ON THE IMPORTANCE OF "REAL" AUDIO DATA FOR MIR ALGORITHM EVALUATION AT THE NOTE-LEVEL – A COMPARATIVE STUDY

Bernhard Niedermayer[1], **Sebastian Böck**[1]
[1]Dept. of Computational Perception
Johannes Kepler University Linz, Austria
music@jku.at

Gerhard Widmer[1,2]
[2]Austrian Research Institute for Artificial Intelligence
Vienna, Austria
music@ofai.at

ABSTRACT

A considerable number of MIR tasks requires annotations at the note-level for the purpose of in-depth evaluation. A common means of obtaining accurately annotated data corpora is to start with a symbolic representation of a piece and generate corresponding audio data. This study investigates the effect of audio quality and source on the performance of two representative MIR algorithms – Onset Detection and Audio Alignment. Three kinds of audio material are compared: piano pieces generated using a freely available software synthesizer with its default instrument patches; a commercial high-quality sample library; and audio recordings made on a real (computer-controlled) grand piano. Also, the effect of varying richness of artistic changes in tempo and dynamics or natural asynchronies is examined. We show that the algorithms' performance on the different datasets varies considerably, but synthesized audio, does not necessarily yield better results.

1. INTRODUCTION

Onset Detection, Automatic Transcription, or Audio Alignment are only a small number of examples of MIR tasks that require ground truth data at the note-level for an in-depth evaluation. However, such data corpora are rare for several reasons. Starting from an audio recording, manual annotation is not only highly time consuming but also has certain limits in terms of accuracy and level of detail. On the one hand, it is questionable how precisely or consistently a human annotator can determine note onsets – particularly "soft" ones. On the other hand, aspects like the loudness of an individual chord note might not be distinguishable even for experienced listeners.

	Man.	Synth.	Diskl.
Audio Onset Detection	X		
Real-Time Audio to Score Alignment	X		
Audio Melody Extraction	X		
Multiple f_0 Estimation and Tracking	X	X	X
Audio Chord Estimation	X		
Audio Beat Tracking	X		

Table 1. Overview of MIREX tasks and the respective sources of test data (manual annotation, synthesized from MIDI, playback on a Disklavier)

Starting from a symbolic representation implies its own challenges. To obtain a realistic audio representation, two aspects have to be taken into account. First, the symbolic data should describe a human-like performance, i.e. contain artistic variations in tempo, dynamics, or playing style and also more subtle ones such as slight arpeggiations or asynchronies.

The second important aspect is the quality of the conversion from the symbolic to the audio domain. One option is to use computer controlled musical instruments (e.g. a player piano) preserving the whole acoustic complexity of the sound source. Problems are the availability of such instruments and recording issues. An alternative would be the usage of (software) synthesizers. Although this method is relatively common in the literature (see [2, 3, 8] for example), it is not clear if and to what extent such data yields different results in an evaluation process.

Table 1 gives an overview of MIREX [4] tasks which require note- or at least beat-level annotations for evaluation purposes. With the exception of one single task, where audio material is generated from a symbolic ground truth representation, there is a clear preference towards the usage of "real" audio recordings and human annotations. However, it is not clear if this under-representation of evaluation data generated from a known ground truth is due to a lack of

such symbolic data and adequate rendering mechanisms, or if such audio material would indeed adulterate evaluation results.

This work presents a study on different approaches for the generation of audio data from symbolic representation and their influence on evaluation results of two MIR algorithms – Onset Detection and Audio Alignment. To this end, MIDI data from real piano performances were turned into audio recordings in three ways: (i) by recording the sound produced by a computer-controlled piano when playing the MIDI files; (ii) by synthesizing the data using a commercial high-quality sample library; and (iii) by using a freely available sound patch library. Also, since performances of professional musicians are rarely available in a symbolic representation, the influence of changes in the richness of artistic variations (i.e. changing tempo, dynamics, pedal pressure) was studied. The piano was chosen due to the availability of computer controlled instruments and thus the opportunity to obtain highly accurate audio data other version can be compared to. Also, piano music is a common means of note-level evaluation in literature.

