TOWARDS A (BETTER) DEFINITION OF THE DESCRIPTION OF ANNOTATED MIR CORPORA

Geoffroy Peeters
STMS IRCAM-CNRS-UPMC
Paris, France

Karën Fort
INIST-CNRS & Université Paris 13,
Sorbonne Paris Cité, LIPN
Nancy, France

ABSTRACT

Today, annotated MIR corpora are provided by various research labs or companies, each one using its own annotation methodology, concept definitions, and formats. This is not an issue as such. However, the lack of descriptions of the methodology used—how the corpus was actually annotated, and by whom—and of the annotated concepts, i.e. what is actually described, is a problem with respect to the sustainability, usability, and sharing of the corpora. Experience shows that it is essential to define precisely how annotations are supplied and described. We propose here a survey and consolidation report on the nature of the annotated corpora used and shared in MIR, with proposals for the axis against which corpora can be described so to enable effective comparison and the inherent influence this has on tasks performed using them.

1. INTRODUCTION

The use of annotated data usually corresponds to increasing performances in a field of research, as has been seen in the cases of speech and language processing. The accessibility of novel annotated data usually corresponds to the initiation of a number of research activities in a field. This is the case of music genre, chord recognition, and music structure in music information retrieval (MIR). For this reason, annotated data can be considered to be a major issue in MIR. In MIR, there is currently no dedicated institution responsible for providing music corpora comparable to ELRA [1] or LDC [2] in the speech and natural language processing (NLP) community. Instead, corpora are provided individually by various research labs and companies. While recent years have seen a large increase in corpora creation initiatives (e.g. Isophonic, SALAMI [3], Billboard, and Quæro), each research lab or company uses its own annotation methodology, concepts definition, and format. This is not a problem in and of itself, but the lack of descriptions of the methodology used, i.e. how the corpus was actually annotated, or of the concepts annotated, i.e. what is actually described, presents problems with respect to the sustainability, usability, and sharing of corpora. Therefore, it is essential to define exactly what and how annotations of MIR corpora should be supplied and described. We propose here an avenue to improve this situation by defining a methodology for describing MIR corpora and the implicit or explicit assumptions made during their creation. It should be noted that similar initiatives have been taken in the speech and NLP community to favor sharing and exchange of corpora (see for example [1], [2] [3]) leading to descriptions close to the one proposed here.

[1] `http://www.elra.info/`
[2] `http://www.ldc.upenn.edu/`
[3] Structural Analysis of Large Amounts of Music Information

2. DEFINING AN ANNOTATED MIR CORPUS

In the following, by annotated corpora, we mean "musical audio data with annotations". Such corpora can be used for research purposes to derive knowledge or train systems, or for benchmarking and evaluation projects, both internal and public, as in MIREX [4]. Creating an annotated MIR corpus involves:

(A) choosing or creating a set of audio items (denoted by "raw corpus" in the following),
(B) creating and/or attaching related annotations, and
(C) documenting and storing the results to ensure sustainability and sharing.

While these points may seem obvious, each of them involves making choices that in the aggregate will define what exactly the corpus is about, what use it is for, and what the underlying assumptions behind it are. In the following, we provide insights about the choices that must be explicitly or implicitly made for each of these points, and the implications of those choices. Figure 1 summarizes the various aspects of the proposed description.

2.1 (A) Raw Corpus

In the case of "audio MIR," the annotations describe audio items [5], which we denote here by the term "raw corpus", as opposed to the "annotated corpus". The choice of these audio items defines the domain, or musical area, for which the **results derived from** the annotations—results

[4] MIREX: Music Information Retrieval Evaluation eXchange
[5] Whether the annotations are distributed with or without the audio items, the following remains true.

