AUTOMATIC LEGATO TRANSCRIPTION BASED ON ONSET DETECTION

Simon Falk & Bob L. T. Sturm
Division of Speech, Music and Hearing,
KTH Royal Institute of Technology, Stockholm
simonfal@kth.se, bobs@kth.se

Sven Ahlbäck
DoReMIR Music Research AB
sven.ahlback@doremir.com

ABSTRACT

This paper focuses on the transcription of performance expression and in particular, legato slurs for solo violin performance. This can be used to improve automatic music transcription and enrich the resulting notations with expression markings. We review past work in expression detection, and find that while legato detection has been explored its transcription has not. We propose a method for demarcating the beginning and ending of slurs in a performance by combining pitch and onset information produced by ScoreCloud (a music notation software with transcription capabilities) with articulated onsets detected by a convolutional neural network. To train this system, we build a dataset of solo bowed violin performance featuring three different musicians playing several exercises and tunes. We test the resulting method on a small collection of recordings of the same excerpt of music performed by five different musicians. We find that this signal-based method works well in cases where the acoustic conditions do not interfere largely with the onset strengths. Further work will explore data augmentation for making the articulation detection more robust, as well as an end-to-end solution.

1. INTRODUCTION

Automatic Music Transcription (AMT) is the process of algorithmically converting musical audio into a symbolic representation, such as common practice notation or "piano roll" [1]. There exist several systems that perform AMT [2, 3], but these focus on detecting what pitches that have been played and leave aside *how* the notes are musically expressed, individually and together. Figure 1 shows several examples of musical expression, including legato or staccato phrasing, dynamics, and ornamentation. An analogy with automatic speech transcription is looking at more than what words have been spoken, but how.

While four decades of work has been published about AMT [2], relatively little has been devoted to detecting expressions, and much less to the transcription of expression [3]. The state-of-the-art system for piano transcription [4] predicts note dynamics in piano roll notation. Dannenberg et al. [5] investigates the detection of idiosyncratic playing styles of a musician – the purpose of which is not for AMT, but instead to enable a natural mode of control of a computer music synthesis system. Some work has looked at extracting expressive tone parameters from recorded performances of solo instruments [6–9]. Several systems have been proposed for detecting a variety of expressions and playing techniques for bowed instruments [10–14] voice [15–17], bamboo flute [18], clarinet [19], and guitar [20, 21]. Applications of the work above are in synthesizing expressive performance, or teaching assistance; however, to the best of our knowledge none of it has been applied transcribing expression of bowed violin performance.

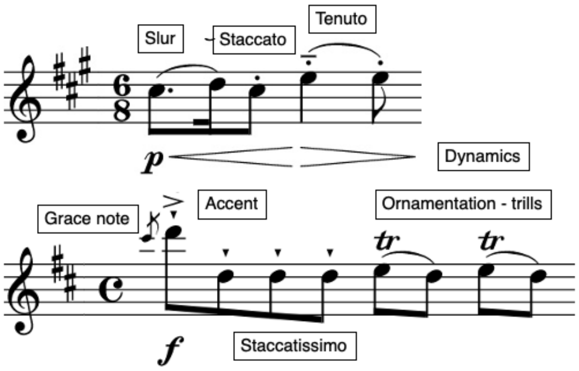

Figure 1. Examples of various expression markings specified in a musical score.

Legato slurs are prevalent in scores for violin (and other strings), and specify how the tones of phrases are connected. In this paper, we seek to automatically detect and demarcate phrases played legato on the bowed violin, and to transcribe the expression along with the tones played. Unlike in prior work detecting legato notes, e.g., [7, 8], we look to detect in a music audio recording the articulated onsets that accompany a change of bow pressure on the part of a performer beginning a legato phrase. Then we make inferences for an entire phrase where the legato slur should be placed. The results of this work can be used to enrich the AMT systems, such as ScoreCloud [22], with expression markings, as well as improve transcription performance.

