PERCEPTUALLY-BASED EVALUATION OF THE ERRORS USUALLY MADE WHEN AUTOMATICALLY TRANSCRIBING MUSIC

Adrien DANIEL, Valentin EMIYA, Bertrand DAVID
TELECOM ParisTech (ENST), CNRS LTCI
46, rue Barrault, 75634 Paris cedex 13, France

ABSTRACT

This paper investigates the perceptual importance of typical errors occurring when transcribing polyphonic music excerpts into a symbolic form. The case of the automatic transcription of piano music is taken as the target application and two subjective tests are designed. The main test aims at understanding how human subjects rank typical transcription errors such as note insertion, deletion or replacement, note doubling, incorrect note onset or duration, and so forth. The Bradley-Terry-Luce (BTL) analysis framework is used and the results show that pitch errors are more clearly perceived than incorrect loudness estimations or temporal deviations from the original recording. A second test presents a first attempt to include this information in more perceptually motivated measures for evaluating transcription systems.

1 INTRODUCTION

In the benchmarking of Information Retrieval systems, performance is often evaluated by counting and classifying errors. Classically the ratio of relevant items that are returned out of the full set of original ones, referred to as *recall*, measures the completeness of the system performance whereas the proportion of relevant items that are retrieved, or *precision*, indicates the correctness of the answer. The F-measure, combining precision and recall, offers a single score to assess the performance. When music processing systems are involved, the question arises as to how to complement such a quantitative assessment by incorporating a certain amount of perceptually motivated criteria or weights.

This paper investigates the perceptual importance of typical errors occurring when transcribing polyphonic music excerpts into a symbolic form, *e.g.* converting a piece recorded in a PCM (.wav) format into a MIDI file. This particular Music Information Retrieval (MIR) task and its related subtasks (onset detection, multipitch estimation and tracking) have received a lot of attention [9] from the MIR community since the early works of Moorer [14] in the mid 70s. The approaches used to accomplish the goal are very diverse [4, 5, 14, 15, 16] and the evaluation of the performance for such systems is almost as varied. Some papers [4, 14] focus on a couple of sound examples, to probe typical errors such as octave errors, or deviations from ground truth such as duration differences, and so forth. However, the most widely used criteria for assessing automatic transcription are quantitative, even if the evaluation framework is not always similar (frame-based [15], note-based [16] or both [1]).

In the practical context of piano music for instance, the evaluation task is often handled by generating the PCM format piece from an original MIDI file which makes it possible to compare the input (ground truth) and output MIDI files. For that particular case, in this study, a perception test has been designed for subjectively rating a list of typical transcription errors (note insertions, deletions, incorrect onsets or duration...). The test is based on pairwise comparisons of sounds holding such targeted errors. The results are then analyzed by means of the Bradley-Terry-Luce (BTL) method [3].

In a second step, the question emerged of finding a way to take into account the perceptual ranking of the discomfort levels we obtained. Another test was designed to subjectively compare transcriptions resulting from different systems. It aimed at deriving more perceptually relevant metrics from the preceding BTL results by synthetically combining their main findings, and at checking their compliance with the test results. We worked in two directions: perceptually weighting typical errors, countable by comparing the input and output MIDI files, and adaptating similarity metrics [17].

2 THE EVALUATION MEASURES

The commonly-used F-measure is defined by:

$$f \triangleq 2\frac{rp}{r+p} = \frac{\#\text{TP}}{\#\text{TP} + \frac{1}{2}\#\text{FN} + \frac{1}{2}\#\text{FP}} \quad (1)$$

The authors thank all the subjects involved in the perceptive test for their participation. They also thank M. Castellengo, A. de Cheveigné, D. Pressnitzer and J. Benenson for their useful remarks, and M. Marolt, N. Bertin and P. Leveau for sharing their programs. The research leading to this paper was supported by Institut TELECOM under the *Automatic Transcription of Music: Advanced Processing and Implementation - TAMTAM* project and by the French GIP ANR under contract ANR-06-JCJC-0027-01, *Décomposition en Éléments Sonores et Applications Musicales - DESAM*.

where r denotes the recall, p the precision, $\#$TP the number of true positives (TP), $\#$FN the number of false negatives (FN) and $\#$FP the number of false positives (FP). f is equivalent to the quantity a, that is referred to as either accuracy or score [5], since $f = \frac{2}{\frac{1}{a}+1}$. The F-measure is useful to obtain the error rate for individually counted errors, but does not consider aspects like sequentiality, chords, harmonic or tonal relationships, etc.

