A HARDANGER FIDDLE DATASET WITH PERFORMANCES SPANNING EMOTIONAL EXPRESSIONS AND ANNOTATIONS ALIGNED USING IMAGE REGISTRATION

Anders Elowsson **Olivier Lartillot**
RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion; University of Oslo
anderselowsson@gmail.com olivier.lartillot@imv.uio.no

ABSTRACT

This paper presents a Hardanger fiddle dataset “HF1” with polyphonic performances spanning five different emotional expressions: normal, angry, sad, happy, and tender. The performances thus cover the four quadrants of the activity/valence-space. The onsets and offsets, together with an associated pitch, were human-annotated for each note in each performance by the fiddle players themselves. First, they annotated the normal version. These annotations were then transferred to the expressive performances using music alignment and finally human-verified. Two separate music alignment methods based on image registration were developed for this purpose; a B-spline implementation that produces a continuous temporal transformation curve and a Demons algorithm that produces displacement matrices for time and pitch that also account for local timing variations across the pitch range. Both methods start from an “Onsetgram” of onset salience across pitch and time and perform the alignment task accurately. Various settings of the Demons algorithm were further evaluated in an ablation study. The final dataset is around 43 minutes long and consists of 19 734 notes of Hardanger fiddle music, recorded in stereo. The dataset and source code are available online. The dataset will be used in MIR research for tasks involving polyphonic transcription, score alignment, beat tracking, downbeat tracking, tempo estimation, and classification of emotional expressions.

1. INTRODUCTION

1.1 Hardanger Fiddle Music

The Hardanger fiddle is a traditional stringed solo instrument played in the southern parts of Norway. It features resonance strings producing a characteristic resonating sound. The flat fingerboard and bridge enable the performer to play several strings simultaneously and the polyphony level of the music is generally 2. Fast trills are frequently used as ornaments. Lack of annotated audio excerpts makes data-driven research on Hardanger fiddle music hard and this study is an attempt to remedy the situation. Our vision is to create a dataset with annotated pitched onsets and offsets so that accurate polyphonic transcription systems can be trained in future studies, enabling researchers to transcribe vast existing libraries of historical audio recordings.

1.2 Transcription Datasets in MIR

Researchers have used many different techniques to create annotated datasets for polyphonic transcription in the past. One method is to record individual voices in isolation to facilitate easier annotation. Examples include the four-voiced *Bach10* dataset [1], the *TRIOS* dataset [2] consisting of musical trios, a five-voiced woodwind recording [3], the audio-visual URMP dataset [4], and the *MedleyDB* multitracks dataset [5]. For polyphonic instruments, the annotation of many simultaneous notes can be cumbersome and time-consuming. Another method for those kinds of instruments has therefore been to generate the sounds and annotations directly from MIDI. The technique has been used for piano datasets [6-8], but has also been applied across the full range of the general MIDI instrument specification [9]. To increase the variability and the size of the dataset, researchers can use data augmentation, varying tempo, pitch, dynamics, and timbre during synthetization [9].

Although the MIDI generation strategy is appealing because of its efficiency, synthesized MIDI often lacks the full range of variation and complexities found in real performances. Researchers can in this case instead create datasets by synchronizing sheet music with an associated recording. This approach was adopted by Thickstun, et al. [10] who used dynamic time warping (DTW) applied to log-frequency spectrograms focused on lower frequencies.

1.3 Mood Datasets in MIR

Datasets spanning different moods/emotions are developed to enable researchers to train and test music emotion recognition (MER) systems. Many MER datasets use the valence-arousal model [11], with the valence and arousal variables annotated by human listeners. Examples include the *MoodSwings* [12], *Emotion in Music* [13], *AMG1608* [14], *DEAM* [15], and *PMEmo* [16] datasets.

For a few datasets, performers have been asked to play the *same* piece of music with *different* emotional expressions. Li, et al. [17] asked violinists to perform classical compositions according to different expressive musical terms (e.g., *tranquillo*) and used the resulting dataset for modeling. Gabrielsson and Juslin [18] asked performers to play with the emotional expressions “happy”, “sad”,

Attribution: A. Elowsson and O. Lartillot, “A Hardanger Fiddle Dataset with Performances Spanning Emotional Expressions and Annotations Aligned using Image Registration”, in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, Online, 2021.

"angry", "fearful", "tender", "solemn", and "no expression", analyzing the recordings both quantitatively and through listening tests. Performers control the musical expression by varying, e.g., phrasing, tempo, timing, articulation, and dynamics [19-24] and the perceptual aspect of such features has also been modeled extensively [25-27]. Note that these types of features are among those varied for data augmentation applied to MIDI (or audio files), but when they are introduced by real musicians, they will be richer in scope and better capture the variability that can be expected in other real performances. It is therefore appealing to create a dataset where each song is performed with several musical expressions, using music alignment to transfer annotations between the different performances. Not only will this bootstrap the annotation effort while retaining variation in the annotated notes, it will also introduce a new dataset for emotional expression, where researchers can, in extension to analyzing the audio files, utilize the annotations as a symbolic representation for MER. This strategy is therefore explored in this study.

