Forming a Corpus of Voice Queries for Music Information Retrieval: A Pilot Study

David Bainbridge
Department of Computer Science
University of Waikato
Hamilton
New Zealand
davidb@cs.waikato.ac.nz

John R. McPherson
Department of Computer Science
University of Waikato
Hamilton
New Zealand
jrm21@cs.waikato.ac.nz

Sally Jo Cunningham
Department of Computer Science
University of Waikato
Hamilton
New Zealand
sallyjo@cs.waikato.ac.nz

ABSTRACT

The use of audio queries for searching multimedia content has increased rapidly with the rise of music information retrieval; there are now many Internet-accessible systems that take audio queries as input. However, testing the robustness of such a system can be problematic, as there is currently no standard test-bed of queries and music files available. A corpus of audio queries would aid researchers in the development of both audio signal processing techniques and audio query systems. Such a corpus would also be essential for making empirical comparisons between different systems and methods. We propose a pilot study that will field test a procedure for collecting audio queries. The lessons learned in the pilot study will guide us in refining the collection methodology, and we will make a final set of queries freely available to MIR researchers. The participants for this pilot study will be attendees of the ISMIR 2002 Conference.

1. BACKGROUND

At the 2001 International Symposium on Music Information Retrieval, the need for "standardised MIR test collections" [1] was discussed. To our knowledge, there are no sets of query recordings currently available to the community at large. We propose laying the groundwork for the creation of a corpus of audio queries that is designed with the needs of music information retrieval researchers and practitioners in mind. Because the intended use of the recordings will be formally declared to participants, there should not be privacy concerns associated with the release of the corpus.

2. INTRODUCTION

An increasing number of music information retrieval practitioners use human voice input in their research; applications include query-by-demonstration systems such as Themefinder [2], MusArt [3] and Meldex [4], and audio transcription services.

Various factors must be accounted for when processing raw audio; often there is a trade-off between the level of processing needed and the burden placed on the person performing a query. The *timbre*, such as whether the query has been sung, hummed, or whistled, is nearly always important. Different systems expect queries in a certain style, while individual users' preferences may affect the quality of the audio query.

Another major factor is note separation. Human speech and song is often *legato*, with each note trailing off and blending into the following note. It is easier to process audio that has a discernible silence between each note, and some systems require this of the singer (for example, [4]).

We recognize that there are many different types of retrieval tasks, each requiring a different corpus of test queries. In this pilot study we will examine the problem of developing test queries for known-item searches — that is, tasks in which the desired result is a piece of music that is known to be in the collection and which the user wishes to retrieve. Other music retrieval tasks — such as locating similar musical items, or automatic genre classification — will require different types of test queries, and may require a different process for developing a test corpus.

Given that target task, the corpus of test queries must be specific to a particular test-bed of music documents. We will design a set of queries for a test-bed of pop/rock recordings, where each query will be based around the retrieval of a single recording. Each query should be sufficiently difficult in that there must exist at least one other recording that is relatively 'similar' to the target song, where similarity may be measured in several dimensions, such as key, rhythm, melodic contour, or shared note sequences.

In text retrieval evaluations it is sufficient to have a single version of each test query, since the query itself remains the same no matter who types it in. The querier plays a more central role in MIR tasks, however, as each individual may be expected to vocalize the query differently. An MIR query corpus, then, should contain several instantiations of each query, as expressed by different individuals. We further believe that the corpus should also include several versions of a query by a single individual — for example, a hummed version, a sung version, and so forth. A corpus that contains queries in multiple formats would:

- allow implementors to test systems with the sub-set of queries that they have allowed for (such as "hummed queries"),
- allow comparison between systems that require different qualities in the audio input, and
- allow testing to see the effect in isolation that each of the above factors has on retrieval performance.

A secondary purpose of collecting a set of queries is to gain further insight into the variation that we can expect when different users attempt to pose a query to an MIR system. For example, an earlier, small-scale study of how people 'natively' prefer to generate queries indicates that sung queries will often drift in pitch unless the query is very brief, and identified a tendency for participants to add or drop consecutive syllables having the same pitch [5]. Substantive evidence of the types and degrees of query variation that an MIR system may be exposed to would be invaluable for researchers exploring techniques for improving search precision.

3. METHODOLOGY

We outline a methodology for our pilot study, which will solicit participants from the attendees of ISMIR 2002. Participation in the pilot study, and indeed for the ensuing development of query corpora, will be on a voluntary basis. No identifying data will be made available about the participants, and the query collection

process will comply with the guidelines of the University of Waikato Ethics Committee.

Given the commercially sensitive nature of music collections, copyright issues must be considered in the construction of a music query corpus. Article 10 of the Berne Convention 1971 allows for "fair practice" of copyrighted material, although each participating country defines the extent permissible in its jurisdiction. Many jurisdictions have exceptions in their copyright laws that allow the use of portions of a copyrighted work for various purposes, including scholarly work.

