Segmentation, Transcription, Analysis and Visualisation of the Norwegian Folk Music Archive

Olivier Lartillot
olivier.lartillot@imv.uio.no
RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, University of Oslo
Oslo, Norway

Anders Elowsson
anderselowsson@gmail.com
RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, University of Oslo
Oslo, Norway

Mats Johansson
Mats.S.Johansson@usn.no
University of South-Eastern Norway
Rauland, Norway

Hans-Hinrich Thedens
Hans-Hinrich.Thedens@nb.no
National Library of Norway
Oslo, Norway

Lars Monstad
lars.monstad@gmail.com
RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, University of Oslo
Oslo, Norway

ABSTRACT

We present an ongoing project dedicated to the transmutation of a collection of field recordings of Norwegian folk music established in the 1960s into an easily accessible online catalogue augmented with advanced music technology and computer musicology tools. We focus in particular on a major highlight of this collection: Hardanger fiddle music. The studied corpus was available as a series of 600 tape recordings, each tape containing up to 2 hours of recordings, associated with metadata indicating approximate positions of pieces of music. We first need to retrieve the individual recording associated with each tune, through the combination of an automated pre-segmentation based on sound classification and audio analysis, and a subsequent manual verification and fine-tuning of the temporal positions, using a home-made user interface.

Note detection is carried out by a deep learning method. To adapt the model to Hardanger fiddle music, musicians were asked to record themselves and annotate all played note, using a dedicated interface. Data augmentation techniques have been designed to accelerate the process, in particular using alignment of varied performances of same tunes. The transcription also requires the reconstruction of the metrical structure, which is particularly challenging in this style of music. We have also collected ground-truth data, and are conceiving a computational model.

The next step consists in carrying out detailed music analysis of the transcriptions, in order to reveal in particular intertextuality within the corpus. A last direction of research is aimed at designing tools to visualise each tune and the whole catalogue, both for musicologists and general public.

DLfM2022, July 28, 2022, Prague, Czech Republic

ACM ISBN 978-1-4503-9668-4/22/07.
https://doi.org/10.1145/3543882.3543883

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**.

KEYWORDS

music archive, folk music, transcription, audio segmentation, beats, visualization

ACM Reference Format:
Olivier Lartillot, Anders Elowsson, Mats Johansson, Hans-Hinrich Thedens, and Lars Monstad. 2022. Segmentation, Transcription, Analysis and Visualisation of the Norwegian Folk Music Archive. In *9th International Conference on Digital Libraries for Musicology (DLfM2022), July 28, 2022, Prague, Czech Republic.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3543882.3543883

1 INTRODUCTION

There is a growing amount of research dedicated to the development and application of Music Information Retrieval (MIR) technologies dedicated to the analysis of music archives [3]. The study presented in this paper is being carried out in the context of the MIRAGE project[1], funded from 2020 to 2023, and aimed at designing new technologies for computational music analysis, with close interaction with musicology needs, with application in music cognition, for the public and for music libraries. The aim of this present work is to complete the digitalisation of the Norwegian Folk Music Collection—or more precisely the subset that can be considered in the public domain—and to augment the audio recordings with automated score transcription and music analysis. The objective in a longer term is to provide a large range of music analysis and visualisation and navigation capabilities, in order to make the music largely accessible to both scholars and the public. We are also making the technologies and interfaces available to the research community.

[1] https://www.uio.no/ritmo/english/projects/mirage/index.html

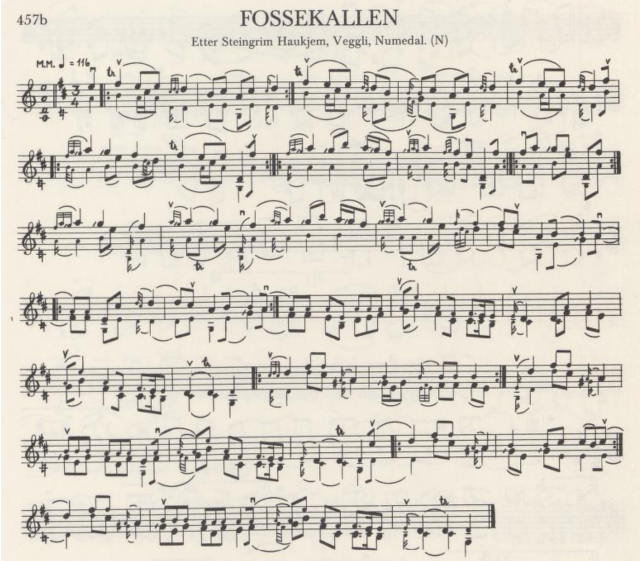

Figure 1: Transcription of a Springar tune called *Fossekallen* [2]. © The Norwegian Research Council for Science and the Humanities 1979

2 CORPUS

2.1 Hardanger Fiddle Music

The Hardanger fiddle is a variety of the violin used in the folk music of the western and central part of southern Norway. It is played as a solo instrument for couple dancing. It spread from the area of Hardanger to many of the valleys in southern Norway in the 18th century. The intricacy of the performance style makes machine transcription difficult. Publishing transcriptions of fiddle tunes has been an important part of Norwegian folk music research in the 20th century. Between 1958 and 1981 the Norwegian Folk Music Institute published seven volumes of such transcriptions covering most of the traditions of Hardanger fiddle playing, organized by melody type and in tune families[2]. Figure 1 shows the transcription of one of the pieces.

