0.5B}$
HMM, 2-State	0.4244	0.3876	0.4052
HMM, 5-State	0.4039	0.4472	0.4245
HMM, 7-State	0.4212	0.4521	0.4361

(c) Audio models, true positive within half beat of annotation

Table 1. Segmentation results

annotation provided by *WJazzD*. For reference, the mean time per beat in the 90 aligned examples is 0.394 seconds, with a standard deviation of 0.178 seconds. All beat durations were less than 1 second.

We report precision, recall, and f-score computed over all examples, all folds, and 5 trials. Reported results use 30 Gaussian components as discrete observations in the audio-based models, and 5 components for the symbolic model, and are summarized in Table 1. For greater insight into the model's performance in different stylistic contexts, we also present the cross-validation results across the six styles in Figure 4.

8. DISCUSSION

ANOVA and post-hoc analysis reveals both multi-state models yielding increased recall over the 2-state model ($F_{2,1332} = 30.62$, $p < 10^{-13}$), and increased f-score ($F_{2,1332} = 11.28$, $p < 10^{-4}$) with no significant difference in precision. Interestingly, the most significant recall increases from addition of the melodic contour states within styles include hardbop ($F_{2,297} = 15.68, p < 10^{-6}$), postbop ($F_{2,432} = 12.22$, $p < 10^{-5}$), and swing styles ($F_{2,177} = 6.73, p < 10^{-3}$).

These increases in recall within a style also correlate well with a high proportion of occurrences of phrase boundaries with no temporal discontinuity. These account for 22% of all phrase boundaries in hardbop, 18% in postbop, and 29% in swing, while accounting for 17%, 7%, and 4% in bebop, cool, and free jazz, respectively. We believe this suggests that incorporating ground-truth melodic contour allows the model to account for the relationship between contours indicative of phrase boundaries and their associated timbral and harmonic shifts.

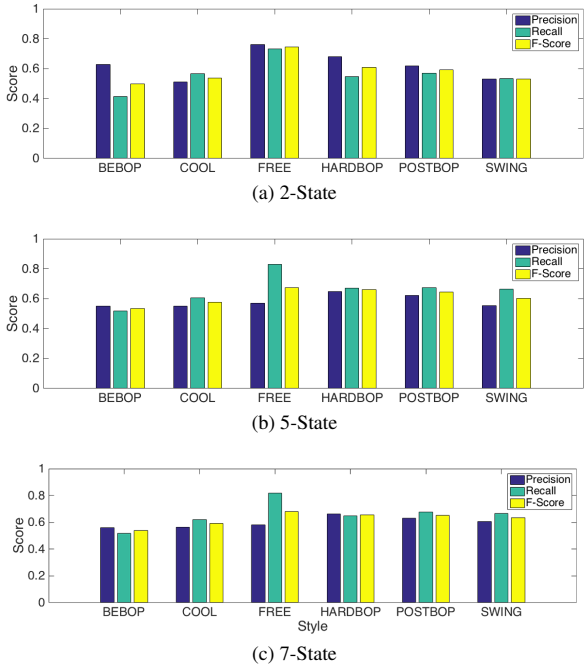
Manual inspection of segmentation results tend to reinforce this idea, as shown in Figure 5. The 2-state model fails to identify four phrase boundaries preceded by very small inter-onset intervals (6th, 15th, 18th, and 21st phrases), while the 7-state model correctly identifies three (6th, 18th, and 21st), at the cost of some tendency toward over-segmentation (in this case).

9. CONCLUSIONS

Evaluation of a 2-state HMM established a baseline phrase segmentation accuracy by detecting the presence or absence of the lead instrument, which presents some difficulty in predicting phrase boundaries based on harmonic and melodic cues with little to no temporal discontinuity. Incorporating a ground-truth state sequence in the multi-state HMMs using melodic contour information derived from the transcription yielded statistically significant increases in recall in styles containing a high proportion of these phrase boundaries.

Although our feature set does not attempt to predict pitches of individual notes, we believe the increased recall associated with the multi-state models indicates the model is exploiting a relationship between timbral and harmonic observations and melodic contours associated with phrase boundaries. These precise relationships are undoubtedly dependent on the timbre of the instrument, yet demonstrate some general utility when trained on a range of lead instruments.