Another evaluation approach comes from the problem of finding the similarity between two (musical) sequences. At the moment, these methods are commonly used to search for similar melodies in large databases, rather than in the field of the evaluation of transcriptions.

Let us assume that one must compare two sequences of symbols, A and B. The Levenshtein's distance, or edit distance [11], is a metric that counts the minimal number of operations necessary to transform A to B. The possible operations on symbols are: deletion from A, insertion into B, or replacement of a symbol in A by another one in B.

Mongeau and Sankoff [13] proposed adapting this distance to the case of monophonic musical sequences, in order to define a similarity metric between two melodies. The two sequences of notes are ordered according to the onset of each note. Each note is characterized by its pitch and duration, which are used to compute the cost of the following possible operations: insertion, deletion, replacement, with costs depending on tonal criteria, fragmentation and consolidation of several notes with the same pitch. These operations reflect typical mistakes in transcriptions. The minimum distance between the sets of notes is then estimated using the edit distance framework.

This melodic edit distance being applicable only to monophonic sequences, an extension to the polyphonic case has been recently proposed [8]. In order to represent the polyphonic nature of musical pieces, quotiented sequences are used. So far, this representation has only been applied to chord sequences, which constitute a restricted class of musical pieces: the notes within a chord must have the same onset and duration.

Another way to compute the similarity between two musical sequences [17] consists in considering each set of notes as points in a multidimensional space, *e.g.* the pitch/time domain. The algorithm is based on two choices. First, each point must be assigned a weight, *e.g.* the note duration. Second, a distance between a point in the first set and a point in the second one is defined, *e.g.* the euclidian distance in the time/pitch space. Then, the overall distance can be computed with the *Earth Movers Distance* (EMD) or the *Proportional Transportation Distance* (PTD). It is related to the minimum amount of work necessary to transform one set of weighted points to the other using the previously-defined distance, making it possible to transfer the weight of a source note towards several targets.

In all of these methods, the setting of the parameters is a crucial point. Indeed, the weighting between the time and the pitch dimensions, for instance, depends on music perception. The tests presented in this paper aim at assessing the trends of the perceptive impact of typical errors and the distribution of their related weights.

3 EXPERIMENTAL SETUP

3.1 Overview

The perception test consists of two tasks, which are detailed below. It was available on the Internet in the spring of 2007 for two weeks and was announced by e-mail. Before accessing the tasks, the subject is given instructions and information on the recommended audio device (high-quality headphones or loudspeakers, and a quiet environment) and on the estimated duration of the test. He or she is then invited to complete the tasks. Both of them consist in hearing a musical excerpt and several transcriptions of it, and in focusing on the discomfort caused by the transcriptions, with respect to the original. Task 1 uses artificial transcriptions, *i.e.* some copies of the original piece into which errors were inserted whereas task 2 uses transcriptions obtained by automatic transcription systems. In both cases, the transcriptions are resynthesized in the same recording conditions as the original piece in order to be heard and compared by the subject. At the end, the subject was asked to describe the criteria he used to compare files and to add any comments. Due to the total duration of the test a subject can possibly endure (about 40' here), we limited the scope of the study to pieces of classical piano music, from different periods, with different tempi and harmonic/melodic content.

3.2 Test 1: Subjective Evaluation of Typical Transcription Errors

3.2.1 *Principle*

Test 1 aims at obtaining a specific score for typical transcription errors. In order to achieve this, the transcriptions to be evaluated are made by inserting one and only one kind of error into an original excerpt. The error is chosen among the following list of typical errors: note deletion, random-pitched note insertion (1 to 11 half-tones), random-pitched note replacement (1 to 11 half-tones), octave insertion, octave replacement, fifth insertion, fifth replacement, note doubling, onset displacement, duration change (offset modification) and loudness modification (MIDI velocity).

These errors are inserted into three excerpts from *Studies, op 10 / Study 1 in C Major* by Chopin (8 seconds), *Suite Bergamasque / III. Clair de Lune* by C. Debussy (20 seconds), and *Sonata in D Major KV 311 / I. Allegro con Spirito* by W.A. Mozart (13 seconds).