1.4 Score Alignment

The task of aligning a musical score with an associated audio file has been fairly widely studied, with researchers often opting for various flavors of DTW. Implementations differ regarding how they compute a similarity metric/feature space for alignment. Researchers can either synthesize or add harmonics to the score [10, 28-30], convert both score and audio to a chroma-space [31], or alternatively learn the feature space for alignment [32-35], casting the task as an optimization problem.

The aforementioned strategies are aligning across full note lengths, but it is mainly the onsets that provide information about timing [36]. It has therefore been suggested that they can be improved by detecting onsets in the audio [30]. One strategy in this direction is to apply DTW to a half-wave rectified spectral flux (SF) [36]. Ewert, et al. [37] instead start from a chroma before computing the flux. Kwon, et al. [38] used a polyphonic pitch tracker to compute the feature space and found that the best results were achieved when including pitched onsets across the full 88-note range. This strategy concerning the feature space is the closest to our implementation, but we decided to forego DTW. Our motivation for, and implementation of, image registration techniques for music alignment are described in Section 3.

2. OVERVIEW AND MOTIVATION

Our primary objective with this study was to create a dataset of Hardanger fiddle music with annotated onsets and offsets. In particular, our focus was on the annotated onsets. Annotating Hardanger fiddle music is non-trivial. It is polyphonic and contains ornaments with very fast tone sequences. In our preliminary studies, we learned that it is rather time-consuming for Hardanger fiddle musicians to produce annotations for tunes that they are unfamiliar with, and accuracy may sometimes be lacking. Furthermore, our overarching project also strives to collect additional data on expressive Hardanger fiddle performances. These circumstances led to the following design:

1. Hardanger fiddle performers are tasked to record five versions of songs they are familiar with, using the expressions: *normal, sad, angry, happy,* and *tender*.
2. They annotate notes in the *normal* recording from scratch, using computer assistance tools as aid.
3. The *normal* recording is aligned with the expressive recordings using music alignment, so that the *normal* annotations can be automatically transferred to them.
4. Performers go through the aligned annotations and make adjustments to ensure that they are correct.

The strategy gives us a few advantages:

- *Does not introduce bias concerning timing*. Since the *normal* recording is annotated from scratch, and the score alignment only used for aligning the two audio recordings, we do not impose priors regarding the exact location of, e.g., onsets in the music, which would have been the case if an algorithm produces the initial annotations.
- *Ensures that annotators annotate songs they are familiar with.* It is easier to be accurate and efficient when annotating a song that you are familiar with, and note sheets are not exhaustive since they do not cover the rich ornamentation in Hardanger fiddle music.
- *Provides five times the training and testing data for polyphonic transcription.* With real performances of bowed instruments, the sound characteristics will vary each time a phrase is played. Thus, repeated sequences, particularly of ornaments, still provide training and testing data with high "entropy".
- *Creates a dataset that can be used for additional tasks in future studies.* Our experimental design provides us with both audio and symbolic data of performances with varying emotional expressions. This data can be used to study how mood is expressed on the Hardanger fiddle and to develop music alignment systems.
- *Enables us to scale future annotation tasks within the same framework.* The method will connect each note in the expressive performances with the notes in the normal performance. Thus, if we assign higher-level features to these notes, such as their metrical position, we can automatically transfer that information to the expressive performances.

3. MUSIC ALIGNMENT ALGORITHMS

Tempo variations in music are often observed and modeled as gradual changes developing over several successive notes. Friberg [39] fitted ”phrase arches” to piano performances, with *accelerando* in the start and ritardando in the end of the phrases. Other researchers fit their observations using spline-shaped profiles [40] or fit the final ritardando using a quadratic polynomial [41].

The DTW algorithm is "local" in scope and will not model differences in tempo and gradual tempo variations observed across longer sections. This means that it can, e.g., fail to accurately stretch matched notes of different lengths or, when the feature space is focused on onsets, fail to produce convincing tempo curves for sections where the feature space is empty. The resulting warping path can therefore become rather irregular and is also discrete, not fitting to a finer scale than the time frame hop length. Various remedies have been proposed to alleviate these issues, for example introducing special silence frames to "stretch out" pauses between notes [28] or trying to smooth the warping path in post-processing [29]. This study explores if techniques developed for image registration can be useful as an alternative approach. Through a free form deformation with a B-spline grid [42] (Section 3.2), we optimize across multiple frames, utilizing a smoothness penalty to constrain neighboring grid points from moving independently while achieving sub-frame resolution. By adopting the Demons algorithm to music alignment (Section 3.3), we instead also test a 2-dimensional alignment approach, where individual pitch bins are allowed to diverge somewhat from the warping path in order to account for natural variations in timing between concurrent notes.

