EMOTION BASED SEGMENTATION OF MUSICAL AUDIO

Anna Aljanaki
Utrecht University
A.Aljanaki@uu.nl

Frans Wiering
Utrecht University
F.Wiering@uu.nl

Remco C. Veltkamp
Utrecht University
R.C.Veltkamp@uu.nl

ABSTRACT

The dominant approach to musical emotion variation detection tracks emotion over time continuously and usually deals with time resolutions of one second. In this paper we discuss the problems associated with this approach and propose to move to bigger time resolutions when tracking emotion over time. We argue that it is more natural from the listener's point of view to regard emotional variation in music as a progression of emotionally stable segments. In order to enable such tracking of emotion over time it is necessary to segment music at the emotional boundaries. To address this problem we conduct a formal evaluation of different segmentation methods as applied to a task of emotional boundary detection. We collect emotional boundary annotations from three annotators for 52 musical pieces from the RWC music collection that already have structural annotations from the SALAMI dataset. We investigate how well structural segmentation explains emotional segmentation and find that there is a large overlap, though about a quarter of emotional boundaries do not coincide with structural ones. We also study inter-annotator agreement on emotional segmentation. Lastly, we evaluate different unsupervised segmentation methods when applied to emotional boundary detection and find that, in terms of F-measure, the Structural Features method performs best.

1. INTRODUCTION

Improving automatic music emotion recognition (MER) methods is crucial to enhance accessibility of large music collections for both personal and commercial use. Driven by this interest, the MER field greatly expanded in the last decade. One of the fundamental MER problems is tracking emotion over time, or music emotion variation detection (MEVD). This problem is usually approached by a continuous approach to MER (dynamic MER), when the emotion of a piece of music is predicted on a second-by-second basis. Though dynamic MER does not actually assume that emotion in music should change every second, the current methods tend to work on very low time resolutions both by choosing rather short excerpts where no serious musical development could occur (e.g., 15 seconds) and by collecting the ground truth with certain task demands on the annotators. It has been notoriously difficult to collect a ground truth for MEVD with a reasonable inter-annotator agreement, and the reason may lie in the fact that musical meaning is usually communicated during bigger time spans than several seconds, and it is therefore difficult and unnatural for the listeners to evaluate their emotional response to music in such a way. Though it might still be interesting and important to track musical change over time, the question should be raised whether change on such a short time scale is actually an expression of musical emotion or *the means of creating* emotional expression on a higher level (e.g., accelerando or crescendo).

A bordering MER field (static MER) studies identification of emotion in somewhat longer musical segments. Static MER methods usually deal with excerpts of 15 to 30 seconds. It is natural for listeners to describe musical content by applying emotional labels to musical excerpts or complete pieces. This kind of labels are used by most music services to categorize their data. However, the real world problem of MEVD requires music to be presegmented into fragments with stable emotion. This problem is usually just neglected by static MER methods, which often use ground-truth excerpts picked by randomly sampling the audio and filtering out the excerpts that receive contradictory ratings from experts. Also, sometimes the problem is solved (or rather avoided) by trying to pick the most representative part of the song for classification (e.g., chorus).

Hence, many questions about emotional segmentation of music remain unsolved. What is a typical length of an emotionally stable fragment in music? (Ironically, both static and dynamic MER methods usually deal with musical excerpts of more or less the same lengths, ranging from 15 to 45 seconds in an attempt to cover as much different music as possible while reducing the annotation burden.) Is emotional segmentation explained by structural segmentation? How many emotional boundaries are there typically in a piece of music? Which segmentation methods work best when applied to emotional boundary detection?

These are the questions that we are going to deal with in this paper. For these purposes we assemble a dataset of 52 double-annotated pieces from the RWC music database [6], which also have structural annotations in the SALAMI dataset [13]. We obtain a little under 2000 annotated emotional boundaries (around 630 from each of the annotators). We compare emotional and structural segmentation of music, analyze the inter-annotator agreement and the

average stable segment length. Then we apply four segmentation algorithms to emotional segmentation problem and benchmark them on our dataset. Though the dataset is not big, a formal evaluation of emotional segmentation performance has never been conducted before.

In this work, we are not going to deal with MER in a traditional sense (predicting emotion from a musical excerpt). There already exist numerous state-of-the-art approaches to this problem [20]. Here we will address the question how to do the preprocessing step before static MER, i.e., emotional segmentation of music.

