CMMSD: A Data Set for Note-Level Segmentation of Monophonic Music

Henrik von Coler[1,2], Alexander Lerch[3],
[1]*SIM - Staatliches Institut für Musikforschung Preußischer Kulturbesitz, Department III for Acoustics and Music Technology*
[2]*Audio Communication Group, Technical University Berlin*
[3]*Georgia Tech Center for Music Technology*
Correspondence should be addressed to Henrik von Coler (vonColer@sim.spk-berlin.de)

ABSTRACT

A musical data set for note-level segmentation of monophonic music is presented. It contains 36 excerpts from commercial recordings of monophonic classical western music and features the instrument groups *strings*, *woodwind* and *brass*. The excerpts are self-contained phrases with a mean length of 17.97 seconds and an average of 20 notes. All phrases are played in moderate tempo, mostly with significant amounts of expressive articulation. A manually annotated ground truth splits each item into a sequence of the three states *note*, *transition* and *rest*. The set is designed as an open source project, aiming at the development and evaluation of algorithms for segmentation, music performance analysis and feature selection. This paper presents the process of ground truth labeling and a detailed description of the data set and its properties.

1. INTRODUCTION

The segmentation of monophonic music into separate note events is required in many audio content analysis tasks. Especially for the extraction of semantic information, single notes have to be extracted from a performance. Relevant applications include, for example, music transcription [1], audio-to-score- and audio-to-audio alignment [2, 3] or score following [4, 5]. Various methods have been proposed for the extraction of performance parameters, related to both vibrato and intra note articulation, from isolated notes [6, 7, 8, 9, 10, 11]. There also exist several models for the analysis of note transitions and inter note articulation [12, 13, 14, 15].

Although the problem of automatic segmentation of monophonic music is often assumed to be solved, it is the authors' opinion that non-percussive instruments with weak transients and large intra-note variations still pose a problem to concurrent approaches (compare also Toh et al. [16]). This is especially true if the requirements for temporal accuracy are very strict. For the extraction of performance parameters, for example, even minor shifts in the note boundaries can significantly change the results. It is thus important to further design improved algorithms and introduce new data sets for the training and evaluation of accurate and robust *note-level segmentation*.

Most data sets related to note-level segmentation originate from the field of note onset detection. Toh et al. [16] used a set with 1127 onsets for the evaluation of onset detection algorithms. However, their set is restricted to singing voice recordings. Another data set for onset detection, presented by Bello et al. [17], consists of 1065 onsets of which 93 originate from recordings of excitation-continuous instruments. The remaining onsets belong to percussive instruments or polyphonic sources. Leveau and Daudet [18] presented a set with ground truth, containing monophonic and polyphonic music, along with a Matlab tool for onset detection. Unfortunately, only six excerpts were taken from monophonic performances. Many other data sets for onset detection feature polyphonic music, exclusively.

While onset-based segmentation is sufficient for score following and alignment applications, the detailed analysis of musical performances also requires accurate information on the boundaries of the transitions between the notes. For such purposes, the authors propose segmenting the audio signal into the three states *note*, *rest* and *transition*. Since no data set exists for this task, the *Classical*

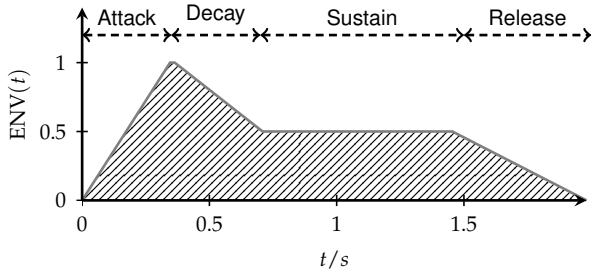


Fig. 1: ADSR model for the description of temporal envelopes

Monophonic Music Segmentation Dataset (CMMSD) is introduced. The data set, consisting of 36 excerpts with a total of 718 notes, is designed for the performance analysis of single-voiced instruments and is suitable for both, training and evaluation of segmentation algorithms. It is easily extensible for usage in related areas such as onset and offset detection. Since annotation of ground truth data is an error-prone process, the complete data of the set is provided in an open source repository. Thus, an improvement of the ground truth can be achieved by allowing professionals to contribute through submitting changes and participating in a discussion, hosted on the project-related website[1].

The remainder of this paper is organized as follows: The definition of the segments and their signal properties is described in Sect. 2. Rules and procedures for the creation of the ground truth for the musical data are presented in Sect. 3. Section 4 details the musical recordings in the set and provides relevant statistical properties. A conclusion with an outlook and future work is presented in Sect. 5.

2. DEFINITION OF THE SEGMENTS

2.1. The ADSR Model

The definition of the three segments (note, rest, transition) is based on the well known ADSR model. Within this model, the trajectory of temporal envelopes is divided into the sections *Attack*, *Decay*, *Sustain* and *Release*, as shown in Fig. 1. In the analysis of musical sounds it is used to model the characteristics of energy envelopes [19, 20]. In performance analysis, the ADSR model and derivatives have been proposed for the characterization of articulation styles and note transitions [12]. The definition of the individual sections in the ADSR model is not consistent in the literature. However, the proposed segmentation is based on the assumption that the sections of the ADSR model each have different spectral and temporal properties as, for example, reported by Every [21].

