EXEMPLAR-BASED ASSIGNMENT OF LARGE MISSING AUDIO PARTS USING STRING MATCHING ON TONAL FEATURES

Benjamin Martin, Pierre Hanna, Vinh-Thong Ta, Pascal Ferraro, Myriam Desainte-Catherine
LaBRI, Université de Bordeaux
firstname.name@labri.fr

ABSTRACT

We propose a new approach for assigning audio data in large missing audio parts (from 1 to 16 seconds). Inspired by image inpainting approaches, the proposed method uses the repetitive aspect of music pieces on musical features to recover missing segments via an exemplar-based reconstruction. Tonal features combined with a string matching technique allows locating repeated segments accurately. The evaluation consists in performing on both musician and non-musician subjects listening tests of randomly reconstructed audio excerpts, and experiments highlight good results in assigning musically relevant parts. The contribution of this paper is twofold: bringing musical features to solve a signal processing problem in the case of large missing audio parts, and successfully applying exemplar-based techniques on musical signals while keeping a musical consistency on audio pieces.

1. INTRODUCTION

Audio signal reconstruction has been of major concern for speech and audio signal processing researchers over the last decade, and a vast array of computational solutions have been proposed [6, 7, 9, 10]. Audio signals are often subject to localized audio artefacts and/or distortions, due to recording issues (unexpected noises, clips or clicks), or to packet losses in network transmissions, for instance [1]. Recovering such missing data from corrupted audio excerpts to restore consistent signals has thus been challenging for applicative research, in order to restore polyphonic music recordings, to reduce audio distortion from lossy compression, or to bring network communications robustness to background noise, for example [10].

The problem of missing audio data reconstruction is usually addressed either in the time domain, aiming at recovering entire gaps or missing excerpts in audio pieces, or in the time-frequency domain, aiming at recovering missing frequencies that cause localized distortions of audio pieces [18]. A typical trend for the latter one, often referred to as audio inpainting, is to treat distorted samples as missing and to attempt to restore original ones from a local analysis around missing parts. Common approaches include linear prediction for sinusoidal models [9], Bayesian estimators [7], autoregressive models [6] or non-negative matrix factorization solving [10]. These studies usually either base on the analysis of distributions of signal features around missing samples, or use local or global statistical characteristics over audio excerpts [18].

However, missing data problems are usually addressed on relatively small segments of audio data at the scale of audio piece duration. Indeed, most audio reconstruction systems proposed so far are based on signal features. The non-stationary aspect of such features makes it particularly difficult to assign data for large missing parts. Thus, audio gaps are generally reduced to a maximum duration of 1 or 2 seconds under particular conditions for the recovered quality to remain satisfying (see [9] for instance). In this paper, we address the challenging problem of reconstructing larger missing audio parts, namely audio gaps over several seconds (from 1 up to 16 seconds of missing data), in music audio pieces.

A similar problem is already addressed in image processing. Indeed, image inpainting aims at restoring and recovering missing data in images in a not easily detectable form (see for instance [2] and references therein). A common and simple approach, from texture synthesis, uses the notion of self-distance by considering that an image has a lot of repetitions of local information. This approach can be seen as an exemplar-based copy-and-paste technique [3, 5].

Similarly to exemplar-based image inpainting approaches, the proposed method analyses perceived repetitions in music audio to recover large missing parts. Note that while potentially allowing the reconstruction of large parts, such an exemplar-based approach induces the limit of reconstructing exclusively parts that are approximately repeated to maintain a musical consistency. To restore such an amount of

missing information, we consider the signal not only as audio excerpts but also as music pieces, therefore taking into account that sounds are temporally organized and may feature redundancies. Indeed, it is the organization and relationships between sound events in music that make music differ from random sound sequences [14]. In Western popular music, for instance, choruses and verses often are approximately repeated parts whose occurrences share a high degree of perceptual similarity. Other examples include classical music pieces, where the repetition of musical phrases structures the forms, or electronic music where repetitive loop techniques are frequently employed. We propose to use this kind of musical redundancy in order to recover missing data. Note that the method described in this paper aims at assigning a musically consistent part, and could be easily combined with signal-based approaches to be used for practical signal reconstruction of large missing parts.

