WEARABLE ASSISTANCE FOR THE BALLROOM-DANCE HOBBYIST – HOLISTIC RHYTHM ANALYSIS AND DANCE-STYLE CLASSIFICATION

Florian Eyben, Björn Schuller, Stephan Reiter, and Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München, Germany
[eyb, sch, res, ri]@mmk.ei.tum.de

ABSTRACT

Automated retrieval of high level information from ballroom dance music is challenging, but has many practical applications. These include, for example, a fully automatic ballroom dance D.J., robots capable of performing ballroom dances, or wearable dance-assistance, as considered herein. It is necessary, for such a system, to retrieve information about the song's quarter note tempo, meter and beat positions. Further, the system must be able to discriminate between the nine Standard and Latin ballroom dances. In this paper we present a model that combines all these requirements in one holistic approach. The polyphonic input is processed by a simplified psychoacoustic model. Tatum, tempo and meter features are extracted using resonant filters. The filter output is used for beat tracking. The extracted features are used for a ballroom dance- style classification by Support-Vector-Machines. To show the high effectiveness regarding dance-style recognition and beat tracking, test-runs are carried out on a database containing 1.8k titles.

1. INTRODUCTION

Especially for the ballroom dance beginner or hobbyist a wearable ballroom dance assistance would be beneficial. Such a device should be able to recognize the ballroom dance-style, quarter note tempo and meter. Further, beat positions and the beginning of a meter must be identified by the assistant, in order to give the dancer hints about the current dance-style, the counting, i.e. when to start, or which dance steps the dancer has to perform on a certain beat. These hints can be delivered, for example, via an ear-plug from an individual wearable computing device or as a public announcement through the ballroom speaker system.

Tempo and meter detection is the basis of such a ballroom dance assistant. Therefore, a few existing tempo detection and meter recognition approaches [2-5] are briefly summarized in section 2, before we present our model that combines both capabilities in one model. As described in section 3 in more detail, the model uses comb filters as resonators [4] to detect tempi on several metrical levels, extract features for ballroom dance-style classification and perform beat tracking using the comb filter output. Test-runs are carried out on a 1.8k public database [1], introduced in section 4. Ballroom dance-style classification is discussed in section 5. Finally, results and conclusion are found in section 6 and 7, respectively.

2. GENERAL APPROACH

Information about musical meter is essential for ballroom dance-style recognition. By using only the meter, it can be distinguished between Waltzes (triple meter) and the other seven dance-styles (duple meter or common time). Meter detection requires tempo independent information about the rhythmic structure of the audio track [3]. Thus, it is necessary to reliably detect the tempo of the song in beats-per-minute. There are basically three different methods that are commonly used for tempo detection: correlation methods, note onset detection followed by computation of the most common inter-onset-interval (IOI) via a histogram method [5] and a multiple resonator based approach using IIR comb filters [4]. Our tempo detection method is based on [4], except for a few improvements and performance enhancements. As it is very difficult, even for a human listener, to determine if the quarter notes of a song are better grouped by 2, 4 or 8, we restrict the meter recognition problem to a simple duple or triple decision, as done in [2]. This is sufficient for ballroom dance music. The algorithms mentioned in [2] and [5] try to determine the metrical grouping by explicitly identifying downbeats. A downbeat thereby is a stronger accented beat, which might indicate the first beat in a meter. However, the task of reliably finding a downbeat is challenging, even for the non-experienced human listener. It must also be considered that downbeats need not always occur at the beginning of a meter. We therefore focus on a reliable two-class decision between duple and triple meter without any knowledge about note onset or downbeat positions. Moreover, other approaches to meter detection (such as [2]) require manual adjustments of the automatically extracted quarter note tempo. Our approach is able to discriminate between duple and triple meter without knowledge of the quarter note tempo. Instead, the fastest tempo present in the song, called the tatum tempo, is detected. Then, the distribution of resonances across 19 fractions of this tatum tempo is analyzed. This distribution combined with information from the tatum detection stage reveals highly relevant information about the metrical structure, that we use for quarter note tempo detection and ballroom dance-style classification.

Once the quarter note tempo, the meter and the ballroom dance-style are known, finding beat positions and downbeat locations is made much easier. With the assumption that a track of typical ballroom dance music has a roughly constant tempo throughout, the most likely phase for a grid of nearly equidistant beats is determined for small excerpts of the song by analyzing the output of a comb filter tuned to the detected quarter note tempo. For deciding which of the labeled beats is a downbeat, loudness information from the psychoacoustic model is used.

