Towards large databases for Music Information Retrieval systems development and evaluation

Gaël RICHARD
Ecole Nationale Supérieure des Télécommunications (ENST)
Signal and Image processing department (TSI)
46, rue Barrault
75013 Paris, France
+33 1 30 70 84 27
Gael.Richard@enst.fr

ABSTRACT

In the context of MIR/MDL evaluation, a key component for evaluation would be the availability to the research community of a large corpus of test data consisting of both audio and structured music data. This paper proposes a possible path towards this goal by following the basic principles of the *SpeechDat* projects. *SpeechDat* refers to successive EC supported projects of large scale multilingual data collection for developing and testing several classes of speech recognition algorithms. Even if the domain of speech recognition differs from the domain of Music Information Retrieval, it is suggested that some of the *SpeechDat* experience could be applied to the collection of large audio/music database that would be suitable for MIR/MDL development and evaluation.

1. INTRODUCTION

The need for MIR/MDL evaluation testbeds is now clearly recognized in the Music Information Retrieval (MIR) and Music Digital Library (MDL) communities. Recent initiative towards the development of MIR/MDL evaluation frameworks are summarized in [1]. These efforts lead to a collection of white papers (see [2]) expressing various ideas and possible paths towards the ultimate goal: the formation of meaningful and comprehensive MIR/MDL evaluation through the identification and/or creation of standardized test collections, retrieval tasks and performance metrics.

As summarized by the ISMIR 2001 resolution on MIR Evaluation [3], a key component for evaluation would be the availability to the research community (with international clearance of relevant copyright) of a large corpus of test data consisting of both audio and structured music data. Indeed, if private or specific data may be used to compare systems and to indicate performances progress of a given system (see [4] for the minimal standard of MIR evaluation), it is clear that reproducible and large scale evaluation can only be tackled on large common databases using common evaluation protocols and metrics.

TREC type evaluation clearly represents an exciting and very valuable direction that would lead to meaningful MIR system evaluation on given tasks ([5]). The objective of this paper is to present a concurrent (but not at all exclusive) path to TREC type evaluation. This other direction would aim at building a database on common specifications and by sharing the cost and effort of the data collection. The basic principle for such an initiative could follow the approach adopted in *SpeechDat* Projects which have lead to large scale multilingual databases for speech recognition.

The paper is organized as follows: the next section is dedicated to the presentation of *Speechdat* projects. Section 3 will then propose some directions to apply the principles of *SpeechDat* projects to audio data collection and then a brief conclusion will be given in section 4.

2. SPEECHDAT PROJECTS OVERVIEW

The goal of the *Speechdat* projects is to build large scale and multilingual corpora for developing and testing several classes of speech recognition algorithms. Such speech recognizers include isolated word systems, word spotting systems and vocabulary independent systems which use either whole word or subword model approaches. More precisely, *SpeechDat* refers to successive projects supported by the European Commission of data collection for speech recognition. To name a few, *SpeechDat(M)* included 1000 speakers over the telephone in 8 languages, *SpeeachDat(II)*, included 28 different databases (with up to 5000 speakers for fixed network and up to 1000 speakers for mobile networks), and *SpeechDat-Car* included 300 speakers recorded in cars in various driving conditions for 10 languages. More information may be found on the *SpeechDat* Web site about past and on-going projects [6] or in the following papers ([7],[8],[9],[10]).

The basic principle behind these projects is to build a consortium of partners who will share the effort to build the targeted database. For example, each partner records a database in its own language and will then exchange it to all other databases recorded in other languages by the other consortium members. As a consequence, for the price of one database, each partner has at the end of the project as many databases as the number of partners.

To ensure consistency and homogeneity across the individual databases, a *Speechdat* project is structured around 5 main phases:

1. A specification phase where all partners define the basic characteristics of the database (i.e. contents, recording equipments, structure and format of the databases, compression format and type of annotation desired, rules for validation…).
2. A recording phase where each partner records his own database.
3. An annotation phase where each partner annotates his database according to the general specifications.
4. A validation phase where each database is checked by an independent center against the specifications.
5. Database exchange phase: after validation each partner exchanges his database with other partners databases.

2.1 Database specifications

It is clearly one of the most difficult phase to complete since the specifications should result from a consensus of a sometimes large consortium of partners. The specifications, also called "*database design*") include different aspects that are briefly described below.