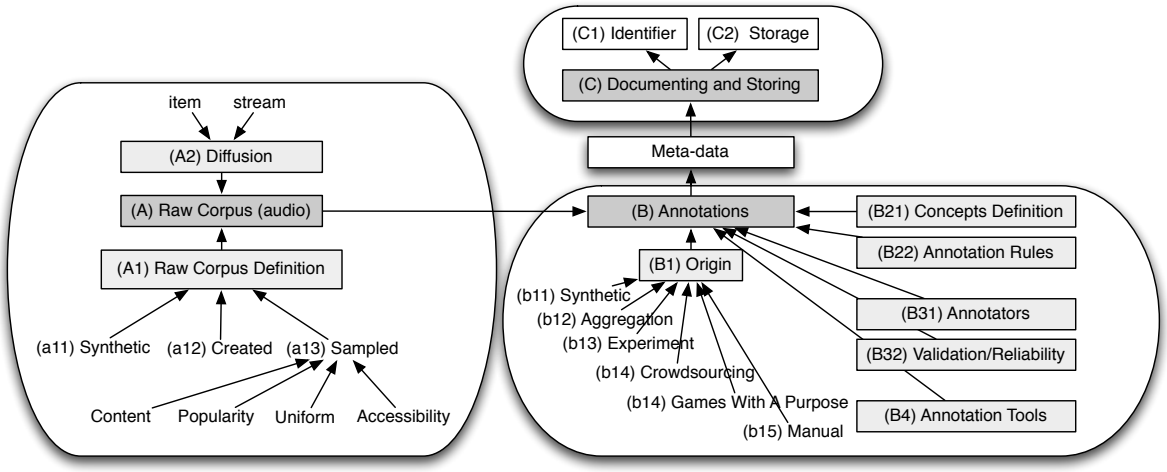

Figure 1. Decomposing the creation of a MIR annotated corpus into the tasks and sub-tasks involved.

from an experiment, training, or evaluation—are valid. For example, results derived from the music genre defined for the Tzanetakis test-set [4] do not generalize to the Million-Song test-set. The choice of audio items also determines the domain for which the **concepts defined by** the annotations are valid. This is specific to the way annotation is performed in the MIR field: while in other domains, concepts are first defined, and then used for the annotation of items, in MIR, the concepts are (in most cases) defined by the annotations themselves. For example, "music genre" is not defined textually, but rather is defined by its application to a specific test-set, such as Tzanetakis or Million Song. The same is true for "chords," whose meaning may differ in the context of the data to which it is applied: in the Beatles [5], it refers to guitar chords, but when applied to Billboard songs, it is a reduction of the overall harmonic content. Because of this, special care must be taken when selecting audio items for an annotated MIR corpus.

It is clear that results obtained from experiments on (i) synthesized MIDI files, (ii) audio recorded for the purposes of an experiment, or (iii) audio as sold by online music services, do not have the same impact. Note that this in no way means that one is better than the others.

To help describe the choices made so far in MIR, we propose to distinguish between three categories:

- a11 - artificial audio items made specifically for the purpose of representing an annotation
- a12 - real audio items recorded specifically for the purpose of the creation of a corpus
- a13 - real audio items sampled from the real world

2.1.1 (a11) Corpus of Synthetic Items

This kind of corpus is specific to the music research community, based on the assumption that, within certain limits, rendering a MIDI file can create a music audio signal. Examples of this are [6] for the creation of a multi-pitch estimation corpus, [7] for the case of chords, and the MIREX corpus used for key estimation.

There are several *advantages* to this approach. It allows (i) having close to perfect annotations very easily, since audio can be partly considered to be a direct instantiation of the annotation; (ii) having full control over the audio rendering process, such as testing the influence of instrument changes, reverberation or delay; and (iii) rapidly creating a large corpus. Its major *drawback* is the lack of realism due to (i) the absence of interpretation by musicians; (ii) the absence of realism due to sound variations, propagation, and capture by microphone; and (iii) the absence of "production" as made in recording studios.