3. FEATURE EXTRACTION

In the process of extracting the features tempo and meter, several other features are extracted, that will be used in the later classification step. Before any feature extraction can be performed a perceptive preprocessing using a simplified psychoacoustic model is applied to the polyphonic PCM input data.

3.1. Perceptive preprocessing

Input audio data is reduced to a sample rate of 11.025 kHz and converted into monophonic by stereo-channel addition in order to reduce computation time. Audio is then split into frames of 256 samples with an overlap of 0.57 so the resulting envelope frame-rate is 100 frames-per-second (fps). A half-wave sine window is used as windowing function when computing the FFT on each frame. By using N_{mel} overlapping triangular filters, equidistant on the Mel-Frequency scale, the 128 bands are reduced to N_{mel} nonlinear bands. Tests were carried out for different numbers for N_{mel} and the best overall result was achieved with N_{mel}=9. The envelope of each of the N_{mel} bands is then low pass filtered by convolving with a half-wave raised cosine filter having a length of 15 envelope samples [4]. This preserves fast attacks, but filters noise and rapid modulation, most as in the human ear. From the filtered band envelopes a weighted differential d_w is computed in the following way for a sample o_i at position i: A moving average is calculated over one window of 10 samples to the left of sample o_i (left mean $\bar{o}_{i,l}$) and a second window of 20 samples to the right of sample o_i (right mean $\bar{o}_{i,r}$). The differential then is given by (1):

$$d_w(i) = \left(o_i - \bar{o}_{i,l}\right) \cdot \bar{o}_{i,r} \tag{1}$$

Taking into account data before and after the current position was inspired by [5]. The sound level before an onset event and the duration of the note belonging to the onset both contribute to the perceived emphasis of the specific onset event, which corresponds to the perceived accentuation of the note.

3.2. Tatum features

The tatum grid, according to [5], is the lowest metrical level of a song. It corresponds to the fastest tempo in the song, or likewise the smallest inter-onset-interval (abbr. IOI in the ongoing). For detecting the tatum tempo we use a comb filter bank consisting of 57 filters. We assign a fixed gain α to each filter in the bank instead of deriving a gain for each filter basing on a constant half energy time as is done in [4]. Overall best results were achieved with a gain of 0.7. The delays of the filters range from τ=18 to τ=74 envelope samples. The differential d_w of each Mel-frequency-band envelope is processed by the comb filter bank and the total energy over all bands of the output of each filter is computed. This value for each filter is stored in the tatum tempo vector $\underline{T}$, as plotted in Fig. 1. From this vector three more features are extracted considering the quality of the peaks in the vector: *T-ratio* is computed by dividing the highest value of $\underline{T}$ by the lowest. *T-slope* is computed by dividing the first value of $\underline{T}$ by the last. *T-peakdist* is computed as mean of the maximum and minimum value of $\underline{T}$ normalized to the global mean of $\underline{T}$.

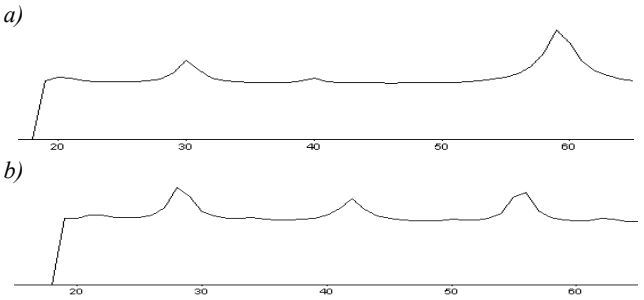

Fig. 1. – Plots of 57 dimensional tatum tempo vectors for *a)* Robbie Williams – Rock DJ, *b)* OMD – Maid of Orleans. Axes are labeled with the delay τ of the comb filter (in envelope samples) corresponding to each tatum vector element.

