A NOVEL MUSIC SEGMENTATION INTERFACE AND THE JAZZ TUNE COLLECTION

Marcelo Rodríguez-López, Dimitrios Bountouridis, Anja Volk
Utrecht University, The Netherlands
{m.e.rodriguezlopez,d.bountouridis,a.volk}@uu.nl

ABSTRACT

In this paper we present MOSSA, an easy-to-use interface for mobile devices, developed to annotate the segment structure of music. Moreover, we present the *jazz tune collection* (JTC), a database of 125 Jazz melodies annotated using MOSSA, and developed specifically for benchmarking of computational models of melody segmentation. Each melody in the JTC has been annotated with segment boundaries by three human listeners, and segment boundary salience by two human listeners. We provide a light analysis of the inter-annotation-agreement of the annotations in the JTC, and also test the likelihood of the annotations been made using 'gap' related cues (large pitch intervals or inter-onset-intervals) and 'repetition' related cues (exact/approximate repetition of the beginning or ending of phrases).

1. INTRODUCTION

Music segmentation refers to a listening ability that allows human listeners to partition music into sections, phrases, and so on. Computational modelling of music segmentation is important for a number of fields related to Folk Music Analysis, such as Music Information Research (for tasks such as automatic music archiving, retrieval, and visualisation), Computational Musicology (for automatic or human-assisted music analysis), and Music Cognition (to test segmentation theories and more generally theories of musical structure).

Research in music segmentation modelling has been conducted by subdividing the segmentation problem into different tasks, most often segment boundary detection and segment labelling. Segment boundary detection is the task of automatically locating the time instants separating contiguous segments. Segment labelling is the task of categorising segments into equivalence classes. Generally, automatic segmentations are evaluated by comparing them to manual (human annotated) segmentations. In this paper we focus on the annotation of segment structure in melodies, which are of special interest in Folk Music Analysis.

1.1 Problem specification

Ideally, a melodic dataset used to test computational segmentation models should have the following two characteristics: first, it should comprise different styles and instrumental traditions, and second, each melody in the dataset should have been annotated by a relatively large number of human listeners.

However, at present most free and readily available annotated databases consist of vocal (mainly european) folk melodies. Furthermore, since the process of annotating segment structure in melodies is time consuming and laborious, participation to melody annotation initiatives is limited, and so melodic datasets are commonly annotated by a single expert annotator (or a small range of annotators that agree on a single segmentation).

Thus, there is a need for easy-to-use tools to avoid discouraging participation to melody annotation initiatives. Moreover, new melody databases are needed to account for stylistic and instrumental diversity when evaluating computational melody segmentation models.

1.2 Paper contributions

In this paper we present MOSSA (in §2) an interface for mobile devices which, aside of its portability, has a fast learning curve. Moreover, we present (in §3) and analyse (in §4) a database of 125 Jazz melodies annotated using MOSSA for benchmarking computational models of melody segmentation.

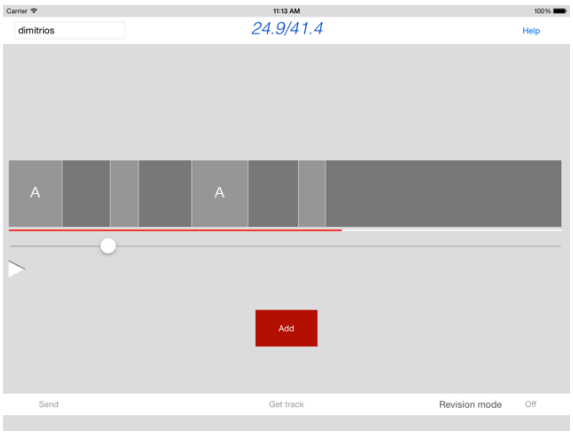

Figure 1: Screenshot of the MOSSA interface

2. MOSSA: MOBILE SEGMENT STRUCTURE ANNOTATION

Figure 1 shows a screenshot of the MOSSA interface. MOSSA is written in Objective-C for iOS. The code is available at http://www.projects.science.uu.nl/music/.