3.1 Onsetgram and Preprocessing

The temporal alignment is performed on a 2-dimensional "*Onsetgram*," consisting of onset activations distributed across pitch and time. The onset activations are first computed using the polyphonic transcription system developed by Elowsson [9], trained on a wide variety of music. In that system, an initial network detects framewise f_0 activations, 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 smoothed thresholded onset activation function was used (cf. [Eqs. A8-A11, 9]). The onset activations were inserted at the corresponding pitch bin and time frame of the Onsetgram, which had a pitch resolution of 1 cent/bin. A Hann window of width 151 bins (cents) was then used to smooth the Onsetgram across pitch. Figure 1 shows the smoothed Onsetgram in green overlaying the f_0 activations in blue.

The pitch range of the Onsetgram was set to 2 semitones below the lowest annotated pitch to 2 semitones above the highest annotated pitch. The pitch resolution was also scaled down to 4 bins/semitone. To speed up processing, the hop size was set to 23.2 ms by keeping only every fourth time frame of the original Onsetgram.

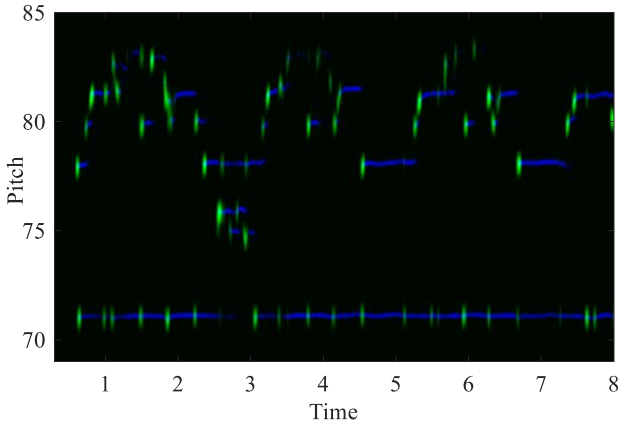

Figure 1. The O*nsetgram* used for music alignment in green overlaying f_0 activations in blue across which the onset activations were computed. The excerpt is from the song Haslebuskane, also featured in Figures 2 and 3.

Before applying the image registration algorithms, a start- and endpoint was computed for both audio files by finding the first and last time frame with a signal level within 10 dB of the average signal level of the audio file, as described by Elowsson and Friberg [43]. 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 using the same transformation.

3.2 B-spline Algorithm

The B-spline music alignment implementation uses low-level MATLAB functions for B-spline image registration from Kroon [44, 45]. The particular non-rigid B-spline alignment method was first introduced by Rueckert, et al. [42]. It is a free-form deformation with a B-spline grid, typically performed at multiple image scales (pyramid levels). For a precise mathematical formalization of the process, cf. [41, p. 64-65]. A multi-scale approach can be beneficial for two reasons – iterations performed at a coarser scale will converge fast, and the risk of reaching local minima is reduced. Since music may contain closely spaced repetitions, it seems reasonable to first align the coarser overall structure, ensuring that repetitions are not misaligned, and to then adjust notes at finer scales.

The temporal grid spacing for the first iteration was 256 frames (5.9 seconds), and at each subsequent iteration, this spacing was halved, ending with a grid spacing of 4 frames (93 ms) at the finest level. To avoid a too local scope with abrupt changes in the tempo curve at the finest level, the smoothness penalty of the B-spline implementation was used [44, 45]. This smoothness penalty constrains neighboring grid points from moving independently, simulating the bending energy of a thin plate of metal [42, 46]. We set the penalty to 0.3 at the finest pyramid level, halving it at each level such that it was 0.005 at the coarsest scale.

The pitch spacing was set such that the whole pitch dimension of the image was contained between two grid points at all pyramid levels, and the pitch dimension of these grid points reset after optimizing at each level.

After fitting the normal Onsetgram to the Onsetgram of the emotional expression, the resulting forward transformation field was applied to the annotations, changing their

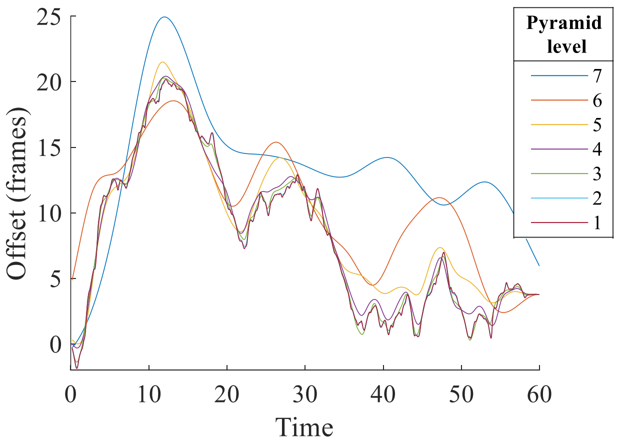

Figure 2. The forward transformation field at each pyramid level for aligning the *sad* and *normal* recordings of the tune Haslebuskane.

timing using linear interpolation. The aligned annotations were finally "tuned" as described in Section 3.4. Figure 2 shows the transformation field for all seven pyramid levels when aligning the *sad* and *normal* recording of the tune Haslebuskane.