2.1.2 (a12) Corpus of Created Real Items

The second trend consists in creating specific audio items for the purpose of research. The first corpora prepared in this way were built for "instrument samples" research—McGill [8] and Studio-On-Line [9]. In this case, the annotation—pitch, instrument name, and playing mode—is added during the recording session. Corpora for multi-pitch, source separation, and event recognition have also been created, such as the ENST Drum database [10], containing audio and video, and the MAP database [11], using a Yamaha disklavier for automatic pitch annotations. The most well-known and used corpus in MIR research is such a data set—the RWC corpus [12, 13].

The *advantages* of this approach are that it allows (i) complete specification of the underlying content property, (ii) easy creation of the annotations at the same time as the occurrence, and (iii) distribution of the corpus with no restrictions, as the creator of the corpus usually owns the audio copyright. The main *drawback* of this is again the lack of realism of the resulting audio items—e.g., RWC is a very valuable resource but does not sound like iTunes music. This is partly due to the recording conditions that a lab can afford—expensive compressors and enhancers remain in the big production studios. Also, the music composition used is often prototypical. All of this frequently creates a bias when using these corpora outside of the context of the experiment for which they were built.

2.1.3 (a13) Corpus of Sampled Real Items

The last trend corresponds to what has long been known as "using a private music collection". These could not be shared, mostly due to copyright issues. Today, because of the possibility of referencing audio items by ID (CD reference or Musicbrainz/ EchoNest/ Amazon/ 7-Digital ID), there is a major trend toward these corpora. The main *advantage* of this type of corpus is that it represents exactly the music people listen to or buy, with artistic interpretation and professional sound production. It also allows the evaluation of concepts that are well-established in the literature for their applicability to everyday music (see the case of the "chord" concept). The major drawback of this type of corpus is the cost of the annotations, which involve either human annotation (by dedicated people or by crowdsourcing) or data aggregation (for example, aggregating guitar-tab collections or music-recommendation sites).

However, underlying a corpus created by sampling the real world lies a major question: how was the sampling done? For which reasons or purposes were the specific music tracks selected? This is actually rarely described, with the exception of [14], which provides an in-depth description of the sampling for the Billboard corpus. We distinguish here between four trends:

Specific-content sampling: The sampling is done in order to highlight specific content characteristics. An example of this is the corpus proposed by [15] for music structure. It consists of a selection of tracks from Eurovision (the European Song Contest), i.e. pop songs with a typical pop structure. Another is the corpus proposed by [5] for chord annotation, which consists primarily of a selection of tracks from The Beatles, essentially made of guitar chords. While this perfectly fits the purpose of their annotations, care must be taken with respect to the validity of the concepts (e.g. the specific definition of structure or chords) outside of the context of these corpora.

Popularity-oriented sampling: The sampling is done according to what people were or are reading, listening to, or watching the most. An example is given in [14], in which the sampling is performed based on the Billboard charts. However, in this case, some music genres might be over-represented.

Uniform sampling: The sampling is done in a uniform way according to a description grid. The dimensions of this grid, as in our project, may represent music genre/style, year, or country [6]. In each resulting cell of the grid, the most popular audio items are selected. In this case, some music styles can be over-represented.

Accessibility-oriented sampling: The last trend consists in selecting items because they are freely available (e.g. Magnatune and Internet Archive), without any other considerations.

2.1.4 (A2) Type of Media Diffusion

Apart from the choice of the sampling process, the type of media diffusion also needs to be decided during the process of corpus creation. Corpora can represent isolated music tracks, but may also include items as diverse as music inside a TV/Radio audio stream (as in the corpus of [16]), the audio part of a video clip or User-Generated-Content videos, a live recording, a bootleg, or a DJ-remix. This implies different audio qualities, and also the possible presence of interfering sounds such as speech, applause, and the ambient atmosphere of live performances.

2.1.5 Definition of the Media Coding Properties

Finally, the audio properties also have to be described, in terms of such variables as frequency bandwidth; the presence of drops, noise, pops, hisses or clicks (due for example to media trans-coding from vinyl); and the number of channels—mono, stereo, or multi-channel.