The main goals for the development of MOSSA, aside from portability, are (a) to avoid visual biases, and (b) to ensure a rapid learning curve. We elaborate into these two points below. (For a more detailed specification of the functionality of MOSSA the reader is referred to the documentation accompanying the code.)

2.1 Avoiding visual biases

Many segment structure annotation studies have used a score representation of the music to be annotated. This is specially true for melody segment annotation, e.g. (Thom et al., 2002; Pearce et al., 2010; Karaosmanoglu et al., 2014). Using a visual representation of musical content results in segment annotation biases. For instance, the geometry of score notation might influence the perception of boundary cues. This in turn might suggest the listener a particular segment structure that (s)he might not have been able to perceive without visual cues.

As seen in Figure 1 MOSSA avoids any visual representation of the music content, depicting music only as a time line. Different playback mechanisms are available for the user to easily examine whether the position of segment boundaries or its equivalent class labels are correctly annotated. For instance, if the user double taps a over a segment, playback starts from the leftmost boundary of the segment.

2.2 Ensuring fast learning

Most freely available interfaces for music annotation are rich in options, e.g. see (Li et al., 2006; Peeters et al., 2008; Cannam et al., 2006). However, the large number of options comes at the expense of user interaction simplicity, and hence may result in a relatively long and steep learning curve. MOSSA has been designed to minimise its learning time, by providing a clean and simple interface, and a visually intuitive way to annotate segment boundaries and label equivalent classes. For instance, as seen in Figure 1 boundaries can be inserted by simply pressing the 'add' button. Alternatively, boundaries can also be inserted by making a downwards swipe gesture over the block region representing the music.

The idea is that MOSSA is used by non-expert users, and then the annotations can be checked by experts in more advanced annotation interfaces, such as Sonic Annotator or Audacity.

3. THE JAZZ TUNE COLLECTION (JTC)

The JTC is a dataset of Jazz theme melodies constructed to evaluate computational models of melody segmentation. A list of global statistics describing the dataset is presented in Table 1.

Total number of melodies	125
Total number of notes	19419
Total time (in hours)	3.103
Approximate range of dataset (in years)	1880-1986
Total number of composers	81
Total number of styles	10

Table 1: Global statistics of the JTC

All melodies are available in MIDI. Each melody in the JTC is annotated with phrase boundaries (by three human listeners) and boundary salience (by two human listeners). [1] In Table 2 we present the total number of phrases and mean phrase lengths (with standard deviation values in parenthesis) per annotation.

Annotation	Number of Phrases	Mean Phrase Length	
		Notes	Seconds
1	1881	10.32 (4.85)	5.94 (3.16)
2	1701	11.42 (6.55)	6.57 (3.93)
3	1682	11.55 (5.78)	6.64 (4.01)

Table 2: Summary statistics of annotated phrases.

All segment boundaries and salience annotations were produced using MOSSA, and are provided in Audacity' label file format. The JTC also provides metadata for each melody. The metadata includes information of tune title, composer, Jazz sub-genre, and year of the tune's composition/release. The JTC dataset can be accessed at: `http://www.projects.science.uu.nl/music/`

3.1 JTC assembly

To assemble the JTC, we consulted online sources that provide rankings of jazz tunes, albums, and composers. [2] We employed a web-crawler to automatically collect MIDI and MusicXML files from a number of sources in the internet. (The majority were crawled from the now defunct *Wikifonia Foundation*. [3]). We cross referenced the rankings and the collected files, and selected 125 files trying to find a balance between tune ranking, composer ranking, sample coverage, and encoding quality. We describe the JTC's sample coverage (in terms of time periods and sub-genres) below, and discuss the encoding quality of the files in §3.3.

