Time-based and Note-based Criteria for Assessment of Music Transcription Quality

Jiří Vass

Czech Technical University
Faculty of Electrical Engineering, Department of Circuit Theory
Technická 2, 166 27 Prague, Czech republic
Tel: (+420) 2 2435 2286
E-mail: vassj@fel.cvut.cz

ABSTRACT

This work is concerned with automatic transcription of monophonic music into the MIDI (Musical Instrument Digital Interface) representation. Specifically, this paper presents a new criteria for evaluating the performance of music transcription systems. The criteria are divided into two separate groups. First, time-based criteria assess the quality of onset and offset detection, and thus quantify the accuracy of time-domain segmentation of the audio signal to be transcribed. These criteria are inspired by evaluation of Voice Activity Detectors (VAD) used in speech processing. Second, note-based criteria evaluate the overall transcription quality using notes as independent units. In particular, notes are classified into several categories according to correctness of fundamental frequency detection and a mutual time overlap between each pair of reference note and transcribed note. This overlap is characterized by two complementary measures: ROT (Reference Overlap Transcription) and TOR (Transcription Overlap Reference). Finally, application of the proposed criteria is shown on a transcription example obtained from real audio recording.

Key Words: transcription of music, monophonic audio, onset detection, fundamental frequency tracking, pitch detection

1 Introduction

Automatic transcription of music is a task of converting a particular piece of music into symbolic representation by means of a computational system. Symbolic representation is generally depicted using the standard music notation which consists of notes characterized by a specific frequency and duration. From the transcription point of view, music can be classified as polyphonic and monophonic. The former consists of multiple simultaneously sounding notes, whereas the latter contains only a single note at each time instant, such as a saxophone solo or singing of a single vocalist.

Transcription of music is related to several fields of science, including musicology, psychoacoustics, and Computational Auditory Scene Analysis (CASA). It belongs to the discipline of *music content analysis* concerning various audio tasks, such as rhythm analysis, instrument recognition, and sound separation. As conversion to symbolic representation significantly reduces the amount of data, music transcription can also be used for compression purposes and hence offers an alternative method to standard audio codecs, such as MPEG-1 Layer 3, e.g. [8].

The state-of-the-art in music transcription is focused on the polyphonic transcription, since the monophonic transcription is considered as practically solved [10], [11]. However, it represents an important case which should be treated separately with much stricter demands on the transcription quality, which still seems to be relatively limited for polyphonic transcribers. Extensive review of published polyphonic systems can be found in [10]. Since monophonic music share various properties with speech, many algorithms suitable for music transcription purposes originate in speech processing, e.g. [12]. Recent works in monophonic music transcription explore the potential of the wavelet transform [6], [9], time-domain techniques based on autocorrelation [2], and probabilistic modelling using Hidden Markov Models [15]. In addition, Bořil [3] developed a simple and robust algorithm for real-time MIDI conversion. This system performs separate monophonic analysis of a signal from each guitar string, and therefore illustrates that monophonic transcribers can be used in special polyphonic transcription systems. Modification of this algorithm, referred to as DFE (Direct Time Domain Fundamental Frequency Estimation), was later successfully applied on speech in noisy conditions [4].

This work presents a criteria for evaluating the performance of music transcription systems. The criteria represent an attempt to numerically quantify transcription results, since pure listening and word commentary is insufficiently informative. Ryynänen [15] published other useful criteria with some properties similar to those described here. The paper is organized as follows. Section 2 describes time-based criteria of transcription assessment. Section 3 presents overlapping measures between a reference and a transcription. Section 4 introduces criteria based on notes. Section 5 provides a transcription example in order to illustrate usefulness of the criteria. The conclusion is given in Section 6.

2 Time-based criteria

Time-based evaluation is inspired by the criteria in [13], which were partially adopted from [14]. These criteria originate in speech processing and were designed to evaluate the performance of Voice Activity Detectors (VAD). In our context, they serve to assess the accuracy of onset and offset detection, which represents an analogy to the task of speech segmentation. Therefore, analogous criteria can simply be obtained by substituting the terms *Front* and *Back* of a speech activity by the note *Onset* and *Offset*, respectively.

