RHYTHM INFERENCE FROM AUDIO RECORDINGS OF IRISH TRADITIONAL MUSIC

Pierre Beauguitte
Dublin Institute of Technology
pierre.beauguitte@mydit.ie

Bryan Duggan
Dublin Institute of Technology
bryan.duggan@dit.ie

John D. Kelleher
Dublin Institute of Technology
john.d.kelleher@dit.ie

ABSTRACT

A new method is proposed to infer rhythmic information from audio recordings of Irish traditional tunes. The method relies on the repetitive nature of this musical genre. Low-level spectral features and autocorrelation are used to obtain a low-dimensional representation, on which logistic regression models are trained. Two experiments are conducted to predict rhythmic information at different levels of precision. The method is tested on a collection of *session* recordings, and high accuracy scores are reported.

1. INTRODUCTION

Our goal is to automatically extract rhythmic information from an audio recording of Irish traditional music (ITM). The large majority of the repertoire can be categorized in a small number of tune types, often related to dance forms (Vallely, 2011). Metres used are:

- simple duple: $\frac{4}{4}$ (reel, hornpipe, fling, barndance) and $\frac{2}{4}$ (polka)
- simple triple: $\frac{3}{4}$ (waltz, mazurka)
- compound duple: $\frac{6}{8}$ (jigs) and $\frac{12}{8}$ (slides)
- compound triple: $\frac{9}{8}$ (slip jigs)

Simple and *compound* refer to the beat subdivision, *duple* and *triple* refer to the grouping of beats. No assymetric metres such as $\frac{5}{8}$ or $\frac{7}{8}$ are found in this musical tradition. Rather than focusing on the metre, we are interested in the tune type. Indeed, inferring a $\frac{4}{4}$ metre would not allow us to differentiate between a reel and a hornpipe, although their rhythm is noticeably different, the latter being interpreted with a clear swing. Melodies in ITM have a lot of repetition, and the majority of the notes have the duration of a quaver. This rhythmic stability makes it possible to extract useful information locally, from short excerpts. Slight tempo deviations can occur in live performances, and the use of a short sliding window will allow us to accomodate for these.

The method we introduce in this article first computes an onset detection function, then uses autocorrelation to extract its periodicity. No prior knowledge such as beat location or tempo is required. Rather than relying on handcrafted decision criteria or predefined templates reflecting musical knowledge, we will use a statistical approach to learn decision functions from a novel representation summarizing the autocorrelation function (ACF). As a first step in this study, we will attempt to predict the beat subdivision (*simple* or *compound*). Then, the same method will be used to predict the tune type of an audio recording.

In Section 2 we present some related work on rhythm inference, not restricted to ITM. We introduce the dataset of recordings used in this study in Section 3. Then we present in Section 4 our proposed method. Results are reported and discussed in Section 5. Section 6 contains closing remarks and ideas for future work.

2. RELATED WORK

Brown (1993) is an early example of using autocorrelation to determine the metre of a piece of music from its score. Decision criteria on the ACF are explicitly defined. Also focusing on symbolic music, Toiviainen & Eerola (2006) use discriminant function analysis to predict the metre of folk tunes. Two experiments are conducted, first to distinguish duple and triple metre, then the actual time signature. As stated in Section 1, our first experiment concerns the distinction of simple vs. compound metre instead. Indeed for the musical genre considered here, it is more natural to keep jigs and slip jigs (both compound) in a same category than *e.g.* jigs and polkas (both duple).

Pikrakis et al. (2004) and Fouloulis et al. (2013) determine the metre of Greek traditional music recordings, including assymetric metres, by hand-crafted decision criteria or template matching on the auto similarity matrix.

In Coyle & Gainza (2007), the time signature is also detected using self-similarity matrix, but the method is based on a prior knowledge of the tempo. The method presented in Gouyon & Herrera (2003) relies on beats extracted in a semiautomatic manner, and uses hand-crafted decision criteria to infer the metre. Gainza (2009) and Varewyck et al. (2013) first extract the beats from the raw audio, then determine the metre by analysing inter-beat similarity.

