STRUCTURAL SEGMENTATION OF HINDUSTANI CONCERT AUDIO WITH POSTERIOR FEATURES

Prateek Verma, Vinutha T.P., Parthe Pandit,Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology Bombay, Mumbai 400076, India
prateek119@gmail.com, {vinutha, parthepandit, prao}@ee.iitb.ac.in

ABSTRACT

Structural segmentation of music involves identifying boundaries between homogenous regions where the homogeneity involves one or more musical dimensions, and therefore depends on the musical genre. In this work, we address the segmentation of Hindustani instrumental concert recordings at the highest time-scale, that is, concert sections marked by prominent changes in rhythmic structure. Tempo features are effectively combined with energy and chroma features motivated by musicological knowledge and acoustic observations. Posterior probability features from unsupervised model fitting of the frame-level acoustic features are shown to significantly improve robustness to local acoustic variations. Finally, two diverse change detection criteria are combined to obtain a superior segmentation system.

Index Terms— music segmentation, structural segmentation, posterior features

1. INTRODUCTION

Structural segmentation of music refers to locating the temporal boundaries between contrasting sections. The contrasts appear along one or more musical dimensions such as melody, harmony, rhythm and timbre. It is therefore expected that the method applied to structurally segment music depends upon the musical genre under consideration. In particular, the features used to detect the change at a segment boundary can be designed to exploit the peculiar musicological and acoustic characteristics of the signal. The homogeneity of a particular region with respect to the chosen features, captured in a self-distance matrix, is a popular method for the detection of the structural boundaries [1]. The structural segmentation of the concert audio can serve important functions including navigation and automatic music summarization [2].

In this paper, we focus on the North Indian classical instrumental concert which follows an established structure via a specified sequence of sections viz. *alap-jor-jhala-gat* [3,4].The first 3 are improvised sections (also together called the "*alap*"). The *gat* or composed section follows the *alap*. The basic dimensions of Indian classical music are melody and rhythm but at the larger time-scales relevant to concert sections, the rhythm is the prominent distinguishing attribute. The structure originated in the ancient style of *dhrupad* singing where a raga performance is subdivided unequally into the improvised and composed sections with the former taking up even up to 80 percent of the performance time. The improvised section (the *alap*), that can last close to an hour, is further divided into *alap*-proper (unmetered, no regular pulse), the *jor* (steady pulsation) and the *jhala* (faster pulsation). We consider solo concerts of stringed instruments, a major component of the Indian instrumental repertoire. The *tanpura* or *drone* is the only accompanying instrument throughout the improvised region, and hence there are no variations in timbral properties as such across the sections. On the other hand, the onset of the *gat* is clearly signaled by the entry of the percussion (*pakhawaj* or *tabla*) and the associated prominent change in timbral texture.

In contrast to timbre and harmony, the use of rhythmic cues in music segmentation has been limited [1]. Jensen [5] used the periodicity of detected note onsets in an autocorrelation similarity matrix to detect rhythm change boundaries. Grosche et al. [6] implemented tempo based music segmentation via a cyclic tempogram that helped resolve octave ambiguities due to metrical hierarchy. Recent work on segmentation for Indian music proposed timbre and rhythm cues for Carnatic vocal concerts [7] where the presence or absence of percussive accompaniment was an important cue to segment type. The segmentation of the North Indian *alap* is more challenging, however, with only the main melodic instrument and the drone prevailing across the sections.
In this paper, we focus on the segmentation of the *alap*, or improvised region, into its sections. Commercially available live concert recordings typically provide the *alap* and *gat* in separate tracks but the *alap* sub-segments are rarely separated. Given the prominence of pulsation rate in discriminating the sections, we start with rhythmic cues, and then explore additional features based on observed acoustic characteristics to further increase the robustness of segmentation. We demonstrate the effectiveness of posterior probability features derived by unsupervised training. The overall system follows the template of a typical music structural segmentation system [1, Fig. 5] with the modules of frame-based feature extraction followed by boundary detection via a self-distance matrix (SDM). The new contributions in this work are the (i) feature design for the genre of interest, (ii) more general transformation to posterior probability features to improve the reliability of the SDM, and (iii) the use of a second pass to reduce over-segmentation based on a statistical model selection criterion applied to the feature sequence. The relative performance improvements are demonstrated through measures of achieved homogeneity of segments and boundary detection accuracy. Since this is an unsupervised approach, it generalizes to any musical piece as long as suitable genre-specific features are employed. Further specific assumptions based on musicological knowledge exploited in the present work are that there are 3 distinct concert sections, each of which is temporally continuous.

