IMPROVING MARKOV MODEL-BASED MUSIC PIECE STRUCTURE LABELLING WITH ACOUSTIC INFORMATION

Jouni Paulus
Fraunhofer Institute for Integrated Circuits IIS
Erlangen, Germany
jouni.paulus@iis.fraunhofer.de

ABSTRACT

This paper proposes using acoustic information in the labelling of music piece structure descriptions. Here, music piece structure means the sectional form of the piece: temporal segmentation and grouping to parts such as chorus or verse. The structure analysis methods rarely provide the parts with musically meaningful names. The proposed method labels the parts in a description. The baseline method models the sequential dependencies between musical parts with N-grams and uses them for the labelling. The acoustic model proposed in this paper is based on the assumption that the parts with the same label even in different pieces share some acoustic properties compared to other parts in the same pieces. The proposed method uses mean and standard deviation of relative loudness in a part as the feature which is then modelled with a single multivariate Gaussian distribution. The method is evaluated on three data sets of popular music pieces, and in all of them the inclusion of the acoustic model improves the labelling accuracy over the baseline method.

1. INTRODUCTION

This paper proposes a method for providing musically meaningful labelling to sectional parts in Western popular music using two complementary statistical models. The first one relies on the sequential dependencies between the occurrences of different parts, while the second models some acoustic properties of the them. A labelling method using the sequence model was proposed earlier by Paulus and Klapuri [9] and this paper proposes an extension that method by including also acoustic information.

In sectional form a music piece is constructed from shorter, possibly repeated *parts*. Especially many Western pop/rock pieces follow this form. The parts can be named according to the musical role they have in the piece, for example, "intro" is in the beginning of the piece and provides an introduction to the song and "verse" tells the main story of the song. Music piece structure analysis aims to provide a description of the sectional form of the piece based on the acoustic signal. Usually the description consists of a temporal segmentation of the piece to occurrences of parts, and of grouping of segments being occurrences of the same part. For a review of methods proposed for the task, see the book chapter by Dannenberg and Goto [2] or the dissertation by Paulus [8]. With the exception of few methods [6, 14], most structure analysis methods do not provide the segment groups with *musically meaningful label*, instead they only provide a *tag* for distinguishing the different groups. However, if the analysis result is presented for a user, providing also meaningful labels for the segments would be valued, as noted by Boutard et al. [1].

A method for musical part labelling given the description with arbitrary tags was proposed by Paulus and Klapuri [9]. It relies on the assumption that musical parts have sequential dependencies which are then modelled with N-grams. The method searches for the labelling that maximises the overall N-gram probability over the resulting label sequence. The obtained results indicate that such a model manages to capture useful information of the music piece structures. This paper proposes to extend that work by including acoustic information in the process. This is motivated by the frequently encountered assumption that the chorus is louder than the other parts. It should be noted that this paper does not discuss the underlying problems in defining the structural description that have been discussed by Peeters and Deruty [11], but instead studies the performance of the proposed models in replicating the labelling in the manual annotations.

The rest of this paper is organised as follows: Sec. 2 describes the labelling problem more formally, revisits the sequential modelling baseline method, and details the proposed acoustic modelling method. Sec. 3 describes the experiments for evaluating the proposed method and presents the obtained results. Finally Sec. 4 provides the conclusions of this paper.

This work was performed when the author was at the Department of Signal Processing, Tampere University of Technology, Tampere, Finland. This work was supported by the Academy of Finland, (application number 129657, Finnish Programme for Centres of Excellence in Research 2006–2011).

2. PROPOSED METHOD

This section provides a more formal definition of the labelling problem, provides a short description of the baseline method relying only on sequence modelling, and details the proposed acoustic modelling extension.