The rest of the paper is organized as follows. In section 2 we describe related research. In section 3 we explain why dynamic MER methods, at least in their current form, might not produce a good solution to the MEVD problem. In section 4 we analyze the obtained emotional segmentation. In section 5 we compare different segmentation methods when applied to a problem of detecting emotional boundaries in music. Section 6 concludes the paper.

2. RELATED WORK

Though the problem of emotional boundary detection has not yet been addressed systematically, there exist MER methods that can be applied to this problem, and we will review them in this section. For a more general overview of MER, [20] can be consulted.

2.1 Static MER for MEVD

The most simple approach to MEVD when using a static MER method is detecting emotion over time using a sliding window. This method would give a distorted result when a sliding window has an emotional boundary in it. In [21], a sliding window of ten seconds and $1/3$ overlap is used to segment a music piece, and a fuzzy classifier is trained to detect the emotion of the segments. In [9] it is suggested that a homogeneous music segment is usually around 16 seconds, and therefore a sliding window of 16s is used to detect the boundaries by comparing feature distributions from neighboring windows. This approach is shown to be viable, though many questions are left open. For instance, only two features — intensity and timbre — are tested, and the evaluation is conducted only on 9 pieces. A similar approach is attempted in [15] to solve a multi-label classification problem (with two sliding windows of 10s and 30s). It is concluded that a more sophisticated emotional segmentation strategy is needed. Multi-label classification approaches recognize that one musical piece can express a variety of emotions and several labels are applicable to one piece. However, the music is often still handled in the same way as in the static MER approach. A short excerpt (e.g., 30s) is selected ([16], [17]), and several labels are applied to it, which addresses the problem of musical ambiguity, but not musical change. As opposed to this approach, in [18] a multi-label classification was applied to whole musical pieces, which were pre-segmented using aligned lyrics annotations on an assumption that most often emotion is stable within one sentence. Then, a hierarchical Bayesian model was applied to a task of multi-label classification. Due to the absence of ground-truth on emotional boundaries in [18], it is left unclear how well the annotated sentences in the lyrics actually correspond to emotional structure of the musical piece.

To answer the question of what is the typical length of musical segments that represent stable emotion, Xiao et al. tried to classify excerpts of different lengths by emotion and found that excerpts of 8 or 16 seconds have a better classification accuracy than excerpts of 4 or 32 seconds [19]. This experiment gives an indirect indication of emotional segmentation resolution.

2.2 Dynamic MER

Dynamic MER methods are usually trained on time-series of annotations, typically with a resolution of 1 or 2 Hz. In Korhonen et al. [7], musical emotion is modeled as a function of musical features using system identification techniques. In [11], conditional random fields were used to model continuous emotion with a resolution of 11×11 in valence-arousal space. A similar strategy was employed in [4], where dynamic texture models were trained corresponding to quadrants of resonance-arousal-valence model and applied to predict musical emotion continuously. When separate models are trained to predict different emotions, emotional boundary detection occurs naturally. This approach might be problematic, however, due to lack of resolution in the emotional space. Also, for boundary detection it might be more important to keep track of the local context and relative changes in musical attributes rather than predict an absolute rating at every moment. This is why unsupervised methods might work very well in this case.

3. MOTIVATION

While static MER methods cannot deal with emotionally non-homogenous music, dynamic MER methods approach this problem by taking the fragmentation to the extreme (the typical resolution of a dynamic MER method is 1 second), which might create even more problems than it solves. Firstly, the output (per-second emotion prediction) produced by a dynamic MER method is not easily interpretable and useful. Secondly, it seems that musical emotion is not conceptualized in this way by listeners.

3.1 Analyzing dynamic MER ground-truth

Dynamic MER relies on human ground-truth in the form of per-second emotional annotations, which are typically recorded from an annotator continuously moving their cursor in a one or two-dimensional space [1, 14]. It seems that this task is extremely difficult for humans, which is, in particular, indicated by a very low inter-annotator agreement as compared to static annotations (where, due to task subjectivity, it is also not very high). For the MediaEval dataset [1], the average Kendalls W is 0.23 ± 0.16 for arousal and 0.28 ± 0.21 for valence, and for the MoodSwings Lite dataset [14] the mean Kendall's W is 0.21 ± 0.14 for arousal and 0.23 ± 0.17 for valence. All these

