Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009) Jyväskylä, Finland
Jukka Louhivuori, Tuomas Eerola, Suvi Saarikallio, Tommi Himberg, Päivi-Sisko Eerola (Editors)

Measuring Music Transcription Results Based on a Hybrid Decay/Sustain Evaluation

Nuno Fonseca,[*1] Aníbal Ferreira,[#2]

[*]*ESTG, Polytechnic Institute of Leiria, Portugal*
[#]*FEUP, Oporto University, Portugal*
[1]nfonseca@estg.ipleiria.pt, [2]ajf@fe.up.pt

ABSTRACT

Although much work is being done in music transcription research, the evaluation of these techniques is less addressed by the research community. The lack of widely accepted metrics and databases presents an obstacle to the assessment of existing music transcription approaches. This paper presents an analysis of existing metrics and proposes a new method for measuring the results of music transcription. Based on the idea that decay and sustained music instruments may have different requirements, a dual process is implemented. On the decay process, a note oriented approach is used, considering pitches and onsets, generating a score for each note. On the sustain process, a time oriented approach is used, measuring the overlap of original and transcribed notes. The final score is produced based on the values obtained in both processes. To evaluate the proposed approach, several music transcription metrics were compared with human tests results. The obtained results show that the proposed method achieves the best correlation with human perception results. Based on the idea that not all transcription errors have the same impact, an effort was made to achieve a metric that is more realistic from the human perception point-of-view.

I. INTRODUCTION

Music Transcription is an important area in Music Information Retrieval (MIR). Its goal is to be able to extract symbolic music information from an audio stream, i.e., to be able to represent note information (e.g. note, onset, offset) that is present on music audio fragments. From the transcription of a monophonic music line (one instrument playing a note at a time) to the transcription of a full symphonic concert, the characteristics and techniques used by the music transcription systems vary immensely, most of them still on a research phase. But there is an aspect of music transcription that doesn't receive much attention and it is even considered (Wang & Brown, 2006) perhaps the weakest aspect of research on the subject: evaluation of music transcription systems. It is fundamental to measure the results obtained by such systems, not only as a tool to improve them, but also as a way to compare the results of different approaches. Although there is much work being done on music transcription techniques, the evaluation of its results does not get much research attention, besides brief paper paragraphs explaining how results were obtained.

To evaluate music transcription techniques, two concerns are in order: which music databases to use and which metrics. This paper will focus on the metrics problem, by presenting an analysis on measuring music transcription results and presenting a different approach based on two different processes that complement each other.

II. MEASURING MUSIC TRANSCRIPTION RESULTS

To measure music transcription results, the metrics must be chosen (besides the choice of music database), and this set of metrics (rules, parameters, etc) will, of course, have an enormous impact on the final results.

A. Current Approaches

There are two main approaches for measuring the results of music transcription. The most common is note oriented - the first step is to identify what are the corrected transcribed notes, and the second step is to use that information for obtaining a final score. The other approach is time (frame) oriented, on which original and transcribed "piano-rolls" are compared, frame-by-frame.

In note oriented approaches, usually notes are considered as corrected transcribed notes or as errors (there is no between). In order to be considered a corrected transcribed note, its parameters (pitch, onset and in some cases offset) must be within some tolerance values. Since most music pieces (especially western music) consider a musical resolution of a semitone, pitch tolerance of ± ½ semitone is widely accepted. Timing tolerances are more heterogeneous. Onset tolerances usually are between ± 25 ms and ± 150 ms.

Regarding offset information, some systems disregard them e.g., (Dixon, 2000). Others consider a tolerance value higher than the onset tolerance, and others may ever use a combination of both absolute and relative tolerance values (e.g. the higher value between 50 ms or 20% of the note duration) (MIREX 2007).