2.1 Labelling Problem

The input to the method consists of a music piece description and the acoustic signal. The description itself is a temporal segmentation of the piece and a grouping of the segments. Each of the groups is assigned with a unique tag r. When the K tags in the description are organised into a sequence based on the temporal locations of the segments, a tag sequence $r_{1:K} \equiv r_1, r_2, \ldots, r_K$ is obtained. The problem of label assignment is to find an injective [1] mapping

$$f : R \rightarrow L \tag{1}$$

from the set R of tags present in the description to the set L of musically meaningful labels. Application of the mapping is denoted with

$$f(r) = l, \tag{2}$$

and it can be done also on sequences:

$$f(r_{1:K}) = l_{1:K}. \tag{3}$$

Since any injective mapping is a valid mapping from tags to labels, the problem is to select the "best" mapping from all the possible choises. The earlier publication [9] proposed a statistical sequence model for the the labels l for selecting the mapping producing the highest model probability. This paper proposes to include acoustic information to the process of selecting the mapping function.

2.2 Markov Model Baseline Method

Some sectional forms are more common in music than the others. An example of this was presented in [9] where it was noted that almost 10% of the songs by The Beatles have the form "intro", "verse", "verse", "bridge", "verse", "bridge", "verse", "outro". Though this cannot be directly generalised to all pieces, some sequences of parts occur more frequently than others and this can be utilised in the labelling.

In sequence modelling the prediction problem is to provide probabilities for the possible continuations of a given sequence. $p\left(s_i|s_{1:(i-1)}\right)$ denotes the conditional probability of s_i to follow the sequence $s_{1:(i-1)}$. Markov models make the assumption that the process has a limited memory and the probabilities depend only on a limited length history. The length of the history is parametrised with N which provides a motivation for the alternative name of N-grams. An N-gram of length N utilises $N-1^{th}$ order Markov assumption

$$p\left(s_i|s_{1:(i-1)}\right) = p\left(s_i|s_{(i-N+1):(i-1)}\right). \tag{4}$$

[1] All tags in input sequence are mapped to a label, but each tag can be mapped only to one label and no two tags may be mapped to same label.

Given a sequence $s_{1:K}$ and the conditional N-gram probabilities the total probability of a sequence can be calculated with

$$p(s_{1:K}) = \prod_{i=1}^{K} p\left(s_i|s_{(i-N+1):(i-1)}\right). \tag{5}$$

For more information on N-grams and language modelling, see [5].

The baseline method proposed by Paulus and Klapuri [9] calculates N-grams using the musical part labels as the alphabet L, and then locates the mapping f_{OPT} maximising the overall sequential probability of (5) while conforming to the injectivity constraint:

$$f_{\mathrm{OPT}} = \underset{\mathrm{f}}{\operatorname{argmax}} \left\{ p_{\mathrm{L}}\left(f|r_{1:K}\right) \right\}, f : R \rightarrow L \text{ injective.} \tag{6}$$

In (6) $p_{\mathrm{L}}\left(f|r_{1:K}\right)$ denotes the Markov probability of the sequence resulting from applying the mapping f

$$p_{\mathrm{L}}\left(f|r_{1:K}\right) = p(f(r_{1:K})). \tag{7}$$

The combinatorial optimisation problem of (6) can be solved, e.g., in a greedy manner by applying a variant of N-best token passing algorithm proposed in [9], or by applying the Bubble token passing algorithm proposed in [10]. Both operate on the same basic principle of creating a directed acyclic graph from the parts and possible labellings, and searching a path through it. Each part in the sequence is associated with each possible label and these combinations form the nodes of the graph. Edges are created between parts that are directly consecutive in the input sequence. Paths through the graph represent label mappings, and the path with the highest probability is returned as the result. Even though the search does not guarantee finding the optimal solution, in small experiments it found the same solution as an exhaustive search with a fractional computational cost. Viterbi or similar more efficient search algorithm cannot be employed here as the mapping has to respect the injectivity and the whole sequence history affects the probabilities instead of only the limited memory of N-grams.

2.3 Sequence Modelling Issues

The number of conditional probabilities $p\left(s_i|s_{1:(i-1)}\right)$ that need to be estimated for N-gram modelling increases rapidly as a function of the model order N and the alphabet size V: there are V^N probabilities that need to be estimated. Usually, the probabilities are estimated from a limited amount of training data, and not all probabilities can be estimated reliably. This problem can be partly alleviated by applying smoothing to the probabilities (assigning some of the probability mass of the more frequently occurring combinations to the less frequent ones), or by discounting methods (estimating high-order models as combinations of lower-order models). Variable-order Markov models (VMMs) [13] attempt solving the model order problem based on the training data by setting the order independently to different subsequences. In other words, if increasing the model order does not bring more accurate information, it is not done.