Ideally, we would like to obtain a ranking of the typical errors. Due to the large number of files, asking the subjects

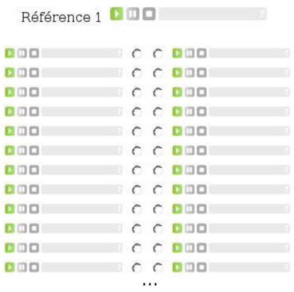

Figure 1. Test 1: for each pair of audio files, the subject selects the one causing more discomfort.

to give a score to each of them is not feasible. We preferred to set up a pairwise comparison task, as shown in Figure 1 and derived the full scale as described in the next section.

3.2.2 *Protocol and Settings*

For each kind of error, several test files are created with various error rates. The number of modified notes is parametered by the Modified Note Rate (MNR), which is set to either 10%, or 33%. For some kinds of error, the error intensity (EI) is also parametrized. This is quantified as a ratio of the note duration for duration changes and onset changes, and as a ratio of the MIDI velocity for loudness modifications. The EI is set to either 25%, or 75%. Modified notes are randomly chosen using the MNR. Intensity changes are made randomly, uniformly in the range centered on the true value and with the EI as radius.

To derive a ranking scale from pairwise comparisons, we choose the BTL method which uses hidden, “true” values associated to the transcriptions, along a given dimension (here, the discomfort). For a given pair of transcriptions, the subject’s answer is a comparison of a noisy version of the two true values, the noise modeling the subjectivity and the variable skill of subjects. Thanks to this statistical framework, the full subjective scale is then obtained by processing all the pairwise comparisons. For this test, 20 pairs out of 812 are randomly chosen and presented to each subject for each musical excerpt. This number has been chosen in order to adjust the test duration and is not critical for the results, as long as the number of subjects is high enough.

3.3 Test 2: Subjective Evaluation of Transcriptions of Musical Pieces

Test 2 aims at obtaining a perceptive score for a series of transcriptions from several pieces of music. Three original excerpts from *Prelude in C minor BWV 847* by J.S. Bach (13 seconds), *Suite Bergamasque / III. Clair de Lune* by C. Debussy (20 seconds), and *Sonata in D Major KV 311 / I. Allegro con Spirito* by W.A. Mozart (13 seconds) were chosen. For each excerpt, five transcriptions are presented, as shown in Figure 2. The subject has to assign a non-negative value to each transcription. These values express the discomfort caused by transcription errors in comparison with its reference. The subject can listen as many times as needed to each transcription and reference.

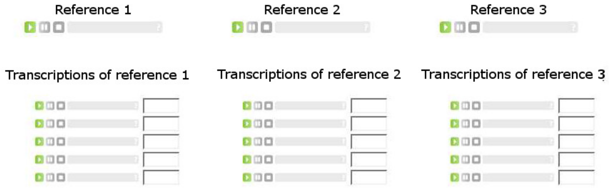
Reference 1
Reference 2
Reference 3
Transcriptions of reference 1
Transcriptions of reference 2
Transcriptions of reference 3

Figure 2. Test 2: the subject scores transcriptions with non-negative values.

In this test, all subjects are presented exactly the same audio files, in random order for each subject. One of the five transcriptions is the original piece in order to check whether the answers are consistent. The other four were obtained by automatic transcription systems, namely SONIC [12], available on the author’s website, Bertin’s system [2], a home-made system by P. Leveau based on [10] and an early version of [7]. The error rates and kinds of error thus depend on the specific behaviors of the transcription systems.

4 RESULTS

Thirty-seven subjects (24 musicians and 13 non-musicians) took part in this test. The results of Tests 1 and 2 are detailed here. The subjects’ comments show that the instructions were understood correctly. They pointed out tone errors as a major cause of discomfort, while they seldom mentioned loudness and duration errors in an explicit way.

4.1 Test 1

Results of Test 1 are given in Figure 3. The BTL method makes it possible to obtain, from the pairwise comparisons of all the subjects, a subjective scale of discomfort for typical errors. A BTL perception value is thus assigned to each modification, which can be ordered according to this scale.

Different forms of evidence show the consistency of the obtained scale. First, increasing scores are obtained with increasing error rates, either MNR or EI, and decreasing harmonicity (octave, fifth, random pitches). Second, a minimum discomfort is obtained for the reference (taking into account its confidence interval). Third, as described in [6], the above 90% confidence intervals are related to a 5% risk. Thus, they are narrow enough to distinguish error types and to assert that the answers make sense, although adjacent error types should be considered perceptually equivalent.