To obtain final score values, performance measures are used. Many authors, e.g. (Ryynanen & Klapuri 2005),(Reis et Al, 2007), use information retrieval parameters like recall and precision. Recall denotes the percentage of original notes that are presented on the transcription (see eq.1).

$$recall = \frac{|OriginalNotes \cap TranscribedNotes|}{|OriginalNotes|} \qquad (1)$$

Precision denotes the percentage of the transcript notes that are presented on the original sequence (see eq.2).

$$precision = \frac{|OriginalNotes \cap TranscribedNotes|}{|TranscribedNotes|} \qquad (2)$$

In many situations (Reis et al, 2008), (MIREX 2007) to end-up with a single global parameter, an F-measure is obtained based on recall and precision values (see eq.3).

$$F\text{-}measure = \frac{2 \times recall \times precision}{recall + precision} \qquad (3)$$

Other method for obtaining a final score is based on accuracy, like in (Dixon, 2000), based on the ratio between correct transcribed notes and the sum of correct notes, missed notes that were left un-transcribed and spurious notes that were added to the transcription (see eq. 4).

$$Accuracy = \frac{|CorrectTranscribedNotes|}{|CorrectTranscribedNotes| + |NotTranscribedNotes| + |AddedTranscribedNotes|} \quad (4)$$

Some authors (Ryynanen & Klapuri 2005), (MIREX 2007) consider additional parameters, like the mean of overlap ratio of each corrected transcribed note (eq. 5), which can be used for analyzing the behavior of offsets (which are disregarded for deciding corrected notes in (Ryynanen & Klapuri 2005)).

$$Overlap\ Ratio = \frac{\min(offsets) - \max\,(onsets)}{\max(offsets) - \min\,(onsets)} \quad (5)$$

In the less common time/frame oriented approaches, original and transcribed notes are represented as notes vs frames matrices, which are used as a mean to calculate their overlapping. The final value can be based on frame-level accuracy (use overlap values at frame level to calculate accuracy), or based on “speaker diarization error score” (Graham et Al, 2007).

More recently, some interesting methods were proposed. (Tavares et al, 2008) present one idea to measure the distance between an original note and a transcribed one (eq. 6) considering two Euclidean spaces. D_t represents the distance from the time point-of-view (considering their onsets and lengths differences); D_f represents the distance from the frequency point-of-view, considering cylindrical coordinates as a way to approximate octave errors; and T_{over} represents their overlap in time.

$$L(N_1, N_2) = \frac{T_{over}(N_1, N_2)}{\sqrt{D_f^2(N_1, N_2) + D_t^2(N_1, N_2)}} \quad (6)$$

After calculating all the possible distances between each original note and each transcribed note, a one-to-one match process would take place. Finally, a performance evaluation array would be created with diverse statistics information. Unfortunately, the method doesn’t output a global measure for the music transcription.

In (Daniel et al, 2008), some human perceived tests were done and a new metric was proposed. It’s still a note oriented approach, but the transcription errors were divide in 6 classes:

- octave error
- fifth error
- other interval error
- deletion error
- duration error
- onset error

Based on their human tests, each class of errors will have an associated weight, depending on their perceptual impact on the music transcription. Although the amount of the error is not considered (still a binary approach: “correct note”/”error”), the impact of that type of error is considered, achieving, according to their tests, a metric closer to the human perception.

B. Impact of Transcription Errors

To be able to find good metrics for measuring music transcription results, it is important to identify what are the contributions to a good transcription.

1) Onset time tolerance. Everyone agrees that notes with onset values within a range of ±1 ms can be considered as successful transcriptions, but applying that small time tolerance to transcript systems will not give us real information about its transcription capabilities, since several “correct” transcribed notes might not fall within such tight time tolerance. On the other extreme, applying a time tolerance of ±200 ms in transcription systems might tell us more information about its capabilities, but if a transcription achieves 100% success, we could not be sure that a perfect transcription was obtained. Some systems tend to use a time tolerance around 25-50 ms, probably because of the margin error of the ground truth and/or due to the proximity effect, also known as Haas effect (Haas, 1972), since human note time resolution may be around 25-50 ms (sounds arriving to the ear with lower time intervals are perceptually merged).