The next section presents the design of discriminatory features. Starting with a standard representation of rhythm, the autocorrelation function of the onsets sequence, new reduced and complementary features are motivated based on genre-specific observations. The feature sets are compared via the purity of

segments obtained by the corresponding SDM based boundary detection presented next. Finally, the best performing features are incorporated in a two-pass segmentation system which is evaluated on a database of commercially available concert *alaps*.

2. FEATURE REPRESENTATION

The change in the pulsation rate is a prominent indicator of the transition between *alap* sections. We present rhythm based features for our segmentation task together with new features motivated by observations on the Hindustani stringed instrument concert audios.

2.1 Rhythm based features

A rhythm representation can be obtained by observing the onsets over a suitably long texture window. The onsets correspond to the string plucks in the audio and appear as time-localized wideband events in the short-time spectrum computed with a 30 ms data window at 10 ms intervals. The spectral flux obtained by the differencing of successive frames' magnitude spectra is half-wave rectified to provide an onset detection function across time [8]. The periodicity inherent in the onset detection function is captured by the autocorrelation function (ACF) computed over a suitable region. Although the tempo ranges are such that the inter-onset duration is no greater than 0.5 s, we compute ACF over 20 s segments with 1 s hop in order to smooth over the brief intermittent pauses that occur throughout the performance. The rhythmogram [9] provides a powerful visualization of the time-varying periodicities captured by the ACF, as shown in Fig. 1 (a) for a sitar concert *alap*. We observe the absence of periodic structure in the *alap*-proper. The *jor* and *jhala* sections are characterized by strong periodicity as seen by the regularly spaced peaks across the ACF lag axis. The decreasing spacing of peaks indicates increasing rate of onsets, or increasing tempo. The boundaries between sections are clearly visible in the rhythmogram suggesting that the ACF could serve as a feature vector for the segmentation.

The ACF feature vector however is high dimensional and contains information about the rate of pulsation only since the rhythm structure is devoid of any metrical hierarchy. It makes sense to reduce it to a tempo value. A reliable method of tempo detection with minimal octave error is peak detection from product of the ACF and DFT [10]. It is necessary in our task to include a confidence measure with the tempo value to signal low reliability such as expected in the *alap*-proper segment. The normalized ACF peak corresponding the detected tempo is used as a measure of salience, and detected tempo values with low salience (< 0.1) are set to zero. We thus obtain the two dimensional vector [tempo, salience] as a compact alternative to the full ACF vector without loss of information, due to absence of metrical structure.

2.2 Additional acoustic features

Observations of the short-time spectra across the musical sections reveal further properties that complement the rhythm features as presented below. The features, computed every 1 s over a 20 s texture window, are normalized to zero-mean and unit variance.

Tempo slope: A characteristic of the *jor* section is the continuously increasing tempo (as seen in Fig 1(a)). A characteristic of the *jor* section is the continuously increasing tempo. A piece-wise linear fit to the tempo trajectory over time

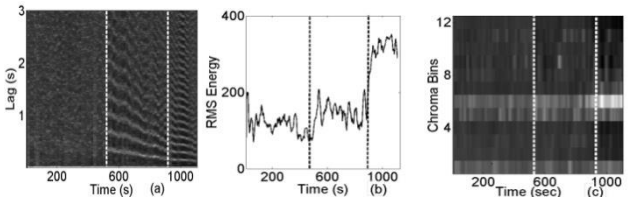

Fig. 1: Analysis of a sitar concert *alap*. (a) rhythmogram of onsets sequence ACF (b) short time energy (c) chromagram. Dashed lines indicate the ground-truth segment boundaries: *alap-jor* and *jor-jhala*.

across the 20 s texture window provides a useful feature for discriminating the *jor* from the two neighboring sections.

Short-time energy: The *jhala* section is characterized by rapid string plucks including the use of *chikari* (drone strings) filling in the intervals in between. This contributes to an audible increase in overall signal intensity. Fig. 1(b) shows the frame-level short-time energy mean over the 20 s texture windows at 1 s hop. We see a sharp increase in energy at the *jor-jhala* boundary.