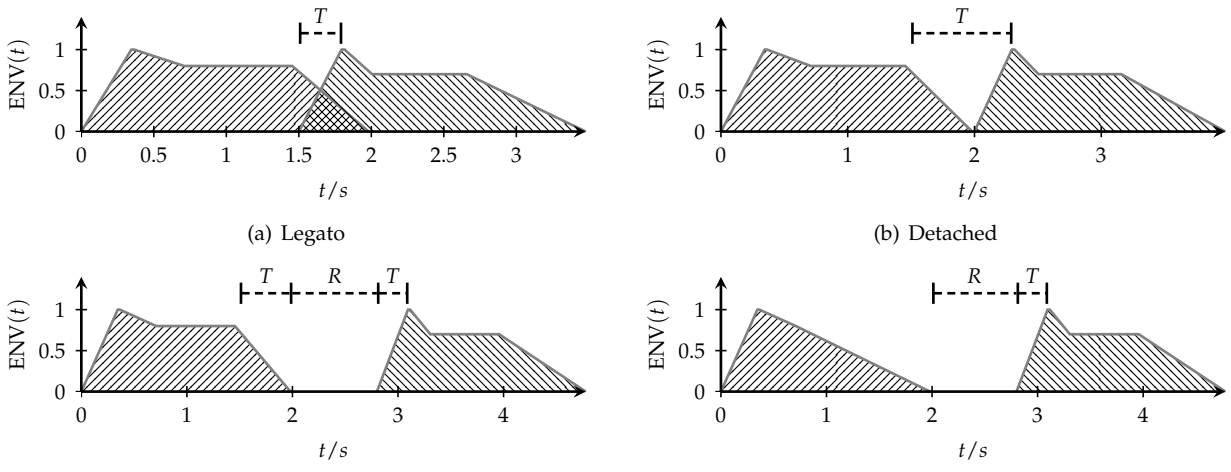
Attack The attack section is most commonly defined as the period of the *onset transient* [22, 23, 24, 25]. Transients are regions with a rapid change of the signals properties with time and contain non-harmonic sinusoidals and noise [17]. These components can be attributed to air flows and bow movements during the initial excitation. The end of the onset transient of pitched instruments is marked by stable tone properties and clearly identifiable partials [26]. Another definition for the attack section is the time span between the the very first rise in energy and the end of the most positive section of the computed slopes [12]. Thus, the attack section is characterized by a rapid increase in signal envelope. Both definitions provide a basis for the annotation of transitions in graphical representations.

Release The release section can be defined as the period of the *offset transient*. Far less standardized than the onset transient, the offset transient or *release transient* [27], marks the section between the end of the excitation and the end of the tone. Similar to the attack, the release is characterized by rapid changes in amplitude. Due to the formants of the instrument's resonant body, individual partials may decrease with different speed during the release segment. This results in significant fluctuations of the spectral envelope.

Decay The decay section has not been widely used to model instrument sounds. For example, Strawn [13] proposes the *Attack / Steady-State / Decay* envelope for modeling notes and transitions. With *Decay*, however, he refers to the section that is defined as release in our context. Since this seems to be sufficient for the modeling of excitation-continuous instruments, we will ignore the decay section for further considerations, too.

Sustain When omitting the decay, the sustain section extends between attack and release, or onset-

[1] http://intelligent-noise-solutions.de/research/cmmsd/



(a) Legato

(b) Detached

(c) Rest-separated

(d) Rest-separated without significant release in the first note

Fig. 2: Note transition types based on ADSR envelopes – identifiers R for rest and T for transition

and offset transient, respectively. It marks the continuously excited, steady-state part of the sound. In contrast to the transient segments, the sustain is characterized by moderate variations in amplitude and frequency content over time. It features a quasi-periodic signal with only minor changes of the partials' amplitudes and frequencies. However, more drastic changes such as expressive intonation and vibrato, however, can be added by the performer.

2.2. Transitions based on the ADSR Model

For the proposed segmentation we define three segment classes:

Note: Steady state, sustain regions

Transition: Transient attack/release regions

Rest: Noise, silence and reverberation

Using the ADSR model, we define four basic transition types, shown in Fig. 2.

Legato Transitions In *legato transitions* the offset of the first note and the onset of the second note occur simultaneously. The excitation is usually not interrupted. Due to the superposition of the energy of both notes there might be no significant changes in overall energy. The comparably short transition is defined as the attack section of the second note, as shown in Fig. 2(a).

Detached Transitions Figure 2(b) shows a detached pair of notes and the annotated transition. In *detached transitions* both the release of the first note and the attack of the second note are clearly identifiable and the excitation is interrupted. However, the notes are closely joined without a detectable rest between them. The transition is defined as the segment containing both the release section of the first note and attack section of the second note.

Rest-Separated Notes If a rest is detectable between consecutive notes, they will be referred to as *rest-separated notes* in the following. A rest segment does not only contain silence but may also include breathing, bowing noise, decaying partials, and reverberation. Rest-separated notes imply two independent transitions, framing the rest segment. Hence, transitions also mark the change from a note to a rest. As Fig. 2(c) shows, one transition marks the offset transient (release) of the first note, the second transition marks the onset transient (attack) of the second note.

