MUSIC STRUCTURE ANALYSIS USING A PROBABILISTIC FITNESS MEASURE AND AN INTEGRATED MUSICOLOGICAL MODEL

Jouni Paulus and Anssi Klapuri
Institute of Signal Processing
Tampere University of Technology, Tampere, Finland
{jouni.paulus,anssi.klapuri}@tut.fi

ABSTRACT

This paper presents a system for recovering the sectional form of a musical piece: segmentation and labelling of musical parts such as chorus or verse. The system uses three types of acoustic features: mel-frequency cepstral coefficients, chroma, and rhythmogram. An analysed piece is first subdivided into a large amount of potential segments. The distance between each two segments is then calculated and the value is transformed to a probability that the two segments are occurrences of a same musical part. Different features are combined in the probability space and are used to define a fitness measure for a candidate structure description. Musicological knowledge of the temporal dependencies between the parts is integrated into the fitness measure. A novel search algorithm is presented for finding the description that maximises the fitness measure. The system is evaluated with a data set of 557 manually annotated popular music pieces. The results suggest that integrating the musicological model to the fitness measure leads to a more reliable labelling of the parts than performing the labelling as a post-processing step.

1 INTRODUCTION

Western popular music pieces tend to follow a sectional form where the piece can be thought to be constructed of smaller parts (e.g., verse or chorus) which may be repeated during the piece, often with slight variations. The analysis of music structure is the process of recovering a description of this kind from audio input.

Automatic music structure analysis enables several applications, including a structure-aware music player [4] and music summarisation or thumbnailing [3, 12, 9] (see [8] for a discussion of further applications). A popular aim has been to locate a representative clip of the piece, such as the chorus, but there are also systems which aim to describe the structure of the whole piece, for example [2, 7].

In the proposed method (block diagram illustrated in Figure 1), the audio content is described using three sets of features: mel-frequency cepstral coefficients (MFCCs), chroma, and rhythmogram, to address different perceptual dimensions of music. [1] The use of multiple features is motivated by [1], which suggests that changes in timbre and rhythm are important cues for detecting structural boundaries in addition to repetitions. A large set of candidate segmentations of the piece is first constructed. All non-overlapping segment pairs are then compared based on the features and two different distance measures: one based on the average value of a feature over a segment and the other matching the two temporal sequences of features in the segments. Utilising the distance values we calculate the probability of the two segments to be occurrences of same musical part (i.e., repeats). The values are used as terms in a fitness measure which is used to rank different candidates for the piece structure description.

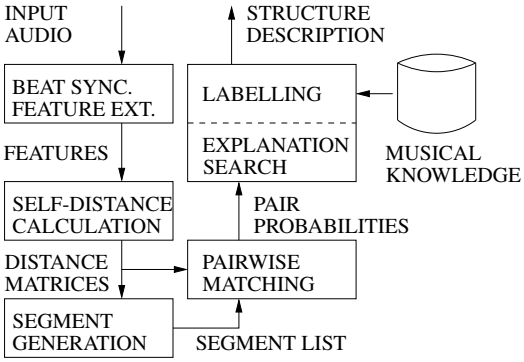

Figure 1. Analysis system block diagram.

The found structural description consists of subdivision of the piece into segments and of forming *groups* of segments that are occurrences of the same musical part. To make the description more informative, the segment groups are automatically named using musical part labels, such as "verse" or "chorus". The labelling is done by utilising a musicological model, in practice an N-gram model for musical parts estimated from a manually annotated data set. Two different strategies for the labelling are evaluated: either per-

This work was supported by the Academy of Finland, project No. 5213462 (Finnish Centre of Excellence pr ogram 2006 - 2011).

[1] A similar set of three features have been used earlier in [5], but the features were considered individually instead using a combination of them.

forming it as a post-processing step after the segments and their grouping has been decided, or by integrating it into the fitness measure for the descriptions. The latter method allows assigning the labels already while searching for the description, and enable utilising the potentially useful musical information of the part N-grams in the search.

The main original contributions of this paper concern the last two blocks in Figure 1, the pairwise matching of segments, the search algorithm, and musicological model for naming the found parts.

The performance of the proposed system is evaluated using a data set of 557 popular music pieces with manually annotated structure information. Different configurations of the overall system are studied in order to determine potential points for improvement.

2 PROPOSED METHOD

Different parts of the proposed method are described now in more detail.