Our method differs from previous work in several ways. In contrast to earlier work that classifies single notes or

Ref.	Expression	Data	Task	Method	Evaluation
[8]	Dynamics, articulation	saxophone, monophonic, two short excerpts	classification of excerpt as staccato/legato	note and intra-note segmentation, geometric envelope characterization	legato measure correlates with articulation, no quantitative metric
[6, 7]	“tone parameters” (including articulation and vibrato)	different instruments, synthesized audio and recorded performances	onset/offset detection; estimation of tone parameters of notes	note segmentation, characterization of notes and transitions, feature extraction	precision and recall for onset detection only
[10]	five playing techniques	violin, monophonic, RWC database + recorded audio	note-wise classification	Decision tree with envelope times, pitch, spectral width	accuracy
[20, 21]	legato, appoggiatura, glissando	guitar, 72 note-pairs of one performer	note-pair classification	onset detection, pitch extraction	classification accuracy
[23]	sustain pedalling	synthesized, recorded from a single MIDI-controlled piano	detection by frame-level classification	convolutional neural network	accuracy
[14]	eleven playing techniques	violin, viola, cello, contrabass, instrument samples from RWC	multi-class classification	SVM on temporal, spectral and cepstral features	macro-average F-measure

Table 1. Summary of relevant work in expression detection from recorded music performance.

pairs thereof, our method can classify an arbitrarily long sequence of notes. In contrast to earlier work that outputs an expression measure at phrase level, our method will output an expression property of each note. In contrast to previous work on expression detection in general, our method addresses the transcription of expression markings in common practice notation, and we create a dataset where we have annotated the usage of performance expression.

In the next section, we review music expression detection work that is particularly relevant to our aims. The third section presents an approach to legato transcription. We then present the data we have collected and made available. In the fifth section, we evaluate our system as a whole. We finally discuss results, and identify the challenges that remain for this task, as well as possible ways forward.

2. REVIEW OF RELEVANT WORK IN MUSIC EXPRESSION DETECTION

Table 1 summarizes the details of several publications approaching expression detection that are relevant to our work here, i.e., detecting expression from the acoustic signal. For each reference, we look at the music expressions of interest, the data used, the task performed, the method proposed, and how the results are evaluated. In particular, this section provides details contrasting past work against our method for legato transcription of bowed violin performance.

Friberg et al. [7] estimate articulation from the gap between two notes relative to the inter-onset-interval. The authors detect onsets and offsets by computing the crossing positions of filtered versions of the sound level curve. They also estimate vibrato in terms of rate and extent from detecting oscillations of a sustained tone, preceded by fundamental frequency estimation. They evaluate the onset- and offset detection accuracy, as well as the accuracy of the sound and frequency levels on synthesized and human performances of electric guitar, piano, flute, saxophone, clarinet and violin. However, they do not evaluate their estimated articulation or vibrato parameters.

Svensson [6] did a series of listening tests to find that the perceived articulation correlates with the articulation measure used in [7]. The author defines perceived articulation globally: that is, on all the notes in a given recording. Furthermore, she determines ranges of the articulation parameter corresponding to legato, ”nominal articulation” and staccato, when analyzing the CUEX system on eight different children’s songs performed on the recorder, by one teacher and five students.

Maestre & Gomez [8] detects dynamics and articulation by note segmentation with subsequent envelope and F0-contour characterization. Their note segmentation method combines early-day approaches to onset detection and f0 estimation, which after a series of processing steps give estimates of the split-points of the envelope. With this approach, there is a need to do note-by-note fine-tuning of many of the parameters. The authors propose two legato measures based on a geometric interpretation of the envelope at note transitions. They report the success of their articulation detection on a single saxophone recording, played by a professional performer. One of the legato measures is observed to correlate with the chosen articulation of a phrase, but individual notes are not evaluated. The authors treat dynamics by approximating the energy envelope, reporting the approximation error. However, the detection of loud notes is missing in this study.

Note segmentation and subsequent frequency contour analysis were used to detect legato and grace notes [20] as well as glissando [21], specific to classical guitar. These papers were published within a year and roughly contain the same methodology and material: recordings of 24-note guitar exercises. Their method for evaluating the system is restricted to pairs of notes, and measures the success rate of a correctly classified note pair as legato/glissando.

In the case of piano, sustain pedalling can be used to convey a legato effect. Liang et al. [23] approach frame-level binary pedal classification by fusing the outputs of two convolutional neural networks. The authors create a dataset by pairing MIDI files from a large dataset (1570 files) with corresponding audio recordings from a Disklavier piano. They derive a pedal-on and pedal-off label from the sustain-pedal MIDI message for each of the frames, and use this label to evaluate the model.