Decaying Notes Notes do not necessarily feature a release segment if they are faded out gradually,

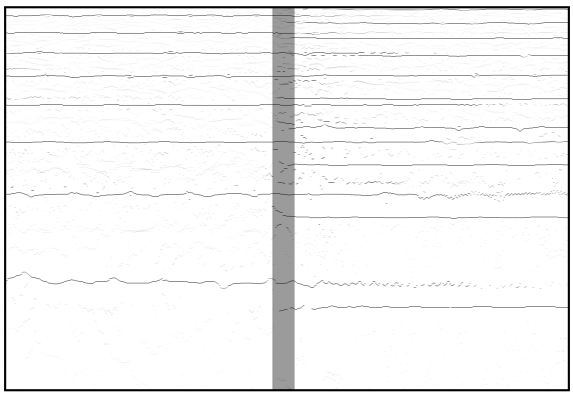


Fig. 3: Labeling example for a legato transition (gray segment) in the spectral representation of SV (t = horizontal, f = vertical)

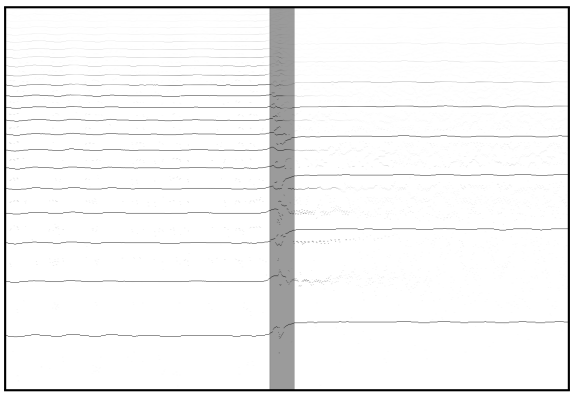


Fig. 4: Labeling example for a portamento transition (gray segment) with continuous partials in the spectrogram representation

especially in reverberant environments. Thus, the space between two notes can also consist of one rest and an onset transition, as shown in Fig. 2(d).

3. GROUND TRUTH LABELING

Based on the segment definitions above, this section outlines the tools and procedures to generate the segment annotations for the data set. Segmentation guidelines, as well as the choice and setup of the labeling environment were developed in a seminar on performance analysis, with graduate engineering students and musicologists. The segmentation and annotation was performed by the first author, using the free software Sonic Visualiser (SV) [28]. Leech-Wilkinson [29] explains the use of this tool in different applications of computer-aided musical analysis. The annotation data and the SV projects for all files can be downloaded from the project-related repository[2]. All ground truth data has been exported as tab-separated text files with two columns, of which the first contains time instants of segment changes (in seconds) and the second lists the associated segment-class, using the labels *0* for rest, *1* for note and *2* for transition segments.

3.1. Tools and Setup

Visual Cues The *peak frequency spectrogram* of SV, shown in Fig. 3, and the *waveform* representation, shown in Fig. 5, provided visual orientation for the segmentation. Both layers were aligned horizontally and could be used at the same time. The parametrization of the peak frequency spectrogram is important for the time and frequency resolution. In order to be able to annotate at maximum accuracy, the parameters were slightly adapted to the particular audio excerpt, using a window length of 1024 or 2048 frames (at a sample rate of 44.1 kHz), and an overlap ratio between 50 % and 93.75 %.
The peak frequency spectrogram allows the evaluation of the spectral content and the partials' trajectories. Changes in signal amplitude can best be determined in the waveform layer. The combination of both representations allows the accurate labeling of all transitions. The exact settings are included in the SV project files, which are part of the repository.

Auditory Cues Although the labeling based on visual cues usually allows a precise segmentation, each annotation has been confirmed by listening. If there are doubts about the correctness of a segment boundary, best practice is to start slow playback from the estimated segment boundary. Sonic Visualiser offers adequate setting of the playback rate for doing that. Audiovisual evaluation of the annotation was not possible due to the delay between graphic display and audio playback.

3.2. Segmentation Guidelines

The accuracy and consistence of the ground truth annotations define the usefulness of a dataset in

[2] http://sourceforge.net/projects/segmentationgt/

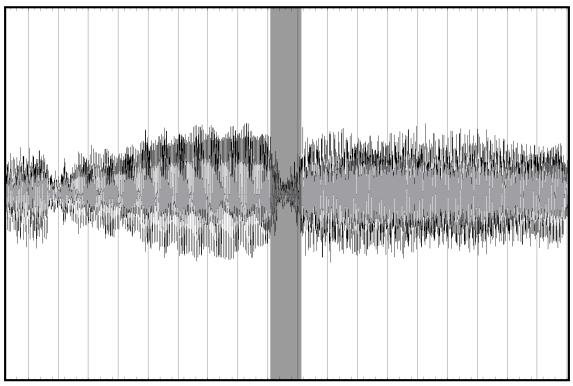


Fig. 5: Labeling example for a detached transition (dark gray) in the waveform representation

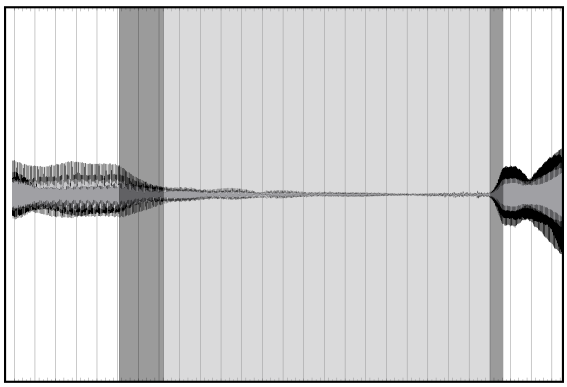


Fig. 6: Labeling example of rest-separated notes (white segments) in the waveform representation with transitions (dark gray) an rest (light gray)

practice. Since no generally approved set of rules exists for this purpose, we propose the following guidelines, based on the findings from Sect. 2 and previous experiments. As explained above, we use the segments *note*, *transition* and *rest*. Based on the segment definitions we establish the following basic rules:

1. Excerpts can start with either a rest or transition segment.
2. Every note has to start with a transition.
3. Notes can end either with or without a transition
4. Every passage is terminated with a transition label.