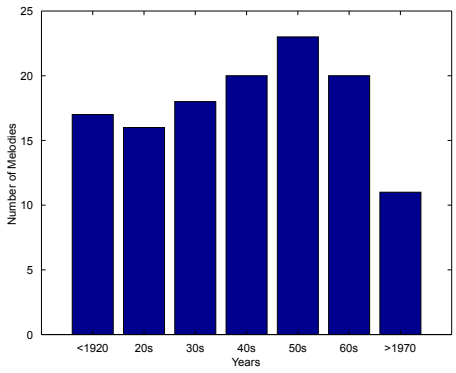

Figure 2: JTC: number of melodies per time period

The JTC can be divided in seven time periods (see Figure 2). Each time period contains between 11 and 23 tunes from representative sub-genres (see Figure 3) and influential composers/performers of the period. The year of release/composition, Jazz sub-genre, and composer metadata was obtained by consulting online sources. [4]

[1] We use the term 'boundary salience' to refer to a binary score that reflects the relative importance of a given boundary as estimated by a human annotator.

[2] The main sources consulted were: `www.allmusic.com`, `www.jazzstandards.com`, `en.wikipedia.org`

[3] `www.wikifonia.org`

[4] in most cases `en.wikipedia.org` and `www.allmusic.com`

Class Label	Sub-Genre
C1	Bebop
C2	Big Band, Swing, Charleston
C3	Bossa Nova, Latin Jazz
C4	Cool Jazz, Modal Jazz
C5	Dixieland
C6	Early, Rag time, Folk Song
C7	Electric Jazz, Fusion, Modern
C8	Other
C9	Musical, Film, Broadway
C10	Post Bop, Hard Bop

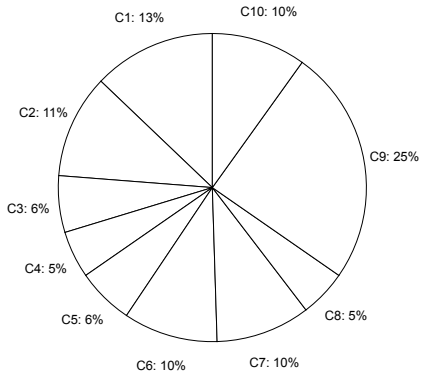

Figure 3: Distribution of sub-genres in the JTC

3.2 Melody encoding quality and corrections

From the 125 melodies making up the JTC, 64 correspond to perfomed MIDI files, 4 to manually encoded MIDI files, and 57 to manually encoded lead sheets in MusicXML format. In most cases the performed MIDI files encoded polyphonic music, so the melody was extracted automatically by locating the MIDI track labelled as 'melody'. [5]

All melodies were exported as MIDI files, using a resolution of 480 ticks-per-quarter-note, which successfully encoded the lowest temporal resolution of the melodies. All melodies were inspected manually, and, if needed, corrected. Correction of the melodies consisted in adjusting note onsets, as well as removing ornamentation. Notated leadsheets from the Real Book series [6] were used as reference for the correction process. Is important to notice that not all ornamentation was removed, only that which was considered to severely compromise the understanding of segment structure. Also, while JTC melody encodings might contain information of meter, key, and dynamics, this information was not checked nor corrected, and thus its use as 'a priori' information for computational modelling of segmentation is discouraged.

3.3 Segment structure annotation process

For each melody, segment boundaries and salience were annotated by one amateur musician and one degree-level musician. These are referred to, respectively, as 'annotation 1' and 'annotation 2' in the Tables and Figures of this paper. For each melody there is also a third annotation of segment boundaries, produced by one of a group of extra annotators. This annotation is referred to as 'annotation 3' throughout the paper.

The group of extra annotators consisted of 27 human listeners (18 male and 19 female), ranging from 20 to 50 years of age. In respect to the level of musical education of the extra annotators, 6 reported to be self taught singer/instrumentalist, 10 reported to having some degree of formal musical training, and 11 reported to having obtained a superior education degree in a music related subject. Moreover, extra annotators were asked to rate their degree of familiarity with Jazz (on a scale of 1 to 3, with 1 being the lowest, and 3 the highest), 12 annotators rated their familiarity as '1', 7 rated their familiarity as '2', and 8 rated their familiarity as '3'. Lastly, none of the extra annotators reported to suffering from any form of hearing impairment, and 2 reported having perfect pitch.