OON	Overlap at Onset
OOF	Overlap at Offset
TON	Truncation at Onset
TOF	Truncation at Offset
ROT	Reference Overlap Transcription
TOR	Transcription Overlap Reference
CTN	Completely Transcribed Notes
PTN	Partially Transcribed Notes
FER	Frequency Errors
OER	Octave Errors
MIN	Missed Notes
FAN	False Notes
NDA	Note Detection Accuracy

Table 1: Terminology of the proposed criteria

Table 1 summarizes abbreviations and names of all proposed criteria. Since it is somewhat difficult to mathematically define the time-based criteria, Fig. 1 provides an illustrative example. As can be seen, each picture shows the "Piano Roll" representation [16] of the *reference note* (gray) and the corresponding *transcribed note* (black). In all four cases, the frequency of both notes is 440 Hz corresponding to the MIDI note number 69. Note that transcribed notes are narrower only to be visually distinguishable from reference notes.

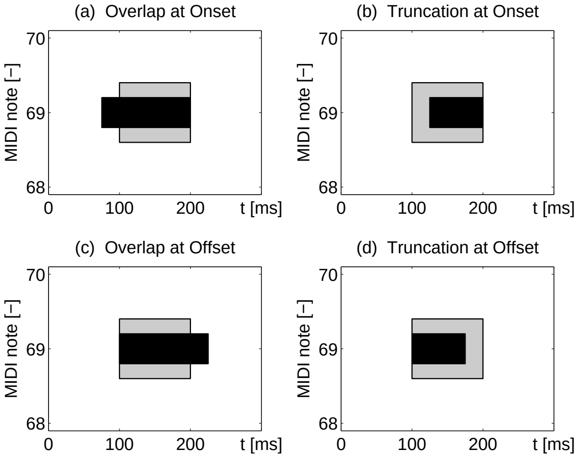

Figure 1: Illustration of time-based criteria

Fig. 1(a) and 1(b) depict the *onset errors* OON and TON, respectively, which represent too early and too late detection of a note onset, whereas Fig. 1(c) and 1(d) depict the *offset errors* OOF and TOF, respectively, which represent too late and too early detection of a note offset.

Each criterion counts the number of samples between the reference and the detected event (i.e. onset or offset) and sums the contributions from all notes to obtain the total amount of a particular error in the transcription. Alternatively, this amount can be divided by the total duration of notes in the reference MIDI recording to obtain the proportional time error in percentage (as in Section 5).

It should be emphasized that these criteria also add contributions from the notes transcribed with a frequency error, provided that certain minimum time overlap is satisfied. The meaning of the overlap is clarified in the following text.

3 Overlapping measures

This section introduces two overlapping measures expressing the mutual time overlap between each pair of reference note and transcribed note. When used in conjunction with correct pitch detection, these measures indicate the overall transcription correctness bilaterally in the context of a reference and its transcription. Therefore, these measures are applied in Section 4 for classification of notes into specific categories constituting the note-based criteria.

Section 3.1 presents *Reference Overlap Transcription* (ROT) measure, whereas Section 3.2 describes *Transcription Overlap Reference* (TOR) measure. Definitions of both measures are given in words only, supported by examples displayed in Fig. 2.

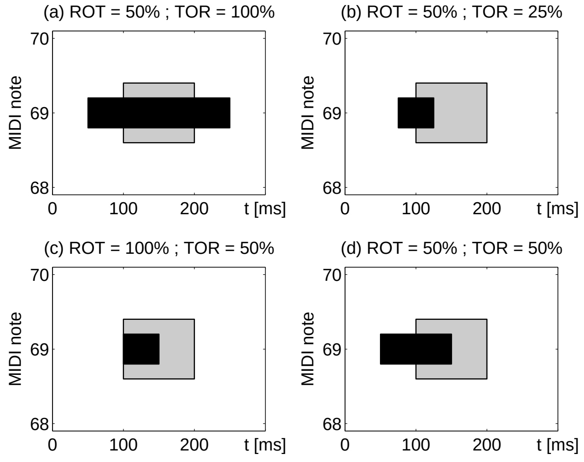

Figure 2: Overlapping measures ROT and TOR

3.1 Reference Overlap Transcription

This parameter indicates how much reference note covers the transcribed note, i.e. how large portion of the transcribed note is indeed correct. If ROT is greater than 75%, for instance, it means that the transcribed note is correct in itself. Although it may not entirely represent the reference note, it at least corresponds to its certain portion, and thus increases the transcription quality. If ROT = 100%, it merely means that the transcribed note is completely located inside the reference note (see Fig. 2(c)), but does not indicate the transcribed area of the reference (which is the purpose of the complementary measure TOR). It is important to mention that ROT penalizes both too long and too short transcriptions, as demonstrated in Fig. 2(a) and 2(b), respectively.