2.4 Acoustic Modelling Method

The baseline method operates only on the sequential information of the musical parts and has no information of the actual content of them. However, if the acoustic signal is available, it can be utilised in the labelling. Naturally, the parts of a song differ from each other in view of the acoustic properties. This is closely related to the definition of sectional form. However, the assumption made here is that there exists acoustic properties that exhibit similar behaviour in large body of the pieces, e.g., it is often stated that "chorus" is the most energetic, or the loudest, part in a song. In addition to "chorus" being most energetic, very few other parts can be said to have any typical acoustic property. Still, e.g., "break" or "breakdown" often has considerably reduced instrumentation, thus it is expected that it exhibits a lower average loudness than the other parts. Despite this, the acoustic modelling is applied to all parts even though is might not produce meaningful information for all labels.

The proposed acoustic modelling represents the acoustic information by associating a single observation vector $\mathbf{x}_i$ to each of the musical parts, thus utilising a highly condensed representation. The input to the labelling now consists of the tag sequence $r_{1:K}$ and acoustic observations $\mathbf{x}_{1:K}$, one vector $\mathbf{x}_i$ for each part. The acoustic model considers now the likelihoods $p_{\mathrm{A}}(\mathbf{x}_i|l)$ of observing $\mathbf{x}_i$ if the musical part label is l. The overall likelihood of the mapping definition f in view of the acoustic observations $\mathbf{x}_{1:K}$ is now calculated with

$$p_{\mathrm{A}}(f|r_{1:K},\mathbf{x}_{1:K}) = \prod_{i=1}^{K} p_{\mathrm{A}}(\mathbf{x}_i|f(r_i)). \tag{8}$$

2.5 Combined Method

Assuming statistical independence, combining the two models (7) and (8) in the same function produces a new likelihood function for the mapping f

$$p(f|r_{1:K},\mathbf{x}_{1:K}) = p(\mathbf{x}_{1:K}|f(r_{1:K}))\,p(f(r_{1:K})) \tag{9}$$

$$= \prod_{i=1}^{K} p(\mathbf{x}_i|f(r_i)) \prod_{i=1}^{K} p\left(f(r_i)|f(r_{1:(i-1)})\right), \tag{10}$$

where the first term is from the acoustic observations and the latter from the N-gram models. The labelling problem can be expressed as the optimisation task

$$f_{\mathrm{OPT}} = \underset{\mathrm{f}}{\operatorname{argmax}} \{p(f|r_{1:K},\mathbf{x}_{1:K})\}, f : R \to L \text{ injective}. \tag{11}$$

The optimisation of (11) can be done with the same algorithm as the optimisation of the sequential model alone. The only required modification is to include the acoustic observation likelihoods. It should be noted that even though the problem resembles hidden Markov model decoding, the injectivity requirement violates the Markov assumption thus prohibiting the use of Viterbi decoding.

2.6 Acoustic Features

As the assumption about the globally informative acoustic property was related to the energy level or loudness, they were tested for the acoustic modelling. The energy is measured by calculating the root-mean-squared value of the signal within the part. However, in preliminary experiments it was noted that using perceived loudness instead produced better results. This is presumably because the loudness calculation addresses also the non-linear properties of human auditory system in amplitude, frequency, and temporal dimensions, the main difference being in the dynamic amplitude scale compression from representing the data in logarithmic decibel scale. [2] The calculation is done using the function `ma_sone` from the MA Toolbox by Pampalk [7]. The loudness is calculated in 11.6 ms frames with 50% overlap and the part loudness is approximated by the mean loudness of the frames within the part in question. In addition to the mean loudness also standard deviation of the framewise loudness values over the part is used to describe the dynamics of the signal. The features are normalised by dividing them by the mean over the piece making the mean over the piece to be 1. An illustration of the feature distributions is provided in Fig. 1.