2. EVALUATION TASKS

To examine the effect of different sound sources on the performance of MIR algorithms, two sample subfields have been selected – (i) Onset Detection and (ii) Audio Alignment. These two task are representative insofar as they allow certain conclusions to be drawn about various other MIR tasks they are either integral parts of (such as Audio Transcription or Cover Version Detection) or share crucial subroutines or features (such as Score Following, Structural Analysis, or Beat Tracking).

2.1 Onset Detection

The chosen algorithm for Onset Detection is the one that yielded the highest average f-measure in the MIREX 2010 [1] algorithm comparison [5].

2.1.1 Features

Features are extracted in the spectral domain. The signal is therefore transformed using two parallel STFTs with Hamming windows of lengths 1024 (23 ms) and 2048 (46 ms) respectively. The hop size, however, is 441 samples in both cases yielding a common time resolution of 10 ms per frame. According to the human perception of sounds, the (power) spectrograms are then converted to the Mel-scale using a filterbank consisting of 40 triangular filters spread equidistantly on the Mel-scale. In a last step, the logarithm is taken to obtain the final feature values.

In addition to the absolute values, the half-wave rectified first order difference is calculated as an indicator for new spectral components.

2.1.2 Algorithm

As most other Onset Detection algorithms, the one used here works in two steps. In the first one, a detection function is calculated, representing novelty within the signal. In a second pass, peaks in the detection function are picked and classified as onsets.

To obtain the detection function, a bidirectional neural network with Long Short-Term Memory (LSTM) units is applied. Its number of input units is 160, corresponding to the feature values as described above. The actual neural net consists of six hidden layers – two for each direction – with 20 LSTM units each and two output units y_o and y_n representing the classes 'onset' and 'no onset' respectively. These outputs are normalized such that the range of values is $[0, 1]$ and the sum of y_o and y_n is 1.

Training of the network was done iteratively by gradient descent with error backpropagation until no more improvement has been observed for 20 epochs. The training and validation sets used consist of samples from the dataset introduced be Bello et al. [1] and the ballroom dataset by Gouyon et al. [7].

The peak picking on the detection function applies a simple thresholding approach where a fixed threshold depending on the median of the detection function is determined for each piece. Each remaining peak is finally reported as an onset.

2.2 Audio-to-Score Alignment

Concerning audio alignment, a simple algorithm based on Dynamic Time Warping (DTW) and Chroma vectors has been chosen. Although this approach dates back several years and improvements concerning aspects like robustness or accuracy have been proposed, it is still used not only for Audio-to-Score Alignment itself but also for Structural Analysis, Cover Version Detection or Retrieval Tasks. For simplicity reasons, the Audio-to-Score Alignment task will be referred to as Audio Alignment only in the remainder of this work.

2.2.1 Features

Due to their robustness to timbre, certain recording conditions, and varying degrees of polyphony, chroma vectors are commonly used for synchronization tasks. They consist of a 12-dimensional vector for each time frame, where each element represents the relative energy of a pitch class (i.e. C, C#, D,...). The extraction from audio signals is done in the spectral domain based on a mapping of each bin to the note where the fundamental frequency is closest to the bin's center frequency. In a second step, coefficients of all bins

[1] MIREX 2010 – Onset Detection Results
http://nema.lis.illinois.edu/nema_out/mirex2010/results/aod/summary.html

mapped to notes of the same pitch class are summed up. Finally, the vector is normalized by linear scaling such that its maximum is equal to 1.

The (mechanic) score representation is segmented into time frames such that the number of time frames and the overlap ratio are the same as for the corresponding audio data. The energy of a pitch is then set to the fraction of the window length in which it is played. The octave folding and normalization is then performed in analogous manner as for the audio data.

2.2.2 Algorithm

To compute the actual alignment, the approach described in [10] is used. In a first pass, features are computed on windows with a length of 4096 samples and an overlap ratio of 50%. Dynamic Time Warping is then performed to obtain an initial alignment. The resulting time resolution is relatively low. However, since the Dynamic Time Warping algorithm is of quadratic complexity in time and also in space, this is necessary to also process long pieces.

To circumvent this tradeoff, a second pass is performed at a higher time resolution. Here, the features are calculated using a window length of 1024 samples and a hop size of 256 samples. Computational costs are kept low by restricting the search for an optimal alignment to a certain area around the coarse initial alignment. Here, a radius of ± 1000 frames has been chosen.