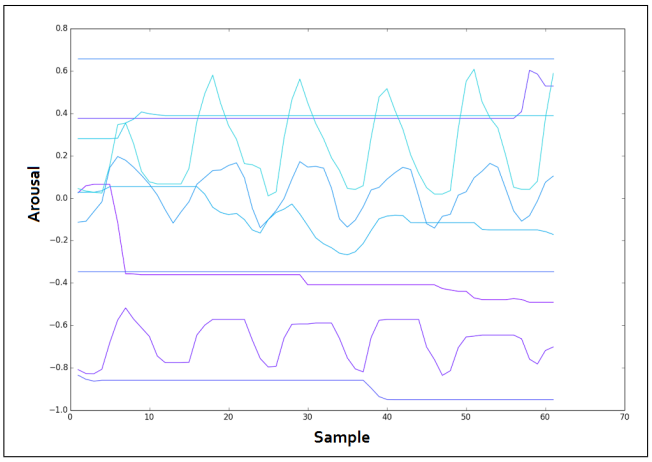

Figure 1. Dynamic annotation of 45 seconds of audio from [1]. One third of the annotators react to every beat of slow music by a peak in arousal.

numbers indicate weak agreement. There are several typical problems arising when annotating music continuously:

1. A dimensional annotation interface has an absolute scale. For instance, on an axis with a slider controlling valence, the leftmost side represents the most miserable music imaginable, and the rightmost the most ecstatic one. Giving absolute ratings is relatively easy when evaluating music statically (comparing a piece to all existing music). When comparing piece with itself over time, humans tend to think of occurring changes relatively. This leads to a huge difference in magnitude of given ratings, though the direction of change can be indicated uniformly (e.g., see Figure 1).

2. Though it is not explicitly requested from the annotators to move their cursor at all times, the task demands (short excerpt, necessity to track and respond continuously) lead to some of the annotators evaluating every single musical event (e.g., see Figure 1). This results in annotations on widely different 'zoom level'.

We argue that continuous annotation is so difficult (albeit through training in the lab and a careful selection of complete music pieces it is possible to obtain satisfying results [3]) because it is unnatural for humans to evaluate their emotional response on a per-second basis, since emotional expression occurs on a much larger time-scale. Though through years of exposure to music listeners acquire an ability to associate certain timbres with genre and emotion, and a crude emotional interpretation is possible even from short sounds snippets of 300ms [8], we believe that real-life emotional interpretation of music is much more complex and happens during longer time spans, most certainly when it concerns induced emotion.

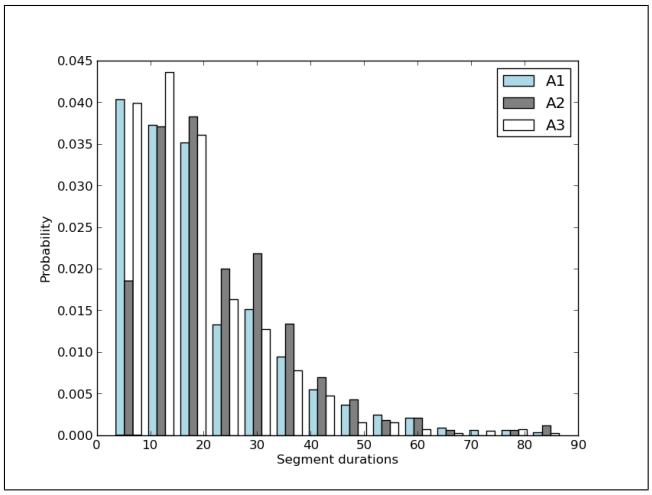

Figure 2. Histogram of segment durations for the three annotators separately.

4. ANALYSIS OF EMOTIONAL BOUNDARIES

4.1 Data

The dataset consists of 52 complete pieces [6] from Pop, Jazz and Genre (the latter contains rock, soul, world etc. music) collections of RWC music database. We picked the pieces that already had SALAMI [13] annotations in order to compare structural and emotional segmentation. The SALAMI annotations for these pieces are single-keyed, our annotations are triple-keyed in order to enable measuring agreement.

The three annotators received instructions to mark when emotion of the piece changes. There were no explicit instructions as to what could be interpreted as an emotional boundary. They were also instructed to mark the transitions between stable emotional states as separate sections, in case those were long enough to be perceived as separate 'transition states'. In practice, this meant for instance marking long diminuendo (fade-out) at the end of a musical piece as a separate section.