In a music transcription, if a note is shifted by ±40 ms (regarding the original note), probably that shift isn’t even perceived. But it’s important to analyze the impact of such time shifts, not only between original and transcribed notes, but also within transcribed notes. For instance, if two notes that should be played simultaneously are shifted respectively +40ms and -40ms, although they stay within a ±40 ms range of the original ones, they became apart by 80 ms, which can be easily perceived as a small transcript error.

2) Impact of the instrument time envelope. The behavior of the instrument sound intensity over time (time envelope) can be one important factor in music transcription process. For a better understanding of these issues, let us focus in two (extreme) types of musical instruments, which we will designate as “decay” instruments and “sustain” instruments (see Figure 1). Let’s define “decay” instrument as a musical instrument with a very fast decay behavior (almost like percussion, high notes from a harp, pizzicato violin, etc), and with a release behavior very similar with its decay. And let’s define “sustain” instruments as the ones characterized by having a constant energy/timbre behavior over the note duration (almost like woodwinds, strings, organ, but even with less attack timbre variations).

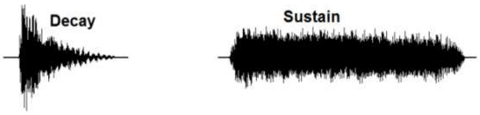

Figure 1. A decay instrument and a sustain instrument

When transcribing music pieces of “decay” instruments, and depending on the instrument, the concept of offset might not be applied (e.g. glockenspiel), which means that to measure the transcription we should only focus on the note and on the onset. Even when the offset concept can be applied (instruments with medium-fast decays), and since most of the energy is concentrated on the initial moment of each note, from the perception point of view, the offset time value will not be as important as its onset time value. Also, in this type of instruments, the offset time is only important in a limited period of time, because after some point the sound energy will be below the threshold of hearing (Zwicker & Fastl, 1999), even if from the musician point of view the note is still playing. For instance, it's almost impossible to detect or perceive if the higher C of a piano plays for 3 seconds or 30 seconds.

When transcribing "sustain" instruments, the importance of offset values highly increase (although might continue to have slight less importance than onset values).

So, if, generally speaking, offset values are not as important as onset values, how can we consider offset values, that are important in music transcription, but without giving them the same amount of importance as onset or pitch? By increasing their tolerances? By adding a note-duration based tolerance? By doing separated tests with and without offset analysis, leaving for the reader to take his own conclusions?

If we compare these differences between the transcription of decay and sustain instruments, with the differences of measuring transcription results using a note oriented approach or a time oriented approach, it may seem that decay instruments transcription may be better measured with note-oriented approaches, but sustain instruments transcription may be better measured with time/frame oriented approach. But the major issue is that the majority of musical instruments are not fully "decay" or fully "sustain" based. Most instruments with a decay envelope end-up having a not so short decay time, and most "sustain" instruments, that may even have a constant energy behavior, end-up having an important timbre variation at the attack that may perceptually increase their impact compared with the rest of the note duration.

3) One-to-one mapping. Some works (Ryynanen & Klapuri, 2005), (MIREX 2007) refer the need of a one-to-one mapping when comparing transcription results, i.e., if a match is detected between an original note and a transcribed one, none of those can be used again on other note matches. The main idea is to avoid situations on which 2 different original notes could be matched to a single transcribed one (for instance, two quick notes been transcribed as one), or vice-versa, allowing a perfect score even in situations that have different number of notes. This rule is important, but it's usually incomplete, since it is not specified what should happen if an original note could be matched against 2 transcribed notes (or vice-versa).

For instance, on a 50 ms tolerance system, imagine an original note (W) at 960 ms and other original note (X) at 1020ms, that are transcribed as a 1000 ms note (Y) and a 1060 ms note (Z), like in figure 2 (let's ignore offset values and consider same pitch for all notes). What should be the transcribed note pair of note X? The best match (Y – a 20 ms difference) or the first available (Z, since Y was first mapped against W – a 40 ms difference)?