3. EVALUATION DATA

The data set used throughout this study comprises the first movements of 13 piano sonatas by W. A. Mozart. Those pieces have been performed by a professional pianist on a computer monitored grand piano (Bösendorfer SE 290), yielding an exact ground truth of all performance parameters including timing, dynamics, and pedal pressure. The data was originally represented in a proprietary, symbolic format which was then converted into MIDI. As shown in Table 2, it covers almost 42000 notes and a performance time of more than 80 minutes.

For the purpose of evaluation, the performance data was matched to a symbolic score representation. Manual correction was done, to ensure that playing errors and also short sections where the pianist did not stick to the score at all are annotated accordingly.

Audio recordings were then obtained from this performance data using three different sources – playback on the Bösendorfer 290 SE from which the symbolic data originated, synthesizing using high quality instrument samples produced by the *Vienna Symphonic Library*, and rendering using the free synthesizer *Timidity* and its default instrument patches provided by the *Freepats* project.

3.1 Bösendorfer SE 290

The Bösendorfer SE 290 is the computer controlled grand piano which was used to obtain the symbolic performance data. It relies on optical sensors to detect movements of individual keys and hammers. One such sensor consists of a phototransistor and a coupled LED about 3 mm apart. Precision-cut aluminum shutters attached to the keys and hammers discontinue the corresponding beam of light and thus trigger a sensor event. The system is set up such that a key movement is reported as soon as it is minutely depressed. A hammer movement and its velocity, on the other hand, are detected at the instant a hammer hits the string [9].

The playback mechanism is based on small linear motors underneath the key bed actuating the keys. They are constructed such that the only contact between key and actuator is during playback mode and no interference occurs while a pianist is playing the instrument.

In [6] the SE 290 was compared to the Yamaha Disklavier grand piano – another system commonly used in performance research. It has been found that the SE 290 is more accurate than the Disklavier at monitoring and also at playback. Both systems were affected by systematic timing deviations (linearly increasing over time) likely to be caused by inaccuracies of the internal clock-pulse generators. This flaw aside, the residual mean timing errors in monitoring mode accounted for 0.2 ms (stddev: 2.1 ms) for Bösendorfer's and for 1.4 ms (stddev: 3.8 ms) for Yamaha's grand piano. Considering reproduction accuracy, the Disklavier was again clearly outperformed by the SE system where timing deviations rarely exceeded 3 ms.

The recordings on this instrument were made at 44.1 kHz using a single high-quality microphone near the corpus of the piano and a DAT recorder.

3.2 Vienna Symphonic Library

The *Vienna Symphonic Library* [2] (VSL) is a commercial vendor of high quality instrument samples not only covering a wide range of musical instruments but also different playing styles. While synthesizing MIDI data, a special sequencer plug-in analyzes the stream of events for repeated notes and other certain patterns and determines the appropriate articulation or nuance in real-time. An example are passages played in legato on wind or string instruments, where not only tones themselves but also real note transitions are sampled to yield a more natural sound.

The *Special Edition – Standard* of the sample library contains the Bösendorfer 290 "Imperial", which is the same type of grand piano the SE system, as described above, was integrated into. This provides the opportunity to compare the authentic sound of the grand piano to its generated reproduction. The objective is to show if and how potential devi-

[2] http://vsl.co.at/

ations influence MIR algorithms and their respective evaluation results.

Since the software is not a sequencer of its own, GarageBand [3] was used for synthesizing. Although GarageBand can not be considered a high-end product, the audio material obtained as described above benefits from the plug-in provided by the VSL.

3.3 Timidity++/Freepats

Timidity++ [4] is a free software synthesizer distributed under the GNU *General Public License* and available for a variety of operating systems. Although it can be configured to work with any set of instrument samples given in GUS/patch format, it, by default, uses the voice data provided by the *Freepats* [5] project. Timidity has been included in this comparison because, on the one hand, the software as well as the instrument samples are freely available and, on the other hand, it has been used in recent MIR research (e.g. [2, 3, 8]).