2.2 (B) Attaching Annotations to Audio Items

Although this is probably the most important aspect of an annotated corpus, it is often the one that is least described (except if the annotations were the subject of a dedicated publication, as in the case of the results of a listening experiment [17]). The main points to detail are the following:

- Where do the annotations come from?
- What do they represent? How are they defined?
- What is their reliability?

2.2.1 (B1) Origin of the Annotations

The central question is the origin of the annotations. We distinguish here between four different cases:

Automatic annotations

• (b11): The annotations are obtained by the **synthesis** parameters [6] (a11), as scores given during the recording process or analysis of the individual tracks of the recordings [13] (a12). In this case, the generative process of the music defines both the labels used for the annotation and the annotation itself. Its reliability is very high.

• (b12) The annotations are obtained by **aggregation** of diverse extant content. Examples of this are the Million Song Test-Set [18] and the use of Guitar-Tab in [19]. In this case, each annotation and its definition and reliability are defined by its provider: Last-FM data are obtained through crowdsourcing, Echo-Nest data are algorithm estimations, and MusicXMatch contains official lyrics.

Manual annotations:

• (b13): The annotations are the results of an **experiment**. In this case, the definition of the annotation is provided by the guidelines of the experiment. The reliability of the annotation is derived from the experimental results, either in a summarized form (e.g. two major peaks of the tempo histogram in [17]) or from the whole set of annotations, letting the user decide the way to summarize it (e.g. perception of tempo and speed in the case of [20]).

• (b14): **Crowdsourcing**, in particular **Games With A Purpose** (GWAP). In this case, annotations are obtained using various game processes [21–24]. The labels used for the annotation are either determined before the game, providing an existing frame of reference; or determined by the users during the game, allowing free input. In both cases, the definitions of the labels are not provided (although they

[6] These meta-data can be provided, e.g. by AMG. It should be noted, however, that the source of meta-data can create a bias.

may be inferred by another gamer's choices), but rather are defined by the use that gamers make of them. In this context, when a reliability measure of the annotation is proposed, it is usually derived from the number of occurrences of a label [24].

• (b15): Traditional **manual** human annotations. Examples of these are [5, 25, 26].

(b13), (b14) and (b15) are the most interesting for us here, since they involve thinking about the definition of the annotation concepts and the techniques for performing the annotations and measuring their reliability. Manual annotation is (very) costly, so the annotation process should ensure quality and reusability. In the field of natural language processing, the authors of [27] show that "corpora that are carefully annotated with respect to structural and linguistic characteristics and distributed in standard formats are more widely used than corpora that are not".

2.2.2 (B2) Definitions

(B21) Concepts: The term *annotation* refers both to the process of adding a note or a label to a flow of data (such as audio music, speech, text or video) and to the result of this process—the notes themselves, anchored in the source flow. The annotations are all the more useful to the extent that they are designed for a specific application [28]. Depending on the final application, the labels may not carry the same semantics. The semantics may even be completely different—for example, annotating football matches with the intent of producing an automatic summary [29] is very different from annotating football matches for purposes of linguistic analysis. In speech and natural language processing, saying that we may find as many annotation models as there are annotation projects is not too far from reality. In MIR, it seems that the same concepts are always used, with different meanings that are sometimes only implicit.

In the case of manual human annotations, the concepts to be annotated must be defined. The absence of definition is clearly a problem for a set of tasks in MIR (beat [7], chord [8], and structure [9], to name just a few). Recently, efforts have been made to clarify the concepts being annotated through dedicated papers or through the on-line availability of so-called "annotation guides" [15, 26]. Those efforts should be encouraged. It must be noted that the use of annotation guidelines has been considered part of "best practices" in speech and natural language processing for some time, following the trend in this direction in corpus linguistics [28].

