Audio Engineering Society

Convention Paper

Presented at the 129th Convention
2010 November 4–7 San Francisco, CA, USA

Perceptually motivated scoring of musical meter classification algorithms

Matthias Varewyck[1], and Jean-Pierre Martens[1]

[1] *Department of Electronics and Information Systems, Ghent University*

Correspondence should be addressed to Matthias Varewyck (matthias.varewyck@ugent.be)

ABSTRACT

In this paper, perceived confusions between the four most popular meters 2/4, 3/4, 4/4 and 6/8 in Western music are examined. A theoretical framework for modeling these confusions is proposed and translated into a perceptually motivated objective score that can be used for the evaluation of meter classification algorithms with respect to meter labels that were elicited from a single annotator. Experiments with three artificial and two real algorithms showed that the new score is preferable over the traditional accuracy since the score rewards algorithms that make reasonable errors and seems to be more robust against different annotators.

1. INTRODUCTION

In most Western songs repetition appears at different time scales or *levels* that are strictly aligned to each other. Together these levels constitute a *metrical structure* that is very characteristic for the song. Because some repetition may temporarily (dis)appear, the metrical structure can locally change in the course of a song. The most continuous or regular level is defined as the *beat* level. The level that typically corresponds to the rhythmic patterns and harmonic changes in the song is defined as the *bar* level. All levels from the beat to the bar constitute the *meter* of the song.

Algorithms that classify songs according to the meter are called *meter classification algorithms*. The earliest algorithms [3, 1] distinguish between *duple* and *triple* meters, referring to the existence of a salient period of two or three beats. More recent algorithms [6, 4, 2, 5] consider up to ten distinct meters. Until now, the quality of a meter classification algorithm was quantified by the percentage correctly classified songs, defined as the *accuracy*. However, such a measure ignores the perceptual relations that exist between some of these meters. The aim of this paper is to investigate these relations, and to propose an objective score that is more representative

of the perceptual score a human listener would assign to the output of an algorithm.

We experimentally and theoretically verify the relations between the regular meters 2/4, 4/4, 3/4 and 6/8 which are the dominant meters in popular Western music. Therefore, the meters 2/4 and 3/4 are represented by two-level structures composed of a beat and a bar level, with the bar level having a period of 2 and 3 beat periods respectively. The meters 4/4 and 6/8 are represented as three-level structures which are obtained by the presence of an extra level with a period of 4 and 6 beat periods respectively. The meters 2/4 and 4/4 can be labeled as duple meters and the meters 3/4 and 6/8 as triple meters.

In what follows each of the four meter *classes* will be represented by a single *label* representing the numerator of the full meter description.

In the remainder of this paper we describe a listening experiment (section 2), we introduce a novel meter perception model whose parameters are optimized to maximally fit the results of the listening experiment (section 3), and we finally apply the model to a perceptually based score for evaluating meter classification algorithms (section 4). The paper ends with a conclusion and a glimpse on future work.

2. LISTENING EXPERIMENT

To experimentally find relations between the studied meters, a small listening experiment was set up.

Ten persons with a musical background were asked to listen to hundred 30-seconds long music excerpts forming a balanced set with equal frequencies for all four meters. Furthermore, the persons were asked to identify the meter of each excerpt as one of the four meters. The label of each person was considered as a vote for a particular meter and the meter with the maximum number of votes was selected as the target label.

By comparing the annotated labels of the individual persons to the target labels one can construct a *confusion matrix* for each person. The elements in this confusion matrix represent the fraction of songs with a certain annotated and a certain target label. In all confusion matrices reported hereafter, the rows and columns represent the annotated and the target labels $2, 4, 3$ and 6 (in that order) respectively.

By accumulating columns of the individual matrices, one can derive the a priori probabilities of the target classes. The obtained probabilities are given by

$$P(t) = \begin{bmatrix} 0.28 & 0.38 & 0.21 & 0.13 \end{bmatrix} \qquad (1)$$

By averaging the matrices of the individual listeners and by normalizing the columns, we obtained the following experimental confusion matrix

$$P(a \mid t) = \begin{array}{c} \begin{matrix} 2 & 4 & 3 & 6 \end{matrix} \\ \begin{bmatrix} 0.72 & 0.13 & 0.02 & 0.05 \\ 0.25 & 0.80 & 0.00 & 0.15 \\ 0.00 & 0.01 & 0.83 & 0.12 \\ 0.02 & 0.07 & 0.15 & 0.69 \end{bmatrix} \end{array} \begin{matrix} t/a \\ 2 \\ 4 \\ 3 \\ 6 \end{matrix}$$

The gray rectangles indicate the errors within each group (duple-triple). Each element in this matrix can be interpreted as a probability $P(a \mid t)$ of an annotated class (a) given the target class (t). Let us briefly discuss the two obtained results now.