2.2 Norwegian Folk Music Archive

The Norwegian Folk Music Archive, now at the Norwegian National Library, was founded in 1951 with the purpose of building a corpus of audio recording for research. First musicians were invited to Oslo to record and later collectors went for trips to areas with strong folk music traditions. The recordings were made on reel-to-reel tape machines.

We are preparing the online publication of on one subset of the recordings, those taken from the years 1953 to 1968, in order to avoid music copyrights issues related to more recent recordings. This corresponds to 900 hours of recordings, which were made on 600 reel-to-reel tape machines. Detailed metadata has been associated with the recordings, including the name of the tunes and information related to the performers.

[2]https://www.nb.no/forskning/feleverkene/

3 SEGMENTATION OF TAPES INTO TUNES

The tape recordings have not previously been segmented into tunes and songs, so this needed to be taken care of. For a third of these recordings, rough indication of temporal location of each tune in the corresponding tape is indicated in the metadata, but not with sufficient details to be used directly for audio segmentation. Due to the very large number of tunes to extract (ca. 20.000), it was necessary to automate or at least semi-automate the process.

3.1 State of the art

Automated segmentation of audio recordings is a research topic that has been investigated for several decades [11], with a particular focus on the discrimination between speech and non-speech—and in particular music—parts in radio broadcasts, as well as a discrimination between foreground and background music. More closely related to our study, a recent study has developed a tool aimed at labelling and segmenting field recordings into individual units labelled as speech, solo singing, choir singing, and instrumentals [11]. The software, SeFiRe, is available for free, and offers the possibility of visualising the recordings and the segmentation as well as modifying the results if needed. The approach is aimed at ignoring particularly short speech segments within longer music sections, focusing rather on larger scale segmentation. There is no attempt to precisely find the starting time of each music section. In our case, we would need to detect very short speech sections that sometimes separate successive tunes, as well as to get a precise estimation of the starting time of each tune.

Commercial sound classification software solutions are also available. Apple's SoundAnalysis framework [3] includes a sound classifier that can identify over 300 different types of sounds. This framework can be integrated into software running on Apple devices and be distributed for free.

3.2 Proposed solution

We use the sound classifier included in Apple's SoundAnalysis framework, focusing on the sound classes Music, Speech, Singing, Bowed and Violin. Sound identification is performed on a moving window of duration 1.5 seconds with a step of 0.75 seconds. For each successive time window, for each sound class is associated a score. We obtain a superposition of time curves, shown in various colors in Figure 2. Bowed and Violin are further fused together, by taking the maximum score. Music score is used to characterised moments where there is music detected but without a clear detection of either singing or bowed instrument, or for other types of music.

Further process is necessary to turn this multidimensional continuous data into a clear temporal segmentation of the audio based on those scores. We design a solution consisting of a graph of states, as shown in Figure 3, corresponding to various configurations related to the time curves, and where transition between segments is modelled by particular transitions between states, themselves conditioned by threshold detection on one or several time curves.

For instance, from the Start state, the transition to the Speech state requires a Speech score higher than .85, on a scale between 0 and 1. If the Speech score further decreases below .3 (and if the Singing score is low), there is a transition towards the Speech? state,