2.1 Feature Extraction

The system uses three sets of acoustic features, all of them with two different time scales to provide the necessary information for further analysis. The extraction starts by estimating the locations of beat (tactus) pulses using the method from [6]. The pulse periods are halved by inserting an extra pulse between the estimated location to alleviate problems due to possible π-phase errors. The pulses are used to create a beat-synchronised time grid for making the features less sensitive to tempo variations.

MFCCs are used to describe the timbral content of the signal. They are calculated using 42-band filter bank energies which are decorrelated with discrete cosine transform. The lowest coefficient is discarded, and the following 12 coefficients are used as the feature. The MFCCs are calculated with 92.9 ms frames with 50% overlap.

The harmonic content is described with chroma, which is calculated using the method described in [14]. First, the saliences of fundamental frequencies in the range 80–640 Hz are estimated. The linear frequency scale is transformed to a musical one by retaining only the largest salience value in each semitone range. The chroma vector of length 12 is obtained by summing the octave equivalence classes. The frame division is the same as with MFCCs.

The rhythmic content is described with *rhythmogram* proposed in [5], but replacing the original perceptual spectral flux front-end with an onset accent signal obtained as a by-product of the beat estimation. After removing the mean value from the accent signal, it is divided into several seconds long, overlapping frames with spacing corresponding to the 46.4 ms frame hop on MFCCs and chroma calculation. From each frame, autocorrelation is calculated and the values below the lag of 2 seconds are retained. The values are normalised to produce value 1 on lag 0.

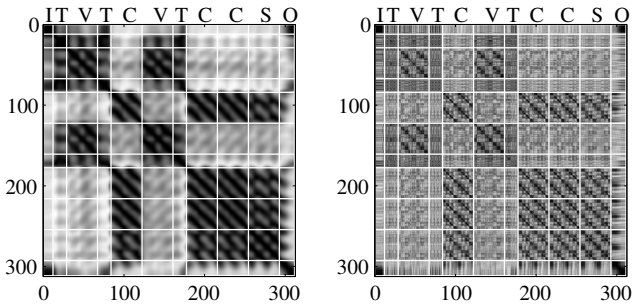

Figure 2. Examples of SDMs calculated using chroma features with two different time-scale parameters. The SDM on the left was calculated with a low cut-off frequency making different parts resemble blocks, and the SDM on right was calculated with a high cut-off making -45° stripes visible. All axes show time in seconds and a darker pixel value denotes lower distance. The overlaid grid illustrates the annotated part borders. The part labels are indicated as: intro (I), theme (T), verse (V), chorus (C), solo (S), outro (O).

Beat-synchronised feature vectors are obtained by calculating the average value of a feature between each two beat pulses. From each feature, two versions focusing on different time scales are produced. With MFCCs and chroma, this is done by low-pass filtering the feature vector series over time with second order Butterworth IIR filters. The filter cut-off frequency ω is determined by $\omega = 1/\tau$, where τ is the time scale parameter. The used values of τ for MFCCs are 8 for finer time scale and 64 for coarse time scale, for chroma the values are 0.5 and 64. The filtering is done twice along time, first forward and then backwards to double the magnitude response of the filters and to cancel out phase distortions. The rhythmogram feature is not filtered, but instead the length of the autocorrelation window is adjusted to a length which corresponds to 4 beat frames for the shorter time scale and 32 beat frames for the longer one. Each feature is normalised to zero mean and unity variance over the whole piece. The feature vector in beat frame k is denoted by $\mathbf{f_k}$. As the remaining operations are similar for all features and time scales, the same notation is used for all feature vectors. The above-mentioned time scale parameter values were selected based on the results from [10].

2.2 Self-distance Matrices

Using the extracted feature vector series, self-distance matrices (SDMs) are calculated. Each element in the SDM denotes the distance between feature vectors in frames k and l, and is calculated with the cosine distance measure $D_{k,l} = d(\mathbf{f_k}, \mathbf{f_l})$. Figure 2 illustrates the SDMs calculated for an example pieces using the chroma features.