Kruger et. al. [14] classify between playing techniques used for bowed string instruments by training a Support Vector Machine (SVM) on temporal, spectral and cepstral features [14]. 11 different techniques such as vibrato, pizzicato, tremolo and glissando are accounted for, while legato is not distinguished from non-legato notes. The authors use note samples on violin, viola, cello and contrabass from the RWC database. Similar works include using envelope- and spectrum-based decision trees [10] or sparse features from the magnitude spectrum [11] for playing technique classification.

3. TRANSCRIBING LEGATO EXPRESSIONS

Legato markings on a score, such as those shown in Figure 4, indicate to a performer how a melodic phrase is to be executed. For bowed violin, notes grouped by a slur are to be played such that they smoothly connect. Consecutive notes that are not grouped with a slur are performed such that they are perceived separated from each other. The first note of a slur is articulated to stand apart from the note that comes before, but the tones inside a slur are not articulated. Since articulated onsets are created when articulating a note, this motivates a signal-based approach to detecting slurred notes, and thus transcribing legato expressions.

We detect articulated notes by combining the output from the ScoreCloud AMT system (see Fig. 2a) with the output from our trained model for articulated onset detection (see the "X" marks in Fig. 2d). We first retrieve notes in piano-roll notation by exporting a performance MIDI file from the development environment of ScoreCloud. [1] As seen in Fig. 2c, these notes are described by the start time, end time and MIDI pitch, respectively. Given N such notes, denote $\alpha_{i=1}^{N}$ as the series of binary articulation variables, and consider a tolerance window of $\pm w$ seconds. We set the values of the articulation variables by the following rule, which is illustrated in Fig. 2d-f. For each note, if there is at least one detected articulated onset, within a distance less than w from the note start, then $a_i = 1$, otherwise we set $a_i = 0$. Fig. 2e shows a close-up example of an articulated onset found within the window.

We demarcate slurs wherever there is a 1 followed by one or several 0:s in the sequence $\alpha_{i=1}^{N}$. Thus, an articulated note proceeded by non-articulated notes is a start-of-slur (SOS). Conversely, a non-articulated note preceding an articulated note is an end-of-slur (EOS). As for the edge cases, we stipulate that slurs cannot end on the first note, and cannot start on the last note. Connecting the start-of-slur notes with the end-of-slur notes demarcates the slurs.

Finally, we create automatic legato transcriptions as follows. We save the ScoreCloud transcription of the tune in MusicXML. Then, we convert this file to ABC notation [2]. We save the tune header for later and clean the string inside the nonempty voice field by removing bar line symbols and comments so that only notes and rests remain. We index all note symbols from 1 to N, inserting a "(" symbol before each SOS note and a ")" symbol after each EOS note (see Fig. 2g). In a last step, we assemble the tune header together with the processed tune body. We render the ABC notation with music21 [3] and MuseScore [4] to yield a legato transcription of the excerpt as Fig. 2h shows.

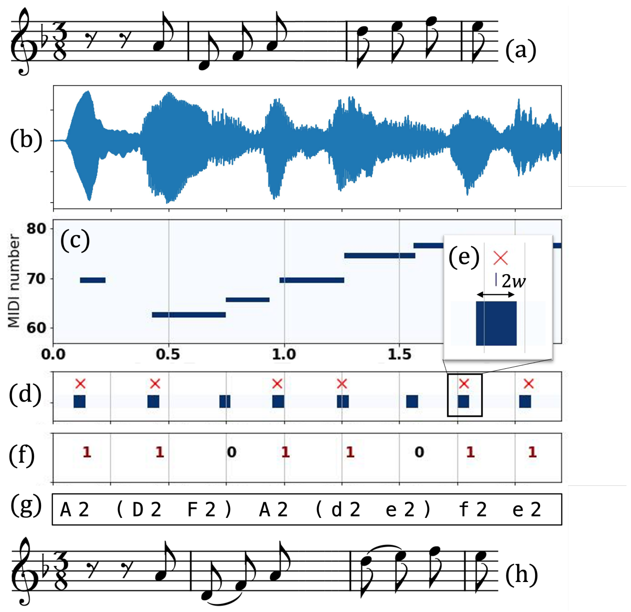

Figure 2. (a) Note input as score. (b) Audio input signal. (c) Performance MIDI as piano roll. (d) Detected articulated onsets (X) and windows (□). (e) Close-up version of (d). (f) Articulation variables. (g) Transcribed slurs in ABC notation. (h) Transcribed slurs as score.