3. DATASET

We will use as our dataset for this study the collection of recordings accompanying the Foinn Seisiún books published by the Comhaltas Ceoltóirí Éireann organisation. They offer good quality, homogeneous examples of the heterophony inherent to an Irish traditional music session, although some solo recordings are also present. Instruments in the recordings include flute, tin whistle, uillean pipes (Irish bagpipes), accordion, concertina, banjo, piano, guitar, bodhran (drum). The whole collection consists of 3

CDs, representing 326 unique recordings. The first 2 CDs (273 tunes) are available under a Creative Commons Licence, while the third is commercially available.

We label each recording by the type of tune played. In most cases the type can be easily identified, but two notable exceptions need mentioning: *Fanny Power* was written as a jig by Turlough O'Carolan ; *Brian Boru's march* is written as a $\frac{6}{8}$ march. However, in the recordings they are arguably played as waltzes, and we decide to label them as such. Two songs are present, both with a duple simple metre. The distribution of tunes per tune type is given in Table 1, as well as the beat subdivision of the metre.

Type	number of tunes	beat subdivision
reel	139	simple
jig	104	compound
polka	28	simple
hornpipe	18	simple
slide	14	compound
barndance	6	simple
waltz	5	simple
mazurka	4	simple
slip jig	3	compound
fling	3	simple
song	2	simple
	205	simple
	121	compound

Table 1: Distribution of tunes per type

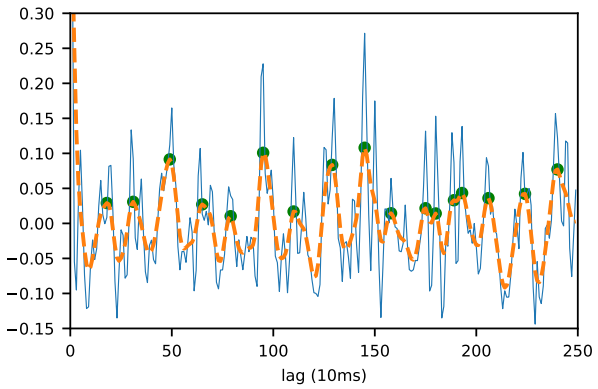

Figure 1: Peak picking of the ACF function. Solid line: ACF function. Dashed line: smoothed function

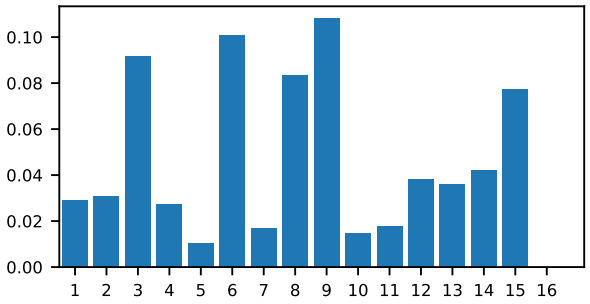

Figure 2: Quantized lag vector

4. METHOD

We now give the details of our proposed approach. First we explain how the audio files are processed to obtain *quantized lag vectors*, then how we train a logistic regression model to infer rhythm features from these vectors.

4.1 Audio processing

The audio files we consider are sampled at 44100Hz. A magnitude spectrogram is generated, with window size of 2048 and step size of 10ms, or 441 samples. This spectogram is then filtered through a filter bank of triangular filters centered at Bark frequencies, resulting in a Bark spectrogram $X_k(t)$ where $1 \leq k \leq 24$ is the Bark index. Following Bello et al. (2005), we obtain an onset detection function by a method of spectral difference:

$$SD(t) = \sum_{k=1}^{24} (H(X_k(t) - X_k(t-1)))^2 \quad \text{for } t > 0$$

where the rectifier $H(x) = (x + |x|)/2$ has the effect of ignoring decreases of energy, because it is equal to zero for negative values. Thus it emphasises onsets more than offsets. As the energy difference is computed in each spectral band before being summed, the presence of percussive instruments is not required to detect onsets.

The autocorrelation functions is then computed on a 5s window of the *SD* function $(w_t) = (SD(t_0+t))_{0 \leq t < N=500}$ (where t_0 is the start of the window) using Pearson correlation coefficient. The autocorrelation for a lag l is:

$$ACF(l) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \text{ where } \begin{cases} X = (w_t)_{0 \leq t < N-l} \\ Y = (w_t)_{l \leq t < N} \end{cases}$$

where cov is the covariance and σ designates standard deviation. We smooth this function by Gaussian filtering with a standard deviation of 20ms, and find the local maxima of this smoothed curve, ignoring the always present peak at $l = 0$. Figure 1 shows and example of the peak picking procedure on a window of a jig. Each peak p has a lag and a value, represented by p_l and p_v respectively. For the goal of this study, what matters is not the actual locations of the peaks p_l, but their relative positions from each other. By abstracting our representation from the actual lag values, we will obtain a form of tempo invariance. The quaver duration will be extracted from the peaks locations and then used to compute a *quantized* representation.