Chromagram variance: The near continuous presence of the *chikari* (drone strings) in the *jhala* results in a strong presence at the tonic note of the concert. This can be captured by a harmony representation such as the chroma vector [9]. Fig 1(c) shows the chromagram of the concert where a high peak is seen to set in at bin 6 of the 12-bin chroma vector in the *jhala* section. The peakiness of the chroma vector is represented by its variance.

Time: To exploit the fact that each concert section is temporally continuous, we add to the above frame-level features, the temporal location with respect to concert duration. Finally, we have a 6-dimensional feature vector (5 acoustic features + time value) obtained every 1 s across the audio signal.

2.3 Posterior features

In order to detect homogeneity, it is necessary that the feature representation is stable and not subject to noisy fluctuations. Transforming the acoustic features to class-conditional probabilities (known as posterior features) could help achieve this, as has been observed in the context of speech recognition [11]. This does not appear to have been previously applied in music segmentation. We use unsupervised ML training, via the EM algorithm, of a 3-mixture GMM (motivated by the presence of 3 acoustic classes in our data: *alap*, *jor* and *jhala*) on the given concert feature vector sequence. The low dimensionality of the feature vector facilitates the concert based unsupervised training of models.

A posterior probability vector, q_i, corresponding to the frame X_i of the audio signal is obtained, as in Eq. (1), by computing the posterior probability of the feature vector with respect to each of the three Gaussian distributions (C1, C2 and C3) of the trained 3-component GMM [11.12].

$$q_i = (P(C_1|X_i), (P(C_2|X_i), \big(P(C_3|X_i)\big) \qquad (1)$$

3. SEGMENT BOUNDARY DETECTION

Fig. 2 shows the modules of our segmentation system. The self-distance matrix (SDM) contains elements that represent the "distance" between feature vectors of two frames (corresponding to the row and column indices of the element) [13]. A homogenous segment of length N would thus appear as an *N*x*N* block of low

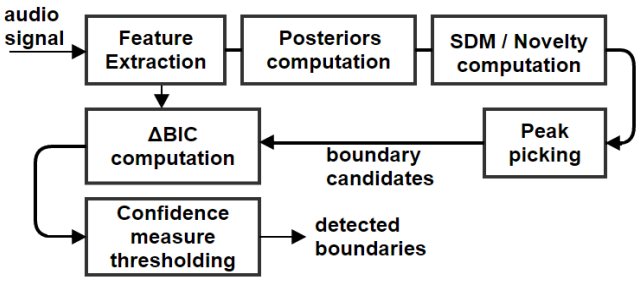

Fig. 2: Concert audio segmentation system

distance values. The distance measure typically used is either an element-wise dot product or the Euclidean distance between feature vectors [9, 13]. Next, points of high contrast in the self-distance matrix are captured by convolution along the diagonal with a checker-board kernel of dimensions matched to the time-scale of interest [1, 13].The one-dimensional plot resulting from the convolution is called a novelty function (NF) whose peaks indicate the contrasting or boundary points in the feature vector stream. Peak detection from the novelty function can result in over-segmentation however. This can be counteracted by a second pass that selectively merges segments based on statistical modeling of entire segments. We describe each stage of the segmentation system next.

3.1 Novelty peaks from self-distance matrix

Fig. 3 shows an SDM computed using the Euclidean distance between feature vectors for each of our feature sets. (The dot product showed no particular advantage over the Euclidean distance for our feature sets.) We observe the *alap-jor-jhala* concert structure corresponding to the 3 homogenous square regions around the diagonal. The section transitions get clearer as we move from the ACF to the proposed feature set, and further, a marked improvement is seen in the SDM of the posteriors as seen from the block boundaries with respect to the ground-truth segments. Convolution with a kernel of width 100 s (± 50 s, chosen considering that the section durations are on the order of few hundreds of seconds) leads to the corresponding smoothened novelty functions shown in the lower half of Fig. 3. Consistent with the improvement in the SDM, the posteriors' NF is characterized by more prominent peaks compared to the other two.