4. DIFFERENT RENDERING METHODS

In a first experiment, the influence of the rendering method was examined. Therefore, audio signals were obtained from the three sources as described above – the computer controlled Bösendorfer SE 290 grand piano, the Vienna Symphonic Library, and Timidity using its default sound patches. The results yielded by the Onset Detection and the Audio-to-Score Alignment are shown in Table 2. The Onset Detection performance is determined analogous to the MIREX evaluation. The reported onsets are compared to the ground truth allowing a timing deviation of ± 50 ms. The quality of the result is then given in terms of the f-measure. The accuracy of the Audio Alignment is expressed by the percentage of individual notes for which the onset time in the alignment deviates by also less than 50 ms from the ground truth.

The evaluation presented here deviates from the one performed at MIREX in one aspect, which is, however, justified by the nature of the ground truth data. Merged onsets, i.e. two adjacent onsets are reported as one single onset, are not penalized here. Since each individual note's onset time is known, it occurs that there is more than one onset within a single or two adjacent audio frames. Such onsets cannot be distinguished without also transcribing the notes' pitches.

Concerning the Onset Detection, the performance on the data synthesized using the Vienna Symphonic Library is the highest on all individual pieces with only one exception – k283-1 – where the signal from the SE 290 yields the highest f-value. On the other hand, the audio data obtained from Timidity results in the lowest f-measure for each piece. This contradicts the possible speculation that lower quality synthesizers (instrument patches) would produce somehow "artificial" sounds and in doing so reduce the complexity of the resulting audio file. Looking at the spectra of two tones – one played on the SE 290 and one generated by timidity – reveals that the tone obtained from timidity contains a significant proportion of noise in the high frequency bins (see Figure 1). This phenomenon was observed to be consistent throughout the whole pitch range and is therefore a likely explanation for the worse performance of the Onset Detection on the timidity dataset.

Although the evaluation of the Audio Alignment does not draw such a clear picture, some of the results are confirmed. Again, the performance on the timidity dataset was significantly lower than the one on the "real" audio from the SE 290. However, the VSL dataset results in the lowest overall accuracy. Comparing the spectra of tones generated by the VSL to those played on the SE 290 shows differences in the relative strengths of individual harmonics. This will influence the chroma feature and is therefore a likely explanation for the discrepancy in the results.

5. VARYING RICHNESS OF EXPRESSIVE DETAILS

The symbolic representation used to obtain the audio materials for the above experiment derives from a real performance (on the Boesendorfer SE290) by a skilled concert pianist. It thus contains detailed information about expressive performance aspects (expressive timing, dynamics nuances, exact pressure on the pedals). In many controlled MIR experiments, the starting MIDI data will be based on a score instead of real performances, and will therefore be impoverished in the sense that it will not correspond to the kind of musical material usually encountered in practice.

In order to find out whether the lack (or presence) of expressive timing etc. significantly impact MIR algorithms, our MIDI files were deliberately "cleaned" from such expressive performance aspects. Specifically, the usage of the pedals, varying dynamics, and intra-chord timings (i.e. arpeggiations and asynchronies) were suppressed by deleting the according events, setting velocities to a constant, and assigning asynchronous chord notes a uniform onset time.

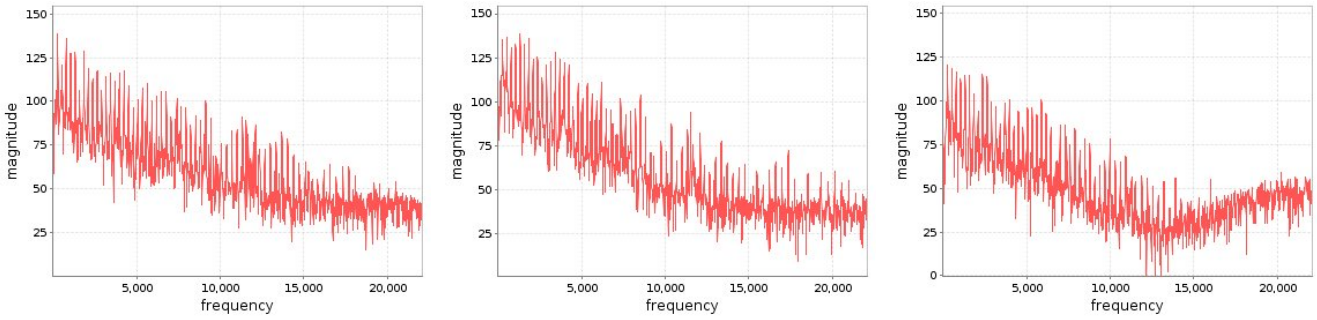
The means of synthesizing was chosen to be timidity for two reasons. First, we assumed that if a computer controlled instrument were available, it could be used to obtain the complete performance information. Second, the VSL software and its mechanism to use different samples according to the musical context would interfere with the experiment.