2.3 Formal Model of Structure Description

The method searches for a *description* E for the structure of a piece S, which is represented as a sequence of *beat frames* $S = c_1, c_2, \ldots, c_K$, K being the length of the piece in beats. The sequence S can be divided into M subsequences of one or more consecutive frames $S = s_1, s_2, \ldots, s_M$, where each subsequence is a *segment* $s_m = c_k, c_{k+1}, \ldots, c_{k+L-1}$. A segment represents an individual occurrence of a musical part. For each segment s_m, there is also associated information about the *group* g_m where it belongs to. The combination of segment and group information is denoted by $(s, g)_m$. All possible segmentations and groupings of the piece form a set $\mathbb{S} = \{(s, g)_1, (s, g)_2, \ldots, (s, g)_Z\}$, where Z is the total number of all segments and groupings. The set of segments with the same group assignment represents all occurrences of a musical part (for example a chorus). A structure description $E \subset \mathbb{S}$ consists of the division of the entire piece into non-overlapping segments and of the grouping of the segments.

2.4 Border Candidates

Because the number of possible segmentations is very high, a set of potential segment border candidates is generated to narrow the search space. Not all of the candidates have to be used in the final segmentation, but the final segment borders have to be selected from the set. The main requirement for the border candidate generation method is to be able to detect as many of the true borders as possible while keeping the total amount of the borders still reasonable.

The border candidates are generated using the novelty detection method from [3]. A Gaussian tapered checkerboard kernel is correlated along the main diagonal of the SDMs to produce novelty vectors. The vectors from different features are summed and peaks in the resulting vector are searched using a median based dynamic thresholding.

2.5 Segment Distance Measures

All segments between all pairs of border candidates are generated. These segments, when replicated with all possible groupings, form the set $\mathbb{S}$, from which the final description is a subset. The fitness measure for the structure description operates on probabilities that two segments in the piece are occurrences of the same musical part. The probability is obtained by matching the features of the segments.

The matched two segments s_m and s_n define a submatrix $\tilde{D}_{[m,n]}$ of the SDM. Two distance measures for the segment pair are defined using this submatrix: *stripe* and *block* distances. The stripe distance $d_S(s_m, s_n)$ is calculated by finding the minimum cost path through $\tilde{D}_{[m,n]}$ with dynamic programming. No transition cost is applied, but instead the total path cost is the sum of the elements along the path. The local path constraint forces the path to take one step in one or both directions at a time. The distance measure is obtained by normalising the total path cost with the maximum of the two submatrix dimensions. The block distance $d_B(s_m, s_n)$ is calculated as the average element value in the submatrix $\tilde{D}_{[m,n]}$.

2.6 Probability Mapping

Given the acoustic distance $d(s_m, s_n)$ [2] between two segments s_m and s_n, it is possible to define the probability $p(s_m, s_n)$ that the segments belong to the same group (are occurrences of the same part) using a sigmoidal mapping

$$p(s_m, s_n) = p(g_m = g_n | d(s_m, s_n) = \delta) \quad (1)$$

$$= \frac{1}{1 + e^{A\delta + B}}, \quad (2)$$

where δ is the observed distance value. The sigmoid parameters A and B are determined using two-class logistic regression with Levenberg-Marquardt algorithm [13].

The probabilities obtained from the mapping of all six distance values are combined with geometric mean. Heuristic restrictions on the segment groupings can be enforced by adjusting the pairwise probabilities. An example of such restriction is to set the segment pair probability to zero if the segment lengths differ too much (ratio 6/5 was used as the limit in the evaluations). The aggregated probability value after adjustments is denoted by $\hat{p}(s_m, s_n)$.

2.7 Fitness of a Description

A probabilistic fitness measure for different structure description candidates is defined using the segment pair probabilities in the description E by

$$P(E) = \sum_{m=1}^{M} \sum_{n=1}^{M} A(s_m, s_n) L(s_m, s_n), \quad (3)$$

where

$$L(s_m, s_n) = \begin{cases} \log(\hat{p}(s_m, s_n)), & \text{if } g_m = g_n \\ \log(1 - \hat{p}(s_m, s_n)), & \text{if } g_m \neq g_n \end{cases}. \quad (4)$$

The weighting factor $A(s_m, s_n)$ is the number of elements in the submatrix $\tilde{D}_{[m,n]}$ defined by the two segments. It is used to enable comparing descriptions with different number of segments, and its intuitive motivation is "the need to cover the whole area of the SDM".

The main concepts related to the calculation are illustrated in Figure 3. The short ticks in top and left side of the figure are the locations of chunk borders. The description claims that the piece consists of five segments (A_1, A_2, B_1, B_2, C) of varying lengths, and that segments A_1 and

[2] Similar definition applies to all three feature sets and two distance measures.

