RECOMMENDATION ITU-R BS.1693

Procedure for the performance test of automated query-by-humming systems

(Question ITU-R 8/6)

(2004)

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 query-by-humming systems enable a natural way to query audio databases;

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

e) that Recommendation ITU-R BS.1657 – Procedure for the performance testing of automated audio identification systems, describes a procedure for the performance test of automated audio identification systems;

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

g) that no quality assessment procedures for audio metadata extraction schemes regarding melody recognition have been standardized until now,

recommends

1 that the procedure described in Annex 1 should be used to evaluate the performance of automated query-by-humming systems.

Annex 1

Procedure for the performance test of automated query-by-humming 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 these masses of data is also growing more urgent day by day. This desire is not only expressed 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 can be discerned here:

– The search for high-level metadata as a human listener would describe contents, e.g. melody, rhythm, timbre, instrumentation or genre. An example application for this would be a query-by-humming system, which can be used in recommendation engines.

– Extracting mid-level metadata for automatic identification of certain interpretations of musical contents. Descriptions of the technical features of the audio data (spectral contents, etc.) are distilled and compared to a database of known material, thus creating a link to relevant metadata such as artist, song name, etc.

For an overview of current state-of-the-art query-by-humming systems please refer to document, ISMIR 2002 (3rd International Conference on Music Information Retrieval, IRCAM – Centre Pompidou Paris, France, October, 2002).

2 Motivation

To meet the demands of the music business, the recognition rate of the applied query-by-humming technology must be high and withstand common alterations to the stored representations in the song database.

This problem is tackled by a number of different, often proprietary, solutions that have arisen recently ([Clarisse *et al.*, 2002], [Ghias *et al.*, 1995], [Haus and Pollastri, 2001], [Heinz and Brückmann, 2003]). All methods, however, face the same problems regarding their robustness to modifications of the original material. This leads to the proposition that automated query-by-humming systems should ideally be as precise and tolerant to signal modifications as human perception and recognition. Therefore, a sophisticated query-by-humming system has to be robust against different distortions regarding signal quality and variations from ideal melody inputs. Also a reliable handling of large song databases consisting of several thousands of songs should be provided.

Consequently, in order to assess the quality of an automated query-by-humming system, a test environment has to be defined that covers different types of signal modifications and describes how to determine other essential system parameters. To allow the objective evaluation of query-by-humming systems, a unified test procedure is needed.

3 Quality parameters

For the evaluation of query-by-humming systems the following quality parameters have to be considered:

Required audio input:

– Is it necessary to sing a certain part of the song or is it possible to sing any part?

– What is the minimal length of the input to provide a reliable result?

Size of data representation:

– How many data (bytes) per song have to be stored in the music database?

Size of the music database:

– How many songs can be held in the music database?

Mode of identification:

– How does the kind of input, such as singing in mother language, humming or singing modes like "na na na", etc., any kind of instrument, influence the recognition rate and performance?

Melody recognition speed:

– How long does it take to identify a melody?

– How does this scale with the number of songs in the music database?

– How does this scale with the quality of the input data?

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 music database (see § 4);

– size of the query input (referring to the playing duration) and number of the test items (see § 4);

– exact modification rules for the test items (see § 5 and 6); and

– computing platform, which includes specification of the central processing unit (CPU), memory, and operating system (see § 7).

4 Selection of test material and size of the music database

A reference music sample database on which all systems perform their query should be defined. It should contain a mixture of different musical styles (pop songs from different countries, classic, ...) with worldwide prevalence in numbers on the most familiar songs. Special care should be taken to avoid duplicate items within the database (cover versions, etc.).

A music database size of 500-1 000 songs is suggested for a statistically reliable and relevant evaluation.

As the preparation of abstract high-quality representations of musical songs as they are needed for the database search is a troublesome and expensive procedure, the construction of the musical reference database is left to the participants. This will lead to an implicit quality criterion which will find its meaning in the obtained test results. All participating parties are free in the choice of the inner database format as this is dependent on the search algorithm.

A set of test items (query sample database) should be defined to meet the following requirements: in order to avoid any calibration regarding a special set of queries every participating party should contribute a totality of 200 query melodies. An adaptation of the parameters of the query-by-humming-systems to a subjectively provided query database can be ruled out by this demand. The test items should be of good audio quality including ideally no signal distortion. The inputs should contain a variety of different types, such as sung lyrics, hummed melodies (da, na, ta, la, ...) and instrumental inputs. These should be performed by a representative distribution of various singers and instrumentalists.