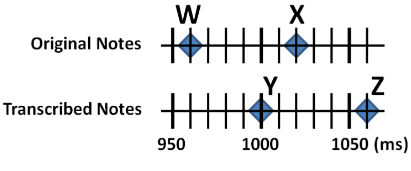

Figure 2. Example of transcription (onsets only)

If we map X-Y, W and Z cannot be mapped because there is a 100 ms gap. If we map X-Z and W-Y, we get full mapping, although not mapping the best notes with X and Y.

Of course, these types of situations are probably rare on a global perspective, but they could be less rare within a special transcription system or within a special audio fragment.

4) Binary approach to transcription. Almost all metrics used on the subject employ a binary approach regarding the transcription of a note: the original note is perfectly transcribed or it isn't transcribed at all. This discrete approach has the advantage of been the simplest one, nevertheless, it might create some unfairness (regarding time tolerances, but also in others situations). Should all errors be treated equally? If some transcription errors have lower impact than others, should metrics take that in consideration?

For instance, splitting or merging notes are common situations: one note that is transcribed as a sequence of shorter notes (with same pitch), or a sequence of notes that are transcribed as only one note. Although these errors are easily perceived on decay sounds, it might be difficult to detect and perceived on sustain sounds (especially sounds with high release time values).

Octave errors are also very frequent in music transcription systems and are usually treated as common errors. But although these octave errors are perceived by humans, their impacts on perception are lower when compared with other pitch errors.

III. PROPOSED METHOD

Taking into consideration all of these issues regarding music transcription metrics, a new method is proposed, on which the final measure of a music transcription is obtained by the means of two different processes, based on the idea that "decay" and "sustain" sounds have different requirements regarding music transcription, and that most musical sounds have a combination of decay and sustain behavior. In the first process, named "decay process", the transcription is evaluated as if the musical sounds had all decay behavior, so only pitch and onsets are considered. On the sustain process, we consider that sounds were "sustain" based, and overlapping of both "piano-rolls" (original and transcribed) are obtained (considering pitches, onsets and offset).

A. Decay Process

In the decay process, only pitches and onsets are considered. Instead of a binary approach, made with "corrected" transcribed notes and "incorrect" transcribed notes, the system will produce a score for each pair of original/transcribed notes that will range from 0 ("completely incorrect" note transcription) to 100% ("completely correct" note transcription). The goal of this note score is to be able to obtain the maximum amount of information about the transcription performance of the system. Applying a threshold only, doesn't take into account several important aspects, so our approach considers that some transcription errors can be smaller than others.

If the pitch is correct (± ½ semitone), and if onsets are within a 25 ms interval, a full score is obtained (100%). If the onsets are separated more than 200 ms, the score is 0. For situations between, a linear value is obtained as can be seen in eq. 7 and in

Figure 3. In the case of octave errors (differences of ±1 octave between original and transcribed notes) a note score is also produced, but considering 30% of the values obtained with eq. 7.

$$NoteScore = \begin{cases} 1 & , \Delta onsets \leq 25ms \\ \dfrac{200 - \Delta onsets}{200 - 25} & , 25ms < \Delta onsets \leq 200ms \\ 0 & , \Delta onsets > 200ms \end{cases} \qquad (7)$$

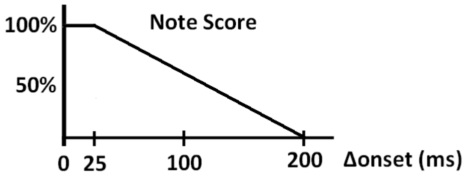

Figure 3. Note Score

To prevent the issues of one-to-one mapping already mentioned in section II.B.3, all original notes are mapped with the transcribed note that generates the higher NoteScore. On a separated process, all transcribed notes are mapped with the original note that generates the higher NoteScore. Mappings of n-to-1 and 1-to-n are permitted, and there could be situations on which the mappings are not bidirectional (see Figure 4).