While the attempted representation of melodic contour

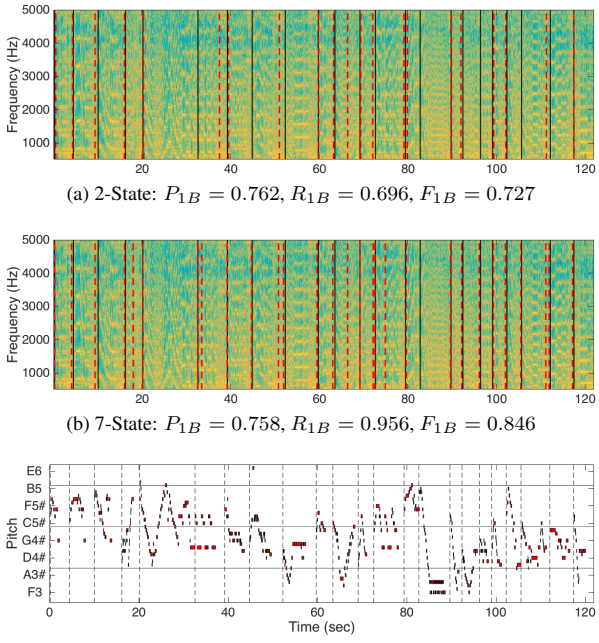


(a) 2-State

(b) 5-State

(c) 7-State

Figure 4. Audio-based HMM segmentation results by style. Significant ($p < 0.005$) increases in recall and f-score observed in Hardbop, Postbop, and Swing.



(a) 2-State: $P_{1B} = 0.762$, $R_{1B} = 0.696$, $F_{1B} = 0.727$

(b) 7-State: $P_{1B} = 0.758$, $R_{1B} = 0.956$, $F_{1B} = 0.846$

(c) MIDI Transcription

Figure 5. Segmentation of Freddie Hubbard's solo in the Eric Dolphy track "245". Black lines indicate ground-truth annotations, and red lines show predicted boundaries.

in the model topology indicates some promise, we believe there are likely better alternatives to modeling contour than arbitrary quanitzation of ground truth inter-pitch intervals. Future work should examine the potential of assembling observed contours from a smaller set of contour primitives over longer time scales than note pair transitions. Furthermore, though our approach avoided relying high-level pitch estimates derived from the audio because of strong potential for propagation of errors, we will investigate the use of mid-level pitch salience functions in future feature sets.

More generally, we believe that the availability of well-aligned audio and symbolic data can allow the use of supervised methods as a precursor to more scalable audio-based methods, and aid in the creation of mid-level features useful for a wide range of MIR problems.

10. REFERENCES

[1] Jakob Aeßler, Klaus Frieler, Martin Pfleiderer, and Wolf-Georg Zaddach. Introducing the jazzomat project - jazz solo analysis using music information retrieval methods. In *In: Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR) Sound, Music and Motion, Marseille, Frankreich.*, 2013.

[2] Emilios Cambouropoulos. *Music, Gestalt, and Computing: Studies in Cognitive and Systematic Musicology*, chapter Musical rhythm: A formal model for determining local boundaries, accents and metre in a melodic surface, pages 277–293. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997.

[3] Tuomas Eerola and Petri Toiviainen. *MIDI Toolbox: MATLAB Tools for Music Research*. University of Jyväskylä, Jyväskylä, Finland, 2004.

[4] D.P.W. Ellis. Aligning midi files to music audio, web resource, 2013. `http://www.ee.columbia.edu/~dpwe/resources/matlab/alignmidi/`. Accessed 2016-03-11.

[5] Sheng Gao, N. C. Maddage, and Chin-Hui Lee. A hidden markov model based approach to music segmentation and identification. In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 3, pages 1576–1580 vol.3, Dec 2003.

[6] M. Levy and M. Sandler. Structural segmentation of musical audio by constrained clustering. *Trans. Audio, Speech and Lang. Proc.*, 16(2):318–326, February 2008.

[7] Marcelo Rodrguez Lpez and Anja Volk. Melodic segmentation: A survey. Technical Report UU-CS-2012-015, Department of Information and Computing Sciences, Utrecht University, 2012.

[8] Jouni Paulus, Meinard Müller, and Anssi Klapuri. State of the art report: Audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, August 9-13, 2010*, pages 625–636, 2010.

[9] Geoffroy Peeters, Amaury La Burthe, and Xavier Rodet. Toward automatic music audio summary generation from signal analysis. In *In Proc. International Conference on Music Information Retrieval*, pages 94–100, 2002.

[10] Mark Sandler and Jean-Julien Aucouturier. Segmentation of musical signals using hidden markov models. In *Audio Engineering Society Convention 110*, May 2001.

[11] David Temperley. *The cognition of basic musical structures*. MIT Press, 2004.