Our method consists in representing each music piece as a sequence of tonal features employed to describe the perceived harmonic progressions. Then, a string matching technique is applied to retrieve the part that best fits the missing segment, according to its left- and right-sided tonal contexts. The identified repetition is finally used as a reference to fill-in missing data. Technical details of the method are described in Section 2. We detail in Section 3 the test protocol employed for evaluating the effectiveness of the system on human listeners and present the results obtained on musician and non-musician subjects. Section 4 finally brings concluding remarks and depicts future work.

2. METHOD

2.1 Musical representation

In a first step, audio signals are represented on musical-based criteria. The key to a well-suited representation in the particular application of finding perceived repetitions is to characterize some meaningful local variations in music while being robust to musical changes. As such, pitch content is particularly adapted to retrieve musical repetitions in the context of analyzing Western music. Indeed, harmonic and melodic progressions are constantly identified by listeners, consciously or not, and composers classically organize the whole structure of their pieces around such progressions and their variations or repetitions. Most state of the art methods dealing with musical structure analysis [16] or related to the detection of musical repetitions [11] rely on the richness of tonal information to retrieve similar segments. We therefore chose to use pitch-related features to represent audio pieces on their musical structure.

Harmonic Pitch Class Profiles (HPCP) are often used to describe this type of musical informations [8]. These features can be summarized as a classified representation of spectral energies into separate bins that correspond to the frequency class where they appear. The considered frequency classes take into account the cyclical perception of pitch in human auditory system: thus, two harmonic sounds contribute to the same chroma bin, or pitch class. Moreover, HPCP features were proven to be rather insensitive to non-pitched variations in noise, timbre, dynamic, tuning or loudness for instance, which makes them very efficient in qualifying only tonal contexts in audio pieces [8].

2.2 Tonal features extraction

Audio signals are first divided into n segments, or audio frames. We chose to use constant-length frames (as opposite to beat-synchronous windows, for instance) in order to optimize the proposed mono-parametric signal representation and to enable our system to be potentially used on diverse musical genres. Each frame is represented by a B-dimensional vector $h = (h_1, \cdots, h_B)$ that corresponds to a HPCP holding its local tonal context. The dimension value B stands for the precision of the note scale, or tonal *resolution*, usually set to 12, 24 or, in our case, 36 bins. Each HPCP feature is normalized by its maximum value; each vector h is thus defined on $[0, 1]^B$. Hence, each audio signal can be represented as a sequence $u = h^1 h^2 \cdots h^n$ of n B-dimensional vectors.

In the following process, we need a similarity measure to compare audio features between each other. The Pearson correlation measure r is better adapted to pitch class profiles comparisons than Euclidean-based measures, for instance, because it provides invariance to scaling. Such a measure then yields a good estimation of tonal context similarities [20], and is used in the following. It is defined as:

$$r(h^i, h^j) = \frac{\sum_{k=1}^{B} (h_k^i - \overline{h^i})(h_k^j - \overline{h^j})}{\sqrt{\sum_{k=1}^{B} (h_k^i - \overline{h^i})^2} \sqrt{\sum_{k=1}^{B} (h_k^j - \overline{h^j})^2}} \quad (1)$$

where $\overline{h^i}$ and $\overline{h^j}$ denote the mean value over the vectors h^i and h^j, respectively.