In total, annotators marked 562, 602 and 746 emotional boundaries, respectively. The dataset is available from the website `osf.io/jpd5z`.

Evaluation metric	A2→A1	A3→A2	A1→A3
Precision @ 0.5	0.47	0.43	0.52
Recall @ 0.5	0.48	0.33	0.55
F-measure @ 0.5	0.46	0.37	0.67
Precision @ 3	0.73	0.88	0.72
Recall @ 3	0.76	0.79	0.88
F-measure @ 3	0.73	0.77	0.78

Table 1. Inter-annotator boundary retrieval with a tolerance window of 0.5 and 3 seconds.

4.2 Inter-annotator agreement

The mean number of boundaries per piece was 12.2 (median = 11.5). The average segment length was 19.5 ± 18s.

Figure 2 shows the histograms of segment lengths from the three annotators. We can see that the distribution is skewed, 90% of intervals are shorter than 37 seconds. Annotators 1 and 3 have annotated more short segments than annotator 2, which was caused mostly by their different decisions about short (1–3 seconds) transition segments in music (e.g., short pauses between verse and chorus).

Unfortunately, segmentation tasks are not well-adapted for formal inter-annotator agreement calculation. We perform the standard F-measure evaluation as is common in the literature [13]:

$$F_1 = 2\,\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}\,. \tag{1}$$

Table 1 shows the F-measure at 0.5 and 3 seconds. The metrics are similar to those obtained for the structural segmentation task, though a bit lower for a 0.5s window [13]. It seems that 0.5s window is too strict for these particular annotations. This might be caused by the nature of the task. Though some emotional boundaries are rather abrupt, others are smeared by a transitional musical process necessary for an emotion to modulate from one state to another.

4.3 Structural segmentation explaining emotional segmentation

In order to check how well emotional segmentation is explained by structural segmentation we compared the emotional boundary annotations to structural boundaries in the SALAMI dataset. The SALAMI dataset contains hierarchical annotations on multiple levels — musical function (verse, chorus, etc.), lead instrument, and musical similarity on large and small scale. Table 2 shows the precision, recall and F-measure obtained when predicting emotional segmentation from structure. From the table we can see that about 69 to 80% of the emotional boundaries coincide with large section boundaries. More than a half of the boundaries coincide with the lead instrument change. Small-scale similarity was not included in the table because of the abundance of small-scale boundaries (meaning close to 100% recall and very low precision). We also didn't include the 0.5s time resolution, because emotional segmentation seems to be less precise than structural and 0.5s time resolution is too detailed.

It is important to note that, with regard to F-measure, the emotional annotations when retrieved from each other have a bigger score than with any of the structural segmentation annotations.

5. SEGMENTATION METHODS EVALUATION

Segmentation methods are usually categorized into homogeneity, novelty and similarity based methods. We argue that for emotional boundary detection only the first two categories are relevant, because an emotional boundary is usually signified by changes in loudness, timbral properties, harmony, instrumentation, etc., and though it might coincide with repetitive segments (i.e., chorus), there is no straightforward connection between them. Hence, in this section we are mostly going to evaluate homogeneity and novelty based methods, namely Convex NMF [10], Mood Tracking [9], the classic method by Foote [5] and Structural Features [12]. We implemented the Mood Tracking method as described in the article, and adapted an implementation [1] of the rest for our purposes (i.e., feature extraction, thresholds etc. as described below).

All of these methods are unsupervised and take as input time-series of features extracted from audio. We extract both low (mfcc, chroma, energy, dissonance and other spectral features) and high-level (scale, tempo, tonal stability) beat-synchronized audio features using Essentia [2]. Beats are determined using the Essentia BeatTracker algorithm. All the music files have 44100 Hz sampling rate and are converted to mono. To extract low-level timbral features we use a half-overlapping window of 100ms, and a window of 3 seconds for high level features. The features are smoothed with median sliding window, normalized, and resampled according to detected beats (see Figure 3a).

We use the same feature set to evaluate all the algorithms. Many segmentation algorithms limit themselves to using only MFCC or chroma features, but through experimentation with different feature sets we found that adding other spectral and high-level features significantly improves the performance on our dataset.

To combine the annotations, we decided to select only the boundaries which were marked by all the annotators with a tolerance window of 3 seconds. We will call them *common*. It can be assumed that the boundaries present in all the three annotations are the most prominent and important ones.