4. ANALYSIS OF PHRASE ANNOTATIONS

In this section we analyse the phrase annotations. In §4.1 we analyse two global properties of the annotated phrases: length and contours. In §4.2 we analyse inter-annotator-agreement using two different measures that score agreement. Finally, in §4.3 we check the vicinity of annotated phrases for evidence of two factors commonly assumed to be of high importance to segment boundary perception: *gaps* (in duration and pitch related information) and phrase start *repetitions* (also in duration and pitch related information).

4.1 Phrase Lengths and Contours

The mean phrase duration lengths presented in Table 2 and the box plots presented in Figure 4 show that the phrases of annotations 2 and 3 tend to be larger than those in annotation 1. Both boxes and whiskers of box plots 2 and 3 tend to be larger than those of box plot 1, indicating a larger spread skewed towards longer phrases. Furthermore, the notch of box plot 1 does not overlap with those of box plots 2 and 3, which indicates, with 95% confidence, that the difference between their medians is significant.

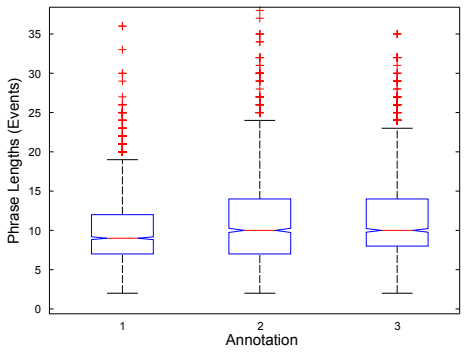

Figure 4: Annotated phrase lengths

To get further insights into these apparent preference for longer phrases, we consulted the degree-level musician of annotation 2 and some of the extra annotators for

[5] If no such track was found the file was automatically filtered from the selection process.

[6] The Real Book editions used as reference for editing are published by `www.halleonard.com`.

their choice of phrase lengths. The most common reply was that on occasion relatively long melodic passages suggested multiple segmentations, where phrases 'seemed to merge into each other' rather than having clear boundaries. For these passages the consulted annotators reported choosing to annotate just one long phrase with 'clear' boundaries rather than attempting to segment the melodic passage into multiple segments.

We also manually checked the outliers identified in Figure 4 for the presence of potential annotation errors. In most cases outliers simply correspond to melodic passages with high tempo and high note density, and are not particularly large in terms of time in seconds. Two examples of these type of outliers (common to all annotations) are phrases in the melodies of Dexterity and Ornithology of Charlie Parker.

	Annotation		
Huron's Contour Classes	1	2	3
convex	33.86	35.10	36.15
descending	23.71	24.99	24.14
ascending	19.30	20.16	19.62
concave	19.99	16.34	17.06
ascending-horizontal	1.33	1.00	1.13
horizontal-descending	0.58	0.88	0.54
horizontal-ascending	0.37	0.59	0.48
descending-horizontal	0.48	0.47	0.42
horizontal	0.37	0.47	0.48

Table 3: Contour class classification of annotated phrases

We classified the annotated phrases in respect to their type of gross melodic contour using the contour types of Huron (1996). Table 3 shows the classification results, expressed as a percentage of the total number of phrases per annotation. The results show that all annotators agree in the ranking given to the four dominant contour classes, namely convex, descending, ascending, and concave (these four contour classes describe ∼96 percent of the phrases in each annotation). The ranking of the four dominant classes also matches the ranking obtained by Huron (1996), who performed phrase contour classification on ∼36000 vocal melodic phrases.

4.2 Inter-annotator-agreement (IAA) analysis

We checked the inter-annotator-agreement for each melody annotation using Cohen's κ (1960). Table 4 shows the mean pairwise agreement $\overline{\kappa}$, with standard deviation σ_κ in parenthesis. According to the scale proposed by Klaus (1980) the mean agreement on phrase boundary locations between annotations can be considered 'tentative', and according to the scale of Green (1997) it can be considered 'fair'. However, if for each melody we consider only the two highest κ scores, then $\overline{\kappa} = 0.86$, which can be considered by both the Klaus and Green scales as 'good/high'. Moreover, this 'best two' mean agreement also shows a substantial reduction in σ_κ. This indicates that, for any melody in the JTC, is likely that at least two segmentations have good agreement.