In the particular case of comparing HPCP features, an enhanced measure was proposed by Serrà *et al.* [17] based on the *Optimal Transposition Index* (OTI). The principle is to compute the local similarity measure, here r, between the first HPCP vector and each musical transposition (*i.e.*, circular shift) of the second compared vector. The OTI denotes the transposition index of the lowest distance found. Finally, according to the OTI, a binary score is assigned as the result of the comparison. In the case of a 12-split note scale ($B = 12$), for instance, a low cost is assigned to the OTI equals to 0 (no transposition was necessary: the local tonal context is similar) whereas a higher cost is given for any greater value of the OTI. Authors highlighted in their paper the superiority of such a binary measure over usual similarity metrics for HPCP. Based on this comparison technique, the similarity measure s employed for our system is:

$$s(h^i, h^j) = \begin{cases} \mu_+ & \text{if OTI}(h^i, h^j) \in \{0, 1, B-1\} \\ \mu_- & otherwise \end{cases} \quad (2)$$

where μ_+ and μ_-, are two possible scores assigned for the comparison of h^i and h^j.

The first representation step of our system thus computes an HPCP vector for each frame, which provides a sequence of chroma features that can now be treated as an input for string matching techniques.

2.3 String matching techniques

A *string* u is a sequence of zero or more symbols defined on an alphabet Σ. In our context, each HPCP vector represents a symbol. We introduce a particular "joker" symbol ϕ assigned to each frame that contains at least one missing audio sample. Thus, the alphabet considered in our context is denoted by $\Sigma = [0, 1]^B \cup \{\phi\}$. We denote by Σ^* the set of all possible strings whose symbols are defined on Σ. The i^{th} symbol of u is denoted by $u[i]$, and u can be written as a concatenation of its symbols $u[1]u[2]\cdots u[|u|]$ or $u[1\cdots|u|]$ where $|u|$ is the length of the string u. A string v is a *substring* of u if there exist two strings w_1 and w_2 such that $u = w_1 v w_2$.

Needleman and Wunsch [15] proposed an algorithm that computes a similarity measure between two strings u and v as a series of elementary operations needed to transform u into v, and represent the series of transformations by displaying an explicit alignment between strings. A variant of this comparison method, the so-called *local alignment* [19], allows finding and extracting a pair of regions, one from each of the two given strings, which exhibit the highest similarity. In order to evaluate the score of an alignment, several scores are defined: one for substituting a symbol a by another symbol b (possibly the same symbol), denoted by the following function $C_m(a, b)$, and one for inserting or deleting symbols, denoted by the function $C_g(a)$. The particular values assigned to these scores form the *scoring scheme* of the alignment.

The local alignment algorithm [19] computes a dynamic programming matrix M such that $M[i][j]$ contains the local alignment scores between the substrings $u[1\cdots i]$ and $v[1\cdots j]$, according to the recurrence:

$$M[i][j] = max \begin{cases} 0 & \\ M[i-1][j] + C_g(u[i]) & (\alpha) \\ M[i][j-1] + C_g(v[j]) & (\beta) \\ M[i-1][j-1] + C_m(u[i], v[j]) & (\gamma) \end{cases} \quad (3)$$

where u and v represent the two strings (HPCP sequences) to be compared, and with the initial condition $M[0][0] = M[i][0] = M[0][j] = 0, \forall i = 1 \ldots |u|, \forall j = 1 \ldots |v|$. (α) represents the deletion of the symbol $u[i]$, (β) represents the insertion of the symbol $v[j]$, and (γ) represents the substitution of the symbol $u[i]$ by the symbol $v[j]$.

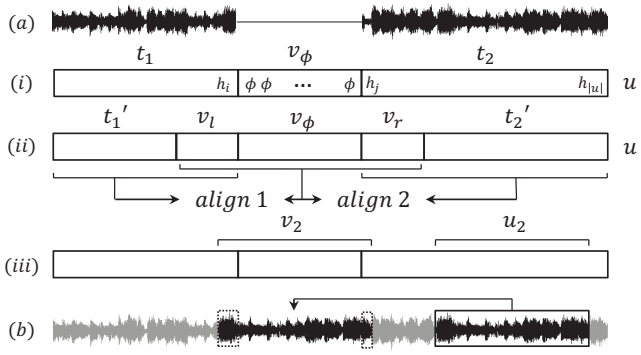

Figure 1. Overview of the algorithm. (a): audio waveform with missing data. (i): string provided by the musical representation step (Section 2.2). (ii): string alignments performed by our algorithm. (iii): aligned strings (Section 2.4). (b): reconstructed audio waveform. Dashed-circled regions correspond to an overlap-add reconstruction (Section 2.5).