3.3 Accelerated "Demons" Algorithm

The diffusion model known as the *Demons* algorithm for non-rigid image registration was introduced by Thirion [47]. It uses the gradient $\vec{\nabla} f$ from the fixed image f to compute a "demons" force for deforming a moving image m. Wang, et al. [48] modified the algorithm by also including the gradient of the moving image $\vec{\nabla} m$, using bi-directional forces,

$$\vec{u} = (m-f)\times\left(\frac{\vec{\nabla} f}{\left|\vec{\nabla} f\right|^2+\alpha^2(f-m)^2}+\frac{\vec{\nabla} m}{\left|\vec{\nabla} m\right|^2+\alpha^2(f-m)^2}\right). \qquad (1)$$

The normalization factor α introduced by Cachier, et al. [49] allows the force strength to be adjusted adaptively in each iteration. The displacement field $\vec{u}$ is computed for both time ($\vec{u}_x$) and pitch ($\vec{u}_y$) deformations in each iteration and added to the corresponding overall displacement fields T_x (time) and T_y (pitch). We used this "accelerated Demons" algorithm, operating over 7 pyramid-levels with 70 iterations at each level, setting α to 0.4 as proposed by Wang, et al. [48], using the basic demon example code from Kroon and Slump [50] as a starting point but adapting the registration to the music alignment task. The Onsetgram of the recording with an emotional expression was used as the moving image and the Onsetgram of the normal recording used as the fixed image. The computed displacement field could then be used as a backward transformation to transfer the annotations to the recordings with emotional expressions.

In its original formulation, the computed displacements $\vec{u}_x$ and $\vec{u}_y$ for each iteration is smoothed before being added to the overall displacement fields T_x and T_y. We instead opted to smooth T_x and T_y directly in each iteration. To understand why this improves performance, recall that the Onsetgram is sparse and that we must be able to accurately move annotations between locations in the moving and fixed image that contain no salience information (e.g., offsets). By applying the smoothing operator directly to T_x and T_y, we iteratively "saturate" the displacement field with deformations also at locations where no gradients can be found in the Onsetgrams. This process also helps us smooth out irregular displacements resulting from erroneous transcriptions. The smoothing was done using Hann windows of length 33 across time and length 3 across pitch for T_x and length 17 and 3 for T_y. The reader is further referred to Cachier, et al. [49] for a discussion concerning the benefits of smoothing operations applied at various stages of the process.

Restrictions were set on T_x and T_y to ensure that the deformations were not bigger than desirable from a music-theoretical standpoint. For T_x, during each iteration before smoothing, we thresholded the displacement at each bin to not diverge more than 100 ms from the average displacement in each time frame. This means that annotations at different pitches can be moved freely but not diverge relative to each other too much. Thus, an annotation of a bass note and a note in the treble where the bass note is played slightly before the treble note in the fixed image, but where circumstances are reversed in the moving image, can be transferred receiving correct timing, but never to such an extent that the interpretation of the score would be vastly different (>100 ms). For T_y, a fixed threshold of 70 cents was instead used, such that the pitch could not be displaced more than this.

The displacements fields (backward transformations) were applied to the annotations, changing their timing using linear interpolation. Since the incorporation of a threshold on T_x could hinder the algorithm from displacing time globally, the mean displacement for $\vec{u}_x$ across all pitch bins is also added to T_x before thresholding and smoothing. Furthermore, since the Onsetgram only activates at onsets, T_y may not be particularly suitable for tuning the annotations. As a default, the post-processing step for tuning (Section 3.4) was instead applied. However, applying T_y directly for tuning was tested in the ablation