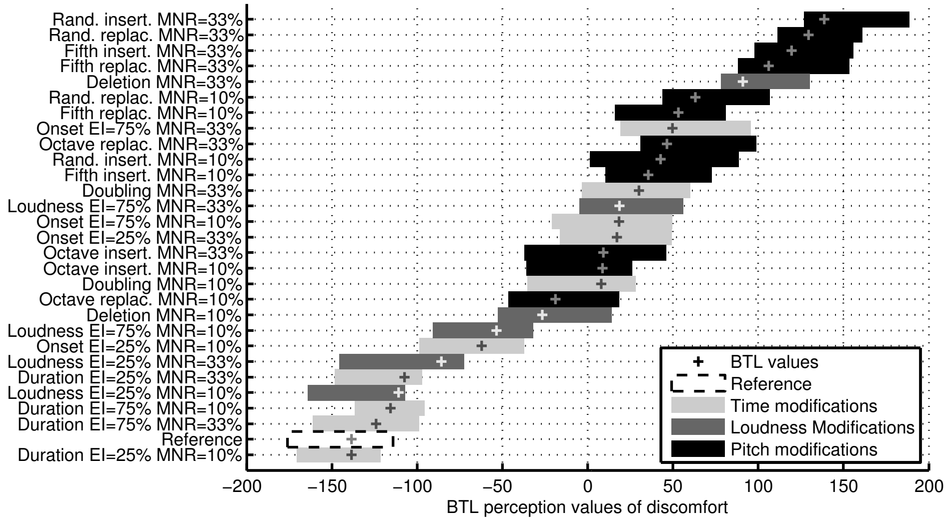

Figure 3. Test 1 : perceptive scale for typical errors. Crosses account for the related BTL value. Horizontal bars depict the 90% confidence intervals, obtained by a bootstrap method [6] using 100 resamplings of the data (because the data is not gaussian, confidence intervals may not be centered on BTL values).

Globally, as expected from the subjects' comments, the highest discomfort values are obtained with pitch modifications; loudness and time modifications cause low to medium discomfort. Regarding pitch changes, octave errors are judged much less serious than fifth changes, which cause a slightly lower discomfort than random changes. In each case, replacements and insertions are judged as equivalent, which would indicate that the discomfort is more induced by the added note than by the deleted note of a replacement. The lower values obtained for deletions confirm this hypothesis, which is commonly observed when working on transcription systems: a false negative usually sounds better than a false positive.

Results with time errors show that the modified onsets cause much more discomfort than duration changes. While one can expect that moving the beginning of an event causes a significant subjective change, it seems that subjects just did not perceive most of the modifications of duration. This can be explained by a specific feature of the piano: the ends of sounds generated from its freely-vibrating strings are less perceptible than for a musical instrument with forced vibrations. Thus current results for perception of note duration cannot be generalized to all instruments.

Finally, additional analysis of the results should be reported, which are not represented in Figure 3. First, similar results were obtained from subjects that were musicians and from non-musicians. Second, the three scales obtained for the three excerpts are also similar, with a little difference for the excerpt by Debussy in which deletions have a lower score and duration changes cause higher discomfort, probably because of the long durations in this slow piece of music.

4.2 Test 2

For each subject, scores of Test 2 are normalized by the maximum score he/she gave. Six subjects were removed since they scored a discomfort greater than 20% for the reference. Average scores and variances were then computed, with respect to the results from all the remaining subjects.

Results are represented in Figure 4. As the test is not a contest between existing algorithms, the systems were made anonymous, numbered from 1 to 4. The confidence in the results is assessed thanks to a 3 (composers) $\times$ 5 (algorithms) factorial ANOVA test, passed for each dimension and for interactions using a $p = 0.01$ test level. Thereby, the scores and the ranking of the algorithms are very dependent on the piece of music. This confirms that the performance of a transcription system is related to the musical content of pieces and thus depends on the test database. Large standard deviations indicate that the evaluation of a musical transcription depends greatly on proper subjective criteria. An important overlap between the answers makes it impossible to obtain a definitive ranking among the different algorithms even if for each excerpt, systems 2 and 3 are judged as the worst and the best one respectively.