(B22) Rules: Beyond the definition of the concepts being annotated, the annotations are performed using a set of rules. This set of rules should also be described. For example, what is the temporal precision used for segment annotations? Which type of dictionary was used for the labels? Are there equivalences between labels? To exemplify the difference between *concept* and *rules used to annotate this concept*, consider an experiment in a recent project to annotate beat/tempo. Two different rules were used. The first was to do annotation of beats and then infer the tempo curve from that; the second was to adjust a tempo curve so as to align a sequencer grid to an audio track, and then infer beat positions. The two methods describe the same concept, but lead to different results (data not shown).

2.2.3 (B3) Actors and Quality

(B31) Who are the annotators? Annotators may be students or researchers, creating a corpus that will directly fit their research, with the model of their algorithm in mind while annotating; musicians, with a strong ability to apply the concepts with respect to detailed musical structure, sometimes losing sight of overall perception; or everyday people. This choice influences the way the annotation is performed.

(B32) Reliability of the annotation? Although they are considered to be able to generate "gold standards", humans are not perfect annotators. The definitions of the concepts to be annotated might not have been defined clearly, they may not fit the content of a given audio file, there might be several plausible possibilities for a particular annotation, or the annotator may lose concentration. The question of the reliability of the annotation is therefore another major issue. For this reason, it is common practice to do cross-validation of the annotations. This can be done by applying either or both of two scenarios. In the first scenario, an annotated track is validated or corrected by a second annotator. In the second scenario, the same track is annotated independently by at least two annotators. The resulting annotations are then compared by computing the inter-annotator agreement (using the Kappa coefficient [31] or other measures [10]). A decision is then made whether the annotation is sufficiently reliable, or whether it should be redone using the same definitions and rules, or whether the definitions or rules should be modified. In speech and natural language processing, computing the intra-annotator agreement (agreement of an annotator with him/herself as the project progresses) is also considered to be good practice and allows the detection of potential issues with the annotators [33]. This is already done in sound perception experiments, and could be extended to annotation projects.

Overall, the methodology used should be documented and detailed. In speech and natural language processing, the typical methodology includes early evaluation of the annotation guidelines using inter-annotator agreement, the update of these guidelines with the help of the annotators' feedback, regular checking, continuous use of inter- and

[7] Given that beat is mostly a perceptual concept, what is the metrical level being annotated?

[8] In the case of chord annotations, what is the definition of chords? Are we considering the perceived chord of the background accompaniment? Do we also consider the lead vocal? Are the chords derived from the guitar part?

[9] The case of music structure is even less defined. What is a chorus? A segment? Why could a segment not be further divided into sub-segments or grouped into meta-segments? Considering this, the proposal made in [30] to store the various possible annotations is worth mentioning.

[10] It must be noted, however, that the resulting coefficient of agreement (Cohen's Kappa or others) is far from being wholly sufficient as a metric when used in isolation, and should be accompanied by details of the choices that were made to compute it. In this respect, the contingency table provides more interesting information about the annotation reliability than the inter-annotator agreement itself [32].

intra-annotator agreement and/or precision measures (going so far as the so-called "agile annotation" [34]), and a final evaluation of the resource that has been produced.

2.2.4 (B4) Annotation Tools

Caution is necessary in selecting the appropriate annotation tool, as the limitations of the tool will impact the annotation model. For example, there may be relations that are impossible to annotate, or the interface may contain a feature that is difficult to access and hence seldom used.

2.3 (C) Documenting and Storing the Results to Ensure Sustainability and Sharing

Corpus sharing and distribution does not simply require putting all of the audio data and annotations into an archive file. From our point of view, it implies providing information on all of the above-mentioned points (A* and B*). We provide here some additional recommendations for improving the distribution process.