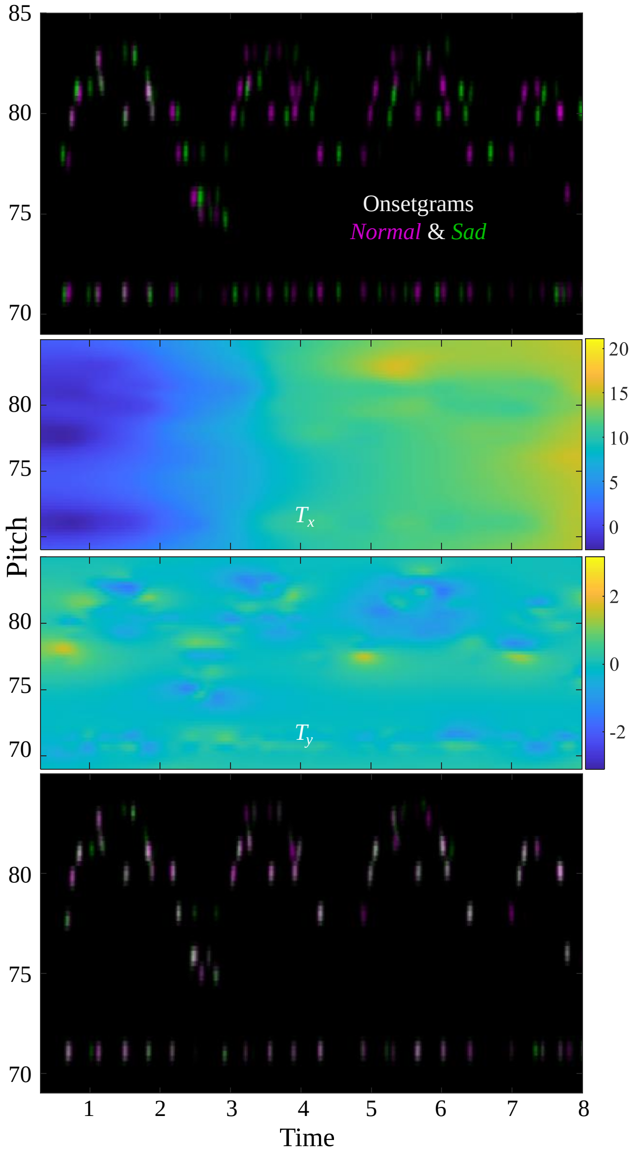

Figure 3. The Onsetgrams of both the *normal* and *sad* recordings of the tune Haslebuskane (pane 1), the backward transformation (displacement) fields T_x and T_y (panes 2 and 3), and the aligned Onsetgrams (pane 4).

study in Section 5.2. Figure 3 shows the Onsetgrams of both the *normal* and *sad* recordings of the tune Haslebuskane (pane 1), the displacement fields T_x and T_y (panes 2 and 3), and the aligned Onsetgrams (pane 4).

3.4 Tuning

A method for adjusting the pitch of each note was added as a post-processing step, motivated by the fact that the fiddle does not have fixed frets and the pitch of individual notes can vary relatively much. For each annotation to be tuned, a rectangular area was first extracted from the f_0 activations (blue in Figure 1), bounded by the onset and offset and extending 100 cents in both pitch directions from the annotated pitch. The average across time was computed and the resulting pitch vector smoothed with a Hann filter 41 cents wide. Only smoothed parts computed without zero-padded edges were kept, making the length 80 cents in both pitch directions. Peaks were detected and weighted based on how close they were to the annotated pitch as well as their pitch salience magnitude, opting to select the peak with the highest computed weight, and moving the annotation to its pitch.

Performances may drift “locally” in pitch through intonation on the fingerboard, such that the pitch of notes in short phrases with no open strings all are a bit higher or lower than in another recording of the same song. The tuning algorithm adapts to this by not allowing one note to be changed more than 45 cents in relation to the weighted average tuning change of other notes close in time and pitch. Due to space constraints, the reader is referred to the MATLAB implementation and its corresponding help text for precise details on all settings for the tuning algorithm.

4. DATASET

4.1 Recording and Annotation

The recordings were done by two Hardanger fiddle musicians, Henrik Nordtun Gjertsen (HNG) and Astrid Garmo (AG), who were students at the Norwegian Academy of Music. They recorded well-known Hardanger fiddle tunes in a relatively dry room in stereo using a Zoom H6 recorder.

The annotations were done by the same musicians using the software Annotemus[1] developed in MATLAB. Annotemus has a graphical user interface and provides functionality for creating annotations on top of a graphical representation of the audio file. We used the f_0 activations shown in blue in Figure 1 for this purpose. The aligned annotations were all initially created using the B-spline method which was being developed in conjunction with the annotation process.

The performers could use various key commands as an aid during annotation. This includes audio playback of the current window, playback between the start and end of one or several selected notes, playback that starts prior to a selected annotated note and ends at the annotated onset position, playback with a click at each annotated onset position, and playback with a synthesized version of the annotated score played in one of the stereo channels. The performers were instructed to first try the playback that ends at the annotated onset position for locating the exact onset times for the *normal* recording and the click and synthesized functionality for verifying annotations, but were free to use whichever method they felt most comfortable with.

All playback functionality is offered with the option of slowing it down to an arbitrary speed selected by the annotator. Since Hardanger fiddle music contains frequent sequences of very fast note successions, the slowdown functionality was used extensively during the annotation process. The onset timing evaluation condition for polyphonic transcription is usually set to 50 ms. This means that we can only allow a very narrow margin of error for the annotations to ensure that they can be reliably used for evaluation. We encouraged performers to be very careful regarding onsets, and try to keep errors within 20 ms. Listeners notice time-displacements of just 10 ms on average [51], but since fiddle music has rather undefined transients at onsets, a narrower margin than 20 ms is very hard to achieve. For both annotators, their first annotations were rejected, and they were encouraged to improve the quality regarding aspects that did not meet our high standards.