5 EVALUATING THE TRANSCRIPTIONS

When comparing the results given by one of the objective evaluation methods proposed in Section 2 to the perceptive results of Test 1, several aspects are differentiated in the former case while they are not in the latter case, and vice versa. For instance, octave, fifth and random pitch changes have

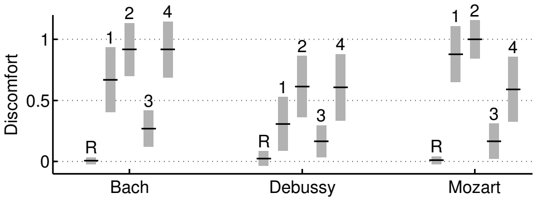

Figure 4. Results of Test 2: perceptive evaluation of the reference (R) and of transcriptions from four systems (1-4), with standard deviations (gray bars).

similar F-measures but cause an increasing discomfort. On the contrary, perceptive results are equivalent for replacements and insertions, whereas the F-measure is higher for insertions. Besides, perceptive results provide a balance between time and pitch influences, which is not taken into account in objective evaluation methods.

In this part, we estimate weighting coefficients of typical errors from Test 1 results, and we then apply them to adapt two existing metrics: the F-measure and the PTD. These modified methods are validated by applying them to the excerpts used in Test 2 and by comparing the results with the discomfort expressed by the subjects.

5.1 Extraction of the Weighting Coefficients

To extract the weighting coefficients, the results of Test 1 are normalized between 0 and 1. We only used results with MNR 33%, and the results were averaged for pitch modifications, insertions and replacements. Six criteria[1] are obtained, to be integrated into metrics. Their related weighting coefficients are given in the following table:

Octave	Fifth	Other intervals	Deletion	Duration	Onset
$\alpha_1 =$ 0.1794	$\alpha_2 =$ 0.2712	$\alpha_3 =$ 0.2941	$\alpha_4 =$ 0.2475	$\alpha_5 =$ 0.0355	$\alpha_6 =$ 0.4687

The coefficients are normalized so that $\frac{1}{3}\sum_{i=1}^{3}\alpha_i + \sum_{i=4}^{6}\alpha_i = 1$, since octave, fifth and random pitch represents alternative false positive errors.

5.2 Perceptive F-measure

In eq.(1), errors are the number of FP and FN, with an equal weight ($\frac{1}{2}$). We thus define the perceptive F-measure by:

$$f_{\text{percept}} \triangleq \frac{\#\text{TP}}{\#\text{TP} + \sum_{i=1}^{6}\alpha_i w_i \#E_i} \tag{2}$$

where $\#E_i$ is the number of errors of type i (see below), $w_1 = w_2 = w_3 = w_4 = 1$, w_5 is the average duration error in the transcription, and w_6 is the average onset error. (The average errors are computed as the square root of the mean square error.) Note that a similar perceptive accuracy could be defined by using the equivalence mentioned in Section 2 and that the expression (2) equals the F-measure in the case $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \frac{1}{2}$ and $\alpha_5 = \alpha_6 = 0$.

Errors from MIDI files are extracted as follows:

1. TP are estimated as notes with correct pitch (rounded to the nearest semitone) and onset deviation lower than 150 ms. For each TP, the relative onset deviation and the relative duration deviation (both with respect to the original note parameters) are extracted. Then, let $\#E_5 = \#E_6 = \#\text{TP}$.

2. FP are transcribed notes which are not TP. The set of FP is split as follows: for each FP, (a) if there is an original note at the octave or sub-octave, at the same time (*i.e.* with any overlap of both time supports), the FP is added to the set E_1 of octave FP; (b) otherwise, if there is an original note at the upper or lower fifth at the same time, the FP is added to the set E_2 of fifth FP; (c) otherwise, the FP is added to the set E_3 of other pitch FP.

3. FN are the set E_4 of original notes that are not associated with one TP.

5.3 Perceptive PTD

The PTD is originally used to evaluate melodic similarities (see Section 2). In this context, note duration as weights to transfer and the euclidian distance in the time/pitch space seem to be appropriate choices. Nevertheless, when comparing generic musical pieces, both of these choices should be changed. PTD weights should be defined in a musical sense but this is beyond the scope of the current work and we thus chose to assign an equal and unitary PTD weight to each note. Using the perceptual coefficients introduced in Section 5.1, the distance between two notes is then defined in the multidimensionnal space of criteria composed of pitch (octave, fifth or others), duration and onset modifications.

5.4 Results

Figure 5 shows the results of the application of the original two objective measures and of their perceptive versions to the musical excerpts from Test 2. In order to compare them to the discomfort results, F-measures were changed by applying the function $x \mapsto 1 - x$. In order to best fit the discomfort scale, all results were scaled by a multiplicative coefficient obtained by minimizing the mean square error.