In addition to these basic rules, precise guidelines have been formulated for the labeling of the specific note transitions which were presented in Sect. 2.

Legato Transitions In legato transitions, the notes are linked very closely without a significant decrease of energy between them. The beginning of the second note triggers the end of the first note. The overall energy might change between the notes, but clear onsets and offsets can not be localized in the waveform. Hence, the waveform information is not useful. Instead, the transition is best visualized in the spectral representation, as shown in Fig. 3. The beginning of the transition is positioned slightly before the occurrence of the first partials from the second note. Setting the transition end is more problematic. Here, the estimated end of the second note's onset transient is chosen. Indicators for the transient region are a non-harmonic content, as well as apparent instability of the partials. Fig. 3 also shows how partials of the first note are still present in the following note. Such overlaps are caused by reverberation. This occurrence of multiple pitches in monophonic music is a well known problem in many analysis tasks [20].

Portamento Transitions Portamento transitions are a special case of legato. Two consecutive notes with different fundamental frequencies are connected without an interruption of the tone. This causes a gliding from the first note's pitch to the second, as shown in the partial trajectories in Fig. 4. Thus, the spectrogram is the best representation for annotation. The beginning of the gray transition segment is labeled as the beginning of the slope in the partials' frequencies. The end is marked by the termination point of the slope, respectively.

Detached Transitions Detached note transitions (sometimes referred to as *non-legato*), are caused by closely linked, yet clearly separated notes. Technically, these transitions are accompanied by an interruption of the excitation. The transition is characterized by a significant decrease of energy in the end of the first note and a significant increase in the beginning of the second note. However, there is no rest between the notes. A waveform representation as shown in Fig. 5 offers the best visual cues. The transition (gray) is labeled as the segment between

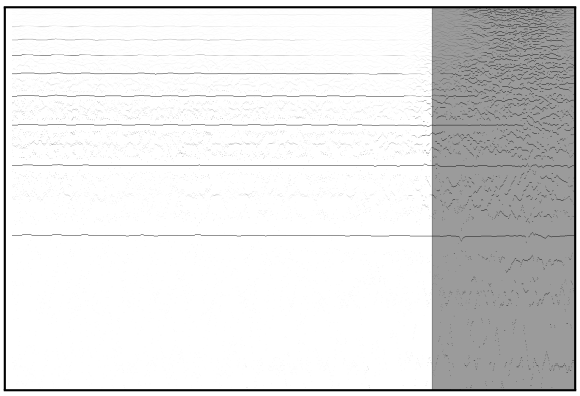


Fig. 7: Labeling example for a decaying note in the spectrogram representation. The white area marks the note - the gray area marks the rest segment

the beginning of the decrease and the end of the increase in energy.

Rest-Separated Notes Compared to legato and detached transitions, this type occurs less frequently in the chosen excerpts. The waveform visualization is most appropriate for labeling, as shown in Fig. 6. The offset transition matches the area of the most prominent decrease in energy of the first note, whereas the onset transition covers the area of the most positive section in the envelope of the second note.

Decaying Notes If a note is fading out slowly, a point has to be determined at which a change from note to rest occurs. This is not trivial since there is no detectable offset, especially if reverberation is involved. The peak frequency spectrogram might still show dominant partials, although they can hardly be perceived among the ambient noise of the recording. SV offers the possibility of normalizing the spectrogram column-wise, so that the signal to noise ratio is clearly visualized. Figure 7 shows the normalized spectrogram for a decaying note, as used for labeling these passages. The beginning of the rest segment (gray) is located at the point where the noise predominates the decaying partials.

4. THE DATA SET

4.1. Instruments in the Set

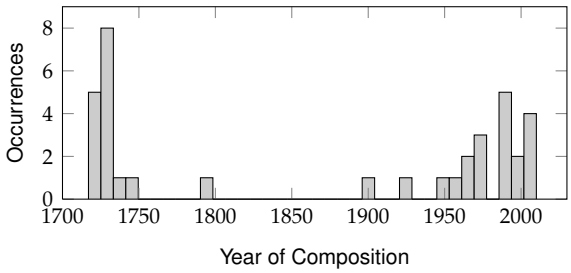


Fig. 8: Occurence of composition dates in the data set

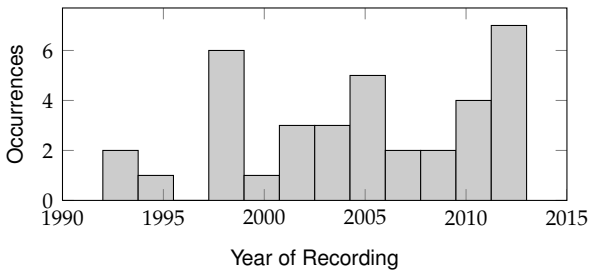


Fig. 9: Occurence of recording dates in the data set

The corpus of 36 items covers a sufficient variety of classical instruments from the established instrument groups *string*, *woodwind* and *brass*. Within each group, three instruments with different frequency ranges were selected:

- String group: double bass, cello, violin
- Woodwind group: bassoon, flute, oboe
- Brass group: trombone, horn, trumpet

Individual instruments are represented with four items in the data set, resulting in 12 audio excerpts per group. All instruments in the data set are driven by a continuous excitation and are hence called *excitation-continuous instruments* (ECI). ECIs are also classified as pitched non-percussive (pnp) instruments in the literature [30, 16]. They offer extended expressive possibilities and means of articulation, since the tone is directly influenced by the musician throughout its duration. The continuous excitation implies a direct contact to the oscillation, and thus makes it highly controllable. Performance analysis and note level segmentation is especially demanding for ECIs, since onset transients are weaker than in percussive instruments [31].

Table 1: Source compositions for the excerpts

Nr.	Composer	Piece	Nr.	Composer	Composition
String					
		Bass			
1	J.S. Bach	Cello Suite Nr.1-2, Allemande	2	J.S. Bach	Cello Suite Nr.1-5, Menuet II
3	W. Jentzsch	Sonata für Kontrabaß, II. Andante	4	G. Noeldeke	Skizze, für Kontrabass solo
		Cello			
5	J.S. Bach	Cello Suite Nr. 5, Sarabande	6	J.S. Bach	Cello Suite Nr. 2, Prelude
7	L. v. Beethoven	Cello Sonata 2, Opus 69 -1	8	B. Britten	Suite No. 3 For Cello, Lento Solenne
		Violin			
9	S. Prokofiev	Sonata for Solo Violin, Op. 115, 2	10	G.P. Telemann	12 Fantasias for Violin, Nr.9
11	G.P. Telemann	Fantasia No. 10 in D Major, II. Largo	12	J.S. Bach	Partita No.2, Allemande
Woodwind					
		Flute			
13	I. Loudova	Suite for Solo Flute, I. Adagio	14	G.P. Telemann	12 Fantasias (Flute), Nr 3
15	C.P.E. Bach	12 Fantasias (Flute), Nr 3	16	G.P. Telemann	12 Fantasias (Flute), Nr. 12
		Oboe			
17	G.P. Telemann	12 Fantasias (Oboe), Nr. 8	18	G.P. Telemann	12 Fantasias (Oboe), Nr. 2
19	L. Durey	Trois pièces brèves for oboe solo: Nr. 1,	20	B. Britten	Six Metamorphoses Op. 49, I.
		Bassoon			
21	S. Adler	Canto XII for Bassoon solo: Sacre Serenade	22	Persichetti	Parable for Solo Bassoon, Op 110
23	G.P. Telemann	12 Fantasias for Bassoon, Nr.6	24	O. Wilenski	Solo for Bassoon
Brass					
		Horn			
25	D. Lyon	Partita: IV. Aria	26	R. Baborak	12 Preludes for Horn, 10 (Largo)
27	J. Labor	Theme and Variations, Op. 10	28	J.S.Bach	Cello Suite No.3, Bourre 1,2
		Trumpet			
29	S. Beamish	Fanfare for Solo Trumpet	30	S. Adler	Canto I for Trumpet Solo
31	S. Wolpe	Solo Piece for Trumpet: I.	32	R. Holloway	Trumpet Sonata: II. Melody
		Trombone			
33	S. Adler	Canto 2 for Trombone , Slowly	34	J.S. Bach	Cello Suite, BVW 1008,4-Sarabande
35	N. Woud	Serenade	36	J.-M. Damase	Sonata for Trombone: III. Postlude

Table 2: Composition dates of source material

Nr.	Date	Nr.	Date	Nr.	Date	Nr.	Date	Nr.	Date	Nr.	Date	Nr.	Date	Nr.	Date	Nr.	Date
1	1717	**2**	1717	**3**	1966	**4**	2005	**5**	1717	**6**	1717	**7**	1797	**8**	1971	**9**	1922
10	1735	**11**	1732	**12**	1717	**13**	1959	**14**	1732	**15**	1747	**16**	1732	**17**	1733	**18**	1733
19	1974	**20**	1951	**21**	1992	**22**	1970	**23**	1727	**24**	2010	**25**	2010	**26**	2006	**27**	1899
28	1726	**29**	1999	**30**	1992	**31**	1966	**32**	1999	**33**	1992	**34**	1726	**35**	1992	**36**	1993