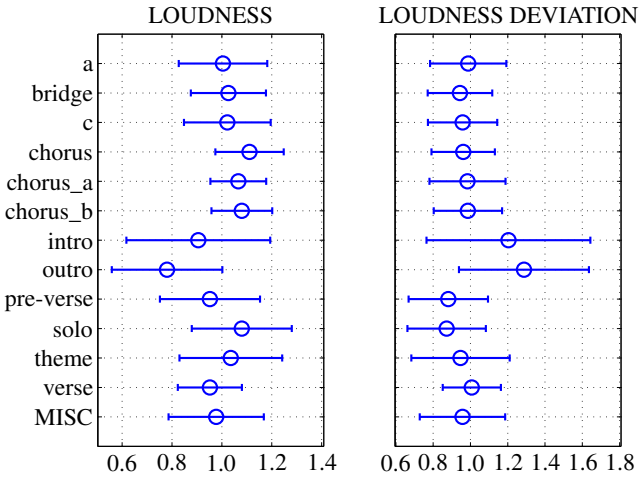

Figure 1. Statistics of the features used in data set *TUT-structure07*. The mean of all occurrences of the part is indicated with circle and the surrounding error bars illustrate the standard deviation over the occurrences. Note that the mean loudness of "chorus" and it's variations support the original assumption.

The acoustic observation likelihoods $p_{\mathrm{A}}(\mathbf{x}|l)$ are modelled as a single multivariate Gaussian distribution

$$p_{\mathrm{A}}(\mathbf{x}|l) = \frac{1}{\sqrt{(2\pi)^D|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right), \tag{12}$$

where D is the feature vector dimensionality, Σ and μ are the covariance matrix and mean vector of the estimated distribution of the part label l.

[2] The preliminary experiments included also acoustic features corresponding to the brightness (spectral centroid) and bandwidth of the signal. The various combinations of different features were tested and based on the results of the small-scale experiments, the set used was limit to loudness and it's deviation.

3. EVALUATIONS

The proposed extension is evaluated with three data sets of popular music pieces. The first set *TUTstructure07* consists of 557 pieces from various genres, mainly from pop and rock, but including also pieces from metal, hip hop, schlager, jazz, blues, country, electronic, and rnb. The pieces have been manually annotated at Tampere University of Technology (TUT). [3] The second data set *UPF Beatles* consists of 174 pieces by The Beatles. The piece forms were analysed by Alan W. Pollack [12], and the part time stamps were later added at Universitat Pompeu Fabra (UPF) and TUT. [4] The third data set *RWC pop* contains 100 pieces from the Real World Computing Popular Music collection [3, 4] aiming to represent typical 1980's and 1990's chart music from Japan and USA.

3.1 Evaluation Setup

Since the ground truth annotations in the data sets originate from different sources, the used labels also differ. For this reason the evaluations are run separately for each data set. The data sets contain relatively large number of unique part labels (e.g., *TUTstructure07* has 82 unique labels) some of which occur very rarely making the modelling more difficult. To alleviate this problem only the most frequent labels contributing to 90% of all part occurrences are retained, and the rest are replaced with an artificial label "MISC". This reduces the number of labels considerably (e.g., to 13 in *TUTstructure07*). The evaluations are run in leave-one-out cross-validation scheme and the presented results are calculated over all folds.

The performance is evaluated with per-label accuracy, which is the ratio of the sum of durations of correctly identified label occurrences to the sum of durations of all occurrence of the label, calculated over the entire data set. Similarly, the total accuracy describes how much of the entire data set duration is labelled correctly, effectively applying weighting to the more frequently occurring labels, such as "chorus".

It should be noted that the segmentation to the input tag sequence $r_{1:K}$ is obtained from the ground truth annotations instead of an automatic signal-based analysis method. This is done to be enable evaluating the accuracy of the labelling method independent of the segmentation performance.

