RECOMMENDATION ITU-R BS.1657

Procedure for the performance testing of automated audio identification systems

(Question ITU-R 8/6)

(2003)

The ITU Radiocommunication Assembly,

considering

a) that metadata will be accompanying most audio broadcast transmissions in the future;

b) that the automatic generation of metadata will be necessary to offer a complete cost-efficient service in future;

c) that automatic identification of audio items enables tracking of transmitted programme material;

d) that different schemes for extraction of audio metadata are developed today;

e) that ISO/IEC JTC 1/SC 29/WG 11 is currently finalizing schemes for the coding of metadata for multimedia data;

f) that no quality assessment procedures for audio metadata extraction schemes have been standardized until now,

recommends

1 that the procedure described in Annex 1 should be used to evaluate the performance of automated audio identification systems.

Annex 1

Procedure for the performance testing of automated audio identification systems

1 Introduction

In a time of ever-increasing databases filled with musical content, be it genuine audio material or associated metadata ("data about data"), the demand for tools to maintain this mass of data is also growing more urgent day by day. This desire is not only voiced by professionals, but also by the common Internet user and music-lover, who searches the web on numerous errands for her or his preferred musical style. In order to facilitate the retrieval of the desired material two different levels of abstraction are here discerned:

– The search for metadata that can more or less be extracted automatically from the audio content, such as instrumentation, melodic theme, rhythm, etc. An example application for this would be a query-by-humming system or the classification into genres, which is commonly used in recommendation engines.

– Automatic identification of titles, where only insufficient, unreliable or no metadata at all is available. An "essence" of the audio data is distilled and compared to a database of known material, thus creating a link to relevant metadata such as artist, song name, etc.

While the first mentioned class contributes mainly to the human interaction interface, the second topic finds its application also in the protection of rights by tracking radio programmes and Internet transactions. It is foremost in this latter context that algorithms fitting that profile are referred to as "fingerprinting" techniques.

2 Motivation

To meet the demands of the music business, the recognition rate of the applied fingerprinting technology must be high and withstand common alterations and modifications of the original audio content. For this purpose, the music business has acknowledged the need of quality assurance for audio identification systems by recently formulating a request for information on audio fingerprinting technologies.

The severity and urgency of this problem is also underlined by the fact that a number of different, often proprietary, solutions have appeared recently. All methods, however, face the same problems regarding their robustness to modification and deterioration of the original material. Although the original material may have changed by a number of processing steps or degradations, it nevertheless shall be recognized as the intellectual property of the artist and composer. This leads to the proposition that automated music identification should ideally be as precise and tolerant to signal modification as human perception and recognition. Beyond robustness to signal alterations, a good fingerprinting system should exhibit a small fingerprint size (considering that certain applications might require the storage of millions of fingerprints), fast fingerprint extraction and recognition and further desirable properties. It should be noted that robustness to signal alterations and compactness of fingerprint representation are two conflicting requirements which have to be reconciled by systems.

Consequently, in order to assess the quality of an automated audio identification system, a test environment has to be defined that covers different types of signal degradation in multiple degrees of severity and describes how to determine other essential system parameters. To allow the objective evaluation of identification systems, a unified test procedure is needed.

3 Quality parameters

For audio identification systems the following quality parameters have to be considered:

- Segment size of the audio material to be identified.
  - What portion of an item is necessary for the identification?
- Size of the fingerprint.
  - How many data (bytes) per item have to be stored in the database?
  - Is the size of the fingerprint constant or variable (with respect to the length of the item)?
- Size of the database.
  - How many items can be handled simultaneously by the system?

- Mode of identification.
  - Does the system allow identification of randomly chosen subsets of audio material (continous fingerprinting) or is identification tied to short fingerprinted segments? If the latter: What is the segment size?
- Identification speed.
  - How long does it take to identify an item?
  - How does this scale with the number of items in the database?
- Identification performance with original and altered material.
  - How much distortion can be introduced without significantly affecting the recognition rate?
  - How does this scale with the number of items in the database and the amount of distortion?
- Fingerprint generation speed.
  - How fast can a fingerprint be generated on a given platform?
  - How many resources are necessary to generate a fingerprint (e.g. central processing unit (CPU) speed, amount of random access memory (RAM), floatingpoint processing (FPU) unit necessary)?
- Training speed.
  - How long does it take to add items to the database? How does this scale with the number of items already in the database?

To assess these properties in a sensible fashion and thereby to show the suitability of a system for real-world application, a test environment must exhibit constant boundary conditions regarding the characteristics under test.

Relevant test conditions are the size and content of the reference database, size (referring to the playing duration) and number of the test items, exact modification rules for the test items, and computing platform, which includes specification of the CPU, memory, and operating system. A number of control titles should also be included with the set of test items that are not contained in the reference database in order to properly test rejection behaviour of the system under test.

4 Selection of test material and size of database

All different musical styles and genres should be present in the reference database with prevalence in numbers on the most heard genres. A database size of 10 000 to 100 000 pieces is suggested for a realistic evaluation.

Definition of terms:

- An item is called a duplicate item with respect to another audio item if it consists of the same recording as the original one with the exception that it might have a certain amount of zero valued leading or trailing samples added. This circumstance can sometimes be observed if the “same” song is located on different compilations or albums.
- A similar item represents a different (re)mix, cover version or (live) recording of another database item.

Requirements for the selection of test material:

– Special care should be taken to avoid duplicate items within the database.