For onset detection we use the convolutional neural network designed by Schlüter and Böck [24]. The input to the network is a multiscale dB Mel spectrogram with 80 frequency bands, and 15 contiguous analysis frames computed from 150 ms of audio data. The network infers if an articulated onset occurs in the center frame. Table 2 shows the architecture. More technical details can be found in [24].

Prior to both fully connected layers we use a dropout layer with a dropout rate set to 0.5. We initialize the network with weights trained for the original Onset Detection task [5], and train all layers in a fine-tuning process. We use a batch size of 256, and a learning rate of 0.05. In contrast to [24], we use Adam as optimizer and a weighted binary cross-entropy loss

$$\mathcal{L}(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[\beta y_i \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})\right] \quad (1)$$

where N is the number of samples in a batch, and $\hat{y}_i$ is the output from the network. We set the weight β empir-

[1] The development environment is the latest software version that is not publicly available.

[2] For a description of the ABC notation, see https://abcnotation.com/wiki/abc:standard:v2.1

[3] http://web.mit.edu/music21/

[4] https://musescore.org/sv

[5] Model parameters and peak-picking function can be found at https://madmom.readthedocs.io/en/v0.16.1/_modules/madmom/features/onsets.html

Input: Mel-spectrogram (15x80x3)
Convolutional 2D (3x7x10), tanh
Max pooling (1x3)
Convolutional 2D (3x3x20), tanh
Max pooling (1x3)
Dropout
Fully-connected (256 units), sigmoid
Dropout
Fully connected (1 unit), sigmoid
Output

Table 2. Architecture for the convolutional neural network used for articulated onset detection.

ically to account for the imbalance of frames labelled as articulated onsets.

The output is post-processed by the peak-picking function designed by Böck et al. [25], which has an implementation in the Madmom software [5]. As parameter settings, we use a threshold of 0.5 and a smoothing parameter of 5. One frame backwards and five frames forward are used for moving average and maximum. The output from the peak-picking function is a list of predicted articulated onset times.

4. THE *SLURTEST* DATASET

To explore legato transcription we create a dataset in the following way. [6] First, we prepare scores of different violin exercises and tunes. We ask three different violin performers, identified with the initials SA, FK and IR, to play from the score and record the performance. SA is a professional musician and professor in folk music performance, while FK and IR are music students. FK and IR record their performances in small dry rooms with a Nikabe USB microphone, whereas SA records in a larger living room using an iPhone. The resulting recordings are monophonic wav files with a sample rate of either 44.1 kHz or 48 kHz.

Our dataset consists of 49 recordings. The total duration is 31 minutes and individual recording lengths span from 16 to 101 seconds. Some exercises are played by multiple performers and some exercises are recorded in multiple takes by one and the same performer. Each performer contributes different amounts as seen in Table 3. The exercises are written by Sven Ahlbäck (SA) and based on general violin exercises, containing repeated phrases, arpeggios, scales and similar with limited melodic diversity. We let the players themselves choose the tunes, resulting in mostly folk music tunes except one classical piece written by J.S. Bach. The contributions vary with regard to the melodic diversity as seen in Table 3.

Annotations of hard onsets are done by listening, visualizing the waveform and adding labels at hard onsets using the open-source software Audacity. [7] The resulting label track is then exported to a text file containing time points of the annotated onsets with a precision of 0.01 s. Annotations are done by the three violin performers and the main author, with no annotator labelling his/her own recording.

Player	Total time	Exercises	Tunes
SA	19 min 3 s	10 min 20 s	8 min 43 s
IR	8 min 48 s	3 min 38 s	5 min 10 s
FK	3 min 2 s	2 min 25 s	37 s
All	30 min 53 s		

Table 3. Durations of recordings in the "slurtest" dataset contributed by each performer playing exercises and tunes.