2.3.1 (C1) Corpus Identification

Currently, most corpora in MIR research or MIREX benchmarking have no identifier (except RWC, Isophonic, or Million Song Test-Set). They are referred to as "the corpus used in the publication of [reference]". A unique identifier should be assigned to each corpus, including versioning of the annotations and annotation guidelines (see [35]). This could take the form of a simple URI (example of this would be corpus:MIR:qmul:2004:beatles:chords:version1.0) or used the more elaborated Vocabulary of Interlinked Datasets [11] . This would solve some ambiguity issues, such as when a corpus is updated over time (for example the SALAMI corpus), or when one set of annotations is revised by another lab and later included in a new corpus (for example the structure annotations of The Beatles).

2.3.2 (C2) Storage of the Created Annotations

Annotations must be sustainable. We therefore recommend that the storage of the data make their semantics explicit. Up to this point in time, many annotations of local-in-time concepts such as beat, chord, and structure were done in formats where the semantics is implicit in the corpus. In particular, the so-called ".csv" or ".lab" formats (one row for time, one row for labels) would not be sustainable outside of the context of a given corpus [12] . RDF (as used by QMUL [5]) or XML (as used by [30]) seem good choices. For the later, the MPEG-7 xml shema [36] already proposes a full-range of description with inherent semantic and the possibility to define new semantics using Classification Schemes (CS). Whatever choice, the definition of and the reference to a controlled list of labels is necessary. It also allow to define the width of the description-space [13] .

Providing a precise reference to the audio items being described is also crucial. Considering that recent annotated corpora were distributed without audio media, this is clearly a major issue. Several linkage mechanisms between annotations and audio media have been proposed so far: CD reference (as in Isophonic), Musicbrainz ID (as in the Million Song) or the EchoNest ID, Amazon ID, 7-Digital ID. The reference should also allow referencing time inside the files. The example of the alignment problem of the Beatles annotations to the various possible audio instances is notorious in the MIR community. Inclusion in the annotation of time-stamped identification, such as is provided by audio-ID techniques, would help.

(C1) Corpus ID: corpus:MIR:AIST:RWC:2006:version1.0
(A) Raw Corpus
(A1) Definition: (a12) created real items; 315 tracks created for the specific purpose of having a copyright-free test-set for MIR research representative of the various genres, styles, instrumentation, vocal types (see [12] for details)
(A2) Type of media diffusion: full tracks stereo high-quality
(B) Annotations
(B1) Origin: (b11) synthetic—obtained during creation and (b15) manual annotations
(B21) Concepts definition: only defined by the annotation rules
(B22) Annotation rules: - Standard MIDI Files (SMF) transcribed by ear, - Lyrics of songs obtained during creation, - Beat/ downbeat annotated using metronome clicks of recording and manual editing, - Melody line annotated using fundamental frequency estimation on the melody track and manual editing, - Chorus sections method is not indicated, - Audio synchronized MIDI Files using the annotated beat positions
(B31) Annotators: a music college graduate with absolute pitch
(B32) Validation/ reliability: not indicated
(B4) Annotation tools: "Music Scene Labeling Editor"
(C) Documents and Storing
(C2) Audio identifier and storage: RWC-specific audio identifiers, audio files are available through audio CDs, annotations available through archive files in CSV format

(C1) Corpus ID: corpus:MIR:LastFM:Tempo:2011:version1.0
(A) Raw Corpus
(A1) Definition: (a13) sampled real items. Sampling method: somehow uniform—4006 tracks chosen "essentially at random" among several thousands
(A2) Type of media diffusion: 30s extract of music items
(B) Annotations
(B1) Origin: (b13) Experiment and (b14) Crowd-Sourcing
(B21) Concepts definition: the concepts are defined by the results of the experiments, itself defined by the instructions provided to the annotators: "tap along to each except", "describe its speed on a three point scale", compare two tracks in terms of speed.
(B22) Annotation rules: defined by the experiment protocol (see [20] for details)
(B31) Annotators: 2141 users of Last-FM (not all tracks are annotated by all the annotators)
(B32) Validation/ reliability: for each track, all the annotations are provided, it is let to the user of the corpus to compute inter-annotator agreement
(B4) Annotation tools: Web-interface
(C) Documents and Storing
(C1) Audio identifier and storage: no audio identifiers are provided (except the artist, album and track name); annotations distributed as an archive file accessible through an URL, files in TSV format (Tab Separated Values File).