[3]https://developer.apple.com/documentation/soundanalysis

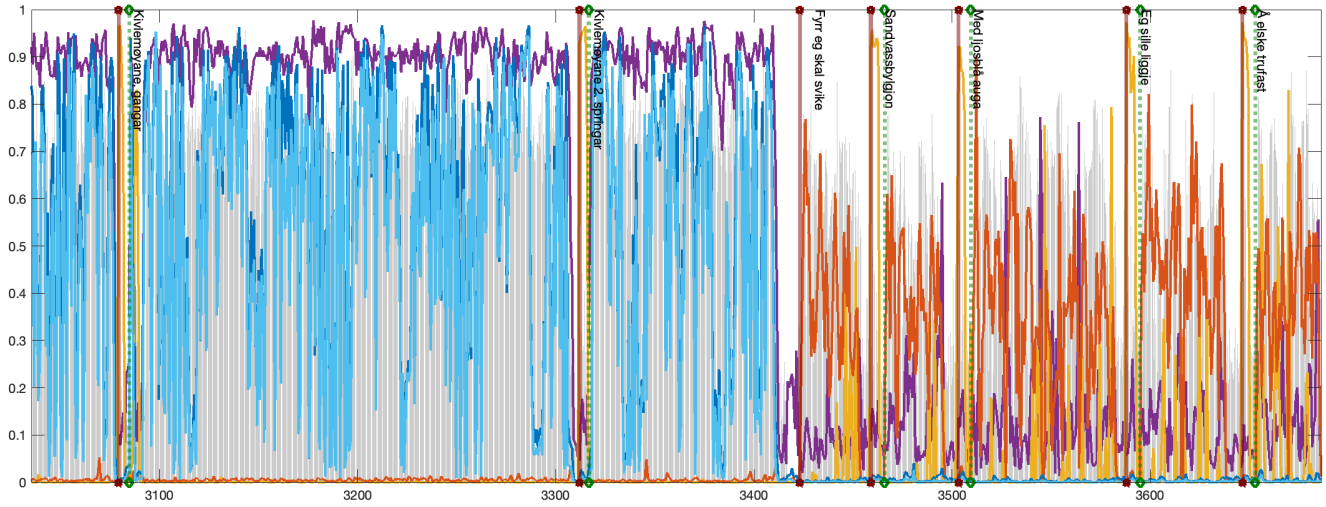

Figure 2: Screenshot of the AudioSegmentor GUI. Coloured curves correspond to the scores, for each successive temporal frame from left to right, related to the classifiers Speech (yellow), Singing (red), Music (purple), Bowing (dark blue) and Violin/fiddle (light blue) of an excerpt of a tape recording. Audio dynamics is represented in grey. Red vertical lines indicate the beginning of tunes, as detected by the automated segmentation system, generally starting with a little speech, and then followed by the actual music, starting at the green vertical line. The title of each tune is shown on the right of each read line.

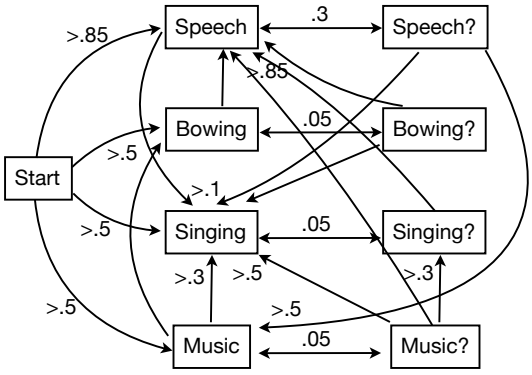

Figure 3: Speech and music type states and transitions between them. For each transition is indicated the required threshold for the score related to the destination state. See the text for further explanation.

and a further transition back to the Speech state if the score exceeds .3 again.

As explained above, the Music state is used to denote a music segment where neither singing nor bowing are detected. This state can turn into one of these two more specific states at a later stage if the corresponding threshold is reached. Short speech segments of less than 2 second in between Singing segments are removed and the two Singing segments are merged.

By dividing each identification type (Speech, Bowing, etc.) into two separate states ("Speech" and "Speech?"), this enables to avoid false transition between identification types. For instance, when in Singing state with high Singing score, a high Speech value does not trigger a transition to the Speech state. Only when the Singing score gets lower than .5—entering the "Singing?" state—could such transition happen.

The starting position of each segment is fine-tuned through a detection of the rising phase from the dynamics in the audio recording, so that the start time is reassigned half a second before the lowest amplitude point, at the very beginning of the rising phase.

The positions of the segments are then verified and, if needed, corrected by an expert, with the help of a graphical user interface, called AudioSegmentor, that we have designed for that purpose (cf. Figure 2). The tune names are retrieved from the metadata and displayed. The successive segments can be quickly listened to, translated along time, or deleted. The resulting segments positions are finally exported back into the metadata database, and the tune recordings are collected.

3.3 Current state

In the middle of June 2022, 70 tape have been segmented (corresponding to more than 5000 tunes), with an objective of 120 tapes at the end of July. Until June it took from 10 minutes (for the easiest cases where tapes only contained solely fiddle music separated by little speech) to to one hour for more complex cases, in particular those involving lullabies songs. The objective is to significantly reduce further this time spent on manual editing, by improving the editing capabilities and improving the automated presegmentation.

4 NOTE DETECTION

Automated transcription can be decomposed into two separate tasks: first, detection of individual notes and their intrinsic characterization, in particular corresponding to their time localisation

(in seconds) and their pitch height (in Hertz). The subsequent task, described in the following sections, incorporates these individual notes within the musical context, defined in particular by the metrical and modal or harmonic structures.

The first step is to detect all the “notes" played by the musician. The playing style in Hardanger fiddle music is generally very ornamented, each ornamented “note" is often decomposed into a series of very short discrete events with a starting point (or onset time) and an ending point (or offset time). In this style of music, each of those discrete events generally features a rather stable pitch height (expressed as frequency in Hz). The objective of the first step of automated transcription consists in detecting the notes at this micro-level and measuring their onset, offset and pitch as precisely as possible.

4.1 State of the art in automated transcription

Commercial music technology plugins such as, Celemony Melodyne 5, Logic Pro X Flex Pitch and Cubase Pro 12 VariAudio, have been tested on a few pieces of our corpus, with unsatisfying results. The main issue with pitch tracking plugins such as Melodyne is that the software fails to filter out noise and the pitch harmonics of a single note are often transcribed as individual polyphonic notes.

Automated music transcription remains a nascent research topic. Robust technologies might exist for particular types of music such as solo piano, but in more general cases, the results are less reliable. State-of-the-art approaches are based on deep learning framework, although Non-negative Matrix Factorization still showing some advantages sometimes [1]. Machine learning methods requires a sufficiently large number of manual annotations of music recordings, used as training data.

Our approach is based on the “deep layered learning" [4]. In the level of onset detection this consists in the establishment of a 2-dimensional “onsetgram," consisting of onset activations distributed across pitch and time. The onset activations are first computed using a polyphonic transcription system, whereby an initial network detects framewise f0 activations, in the form of a so-called “pitchogram", which are used to identify the contours of the music. An additional network then operates across each detected contour, computing an onset activation at each time frame of the contour. The onset activations are inserted at the corresponding pitch bin and time frame of the onsetgram.