4.2 Dataset Overview

The final dataset consists of 19 734 annotated notes across 40 stereo recordings of 8 tunes. The audio recordings and annotations are available online,[2] as well as MATLAB source code.[3] The dataset is summarized in Table 1.

Title	Notes	Length	ID
Haslebuskane	2 828	4:35	HNG
Havbrusen	4 114	8:50	HNG
Ivar Jorde	1 665	3:52	AG
Låtten som bed om noko	1 819	4:51	AG
Signe Uladalen	2 177	4:30	AG
Silkjegulen	2 906	5:38	HNG
Valdresspringar	1 692	3:49	AG
Vossarull	2 533	6:34	HNG
Total	**19 734**	**42:38**	

Table 1. The eight tunes of the dataset, each performed with five different emotional expressions. The number of notes and the length of the recordings are computed as the total across the five variations. The ID identifies the musician. The last row provides totals across the dataset.

5. MUSIC ALIGNMENT EVALUATION

5.1 Main Results

The performance of the two methods was evaluated by matching onsets aligned from the *normal* version with the human-verified onset of the *expressive* version and measuring their distance. The two aligned recordings frequently vary, e.g., in ornaments, which means that many notes will not have a counterpart in the other recording. To account for this, we used weighted bipartite matching to first

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

[2] https://www.uio.no/ritmo/english/projects/mirage/databases/

[3] https://github.com/aelowsson/music-alignment

connect onsets of the two recordings, where the weight for how well a pair matches falls using a half Hann window up to a distance of 5 seconds and 70 cents respectively. Regular unweighted bipartite matching is not ideal in this circumstance since it can create incorrectly matched pairs containing ornaments with no counterparts, whenever two such ornaments, one in each recording, are within 5 seconds of a real correct pair of onsets with a similar pitch. The F-measure $\mathcal{F}$ was measured for the matched onset pairs only, leaving out the around 3 % of onsets with no counterpart that were unmatched.

Table 2 shows the results, with the F-measure for onsets within 80 ms ($\mathcal{F}_{80}$) highlighted in bold. We note that the Demons algorithm was more accurate even though the B-spline method was used as a starting point for the aligned expressive performances. Since this algorithm is also faster (the full dataset aligned in 2.5 minutes on an i7-6700K processor), it was our focus in the ablation study.

	$\mathcal{F}_{50}$	$\mathcal{F}_{80}$	$\mathcal{F}_{150}$	$\mathcal{F}_{300}$	*Avg*
B-spline	91.1	**95.9**	98.2	99.2	28.9 ms
Demons	95.4	**98.3**	99.1	99.5	23.0 ms

Table 2. F-measures at different distance metrics as well as the average distance (*Avg*) between matched onsets for the B-spline and Demons music alignment methods.

5.2 Ablation Study

Various settings of the Demons algorithm were tested in an ablation study:

- ***T_x* Thresh**: Instead of a 100 ms threshold we tested a strict zero threshold (*0*) or used no threshold (*None*).
- ***T_y***: Foregoing the use of T_y completely (*No T_y*), also skipping the tuning stage (*No TT*), applying T_y to the annotations instead of using the tuning algorithm (*Apply*), or using the default setting but without thresholding (*No Th*).
- ***T_x* Mean**: Testing to *not* add the mean displacement for $\vec{u}$ to T_x before thresholding and smoothing (*None*).
- **$\vec{u}_x$**: Smoothing $\vec{u}_x$ instead of smoothing T_x, tested across time (*$\vec{u}_x$ Ti*), time and pitch (*$\vec{u}_x$ TP*), or across pitch only (*$\vec{u}_x$ Pi*).
- ***T_x* Smooth**: Smoothing T_x across time with shorter or longer Hann windows (*15* or *45*).

Figure 4 shows the results of the ablation study as the difference in performance at $\mathcal{F}_{80}$. The 95 % confidence intervals (CIs) illustrated with black bars were derived from the difference in $\mathcal{F}_{80}$ for individual tunes between the default setting and the tested setting. This difference was sampled with replacement from the tunes 8 × 4 = 32 times to compute a single overall outcome, and the procedure repeated 10^6 times to compute a distribution of possible outcomes, from which the 5th and 95th percentile could be extracted.

6. CONCLUSIONS

We have created an annotated Hardanger fiddle dataset with performances spanning five emotional expressions. The process of creating accurate note annotations for real polyphonic instrument recordings can be cumbersome, and we hope that the developed techniques and source code can be useful to other researchers in the field.

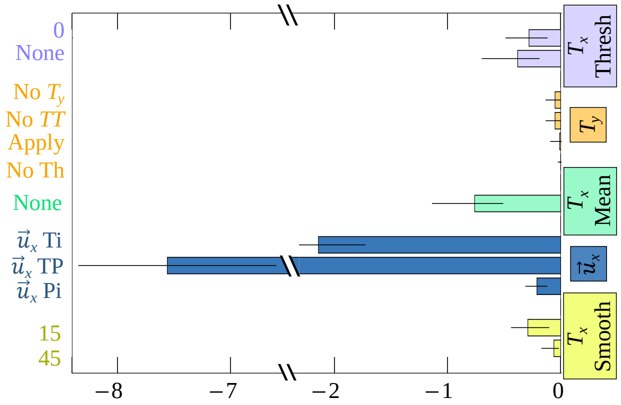

Figure 4. The results of the ablation study for the Demons method, showing the change in F-measure relative to the default setting. Black bars indicate 95 % CIs. Note that the x-axis has been spliced to accommodate the lower result for the $\vec{u}_x$ *TP* setting.