2.1.1 Database content

This is in fact a fundamental feature of a database. In the context of Automatic Speech Recognition (ASR), it is important to recall that most modern speech recognizers are based on statistical pattern recognition approaches amongst which Hidden Markov Models (HMM) are the most popular. The first phase of a speech recognizer, called training, builds internal models that will be used to recognize words or subdivisions of words (syllables or phonemes for example). To be efficient, speech recognizers need to be trained on a large number of speech utterances pronounced by a sufficiently high number of speakers (to obtain "speaker independent recognizers") with a good coverage of the acoustic environments of the targeted application. In addition, a sufficiently high number of repetitions per vocabulary words (or of each subdivisions of words such as phonemes) and per speaker is necessary. Therefore, the database content phase will aim at defining the number of speakers and a complete list of words, sentences, number, dates etc… with the corresponding number of repetition per speaker that is desired. In the context of multilingual speech recognition, it is also necessary to record the database in all targeted languages.

2.1.2 Acoustic environments

In speech recognition, it is acknowledged that best performances are obtained when the data are recorded in the context of the future applications (i.e. recordings in office for dictation products, recordings in car for automotive applications etc.)[1]. In addition, for training, it is sometimes desired to have high quality signals that could be acquired in an anechoic chamber. Nevertheless, it is clear that an infinite number of situations may occur in real life, even in well defined application scenarios. It is therefore necessary to first reduce the number of possible environments to typical situations and acoustic conditions to acquire the data (For example, in *SpeechDat-Car*, 7 different conditions were defined ranging from "car stopped engine running" to "high speed with windows open").

2.1.3 Recording platform

The choice of the recording platform is also important and includes the choice and number of microphones used, the distance of the microphones to the speaker, the type of AD converter, the sampling rate of the audio signals, etc. The idea is to define a common platform to all partners to first ensure comparable quality but also to optimize recording problem resolution.

2.1.4 Database format and structure

In *SpeechDat* projects, the directory structure is independent of the content of the speech files. Since it does not contain any semantics regarding for example the speaker or the recording environment characteristics it allows a fully automatic creation of a file system during recordings. Each speech file is accompanied by an annotation file that is stored in the same directory. The annotation file describes the content of the corresponding speech signal and gives therefore information about the speakers (age, sex, region in the country), the acoustic environment of the recordings, the text pronounced etc…

These annotation files (or *label files*) adhere to a modified SAM format ([11]):

ABC: item1,item2,item3,

where ABC is a three letter mnemonic followed by a column and itemY are information corresponding to the label. (For example, a file recorded with a single microphone XXX will have a line in the corresponding annotation file

MIT: XXX

where MIT stands for MIcrophone Type).

Each *SpeechDat* database includes in addition of the speech files, four types of accompanying files that are built from the label files to permit rapid search of specific information without the need for a full browsing of the label files.

2.2 Annotation

The annotation phase follows the recording phase. It consists in producing the annotation file (in ASCII format) for each of the speech file previously recorded. For *SpeechDat* projects, most of the labels are automatically generated. This is the case of the labels that are the same for the whole database (sampling frequency, number of bytes per sample,…) and of the labels that can be generated from information entered at the beginning of a recording session (speaker sex and age, recording environment etc.).

However, some of the most important labels have to be filled manually after having listened to the corresponding speech file. This is the case for the *"pronounced text"* label, or all noise and mispronunciation markers that are inserted in the *"text pronounced"* field. This phase is of primary importance since it gives all the value to the database. It is clear, though, that it is one of the most tedious part of the database collection. Extensive information about annotation and labels format may be found in [7].

2.3 Database validation

In the context of *SpeechDat* projects, validation refers to the process in which a database is checked against the specifications and corresponding tolerance intervals (also called "*validation criteria"*) defined in the specification phase. The validation is performed by an independent center[2]. A large part of the validation is done automatically. For example, file format, missing files, transcription symbols used, speaker and environment balances etc.. are checked automatically on the label files. For all files, information such as clicking rate or average SNR are also automatically extracted. Then, in a second step, a number of files of the database are randomly selected and manually checked. All errors of annotation are then counted and the database is declared valid if the error rate is below a pre-defined threshold. In practice, it is often necessary to define threshold for several variables that characterize the quality of a database. If a database does not satisfy the validation criteria, the corresponding partner needs to rework the annotation, or to re-record part of his database before it could be revalidated. In the *SpeechDat(II)* project, it is worth to note that about one out of six databases needed a revalidation.

[1] *This also explains why several "SpeechDat" projects were built where each project focuses on more specific acoustic environments.*

[2] *For the SpeechDat projects, the independent center is the Speech Processing Experise center (SPEX),* http://lands.let.kun.nl/home.en.html

2.4 Database exchange

The validation phase is essential for the quality control of the database. However, an another motivation for this quality check is the free exchange policy of databases within the consortium. The principle is simple. Each partner builds his database on a given language and then exchanges his database with all other partners. As a result, for the price of building a database in one language, each partner obtains, at the end of the project, the corresponding databases in all other languages included in the project. Obviously, this is tractable if all single databases have similar value and quality. Also to further justify the European funds for this kind of projects, the databases become, after a pre-defined lead time, publicly available to all with favorable financial conditions for research. The distribution of the SpeechDat is done by ELRA [12].