4.2 Proposed approach

The model presented in [4] was initially trained on a wide variety of music, but not on Hardanger fiddle music. The proposed approach is based on the constitution of annotated music, with the desired level of details and precision, from our corpus of study, to be used as training data for the machine learning models, and in particular for the deep layered learning approach mentioned above.

To facilitate the development of high-quality transcription algorithms, we need examples of transcriptions that could be considered as references. Concerning the first step, we need to collect a set of high-quality annotations of recordings of music performances, listing the temporal location (in seconds) and pitch height of each note.

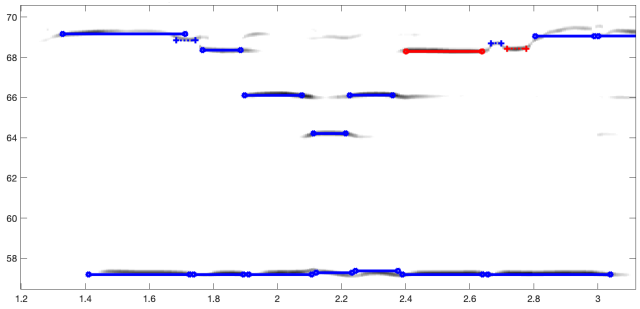

Figure 4: Screenshot of the Annotemus software, showing the pitchogram (in grey) of an excerpt of a Hardanger fiddle tune, and the manual note annotations, shown with vertical blue and red lines. Red lines correspond to selected note annotations.

The validation of the machine learning predictions is typically ruled by very restrictive criteria, tolerating no more than 50 ms of difference between the onset (or offset) positions given by the machine and by the human annotator, and no more than 50 cents in pitch height. We therefore needed to ask the annotators to indicate all the micro-events and determine their onset, offset and pitch with this very high level of precision, which can easily be a cumbersome task. It is also preferable to ask the annotators to annotate from scratch, without relying on any suggested annotation given by any prior transcription system, to avoid any bias towards that prior transcription.

To enable our annotators to work efficiently, they therefore should work on material that they are familiar with. For these reasons, we hired three skilled musicians to record songs that they are deeply familiar with. Olav Luksengård Mjelva is a well-known professional musician fidler and Astrid Garmo and Henrik Nordtun Gjertsen were students at the Norwegian Academy of Music at the time of the recording. Each musician also annotated each of their recordings. The students’ recordings and annotations have been published [5].

4.3 Annotemus, a new annotation software

To ensure that the annotation be carried out with minimum time and effort while obtaining data of the highest quality, it is important to make sure that the interface used to perform the annotation be as convenient and easy to use as possible. For that purpose, we designed—in the MATLAB programming language—our own software, called Annotemus, aimed at facilitating manual annotation of note onset, offset and pitch. First a graphical representation of the audio recording is displayed as a visual support for the annotation task. A simple solution would consist in displaying a spectrogram. We chose to display a more refined representation, showing only the fundamental frequencies and filtering out the harmonics of each sound. The chosen representation is the pitchogram, part of the deep learning system presented above, section 4.2.

It is possible to zoom in by drawing a box to indicate the chosen time and frequency region and to listen to the selected time region in the audio recording. This seems to be a convenient method for annotating complex music: by focusing on short time sequences,

listening repetitively the corresponding audio excerpt, completing and correcting the annotation accordingly, and then moving the focus on a slightly later part. Horizontal lines, indicating the temporal location and pitch height of notes, are created by simply clicking and drawing. Lines can then be modified.

In some cases, it may be up to interpretation if a sound event should be included as a note or defined as noise. This may happen, for example, during very short notes, or if the player hesitates during bowing, creating the impression of an extra onset, or if a string is faintly touched during the performance. The annotator can therefore flag any note as uncertain.

The annotation can be played back alone, or with the audio recording. The temporal precision of the onset and offset positions can be checked aurally through various playback methods: The options are to play:

- The audio from the onset time to the offset time of the selected note. This enables users to listen to if the annotation spans the full extent of the note.
- The audio from 600 ms before the onset position up until the exact time point of the annotated onset. This can be used to locate the exact position where the onset of a new note becomes just barely audible.
- The audio from 500 ms before the onset position until the offset position, with a burst of noise indicating the onset position. In this way, annotators can listen to the start of the note, including a brief part before the start, and check whether the location of the burst of noise corresponds to what we would consider as the onset position.
- The audio from 500 ms before the onset position until 1 s after the offset position, with bursts of noise indicating both the onset and offset positions. Same as for the previous point, but this time checking both onset and offset times.

All these options (except the second one) can be performed on a sequence of selected annotations: the whole audio from the onset (or 500 ms before) of the first note to the offset (or 1 s after) of the last note is played, with, when appropriate, the bursts of noise corresponding to the selected notes onset and/or offset.

The system offers a rudimentary function for separating individual tones, so that the annotators can listen to them with less interference from other concurrent tones. The users can also select to perform the opposite operation, i.e., filter out the currently selected tones and only playback the residual. This is useful for determining if the currently annotated notes cover all partials present in the audio recording. More precisely, the filtering of notes is performed on the spectral domain (computed via a Constant-Q transform, or CQT) by applying masks at frequencies related to the fundamental and a series of harmonies for each annotated note. More details in [5].