The novelty functions obtained from the ACF and that from the features vector are both rather noisy. Clear peaks at the ground-truth section boundaries are hard to discern. This can be attributed to the fluctuations in the distance values within the blocks of the SDM corresponding to the local variations in ACF and in the features. The musical sections are performed by a human and therefore bound to exhibit local acoustic variabilities. Apart from this, random variations can be expected in specific features irrelevant in a given section (e.g. tempo and slope in *alap*-proper). The texture window of 20 s is already large, and any increase to achieve further smoothing of the features would be undesirable in other to preserve time resolution in the boundary localization. The posterior features help alleviate this trade-off as follows. While the noisy fluctuations occur in some or other feature values, thus adversely affect the homogeneity of the features SDM, these would ideally map into uniformly low class-conditional probability values in the posteriors vector. Thus distances in posterior probabilities vector space can be expected to be much more uniformly dependent on the underlying musical structure. Peak picking from the novelty function is the next crucial step. In order to eliminate

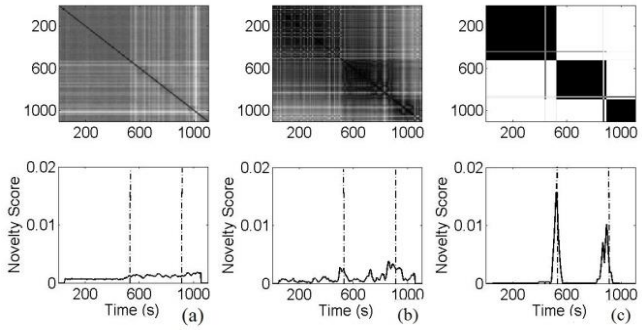

Fig. 3 : SDM (top) and novelty functions (bottom) for the sitar concert of Fig. 1. (a) ACF vectors (b) acoustic feature vectors (c) posterior feature vectors

spurious and closely-spaced peaks from consideration, we apply a "local-peak-local-neighborhood" smoothing with the cascade of a median filter and moving average [14]. The neighborhood is chosen to be ±10 s, and the final peaks and their strengths are chosen from the unsmoothened NF with a threshold of 0.1 applied to the normalized strengths.

3.2 Segment merging with second pass

The detected NF peaks of the first pass are considered as candidate boundaries in a second pass where a different method based on a statistical model selection criterion is used to compute a confidence measure. In the second pass, we use a statistical model selection criterion to decide whether a particular instant is a point of change.

The ΔBIC has often been used to solve the problem of acoustic change detection by comparing the penalized likelihoods of data modeled with two separate models (indicating change) against just a single model (indicating homogeneity) [15]. Our statistical model is a diagonal covariance multivariate Gaussian distribution for the feature vectors (comprising the 5 acoustic features: tempo, salience, tempo slope, short-time energy, chromagram variance). As shown in Fig. 2, for every candidate boundary (first pass prediction from the SDM-NF), we consider the 20s vicinity for positive local maxima in ΔBIC over windows of duration 300 s to 1200 s. If no local maximum is found in this neighborhood, the candidate is rejected. We next obtain a confidence measure for selected candidates as the product of novelty score and the ΔBIC score, normalized over the concert. A threshold of 0.05 is applied to further reject spurious boundaries. For the surviving candidates, the final estimated boundary is located at the corresponding ΔBIC local maximum.

The uncorrelated behaviors of the SDM-NF and ΔBIC boundary predictions are evident in Fig. 4 (where, for illustration, we have shown the ΔBIC local maxima computed across all frames of the sitar concert). We see that both methods produce false alarms but that these are non-overlapping. The true boundaries are represented in both methods of segmentation. This diverse behavior allows us to combine the two predictors to achieve a better performance over either one as is demonstrated in the experiments.

4. EXPERIMENTS

4.1 Datasets and evaluation measures

Full length *alaps* were extracted from commercial recordings of live concerts for our segmentation task with a goal to segment the

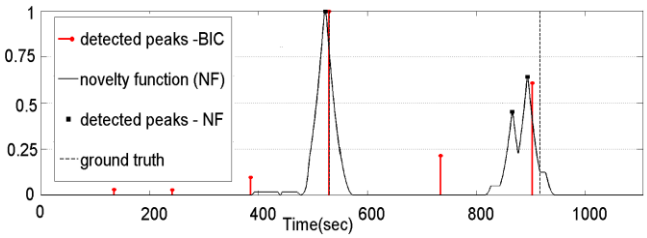

Fig. 4: Novelty function and detected boundaries from the SDM shown with detected boundaries from ΔBIC local maxima.