3. TOWARDS AUDIO DATABASES FOR MIR/MDL EVALUATION

SpeechDat databases described above are designed for speech recognition evaluation and/or development. Obviously, the domain of speech recognition is quite different from the domain of Music Information Retrieval. However, similar approaches than those used for setting up *SpeechDat* databases could be applied to build large audio/music database suitable for MIR/MDL development and evaluation. For example, the free exchange policy of databases built on common specification should lead to the availability of large databases at reduced production cost.

Such an initiative would therefore contain similar phases as those of *SpeechDat* projects (i.e. specification, recordings/acquisition, annotation, validation and database exchange phases). In addition, one may think of a test protocol definition phase where specific test protocols would be defined on these databases. Some direction for these phases are outlined in the following paragraphs.

3.1 Specification

As in *SpeechDat* projects, the first important phase would be to define the specifications of the future large audio database. This phase is always tricky due to the diversity of possible applications and therefore of potential needs of researchers or developers. This diversity is in fact clearly present in audio where there is a large variety of potential users (as pointed for example by [13]) and of problems in real word MIR applications (see [14]).

The goal of the specification would then be to define the content of the database in terms of audio signals, annotation content and targeted usage.

3.1.1 Audio signals and targeted usage

There is already a huge amount of available audio signals that could be included in a large scale audio database. These signals can be natural (car noise, door slam, music, speech, etc..) or synthetic. Even in one selected domain such as music signals, an infinite variety of signals can be found (classical music, jazz, rock, funk, etc…) . It is clear that the selection of the desired classes of signals can only be made once the targeted usage of the database has been defined (or in other words that the targeted application is well defined). For example in a "query by humming" application, two sub-databases should co-exist:

- A *query database* that would contain recorded signals of the query. It is clear in this case that we are getting quite close to "speech recognition application scenarios" where the type of signal may vary significantly depending on the surrounding acoustic environment (query using a mobile phone, at home, in a car, etc..). Also, even if it should not be to the same extent as in speech, it is probable that some variation in the way the song is hummed should depend on the native language of the user (humming using the "words" /mmm/, /la la la/ or /da da da/…" etc…). Finally, to obtain a meaningful query database, it is clear that a sufficiently high number of similar query are needed.
- A *real audio signals database*, on which the queries will be tested. This audio database would need to have a significant overlap with the query database (i.e. the query should have, in most cases, the targeted music signal in the real audio signals database).

The specification would also need to specify other signal characteristics such as compression rate, sampling rate, number of quantization bits etc….. Although, it may seem natural to opt for the highest possible sampling frequency and the best possible quality (no compression), a database with signals that would not be of top quality may prevent illegal use of these databases and may facilitate negotiation with rights owner to build such databases.

3.1.2 Annotation files

Another difficult choice concerns the annotation format. In speech databases, the problem of granularity of the annotation is crucial (should we annotate everything that is present in the speech files down to the phonemic representation with a description of all other acoustic events ? or should we only give the most important information, i.e. the text effectively pronounced and some markers for extraordinary events ?). Obviously, the targeted application should drive this choice but it is always important to keep in mind that a rudimentary annotation can always be completed afterwards when needed where the reverse could result in dramatic waste of energy to obtain the full, very detailed information.

In music, of course, the granularity of possible annotation is even wider due to the diversity of music signals (Should we annotate each note that was actually played with all nuances and subtle expression ? or merely attach a standard music score with an overall tempo indication ?). For music, an appropriate notation could be obtained with one of the already existing formats for encoding scores or musical performances such as MIDI, CMN or MPEG4-SA. It seems to me that an appropriate choice would be a format that does not go much beyond a slight adaptation of the basic score (for example updating only overall tempo or transposition of the performed piece compared to the original score). It is worth to mention that this phase is particularly important since the value of the database for researchers and developers heavily depends on the annotation information.

3.2 Database validation

The validation process as set up in *SpeechDat* projects is central to the free database exchange policy but is also important for controlling the quality and homogeneity of the data. Quality control of audio databases is also essential to obtain useful data (as noted in [14]). Though, the validation criteria can only be defined once the content specification and targeted applications or usage have been identified. For audio signals, such criteria could be very similar to those used in *SpeechDat* projects (this would be especially relevant for "query by humming" databases) but could also be more specific to music signals (one such criteria could be: the tempo indicated should not differ from the true tempo of more than 5%).