Annotation	$\overline{\kappa}$
1 vs 2	0.72 (0.22)
1 vs 3	0.71 (0.24)
2 vs 3	0.69 (0.26)
Best two	0.86 (0.15)

Table 4: Mean pairwise IAA ($kappa$)

Manual inspection of the boundary annotations showed that, even in cases when the annotators roughly agree on the total number of boundaries for a melody, constructing histograms of boundary markings results in clusters of closely located boundaries. We observed that these boundary clusters are in cases a side effect of dealing with ornamentation during segmentation (i.e. deciding whether grace notes, mordents, or fills should be part of one or another segment). We argue that boundary clusters are examples of 'soft' disagreement and should not be harshly penalised when estimating agreement.

The κ statistic does not take into account the possibility of, nor is able to provide partial scores for, points of 'soft' disagreement when estimating agreement. Hence, to investigate the effect of soft disagreement in the JTC we employed an alternative measure, namely the Boundary Edit Distance Similarity (B), recently proposed in (Fournier, 2013). One of the parameters of the B measure is a tolerance window (in notes). Within this tolerance window boundaries are given a partial score proportional to their relative distance. We tested the effect of soft disagreement by computing the B for each melody in the JTC using two tolerance levels: one note (giving score only to points strong agreement) and four notes (giving score also to points of soft agreement). We then computed whether the differences between the medians of the two sets of scores is statistically significant using a paired Wilcoxon Signed Rank test (WSRT). The results of this analysis are presented in Table 5. The WSRT confirms that the difference in medians is significant ($p < 0.001$), with medium effect size ($r = 0.41 - 0.47$). These results suggest that the number of points of 'soft' disarrangement is not negligible and it should be taken into consideration when benchmarking computational models of segmentation.

4.3 Analysis of Segment Boundaries

In this section we check annotated phrase boundaries and their immediate vicinity for the presence of two cues commonly assumed to be of high importance to segment boundary perception: melodic *gaps* and phrase start *repetitions*.

Melodic gaps can be defined as overly large changes in the temporal evolution of a given attribute used to describe a melody. Phrase start repetitions can be defined as an exact or approximate match of the attributes representing the starting point of two or more phrases. Our goal is to test to what extent gaps and repetitions can be considered a defining feature of the annotated phrase boundaries of the JTC. To that end, we make two complementary hypotheses: (a) the probability of detecting a gap at annotated phrase boundaries in a melody should be relatively high, which provides evidence that phrase boundaries of-

Annotation	$\tilde{B}$ (tolerance = 1 note)	$\tilde{B}$ (tolerance = 4 notes)	WSRT
1 vs 2	0.67	0.70	h: 1, Z: 4.54, $p < 0.001$, r: 0.41
1 vs 3	0.62	0.67	h: 1, Z: 5.23, $p < 0.001$, r: 0.46
2 vs 3	0.60	0.65	h: 1, Z: 5.23, $p < 0.001$, r: 0.47

Table 5: WSRT of B scores, tilde is used to denote the median, for the WSRT see Appendix A.1.

ten contain gaps, and (b) the probability of detecting a gap at non-boundary points in a melody should be relatively low, which provides evidence that gaps might be unique-to or distinctive-of phrase boundaries. The same pair of complementary hypotheses can be made for phrase start repetitions.

4.3.1 Computing per-melody detection probabilities

We compute the probability of detecting gaps/repetitions at/following boundaries:

$$P_B = \frac{A_D}{A}, \tag{1}$$

where A_D is the number of annotated boundaries containing/preceding detected gaps/repetitions, and A is the total number of annotated boundaries in the melody. Likewise, we can compute the probability of detecting gaps/repetitions at/following non-boundaries:

$$P_N = \frac{N_D}{N}, \tag{2}$$

where N_G is the number of non-boundaries containing/preceding detected gaps/repetitions, and N is the total number of non-boundaries in the melody.