The complementary aspects of the proposed method are evaluated: sequence modelling alone (effectively reproducing the results from [9]), acoustic modelling alone, and the two combined. The sequence modelling is attempted with N-gram length of 1 to 5 (from only prior probabilities to utilising history of length 4), and with a variable-order Markov model. The VMM method employed was decomposed context tree weighting after the earlier results, and the implementation was from [13]. These results operate as the baseline on top of which the acoustic modelling is added. The sequence modelling choises were done to follow the experiments in the earlier paper, thus providing a clear baseline for comparing the effect of the added acoustic model.

3.2 Results

The evaluation results are presented in Tables 1–3, each table containing the results for a different data set. The column denoted with "N=0" provides the result for using only the proposed acoustic model, while the other columns contain the results of the combined modelling with different N-gram lengths. The results of using only the sequence model are provided in parentheses.

The results indicate that including the acoustic information into the labelling model improves the result in some cases. In all data sets the best overall result is obtained by including the acoustic information, though the improvement in *UPF Beatles* is so small that it may not be statistically significant. [5] The same relatively small obtained improvement is observed in the results for individual labels in *UPF Beatles*. This may be because the pieces are from a single band mainly from the 1960's and thus may not exhibit all the stereotypical properties found in more modern pop music, as noted also by Peeters [11]. The improvement in *TUTstructure07* is slightly larger. It is assumed that the lower impact of the acoustic model is partly caused by the large variety of musical styles present in the data, thus the modelling assumption may not hold in all cases. The improvement due to the inclusion of the acoustic model is most prominent with the *RWC pop* data which represents more typical chart music.

4. CONCLUSIONS

This paper has presented a method for assigning musically meaningful labels music piece structure descriptions. The baseline method utilises the sequential dependencies between musical parts. This paper proposes a simple acoustic model for the labelling and combines it with the sequential modelling method. The proposed method is evaluated on three data sets of real popular music. The obtained results support the original assumption that musical parts differ in their loudness, and the acoustic information alone can be used to some extent to label the parts. The acoustic information alone has the labelling performance in par with using only part occurrence priors. Combining the acoustic model with the baseline sequential model provides in most cases a improvement in the accuracy. However, the improvement cannot be obtained with all data, because typical loudness relations between different parts seem to depend on the musical genre. Finally, the same search algorithm as with the baseline method can be used for the combined model with very small modifications.

[3] A full list of pieces is available at `http://www.cs.tut.fi/sgn/arg/paulus/TUTstructure07_files.html`.

[4] The annotations are available at `http://www.iua.upf.edu/%7Eperfe/annotations/sections/license.html`, and including some corrections at `http://www.cs.tut.fi/sgn/arg/paulus/structure.html#beatles_data`.

[5] As the entire data set forms one instance in the evaluation measure calculation, no statistical measure could be calculated for proper comparison.

	N=0	N=1	N=2	N=3	N=4	N=5	VMM
a	0.0	0.0 (0.0)	0.8 (0.0)	22.2 (**34.9**)	23.8 (31.7)	27.8 (27.0)	24.6 (29.4)
bridge	18.6	25.8 (17.9)	47.4 (38.6)	51.1 (45.4)	**50.2** (47.4)	49.1 (43.9)	47.7 (41.4)
c	3.3	13.5 (3.6)	41.6 (38.3)	44.6 (42.1)	47.4 (47.7)	**56.2** (54.8)	49.6 (48.5)
chorus	29.5	75.5 (67.9)	83.4 (76.3)	**85.0** (80.6)	82.7 (76.6)	79.8 (75.3)	82.4 (77.9)
chorus_a	11.9	0.0 (0.0)	0.0 (0.0)	8.2 (7.5)	**15.7** (**15.7**)	**15.7** (11.2)	0.0 (3.0)
chorus_b	4.4	0.0 (0.0)	0.9 (0.9)	8.8 (5.3)	**12.4** (**12.4**)	**12.4** (7.1)	0.9 (2.7)
intro	32.7	43.4 (22.7)	97.2 (97.6)	97.6 (98.2)	97.0 (97.8)	**98.2** (97.8)	97.0 (96.8)
outro	52.4	47.4 (9.9)	98.3 (98.3)	**98.6** (**98.6**)	98.3 (97.6)	95.9 (90.9)	98.1 (98.3)
pre-verse	30.1	10.0 (3.7)	51.4 (40.5)	**55.6** (45.6)	50.7 (43.3)	46.8 (41.7)	52.5 (42.6)
solo	23.8	2.2 (0.0)	6.6 (4.4)	6.6 (7.2)	13.8 (15.5)	**23.2** (18.8)	21.0 (16.0)
theme	5.5	0.0 (0.0)	2.7 (0.0)	2.7 (2.7)	2.7 (4.4)	**6.6** (3.3)	3.3 (0.5)
verse	46.5	59.0 (38.4)	72.5 (62.6)	71.0 (64.6)	70.6 (64.5)	72.1 (64.7)	**74.5** (65.4)
MISC	7.8	21.4 (11.7)	35.5 (29.2)	**44.4** (38.6)	42.8 (37.9)	41.7 (40.6)	40.3 (37.3)
total	27.7	42.1 (29.6)	61.7 (55.6)	**64.3** (60.2)	63.6 (59.9)	63.6 (59.3)	63.7 (59.2)