recording into *alap*-proper, *jor* and *jhala* sections. A total 15 concerts across 2 Indian classical stringed instruments: the plucked sitar (7 concerts) and the hammered santoor (8 concerts) performed by 7 well-known artists were manually annotated independently by 3 musicians. The concert sections are of unequal duration with *alap*-proper typically being the longest, and *jhala* the shortest. The durations vary considerably across concerts, and the median values are: *alap*-proper (900 s), *jor* (600 s) and *jhala* (300 s).The judges differed from each other in the precise location of the boundary due to the nature of the style. When moving from one section to the next, the artist often starts with intermittent hints of the coming section before settling in completely several seconds later. The inter-judge differences in labeling were found to be within a 20 s interval for all segment boundaries except for two sitar concerts where the ambiguous regions were more extended. The features proposed in this work were designed for stringed instruments. Given the similarity of the concert structure with that of *dhrupad* vocal genre (where the structure originates), we also tested the system on an available dataset of 10 vocal music concert *alap*s in *dhrupad* style. The rhythmic pulsation cues correspond to syllable rates in dhrupad vocal [3]. Our acoustic features were restricted to tempo and salience alone since the additional features of Sec. 2.2 are not applicable to vocal music. Further distinctive features of dhrupad vocal concert segments are to be researched.

With homogeneity being the basis of the structural segmentation, the feature sets are first compared in terms of the obtained quality of clustering of the frames [16]. We use the "average cluster purity" (acp) and "average annotation purity" (aap) measures proposed for song segmentation [17] that consider to what extent the frames of an estimated (or annotated) segment bear a consistent correspondence with an annotated (or estimated) segment. Next, the overall system performance is measured in terms of the hit rate (HR) given by the ratio of number of detected true hits to the number of ground truth boundaries, and the false alarm rate (FA) given by the number of detected boundaries that are not ground truth boundaries. Based on the observed inter-judge differences, we consider the detected boundary to be a true hit if it is the strongest peak within 20 s of the average subjective marking, and making a relaxation to 30 s for the 2 boundaries with larger ambiguous zones as observed in the subjective labeling.

4.2 Segmentation performance

Table 1 compares the average achieved purity for the three different feature sets based on the novelty function peaks obtained in the first pass. We observe that all feature sets attain high cluster purity (acp) indicating that under-segmentation is minimal. The aap, on the other hand, reflects the extent of over-segmentation. We see a significant improvement with the posterior features, while the direct use of features is somewhat better than the ACF. Based on this observation, we evaluated boundary detection

	ACF		Features		Posteriors	
	acp	**aap**	**acp**	**aap**	**acp**	**aap**
I	0.97	0.11	0.97	0.13	0.97	0.77
V	0.97	0.13	0.95	0.24	0.97	0.60

Table 1: Purity measures on instrumental (I) and vocal (V) concerts with different feature sets after the first pass.

Concert type (Number)	**Total Duration**	**1st pass**		**2nd pass**	
		HR	**FA**	**HR**	**FA**
I(15)	433 min	30/30	64	29/30	8
V (10)	406 min	16/20	63	14/20	10

Table 2: Segmentation results after each pass using posterior features in the first pass, and the acoustic features in the second.

accuracy on the first and second pass outputs using posterior features alone. The goal of the second pass is reduce over-segmentation further by eliminating the false boundaries of the SDM-NF peak detection.

Table 2 indicates that the second pass is very effective in reducing the over-segmentation errors of the first pass in both instrumental and vocal concerts. This was consistent with an observed improvement in the aap score to 0.91 (from 0.77) and to 0.84 (from 0.60) after the second pass. There is a single missed detection introduced in the instrumental concerts by the second pass. This was identified as being due to an unusually short *jhala* section (70 s) in one of the *santoor* concerts, making for insufficient data for the pdf modeling for the ΔBIC. The surviving false alarms were observed to lie in regions where the audio recording of the live concert was reverberant. The vocal concerts show a lower overall hit rate indicating a need for more distinctive features as well as better sung syllable onset detection.