4.3.2 Defining non-boundary points

We selected random non-boundary points with the following constraints: First, for each melody there should be an equal number of boundaries and non-boundaries. Second, non-boundary points should result in a set of segments of comparable length and standard deviation than that of the annotated phrases. With these two constraints, non-boundaries were drawn with uniform probability over eligible portions of the melody.

4.3.3 Gap analysis procedure

For gap detection we represent melodies as sequences of pitch or duration intervals. In this paper we measure pitch intervals (PI) in semitones, and measure duration using inter-onset-intervals (IOI) in seconds.

We classify (non-)boundaries as either containing or not containing a gap separately for PI and IOI using four different models of gap detection:

> T (Tenney & Polansky, 1980) in which a gap is detected if the interval at the (non-)boundary is larger than the intervals immediately preceding and following it.
>
> C (Cambouropoulos, 2001) in which a gap is detected if the interval at the (non-)boundary has a larger 'boundary strength score' than intervals immediately preceding and following it.
>
> R in which a gap is detected if the interval at the (non-)boundary is (a) equal or larger that four times the mode IOI of the melody, or (b) equal or larger than the mean PI of the melody plus one standard deviation.
>
> L in which a gap is detected if the interval at the (non-)boundary has (a) an IOI equal or larger than 1.5 seconds, or (b) a PI equal or larger than 9 semitones.

4.3.4 Repetition analysis procedure

For repetition detection we represent melodies as sequences of pitch intervals or inter-onset-interval ratios. We measure pitch intervals (PI) in semitones, and measure inter-onset-interval ratios (IOIR) in nats. [7]

We used the edit distance (Levenshtein, 1966) to compute similarity values S between the starting point of all phrases per melody. The similarity obtained per melody is normalised so that $S \in [0, 1]$. Pairwise phrase S values were computed separately for the PI and IOIR representation of the melody.

We define the start of a phrase according to the following rules. First, for an annotated segmented to be considered a valid phrase, we required segments to be longer than 2 intervals. Second, each valid phrase is divided in two (rounded to the nearest integer down) and the first half is used as a phrase start. If the first half is longer than 9 intervals truncation is applied. The maximum length of phrase start was chosen so that phrase starts are not longer than approximately the mean phrase size of the JTC (which according to Table 2 ranges between ∼10-11 notes).

For our experiments we classify phrase starts as either being repeated or not by considering three thresholds: similar ($S > 0.6$), closely similar ($S > 0.8$), and exact match ($S = 1$). [8]

4.3.5 Results

The results of the gap analysis are presented in Table 6. The results of the repetition analysis is presented are Table 7. To test if the differences between the medians of the obtained P_B and P_N scores are significant, we used once again the WSRT.

Our results show that all annotations seem to roughly rank the tested cues in the same way. That is, IOI gaps are at the top of the ranking, with a P_B peaking at $\sim 0.95 - 1.00$, showing large and significant differences in respect to P_N scores. IOIR and PI repetitions are second, with

[7] The IOIR are computed using the formula and parameters proposed in (Wolkowicz, 2013, p. 45).

[8] For the exact match threshold we used the raw (not normalised) values of S.

P_B scores ranging between $\sim 0.30 - 0.66$, also showing relatively large and significant differences in respect to P_N scores. PI gaps are at the bottom of the raking, with P_B scores ranging $\sim 0.01 - 0.43$, showing in various cases non-significant differences in respect to P_N scores.

5. CONCLUSIONS

In this paper we have presented MOSSA a music segment structure annotation interface for mobile devises. We have discussed some of the benefits of MOSSA in respect to existing segment structure annotation interfaces, such as its fast learning curve and avoidance of visual biases. In addition, we presented and analysed the *jazz tune collection* (JTC), a database of 125 Jazz melodies annotated using MOSSA, developed for benchmarking of computational models of melody segmentation. Our analysis of the JTC is aimed at investigating the inter-annotation-agreement of the annotations in the JTC, and also test the likelihood of the annotations been made using 'gap' related cues (large pitch intervals or inter-onset-intervals) and 'repetition' related cues (exact/approximate repetition of the beginning or ending of phrases).