In the following, the local alignment algorithm is denoted by the function $align(u, v)$. As a result, it yields a triplet (x, u', v') where x is the best similarity score between two strings, and u' and v' are the two aligned substrings respectively in u and v.

Considering two HPCP features h^i and h^j, the scoring scheme used in our experiments is defined as follows:

$$\begin{array}{l} \mu_+ = 1 \\ \mu_- = -0.9 \\ C_g(h^i) = -0.7 \quad \text{if } h^i \neq \phi, 0 \text{ otherwise} \\ C_m(h^i, h^j) = \begin{cases} s(h^i, h^j) & \text{if } h^i \neq \phi \text{ and } h^j \neq \phi \\ 0.1 & h^i = \phi \text{ xor } h^j = \phi \\ 0 & \text{otherwise} \end{cases} \end{array} \quad (4)$$

Numerical values were obtained empirically on a subset of 80 songs from the datasets presented in Section 3.2. The disjunction case for symbol ϕ is motivated by constraints over the alignment of frames that correspond to frames of missing data.

2.4 Algorithm

The general principle of our exemplar-based method is to identify in the partially altered music piece sequence the part that best fits the missing section. We call this best-fitting part the *reference part*. We denote as *local tonal context* tonal progressions that occur prior and after the missing part. More formally, we introduce a threshold δ that corresponds to the size of tonal contexts considered before and after the missing segment, as a number of frames.

Figure 1 depicts an overview of the applied algorithm. Formally, the computation is performed as follows:

(i) Let u be the string representing a music piece, *i.e.*, the HPCP sequence obtained from the signal representation step. By hypothesis, u contains a string $v_\phi = \phi \cdots \phi$ of joker symbols, and there exists t_1, t_2 in Σ^* such that $u = t_1 v_\phi t_2$.

(ii) Define as the left (resp. the right) *context string* v_l (resp. v_r) of v_ϕ the unique string of length δ such that there exists t_1' and $t_2' \in \Sigma^*$ verifying $t_1 = t_1' v_l$ and $t_2 = v_r t_2'$. Compute (x_1, u_1, v_1) as the result of $align(t_1, v_l v_\phi v_r)$ and (x_2, u_2, v_2) as the result of $align(t_2, v_l v_\phi v_r)$.

(iii) If $x_1 > x_2$, then keep u_1 as the reference part, u_2 otherwise.

This process provides both a *reference part* u' (u_1 or u_2) corresponding to the excerpt that best fits the missing section, and a *destination part* v' (v_1 for u_1, v_2 for u_2) that was aligned with u'. Note that the scoring constraints described in Eq. 4 ensure that the identified part v' contains the missing segment v_ϕ.

2.5 Audio data assignment

In order to fill-in missing data, the method consists in assigning data from the identified reference part into the destination part. Since the identified destination part v' may be longer than the missing data segment v_ϕ, the samples assignment may overlap existing samples in the audio piece. In order to ensure a smooth audio transition, overlap-add reconstructions are performed [4].

Note that we deliberately chose not to implement any beat, onset or any kind of synchronization, in order to avoid the addition of potential analysis errors and to enable the strict evaluation of this exemplar-based audio alignment method. We leave as a perspective such more advanced audio synchronizations or overlapping techniques.

3. EXPERIMENTS AND RESULTS

Our alignment system is based on musical features. The identified repetitions only depend on a musical criterion: pitch content. Therefore, variations in timbre, rhythm or lyrics may appear between occurrences of an identified repetition and original and reconstructed audio signals may be completely different. Hence, standard signal processing metrics such as SNR seem inadequate to the evaluation of musical resemblance. Since it works on a musical abstraction, the aim of the method is to produce perceptually consistent results, *i.e.*, reconstructions satisfactory for human listeners. The proposed experiments are therefore based on human subjective evaluation of reconstructed audio files.