[1] Loudness was not considered since the results were not satisfying, probably due to the difficulty of having a trustworthy loudness scale. Doubled notes were not used either because they could not be integrated into metrics.

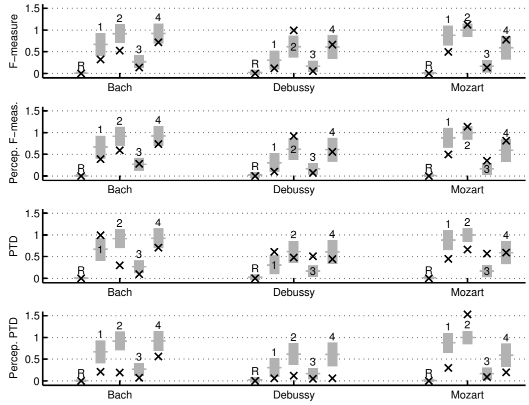

Figure 5. Transcription evaluation results with several objective and perceptive measures: in each case, crosses show the normalized error related to a measure, and the gray bars indicate the discomfort obtained in Test 2.

Results with the perceptive F-measure are slightly closer to the discomfort values than the original F-measure. Moreover, the ranking of the 15 excerpts is also closer to the discomfort-based ranking. Results of the perceptive PTD do not look better than the original, due to a high isolated value for the excerpt with highest discomfort (Mozart, System 2), that makes it difficult to scale the results adequately. However, the achieved ranking is dramatically better than the ranking by the original PTD, and also slightly better than the ranking by the perceptive F-measure. Thus, even if the relation between the discomfort and the perceptive PTD may be non-linear, the latter is appropriate in a ranking task.

6 CONCLUSIONS

The main idea of these tests was to get a ranking of the typical automatic transcription errors, to extract perception weights, and to integrate them into several musical sequence distance metrics. These primary results are consistent and the proposed perceptive metrics give satisfying results.

However further investigations should focus on a number of aspects, such as non-linear relations between specific error rates and discomfort, musical-based typical errors (taking into account tonality, melody, chords, etc.), and more specific algorithms to identify them.

References

[1] Multiple fundamental frequency estimation & tracking. In *Music Information Retrieval Evaluation eXchange (MIREX)*, 2007.

[2] N. Bertin, R. Badeau, and G. Richard. Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark. In *Proc. of ICASSP*, Honolulu, Hawaii, USA, April 2007.

[3] R.A. Bradley. Some statistical methods in taste testing and quality evaluation. *Biometrics*, 9(1):22–38, 1953.

[4] A.T. Cemgil, H.J. Kappen, and D. Barber. A generative model for music transcription. *IEEE Trans. Audio, Speech and Lang. Proces.*, 14(2):679–694, 2006.

[5] S. Dixon. On the computer recognition of solo piano music. *Australasian Computer Music Conf.*, 2000.

[6] B. Efron and R. J. Tibshirani. An introduction to the bootstrap. In *London: Chapman & Hall*, 1993.

[7] V. Emiya, R. Badeau, and B. David. Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches. In *Proc. of EUSIPCO*, Lausanne, Switzerland, August 2008.

[8] P. Hanna and P. Ferraro. Polyphonic music retrieval by local edition of quotiented sequences. In *Proc. of CBMI*, Bordeaux, France, June 2007.

[9] A. Klapuri and M. Davy. *Signal Processing Methods for Music Transcription*. Springer, 2006.

[10] P. Leveau, E. Vincent, G. Richard, and L. Daudet. Instrument-specific harmonic atoms for mid-level music representation. *IEEE Trans. Audio, Speech and Lang. Proces.*, 16(1):116–128, January 2008.

[11] V. I. Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1(1):8–17, 1965.

[12] M. Marolt. A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Trans. on Multimedia*, 6(3):439–449, 2004.

[13] M. Mongeau and D. Sankoff. Comparison of musical sequences. *Computers and the Humanities*, 24(3):161–175, 1990.

[14] J.A. Moorer. *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*. Dept. of Music, Stanford University, 1975.

[15] G. Poliner and D. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, 8:1–9, 2007.

[16] M. Ryynänen and A.P. Klapuri. Polyphonic music transcription using note event modeling. In *Proc. of WASPAA*, pages 319–322, New Paltz, NY, USA, 2005.

[17] R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering, and R. van Oostrum. Using transportation distances for measuring melodic similarity. In *Proc. of ISMIR*, Baltimore, Maryland, USA, 2003.