3.3 Database exchange and distribution

3.3.1 Free exchange policy

Once common specifications have been defined and accepted by the consortium members and that a comprehensive validation procedure is set up, it is feasible to agree on a free database exchange policy. Similarly, to *SpeechDat* project, each partner will be responsible for the recordings/annotation of part of the final complete database. In a sense, each partner will own his sub-database and will therefore be able to exchange it with all other partners sub-databases as soon as similar value and quality can be guaranteed (which is in fact the role of the common specifications and validation phases). Such sub-databases could be French jazz songs, folk Romanian music of the nineties, etc….

3.3.2 Rights

Having international clearance of all rights on the audio signals of these database for research purposes is a necessity. There is probably no straightforward solution to this problem. To split the complete database into sub-databases may simplify the negotiation of such rights for research purposes. Also, since the main threat for rights owner is to loose the control of the diffusion of the music pieces under protection, a solution may be obtained if the database does not contain top quality signals (this may be acceptable since most MIR application may not require full bandwidth signals). It should also be possible to push forward the fact that MIR research/development may clearly benefit rights owners protection in developing efficient and automatic music pieces identifier.

3.3.3 Distribution and Support

Such databases could also become publicly available after an eventual lead time as in *SpeechDat* projects. This kind of project due to their extreme collaborative nature, and due to the considerable progress it could bring to the field should definitely be eligible for international funding (funding from European Commission for European partners, funding from American agencies for American partners etc….). This would result in the acquisition of a large database at reasonable cost for each partner.

3.4 Development of test protocols

Having a large scale and annotated audio database would represent a clear step forward towards meaningful evaluation of MIR systems. It is clear also that in following sensitive criteria, any researcher can give solid indication on the relevance of a given MIR approach ([4]). However, the ultimate solution that would allow a meaningful and thorough comparison of different approaches for a given task would follow a predefined common test protocol. For example in a problem of "music style classification", a test protocol would specify which part of the database is used for training, and which part of the database is used for testing. Furthermore, it should specify the length of the test segments and how they are selected in the music piece (for example 10 s from the middle of the file, etc…). But even in this case, the honesty of the researcher as pointed by [4] will be needed.

4. CONCLUSION

This paper proposed an alternative (but not at all exclusive) path towards the availability of a large scale audio database suitable for MIR/MDL system evaluation. Some directions on how the basic principle of *SpeechDat* projects (Collaborative large scale data collection for testing and developing multilingual speech recognition algorithms) could be applied to obtain such large scale audio databases.

5. REFERENCES

[1] S. Downie, *"Who, What, When, Where and Why: Introduction and Acknowledgments (1st edition)"*, The MIR/MDL Evaluation Project white paper collection #1, http://music-ir.org/evaluation/wp1/wp1_downie_intro.pdf

[2] The MIR/MDL Evaluation Project White Paper Collection, Edition 1, http://music-ir.org/evaluation/wp1/

[3] ISMIR 2001 Resolution on MIR Evaluation, http://music-ir.org/mirbib2/resolution

[4] J. Futrelle, *"Three Criteria for the Evaluation of Music Information Retrieval Techniques Against Collections of Musical Material"*, The MIR/MDL Evaluation Project white paper collection #1, http://music-ir.org/evaluation/wp1/wp1_futrelle.pdf

[5] E. M. Voorhees, *"Whither Music IR Evaluation Infrastructure: Lessons to be Learned from TREC [Keynote Address]"*, The MIR/MDL Evaluation Project white paper collection #1, http://music-ir.org/evaluation/wp1/wp1_voorhees.pdf

[6] *The SpeechDat projects*, http://www.speechdat.org/

[7] Van den Heuvel H., Boves L., Moreno A., Omologo M., Richard G., Sanders E., *"Annotation in the speechdat projects'*, accepté pour publication dans l'International Journal of Speech Technology (IJST). (2001).

[8] H.Höge, "Speech database technology for commercially used recognizers : status and future issues", Very Large Telephone Speech databases Workshop, LREC2000, Greece, May 2000.

[9] C. Draxler, H. Van den Heuvel, H. Tropf, "SpeechDat experiences in creating large multilingual speechdatabases for teleservices", in Proc. of LREC98.

[10] Richard G., "The SpeechDat-Car Project: Overview of a very large multilingual speech database recorded in cars", XLDB 2000 (satellite workshop to LREC2000), Athens, May 2000.

[11] SAM. *"User guide to ETR tools. SAM: Multi-lingual speech Input/output Assessment, Methodology and Standardisation*" Ref: SAM-UCL-G007. (1992).

[12] ELRA/ELDA: http://www.elda.fr/

[13] S. Cunningham, *"User Studies: A First Step in Designing an MIR Testbed"*, The MIR/MDL Evaluation Project white paper collection #1, http://music-ir.org/evaluation/wp1/wp1_cunningham.pdf

[14] D. Byrd and T. Crawford, *"Problems of music information retrieval in the real world"*, Information Processing and Management, Elsevier, **38**, (2002), pp 249-272.