Table 1. Application of the proposed description to the corpus of [12, 13] and [20]

[11] http://www.w3.org/TR/void/

[12] Consider the question of how a user will interpret the "1" label ten years from now.

[13] This would for example make it possible to decide whether a C-Maj chord is really a C-Maj or a reduction of a C-Maj7+9 chord.

3. EXAMPLES OF DESCRIPTIONS

As examples of the application of the proposed description, we illustrate in Table 1 its use for the (short) description of two corpora [12, 13] and [20]. It should be noted that these descriptions are solely based on the information provided with the distributed corpora and the respective publications and should ideally be complemented and corrected by the respective authors themselves. Based on this, a comparative table of the corpora can easily be made [14] .

4. CONCLUSION

MIR should benefit from the "best practices" that have been evolving for decades in the speech and natural language processing communities. Among these practices, we attempt here to provide insights into the choices currently made when creating a MIR annotated corpus, their implications, and the resulting necessity to better describe them when distributing an annotated corpus. We presented them in the form of a numbered list—A*, B*, C*—to highlight the fact that all of these choices must be described. Considering the importance that the distribution of annotated corpora will have to the development of MIR research, we hope that providing this list will facilitate the sharing and re-use of annotated corpora.

Ackowledgements: This work was party supported by the Quaero Program funded by Oseo French State agency for innovation and by the MIReS project funded by EU-FP7-ICT-2011.1.5-287711. Many thanks to Kevin B. Cohen for proof-reading.

5. REFERENCES

[1] F. Schiel, C. Draxler, *et al.*, "Production and validation of speech corpora," *Bavarian Archive for Speech Signals. Munchen: Bastard Verlag*, 2003.

[2] A.-J. Li and Z.-g. Yin, "Standardization of speech corpus," *Data Science Journal*, vol. 6, no. 0, pp. 806–812, 2007.

[3] M. Wynne, ed., *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 2005.

[4] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[5] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Klozali, D. Tidhar, and M. Sandler, "OMRAS2 metadata project 2009," in *Proc. of ISMIR (Late-Breaking News)*, (Kobe, Japan), 2009.

[6] C. Yeh, N. Bogaards, and A. Roebel, "Synthesized polyphonic music database with verifiable ground truth for multiple f0 estimation," in *Proc. of ISMIR*, pp. 393–398, 2007.

[7] T. Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music," in *Proc. of ICMC*, (Beijing, China), pp. 464–467, 1999.

[8] F. Opolko and J. Wapnick, "McGill university master samples CD-ROM for SampleCell VOLUME 1," 1991.

[9] G. Ballet, R. Borghesi, P. Hoffman, and F. Lévy, "Studio online 3.0: An internet "killer application" for remote access to ircam sounds and processing tools," in *Proc. of JIM*, (France), 1999.

[10] O. Gillet and G. Richard, "Enst-drums: an extensive audio-visual database for drum signals processing," in *Proc. of ISMIR*, pp. 156–159, 2006.

[11] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical, and jazz music databases," in *Proc. of ISMIR*, (Paris, France), pp. 287–288, 2002.

[13] M. Goto, "Aist annotation for the rwc music database," in *Proc. of ISMIR*, (Victoria, Canada), pp. 359–360, 2006.

[14] J. A. Burgoyne, J. Wild, and I. Fujinaga, "An expert ground-truth set for audio chord recognition and music analysis," in *Proc. of ISMIR*, (Miami, USA), 2011.

[15] F. Bimbot, E. Deruty, S. Gabriel, and E. Vincent, "Methodology and resources for the structural segmentation of music pieces into autonomous and comparable blocks," in *Proc. of ISMIR*, (Miami, USA), 2011.