Although the song excerpts were selected to constitute a balanced set the prior probabilities of the target labels emerging from our listening experiment are unequal. This is due to the fact that we made the selection on the basis of the annotations of a single annotator, and that these annotations are not entirely reliable. Although the set is apparently not as balanced as we anticipated, it still comprises all classes with a sufficiently high frequency.

The obtained confusion matrix reveals that the mean error rate for the classification into four classes amounts to 22.9%, whereas it drops to 6.4% if only duple-triple confusions are being considered as errors. Most of the confusions clearly occur within the same group. Furthermore, the meters of the duple group seem to be more confusable than the meters of the triple group. Finally, cross-group confusions seem to be dominated by confusions between 4 and 6. These phenomena are also reflected in the confusion matrices reported in papers containing evaluations of meter classification algorithms.

3. METER PERCEPTION MODEL

In this section we conceive a simple model with only two tunable parameters that will show to explain the observed confusion probabilities rather well for reasonable values of the associated parameters. We depart from the hypothesis that most of the observed

errors originate from two causes that are considered to be more or less independent from each other. Based on this hypothesis, we construct a probabilistic model that estimates the probabilities $P(a \mid t)$ of an annotated label (a) given a target label (t).

We investigated several options for the two causes and we found that the best results were obtained by considering that the annotator can confuse the target beat level with an adjacent level in the metrical structure (cause 1), and that the annotator can wrongly annotate a meter with three levels when the target meter has two levels and vice versa (cause 2).

We first introduce $d_b = -1$, 0 or $+1$ as the difference between the annotated and the target beat level, and $d_L = 0$ or 1 as the absolute difference between the annotated and the target number of levels in the meter. In order to compute the desired conditional probabilities, we then propose the following two-step procedure:

1. Determine the probabilities $P(d_b, d_L \mid t)$ of the six combinations (d_b, d_L) for each target class t, and record the annotation(s) that would emerge for each combination. Combinations requiring extra levels can have two possible annotations as will be explained later in this section.

2. Derive from this information the envisaged conditional probabilities as

$$P(a \mid t) = \sum_{d_b} \sum_{d_L} P(d_b, d_L \mid t)\, P(a \mid d_b, d_L)$$

   If a certain (d_b, d_L) allows for two annotations, for simplicity, both of them are presumed to be equally likely and thus receive a probability $P(a \mid d_b, d_L) = 0.5$.

Once more the probabilities $P(d_b, d_L \mid t)$ introduced in step one are also achieved by means of a two-step procedure:

1. First consider the two error causes as independent and assign probabilities to the combinations (d_b, d_L) in terms of the prior probabilities P_b and P_L of making a beat-level or a number-of-level error respectively. These probabilities form the two parameters the model depends on.

d_L	d_b	$P(d_b, d_L \mid t)$	a
0	0	$(1-P_b)\,(1-P_L)$	2
0	-1	$P_b/2\,(1-P_L)$	2,3
0	1	$P_b/2\,(1-P_L)$	2,3
1	0	$(1-P_b)\,P_L$	4
1	-1	$P_b/2\,P_L$	4,6
1	1	0	-

Table 1: Illustration of the proposed procedure to compute the possible annotations and their unnormalized probabilities $P(d_b, d_L \mid t)$ for $t = 2$.

2. Replace the probabilities corresponding to annotations that are outside the class set or too different from the target meter (annotation has two missing levels or two extra levels) by zero and normalize the non-zero probabilities in order to compensate for this.

Note that some combinations (d_b, d_L) require an extra metrical level that lies outside the target structure. This extra level cannot be derived uniquely from the target class. If adding a level whose period is two times larger/smaller and adding one whose period is three times larger/smaller both yield an annotation inside the class set, these two annotations get a $P(a \mid d_b, d_L) = 0.5$. By doing so, we avoid the need for a third parameter while still taking all possible extensions of the target structure into account. Table 1 illustrates the whole procedure (except for the normalization) for the case of target label 2.