Since our comb filters tend to stronger resonances at higher tempi, the vector is flattened by considering the difference between the average of the first 6 values and the average of the last 6 values. From the resulting vector the two most dominant peaks are picked as follows: Firstly, all local minima and maxima are detected, then for each maximum n its apparent height D_n is computed by taking the average of the maximum minus its left and right minimum. The indices of the two maxima with the greatest apparent height are considered possible tatum IOI (abbreviated T in the ongoing) candidates (T_1 and T_2). For each of these two candidates a confidence C_n is computed with (2) and the candidate with the higher confidence is chosen as the final tatum IOI T_f:

$$C_n = D_n + \underline{T}(T_x) \tag{2}$$

3.3. Meter and Tempo Features

A meter vector $\underline{m}$ is computed by setting up small comb filter banks with filter delays τ ranging from $T_f \cdot i - i$ to $T_f \cdot i + i$. For each filter bank the filter with the highest energy output is selected as the tempo θ_i and the total energy of this filter summed over all bands is a score value m_i which is stored in the meter vector at position i. This score value m_i indicates how well the tempo $T_f \cdot i$ resonates with the song, i.e. how strong is is present compared to the other tempi. It is sufficient to use $i \in [1;19]$, since all important multiples for meter classification are included in this range. At this point we present a simple rule based method for discriminating between duple and triple meter. Later on we will use Support-Vector-Machines to improve meter recognition accuracies. Two sums s_2 and s_3 are computed ($\underline{m}$ refers to the meter vector):

$$s_2 = \frac{1}{3} \cdot \left[\underline{m}(4) + \underline{m}(8) + \underline{m}(16)\right] \tag{3}$$

$$s_3 = \frac{1}{3} \cdot \left[\underline{m}(6) + \underline{m}(9) + \underline{m}(18)\right] \tag{4}$$

The greater of the two sums determines what we call base meter M_b. If s_3 is greater, then M_b is triple (e.g. 3/4 or 6/8 time signature), otherwise it is duple (e.g. 4/4 or 2/4 time signature).

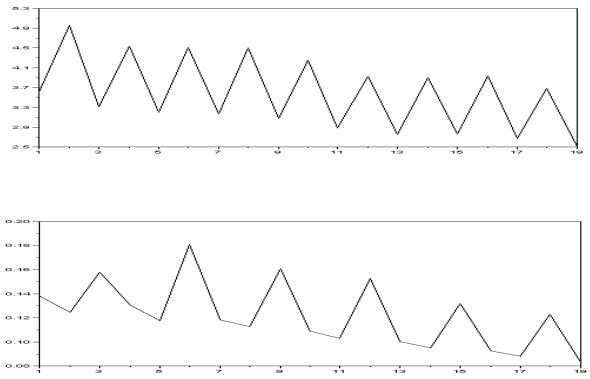

Fig. 2. - Meter vectors for the songs referenced in Fig. 1. Clearly visible duple meter in top plot and triple meter in bottom plot.

We also use the meter vector to identify the quarter note (main) tempo. Bounds within which the tempo is expected are defined. For duple M_b the range is from 60 BPM to 222 BPM and from 75 BPM to 240 BPM for triple M_b. For each θ_i within bounds that satisfies the condition $m_{2i} > m_{3i}$ for duple M_b or $m_{3i} > m_{2i}$ for triple M_b a score $\sigma_i = m_i + m_{2i}$ or $\sigma_i = m_i + m_{3i}$, respectively, is computed and the θ_i for which σ_i is maximal is chosen as main tempo θ_B. If no θ_i within bounds is found that satisfies these conditions, the θ_i within the defined tempo bounds having the highest score m_i is chosen as main tempo θ_B. Should it happen, that no θ_i falls within bounds, the tatum tempo T_f is chosen for the main tempo θ_B.

3.4. Beat tracking

A tracking envelope E_t is computed by filtering d_w of each Mel-Frequency band (see section 3.1.) by a comb filter with α=0.4 and delay $\tau=\theta_B$ and adding the filter outputs over all N_{mel} bands. Multiple pass beat tracking is performed using this tracking envelope. Beats may differ from their expected position by a tolerance t. Starting at frame 0 as the first beat candidate an iterative tracking with a large tolerance window is performed to find possible beat candidates. If p_n is the position of the current beat candidate n, the position of the candidate $n+1$ is determined as the position of the highest peak in a window of E_t between positions $p_n+\theta_B-t$ and $p_n+\theta_B+t$. The same tracking algorithm is now applied for each candidate n as the first beat with a now smaller tolerance t of 8 frames in order to find a reliable first beat and the optimal set of beat positions. This set is chosen by determining for which candidate the sum over the values of E_t at the tracked beat positions is maximal, because the higher the amplitude of a peak in E_t is, the more likely this peak corresponds to a beat. After a reliable first beat has been detected by analyzing at least 15 seconds of recorded audio, the beat tracking algorithm is capable of predicting beats on-line in real-time.