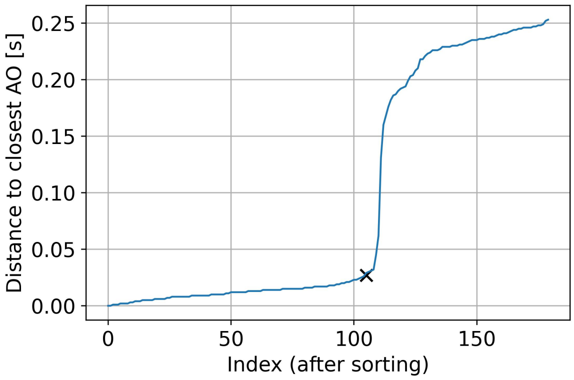

Figure 3. Distances from each note start to the closest articulated onset, computed on one recording in the validation set and sorted from smallest to largest.

The *slurtest* dataset is small compared to the one used to train the onset detection network [24]. There are only three performers, among whose recordings we observed large differences: both acoustically, and with regard to the detection performance when used for training. Because SA and IR recordings are the largest in number, those are used for training, while recordings by FK are placed in the validation set, resulting in a 90%-10% ratio with regard to the audio length.

5. RESULTS

We now describe how we use the "slurtest" dataset in training the onset detection model. Then we describe how we test the final legato transcription method using performances of a brief music passage different from the training data. This provides an assessment of the generalization of the method.

5.1 Parameter Selection

The loss weight β is chosen as the ratio between the number of frames *not* annotated as onsets and the number of frames annotated as onsets. We set this empirically on a subset of our data to $\beta = 28.5$. The next question is when to stop training. We found that the loss function (1) is not suitable for early stopping: partly because it fluctuates heavily, and partly because the weighting causes the $\{\hat{y}_i\}$ function to change its shape during training. The latter causes a negative feedback loop that attenuates changes

[6] Our (*slurtest*) dataset can be found at `https://www.kth.se/profile/bobs/page/research-data`

[7] `https://www.audacityteam.org`

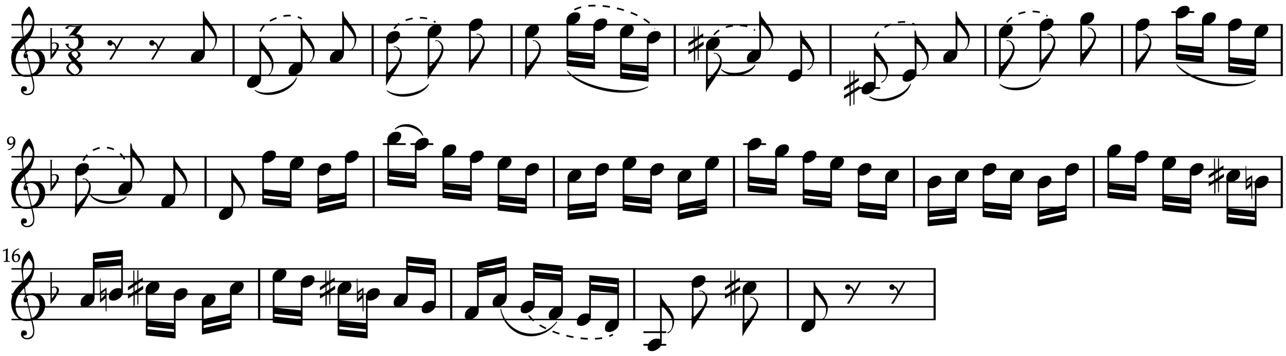

Figure 4. J. S. Bach's "Gigue" from *Partita 2*, with slurs transcribed by an expert (SA, dashed) and slurs transcribed by our legato transcription system (solid)

in $\mathcal{L}$. To decide on a stopping criterion, we first pay attention to at what epoch the F-score is maximized on the validation set. F-score is defined as in the MIREX Onset Detection task and computed with a tolerance window of 0.025 s. [8] Then, we re-train the full model on the training and validation data and stop the training after this many epochs; which we find to be 32 epochs.