5.1 Summary of evaluated methods

5.1.1 Foote

Foote's method [5] relies on a self-similarity matrix (composed using pairwise sample comparisons). A short-time

[1] https://github.com/urinieto/SegmenterMIREX2014

Evaluation metric	Functions			Large scale			Instruments		
	A1	A2	A3	A1	A2	A3	A1	A2	A3
Precision @ 3	0.61	0.68	0.67	0.63	0.63	0.67	0.52	0.50	0.51
Recall @ 3	0.74	0.78	0.75	0.69	0.80	0.75	0.55	0.55	0.58
F-measure @ 3	0.65	0.71	0.69	0.64	0.68	0.69	0.50	0.50	0.55

Table 2. Retrieving emotional segmentation from structural segmentation

	C-NMF				SF				Foote				MT (enh.)			
Evaluation metric	C	A1	A2	A3	C	A1	A2	A3	C	A1	A2	A3	C	A1	A2	A3
P@3	.27	.35	.36	.47	.33	.43	.49	.57	.31	.38	.41	.50	.18	.28	.27	.34
R@3	.71	.67	.69	.67	.67	.61	.68	.61	.72	.67	.72	.66	.43	.47	.47	.41
F@3	.36	.43	.45	.52	.41	.47	.55	.56	.39	.45	.50	.53	.23	.34	.33	.35

Table 3. Performance of investigated methods on emotional segmentation task (F-measure).

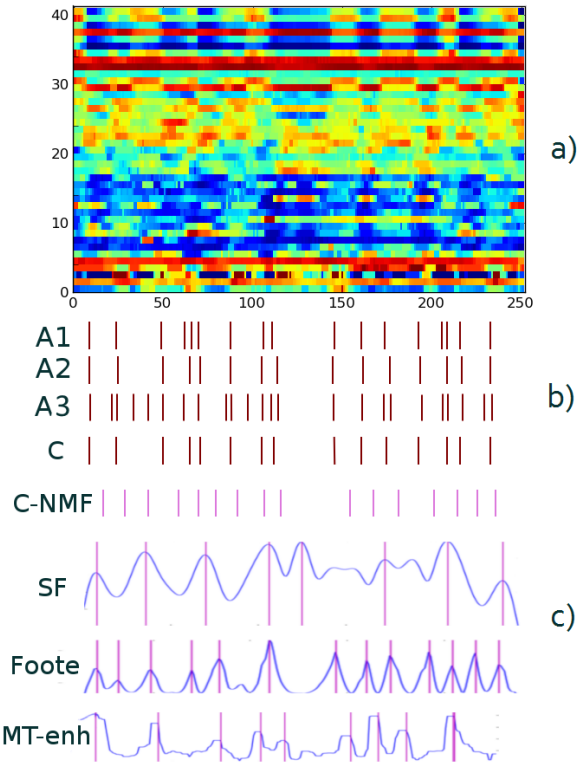

Figure 3. An illustration of the boundary detection process on the *Radetzky March* by J. Strauss Sr.. a) Beat-synchronized features. b) Annotations. c) Novelty curves and detected boundaries.

Gaussian checkerboard-shaped kernel is slided over the diagonal of the matrix, resulting in a novelty curve. The boundaries are detected by picking the peaks on the novelty curve. We experimented with different distance measures to compute the SSM and found that standardized euclidean distance gave the best results, which is computed between two vectors u and v as follows:

$$\sqrt{\sum (u_i - v_i)^2 / V[x_i]}\,, \tag{2}$$

where V is the variance vector; V[i] is the variance computed over all the ith components of the points. We set the size of the checkerboard kernel to the size of the average emotionally stable segment — 20 seconds.

5.1.2 *Convex NMF*

The Convex non-negative matrix factorization method [10] (Convex NMF) uses a convex variant of NMF in order to divide the audio features into meaningful clusters. This algorithm focuses both on finding segments and grouping them by similarity. If a NMF of input feature matrix X is FG, Convex NMF adds a constraint to the columns of the matrix F $(f_1, f_2, ..., f_n)$ that the columns should become convex combinations of the features of X:

$$f_i = x_1 w_{1j} + ... + x_p w_{pj} = X w_j, \quad j \in [1:r], \tag{3}$$

where x_p is a column of matrix X, r is a rank of decomposition, and $w_{ij} \geq 0, \sum_j w_{ij} = 0$. This makes columns f_i interpretable as cluster centroids. We set the rank of decomposition to 2.