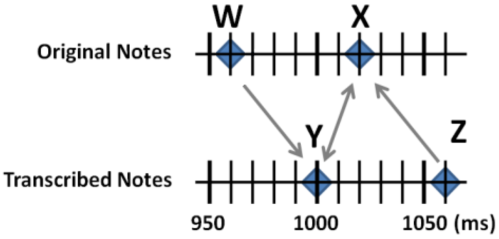

Figure 4. The best mappings between original and transcribed notes

Each note (original or transcribed) will have an individual weight value (W_O or W_T) with a default value of 1. But, if the note (original/transcribed) ends-up never being picked as the best match for any other note (transcribed/original), their weight is reduce to 0.5. For instance, on figure 4, note W and Z will have weights of 0.5 because no other notes have chosen them as best match.

Although the application of the weights (W_O or W_T) might increase the complexity of the algorithm, they are responsible for three important features:

- Note sequences are commutative (exchanging original and transcribed note sequences will produce the same final score).
- Splitting/merging note errors are not so penalized.
- Since the proposed process doesn't have a binary approach to correct/incorrect transcribed notes, forcing one-to-one maps would require the existence of an important optimization process to be able the choose the right one-to-one mappings as a way to get the best overall score.

A variation of recall and precision is calculated (eq. 8 and eq. 9), where N_O and N_T represent the number of original notes and the number of transcribed notes. As can be seen, the DecayRecall just considers the best NoteScore of each original note, and DecayPrecision just considers the best NoteScore of each transcribed note.

$$DecayRecall = \frac{\sum(NoteScore(O_i) \times W_{oi})}{N_o} \qquad (8)$$

$$DecayPrecision = \frac{\sum(NoteScore(T_j) \times W_{Tj})}{N_T} \qquad (9)$$

Although it might look that the obvious choice for a final decay score would end-up on using the f-measure formula, the authors prefer using accuracy as final decay score, since it is a more linear behavior and most of f-measure advantages (Manning 2008) don't apply. By disregarding true negatives (TN), accuracy can be calculated as seen in eq. 10.

$$Accuracy = \frac{1}{\frac{1}{recall} + \frac{1}{precision} - 1} \qquad (10)$$

So, the final decay score is calculated as (eq. 11).

$$DecayScore = \frac{1}{\frac{1}{DecayRecall} + \frac{1}{DecayPrecision} - 1} \qquad (11)$$

B. Sustain Process

In the sustain process, the main idea is to measure the overlapping between the original notes and transcribed ones, almost like analyzing the overlapping of their "piano-rolls", considering the pitch, onset and offset of each note. For each note, up to a 25 ms tolerance on the onset and on the offset are considered. The idea is to be able to get a 100% score even if notes timings are slightly different (within a 25ms range). To achieve that, a value between 0 and 25 ms can be subtracted[1] to the onset value or added to the offset value, in original or transcribed notes, as a way to increase their interception, i.e., the duration of the original note or the duration of the transcribed note can be increased by up to 25 ms on its onset and up to 25 ms on its offset.

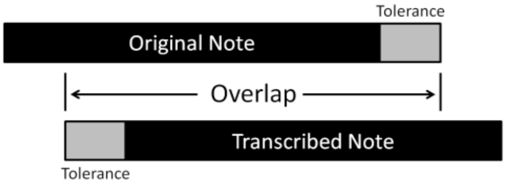

Figure 5. Interception between an original note and a transcribed one

[1] Decreasing the onset value, which means that the onset will begin a little earlier.

After this transformation on the timings of original notes and transcribed notes, a variation of recall and precision values are obtained using eq. 12 and eq. 13.

$$SustainRecall = \frac{\sum Overlaps}{\sum OriginalNoteDurations} \qquad (12)$$

$$SustainPrecision = \frac{\sum Overlaps}{\sum TranscribedNoteDurations} \qquad (13)$$

It's important to refer, that the values used on eq. 12-13 are the ones calculated after timing transformation (with added tolerances), and not based on the real original and transcribed notes.