All test items have to be performed representations of melodies contained in the reference database. Testing a rejection behaviour is not suitable because of gliding degrees of similarities between melodies.

As the number of additionally tested query-by-humming-systems is growing in time there will result an increasing size of the query sample database. Therefore, a repeated testing of the systems will be necessary in order to compare the performances according to the statistically more and more expressive query database. An automated testing procedure is recommended.

5 Modifications

To be more realistic regarding real-world applications, high-quality test items (see § 4) should be modified using common acoustic pollution sources:

– audio compression (mp3, aac, ...);

– bandwidth limitation (telephony, ...);

– quantization (pulse code modification (PCM), A-law, ...);

– GSM distortion (fullrate, ...);

– background noise (audience, restaurant, music store, ...).

A list of exact rules is listed in § 6.

6 Test method

The main parameter for estimating the quality of involved systems shall be the percentage of correctly classified melodies. This can be divided into two categories:

– the searched item is ranked number one in the list of presented results;

– the searched item is among the ten melodies estimated most similar by the system.

These figures as well as the speed of the extraction and search (classification) process have to be measured separately for each experiment.

6.1 Experiment 1

In a first test run, all titles from the query sample database remain unmodified and have to be identified. So optimum conditions regarding audio quality are provided and the results should show a very high rate of correct identifications.

6.2 Experiment 2

For testing the robustness behaviour of the systems under consideration miscellaneous modifications are applied to the items of the query sample database. Modifications representing acoustic distortions occurring in every day life have to be chosen.

– GSM distortion:

  Test items have to be processed by three different speech coding techniques used for mobile telephony (GSM “fullrate”, “enhanced fullrate” and “halfrate”).

– Audio compression:

  The examples have to be compressed/decompressed using "MPEG-1/2 Layer-3" audio codecs applying coding rates of 64, 96 and 128 kbit/s. Recommended is the original Fraunhofer codec.

– Quantization:

  The query items have been subject to a non-linear A-law quantization (8 kHz, 8 bits).

– Bandwidth limitation:

  The input is limited by a bandpass according to conventional telephone quality, i.e. 300-3 400 Hz. The filter characteristics of the used bandpass should fulfil the requirement of a minimum descent of 12 dB/oct.

– Background noise:

  In order to have a quasi-standardized distortion database of real life speech signals and babble noise, the contents of the "ICRA Noise CD" [Dreschler *et al.*,] have to be utilized. Two kinds of different noise signals have to be composed with the original query data, i.e. using the unmodified noise tracks and attenuated versions (–6 dB), respectively:

  – 2 persons babble (normal effort, track 6);

  – 6 persons babble (raised effort, track 8).

7 Test platform

As a recommended computational platform devices and operating systems should be utilized that comply with the state-of-the-art equipment available to the regular user. In 2004 an example of an adequate and easily available platform is a Pentium 4/Athlon XP class machine running at 2.4 GHz with 512 Mbits of RAM using Windows XP™ or Linux.

8 Test report

Test reports should convey, as clearly as possible, the rationale for the study, the methods used and conclusions drawn. Sufficient details 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:

– a specification and selection of the music database and the audio sample database;

– a detailed description of the systems under test;

– a detailed description of all the conclusions that are drawn.

References

CLARISSE, L. P., MARTENS, J. P., LESAFFRE, M., DE BAETS, B., DE MEYER, H. and LEMAN, M. [October 2002] An Auditory Model Based Transcriber of Singing Sequences. ISMIR 2002, 3rd International Conference on Music Information Retrieval, IRCAM – Centre Pompidou Paris, France, p. 116-123.

GHIAS, A., LOGAN, J., CHAMERLIN, D. and SMITH, B. C. [1995] Query By Humming. Musical Information Retrieval in an Audio Database. Procs. ACM Multimedia, p. 231-236.

HAUS, G. and POLLASTRI, E. [2001] An Audio Front End for Query-by-Humming Systems. Procs. ISMIR 2001, p. 65-72.

HEINZ, Th. and BRÜCKMANN, A. [March 2003] Using a Physiological Ear Model for Automatic Melody Transcription and Sound Source Recognition. AES 114th Convention. Amsterdam, Netherlands.

DRESCHLER, W. A., VERSCHUURE, H., LUDVIGSEN, C. and WESTERMANN, S. ICRA Noises: Artificial noise signals with speech-like spectral and temporal properties for hearing aid assessment. *Audiology*, 40, p. 148-157.