Table 3: Recordings and excerpts

Nr.	Artist	Rec.	ASIN	Excerpt (sec)	Nr.	Artist	Rec.	ASIN	Excerpt (sec)
1	E. Meyer	2000	B001UKAFZY	15.48 - 23.17	2	B. Salles	2012	B008IVMSTY	0.00 - 14.83
3	C. Ferenc	1998	B00CN197BC	0.00 - 21.59	4	G. Noeldeke	2011	B005Q5OW80	50.38 - 70.37
5	H. Schiff	2003	B001QGZG5Q	19.74 - 36.45	6	J. Starker	1991	B001SS9TEQ	1.07 - 17.03
7	M. Maisky	2007	B001SSDW44	1.05 - 13.54	8	T. Mork	2001	B001QL03OA	0.00 - 15.91
9	M. Kym	2001	B001SIJIMO	0.00 - 18.50	10	R. Podger	2002	B001S57N8I	0.00 - 16.59
11	O. Kam	2012	B007RK981I	0.00 - 13.20	12	O. Kagan	2005	B0030HPR7Q	0.00 - 13.98
13	K. Stoner	2009	B005P9CYXW	0.00 - 13.46	14	L. Zucker	2011	B006LPPTIG	0.41 - 19.35
15	E. Pahud	1993	B00296NAKK	35.02 - 44.07	16	B. Kuijken	2008	B001RSYS2U	48.89 - 68.20
17	H. Holliger	2010	B00584PFEE	0.00 - 15.62	18	V. Veverka	2013	B00B6B4D3W	0.00 - 17.43
19	L. Lencss	2003	B002S9YE9O	17.08 - 30.66	20	F. Leleux	1995	B00B4LG03A	20.52 - 36.19
21	J. Leclair	1998	B00B4LG03A	28.72 - 50.29	22	D.-Y. Kwon	2006	B002GLH90K	18.38 - 35.86
23	N.M. Jackson	2003	B002HXF6CU	0.00 - 12.49	24	B. Verde	2012	B00933H350	0.00 - 27.07
25	M. Stebleton	2010	B00BNXR7PE	0.00 - 30.45	26	R. Baborak	2006	B00421NDGG	0.00 - 26.08
27	G. Miller	2006	B00AOWOJXS	0.00 - 47.04	28	J.Lipton	2005	B002SS0HZ0	7.85 - 15.29
29	A. Mackie	2010	B00CTL5Y8Q	46.90 - 60.88	30	D. Bilger	1998	B001S4W758	0.00 - 35.36
31	R. Friedrich	1992	B002XZMVF2	0.00 - 06.83	32	H. Hardenberger	2006	B002PS7JIG	0.00 - 14.01
33	C. Vernon	1998	B001S506BY	0.00 - 26.28	34	C. Lindberg	1998	B002X1CQ74	29.51 - 40.05
35	I. Petry	1998	B002WM2YL2	38.75 - 54.17	36	A. Karafezliev	2012	B007J9FUVO	0.00 - 24.48

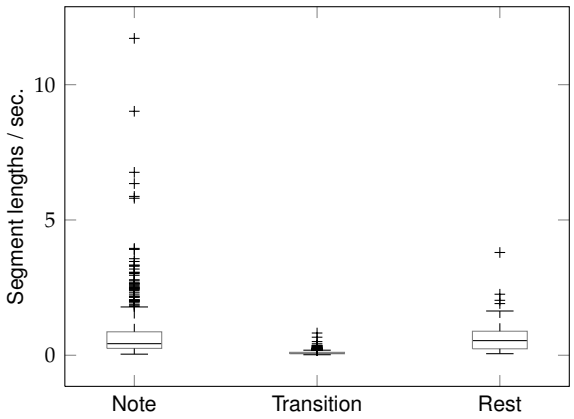


Fig. 10: Boxplot with individual lengths of the three segments (for all items)

4.2. Musical Content

All 36 excerpts are taken from western art music (Tab. 1), with compositions dating from the early 18th century until today. The Baroque period and works from the 20th century are predominant, as shown in Tab. 2 and Fig. 8. A gap is caused since few suitable excerpts from the 19th century were found.

Excerpts were chosen to be in a moderate tempo, trying to avoid very short notes. This allows a precise data annotation. The shorter and manifold the notes, the more problems occur in the labeling process. The chosen content allows the musician to apply various expressive gestures. Although all excerpts were chosen to be played mostly in moderate tempo, it was also taken care that articulation style and note lengths slightly vary throughout the whole set.

All items in the set were recorded between 1991 and 2013, as listed in Tab. 2 and visualized in Fig. 9. Thus, all files benefit from modern recording technology and the set is not representative for historic recordings. All files are encoded in the MP3 format with the corresponding implications for signal quality. Since all excerpts are taken from different recording sessions, the whole set contains a large variety of acoustical scenes and technical setups. As usual for solo performances, most recordings contain a significant amount of reverberation. In contrast to data sets generated in studio or laboratory settings, the CMMSD is thus more demanding in this aspect. It thus represents the predominant conditions of applied performance analysis.