The quaver duration q is determined by the *fuzzy histogram* algorithm, introduced in Duggan (2009), and given in Algorithm 1. The intervals, or lag differences, between the peaks are grouped into bins, allowing for a deviation of a fraction of the bin center, set to $1/3$. The centers of the bins are adjusted for each new interval added. The quaver length is taken as the center of the largest bin.

Knowing the quaver length will now allow us to obtain a tempo-invariant representation of the peaks of the ACF. The following step involves summing peak values.

```
Data: P, list of peaks of the ACF (size l)
Result: quaver length
bins ← {};
max ← 0;
for i ← 1 to l do
    if i = l then
        dur = P[i]_l;
    else
        dur = P[i]_l − P[i − 1]_l;
    end
    found ← false;
    for b in bins do
        bin_start ← b.center∗(1 − 1/3);
        bin_end ← b.center∗(1 + 1/3);
        if dur ≥ bin_start and dur ≤ bin_end then
            found ← true;
            b.center
              ← (b.center∗b.count+dur)/(b.count+1);
            b.count += 1;
            break;
        end
    end
    if found = false then
        newBin.center ← dur;
        newBin.count ← 1;
        bins.add(newBin);
    end
end
for b in bins do
    if b.count > max then
        maxBin ← b;
        max ← b.count;
    end
end
return maxBin.center;
```

Algorithm 1: Fuzzy histogram algorithm, adapted from (Duggan, 2009)

We now introduce the *quantized lag vector* $(ql_i)_{1\leq i\leq 16}$ [1] obtained by first grouping the peaks as:

$$P_i = \{p \in P \text{ where } round(p_l/q) = i\}$$

and averaging across these sets:

$$ql_i = \begin{cases} \left(\sum_{p\in P_i} p_v\right)/|P_i| & \text{if } P_i \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

An example of such a vector is plotted in Figure 2, computed from the ACF peaks shown of Figure 1. The ratio of the first nine peaks is preserved, but the absolute durations of the lags have been discarded, making this representation tempo-invariant. Some of the subsequent peaks are grouped together by the rounding operations. More prominent peaks appear at multiples of 3, as is to be expected from the compound metre of that tune type (jig).

Each 5s window produces a 16 valued vector, and we slide the window with a step size of 0.5s. Choosing such a small step size results in a large amount of examples, which is an advantage for the machine learning methodology we present in the next section.

[1] The number of 16 quavers was chosen empirically. Experiments with alternative values did not lead to significantly different results.

4.2 Model training

Regression analysis in general attempts at modelling the relationship between independent variables x (here the *ql* vectors) and a dependent variable y (here the rhythm information). We will use logistic regression models, or classifiers, because our dependent variables are categorical, *i.e.* they can only take one of a given set of values. A similar methodology will be used to predict, in a first experiment, the beat subdivision and, in a second one, the tune type.

4.2.1 Experiment A: beat subdivision prediction

The dataset consists of pairs (x, y), where x is a *ql* vector and y a label in $\{simple, compound\}$. We use 10-fold cross validation as a way to evaluate how well the models generalize (Kelleher et al., 2015). Each fold is, in turn, kept as a test set, and a binary classifier is trained on the remaining 9. When preparing the folds, we make sure to keep all *ql* vectors from one tune in only one of the folds. This way, the models will be tested on recordings that have not been used during training, thus avoiding a form of cheating.

To account for the fact that the classes *simple* and *compound* are not balanced in the dataset, during training the error on an instance is weighted by the inverse of the relative frequency of the output class of the instance in the training set; *i.e.*, errors on *compound* instances are given a higher weighting than errors on *simple* instances in the calculation of the loss function to account for the fact that *compound* instances are less frequent.