To identify the parameter values that best explain the experimental results, we performed a grid search in the (P_b, P_L) space and we selected the grid point yielding the lowest mean squared difference between the experimental confusion matrix and the modeled probabilities matrix. This results in the optimal values $P_b = 0.14$ and $P_L = 0.18$. The fact that $P_L > P_b$ agrees with the expectations of the participants of the listening experiment. Using these optimal values we obtain the following conditional probabilities matrix.

$$P(a \mid t) = \begin{bmatrix} 0.77 & 0.17 & 0.06 & 0.01 \\ 0.16 & 0.80 & 0.00 & 0.06 \\ 0.06 & 0.00 & 0.78 & 0.17 \\ 0.01 & 0.03 & 0.16 & 0.76 \end{bmatrix} \qquad (2)$$

Apparently, the proposed model largely predicts the right relative magnitudes: the matrix is dominated

by the diagonal elements, followed by within-group confusions and finally by between-group confusions. However the model does not predict the more likely confusions in the duple group compared to the triple group and the prominent confusions between 6 and 4 for between-group confusions.

The latter can be due to the fact that participants were restricted to the class set and thus some may have heard a 12/8 meter but have annotated it as 4/4 while others have annotated it as 6/8. In the proposed model we did not consider meters with more than three levels neither did we try to interpret these meters as meters from the class set. At the other side, given the small size of our experiment, we argue that our data do not provide an incontrovertible proof of the existence of the prominence.

4. PERCEPTUALLY MOTIVATED SCORE

We will now apply the obtained conditional probabilities matrix to conceive a perceptually motivated quality score for the assessment of meter classification algorithms. The outputs of the algorithms are denoted by c. The presumed assessment of such an algorithm is based on an analysis of the discrepancies between the outputs c and the single-annotator annotations a that are available for the dataset the algorithm has processed.

In a first step, we define the *subjective accuracy* (SA) of an algorithm as the probability that its output c is correct, given the available annotations. It is computed as

$$\mathrm{SA} = \sum_c \sum_a P(c,a)\, P(t = c \mid a) \tag{3}$$

with the sums taken over the eligible classes. The first factor describes the observed discrepancies between the outputs of the classifier and the available annotations. The second factor models the possible errors in the annotation and follows from the conditional probabilities $P(a \mid t)$ delivered by our proposed model. It follows from Bayes law that

$$P(t \mid a) = \frac{P(a \mid t)\, P(t)}{\sum_j P(a \mid j)\, P(j)} \tag{4}$$

Substituting eq. (1) and eq. (2) in eq. (4) then gives the following matrix with conditional probabilities that should be used in the subjective accuracy.

$$P(t \mid a) = \begin{bmatrix} 0.73 & 0.22 & 0.04 & 0.01 \\ 0.13 & 0.85 & 0.00 & 0.02 \\ 0.08 & 0.00 & 0.81 & 0.11 \\ 0.01 & 0.08 & 0.23 & 0.68 \end{bmatrix} \tag{5}$$

For a simple classification between duple and triple meters, the four-by-four matrix can be converted to a two-by-two matrix by accumulating the figures in the indicated two-by-two sub-matrices.

Note that if the annotated labels were equal to the target labels, the model would boil down to a diagonal matrix and eq. (3) would provide the standard accuracy that is used in most studies. This accuracy would be one for the algorithm that succeeds in predicting the available annotations ($c = a$). Now, in order to obtain a *subjective score* (SS) that is easier to compare to the standard accuracy, we propose to divide the SA by the SA of this faultless algorithm. The latter value is presumed to represent the maximum attainable value of SA. The described normalization then leads to the following subjective score

$$\mathrm{SS} = \frac{\mathrm{SA}}{\sum_a P(a)\, P(t = a \mid a)} \tag{6}$$

with $P(t = a \mid a)$ being the elements on the principal diagonal of the matrix in eq (5). Note that the prior probabilities $P(a)$ must be derived from the dataset on which the algorithm was evaluated.

5. EXAMPLES

To illustrate the advantages of the new score over the traditional percentage correctly classified songs, different evaluation measures were computed for three artificial and two real algorithms using the dataset described before. In order to make a fair assessment, the output c of an algorithm was compared to the annotated label a provided by one of the ten annotators that participated in our listening experiment and the target label t as well as the applied model were recomputed for the annotations of the remaining nine annotators. The experiment was repeated for each annotator and the outcome was averaged across the annotators. For new datasets, the reader can obviously use the matrix defined in eq. (5).