We found that suppressing the usage of the pedals, changing dynamics, or both had only negligible influence on the overall performance. Likely explanations are that the usage of pedals plays a relatively minor role when performing

[3] http://www.apple.com/de/ilife/garageband/
[4] http://timidity.sourceforge.net
[5] http://freepats.zenvoid.org

piece	# notes	duration	Onset Detection			Audio-to-Score Alignment		
			SE 290	VSL	timidity	SE 290	VSL	timidity
k279-1	2803	4:55	96.31	98.00	92.11	90.37	85.52	87.73
k280-1	2491	4:48	98.08	98.80	95.64	85.27	79.37	85.47
k281-1	2648	4:29	95.83	97.83	92.20	88.37	85.08	86.48
k282-1	1907	7:35	97.70	98.87	96.42	76.68	71.93	74.93
k283-1	3304	5:22	97.08	96.53	92.45	93.89	85.05	90.89
k284-1	3700	5:17	94.82	98.58	93.40	92.08	90.35	86.97
k330-1	3160	6:14	97.19	99.32	95.50	95.13	90.03	90.19
k331-1	6123	13:35	98.02	98.50	95.55	73.00	66.62	70.70
k332-1	3470	6:02	94.84	98.26	94.01	87.61	83.52	81.07
k333-1	3774	6:44	96.83	98.31	93.13	93.51	93.19	92.29
k457-1	2993	6:15	95.92	96.80	92.33	88.31	79.45	80.09
k475-1	1284	4:58	96.69	98.29	95.60	61.21	59.04	43.04
k533-1	4339	8:25	95.30	98.11	94.06	92.90	87.14	89.91
all	41994	1:24.39	96.51	98.18	94.00	86.85	81.93	82.99

Table 2. Performance of the example algorithms on the datasets generated using different rendering methods



(a) (b) (c)

Figure 1. Spectra of a C3 as played on the Bösendorfer grand piano (a) and synthesized by the VSL (b) and timidity (c) calculated applying a Blackman-Harris window of length 8192 starting 50 ms after the note onset

pieces by Mozart. Also, the chroma vectors used for Audio Alignment are normalized to reduce the influence of varying loudness and the neural network seems to have learned a similar concept.

However, the influence of micro timings (i.e. asynchronies) on the Audio Alignment was significant compared to a version where the onsets of all notes of a chord were set to same time (see Table 3). This is partly due to the fact that Audio-to-Score Alignment using Dynamic Time Warping without post-processing at the note-level is inherently error prone as soon as asynchronies occur. The algorithm cannot assign different times to events which are simultaneous in the score.

Although we expected the Onset Detection to also benefit substantially from having one simultaneous onset for a whole chord instead of several onsets of the individual notes, results disproved this assumption. A further inspection showed that while chord onsets have been correctly detected, onsets of notes played one at a time were missed. This is due to a masking effect caused be the exceptionally high values in the detection function caused by the exact concurrence of several notes' onsets.

To get an idea on the actual extent of asynchronies in a natural performance, the time spreads of chords according to their degree of polyphony was determined. Table 4 shows that two notes which are notated concurrently in the score can be up to half a second apart in the actual performance, highlighting that natural timings contribute significantly to the complexity of a musical performance.