All playback functionality is offered with the option of slowing it down to an arbitrary speed. Since Hardanger fiddle music contains frequent sequences of very fast note successions, the slowdown functionality was used extensively during the annotation process. A novel time-shifting algorithm was developed for that purpose. It consists of two parts: a resampling of the audio waveform in the time-domain to speed up or slow down the audio, and a pitch-shifting of the resulting audio waveform in the log-frequency domain to compensate for the change in pitch introduced by the resampling operation. More details in [5].

4.4 Augmenting the dataset by varying the performance style of given tunes

Annotating all the notes in a recording of one single tune remains time-consuming for the musicians, despite the improvements implemented in the Annotemus interface. Because we need to collect a significant amount of tunes annotations to train our machine learning models, we conceived a methodology allowing a reduction in the amount of annotation work [5]. The idea is to ask the musicians to play each tune five times, trying to keep the same ornamentations for each version, but at the same time varying significantly the other aspects related to timing, accentuation, nuances, playing styles, etc. More precisely, we invite the musicians to use a series of emotion-related musical expressions, namely: normal, sad, angry, happy, and tender. For each song, they annotate one of the recordings from scratch, using the Annotemus interface. These annotations can then be automatically transferred to the other performances of the same tunes, using automated music alignment. The annotators just need to check and adjust the preliminary transcriptions if needed.

The temporal alignment of two given audio files is automatically performed on their respective Onsetgrams, i.e., on the note annotation automatically inferred by the deep learning model. First, a start- and endpoint was computed for both audio files and the normal Onsetgram was then re-scaled to have the same length as the Onsetgram of the emotional expression using linear interpolation. The annotations were also re-scaled with the same transformation.

State-of-the-art approaches in temporal alignment are generally based on Dynamic Type Warping (DTW), which is “local" in scope and will not model differences in tempo and gradual tempo variations observed across longer sections. This can produce a rather irregular warping path, which is also discrete, bounded by the time frame hop length. To overcome this, we explored two separate techniques developed for image registration. The first approach was based on a free-form deformation with a B-spline grid [12]. It was performed at multiple different image scales (so-called pyramid levels), starting from a coarser scale to fit the Onsetgrams according to the general structure of the music. The finer scale of the last pyramid levels then accounted for local variations between the performances. The second approach was the Demons algorithm for non-rigid image registration [13]. It uses the gradient from one Onsetgram to compute a “demons" force that deforms the other Onsetgram. With this approach, individual pitch bins are allowed to diverge somewhat from the warping path to account for natural variations in timing between concurrent notes. The method produces displacement matrices for time and pitch that also account for local timing variations across the pitch range. These computed displacement fields are then used as a backward transformation to transfer the annotations to the recordings with emotional expressions.

In our tests, the Demons algorithm was more accurate even though the B- spline method was used as a starting point for the aligned expressive performances [5]. The Demons algorithm is faster and easier to adapt to music and it also produces the best

alignments. A method for adjusting the pitch of each note was applied as a post-processing step.

The dataset is further expanded using data augmentation. Several microphones are used during recording to create additional audio tracks. The tracks are also shifted in tempo and pitch, and randomly equalized. Furthermore, noise and ambiance are added.

In total 20 different tunes (of average duration between around 1 and 25 minutes) have been recorded, with 5 variants for each tune, totaling 100 recorded tunes. The audio recordings and annotations related to the two students' recording (8 out of the 20 tunes in total) are available online, as well as MATLAB source code [5]. A subsequent set of 12 tunes, with their 5 variants, is under preparation by the professional fiddler, Olav Luksengård Mjelva.

4.5 Application to the transcription of the music collection

This collected music annotation dataset of Hardanger fiddle music is used to train the deep learning model. As of June 2022, we are finishing the training phase and are starting testing the new model on tunes from the Norwegian folk music collection. One particular difficulty concerning evaluation is that there is no ground-truth data to which the model's transcription can be compared. Hence the results need to be evaluated manually by music experts. One potential strength of the approach is that the Annotemus interface should facilitate the precise correction of the model's output, which could then be used as additional training data.

5 RHYTHMIC ANALYSIS

The second step of music transcription involves the reconstruction of the metrical structure particular to each piece (discussed in this section), as well as of its modal/harmonic structure and the melody and drone lines (next section).

5.1 Played beats vs. experienced beats

Particularly challenging in Hardanger fiddle music is the determination of the metrical structure of each piece, due to the asymmetrical and fluctuating beat duration structure, and because the beats are not always clearly indicated by accentuated notes. In fact, even Hardanger fiddle music experts might find it challenging to precisely pinpoint the exact localization of beat onsets (i.e., the timepoint where a beat appears in the musical sound). Moreover, following a conversation with the professional musician Olav Luksengård Mjelva and the musicologist Mats Johansson, it appears that a distinction ideally should be made between precise annotation of beat onsets that are associated with onsets of notes—that is, played beats—and the beat that listeners would feel, in a kind of more approximate way, as expressed through, e.g., dance movements, foot tapping and head nodding. This second, more holistic, concept of beat—which could be called experienced beat or groove—is related to but not equal to the unfolding of played beats. Importantly, this does not mean that the played beats are to be understood as syncopations against or deviations from some actual beat onset position. Instead, what should be emphasized is that 1) a specific point in time is not an ecologically valid representation of an experienced beat onset, and that 2) "feeling the beat" is informed by the interaction between several musical parameters—including accentuation/dynamics, attack quality, phrasing, rhythmic subdivision, and melodic and ornamental articulation—and thus cannot be reduced to beat timing alone [8]. For all practical-analytical purposes in the present context, however, the played beats are the beats. Moreover, at this stage, our study is solely focused on music transcription and the most immediate need is to associate each detected note with its corresponding metrical position. Analytically exploring the concept of experienced beat, then, would require a correspondingly more holistic and multidimensional/-parametrical approach (cf. above) complemented by additional music cognition studies.