The window length w should allow for random errors in annotation and detection and be set to the maximum distance that the start of notes and articulated onsets can separate on the time axis and still be associated. Therefore we calculate, for each note of one of the recordings in the validation set, the closest distance to a detected articulated onset. Next, we plot those distances, sorted from smallest to largest, and select as w the largest distance for which the slope has yet not become steep (see Fig. 3). This results in $w = 0.025$ s.

5.2 System Test

We now turn to testing our legato transcription system with violin performances of the beginning measures of "Gigue" from *Partita No. 2* for violin in D minor, BWV 1004 (J. S. Bach). This music is not in the "slurtest" dataset (Sec. 4), and is only used after training our method. We collect from *YouTube* at random five recordings of this passage played by different performers in different acoustic and recording conditions: Hillary Hahn (HH), Sara Övinge (SÖ), Iztak Perlman (IP), Rachel Kolly (RK), and Yehudi Menuhin (YM). We manually transcribe each performance with legato markings as performed, which provides the ground truth for evaluating our method. Figure 4 shows the ground-truth slurs for one of these five recordings.

Following the method in 3, we load the mp3 files in ScoreCloud, and save them both as MusicXML and as performance MIDI (unquantized). We correct a systematic error in those performance MIDI files, where the start time of each note lags 0.07 s relative to the corresponding articulated onsets. We then compute binary articulation variables by matching between articulated onsets and note onsets and subsequently transcribe slurs by inserting slur symbols at SOS and EOS notes. The transcribed performance by HH can be seen in Fig. 4, where solid lines denote predicted slurs.

To measure how well our transcribed slurs match the ground-truth we propose using a string edit distance, or Levenshtein distance. We first convert the predicted and ground truth transcriptions into strings for comparison. Let s_i be a character, which is "(" if the note i is such that a slur starts there; or which is ")" if a slur ends there; or which is "-" otherwise. The string $(s_i)_{i=1}^{N}$ is thus a representation of the positions of slurs of an excerpt. The Levenshtein distances between the predictions and ground truths for all recordings in our test set are reported in Table 5.2 alongside the created strings.

6. DISCUSSION AND CONCLUSION

Our results show that successful legato transcription is possible, but also highlights the challenges coming with this task. Bar 18 of Figure 4 shows an example of when our system correctly identifies a passage as played legato, but fails to identify the correct start- and end-points of the transcribed slur. Unlike previous work such as [6–8] our aim is to detect those details, but we now see why this is particularly challenging. However, in light of difficulties of neural networks to generalize beyond training data, the performance on the Hillary Hahn recording is better than expected.

We only show the score representation for one recording in our test set, while the system performance on the other recordings can be assessed by looking at Table 5.2. Clearly, the slurs of the HH performance are captured successfully, while the prediction on the other recordings is not very impressive. It should be noted that the recording quality of the CW performance is dubious — the violin sound is almost synthetic. The SÖ performance stands out from the rest with a personal bowing style and heavy reverberation. We hear no clear articulated onsets in the SÖ recording, which explains the many long slurs in the predicted string in Table 5.2. We leave for further work to attend to varying recording quality, bowing style and reverberation. We believe data augmentation to be a viable

[8] https://www.music-ir.org/mirex/wiki/2013:Multiple_Fundamental_Frequency_Estimation_%26_Tracking

Perf.	SED	Reference string Hypothesis string
HH	0.081	-()-()--(--)()-()-()------()--(--)---- -()-()--(--)()-()-()--(--)()------()---(-)------
SÖ	0.302	-()--()-(--)()-()-()--(--)()--(--)---- (----)(-----)(---------------)----()---()()---------------------(-)------------(-)--()
IP	0.267	-()-()------()-()-()------()---(---)---- ------(---)-()----()--(-)-----()--()----------()(-)-------()-----------------(---)--()
RK	0.291	-()-()--(--)()-()-()--(--)()------------------------------()-----------------(---)-()- -()-()-(---)()-()(--)-(----)---()---()-----()()--()---------()--()------()-----(-)----
CW	0.453	-()-()--(--)()-()-()--(--)()-- (-)----(--)()--()-----(---)()-(-)-(-)-()()--()(-)---()---()-()--(-)-()(-)()--()-()----

Table 4. Reference string and hypothesis string constructed from transcription of slurs, for each performer in the test set, along with the string edit distance between each.

method, as it has been used to address these challenges within the AMT field [26, 27].