4. BALLROOM DANCE DATABASE

We chose a set of 1,855 pieces of typical Standard and Latin dance music publicly accessible at [1], covering the Standard Dances Waltz, Viennese Waltz, Tango, Quick Step, and Foxtrot, and the Latin Dances Rumba, Cha Cha, Samba and Jive. 30 seconds of each song are available, which we converted to 44.1kHz PCM as required by our preprocessing. The distribution among dance-styles is depicted in tab. 1. The set is abbreviated *BRD* in the ongoing.

Waltz	Viennese Waltz	Tango	Quick Step
293	136	185	242

Foxtrot	Rumba	Samba	Cha Cha	Jive
245	217	188	211	138

Tab. 1: Distribution of dance-styles within *BRD* data-set

The ground truth of tempo, ballroom dance-style and meter for the *BRD* data-set is known from [1]. A part of this set is also used in [7] for ballroom dance-style recognition.

5. CLASSIFICATION

With all 83 features introduced in section 3 we now aim at classifying meter and ballroom dance-style. As classifier we employ Support Vector Machines (SVM) with a polynomial Kernel function basing on our experience in Musical Genre Discrimination [6]. We firstly analyze the dataset by performing a closed-loop Hill-climbing feature selection employing the target classifier's error rate as optimization criterion, namely Sequential Forward Floating Search (SVM-SFFS) [7]. This reveals the following feature-set to yield the best results for ballroom dance-style classification (FS_D in the ongoing): Base meter M_b, *T-ratio*, *T-slope*, *T-peakdist*, meter vector $\underline{m}$ elements 4, 5, 6, 8, 11, 12, 14, 15, 19 and tatum vector $\underline{T}$ elements 1-20, 22-28, 30-57. Features relevant for meter classification (FS_M in the ongoing) are *T-ratio*, meter vector $\underline{m}$ elements 4, 6, 8, 16 and tatum vector $\underline{T}$ elements 3, 4, 7, 13, 16, 19, 20, 22, 23, 25, 26, 28, 29, 33, 36, 37, 39, 48, 50.

6. RESULTS AND DISCUSSION

Test runs have been carried out on the formerly introduced set *BRD*. The accuracy of the features tempo θ_B and base meter M_b compared to the given ground truth of tempo and meter is evaluated. Hereby the tolerance for tempo detection is 2.5 frames absolute IOI deviation. It makes more sense to express the tolerance as allowed deviation of the IOI in frames than as relative BPM deviation, since slight tempo detection inaccuracies are due to roundoff errors because a frame-rate only 100fps is used. However, a quite sufficient accuracy for our purpose is achieved with this frame-rate and computation time is kept at a minimum, considering low CPU resources on a wearable computing device. When the detected tempo is twice or half the ground truth tempo (octave error), it is still treated as correctly detected, since the correct tempo octave can be determined using knowledge about tempo ranges of nine typical ballroom dance-styles.

Additionally, we evaluate the effectiveness for ballroom dance-style and meter classification by Support-Vector-Machines in a 10-fold stratified cross validation (SCV). Thereby evaluation splits are always disjunctive in terms of musical pieces. The data is standardized to have zero mean and unit variance with respect to the training split.

Accuracy [%]	Jive	Samba	Rumba	Cha	Fox	QuickS	Waltz	VWaltz	Tango	MEAN
Tempo θ_B	98.6	94.1	90.3	97.6	94.3	95.0	73.7	75.7	97.3	**90.3**
Base meter M_b	100.0	93.6	96.8	99.5	94.3	95.0	87.4	80.9	94.6	**93.3**
Tempo & base meter	98.6	93.6	88.9	97.6	90.2	89.7	69.6	75.7	93.0	**87.8**
Tracked beats (A)	52.0	76.0	76.0	88.0	84.0	72.0	68.0	84.0	96.0	**77.3**
Tracked beats (B)	28.0	76.0	76.0	20.0	84.0	72.0	68.0	84.0	52.0	**62.2**
Downbeats (rel. %)	14.3	52.6	68.4	40.0	61.9	50.0	88.2	95.2	38.5	**62.9**
BDS Recall	92.8	81.9	78.8	85.8	94.7	92.1	94.9	94.1	91.9	**89.8**
BDS Precision	92.1	91.7	78.4	92.8	96.3	87.8	89.4	96.2	86.7	**89.9**
BDS F_1-Measure	92.4	86.5	78.6	89.2	95.5	89.9	92.1	95.2	89.2	**89.8**

Tab. 2. - Feature extraction accuracy for tempo and meter (Tempo tolerance is IOI±2.5 frames). Beat tracking accuracies (see text for details). Recall, Precision and F_1-Measure of ballroom dance-style classification (BDS) with feature-set FS_M.