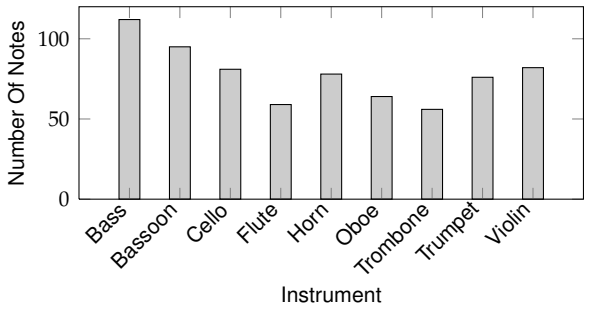


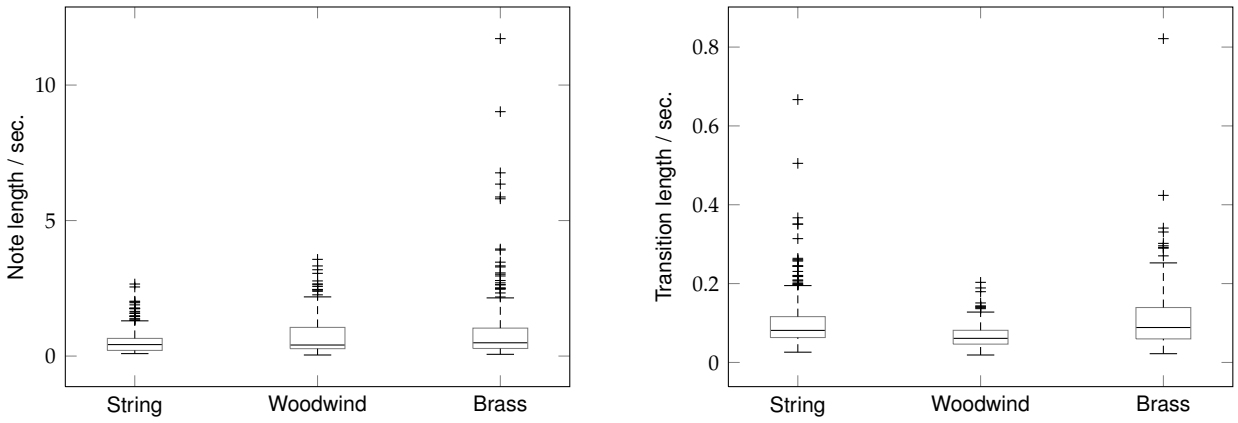
Fig. 11: Number of notes in all items for the single instruments in the data set

Since all items originate from commercial recordings with professional musicians, a certain minimal quality of the performance can be assumed. No artist is represented more than once. This allows for various personal articulation styles and techniques to be included. For allowing a quick access to the audio data, the unique Amazon identifier (ASIN) is listed for all recordings in Tab. 3.

4.3. Data Set Statistics

Excerpts in the presented set have a mean length of 17.97 seconds and an average of 20 notes. The data set contains a total of 718 notes, 736 transitions and 80 rests. Observed among all items, the segments *note*, *rest* and *transition* show intrinsic temporal properties. All note segments together have a length of 523.64 seconds, all rest segments occupy 54.44 seconds and the entirety of transition segments has a duration of 68.97 seconds. This imbalance has to be taken into account when used as a training set. Figure 10 shows box-plots with the duration of the individual segments, calculated from the complete ground truth of all instruments. With a mean length of 94 ms, transitions are significantly shorter than the other segment types. Transitions longer than 200 ms are usually caused by portamento. Notes have a mean length of 0.73 seconds, rests a mean length of 0.68 seconds. Long rests are usually caused by silence in the very beginning of the excerpts.

Of all 718 notes in the set, 275 notes belong to string recordings, 232 to woodwind recordings and 211 to brass recordings. Thus, there is no serious imbalance between the instrument groups. The bar chart displayed in Fig. 11 visualizes the distribution of



(a) Note lengths for instrument groups

(b) Transition lengths for instrument groups

Fig. 12: Individual segment lengths (note, transition) for the three instrument groups

the number of notes per individual instrument.
Figure 12 shows the individual segment durations of notes and transitions for the instrument groups. Especially woodwind instruments tend to have shorter transitions than the other groups, with a mean length of 60 ms. This is likely to be caused by the rather short attack transient time of reed instruments (bassoon, oboe). String instruments, on the other hand, are well known to have long attack transients [31]. String transitions have a mean length of 102 ms, brass transitions a mean length of 111 ms. However, all these observations may be biased by different reverberation times in the recordings and the musical content.

5. CONCLUSION

With the CMMSD, a data set for note-level segmentation of monophonic music has been presented. The ground truth partitions the audio data into the segments note, rest, and transition. The set contains carefully selected excerpts of commercial recordings of pieces from different epochs. It covers an appropriate variety of instruments, artists and acoustic environments to include different types of transitions and articulations. The focus of the data set is on segmentation and performance analysis, but it should be suited for other applications as well.
Future work will include the extensive evaluation of features for the segmentation, in order to determine suitable features for the use in machine learning algorithms. We plan to publish the extracted features as extensions of the data set. Preliminary experiments with hidden Markov models showed the promising applicability of the set for training and evaluation. The findings of these experiments will be subject to a following publication.
Based upon the proposed segmentation, a further extension of the segment classes might be practical. Dedicated labels for different transition types might increase the applicability of the set, as well as a definition of different rest classes. Future work will also include the extension of the data set with ground truth for other analysis tasks in monophonic music. A ground truth for onset detection is currently being created. Eventually, annotations on note transition styles and vibrato parameters might be added, too.

By choosing easily accessible audio files and creating an open source ground truth, the CMMSD is specifically designed to be used by researchers with similar research questions. Although the set is ready for use, it is also intended as a starting point for discussions. Interested researchers are welcome to contribute to the project and share suggestions on improvements of the ground truth. Thus, a generally accepted data set for the segmentation and analysis of monophonic music can be established.

Bibliography

[1] R.J. McNab and L.A. Smith. Evaluation of a Melody Transcription System. In *IEEE International Conference on Multimedia and Expo (ICME)* , volume 2, pages 819–822, 2000.