We proposed a quantitative metric based on the string edit distance. It should penalize missed or superfluous slurs more than misplaced/shifted slurs. We have not yet analyzed if this holds. Comparing RK and IP [ref] in Table 5.2 we observe that the slurs in the first bars (∼ 20 first characters) are better predicted on RK than on IP. In both cases however, the 16:th-note part contains several erroneously predicted slurs. RK thus scores worse than IP because of more errors in the last bars, while it manages to capture the pattern of slurs of the first bars. This is an effect of the way the test data piece is composed. Further work on metric design for expression detection should consider these irregularities.

We did automatic legato transcription using a modular approach, by combining the result of articulated onset detection with ScoreCloud performance MIDI. As model for articulated onset detection we chose to use the CNN onset detection model by Böck et al, since it was seen to outperform signal-based onset detection methods in a preliminary study. Choosing this model was motivated by several factors: it was ranked highest in the Onset Detection MIREX task, its size was well-adapted for our small dataset, and it was available online. However, deep learning research has progressed much since the creation of the CNN model, resulting in for example attention-based models [28], which could further improve the performance of our system.

Taking a modular approach also brings certain limitations. First, the CNN model is tuned to maximize the F-score for the articulated onset detection task. This does not necessarily optimize the system as a whole; it was observed that high precision is more important than high recall for the subsequent matching step. Second, we use a constant matching window size based on the validation set. Music containing vibrato and played in slower tempi will likely require larger matching windows.

Future work should investigate end-to-end approaches, for which the main challenge will be access to large-scale datasets with good quality. This challenge has been addressed with data-mining for other tasks; however, it should be noted that processing a notated representation of the audio is not enough. Our ground-truth do not merely contain the expression markings found in a printing, but annotated by an expert listening to *how* a particular performance is expressed. We hypothesise that the correlations between score and actual performance are rather weak.

Articulation, in the sense of either articulating the attack of a note, or tying notes together, is a general property that can be effectuated on several instruments. The applications for AMT systems range from music education, musicological research, and Human-Computer Interaction. Given the wide application range of AMT, and articulation having such a central role in the grammar of musical language, it is surprising that so little research yet has been devoted to articulation transcription.

In this paper, we presented a method for automatic transcription of legato slurs. We also create a dataset of solo violin performances with annotated articulation onsets. We took advantage of this dataset to train a neural network to recognize articulated onsets, which inform us where slurs are to be placed. Slurs are transcribed by combining the output of ScoreCloud with our trained network. We test our system on unseen data, containing different players and acoustic conditions, and find that we can successfully transcribe legato in cases of clear and accentuated playing styles. Further work is needed for adapting to conditions such as heavy reverberation and personal bowing techniques. However, we present here a working method for the overlooked task of expressive AMT, which we believe requires more attention.

Acknowledgments

This paper is an outcome of MUSAiC, a project that has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 864189). We thank Fanny Källström and Ingrid Rodebjer for contributing to the "slurtest" dataset.

7. REFERENCES

[1] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*, 1st ed. Switzer-

land: Springer International Publishing, 2015.

[2] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, “Automatic Music Transcription: An Overview,” *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, Jan. 2019.

[3] L. Liu and E. Benetos, “From Audio to Music Notation,” in *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity*, E. R. Miranda, Ed. Cham: Springer International Publishing, 2021, pp. 693–714.

[4] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, “Onsets and Frames: Dual-Objective Piano Transcription,” in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, Paris, France, 2018, pp. 50–57.

[5] R. B. Dannenberg, B. Thom, and D. Watson, “A machine learning approach to musical style recognition,” in *Proc. Int. Computer Music Conf.*, 1997, pp. 344–347.

[6] M. Svensson, “Automatic segmentation and classifica tion of articulation in monophonic music,” Master’s thesis, Royal Institute of Technology (KTH), Stockholm, Sep. 2004.

[7] A. Friberg, E. Schoonderwaldt, and P. Juslin, “CUEX: An Algorithm for Automatic Extraction of Expressive Tone Parameters in Music Performance from Acoustic Signals,” *Acta Acustica united with Acustica*, vol. 93, pp. 411–420, May 2007.

