EVALUATING NON-ALIGNED MUSICAL SCORE TRANSCRIPTIONS WITH MV2H

Andrew McLeod
Kyoto University
mcleod@sap.ist.i.kyoto-u.ac.jp

Kazuyoshi Yoshii
Kyoto University
yoshii@kuis.kyoto-u.ac.jp

EXTENDED ABSTRACT

MV2H (for **M**ulti-pitch, **V**oice, **M**eter, note **V**alue, and **H**armony) [4] was introduced to evaluate "complete" automatic music transcription (AMT) systems which transcribe an input audio (or MIDI) piece into a musical score, complete with voice (or instrument) separation, a time signature and metrical structure, note values, and harmonic information such as a key signature and chord progression. Its design was based on the principal of *disjoint penalties*: that a single transcription error should only be penalized once, even if that error causes multiple mistakes (e.g., an error in time signature also causing note value errors).

Traditionally, multi-pitch detection systems (arguably the most important step for an AMT system) have aligned their output in time with the input musical piece (e.g., [3]), and while complete AMT systems are uncommon, the design of MV2H relied on such an alignment being present both for the transcription and the ground truth. However, recently a few complete AMT systems have been proposed which skip this alignment and instead output a musical score (or at least a text-based representation of a musical score) directly in an end-to-end fashion [1, 7], making them impossible to evaluate with MV2H. Cogliati and Duan [2] describe a metric for evaluating complete transcriptions of non-aligned musical scores, using dynamic time warping (DTW; [8]) for alignment. Their metric, however, does not share our principal of disjoint penalties. Nakamura et al. [5] describe a method to evaluate multi-pitch and rhythmic aspects of a non-aligned transcription, using an existing alignment method [6].

This paper describes an alignment algorithm based on DTW to allow for such non-aligned transcriptions to be evaluated with MV2H. This also enables digital musical scores (such as MusicXML files, for which a parser is included) to be used as the ground truth without requiring any additional manual alignment or annotation. Furthermore, we extend MV2H to be able to evaluate transcriptions and ground truths which may contain key and time signature changes.

To perform an alignment between two musical scores, we first consider each musical score (transcribed and ground truth) as a sequence of chords, where each chord is a set of the notes which share an onset position. These sequences are then aligned using DTW, where the distance between two chords is defined as the F-measure between its notes, regarding only pitch: a matched note is a true positive, an unmatched transcribed note is a false positive, and an unmatched ground truth note is a false negative. When a chord contains two notes of identical pitch (for example, in different instruments), each must match with a different note to count as a true positive. We set the penalty for an insertion or deletion to 0.6. This makes the algorithm align chords which share no notes (distance of 1.0 per chord pair) rather than have a series of insertions and deletions (distance of 1.2 per chord pair), ensuring a more linear alignment. It is small enough that the distance between two chords dominates this penalty, ensuring that two chords that share even a single note are still aligned if possible.

After running this DTW, we treat aligned chords as anchor points, and set the relative timing for non-aligned musical features (chords, as well as bars, beats, sub beats, key signatures, chord symbols, etc.) based on the proportional distance between the surrounding anchor points, assuming constant tempo. The relative timing of musical features which lie after the final anchor point or before the first anchor point are set based on the preceding or succeeding anchored section, respectively. Once two musical scores are aligned, the thresholds for matching their note onsets, durations, and metrical groupings are all set to 0 ms (such that they must align exactly to count as a match).

MV2H uses the standard key detection evaluation, used in MIREX.[1] It assigns a score of 1.0 to the correct key, 0.5 to a key which is a perfect fifth off, 0.3 to the relative major or minor of the correct key, 0.2 to the parallel major or minor of the correct key, and 0.0 otherwise. The new version, which allows key changes, evaluates each *continuous key section* (a section of a piece which contains no time signature changes in the transcription and the ground truth) with the standard metric. The final key evaluation is calculated as the sum of all of those scores, weighted by the proportion of the piece for which the given score is assigned.

The metrical structure of a transcription is evaluated in MV2H using metrical F-measure. The metric is based on groupings at the bar, beat, and sub-beat level (because these three levels exactly define a time signature), where a grouping is represented by its start and end time. Transcribed groupings are compared to ground truth groupings, and any whose start and end times are both within 50 ms of each other, regardless of level, count as true positives. Unmatched transcribed groupings are false positives, and unmatched ground truth groupings are false negatives. The final metric is the standard F-measure using those counts. The new metrical F-measure, which allows time signature changes, is similarly based on the same groupings (which are well-defined, even across changes in time signature). However, when using an automatically-aligned transcription and ground truth, two groupings count as a match only if their start and end points are exactly aligned. The rest of the calculation is performed identically.

This paper has introduced an automatic alignment method to allow MV2H (**M**ulti-pitch, **V**oice, **M**eter, note **V**alue, and **H**armony) [4] to be used to evaluate non-aligned transcriptions and ground truth musical scores. This: (1) allows for the evaluation of end-to-end systems which do not produce such an alignment, and (2) allows for (much more widely available) non-aligned musical scores (e.g., MusicXML, for which a parser is included) to be used as ground truth. The metric is now also able to evaluate transcriptions and ground truths with key and time signature changes. The described improvements vastly increase the types of transcription systems which can use MV2H for evaluation. Code is available at `www.github.com/apmcleod/MV2H`.

ACKNOWLEDGMENTS

This work was supported in part by JST ACCEL No. JPMJAC1602, JSPS KAKENHI No. 16H01744 and No. 19H04137, and the Kyoto University Foundation.

REFERENCES

[1] Ralf Gunter Correa Carvalho and Paris Smaragdis. Towards end-to-end polyphonic music transcription: Transforming music audio directly to a score. In *WASPAA*, pages 151–155, 2017.

[2] Andrea Cogliati and Zhiyao Duan. A metric for music notation transcription accuracy. In *ISMIR*, pages 407–413, 2017.

[3] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck. Onsets and frames: Dual-objective piano transcription. In *ISMIR*, pages 50–57, 2018.

[4] Andrew McLeod and Mark Steedman. Evaluating automatic polyphonic music transcription. In *ISMIR*, pages 42–49, 2018.

[5] Eita Nakamura, Emmanouil Benetos, Kazuyoshi Yoshii, and Simon Dixon. Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization. In *ICASSP*, pages 101–105, 2018.

[6] Eita Nakamura, Kazuyoshi Yoshii, and Haruhiro Katayose. Performance error detection and post-processing for fast and accurate symbolic music alignment. In *ISMIR*, pages 347–353, 2017.

[7] Miguel A. Román, Antonio Pertusa, and Jorge Calvo-Zaragoza. An end-to-end framework for audio-to-score music transcription on monophonic excerpts. In *ISMIR*, pages 34–41, 2018.

[8] Hiroaki Sakoe, Seibi Chiba, A Waibel, and KF Lee. Dynamic programming algorithm optimization for spoken word recognition. *Readings in speech recognition*, 159:224, 1990.

[1] `http://www.music-ir.org/mirex/wiki/2017:Audio_Key_Detection`