3.2 Transcription Overlap Reference

This parameter indicates the percentage of the reference note covered by the transcribed note, i.e. how large portion of the reference is correctly detected. If TOR = 100%, it only means that the transcribed note entirely overlaps the reference note, though the ideal result can be a subset of the transcribed note (see Fig. 2(a)).

4 Note-based criteria

This section presents criteria performing the evaluation using the notes as independent units. Although this approach share some ideas with [15], our note-based evaluation is approached mainly from the reference notes perspective and compensates this imperfection by penalizing the transcriber for errors. On the other hand, the symmetry is in fact embedded in our conception as well, since the overlapping measures characterize the bilateral relationship between reference and transcribed notes.

The note-based evaluation classifies each reference-transcription pair into one of the note categories, according to the measures ROT and TOR, as well as the correctness of frequency detection. Each criterion then simply counts the number of notes in the respective note category defined in the sequel.

4.1 Completely Transcribed Notes

A reference note is regarded as *completely transcribed* (CTN) by a note from the transcription, when the MIDI note frequencies agree and the notes exhibit large overlap in time:

$$f_{mid}^{ref} = f_{mid}^{trn} \tag{1}$$

$$\begin{gathered}\mathrm{ROT} + \mathrm{TOR} \geq 150\%,\\ \mathrm{ROT} \geq 60\%, \quad \mathrm{TOR} \geq 60\%\end{gathered} \tag{2}$$

Three joint conditions in Eq. (2) appear to be more flexible and yield better results than simple requirement of ROT or TOR to exceed a specific minimum value. Indeed, all four pairs in Fig. 1 satisfy the above conditions, resulting in classification of the reference notes as CTN. On the other hand, the examples in Fig. 2(a) and 2(c) violate the condition either for ROT or TOR (though satisfying the most difficult condition for the sum), and thus fall to the following category PTN.

4.2 Partially Transcribed Notes

A reference note is considered *partially transcribed* (PTN) by a transcribed note when the frequency condition (1) is met and the notes satisfy less demanding overlap than CTN:

$$\mathrm{ROT} + \mathrm{TOR} \geq 100\%, \quad \mathrm{TOR} \geq 40\% \tag{3}$$

As suggested by an example PTN in Fig. 2(d), this approach may result in a somewhat undervalued score, since listeners would probably regard this transcription as correct due to a hardly perceivable time shift. On the other hand, monophonic transcription is a significantly simpler task compared to polyphonic transcription, hence the quality demands should be much stricter. Moreover, such errors become considerably more audible with increasing note duration and decreasing tempo of the recording.

4.3 Frequency Errors

A reference note is classified as transcribed with a *frequency error* (FER), when the notes exhibit time overlap defined in Eq. (2) or (3), but Eq. (1) is not fulfilled.

4.4 Octave Errors

Octave errors (OER) represent a special case of a frequency error greater than or equal 12 semitones:

$$\left| f_{mid}^{ref} - f_{mid}^{trn} \right| \geq 12 \tag{4}$$

4.5 Missed Notes

A reference note is classified as *missed* (MIN) when no appropriate transcription candidate exists, or when the candidate is too inaccurate in time, regardless of the error in frequency detection. Specifically, the MIN criterion counts the references notes not identified by the previous criteria CTN, PTN, FER or OER.

4.6 False Notes

This criterion counts the notes in the transcribed MIDI sequence not involved in the original recording. In other words, *false notes* (FAN) are constituted by redundantly transcribed notes coupled with no reference note. In addition to that, a transcribed note is considered false (FAN) whenever the corresponding reference note is classified as missed (MIN). An illustrative example of this situation is depicted in Fig. 2(b).