[8] E. Maestre and E. Gómez, “Automatic Characterization of Dynamics and Articulation of Expressive Monophonic Recordings,” in *AES 118th Convention*. Barcelona, Spain: Audio Engineering Society, May 2005.

[9] F. J. M. Ortega, S. I. Giraldo, A. Perez, and R. Ramírez, “Phrase-Level Modeling of Expression in Violin Performances,” *Frontiers in Psychology*, vol. 10, p. 776, Apr. 2019.

[10] I. Barbancho, C. de la Bandera, A. M. Barbancho, and L. J. Tardon, “Transcription and expressiveness detection system for violin music,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Taipei, Taiwan: IEEE, Apr. 2009, pp. 189–192.

[11] L. Su, H.-M. Lin, and Y.-H. Yang, “Sparse Modeling of Magnitude and Phase-Derived Spectra for Playing Technique Classification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2122–2132, Dec. 2014.

[12] P.-C. Li, L. Su, Y.-H. Yang, and A. W. Y. Su, “Analysis of expressive musical terms in violin using score-informed and expression-based audio features,” in *Proc. ISMIR*, 2015, pp. 809–815.

[13] Z. Wang, J. Li, X. Chen, Z. Li, S. Zhang, B. Han, and D. Yang, “Musical Instrument Playing Technique Detection Based on FCN: Using Chinese Bowed-Stringed Instrument as an Example,” *arXiv:1910.09021 [cs, eess]*, Oct. 2019.

[14] A. B. Kruger and J. P. Jacobs, “Playing technique classification for bowed string instruments from raw audio,” *Journal of New Music Research*, vol. 49, no. 4, pp. 320–333, Aug. 2020.

[15] S. S. Miryala, K. Bali, R. Bhagwan, and M. Choudhury, “Automatically identifying vocal expressions for music transcription,” in *International Society for Music Information Retrieval Conference*, 2013.

[16] Y. Ikemiya, K. Itoyama, and H. G. Okuno, “Transcribing vocal expression from polyphonic music,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, May 2014, pp. 3127–3131.

[17] Y. Yamamoto, J. Nam, H. Terasawa, and Y. Hiraga, “Investigating time-frequency representations for audio feature extraction in singing technique classification,” in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2021.

[18] C. Wang, V. Lostanlen, E. Benetos, and E. Chew, “Playing Technique Recognition by Joint Time–Frequency Scattering,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 881–885.

[19] W. L. Coyle and J. D. Gabriel, “A method for automatic detection of tongued and slurred note transitions in clarinet playing,” *The Journal of the Acoustical Society of America*, vol. 146, no. 3, pp. EL238–EL244, Sep. 2019.

[20] T. H. Özaslan, E. Guaus, E. Palacios, and J. L. Arcos, “Attack Based Articulation Analysis of Nylon String Guitar,” 2010.

[21] T. H. Özaslan and J. L. Arcos, “Legato and Glissando Identification in Classical Guitar,” in *Proceedings of the 7th Sound and Music Computing Conference, SMC 2010*, Barcelona, Spain, 2010, pp. 457–463.

[22] Doremir Music Research AB, “ScoreCloud Studio,” 2021.

[23] B. Liang, G. Fazekas, and M. Sandler, “Piano Sustain-pedal Detection Using Convolutional Neural Networks,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, May 2019, pp. 241–245.

[24] J. Schluter and S. Bock, “Improved musical onset detection with Convolutional Neural Networks,” in *2014*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, May 2014, pp. 6979–6983.

[25] S. Böck, F. Krebs, and M. Schedl, “Evaluating the online capabilities of onset detection methods,” in *Proceedings of the 13th International Society for Music Information Retrieval Conference*. Porto, Portugal: FEUP Edições, 2012, pp. 49–54.

[26] J. Schlüter and T. Grill, “Exploring data augmentation for improved singing voice detection with neural networks,” in *International Society for Music Information Retrieval Conference*, 2015.

[27] J. Thickstun, Z. Harchaoui, D. P. Foster, and S. M. Kakade, “Invariances and Data Augmentation for Supervised Music Transcription,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 2241–2245.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762