The final sustain score, once again based on the concept of accuracy, can be calculated using its recall and precision values using eq. 14.

$$SustainScore = \frac{1}{\frac{1}{SustainRecall} + \frac{1}{SustainPrecision} - 1} \qquad (14)$$

Octave errors are also considered, but once again with only 30% of the time value obtained with the interception. An original note fragment and a transcribed note fragment can be considered as an interception if they are one octave apart and if they don't intercept with other notes/fragments. If this is the case, 30% of their time value is considered as the interception, and the remaining 70% of the time value are considered as not intercepted value. As can be easily seen, on the sustain process, longer notes have a bigger impact on the final score than smaller notes. The sustain process also allows us to benefit some errors that otherwise would be considered as ordinary transcription errors like errors in missed onset determination on small-attack sounds, note merging/splitting, etc.

C. Final Score

The final result is the average of the values obtained on the decay process and on the sustain process (eq. 15), and will range between 0 and 100 %.

$$FinalScore = \frac{DecayScore + SustainScore}{2} \qquad (15)$$

From the analysis point of view, the process makes available 7 different parameters that present interesting information about the transcription capabilities of the system: DecayRecall, DecayPrecision, DecayScore, SustainRecall, SustainPrecision, SustainScore and FinalScore.

D. Implementation

An implementation of the proposed method, for the Matlab Framework or C/C++, is freely available at http://www.estg.ipleiria.pt/~nfonseca.

IV. PERCEPTION TESTS

Although it makes sense to use a metric that consider different types of error and different amounts of errors, it is important to compare the proposed method with other methods and with human perception tests.

To validate our approach, the authors used the test materials created by (Daniel et al, 2008), consisting on 3 different music fragments (from Bach, Debussy and Mozart piano pieces) that were transcribed with 4 music transcription systems (15 files, considering the reference files also). A group of 31 listeners were asked to rate the discomfort of each transcription on a scale between 0 and 1.

To evaluate which metric correlates best with the human results, 3 methods were used (as presented in (Emiya, 2008)):

- Pearson linear correlation coefficient (prediction accuracy), which measure the linear correlation between 2 sets of values.
- Spearman rank order coefficient (prediction monotonicity), which disregard the actual values, considering only their order.
- Outlier ratio (prediction consistency), which validates that there isn't any value too far apart from the correct value.

Table 1 shows the obtained results of (Daniel et al, 2008) (marked with *) and the ones of the proposed method (individual decay and sustain scores are also presented).

Table 1. Prediction accuracy, monotonicity and consistency for several metrics when compared with the human perceived results.

Metrics	Prediction accuracy	Prediction monotonicity	Prediction consistency
F-measure*	83.4 %	83.5 %	**0 %**
Perceptive F-measure*	84.1 %	84.9 %	**0 %**
PTD*	60.3 %	61.6 %	**0 %**
Perceptive PTD*	64.8 %	**89.6 %**	6.7 %
Decay Score	89,2 %	84,3 %	**0 %**
Sustain Score	85.0 %	80.0 %	**0 %**
DecaySustain Score	**90.5 %**	82.8 %	**0 %**

V. DISCUSSION

The proposed method presents the best prediction accuracy with the human results with a significant improvement over the other methods. Like most other metrics, the method also does not present outliers (prediction consistency = 0%). The only parameter with lower results is Prediction monotonicity. Since this parameter is particularly used on non-linear situations, and since the linear correlation value is above 90%, it seems, from the authors point of view, that its importance can be slightly reduced.

The results also show that although the human tests were done with piano transcriptions, that might be considered a "more" decay instrument, the combined use of decay and sustain methods gives the best score.

Although the test set could be used as a way to partially validate the proposed method, it doesn't have the necessary size or diversity to truly validate the proposed method, much less to tune its internal parameters. For instance, additional tests were made using the same test set and changing internal values of the proposed method, allowing better results. Nevertheless, those value are not presented since it is the authors opinion that those better results simply represent a best fitting with the small test set, and that would probably not improve the overall capabilities on other test sets.

VI. CONCLUSIONS AND FUTURE WORK

A different metrics approach to music transcription was proposed.