5.2 Beat annotation

Even if played beats can be indicated in a rather tangible way, there can still be possible divergence between annotators on the exact positions of those beat onsets. We collected, for the 12 tunes that have so far been recorded and note-annotated by Olav Luksengård Mjelva, the annotation of the played beats by the musician himself, as well as the musicologist Mats Sigvard Johansson and students with Hardanger fiddle expertise. The Annotemus software offers the capability to annotate beat onsets on top of note annotations, by selecting for each successive beat onset one note for which the annotate considers the onset to be synchronous to the beat onset. "Silent beat onset" for which no note onset is associated are simply ignored.

As aforementioned, Hardanger fiddle springar dances generally follow a triple meter, in the sense that each bar is decomposed into 3 beats. The first beat of a bar will be noted 1/3, the second beat 2/3 and the third beat 3/3. In traditional time meter notation, a quarter note is represented by the denominator 4, and an eighth note is represented by the denominator 8, implicitly setting the reference unit to the whole note, which is four beat long. Since we focus on triple meter, we prefer setting the reference unit to the three-beat bar, hence using the denominators 3 and 6 for the beat ("quarter" note) and the half-beat ("eighth" note) respectively. The first eighth note of a bar, 1/6 starts at the same time position as the first quarter note 1/3, while the second eighth note, 2/6, corresponds to the eighth note between 1/3 and 2/3. Following this convention, the six eighth notes in a bar would have the following metrical positions:

- First downbeat: 1/3(=1/6)
- Upbeat: 2/6
- Second downbeat: 2/3(=3/6)
- Upbeat: 4/6
- Third downbeat: 3/3(=5/6)
- Upbeat: 6/6.

Another aspect related to the transcription of fiddle music concerns the detection of bow strokes—tying succession of notes under single bowing gestures—and of bowing shifts—i.e., the transition between successive bow strokes. We discovered that the note transcription of the drone line implicitly indicates bowing shifts: since a drone line is made of a single pitch, only one note is generally played within a single bow stroke and each bowing shift implies the iteration of a new note of the same pitch. The bow strokes combined with the melodic line provide a dataset that can be further used for designing a machine learning model to detect bow strokes in the absence of a drone line.

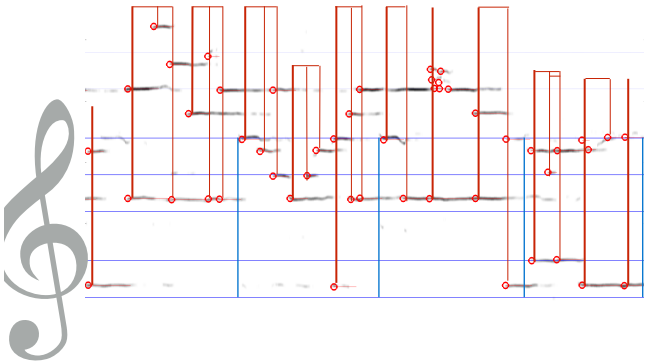

Figure 5: Transcription of an excerpt of a Springar tune called *Vrengja*. The pitchogram is shown with grey lines. A coarse approximation of the pitch levels is shown by the superposed score staff in blue. Detected note onsets are shown with red circles. An attempt of metrical analysis is shown with red stems and beams, as well as bar lines in blue.

5.3 Computational beat tracking of Hardanger fiddle music

A computational model predicting the localization of played beat onsets — and therefore reconstructing the underlying metrical grid — from the note annotations is currently being conceived and experimented. The approach infers beats through a single scanning of the note sequence from start to end. Structures are progressively constructed note after note, based on the local context. The decisions are based on a set of explicit rules, inspired by musicological and cognitive principles, a model that was progressively established while observing the behavior of the system in practical examples. One underlying objective is to establish a rule-based system that is as simple as possible, although at the same time should be able to offer results congruent to the experts' annotations. Although the model progressively augments in complexity due to the complexity of the musical problem itself, the hypothesis is that all specific rules would ultimately be replaced by more general mechanisms. This methodology was previously experimented for other musical problems such as the discovery of motivic patterns [9]. One core advantage of this approach, as we will see below, is that it enables to study the complex interdependencies between various musical dimensions.

When several notes are played nearly at the same time, they are considered as one single cluster with a cluster onset to which is assigned a metrical position. It often happens also that clusters form a quick succession, such as a trill. One of these clusters is considered to corresponds to the onset of the beat (or any other metrical position between two beats), and its metrical position is notated with an additional offset position set to zero. For instance, for the cluster right on the onset of the first beat of a bar, the metrical position is (1/3, 0). If one of the other clusters is just before that beat onset, it has an offset of -1, hence notated (1/3, -1). Same for a cluster just after that metrical onset, with metrical position (1/3, +1), then the subsequent cluster would have the position (1/3, +2), etc.