Table 1. Per-label accuracy (%) on *TUTstructure07* obtained using only acoustic modelling (N=0 column), only sequence modelling (values in parentheses), and combining sequence and acoustic modelling (other values).

	N=0	N=1	N=2	N=3	N=4	N=5	VMM
bridge	6.2	22.0 (24.3)	48.0 (45.8)	**77.4** (76.8)	75.1 (75.1)	70.1 (74.0)	69.5 (69.5)
intro	43.8	50.0 (41.4)	92.6 (92.0)	93.2 (92.6)	**93.8** (**93.8**)	**93.8** (**93.8**)	93.2 (93.2)
outro	73.2	60.6 (0.0)	**99.3** (**99.3**)	**99.3** (**99.3**)	98.6 (97.9)	97.9 (93.7)	**99.3** (**99.3**)
refrain	20.1	30.1 (28.1)	43.8 (45.4)	61.8 (62.2)	69.1 (69.9)	65.1 (67.5)	69.1 (**70.3**)
verse	37.2	73.4 (70.6)	80.9 (81.5)	**88.5** (87.9)	86.5 (85.3)	83.7 (84.5)	87.1 (87.5)
verses	23.2	0.0 (0.0)	8.9 (8.9)	51.8 (**53.6**)	37.5 (37.5)	44.6 (44.6)	42.9 (42.9)
versea	**39.2**	0.0 (0.0)	2.0 (2.0)	7.8 (7.8)	23.5 (19.6)	25.5 (19.6)	11.8 (11.8)
MISC	5.7	3.8 (4.5)	17.8 (17.2)	28.7 (29.3)	**43.9** (37.6)	26.1 (25.5)	30.6 (29.9)
total	31.1	43.7 (36.1)	61.8 (61.8)	73.8 (73.7)	**75.7** (74.6)	71.9 (72.4)	73.6 (73.9)

Table 2. Per-label accuracy (%) on *UPF Beatles* obtained using only acoustic modelling (N=0 column), only sequence modelling (values in parentheses), and combining sequence and acoustic modelling (other values).

	N=0	N=1	N=2	N=3	N=4	N=5	VMM
bridge a	20.1	20.1 (8.2)	72.3 (62.9)	**73.6** (66.7)	64.8 (66.0)	59.7 (49.7)	71.7 (62.9)
chorus a	51.2	70.9 (45.6)	85.3 (76.2)	85.6 (77.6)	80.6 (73.5)	73.5 (71.5)	**86.2** (79.7)
chorus b	28.0	37.5 (6.5)	79.2 (73.2)	**79.8** (71.4)	72.6 (65.5)	72.0 (71.4	76.2 (72.0)
ending	80.6	84.7 (32.7)	**100** (**100**)	99.0 (**100**)	98.0 (94.9)	99.0 (88.8)	**100** (99.0)
intro	50.0	45.1 (10.8)	**100** (**100**)	**100** (**100**)	**100** (**100**)	**100** (**100**)	**100** (**100**)
pre-chorus	7.6	7.6 (3.3)	**64.1** (51.1)	60.9 (52.2)	63.0 (39.1)	48.9 (42.4)	63.0 (45.7)
verse a	35.0	48.1 (20.3)	**85.7** (76.8)	84.4 (78.9)	81.4 (78.1)	77.6 (73.4)	81.0 (76.4)
verse b	19.4	30.3 (17.4)	**85.6** (76.6)	84.1 (80.1)	79.6 (74.6)	73.6 (69.2)	82.1 (76.6)
verse c	41.9	14.0 (0.0)	**60.5** (30.2)	55.8 (39.5)	47.7 (30.2)	32.6 (30.2)	47.7 (33.7)
MISC	29.8	52.4 (8.4)	**84.9** (67.6)	80.4 (73.8)	77.8 (68.4)	69.8 (67.6)	83.1 (74.7)
total	36.0	45.5 (19.1)	**82.8** (72.8)	81.7 (75.3)	77.5 (70.9)	71.8 (68.0)	80.7 (74.1)