5. DISCUSSION AND CONCLUSION

We see that features designed to exploit musicological and acoustic distinctions of a genre have the potential to achieve the accurate structural segmentation of music. This needs to be validated on a large dataset. Posterior features, or the class-conditional probabilities obtained from unsupervised GMM fitting of frame-level acoustic features, were shown to improve the homogeneity of the representation. This general outcome is expected to be applicable to any music segmentation problem where individual features are diverse and motivated by the characteristics of different underlying sections. When the number of musical sections is not known, methods for the optimal selection for number of GMM mixtures need to be incorporated. Finally, the combination of two different change detection methods helped achieve an effective two-pass system. Both methods, the SDM-NF from the posteriors feature sequence and the ΔBIC, were based on the same assumption of a Gaussian pdf for each section's feature vectors. However the precise change detection criteria, one a more local differencing and the other from a global fitting of pdfs, were different enough to cause the errors to be distinct, and so helped in exploiting diversity to achieve a superior combination system.

Acknowledgement:

This work received partial funding from the ERC under the European Union's 7th Framework Programme (FP7/2007-2013)/ ERC grant agreement 267583 (CompMusic).

6. REFERENCES

[1] J. Paulus, M. Müller, and A. Klapuri: "State of the Art Report: Audio-Based Music Structure Analysis", in Proceedings of the International Symposium on Music Information Retrieval (ISMIR), 2010, pp 625-636.

[2] G. Peeters, A. La Burthe and X. Rodet: "Toward Automatic Music Audio Summary Generation from Signal Analysis", in Proceedings of the International Symposium on Music Information Retrieval (ISMIR), 2002, pp 94-100.

[3] B.C. Wade:"Music in India – The Classical Traditions, Chapter 7: Performance Genres of Hindustani Music", Manohar Publishers, New Delhi, India, 2008.

[4] Deepak Raja's World of Hindustani Music, Raga Performance on Sitar and Sarod, http://swaratala.blogspot.in/2013/08/raga-performance-on-sitar-and-sarod.html, accessed 10/2014.

[5] K. Jensen, J. Xu, and M. Zachariasen: "Rhythm-based segmentation of popular Chinese music", in Proceedings of the International Symposium on Music Information Retrieval (ISMIR), 2005, pp 374-380.

[6] P. Grosche, M. Muller, and F. Kurth, " Cyclic tempogram-A mid-level tempo representation for music signals", in Proceedings of ICASSP, 2010, pp 5522-5525.

[7] H.G. Ranjani and T.V. Sreenivas: "Hierarchical classification of Carnatic music forms", in Proceedings of the International Symposium on Music Information Retrieval (ISMIR),2013, pp251-256.

[8] S. Dixon: "Onset Detection Revisited", in Proceedings of the International Conference on Digital Audio Effects (DAFx), Montreal, Canada, 2006, pp 133-137.

[9] K. Jensen: "Multiple Scale Music Segmentation Using Rhythm, Timbre, and Harmony", EURASIP Journal on Advances in Signal Processing, Aug 2006, pp 159-159.

[10] G. Peeters: "Template-based estimation of time-varying tempo", EURASIP Journal on Advances in Signal Processing, 2007, pp 1-14.

[11] S. Soldo, M.M. Doss, J. Pinto and H. Bourlard: "Posterior features for template-based ASR",in Proceedings of ICASSP, 2011, pp 4864-4867.

[12] Y. Zhang, and J. R.Glass: "Towards multi-speaker unsupervised speech pattern discovery",in Proceedings of ICASSP, 2010, pp 4366-4369

[13] J. Foote: "Automatic Audio Segmentation using a Measure of audio Novelty", in Proceedings of International Conference of Multimedia and Expo (ICME), New York, 2000, pp 452-455.

[14] D. Turnbull, G. Lanckriet, E. Pampalk, M.Goto: "A supervised approach for detecting boundaries in music using difference features and boosting", in Proceedings of the International Symposium on Music Information Retrieval (ISMIR), Vienna, 2007, pp 51-54.

[15] C. Scott and P. Gopalakrishnan: "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion", in Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[16] M. Levy, and M. Sandler: "Structural segmentation of musical audio by constrained clustering",IEEE Transactions on Audio, Speech, and Language Processing, 16.2, 2008, pp 318-326.

[17] H.M. Lukashevich: "Towards Quantitative Measures of Evaluating Song Segmentation", in Proceedings of the International Symposium on Music Information Retrieval (ISMIR), 2008, pp 375-380.