Two music alignment algorithms based on image registration were created and analyzed. The Demons algorithm is faster and easier to adapt to music and it also produces the best alignments. It can be noted that the alignment is evaluated using two separate annotations, so if a matched pair of notes have annotations that are 40 ms off each, they may just fail on the $\mathcal{F}_{80}$ evaluation metric even if the alignment is performed perfectly. Furthermore, ornaments with no counterpart (see Section 5.1) may still be erroneously matched if they are within 5 seconds of each other. Thus, even with a few missed notes on the $\mathcal{F}_{80}$ metric, we can still suspect that the alignment is very accurate overall. Informal closer analysis of the alignments also indicates that this is the case.

The ablation study indicates that the proposed default settings for the Demons algorithm are well-adjusted. We note that smoothing across T_x instead of $\vec{u}_x$ is an important ingredient for successful Demons music alignment. The 100 ms threshold for individual pitch bin displacements in T_x relative to the mean displacement is an important addition (T_x Thresh), and should be combined with adding the mean displacement for $\vec{u}$ to T_x before thresholding and smoothing (T_x Mean).

We intend to expand the annotations to also contain higher-level metrical information. Furthermore, we intend to develop models for polyphonic transcription and MER based on the dataset, something that we hope other research groups will do as well.

7. ACKNOWLEDGMENTS

We thank Astrid Garmo and Henrik Nordtun Gjertsen for their fiddle performances and annotations. This study was supported by the Research Council of Norway through its Centers of Excellence scheme, project number 262762, and the MIRAGE project, grant number 287152. The ablation study was performed on resources provided by UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway.

8. REFERENCES

[1] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE TASLP,* vol. 18, no. 8, pp. 2121-2133, 2010.

[2] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 888-891: IEEE.

[3] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of Multiple-F0 Estimation and Tracking Systems," in *ISMIR*, 2009, pp. 315-320.

[4] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia,* vol. 21, no. 2, pp. 522-535, 2018.

[5] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research," in *ISMIR*, 2014, vol. 14, pp. 155-160.

[6] G. E. Poliner and D. P. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. on Adv. in Signal Proc.,* vol. 2007, no. 1, 2006.

[7] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643-1654, 2010.

[8] C. Hawthorne *et al.*, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," *arXiv preprint arXiv:1810.12247,* 2018.

[9] A. Elowsson, "Polyphonic pitch tracking with deep layered learning," *Journal of the Acoustical Society of America,* vol. 148, no. 1, pp. 446-468, 2020.

[10] J. Thickstun, Z. Harchaoui, and S. Kakade, "Learning features of music from scratch," in *International Conference on Learning Representations (ICLR),* 2017.

[11] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology,* vol. 17, no. 3, p. 715, 2005.

[12] J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim, "A Comparative Study of Collaborative vs. Traditional Musical Mood Annotation," in *ISMIR*, 2011, vol. 104, pp. 549-554: Citeseer.

[13] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia,* 2013, pp. 1-6.

[14] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H. Chen, "The AMG1608 dataset for music emotion recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 693-697: IEEE.

[15] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PloS one,* vol. 12, no. 3, p. e0173392, 2017.

[16] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The pmemo dataset for music emotion recognition," in *Proceedings of the 2018 acm on international conference on multimedia retrieval*, 2018, pp. 135-142.

[17] P.-C. Li, L. Su, Y.-H. Yang, and A. W. Su, "Analysis of Expressive Musical Terms in Violin Using Score-Informed and Expression-Based Audio Features," in *ISMIR*, 2015, pp. 809-815.

[18] A. Gabrielsson and P. N. Juslin, "Emotional expression in music performance: Between the performer's intention and the listener's experience," *Psychology of music,* vol. 24, no. 1, pp. 68-91, 1996.

[19] A. Friberg, "A quantitative rule system for musical performance," KTH Royal Institute of Technology, 1995.

[20] A. Friberg, "Generative rules for music performance: A formal description of a rule system," *Computer Music Journal,* vol. 15, no. 2, pp. 56-71, 1991.

[21] R. Bresin and A. Friberg, "Emotional coloring of computer-controlled music performances," *Computer Music Journal,* vol. 24, no. 4, pp. 44-63, 2000.

[22] R. Bresin and A. Friberg, "Emotional expression in music performance: synthesis and decoding," *TMH-QPSR,* vol. 39, pp. 085-094, 1998.