Table 3. Per-label accuracy (%) on *RWC pop* obtained using only acoustic modelling (N=0 column), only sequence modelling (values in parentheses), and combining sequence and acoustic modelling (other values).

5. REFERENCES

[1] Guillaume Boutard, Samuel Goldszmidt, and Geoffroy Peeters. Browsing inside a music track, the experimentation case study. In *Proc. of 1st Workshop on Learning the Semantics of Audio Signals*, pages 87–94, Athens, Greece, December 2006.

[2] Roger B. Dannenberg and Masataka Goto. Music structure analysis from acoustic signals. In David Havelock, Sonoko Kuwano, and Michael Vorländer, editors, *Handbook of Signal Processing in Acoustics*, volume 1, pages 305–331. Springer, New York, N.Y., USA, 2008.

[3] Masataka Goto. AIST annotation for the RWC music database. In *Proc. of 7th International Conference on Music Information Retrieval*, pages 359–360, Victoria, B.C., Canada, October 2006.

[4] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Popular, classical, and jazz music databases. In *Proc. of 3rd International Conference on Music Information Retrieval*, pages 287–288, Paris, France, October 2002.

[5] Daniel Jurafsky and James H. Martin. *Speech and language processing*. Prentice-Hall, Upper Saddle River, N.J., USA, 2000.

[6] Namunu C. Maddage. Automatic structure detection for popular music. *IEEE Multimedia*, 13(1):65–77, January 2006.

[7] Elias Pampalk. A Matlab toolbox to compute music similarity from audio. In *Proc. of 5th International Conference on Music Information Retrieval*, Barcelona, Spain, October 2004.

[8] Jouni Paulus. *Signal Processing Methods for Drum Transcription and Music Structure Analysis*. PhD thesis, Tampere University of Technology, Tampere, Finland, December 2009.

[9] Jouni Paulus and Anssi Klapuri. Labelling the structural parts of a music piece with Markov models. In Sølvi Ystad, Richard Kronland-Martinet, and Kristoffer Jensen, editors, *Computer Music Modeling and Retrieval: Genesis of Meaning in Sound and Music - 5th International Symposium, CMMR 2008 Copenhagen, Denmark, May 19-23, 2008, Revised Papers*, volume 5493 of *Lecture Notes in Computer Science*, pages 166–176. Springer Berlin / Heidelberg, 2009.

[10] Jouni Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1159–1170, August 2009.

[11] Geoffroy Peeters and Emmanuel Deruty. Is music structure annotation multi-dimensional? A proposal for robust local music annotation. In *Proc. of 3rd Workshop on Learning the Semantics of Audio Signals*, pages 75–90, Graz, Austria, December 2009.

[12] Alan W. Pollack. 'Notes on...' series. The Official rec.music.beatles Home Page (http://www.recmusicbeatles.com), 1989–2001.

[13] Dana Ron, Yoram Singer, and Naftali Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, 25(2–3):117–149, 1996.

[14] Yu Shiu, Hong Jeong, and C.-C. Jay Kuo. Musical structure analysis using similarity matrix and dynamic programming. In *Proc. of SPIE Vol. 6015 - Multimedia Systems and Applications VIII*, pages 398–409, 2005.