3.1 Test data generation

The tests of our method consist in erasing random audio parts in a dataset of music pieces, recovering missing data with our system and asking human listeners to evaluate the audio reconstruction. Since our method uses an exemplar-based approach, a part needs to be approximately repeated in the same piece at least once in order for our system to recover it. Thus, we introduce a *repetitiveness hypothesis* prior to the evaluation of the proposed system: every concealed part for audio tests must belong to a repeated structural section, according to a structural ground truth. For instance, for a music piece annotated with the structure ABCAAB, the hypothesis force concealed parts to be chosen within one of the repeated patterns A, B or AB.

The test data generation is performed according to the following process:
1. Select randomly a concealment length l between 5 and 16 seconds.
2. According to an annotated structural ground truth, select randomly a repeated section lasting at least l.
3. Select randomly a beginning time instant d in this chosen part.
4. Perform the concealment: erase every sample between d and $d + l$.
5. Perform the reconstruction using the algorithm described in Section 2.4.
6. Finally, select two random durations t_1, t_2 between 5 and 10 seconds, and trim the reconstructed audio piece between $d - t_1$ and $d + l + t_2$.
The last step is dedicated to reducing the duration of excerpts in order to reduce the test duration. Note that whereas this last step makes the experiment more comfortable (faster) for the testers, it tends to sharpen up their attention around to the reconstructed region, and requires the reconstruction to be specially accurate.

3.2 Dataset

As a test dataset, we elected the OMRAS2 Metadata Project dataset [13] that provides structural annotations for Western popular audio music of different artists [1]. For our experiments, we chose to test on 252 music pieces mostly from *The Beatles* (180 pieces), *Queen* (34 pieces) and *Michael Jackson* (38 pieces). These artists were most likely to be known by listeners, hence reinforcing their judgment. Note that audio pieces were taken from mp3-encoded music collections compressed with a minimum bit-rate of 192 kbps.

In order to compute HPCP features on audio signals, we chose the window size of $46ms$ in order to keep accurate alignment on audio data. Performing preliminary tests on a few songs, the local context threshold value of $\delta = 4$ seconds appeared to be sufficient for consistent alignments.

[1] http://www.isophonics.net/content/reference-annotations

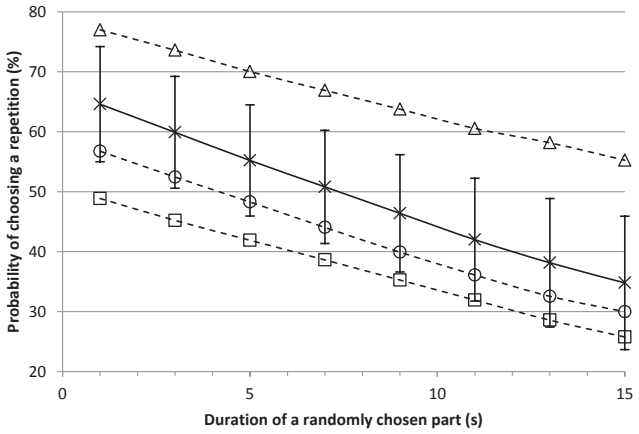

Figure 2. Probability of randomly choosing repeated parts according to the ground truth. Plain line shows the average values over the whole dataset, while dashed lines stand for the different artists' songs: square points for *Queen*, circle points for *Michael Jackson* and triangle points for *The Beatles*.

To evaluate how restrictive the repetitiveness hypothesis may be on this specific dataset, we computed the average percentage of parts in audio pieces that are repeated according to the structural ground truth. Figure 2 shows the average probability of finding a repetition as a function of the size of the randomly chosen part. The plain line shows the average values over the dataset. The graphic shows for instance that a random part that lasts 8 seconds corresponds to a fully repeated section in structural ground truth 48% of the time on average. Repetitiveness seems to vary between artists in the dataset, as suggested by the different dashed lines. Thus, the probability of finding repeated parts in pieces from *The Beatles*, for instance, is between 8.7% and 16.2% higher than on pieces from *Queen*. The hypothesis of deleting exclusively random parts inside repeated sections therefore induces the consideration of 35% of 15 seconds parts in audio pieces, to 65% for 1 second parts on average.