4.2.2 Experiment B: tune type prediction

In this second experiment, we attempt to predict the tune type from the *ql* vector. Because some tune types are too rare in the dataset, we limit ourselves to the 5 types at the top of Table 1, namely reel, jig, polka, hornpipe and slide. There are only 14 slides in the collection, hence using 10-fold cross validation would only result in one or two of them in each fold. To avoid this problem, we use 4-fold validation instead. For each fold, a multinomial logistic regression classifier is trained in a *one-versus-all* manner, meaning that the model actually consists of a set of binary classifiers. As in experiment A, during the training phase, errors are weighted by the inverse of the relative frequency of the output class.

5. RESULTS AND DISCUSSION

We now report the results of our 2 experiments. Accuracy scores are given for aggregate matrices resulting from the k-fold cross validation methodlogy described above.

The models of both experiments predict a label for a 5s window. In addition to the window-level scores, we are also interested in predictions across a span of several consecutive windows. The reason we are interested in this is that rhythm is not as easily identifiable on all 5s sections of a tune. Thus we hope to reach better accuracy by gathering predictions on a longer segment. The prediction over a span of s windows is simply defined as the most frequent of the s predictions. We report performances at window-level, across s windows, and finally over whole tunes.

	simple	compound
simple	26910	1105
compound	2292	15515
overall (%)	92.2	93.4

Table 2: Aggregate confusion matrix at window-level for experiment A (column: reference, line: prediction)

Type	Accuracy (%)
reel	96.5
jig	95.1
polka	88.6
hornpipe	86.5
slide	79.1
barndance	93.4
waltz	65.4
mazurka	68.2
slip jig	99.2
fling	85.0
song	96.1

Table 3: Window-level accuracy score per tune type for experiment A

5.1 Experiment A

The aggregate confusion matrix resulting from the 10-fold cross validation is given on Table 2. The overall accuracy score at the 5s window level is 92.6%. The prediction accuracy is slightly lower on the *simple* class than on the *compound* class. A possible explanation for this is that there are more distinct tune types included in the *simple* class (reel, hornpipe, polka, waltz,...) than in the *compound* class (only jig, slip jig and slide), as can be seen on Table 1. Looking at the score per tune type on Table 3, we see that it is particularly low for waltz and mazurka, both in simple triple metre $\frac{3}{4}$. Mazurkas are also interpreted with a noticeable swing.

When considering spans of successive overlapping windows, the accuracy increases up to 99.3%, as is shown on Figure 3. We can only compute this up to a span size of 87 windows, corresponding to the duration of the shortest tune in the collection.

Lastly, for each tune, we consider the prediction over the span of all windows of its recording. At this tune-level, the prediction only fails on 3 tunes, all of type *slide*. The overall prediction accuracy is of 99.1%.

Although the task tackled in this first experiment is arguably easy, these near-perfect scores are very encouraging and suggest that our *ql* vector representation does capture some useful rhythmic information.

5.2 Experiment B

The aggregate confusion matrix resulting from the 4-fold cross validation is given on Table 4, and the overall accuracy score at the 5s window level is 82.7%. The accuracy per window span length is shown on Figure 4, and reaches a maximum of 92.9% at $s = 87$.

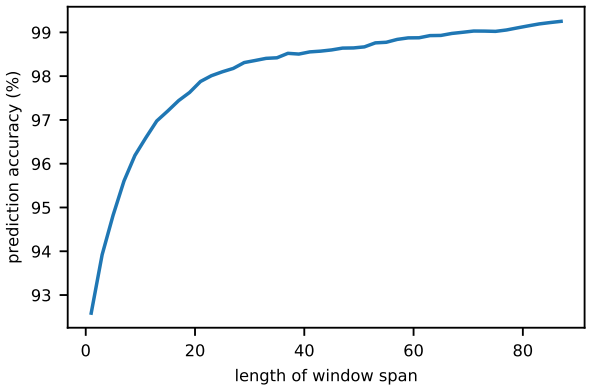

Figure 3: Prediction accuracy by span length for experiment A

	reel	jig	polka	hornpipe	slide
reel	16421	406	617	135	82
jig	296	12577	90	239	1087
polka	339	207	2602	209	200
hornpipe	920	120	198	2537	106
slide	614	1058	115	188	408
overall (%)	88.3	87.5	71.8	76.7	21.7

Table 4: Aggregate confusion matrix at window-level for experiment B (column: reference, line: prediction)

Finally, the confusion matrix for tune-level prediction is given in Table 5. The overall score on slides is low, which is in line with the observation made on tune-level predictions for experiment A. Most of the slides are misclassified as jigs. Both are in duple compound metres, which suggests that the model did manage to capture relevant features, but could not make a good enough distinction between these two tune types. However all 18 hornpipes in the dataset have been correctly classified, despite sharing the $\frac{4}{4}$ time signature with reels. Our method manages to distinguish two tune types having distinct "rhythm signatures" but the same metre.