As mentioned in the previous paragraph, the proposed computational analysis method is based on a single scanning of the notes from start to end. The assignment of notes to clusters and trills can be formalized within that chronological paradigm using temporal thresholds. At the lowest level, a threshold of 20 ms determines whether successive notes are clustered (if the inter-onset-interval (IOI) between successive notes is below the threshold) or not (else). When the IOI falls near the threshold value, the ambiguity leads to a somewhat arbitrary decision but that is not an issue as this distinction does not play a core role in the metrical analysis. The decision whether successive clusters are part of a same trill or instead of successive metrical onsets is based on a temporal threshold, this time of 70 ms, further refined by additional heuristics, as developed below.

Deciding about the metrical position of successive notes (or clusters) cannot be simply based on IOI duration between successive notes, but requires taking into consideration the underlying musical context, in particular the metrical structure already constructed but also aspects related to other musical dimensions. For each new note (or cluster), the metrical structure already inferred might indicate a clear metrical position, or instead there can be several alternative solutions. In such case, all solutions are tried in parallel and progressively extended further, note after note. Each alternative analysis is associated with a score indicating the degree of congruency with respect to stylistic constraints, such as the range of tempi, the regularity of the beats, etc. At the very beginning of a tune, due to the absence of a metrical context, many alternatives might arise, each attempting to consider one of the first played note as anchor of one possible metrical position, but quickly after a few more notes, many of those alternatives are deemed impossible.

Occasionally, a rhythmical regularity can be detected and tracked, and can guide the metrical analysis. For instance, if a given succession of beats is characterized by a very regular binary or ternary rhythm, the simple continuation of this regular rhythm will offer a clear indication of the localization of the beat anchors. In the case of a sequence with very regular successive IOIs, it might be difficult to infer the localization of the beat onsets. It can appear sometimes that the melodic content will indicate the rhythmical subdivision, such as in the example in Figure 5, where we observe the successive repetition of a three-note melodic pattern. We have integrated the capability of detecting such patterns, thus indicating the beat onsets for the beginning of the sequence. Another possible strategy to establish the beat in a regular rhythmical sequence is based on change of chords melodic-harmonic content. Then by simply continuing further this rhythmical and metrical regularity, the beat onsets can be inferred directly even if the melodic pattern is not expressed anymore. The absence of salient characteristic indicating the beat onsets may lead to a somewhat ambiguous rhythmical sequence, although other factors, such as, here again, change of chords, can offer additional cue for the beat onsets. This suggests that robust beat tracking in challenging music cases needs to be integrated within a more comprehensive music analysis system.

6 OTHER ASPECTS RELATED TO MUSIC TRANSCRIPTION AND ANALYSIS

In Hardanger fiddle music, there often is a superposition of two monodic voices, one upper and one lower. One voice plays the

role of the melody while the other is a drone, generally based on successive repetition of a same note emphasizing an important degree in the mode. At times only the melody voice is played on one string, meaning that the drone line stops from time to time, resulting in a solo melody. One challenge is that the melodic and drone voices may sometimes cross, and this needs to be automatically detected. The melodic and drone lines may sometimes join into a unison played on two different strings, and this needs to be represented as such. Since the drone line sometimes stops, a decision needs to be made whether the drone stops in a note specific to the drone line, or instead on a melody note. Sometimes the upper and lower voices change function, the melody switching from one to the other and the drone reversely. We are working on a model that systematizes and automatizes this segregation between the two monodic lines. This is based on a combination of heuristics, focusing on aspects such as: the presence of ornamented notes (which by definition belong to generally emphasize note of the melodic line); whether additional notes are played on one of the voices—which would be the melodic voice—while the other keep a steady note—the drone. One example of crossing is when a given pitch height is repeated while at the same time a note below that pitch is followed by a note above that pitch: the repeated pitch forms the drone while the pitch jump is part of the melody.

A more detailed modal analysis of the music is needed, in order to distinguish, within the melody, the notes that constitute the structural pitches and form the successive chords , and those that are ornamental and thus of lesser importance. We are considering this progressive reduction on multiple hierarchical levels, in order to reveal core articulations in the modal discourse.

Music is often characterized by the repetition of sequences of notes, at multiple levels, starting of very short cells of a few notes, to longer phrases and finally complete sections. The repetition can be identical or with various types of transformations. Automatically detecting those repetitions is a hard problem, due to the combinatorial explosion of possible solutions. One challenge is to offer a clear and synthetic depiction of the motivic and thematic material, while at the same time aiming at revealing as much information as possible related to the richness of the melodic development [9]. One difficulty in the case of Hardanger fiddle music is the presence of ornamentation, which are not necessarily repeated the same way for each repetition of a given phrase. Hence to be able to detect the repetition, the ornamentation needs to be reduced automatically.

Combining these different analyses, while also considering aspects related to phrasing and silence leads to a multi-level segmentation of music [10], which will be further investigated. Other music analysis aspects are also considered, such as the musicological characterization of styles, or the computational prediction of musical/perceptual features from audio or transcription, such as rhythmic entrainment.