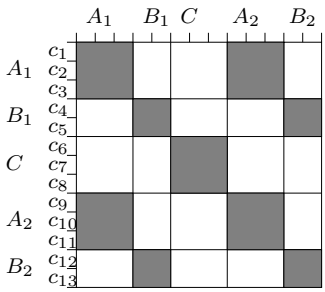

Figure 3. Illustration of the concepts regarding the calculation of the overall probability of a description. See text for details.

A_2 belong to same group as well as B_1 and B_2. There is a probability value assigned to each of the 25 submatrices, denoting the acoustic probability that the segments defining the submatrix belong to the same group. When evaluating Eq. (4), the upper alternative is used with the shaded submatrices, whereas the lower alternative is used for all the other submatrices.

The task to solve is to find the description E maximising the total log-probability of Eq. (3), given the acoustic information embedded to the pairwise probabilities

$$E_{OPT} = \underset{\mathrm{E}}{\operatorname{argmax}} \{P(E)\}. \tag{5}$$

2.8 Segment Labelling

The found description E_{OPT} defines a segmentation of the piece to musical sections and grouping of the segments which are occurrences of the same part. However, the groups have no musically meaningful labels. From the application point of view, the knowledge of the "musical role" of each part would be valuable. Musical pieces tend to follow certain forms when considering the sequences formed by the part names, e.g., "intro, verse, chorus, verse, chorus, chorus". These sequential dependencies are here modelled with N-grams of length 3. The 12 most often occurring musical part labels cover 90% of all part occurrences in the data set and are retained; the other labels are replaced with an artificial label "MISC". The estimated trigrams are smoothed using Witten-Bell smoothing [15]. The musicological information can be used in a post-processing stage to label the segments, or integrated into the fitness measure.

2.8.1 Labelling as Post-processing

In post-process labelling, the description found using Eqs. (3)–(4) is handled as a sequence of parts. Each of the groups in the description is mapped on trial with a unique musical part label in the trained vocabulary. The probability over the resulting label sequence is evaluated by calculating the cumulative Markov probability over the sequence and the most probable labelling is chosen. The post-processing labelling method is described in more detail in [11].

2.8.2 Integrated Labelling Model

As earlier experiments have shown the N-gram model to be informative in labelling the structural descriptions, an attempt is made to utilise this information already in the search of the descriptions. This is done by modifying the fitness measure of Eq. (3) to include a term containing the probabilities p_N from the N-grams:

$$P(E) = \sum_{m=1}^{M}\sum_{n=1}^{M} A(s_m, s_n) L(s_m, s_n) \tag{6}$$
$$+ \frac{w}{M-1} \sum_{o=1}^{M} \log\left(p_N(g_o|g_{1:(o-1)})\right) \sum_{m=1}^{M}\sum_{n=1}^{M} A(s_m, s_n),$$

where w is the relative weight given for the labelling model. [3] In effect, the additional term is the average part label transition log-probability weighted with the total area of the SDM. The labelling model likelihoods are normalised with the number of transitions $(M-1)$ to ensure that descriptions with different number of parts would have equal weight for the musicological model.

3 SEARCH ALGORITHM

Given all segments, segment pairs, and the probabilities that a pair of segments are from the same group, the search algorithm attempts for find the description E_{OPT} maximising the total fitness of Eq. (3) or Eq. (6). This section describes a novel search algorithm *Bubble token passing* (BTP) solving the optimisation task. An exhaustive search over all descriptions as is possible, but it is computationally too heavy for practical applications even with admissible search bounding. The exhaustive search was used to verify the operation of the greedy BTP algorithm presented next.

The proposed BTP can be seen as a variant of the N-best token passing (TP). The states in the algorithm are formed by the set $\mathbb{S}$ of all segments and all possible group assignments to each segment. In other words, the problem is finding the best path in a directed acyclic graph where the vertices are the segments with group information and there is an edge from a segment to the segments that start from the same border the previous ends. Because of the form of the total fitness measure of Eq. (3) or Eq. (6), the edge weights depend on the earlier vertices occupied by the path. Therefore, more efficient dynamic programming algorithms can not be used.