4.7 Note Detection Accuracy

Based on the preceding note criteria, we can characterize the overall quality of transcription by introducing the *Note Detection Accuracy* (NDA) parameter defined as:

$$NDA = \frac{\mathrm{CTN} + \mathrm{PTN}/2 - 2{\cdot}\mathrm{OER} - \mathrm{FAN}}{N} \tag{5}$$

where N is the total number of reference notes. As can be observed, this parameters takes into an account both the reference and the transcription point of view. While the former is represented by the CTN, PTN and OER criteria, the latter is described by the FAN criterion. FER errors are not penalized for two reasons. First, classification of a particular note as FER is automatically reflected as a proportional decrease in CTN or PTN criterion. Second, FER errors express the correctness of onset/offset detection. On the other hand, octave errors OER are strongly penalized since they symbolize gross errors causing especially unpleasant impression of the transcribed melody.

5 Transcription example

This section presents an example of a transcribed audio signal and discusses the results based on the proposed criteria.

Transcription was obtained using the automatic system described in [17], [18]. This system incorporates two separate algorithms in order to extract the necessary musical information from the audio signal. Detection of the fundamental frequency is based on a pattern recognition method [5] applied on the Constant Q Transform (CQT) of a signal. Onset detection is achieved by a sequential algorithm [1] based on a statistical distance measure (Kullback's divergence) between two autoregressive (AR) models. The results of both algorithms are combined by heuristic rules eliminating the transcription errors. Proper settings of algorithm parameters are given in [18].

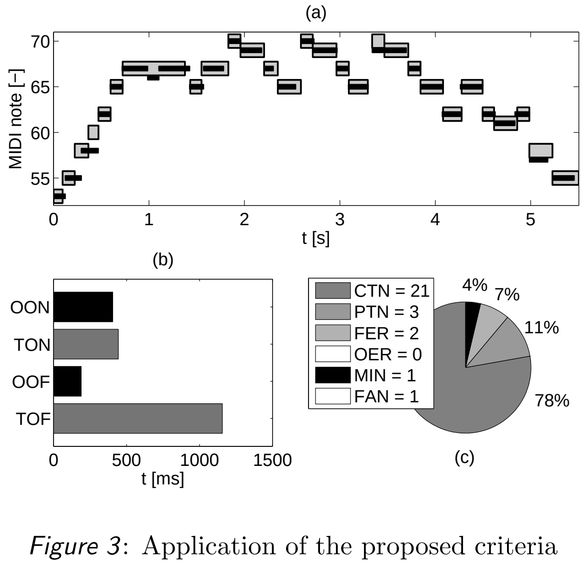

Figure 3: Application of the proposed criteria

The audio recording to be transcribed is a part of a blues solo played on the alto saxophone. Contrary to wind instruments with a relatively pure sound (such as horn and flute), saxophones are characterized by a harmonically rich spectrum containing approximately 14 harmonics. The recording was manually transcribed in order to obtain reference musical notation. First of all, note onsets and offsets were labeled similarly as in speech processing. Then, correct MIDI notes were detected by repeated listening, and the reference MIDI file was generated using the software by [7]. Finally, the resulting MIDI file was played several times to adjust the offsets in such a way, that all notes sound as closely as possible as the original instrument.

As can be seen in Fig. 3, three notes were detected partially, two frequency errors occurred, one false note was generated, and majority of notes were transcribed correctly (CTN = 21). Time criteria reveal that note offsets tend to be truncated (TOF = 16.35%) rather than overlapped (OOF = 4.78%). Large value of TOF is caused by a single note (classified as PTN), whereas OOF is formed by comparable contributions of several notes. Generally, OOF expresses a delay in offset detection, caused in our system by a simple method based on thresholding of signal power. Despite some evident drawbacks, the transcription result can be considered relatively successful (NDA = 76.8%). More transcription examples can be found in [18].

6 Conclusion

This paper introduces two groups of criteria for evaluating transcription systems of monophonic music.

The first group comprises four time-based criteria in order to assess the detection of note onsets and offsets. These criteria measure the amout of overlap or truncation at the beginnings and endings of individual notes. For this reason, time-based criteria consist of Overlap at Onset (OON), Truncation at Onset (TON), Overlap at Offset (OOF), and Truncation at Offset (TOF).

The second group is based on classification of notes into specific categories, thus denominated as note-based criteria. This is achieved by definition of two overlapping measures: Reference Overlap Transcription (ROT) and Transcription Overlap Reference (TOR). The value of both measures determine whether a reference note is classified as Completely Transcribed (CTN) or Partially Transcribed (PTN). Notes with incorrectly detected MIDI frequency are regarded as Frequency Errors (FER) or Octave Errors (OER); missing and redundant notes are classified as Missed Notes (MIN) and False Notes (FAN), respectively. Overall transcription quality is expressed as Note Detection Accuracy (NDA).