[2] Roger B. Dannenberg and Ning Hu. Polyphonic Audio Matching for Score Following and Intelligent Audio Editors. In *Proceedings of the International Computer Music Conference*, pages 27–34, 2003.

[3] D. Schwarz. Corpus-Based Concatenative Synthesis. *Signal Processing Magazine, IEEE*, 24(2):92–104, March 2007.

[4] Marco Fabiani. *Interactive Computer-Aided Expressive Music Performance*. PhD thesis, KTH School of Computer Science and Communication, 2011.

[5] N. Orio and F. Déchelle. Score Following Using Spectral Analysis And Hidden Markov Models. In *Proceedings of the International Computer Music Conference*, pages 151–154, 2001.

[6] Robert C. Maher. Control of Synthesized Vibrato during Portamento Musical Pitch Transitions. *Journal of the Audio Engineering Society*, 56(1/2):18–27, 2008.

[7] Hee-Suk Pang and Doe-Hyun Yoon. Automatic Detection of Vibrato in Monophonic Music. *Pattern Recognition*, 38(7): 1135–1138, 2005.

[8] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.-L. Collette. Vibrato: Detection, Estimation, Extraction, Modifcation. In *Digital Audio Effects Workshop*, pages 175–179, 1999.

[9] Eric Prame. Measurements of the Vibrato Rate of Ten Singers. *Journal of the Acoustical Society of America*, 96(4): 1979–1984, 1994.

[10] Perfecto Herrera and Jordi Bonada. Vibrato Extraction and Parameterization in the Spectral Modeling Synthesis Framework. In *Proceedings of the Digital Audio Effects Workshop (DAFX)*, pages 107–110, 1998.

[11] Ixone Arroabarren, Miroslav Zivanovic, José Bretos, Amaya Ezcurra, and Alfonso Carlosena. Measurement of Vibrato in Lyric Singers. *IEEE Transactions on Instrumentation Measurement*, 51(4):1529–1534, 2002.

[12] Emilia Gomez and Esteban Maestre. Automatic Characterization of Dynamics and Articulation of Expressive Monophone Recordings. In *Audio Engineering Society Convention 118*, 5 2005.

[13] J. Strawn. *Modeling Musical Transitions*. Ph.d., Stanford University, 1985.

[14] A. Friberg, E. Schoonderwaldt, and P. N. Juslin. CUEX : An Algorithm for Automatic Extraction of Expressive Tone Parameters in Music Performance from Acoustic Signals. *Acta Acoustica united with Acustica*, 93(3):411–420, 2007.

[15] J. Abesser, H. Lukashevich, and G. Schuller. Feature-Based Extraction of Plucking and Expression Styles of the Electric Bass Guitar. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 2290–2293, 2010.

[16] Chee Chuan Toh, Bingjun Zhang, and Ye Wang. Multiple-Feature Fusion Based Oset Detection for Solo Siniging Voice. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 515–520, 2008.

[17] J.P. Bello, L. Daudet, S. Abdullah, C. Duxbury, M. Davies, and M. Sandler. A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing*, 13 (5):1035–1047, 2005.

[18] P. Leveau and L. Daudet. Methodology and Tools for the Evaluation of Automatic Onset Detection Algorithms in Music. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)*, pages 72–75, 2004.

[19] Giovanni De Poli and Luca Mion. *From Audio to Content*, chapter 4. 2006. Dipartimento di Ingegneria Dell'Informazione - Università degli Studi di Padova.

[20] Diemo Schwarz. *Data-Driven Concatenative Sound Synthesis*. PhD thesis, Université Paris 6 – Pierre et Marie Curie, 2004.

[21] Mark Robert Every. *Separation of Musical Sources and Structure from Single-Channel Polyphonic Recordings*. PhD thesis, University of York, 2006.

[22] J.C. Risset and M.V. Mathews. Analysis of Musical Instrument Tones. *Physics Today*, 22(2):23–30, 1969.

[23] Paul Masri. *Computer Modelling of Sound for Transformation and Synthesis of Musical Signals*. PhD thesis, University of Bristol, 1996.

[24] G. Peeters. A Large Set of Audio Features for Sound Description. Technical report, IRCAM, Paris, 2004.

[25] J. Devaney, M. I. Mandel, D. P. W. Ellis, and I. Fujinaga. Automatically Extracting Performance Data from Recordings of Trained Singers. *Psychomusicology: Music, Mind & Brain*, 21(1/2):108, 2011.

[26] Giuliano Monti and Mark Sandler. Monophonic Transcription with Autocorrelation. *Conference on Digital Audio Effects (DAFX)*, pages 257–260, 2000.

[27] Ivan Bruno and Paolo Nesi. Automatic Music Transcription Supporting Different Instruments. *Journal of New Music Research*, 34(2):139–149, 2005.

[28] Sonic Visualiser Website, visited 2013. URL `http://www.sonicvisualiser.org`.

[29] Daniel Leech-Wilkinson. *The Changing Sound of Music: Approaches to Studying Recorded Musical Performances*. CHARM, 2009.

[30] Nick Collins. A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions. In *Audio Engineering Society Convention 118*, pages 28–31, 2005.

[31] Nicholas M. Collins. *Towards Autonomous Agents for Live Computer Music: Realtime Machine Listening and Interactive Music Systems*. PhD thesis, University of Cambridge, 2006.