[16] M. Ramona, S. Fenet, R. Blouet, H. Bredin, T. Fillon, and G. Peeters, "A public audio identification evaluation framework for broadcast monitoring," *Journal of Experimental and Theoretical Artificial Intelligence (Special Issue on Event Recognition)*, 2012.

[17] D. Moelants and M. F. McKinney, "Tempo perception and musical content: What makes a piece slow, fast, or temporally ambiguous?," in *International Conference on Music Perception and Cognition*, Evanston, IL, 2004.

[18] T. Bertin-Mahieux, D. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. of ISMIR*, (Miami, USA), 2011.

[19] M. McVicar and T. De Bie, "Enhancing chord recognition accuracy using web resources," in *Proc. of the 3rd international workshop on Machine learning and music*, pp. 41–44, ACM, 2010.

[20] M. Levy, "Improving perceptual tempo estimation with crowd-sourced annotations," in *Proc. of ISMIR*, (Miami, USA), 2011.

[21] Y. E. Kim, E. Schmidt, and L. Emelle, "Moodswings: A collaborative game for music mood label collection," in *Proc. of the International Symposium on Music Information Retrieval*, pp. 231–236, 2008.

[22] D. Turnbull, R. Liu, L. Barrington, and G. Lanckriet, "A game-based approach for collecting semantic annotations of music," in *Proc. of ISMIR*, (Vienna, Austria), 2007.

[23] E. L. M. Law, L. v. Ahn, R. Dannenberg, and M. Crawford, "TagATune: A game for music and sound annotation," in *Proc. of ISMIR*, (Vienna, Austria), 2007.

[24] M. I. Mandel and D. P. W. Ellis, "A web-based game for collecting music metadata," in *Journal of New Music Research*, vol. 37, pp. 151–165, Taylor & Francis, 2008.

[25] C. Harte, M. Sandler, S. Abdallah, and E. Gomez, "Symbolic representation of musical chords: A proposed syntax for text annotations," in *Proc. of ISMIR*, (London, UK), pp. 66–71, 2005.

[26] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and J. S. Downie, "Design and creation of a large-scale database of structural annotations," in *Proc. of ISMIR*, (Miami, USA), 2011.

[27] K. B. Cohen, L. Fox, P. V. Ogren, and L. Hunter, "Corpus design for biomedical natural language processing," in *Proc. of the ACL-ISMB workshop on linking biological literature, ontologies and databases: mining biological semantics*, pp. 38–45, 2005.

[28] G. Leech, *Developing Linguistic Corpora: a Guide to Good Practice*, ch. Adding Linguistic Annotation, pp. 17–29. Oxford: Oxbow Books, 2005.

[29] K. Fort and V. Claveau, "Annotating football matches: Influence of the source medium on manual annotation," in *Proc. of LREC*, (Istanbul, Turkey), May 2012.

[30] E. Peiszer, T. Lidy, and A. Rauber, "Automatic audio segmentation: Segment boundary and structure detection in popular music," in *Proc. of LSAS (Learning the Semantics of Audio Signals)*, (Paris, France), 2008.

[31] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[32] K. Fort, C. François, O. Galibert, and M. Ghribi, "Analyzing the impact of prevalence on the evaluation of a manual annotation campaign," in *Proc. of LREC*, (Istanbul, Turkey), May 2012.

[33] U. Gut and P. S. Bayerl, "Measuring the reliability of manual annotations of speech corpora," in *Proc. of Speech Prosody*, (Nara, Japan), pp. 565–568, 2004.

[34] H. Voormann and U. Gut, "Agile corpus creation," *Corpus Linguistics and Linguistic Theory*, vol. 4(2), pp. 235–251, 2008.

[35] D. Kaplan, R. Iida, and T. Tokunaga, "Annotation process management revisited," in *Proc. of LREC*, pp. 365 – 366, May 2010.

[36] MPEG-7, "Information technology - multimedia content description interface - part 5: Multimedia description scheme," 2002.

[14] It could not be included due to space constraints.