Acknowledgments: Marcelo Rodríguez-López and Anja Volk (NWO-VIDI grant 276-35-001) and Dimitrios Bountouridis (NWO-CATCH project 640.005.004) are supported by the Netherlands Organization for Scientific Research.

6. REFERENCES

Cambouropoulos, E. (2001). The local boundary detection model (lbdm) and its application in the study of expressive timing. In *Proceedings of the international computer music conference*, (pp. 17–22).

Cannam, C., Landone, C., Sandler, M. B., & Bello, J. P. (2006). The sonic visualiser: A visualisation platform for semantic descriptors from musical signals. In *ISMIR*, (pp. 324–327).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (1988). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Fournier, C. (2013). Evaluating text segmentation using boundary edit distance. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, (pp. 1702–1712).

Green, A. M. (1997). Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the 22nd annual SAS User Group International conference*, (pp. 1110–1115).

Huron, D. (1996). The melodic arch in western folksongs. *Computing in Musicology*, *10*, 3–23.

Karaosmanoglu, M. K., Bozkurt, B., Holzapfel, A., & Disiacik, N. D. (2014). A symbolic dataset of turkish makam music phrases. In *Proceedings of the 4th Folk Music Analysis Workshop (FMA)*, (pp. 10–14).

Klaus, K. (1980). Content analysis: An introduction to its methodology.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, (pp. 707–710).

Li, B., Burgoyne, J. A., & Fujinaga, I. (2006). Extending audacity for audio annotation. In *ISMIR*, (pp. 379–380).

Pearce, M., Müllensiefen, D., & Wiggins, G. (2010). Melodic grouping in music information retrieval: New methods and applications. *Advances in music information retrieval*, 364–388.

Peeters, G., Fenech, D., & Rodet, X. (2008). Mcipa: A music content information player and annotator for discovering music. In *ISMIR*, (pp. 243–248).

Tenney, J. & Polansky, L. (1980). Temporal gestalt perception in music. *Journal of Music Theory*, 205–241.

Thom, B., Spevak, C., & Höthker, K. (2002). Melodic segmentation: Evaluating the performance of algorithms and musical experts. In *Proceedings of the International Computer Music Conference (ICMC)*, (pp. 65–72).

Wolkowicz, J. M. (2013). Application of text-based methods of analysis to symbolic music.

A. APPENDICES

A.1 Wilcoxon Signed Rank test (WSRT)

Since the B scores can not be assumed to be normally distributed, we use the Wilcoxon Signed Rank test, which is a non-parametric alternative to the paired Students t-test, and gives the probability that two distributions of paired samples have the same median.

In this paper the results of the WSRT are reported using: h - test result (a value of 1 indicates the test rejects null hypothesis), Z - value of the z-statistic, p - p value, r - effect size. The effect size is computed as $r = {}^{Z}/\sqrt{N}$, where N is the total number of the samples. According to (Cohen et al., 1988), effect size values can be interpreted as small size if $r \leq 0.1$, medium size if $0.1 > r \leq 0.3$, large size if $0.3 > r \leq 0.5$, and very large size if $r > 0.5$.