6. CONCLUSION

We have presented an extensive comparison of different approaches to generate audio material from a symbolic repre-

piece	Onset Detection		Audio Alignment	
	full	time	full	time
k279-1	92.11	98.10	87.73	95.33
k280-1	95.64	99.30	85.47	95.19
k281-1	92.20	82.53	86.48	91.66
k282-1	96.42	92.55	74.93	96.89
k283-1	92.45	97.15	90.89	99.64
k284-1	93.40	99.52	86.97	98.57
k330-1	95.50	89.56	90.19	96.52
k331-1	95.55	98.49	70.70	99.11
k332-1	94.01	99.15	81.07	99.17
k333-1	93.13	99.73	92.29	96.88
k457-1	92.33	99.32	80.09	95.07
k475-1	95.60	91.56	43.04	80.58
k533-1	94.06	92.24	89.91	97.29
all	96.51	96.01	82.99	96.61

Table 3. Performance of the example algorithms on the datasets exhibiting all aspects of expressive variations (full) and with suppressed micro timings (time)

p	# occurrences	min	avg	max	stddev
1	15999	-	-	-	-
2	6742	0.000	0.015	0.286	0.017
3	2732	0.000	0.020	0.471	0.023
4	840	0.001	0.035	0.391	0.051
5	130	0.005	0.125	0.529	0.131
6	46	0.005	0.155	0.511	0.121
7	3	0.010	0.014	0.017	0.003
8	1	-	0.009	-	-

Table 4. Asynchronies and arpeggiations in $[sec]$ for each degree of polyphony p

sentation and its influence on the evaluation results of two representative MIR algorithms. On the one hand, the usefulness of synthesized data for evaluation purposes was proven by the large number of consistencies concerning the ranking of individual results. On the other hand, however, it became evident, that synthesized data can have their own specificities carrying the inherent risk of overfitting.

We have shown that the quality of instrument samples used for synthesizing has a significant influence on evaluation results. Also, natural timings including asynchronies and arpeggiations are a crucial aspect to account for in the ground truth data in order to obtain most meaningful evaluation results. This does not only refer to a algorithms performance on different audio data but also to evaluation itself, where such rich data would allow for criteria more accurate than, for example, the ± 50 ms tolerance threshold commonly used in onset detection.

7. ACKNOWLEDGMENTS

This research is supported by the Austrian Research Fund (FWF) under grants TRP109-N23 and Z159. Special thanks are due to the *Vienna Symphonic Library* and *Bösendorfer* for providing access to instrument samples and the instrument itself respectively.

8. REFERENCES

[1] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler: "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 5, pp. 1035–1047, 2005.

[2] R. B. Dannenberg, and N. Hu: "Polyphonic Audio Matching for Score Following and Intelligent Audio Editors," *International Computer Music Conference (ICMC 2003)*, pp. 27–34, San Francisco, 2003.

[3] S. Dixon: "On the computer recognition of solo piano music," *Australasian Computer Music Conference*, pp. 31–37, Brisbane, 2000.

[4] J. S. Downie, A. F. Ehmann, and J. H. Lee: "The Music Information Retrieval Evaluation eXchange (MIREX): Community-led formal evaluations," *Digital Humanities 2008*, pp. 239–240, Oulu, 2008.

[5] F. Eyben, S. Böck, B. Schuller, and A. Graves: "Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks," *11^{th} International Society for Music Information Retrieval Conference (ISMIR 2010)*, pp. 589–594, Utrecht, 2010.

[6] W. Goebel, and R. Bresin: "Measurement and Reproduction Accuracy of Computer-Controlled Grand Pianos," *Stockholm Music Acoustics Conference (SMAC 03)*, pp. 155–158, Stockholm, 2003.

[7] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano: "An experimental comparison of audio tempo induction algorithms", *IEEE TASLP*, Vol. 14, No. 5, pp. 1832–1844, 2006.

[8] A. Klapuri: "A method for visualizing the pitch content of polyphonic music signals," *10^{th} International Society of Music Information Retrieval Conference (ISMIR 2009)*, pp. 615–620, Kobe, 2009.

[9] R. A. Moog, T. L. Rhea: "Evolution of the keyboard interface: The Bösendorfer 290 SE recording piano and the Moog multiply-touch-sensitive keyboards," *Computer Music Journal*, Vol. 14, No. 2, pp. 52–60, 1990.

[10] B. Niedermayer: "Towards Audio to Score Alignment in the Symbolic Domain", *6^{th} Sound and Music Computing Conference (SMC 2008)*, pp. 77–82, Porto, 2008.