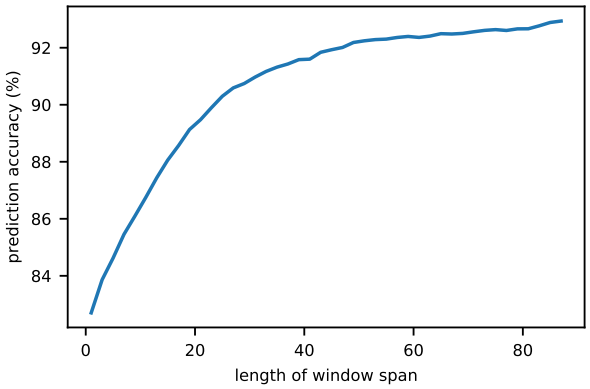

Figure 4: Prediction accuracy by span length for experiment B

	reel	jig	polka	hornpipe	slide
reel	137	0	3	0	0
jig	1	104	0	0	9
polka	0	0	25	0	2
hornpipe	1	0	0	18	0
slide	0	0	0	0	3
overall (%)	98.6	100	89.3	100	21.4

Table 5: Aggregate confusion matrix at tune-level for experiment B (column: reference, line: prediction)

6. CONCLUSION

We introduced a new method for inferring rhythm information from an audio recording, using low-level spectral features and logistic regression classifiers. The performance on the dataset was very good, or even perfect, for some types of tunes (jigs, hornpipes). Other tune types proved to be more challenging (slides), while others were too rare in our collection to be considered.

In future work, we hope to be able to predict more tune types. In order to do so, a larger collection of recordings will have to be used, so more examples can be used to train our models. Testing our models on solo recordings would be useful to further assess the robustness of our proposed approach. Indeed, although our onset detection function relies on spectral content and not on hard onsets from percussive instruments, drums or plucked string instruments (guitar, banjo) are present in most of the recordings in our dataset. Applying our method to flute or fiddle solo recordings could establish to what extent hard onsets help the rhythm inference.

7. REFERENCES

Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. B. (2005). A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 1035–1047.

Brown, J. C. (1993). Determination of the meter of musical scores by autocorrelation. *The Journal of the Acoustical Society of America*, *94*(4).

Coyle, E. & Gainza, M. (2007). Time Signature Detection by Using a Multi-Resolution Audio Similarity Matrix. In *Audio Engineering Society Convention 122*.

Duggan, B. (2009). *Machine annotation of traditional Irish dance music*. PhD thesis, Dublin Institute of Technology.

Fouloulis, T., Pikrakis, A., & Cambouropoulos, E. (2013). Traditional asymmetric rhythms: a refined model of meter induction based on asymmetric meter templates. In *Proceedings of the Third International Workshop on Folk Music Analysis*, (pp. 28–32).

Gainza, M. (2009). Automatic musical meter detection. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, (pp. 329–332). IEEE.

Gouyon, F. & Herrera, P. (2003). Determination of the Meter of Musical Audio Signals: Seeking Recurrences in Beat Segment Descriptors. In *Proceedings of the 114th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. Cambridge, Massachusetts: The MIT Press.

Pikrakis, A., Antonopoulos, I., & Theodoridis, S. (2004). Music meter and tempo tracking from raw polyphonic audio. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, Barcelona (Spain).

Toiviainen, P. & Eerola, T. (2006). Autocorrelation in meter induction: The role of accent structure. *The Journal of the Acoustical Society of America*, *119*(2), 1164.

Vallely, F. (2011). *The Companion to Irish Traditional Music* (Second edition ed.). Cork: Cork University Press.

Varewyck, M., Martens, J.-P., & Leman, M. (2013). Musical Meter Classification with Beat Synchronous Acoustic Features, DFT-based Metrical Features and Support Vector Machines. *Journal of New Music Research*, *42*(3), 267–282.