In the conventional TP, tokens are propagated between states time synchronously. Tokens record the travelled path and the associated path probabilities. At each time instant, the states take the best contained token, copy it and propagate it to all connecting states updating the token path and

[3] Values in the range 0.02–0.3 were noted to be most suitable in experiments with a subset of the evaluation data.

probability. When tokens are arriving to a state, they are sorted and only the best one is retained to be propagated at the next iteration. In N-best TP the N best tokens are propagated. [16]

The BTP operates as follows: an initial token is set to a start state, it is replicated and propagated to all connecting states updating the token information. The states store β best tokens that have arrived and propagate α tokens at the next iteration. If $\alpha < \beta$, the tokens not propagated remain in the state and they are considered for propagation in the next iteration. After some iterations, tokens start arriving to the end state. They contain the found paths (structural descriptions) and the related probabilities (fitness measures). As iterations are continued, more tokens "bubble" through the states to the end state and more description alternatives are found. It is likely that the found paths initially improve, but after a while the improving stops. The iterations can be stopped at any time, e.g., if the solution has converged or there are no more tokens in the system.

The proposed algorithm contains some beneficial properties mostly based on the two parameters α and β. As the tokens are propagated in a best-first order, the algorithm finds some reasonable solution fast and then continues improving it. The search "beam width" can be controlled with α, while β controls the the overall greediness. If the two are equal the algorithm is approximately the N-best TP. If β is set to infinity and the search is run until all tokens are out of the system, the search is exhaustive and guaranteed to find the global optimum. In experiment it was found to be sufficient to propagate a moderate amount ($\alpha = 5 - 50$) of tokens at a time while retaining larger amount of them ($\beta = 50 - 500$). With these values, the BTP was noted to find the same result as the exhaustive search in almost all of the test cases with considerably less computational cost. The exact values depend on the number of possible states and their connectivity.

When the description labelling done as a post-processing step, the search can be optimised by occupying different groups in order which eliminates much of the search space. [4]

4 RESULTS

The proposed algorithm is evaluated using 557 manually annotated pieces. The evaluation metrics are calculated from different evaluation aspects. The effect of the segmentation and border candidate generation is evaluated. In addition to them, both alternatives for the musical part labelling are evaluated. By comparing these two results, it is possible to test if the musicological knowledge is able to provide useful information for deciding which description is the best.

[4] With the number of border candidates used in the evaluations (32), and possible labels (13), there are approximately $3.1 \cdot 10^{35}$ different descriptions E. With post-processing labelling the amount is reduced to $1.1 \cdot 10^{26}$.

4.1 Data

The evaluation data set consists of 557 Western popular music pieces. [5] The pieces were selected to provide a representative sample of the music played on mainstream radio. The pieces are mainly from the pop/rock genre, but also some pieces from jazz, blues and schlager are present. For each piece, the sectional form was annotated by hand by segmenting and labelling the musical parts. The annotations were made by two research assistants with some musical background. The simulations were run using 10-fold cross-validation scheme. At each iteration, the training subset was used to train the labelling model N-grams and the distance to probability mapping function parameters. The presented results are averaged over all folds.

4.2 Evaluation Metrics

Two different metrics are used in the evaluations: frame pairwise grouping F-measure, and total frames labelled correctly. The first measure has been used in evaluation of a structure analysis algorithm in [7]. It considers all beat-frame pairs and whether or not the frames in the pair are assigned correctly in the same group. The recall rate R_r is calculated as the ratio of correctly found frame pairs with the frames assigned to the same group to the number of all segments pair from the same group. The precision rate R_p is the ratio of correctly found frame pairs to the claimed frame pairs. The F-measure is calculated from these two as $F = 2R_pR_r/(R_p + R_r)$.

The second used evaluation metric is motivated by the desired final output of the system: the result should be a segmentation to musical parts and the labelling of each segment with the appropriate musical part name. The measure is the amount of the piece assigned with the correct musical part name.

4.3 Evaluation Results

The evaluation results are given in Table 1. Results are shown for six different system configurations. First, the effect of segment boundary candidate generation (see Sec. 2.4) is studied by considering three different cases:

- "full": Fully automatic system which generates the segment border candidates using the novelty vector. The system has to decide at each candidate border should is be included, group the created segments, and label the groups. (This configuration uses all features except rhythmogram stripes.)
- "bord": The border candidate set is taken from the annotated segmentation points. Otherwise the same as above. (Uses MFCC and chroma stripes.)