5.1.3 *Mood Tracking*

A method by Lu et al. [9] finds boundaries by comparing the audio features extracted from the two consecutive windows of 16 seconds and computing a difference between them. A novelty curve is formed using an obtained difference feature, from which peaks are picked. The difference between the consecutive windows is computed using divergence shape measure:

$$D_{i|i+1} = \frac{1}{2} \operatorname{Tr}\left[(C_i - C_{i+1})(C_{i+1}^{-1} - C_i^{-1})\right], \tag{4}$$

where C_i and C_{i+1} are the covariance matrices of features of windows i and $i+1$. Then, confidence of boundary is computed:

$$\text{Conf}_{i|i+1} = \exp\left(\frac{|D_{i|i+1} - D_{\text{mean}}|}{D_{\text{var}}}\right), \tag{5}$$

where D_{mean} and D_{var} are the mean and variance of all divergence shapes for this song. From a list of boundary confidences the boundaries are retrieved by satisfying conditions of being a local maximum and exceeding a local adaptive threshold.

We implemented the method as it was described in [9], but it didn't work well in its original form on our data. The constraint of 16 seconds was too conservative and adaptive threshold window was too narrow. We describe an optimized version below. The optimized version performs on average about 10% better than the original method, and we only show the performance of the optimized version in Table 3.

5.1.4 Enhanced Mood Tracking

The best results with Lu et al. method were obtained using a window of 4 seconds to compute the divergence shape measure. We smoothed the boundary confidence vector with a median filter before peak picking. To pick the peaks, we select a maximum in a neighbourhood of 10 beats in case it exceeds both of the two threshold – a moving average and half of the global average.

Though the performance of the method improved with modifications, it still performed worse than other methods in our evaluation.

5.1.5 Structural Features

The Structural Features (SF) method is both homogeneity and repetition based. It uses a variant of lag matrix to obtain structural features. The SF are differentiated to obtain a novelty curve, on which peak picking is performed. The SF method calculates self-similarity between samples i and j as follows:

$$S_{i,j} = \Theta\left(\varepsilon_{i,j} - ||x_i - x_j||\right), \quad (6)$$

where $\Theta(z)$ is a Heaviside step function, x_i is a feature time series transformed using delay coordinates, $||z||$ is a Euclidean norm, and ε is a threshold, which is set adaptively for each cell of matrix S. From matrix S structural features are then obtained using a lag-matrix, and computing the difference between successive structural features yields a novelty curve.

5.2 Evaluation results

Table 3 shows the results obtained in evaluation. We only use a tolerance window of 3 seconds, because for our dataset a tolerance window of 0.5s is too strict. From the table we can see that the SF method consistently shows the best results in terms of F-measure. The method proposed in [9] consistently shows the worst results.

6. DISCUSSION

In this paper we discussed the problems associated with dynamic MER and argued that these problems originate from the unnaturally low time resolutions that dynamic MER is usually dealing with (Section 3). We proposed to move to bigger time resolutions by tracking emotionally stable segments over time and identifying transitions between them. We call this problem emotion based segmentation, and conduct a formal evaluation procedure, which has not been done before for this task.

We collected data on emotional segmentation of music; in total about 2000 emotional boundaries were annotated. In general, the annotators could agree rather well when identifying stable emotional segments, the inter-annotator F-measure was comparable to the one obtained for, supposedly less ambiguous, structural segmentation task, except for the very high resolution level (0.5 s). In terms of F-measure the emotional annotations coincide with each other better than any of the structural segmentation levels. That means that there exist some robust and important emotional boundaries which are not explained by structural segmentation.

We compared emotional and structural segmentation and found that emotional boundaries coincide with structural boundaries very often. About half of the emotional boundaries were accompanied by a lead instrument change. Approximately 25% of the emotional boundaries did not coincide with the structural boundaries. For instance, an emotional change can occur within a structural section due to a modulation to a different tonality.

We found that the average length of stable emotional segment is approximately 20 seconds. This finding could be used to calculate a suitable length of musical excerpts to be employed for MEVD algorithms development and evaluation. Namely, we believe that length of such excerpts should be several times bigger than 20 seconds.