[23] R. Bresin and G. Umberto Battel, "Articulation strategies in expressive piano performance analysis of legato, staccato, and repeated notes in performances of the andante movement of Mozart’s sonata in g major (k 545)," *Journal of New Music Research,* vol. 29, no. 3, pp. 211-224, 2000.

[24] A. Gabrielsson and P. N. Juslin, *Emotional expression in music*. Oxford University Press, 2003.

[25] A. Friberg, E. Schoonderwaldt, A. Hedblad, M. Fabiani, and A. Elowsson, "Using listener-based perceptual features as intermediate representations in music information retrieval," *Journal of the Acoustic Society of America (JASA),* vol. 136, no. 4, pp. 1951-1963, 2014.

[26] A. Elowsson and A. Friberg, "Predicting the perception of performed dynamics in music audio with ensemble learning," *Journal of the Acoustic Society of America (JASA),* vol. 141, no. 3, pp. 2224-2242, 2017.

[27] A. Elowsson, A. Friberg, G. Madison, and J. Paulin, "Modelling the Speed of Music using Features from

Harmonic/Percussive Separated Audio," presented at the ISMIR, 2013.

[28] N. Orio and D. Schwarz, "Alignment of monophonic and polyphonic music to a score," in *International Computer Music Conference (ICMC)*, 2001, pp. 1-1.

[29] R. J. Turetsky and D. P. Ellis, "Ground-truth transcriptions of real music from force-aligned midi syntheses," 2003.

[30] F. Soulez, X. Rodet, and D. Schwarz, "Improving polyphonic and poly-instrumental music to score alignment," 2003.

[31] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*, 2003, pp. 185-188: IEEE.

[32] R. Lajugie, P. Bojanowski, P. Cuvillier, S. Arlot, and F. Bach, "A weakly-supervised discriminative model for audio-to-score alignment," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2484-2488: IEEE.

[33] R. Lajugie, D. Garreau, F. R. Bach, and S. Arlot, "Metric Learning for Temporal Sequence Alignment," in *NIPS*, 2014.

[34] Ö. İzmirli and R. B. Dannenberg, "Understanding Features and Distance Functions for Music Sequence Alignment," in *ISMIR*, 2010, pp. 411-416: Citeseer.

[35] C. Joder, S. Essid, and G. Richard, "Learning optimal features for polyphonic audio-to-score alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2118-2128, 2013.

[36] S. Dixon and G. Widmer, "MATCH: A Music Alignment Tool Chest," in *ISMIR*, 2005, pp. 492-497.

[37] S. Ewert, M. Muller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1869-1872: IEEE.

[38] T. Kwon, D. Jeong, and J. Nam, "Audio-to-score alignment of piano music using RNN-based automatic music transcription," presented at the Sound and Music Computing conference (SMC), 2017.

[39] A. Friberg, "Matching the rule parameters of Phrase arch to performances of" Träumerei": A preliminary study," *STL-QPSR*, vol. 36, no. 2-3, pp. 063-070, 1995.

[40] J. Feldman, D. Epstein, and W. Richards, "Force dynamics of tempo change in music," *Music perception*, vol. 10, no. 2, pp. 185-203, 1992.

[41] A. Friberg and J. Sundberg, "Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1469-1484, 1999.

[42] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE transactions on medical imaging*, vol. 18, no. 8, pp. 712-721, 1999.

[43] A. Elowsson and A. Friberg, "Modeling Music Modality with a Key-Class Invariant Pitch Chroma CNN," in *20th International Society for Music Information Retrieval Conference ISMIR*, Delft, Netherlands, 2019, pp. 541-548.

[44] D.-J. Kroon, *Segmentation of the mandibular canal in cone-beam CT data*. Citeseer, 2011.

[45] D.-J. Kroon, "B-spline Grid, Image and Point based Registration," ed. MATLAB Central File Exchange, 2020.

[46] C. R. Meyer *et al.*, "Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin-plate spline warped geometric deformations," *Medical image analysis*, vol. 1, no. 3, pp. 195-206, 1997.

[47] J.-P. Thirion, "Image matching as a diffusion process: an analogy with Maxwell's demons," *Medical image analysis*, vol. 2, no. 3, pp. 243-260, 1998.

[48] H. Wang *et al.*, "Validation of an accelerated ‘demons’ algorithm for deformable image registration in radiation therapy," *Physics in Medicine & Biology*, vol. 50, no. 12, p. 2887, 2005.

[49] P. Cachier, X. Pennec, and N. Ayache, "Fast non rigid matching by gradient descent: Study and improvements of the" demons" algorithm," Inria, 1999.

[50] D.-J. Kroon and C. H. Slump, "MRI modalitiy transformation in demon registration," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009, pp. 963-966: IEEE.

[51] A. Friberg and J. Sundberg, "Perception of just-noticeable time displacement of a tone presented in a metrical sequence at different tempos," *The Journal of The Acoustical Society of America*, vol. 94, no. 3, pp. 1859-1859, 1993.