[5] Full list of the pieces is available at <http://www.cs.tut.fi/sgn/arg/paulus/TUTstructure07_files.html>.

System	F (%)	R_p (%)	R_r (%)	label hit (%)
full w/ LM	61.4	64.2	63.9	39.1
full post-LM	61.7	65.9	63.1	36.1
bord w/ LM	77.0	80.5	77.8	46.8
bord post-LM	77.9	81.8	78.2	45.4
segs w/ LM	86.1	96.0	80.6	49.3
segs post-LM	86.2	95.6	81.1	46.5

Table 1. Evaluation results for six different system configurations and four evaluation measure. See text for details.

- "segs": The segmentation is taken from the ground truth. The system has to group the different occurrences of a part together and label the groups. (Uses MFCC and chroma stripes.)

Secondly, the strategies of using the musicological model are: "w/ LM" when then labelling is done during the description search using Eq. (6), and "post-LM" when the description search is done with Eq. (3) and the labelling is done as post-processing.

The results indicate that the method for generating the border candidates is currently a bottle-neck for the system, and should be considered in future work. The integrated usage of musicological model in the fitness measure does not seem to have a big effect on the result when looking at the F-measure (the difference is statistically insignificant, the level of $p < 0.05$ would require F-measure difference of at least 1.5 %-units). From the point of view of the label hit measure, the integrated labelling improves the results slightly, but the improvement is statistically significant ($p < 0.05$) only in the results obtained from fully automatic system. Comparing the results for "full" with [7] where the same F-measure was used, the performance is very similar, although it should be noted that differring data sets were used.

5 CONCLUSIONS

A method for analysing the sectional form of a music piece was presented. The method uses a probabilistic fitness measure for comparing different structural descriptions. This allows the focus of the development work to be concentrated on the definition of the fitness measure and its terms which is typically more intuitive and conceptually simpler than algorithm development. A musicological model of the sequential dependencies of musical parts was integrated to the proposed fitness measure. The integration improves the performance of the system when measuring the amount of time where the piece is assigned with the correct musical part label. A novel algorithm was presented for searching a description which maximises the defined fitness measure. The algorithm can be controlled with two intuitive parameters and its average- and worst-case performance is considerably better than that of an exhaustive search. Future work will concentrate on improving the method that generates segment border candidates since it seems to be a bottleneck of the current system.

6 REFERENCES

[1] M. J. Bruderer, M. McKinney, and A. Kohlrausch. Structural boundary perception in popular music. In ISMIR, Victoria, B.C., Canada, 2006.

[2] W. Chai. *Automated Analysis of Musical Structure*. PhD thesis, Massachusetts Institute of Technology, 2005.

[3] M. Cooper and J. Foote. Summarizing popular music via structural similarity analysis. In WASPAA, New Platz, N.Y., USA, 2003.

[4] M. Goto. A chorus-section detecting method for musical audio signals. In ICASSP, Hong Kong, 2003.

[5] K. Jensen. Multiple scale music segmentation using rhythm, timbre, and harmony. *EURASIP Journal on Advances in Signal Processing*, 2007.

[6] A. Klapuri, A. Eronen, and J. Astola. Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[7] M. Levy and M. Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 2008.

[8] N. C. Maddage. Automatic structure detection for popular music. *IEEE Multimedia*, 2006.

[9] B. S. Ong. *Structural analysis and segmentation of musical signals*. PhD thesis, Universitat Pompeu Fabra, Barcelona, 2006.

[10] J. Paulus and A. Klapuri. Acoustic features for music piece structure analysis. In DAFx, Espoo, Finland, 2008.

[11] J. Paulus and A. Klapuri. Labelling the structural parts of a music piece with Markov models. In CMMR, Copenhagen, 2008.

[12] G. Peeters. Deriving musical structure from signal analysis for music audio summary generation: "sequence" and "state" approach. In *Lecture Notes in Computer Science*, vol. 2771. Springer-Verlag, 2004.

[13] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*. MIT Press, 1999.

[14] M. P. Ryynänen and A. P. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 2008.

[15] I. H. Witten and T. C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transcations on Information Theory*, 1991.

[16] S. J. Young, N. H. Russell, and J. H. S. Thornton. Token passing: a simple conceptual model for connected speech recognition systems. Tech Report CUED/F-INFENG/TR38, Cambridge, UK, 1989.