– The database shall contain a certain amount of similar items (minimum 20 pairs). Example:

  – ten live recordings of one artist of the same song at different concerts;

  – ten original/remix pairs of one song of different artists;

  – ten original/cover version pairs of one song of different artists.

– The database shall be defined before the first experiment. It is not permitted to modify the database according to the test results.

5 Test method

As the speed of the calculation may depend on the amount of distortion in the test item, it is required to measure the speed of the extraction and search (classification) process separately for each experiment (1, 2, 3a) to 3i)).

5.1 Experiment 1

In a first test run, all titles from the reference database remain unmodified and have to be identified. The performance of the system under test should therefore be 100% for the correctly identified items.

The average fingerprint size is calculated based on the full set of reference items. This results in an average size per item or a size per length of an item depending on the type of fingerprint of the system under test. Data from systems which do not perform continous fingerprinting shall be considered separately from the data of continous fingerprinting systems.

5.2 Experiment 2

Thereafter excerpts of 1 000 items not contained in the reference database and thus unknown to the system with a length of 5 s and 30 s, respectively, shall be added to the test set. These 2 000 excerpts are presented to the system to acquire the rejection behaviour and to test for potential false positives. In this set of 2 000 items there should be at least ten items which are of the type "similar item" (to a corresponding item in the reference database).

5.3 Experiment 3

For testing the recognition robustness with modified musical pieces a set of 1 000 items is chosen from the reference set. The first test shall be conducted as described in 3a). Then all other tests (3b) to 3i)) are based on the excerpts created in 3a), that is, they represent a combination of the specific distortion with the "cropping" effect as described in 3a). The combination of all other distortions with cropping is reasonable to eliminate the unrealistic assumption of perfectly aligned fingerprints.

The following modification procedures are recommended to be used:

3a) Cropping/offset

Taking only small subsegments of the test item. The start sample of the excerpt shall be varied (randomly chosen but fixed for all test systems). The length of the excerpt should be 5, 10 and 20 s, respectively.

3b) Dynamic compression and expansion

Parameters shall be chosen according to customary settings used for broadcasting.

3c) Level adjustment

Scaling the input signal by a certain factor, e.g. –6 dB and 10 dB. Clipping shall be avoided.

3d) Equalization

Using octave band equalization with adjacent band attenuations set to –6 dB and +6 dB.

3e) Addition of noise

Addition of white or pink noise with an overall *S*/*N* of 10 and 20 dB, respectively.

3f) Sampling rate conversion and pitch shifting

Deviations of +5% and –5% in sampling rate shall be used.

3g) Audio coding and watermarking

The effects of audio coding shall be evaluated using an MPEG-1/2 Layer-3 encoded signal with the following bit-rate/channel combinations: 24 kbit/s (mono), 64 kbit/s (stereo), 96 kbit/s (stereo) and 128 kbit/s (stereo).

3h) Band limiting

The input signal shall be band limited to an upper frequency limit of 4 kHz.

3i) Acoustic transmission

The imperfections caused by acoustic playback under moderate acoustic conditions shall be tested: The signal is transmitted using a loudspeaker and recorded again using a microphone. The recommended distance between both is about 50 cm. It is not necessary to choose a high quality loudspeaker and/or microphone. The test should be done within a regular (not acoustically treated or isolated) room.

The parameters of the individual modification tests have been adjusted in a manner that the equivalent human listening perception would rate from “slight alteration” up to “strong alienation” of the original piece. For audio coding this would correspond to encoding in the MP3 format at 128 kbit/s (stereo) for slight alteration of the original material, and to 24 kbit/s (mono) for strong alienation. Encoding to 96 kbit/s (stereo) and 64 kbit/s (stereo) as intermediate steps is recommended, since these bit rates are most commonly used in Internet transactions. Therefore no more than five levels of degradation should be chosen[1].

---

[1] The inclusion of MPEG-1/2 Layer-2, MPEG-2/4 AAC, Dolby-E and others, which frequently are used in broadcast environment is regarded as not necessary because these schemes are usually not misused in a study environment as happens frequently with MPEG-1/2 Layer-3 (MP3).

6 Test platform

As a recommended computational platform devices and operating system should be utilized that comply with the state-of-the-art equipment available to the regular user. In 2002 an example of an adequate platform is a Pentium class machine running at 1 GHz with 512 MB of RAM using Windows 2000™ or Linux.

7 System parameter variation

During the different tests, fingerprinting systems which allow the achievement of varying degrees of robustness/fingerprinting compactness depending on their extraction parameter settings may be adapted in their setting to achieve optimum performance for each individual task/test. However, in this case each system/setting combination shall then be considered a separate system with a limited scope of application, its own fingerprint format and extraction process. This does not apply for systems in which a more compact/less robust fingerprint database can be derived from a less compact/more robust representation by means of a self-contained transcoding process, i.e. when only a single fingerprint extraction process from the reference audio material is sufficient to enable all functions shown in the tests.

8 Test report

Test reports should convey, as clearly as possible, the rationale for the study, the methods used and conclusions drawn. Sufficient detail should be presented so that a knowledgeable person could, in principle, replicate the study in order to check empirically on the outcome. An informed reader ought to be able to understand and develop a critique for the major details of the test, such as the underlying reasons for the study, the experimental design methods and execution, and the analyses and conclusions.

Special attention should be given to the following aspects:

– the specification and selection of the reference and test items;

– the selection of the similar items and the corresponding test results for these special items;

– the detailed description of the parameters of the different distortions;

– the detailed description of the set-up parameters used for the systems under test;

– the detailed basis of all the conclusions that are drawn.