We evaluated different unsupervised segmentation algorithms on the task of emotional segmentation and found that the SF method performed best. This segmentation method is different from the second best Foote's method by incorporation of repetition-based criteria along with homogeneity-based ones. This shows that sequences of emotionally stable segments, probably, repeat in the same way as structural sequences, and therefore repetition-based cues are useful for emotional boundary detection. This finding goes against our initial intuition that novelty and homogeneity cues must be the only ones important to detect emotional change. The Mood Tracking method was demonstrated to be least useful. This method only uses a very narrow local context to find the discontinuities in a feature matrix, which appears to be not enough. We also found that employing higher level audio features, along with traditional chroma features and MFCCs, improves the performance of the methods on emotional segmentation task.

Though SF's method performed reasonably well, its performance was still much worse than the performance achieved by best methods for structural segmentation, which is a more mature area of research now. Developing better emotional segmentation methods is a crucial task to enable applying static MER algorithms to real world problems. We leave this task for future work, which can be facilitated by the data provided in this study.

7. ACKNOWLEDGEMENTS

We thank Kayleigh Hagen and Valeri Koort for assistance with the data annotation. This research was supported by COMMIT/.

8. REFERENCES

[1] Anna Aljanaki, Mohammad Soleymani, and Yi-Hsuan Yang. Emotion in music task at mediaeval 2014. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, 2014.

[2] D. Bogdanov, N. Wack, E. Gomez, S. Gulati, P. Herrera, and O. Mayor. Essentia: an audio analysis library

for music information retrieval. In *International Society for Music Information Retrieval Conference*, pages 493–498, 2013.

[3] Eduardo Coutinho and Angelo Cangelosi. Musical emotions: Predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion*, 11(4):921–937, 2011.

[4] J. Deng and C. Leung. Dynamic time warping for music retrieval using time series modeling of musical emotions. *IEEE Transactions on Affective Computing*, PP(99), 2015.

[5] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference of Multimedia and Expo*, pages 452–455, 2000.

[6] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. Rwc music database: Popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval*, pages 287–288, 2002.

[7] M.D. Korhonen, D.A. Clausi, and M.E. Jernigan. Modeling emotional content of music using system identification. *IEEE Transactions on Systems, Man, and Cybernetics*, 36(3):588–599, 2006.

[8] C. L. Krumhansl. Plink: thin slices of music. *Music Perception: An Interdisciplinary Journal*, 27(5):337–354, 2010.

[9] L. Lu, D. Liu, and H.J. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):5–18, 2006.

[10] O. Nieto and T. Jehan. Convex non-negative matrix factorization for automatic music structure identification. In *Proceedings of the 38th IEEE International Conference on Acoustics Speech and Signal Process- ing*, pages 236–240, 2013.

[11] E. M. Schmidt and Y.E. Kim. Modeling musical emotion dynamics with conditional random fields. In *Proceedings of the 2011 International Society for Music Information Retrieval*, 2011.

[12] Joan Serra, Meinard Muller, Peter Grosche, and Josep Lluis Arcos. Unsupervised music structure annotation by time series structure features and segment similarity. *IEEE Transactions on Multimedia, Special Issue on Music Data Mining*, 2014.

[13] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and J. S. Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 555–560, 2011.

[14] J. A. Speck, E.M. Schmidt, B.G. Morton, and Y.E. Kim. A comparative study of collaborative vs. traditional annotation methods. In *Proceedings of the 2011 International Society for Music Information Retrieval Conference*, 2011.

[15] J.-H. Su, Y.-C. Tsai, and V. S. Tseng. Empirical analysis of multi-labeling algorithms for music emotion annotation. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2013.

[16] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multilabel classification of music into emotions. In *Proc. 9th International Conference on Music Information Retrieval*, pages 325–330, 2008.

[17] A. Wieczorkowska, P. Synak, and Z. W. Ras. Multilabel classification of emotions in music. In *Intelligent Information Processing and Web Mining*, pages 307–315, 2006.

[18] B. Wu, E. Zhong, A. Horner, and Q. Yang. Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 117–126, 2014.

[19] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen. What is the best segment duration for music mood analysis. In *In Proceedings of the IEEE International Workshop on Content-Based Multimedia Indexing*, pages 17–24, 2008.

[20] Y.-H. Yang and H. H. Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology*, 3(3), 2012.

[21] Y.-H. Yang, C.-C. Liu, and H. H. Chen. Music emotion classification: A fuzzy approach. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pages 81–84, 2006.