IOI gaps				
Annotation	Gap Model	$\tilde{P}_A$	$\tilde{P}_N$	WSRT
1	T	0.95	0.20	h: 1, Z: 9.57, $p < 0.001$, r: 0.86
	C	0.94	0.21	h: 1, Z: 9.56, $p < 0.001$, r: 0.85
	R	0.67	0.04	h: 1, Z: 9.05, $p < 0.001$, r: 0.81
	A	0.56	0.02	h: 1, Z: 9.04, $p < 0.001$, r: 0.81
2	T	1.00	0.20	h: 1, Z: 9.64, $p < 0.001$, r: 0.86
	C	1.00	0.21	h: 1, Z: 9.63, $p < 0.001$, r: 0.86
	R	0.79	0.04	h: 1, Z: 9.14, $p < 0.001$, r: 0.82
	A	0.64	0.02	h: 1, Z: 9.08, $p < 0.001$, r: 0.81
3	T	0.96	0.20	h: 1, Z: 9.57, $p < 0.001$, r: 0.86
	C	0.96	0.20	h: 1, Z: 9.58, $p < 0.001$, r: 0.86
	R	0.78	0.04	h: 1, Z: 9.07, $p < 0.001$, r: 0.81
	A	0.61	0.04	h: 1, Z: 8.97, $p < 0.001$, r: 0.80
PI gaps				
Annotation	Gap Model	$\tilde{P}_A$	$\tilde{P}_N$	WSRT
1	T	0.25	0.27	h: 0
	C	0.42	0.36	h: 0
	R	0.29	0.11	h: 1, Z: 7.57, $p < 0.001$, r: 0.68
	L	0.01	0.01	h: 1, Z: 5.17, $p < 0.001$, r: 0.46
2	T	0.27	0.26	h: 0
	C	0.42	0.35	h: 0
	R	0.29	0.12	h: 1, Z: 6.78, $p < 0.001$, r: 0.61
	L	0.01	0.01	h: 1, Z: 5.33, $p < 0.001$, r: 0.48
3	T	0.29	0.26	h: 0
	C	0.43	0.34	h: 1, Z: 2.80, $p < 0.01$, r: 0.25
	R	0.27	0.10	h: 1, Z: 6.72, $p < 0.001$, r: 0.60
	L	0.01	0.01	h: 1, Z: 4.91, $p < 0.001$, r: 0.44

Table 6: Gaps at annotated boundaries and random boundaries, tilde is used to denote the median, for the WSRT see Appendix A.1.

Repetition of Phrase Beginning: IOI Ratio (IOIR)				
Annotation	Threshold	$\tilde{P}_A$	$\tilde{P}_N$	WSRT
1	$S > 0.6$	0.66	0.42	h: 1, Z: 8.56, $p < 0.001$, r: 0.77
	$S > 0.8$	0.50	0.25	h: 1, Z: 8.57, $p < 0.001$, r: 0.77
	$S = 1$	0.33	0.18	h: 1, Z: 7.65, $p < 0.001$, r: 0.68
2	$S > 0.6$	0.63	0.41	h: 1, Z: 8.56, $p < 0.001$, r: 0.77
	$S > 0.8$	0.50	0.26	h: 1, Z: 8.71, $p < 0.001$, r: 0.78
	$S = 1$	0.38	0.18	h: 1, Z: 7.49, $p < 0.001$, r: 0.67
3	$S > 0.6$	0.64	0.40	h: 1, Z: 7.92, $p < 0.001$, r: 0.71
	$S > 0.8$	0.50	0.27	h: 1, Z: 7.60, $p < 0.001$, r: 0.68
	$S = 1$	0.30	0.20	h: 1, Z: 6.90, $p < 0.001$, r: 0.62
Repetition of Phrase Beginning: Pitch Interval (PI)				
Annotation	Threshold	$\tilde{P}_A$	$\tilde{P}_N$	WSRT
1	$S > 0.6$	0.59	0.38	h: 1, Z: 8.46, $p < 0.001$, r: 0.76
	$S > 0.8$	0.46	0.23	h: 1, Z: 8.79, $p < 0.001$, r: 0.79
	$S = 1$	0.33	0.17	h: 1, Z: 7.76, $p < 0.001$, r: 0.69
2	$S > 0.6$	0.60	0.35	h: 1, Z: 8.55, $p < 0.001$, r: 0.76
	$S > 0.8$	0.50	0.24	h: 1, Z: 8.75, $p < 0.001$, r: 0.78
	$S = 1$	0.33	0.18	h: 1, Z: 7.37, $p < 0.001$, r: 0.66
3	$S > 0.6$	0.57	0.38	h: 1, Z: 7.74, $p < 0.001$, r: 0.69
	$S > 0.8$	0.43	0.25	h: 1, Z: 7.93, $p < 0.001$, r: 0.71
	$S = 1$	0.29	0.20	h: 1, Z: 6.84, $p < 0.001$, r: 0.61

Table 7: Repetitions at annotated and random phrase beginnings, tilde is used to denote the median, for the WSRT see Appendix A.1.