Using a P4-mobile with 1.4 Ghz, average computation time for the feature extraction is 1.3s for the 30s excerpts from the *BRD* set, roughly corresponding to a rtf of 0.05. This makes it possible to run the algorithm on a wearable computing device, e.g. a smart-phone, in real-time.

Tab. 2 lists detailed per-dance-style results for quarter note tempo and meter recognition. Comparing these results with results from other work is not possible because other data-sets are used there. However, accuracies of up to 98% for certain dance-styles and mean accuracies around 90% definitely prove that our approach is state-of-the-art.

Beat tracking was evaluated on a subset 225 songs from the *BRD* set, containing 25 songs of each dance-style. Detailed results are also given in tab. 2. The results were evaluated considering three different aspects. (A) refers to the percentage of songs where beats were labeled correctly on at least one metrical level. Songs with tempo detection octave errors are included in this evaluation. (B) refers to the percentage of songs where beats on the quarter note level were labeled correctly. Using information about ballroom dance-style from the following classification-step the numbers for (B) could be improved by far. The relative percentage of identified downbeats indicates on how many songs of those where the quarter note tempo was tracked correctly, the beginning of a meter (the "1") was identified correctly. Due to the off-beat structure of Jive pieces half of them were tracked correctly regarding tempo, but at an off-beat phase. Thus, they were counted as incorrectly tracked.

By using data-driven classification with SVM (feature-set FS_M), meter recognition accuracies (see tab. 3), are improved by 4%. Detailed results for Recall, Precision and F_1-Measure for data-driven ballroom dance-style classification by SVM using feature-set FS_D are printed in tab. 2. A total of **90%** correctly classified instances can be reported for the latter.

Accuracy	Recall triple meter	Recall duple meter
97.5%	96.3%	97.9%

Tab. 3. - Meter classif. by SVM, feature-set FS_M, 10-fold SCV.

7. CONCLUSION AND OUTLOOK

Within this work we presented a holistic approach combining tempo detection, meter recognition, beat tracking, and ballroom dance-style classification. For the latter, a significant absolute increase of 10% considering mean recall rate over existing work on the same data-set [3] can be reported. A working prototype for ubiquitous on-line dance-assistance provides dance-style announcement and counting or step-instructions via a Bluetooth ear-plug. A minimum of 15 sec audio-material of a song is needed for analysis on a wearable computing device – enough time to ask someone for this dance. However, results are slightly more robust having 30 sec, as in our test-results. In future work we aim at more robust location of meter boundaries (i.e. identifying downbeats) by chroma-based Eigen-texture chord-change-modeling and/or BIC-based analysis of quarter note beat positions. Ballroom dance-style classification results shall be improved by integration of additional high-level features, such as timbre information and cepstrum-based rhythmic descriptors [3]. Furthermore we will investigate the effects of ballroom noise conditions and optimal microphone placement for audio signal capturing.

8. REFERENCES

[1] http://www.ballroomdancers.com, 2006.

[2] F. Gouyon, P. Herrera, "Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors," *Proc. of Audio Engineering Society*, 114th Convention, Amsterdam, The Netherlands, 2003.

[3] F. Gouyon, S. Dixon, E. Pampalk, G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," *Proc. AES 25th Int. Conf.*, London, UK, 2004.

[4] E. Scheirer, "Tempo and Beat Analysis of Acoustic Musical Signals," *Acoustic Society of America*, 103(1), p. 588-601, 1998.

[5] J. Seppänen, "Computational models of musical meter recognition," *M. Sc. Thesis*, Tampere Univ. of Technology, 2001.

[6] B. Schuller, F. Wallhoff, D. Arsic, G. Rigoll : "Musical Signal Type Discrimination Based on Large Open Feature Sets," *Proc. ICME 2006*, p. 1089-1093, Toronto, Ontario, 2006.

[7] I. H. Witten, E. Frank: Data Mining, Practical machine learning tools with Java implementations, Morgan Kaufmann, San Francisco, pp. 133, 2000.