The previously described data generation process was performed once for each music piece in the dataset. 252 excerpts were thus generated, each lasting between 10 and 30 seconds, with an average duration of 21.8 seconds over the set. The artificial data concealment durations were randomly generated between 1 and 16 seconds, with an average value of 8.2 seconds.

3.3 User tests

The test protocol employed for evaluating our system is inspired from the MUSHRA audio subjective test method [12]. In order to respect a maximum test duration of approximately 10 minutes, each subject is asked to listen for 26 audio excerpts from the generated test dataset. Among these, 5 excerpts are proposed in every test and correspond to non-altered audio excerpts. These are supposed to observe individual effect, enabling for instance the detection of randomly answering subjects. The 21 remaining excerpts are randomly chosen among the reconstructed database. Each subject is asked to listen to each of these excerpts once, with no interruption, and to indicate whether or not he detected any audio artefact or distortion. If so, the subject is asked to rate the quality of the reconstruction applied: 1) Very disturbing, 2) Disturbing, 3) Acceptable, 4) Hardly perceptible. The rate of 5 is assigned for no distortion heard. Note that the exact meaning of terms in the context of the experiment is not provided to the testers, hence letting them define their own subjective scale. Finally, a few additional information is asked, such as which audio restitution material is used, and whether or not the tester is a musician.

3.4 Results

Tests were carried out on 80 distinct listeners, 34 musicians and 46 non musicians. The average number of observations per audio excerpt is 7.1, values ranging from 1 to 15 observations for altered excerpts. The 5 common non-altered pieces logically led to 400 observations among which 10 were incorrectly evaluated (artefacts perceived). Since all of these invalid rates were attributed by distinct users, we chose to take into account every subject in the evaluation (no abnormal behavior). Table 1 summarizes the results obtained for both classes of testers and for the different artists in the dataset. Note that the rates attributed to the 5 non-altered excerpts were not used for computing these average values. Overall results highlight an average rate of 4.04 out of 5 for the quality of the applied data assignment. More precisely, 30% of reconstructed excerpts were attributed the rate 5 by all of their listeners, which highlights very accurate audio assignments on a third of the dataset. The distribution of other average rates is as follows: 31% pieces rated between 4 and 5, 17% pieces between 3 and 4, 15% between 2 and 3 and 7% between 1 and 2. Reminding that 4 corresponds to a "hardly perceptible" reconstruction and 5 to no distortion perceived, the method therefore seems successful in performing inaudible or almost inaudible reconstructions in 61% of the cases.

As one could expect, musician subjects perceive more distortions with an average rate of 3.92 against 4.13 for non musicians. Scores obtained for each audio material class highlight a slightly better perception of reconstructions for headset restitution, with an average value of 3.98 against 4.05 for other material. However, since all musician testers chose to use headset, musician and headset scores may be closely related. Reported distortions include short rhythmic lags, unexpected changes in lyrics, sudden changes in dynamics or abrupt modification of instruments. Results

	Musicians	Non musicians	Total
The Beatles	3.95	4.13	4.05
Michael Jackson	4.21	4.26	4.24
Queen	3.40	3.94	3.71
Whole dataset	3.92	4.13	4.04

Table 1. Audio test results. Values correspond to average rates on a 1 (very disturbing reconstruction) to 5 (inaudible reconstruction) scale.

also vary between artists; for instance, reconstructions on *Michael Jackson* songs seem to be better accepted, with an average value around 4.24 whether listeners are musicians or not. Contrastingly, reconstructions on *Queen* pieces were more often perceived, with an average value of 3.94, and musicians assigned a 0.5 lower rate on average. An explanation for such gaps between artists may be the more or less repetitive aspect of similar structural sections, such as choruses that tend to vary often along *Queen* music pieces. Moreover, a few pieces such as *We will rock you* by *Queen* were assigned particularly low rates (1.25 in this case for 8 observations) probably because their pitch content is insufficient for the algorithm to detect local similarities.