Its main features are the following:

- music transcription errors are different from each other, either on type and on amount;
- offset has a smaller impact than onset (since it is considered only on “sustain” process);
- timing variations have a more gradual impact (instead of a discrete impact);
- the process is commutative (exchanging original and transcribed note sequences generate the same final score);
- octave errors have a slight lower impact than other pitch errors;
- merging or splitting notes has a slight lower impact than in traditional metrics;
- slow-attack sounds are better handled than traditional metrics.

Nevertheless, the proposed method still presents some issues: it’s more complex than standard metrics (although a free implementation is available) and it doesn’t analyze the impact of note dynamics/loudness, which can be important in some music transcription scenarios.

In the future, much work and discussion must be done regarding music transcription metrics. Larger perception tests are needed as a mean to better understand the impact of each type of transcription error, the impact on different types of instruments, and also was a way to allow tuning of metrics internal parameters. Besides pitch, onset and offset, future metrics should also be able to analyze other music features that future transcription systems may extract, like dynamics/loudness, timbre identification, pitch variations (vibrato, portamento, etc), etc.

Although this metrics are created for music transcription, it can be also used on other areas of MIR, as a measure of similarity between note sequences (e.g. measuring melody extraction, resynthesis/transynthesis).

ACKNOWLEDGMENT

We thank Valentin Emiya and all the authors of (Daniel et al, 2008) for sharing with us their testing materials, allowing us to cross-validate our approach.

The authors also like to acknowledge the Portuguese Science and Technology Foundation for partial financial support.

REFERENCES

Daniel A., Emiya V., David B. (2008). *Perceptually-based evaluation of the errors usually made when automatically transcribing music*, Proc. of the Ninth International Conference on Music Information Retrieval (ISMIR).

Dixon, S. (2000). *On the Computer Recognition of Solo Piano Music*, Australasian Computer Music Conference, Brisbane, Australia, 31—37.

Emiya V. (2008). *Automatic transcription of piano music,* PhD Thesis, TELECOM ParisTech, Paris, France.

Graham E. Poliner, G.E. and Ellis, D.P.W. (2007). *A Discriminative Model for Polyphonic Piano Transcription*, EURASIP Journal on Advances in Signal Processing, Volume 2007, Article ID 48317.

Haas, H. (1972). *The Influence of a Single Echo on the Audibility of Speech*, JAES Volume 20 Issue 2 pp. 146-159.

Manning, C.D. Raghavan, P. Schütze, H. (2008). *An Introduction to Information Retrieval*, Cambridge University Press, http://wwwcsli.stanford.edu/~hinrich/information-retrievalbook.html

MIREX 2007 - Music Information Retrieval Evaluation eXchange (2007), Retrived from http://www.musicir.org/mirex/2007/index.php/Main_Page, 2007.

Reis, G., Fonseca, N. and Fernandez, F. (2007). *Genetic Algorithm Approach to Polyphonic Music Transcription*, IEEE International Symposium on Intelligent Signal Processing (WISP 2007), Spain.

Reis, G., Fonseca, N., Fernandez, F., Ferreira A. (2008). *A Genetic Algorithm Approach with Harmonic Structure Evolution for Polyphonic Music Transcription*; ISSPIT 2008 - IEEE Symposium on Signal Processing and Information Technology; Sarajevo, Bosnia & Herzegovina.

Ryynanen, M.P. and Klapuri, A. (2005). *Polyphonic Music Transcription Using Note Event Modeling*, 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, NY.

Tavares T.F., Barbedo J.G.A., Lopes A. (2008), *Towards the evaluation of automatic transcription of music,* In: VI Congresso de Engenharia de Áudio 2008. p. 96-99.

Wang, D. and Brown, G.J. (2006). *Computational Auditory Scene Analysis, Principles, Algorithms and Applications:* IEEE Press.

Zwicker, E. and Fastl, H.(1999). *Psychoacoustics: facts and models*, Springer series in information sciences, 22. Springer, Berlin ; New York, 2nd updated edition.