At this point, we are designing a *process* for collecting a corpus of queries, with the understanding that each different MIR test-bed will require its own sets of music documents, queries, and retrieval tasks. In the first instance, we will build a set of test queries for a large set of pop/rock recordings. This genre was selected for the pilot study for two reasons. Firstly, several collections of pop/rock recordings currently exist and are already seeing use in MIR experiments, and so a query corpus could be immediately useful to more than one research group. Secondly, the tunes that will form the basis of the test queries are more likely to be familiar to participants than other, more esoteric genres held in other available music document collections (for example, folk songs in a variety of languages), and so the participants should require relatively little coaching or assistance in forming their queries.

The corpus will be created in two parts. The aim of the first part is to focus on transcription issues and allow the separation of various factors, such as query method (whistle, hum) or style (legato or detached). Subjects will be given a short theme from a song (the query), in both audio and notated formats, and will be asked to vocalize the query in a number of prescribed methods — for example, whistled, hummed, or sung with ta syllables, or their natural preferred method — as well as a given articulation (that is, staccato or legato). Each subject will perform several queries, and order of query presentation will be randomised to minimise testing bias. Participants will be allowed to decline to perform queries for a particular song if they feel that they do not know the piece adequately, and another song will be offered.

The test corpus, then, will be based around a set of queries, where each query will be instantiated by several subjects and in more than one format per subject. Each instance in the corpus will consist of an indication of the query that it represents, an audio sample, metadata characterising the relevant attributes of the individual who has provided the sample, metadata describing the query type, and metadata for identifying the audio content.

The audio content will be of compact disc quality — that is, recorded at a sample rate of around 44 kilohertz. A symbolic representation (such as GUIDO format) of the audio query would provide an idealised transcription, allowing implementors to test the effects of the audio-to-symbol translation process separately, if required.

Metadata describing the relevant characteristics of subjects will be gathered through a short questionnaire to collect demographic information such as the subject's musical background, age bracket, gender, etc (for examples, see [6]). Query-specific metadata will include the voice query type (whether this query instantiation is whistled, hummed, sung, etc.), and whether the notes are performed legato or "detached".

Metadata for describing the audio content, such as song title and author, is needed for identifying matches with a query database. Other metadata — such as song genre, and whether or not the query is a main theme of the song — could also be useful, although these qualities are more subjective.

The second part of the corpus will involve subjects choosing their own query for well-known songs. Both an audio recording and a symbolic transcription will be made, so that transcription effects can be independently accounted for. The transcription will be made in conjunction with the experimenter, if assistance is required. The participants may perform the queries in their own choice of style.

We recognise that the queries gathered during this pilot study are unlikely to form a complete, unbiased corpus for our selected test-bed of music documents — for example, the range of musical ability shown by subjects may not adequately represent the queries generated by typical users of query-by-demonstration systems. This pilot study will allow the MIR community to explore the requirements for an effective audio query corpus and to examine the practical issues involved in creating such a corpus. The results of this study will be made available online at <`http://www.cs.waikato.ac.nz/music`>.

4. CONCLUSION

We have outlined a procedure for collecting a flexible corpus of music queries, where the procedure is designed specifically to meet some of the needs of audio query researchers. However, we anticipate that modifications and additions may be necessary in the future. The design discussed here is part of a learning process as our community discovers exactly what is required to make such a corpus successful, and what scope can be reasonably covered.

5. REFERENCES

[1] Matthew J. Dovey. Resolution on the need to create standardized mir test collections, tasks, and metrics. ISMIR 2001. `music-ir.org/mirbib2/resolution`, October 2001.

[2] Andreas Kornstädt. Themefinder: A web-based melodic search tool. *Computing in Musicology*, 11:231–236, 1998.

[3] William P. Brimingham and Roger B. Dannenberg et al. Musart: Music Retrieval via Aural Queries. In *Proceedings of the Second Annual International Symposium on Music Information Retrieval*, Bloomington, Indiana, USA, October 2001.

[4] Rodger J. McNab, Lloyd A. Smith, David Bainbridge, and Ian H. Witten. The New Zealand Digital Library MELody inDEX. *D-Lib Magazine*, May 1997.

[5] Rodger J. McNab, Lloyd A. Smith, Ian H. Witten, Claire L. Henderson, and Sally Jo Cunningham. Towards the digital music library: tune retrieval from acoustic input. In *Proceedings of Digital Libraries '96*, pages 11–18, Bethesda (MD, USA), March 1996. ACM.

[6] J. Stephen Downie. Creating the ideal full-text music database: User assessment survey. Technical report, Graduate School of Library and Information Science, University of Western Ontario, London, Ontario, Canada, 1993.