4. CONCLUSION AND FUTURE WORK

In this paper, we addressed the problem of reconstructing missing data in large audio parts. We used a tonal representation to obtain a feature sequence on a musical criterion, and analyzed it using string matching techniques to extract a musically consistent part as a reference for substitution. We generated audio test data introducing random concealments between 1 and 16 seconds long in repeated structural parts, and tested out our music assignment system in an audio evaluation on 80 subjects. Results highlighted a good performance of the method in recovering consistent parts with 30% random reconstructions undetected, and 31% hardly perceptible.

As a future work, in order to make this method useful in practice, the algorithm may be combined with other signal-based approaches. For instance, audio synchronizations could be applied by aligning assigned beats with original ones. Other possible audio improvements include the correction of dynamics, or the combined use of other musical descriptions (timbre features, rhythm, *etc.*). We also leave as a perspective the improvement of the comparison algorithm, which could retrieve a set of parts locally fitting the missing data section and combine such parts iteratively, or the development of an inspired approach performing real-time audio reconstruction.

5. REFERENCES

[1] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M.D. Plumbley. Audio inpainting. Research Report RR-7571, INRIA, 2011.

[2] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. *Proc. of SIGGRAPH*, pp. 417–424, 2000.

[3] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. on Image Processing*, v. 13, pp. 1200–1212, 2004.

[4] R. Crochiere. A weighted overlap-add method of short-time fourier analysis/synthesis. *IEEE Trans. on Acoustics, Speech and Signal Processing*, v. 28, pp. 99–102, 1980.

[5] A.A. Efros and T.K. Leung. Texture synthesis by non-parametric sampling. *Proc. of ICVV*, p. 1033, 1999.

[6] W. Etter. Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters. *IEEE Trans. on Signal Processing*, v. 44, pp. 1124–1135, 1996.

[7] S.J. Godsill and P.J.W. Rayner. *Digital Audio Restoration - A statistical model based approach*. 1998.

[8] E. Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra, 2006.

[9] M. Lagrange, S. Marchand, and J.B. Rault. Long interpolation of audio signals using linear prediction in sinusoidal modeling. *Journ. of the Audio Engineering Society*, v. 53, pp. 891–905, 2005.

[10] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama. Computational auditory induction by missing-data non-negative matrix factorization. *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, 2008.

[11] B. Martin, P. Hanna, M. Robine, and P. Ferraro. Indexing musical pieces using their major repetition. *Proc. of Joint Conference on Digital Libraries*, 2011.

[12] A.J. Mason. The MUSHRA audio subjective test method. *BBC R&D White Paper WHP*, 38, 2002.

[13] M. Mauch, C. Cannam, M. Davies, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler. Omras2 metadata project 2009. *Proc. of ISMIR, Late-Breaking Session*, 2009.

[14] R. Middleton. *"Form", Key Terms in Popular Music and Culture*. Wiley-Blackwell, 1999.

[15] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journ. of Molecular Biology*, v. 48, pp. 443–453, 1970.

[16] J. Paulus, M. Müller, and A. Klapuri. Audio-based music structure analysis. *Proc. of ISMIR*, pp. 625–636, 2010.

[17] J. Serrà, E. Gómez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Trans. on Audio, Speech and Language Processing*, v. 16, pp. 1138–1151, 2008.

[18] P. Smaragdis, B. Raj, and M. Shashanka. Missing data imputation for time-frequency representations of audio signals. *Journ. of Signal Processing Systems*, pp. 1–10, 2010.

[19] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journ. of Molecular Biology*, v. 147, pp. 195–197, 1981.

[20] D. Temperley. *The Cognition of Basic Musical Structures*. p. 175, 2004.