The artificial algorithms in this study are: (1) *Correct*: an algorithm that produces the target labels

derived from all annotations emerging from our listening experiment, (2) *Random4*: an algorithm that produces a random label of the class set, and (3) *Random2*: an algorithm that produces the correct group according to the target labels derived from all annotations but a random label within that group. Note that it would have made no sense to consider an algorithm whose outputs are equal to the annotations against which to compare because this algorithm would always get a score of 100%. Furthermore, it is more realistic to use the same classifier output for each experiment.

The real algorithms in this study are those of *Klapuri* et al. [4] and *Pikrakis* et al. [5]. Since the algorithm of Klapuri computes one meter label per single bar interval and since this label could be outside the class set, we selected the most frequently observed label that belongs to the set as the final output. The original algorithm of Pikrakis could generate labels that are outside the class set as well, but the algorithm was modified by its author to restrict the output to the four labels investigated here.

The results of our study are displayed in Table 2. First of all the data show that the newly proposed score for the correct algorithm is higher than the traditional 4-class accuracy and therefore more in agreement with the expectation that this algorithm should indeed receive a high score. Furthermore, the spread on the proposed scores is smaller than that on the traditional scores for all except one algorithm. Finally, the new score seems to take the errors into account in a more reasonable way: algorithms that produce less within-group confusions and thus have a higher 2-class accuracy (duple-triple) are rewarded. This explains why the superiority of Klapuri over Pikrakis is more apparent with respect to the new score than to the 4-class accuracy.

Algorithm	4-class acc.	2-class acc.	New score
Correct	77.1 ± 13.1	93.6 ± 2.2	82.4 ± 10.5
Random4	25.0 ± 2.8	50.9 ± 5.1	31.4 ± 3.3
Random2	46.9 ± 3.5	93.6 ± 2.2	58.1 ± 2.9
Klapuri	44.8 ± 8.2	81.4 ± 2.4	54.1 ± 7.2
Pikrakis	41.4 ± 2.6	75.0 ± 1.9	49.9 ± 2.3

Table 2: Mean and deviation of the traditional 4-class accuracy and the newly proposed score obtained for five meter classifiers (see text) distinguishing between 4 meter classes. The 2-class accuracies are added to show the balance between the within- and between-group accuracies of these classifiers.

6. CONCLUSIONS AND FUTURE WORK

We have proposed a parametric model that is based on realistic assumptions and ditto parameters and that can fairly well explain the outcomes of a meter classification experiment involving four meter classes (2/4, 3/4, 4/4 and 6/8). We have then adopted this model to present a new score to evaluate automatic meter classification algorithms. Furthermore we compared the new score obtained with this model for three artificial and two real algorithms to the traditional accuracy score. From this comparison we conclude that the new score provides more intuitive results in the sense that a higher score is assigned to an algorithm making reasonable errors, and that the score depends less on the annotator that provided the reference annotations. We therefore plan to use the new score as the evaluation metric for the meter classification algorithm we are developing.

7. ACKNOWLEDGMENTS

This paper describes work that was conducted under the Ghent University Grant GOA-1250604 (SEMA). The authors would like to thank A. Klapuri and A. Pikrakis for providing their source code and the ten persons for participating in the listening experiment.

8. REFERENCES

[1] S. Dixon, E. Pampalk, and G. Widmer. "Classification of dance music by periodicity patterns," In Proc. of 4th int. soc. of music inf. retrieval conf., 159–165, Baltimore, MD, USA, 2003.

[2] M. Gainza. "Automatic musical meter detection," In Proc. of IEEE int. conf. on acoustics, speech and signal proc., 329–332, Washington, DC, USA, 2009.

[3] F. Gouyon and P. Herrera. "Determination of the meter of musical audio signals: seeking recurrences in beat segment descriptors," In Proc. of 114th audio eng. soc. conv., Amsterdam, the Netherlands, 2003.

[4] A. Klapuri, A. Eronen, and J. Astola. "Analysis of the meter of acoustic musical signals," IEEE trans. on speech and audio proc., no.14(1):342–355, 2006.

[5] A. Pikrakis, A. Iasonas, and S. Theodoridis. "Music meter and tempo tracking from raw polyphonic audio," In Proc. of 5th int. soc. of music inf. retrieval conf., 192–197, Barcelona, Spain, 2004.

[6] C. Uhle, J. Rohden, M. Cremer, and J. Herre. "Low complexity musical meter estimation from polyphonic music," In Proc. of 25th audio eng. soc. conf., London, UK, 2004.