Finally, proposed criteria are applied on a musical recording in order to evaluate results of automatic transcription. Although proved only on a small number of audio signals, the criteria appear to successfully describe the transcription quality in general. Extensive testing was not performed due to absence of a representative database of monophonic music.

Acknowledgement

This work has been supported by the research program MSM6840770012 "Transdisciplinary Research in Biomedical Engineering II" and the GA ČR grant No. 102/03/H085 "Biological and Speech Signal Modelling". This research was supervised by Hadas Ofir, Signal and Image Processing Laboratory (SIPL), Technion, Israel and Ing. Radim Špetík, Department of Circuit Theory, FEE, CTU. The help of Doc. Ing. Petr Pollák, CSc. and Ing. Ladislava Smítková Janků is much acknowledged. This paper is dedicated to musicians from bands Jahzmuzik and ChKO.

References

[1] Basseville, M. - Benveniste, A.: *Sequential Detection of Abrupt Changes in Spectral Characteristics of Digital Signals*, IEEE Trans. on Information Theory, Vol. 29, No. 5, pp. 709-724, 1983

[2] Bello, J. P. - Monti, G. - Sandler, M. B.: *Automatic Music Transcription and Audio Source Separation*, Cybernetics and Systems: An International Journal, 33: pp. 603-327, 2002

[3] Bořil, H.: *Pitch Detector for Guitar Midi Converter*, Proc. Poster 2003 [CD-ROM], CTU FEE Prague, 2003

[4] Bořil, H. - Pollák, P.: *Direct Time Domain Fundamental Frequency Estimation of Speech in Noisy Conditions*, Proc. of EUSIPCO 2004, Wien, Austria, 2004, Vol. 1, pp. 1003-1006

[5] Brown, J.C.: *Musical Fundamental Frequency Tracking using a Pattern Recognition Method*, Journal of Acoustical Society America, Vol. 92, pp. 1394-1402, Sep. 1992

[6] Cemgil, A.T. - Anarim, E. - Caglar, H.: *Comparison of Wavelet Filters for Pitch Detection of Monophonic Music Signals*, European Conference on Circuit Theory And Design, Vol. 2, pp. 711-14, 1995

[7] Cemgil, A.T.: *Software to Download: Midi Toolkit*, [online], [cit. November 2005] `http://staff.science.uva.nl/˜cemgil`

[8] Herrera, M.: *Summary of the Subjective Audio Coding Tests during the period 2003-2005 at the ČVUT*, Proceedings of ISSET 2005, Krakow, Poland, 2005

[9] Jehan, T.: *Musical Signal Parameter Estimation*, Msc. Thesis, CNMAT, Berkeley, 1997

[10] Klapuri A.: *Automatic Transcription of Music*, MSc. Thesis, Tampere University of Technology, April 1998

[11] Martins, L.G.: *PCM to Midi Transposition*, MSc. Thesis, Universidade do Porto, 2001

[12] Medan Y., Yair E., Chazan, D.: *Super Resolution Pitch Determination of Speech Signals*, IEEE Trans. on Signal Processing, Vol. 39, No. 1, Jan. 1991

[13] Pollák P.: *Criteria for VAD Classification*, Internal Research Report R02-1, Czech Technical University in Prague, Dec. 2002

[14] Rosca J. - Balan R. - Fan N.P. - Beaugeant C. - Gilg V.: *Multichannel Voice Detection in Adverse Enviroments*, Proc. of EUSIPCO 2002, Toulouse, France, Sep. 2002

[15] Ryynänen M.: *Probabilistic Modelling of Note Events in the Transcription of Monophonic Melodies*, MSc. Thesis, Tampere University of Technology, March 2004

[16] Twelve Tone Systems, Inc: *Cakewalk MIDI Software (ver.9)*, [online], [cit. Nov. 2005] `http://www.cakewalk.com`

[17] Vass, J. - Ofir, H.: *Automatic Transcription of Monophonic Audio to MIDI*, Proc. Poster 2004 [CD-ROM], CTU FEE Prague, 2004

[18] Vass, J.: *Automatic Transcription of Audio Signals*, Master Thesis, 60 pp., CTU FEE Prague, 2004