Recent studies have shown that beat-level variations in the asymmetrical timing patterns of springar performances seem to be related to “melodic-rhythmic" structures, in the sense that particular motivic segments are associated with particular timing profiles, suggesting that structural and other expressive features influence beat duration patterns [7]. Inspired by these findings, additional graphical user interfaces have been developed in the Max/MSP/Jitter platforms for data visualization and exploration of performance patterns related to the duration of the successive beats, offering structural and multi-dimensional perspectives on the complex rhythmical structuring of springar performances [14]. Interaction between the music, the fiddler’s gestures and the dancer’s movements have also been investigated [6].

Each tune in this catalogue is then transcribed—following the steps described above: note detection, beat and voice tracking, pitch spelling—as well as subject to modal and motivic analysis and segmentation. We are experimenting on this digital catalogue the design of additional tools to make the analyses easily accessible to musicians, musicologists and the general public, through interactive visualization of the music content of each tune as well as a visual distribution of the whole catalogue based on their intrinsic content, guided for instance by stylistic clustering, allowing to navigate into the catalogue, to aid listeners to better understand and appreciate the richness of these catalogues.

7 AVAILABILITY OF THE DATASET AND TOOLS

The following tools are dataset are available for free from the MIRAGE website[4]:

- AudioClassify, generating a Sound Classification File (SCF) from a tape recording, featuring scores for the classes Music, Speech, Singing, Bowed and Violin (cf. Section 3.2). This software needs to be run on either Mac or iOS devices.
- AudioSegmentor (cf. Section 3.2), running on Mac or PC, requiring Matlab Runtime component, which is included and installed for free. It loads the tape recordings altogether with the SCF file and generates a presegmentation that can be further edited.
- Annotemus (cf. Section 4), also running on Mac or PC with Matlab Runtime component. It enables to load and display audio recordings, load corresponding note annotations, or manually add those annotations. Further versions will include automated note annotation as well as beat annotation and music analyses.
- Individual tunes from the subset of the Norwegian Folk Music Archive indicated in Section 2.2 will be progressively added, with metadata and annotations.

ACKNOWLEDGMENTS

This work is supported by the Research Council of Norway through its Centers of Excellence scheme, project number 262762 and the MIRAGE project, grant number 287152. The training of the machine learning model was performed on resources provided by Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway. We thank Unni Løvlid from the Norwegian Academy of Music for advising for this project. Tape segmentation has been carried out by Rasmus Kjorstad and Åsmund Solberg.

REFERENCES

[1] Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. 2019. Automatic Music Transcription: An Overview. *IEEE Signal Processing Magazine* 36, 1 (2019), 20–30.

[4] https://www.uio.no/ritmo/english/projects/mirage/index.html

[2] Jan Peter Blom, Sven Nyhus, and Reidar Sevåg. 1979. *Hardingfeleslåttar Springar i 3/4 takt = Slåttar for the harding fiddle Springar in 3/4 time.* Norsk folkemusikk : Norwegian folk music, Serie 1, Vol. 6. Universitetsforlaget, Oslo, Norway.

[3] Reinier de Valk, Anja Volk, Andre Holzapfel, Aggelos Pikrakis, Nadine Kroher, and Joren Six. 2017. MIRchiving: Challenges and opportunities of connecting MIR research and digital music archives. In *Proceedings of the 4rd International workshop on Digital Libraries for Musicology.*

[4] Anders Elowsson. 2020. Polyphonic pitch tracking with deep layered learning. *Journal of the Acoustical Society of America* 148, 1 (2020), 446–468.

[5] Anders Elowsson and Olivier Lartillot. 2021. A Hardanger Fiddle Dataset with Performances Spanning Emotional Expressions and Annotations Aligned using Image Registration. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference.*

[6] Mari Romarheim Haugen. 2021. Investigating Music-Dance Relationships. A Case Study of Norwegian Telespringar. *Journal of Music Theory* 65, 1 (2021), 17–38.

[7] Mats S Johansson. 2017. Empirical Research on Asymmetrical Rhythms in Scandinavian Folk Music: A Critical Review. *Studia Musicologica Norvegica* 43, 1 (2017), 58–89.

[8] Mats S. Johansson. 2022. Timing-sound interactions as groove-forming elements in traditional Scandinavian fiddle music. *Puls* 7 (2022).

[9] Olivier Lartillot. 2005. Multi-Dimensional Motivic Pattern Extraction Founded on Adaptive Redundancy Filtering. *Journal of New Music Research* 34, 4 (2005), 375–393.

[10] Olivier Lartillot and Mondher Ayari. 2011. Cultural impact in listeners? structural understanding of a Tunisian traditional modal improvisation, studied with the help of computational models. *Journal of interdisciplinary music studies* 5, 1 (2011), 85–100.

[11] Matija Marolt, Ciril Bohak, Alenka Kavčič, and Matevž Pesek. 2019. Automatic Segmentation of Ethnomusicological Field Recordings. *Applied Sciences* 9, 3 (2019), 439.

[12] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes. 1999. Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images. *IEEE Transactions on Medical Imaging* 18, 8 (1999), 712–721.

[13] J.-P. Thirion. 1998. Image matching as a diffusion process: An analogy with Maxwell's demons. *Medical Image Analysis* 2, 3 (1998), 243–260.

[14] Aleksander Tidemann, Olivier Lartillot, and Mats S. Johansson. 2021. Towards New Analysis and Visualization Software for Studying Performance Patterns in Hardanger Fiddle Music. In *Proceedings of the Nordic Sound and Music Computing Conference.*