Benchmarking Music Information Retrieval Systems

Josh Reiss
Department of Electronic Engineering
Queen Mary, University of London
Mile End Road,
London E1 4NS UK
+44-207-882-5528
josh.reiss@elec.qmul.ac.uk

Mark Sandler
Department of Electronic Engineering Queen Mary,
University of London
Mile End Road,
London E1 4NS UK
+44-207-882-7680
mark.sandler@elec.qmul.ac.uk

ABSTRACT

Our goal is to create an accurate and effective benchmarking system for music information retrieval (MIR) systems. This will serve the multiple purposes of inspiring the MIR community to add additional features and increased speed into existing projects, and to measure the performance of their work and incorporate the ideas of other works. To date, there has been no systematic rigorous review of the field, and thus there is little knowledge of when an MIR implementation might fail in a real world setting. Benchmarking MIR systems is currently hindered by the diversity of the systems, by their relatively new and unrefined nature, and by the limited number of accessible systems. Thus most of what will be described here will be introductory and will lay down the framework for future benchmarking and analysis. Particular attention will be paid to the evaluation issues surrounding retrieval of audio in test collections.

1. INTRODUCTION

The Music Information Retrieval (MIR) field is primarily concerned with efficient content-based searching and retrieval of musical information from online databases. The musical data may be stored in a variety of formats ranging from encoded scores to digital audio. MIR systems should be easily operated by users with a wide range of musical ability and understanding and should be controlled by a simple-to-use graphical 'musical' interface, both for search queries and for the presentation of results. The musical databases might range from a modest collection of score-files stored on a single hard disk for an individual research project, to the collection of countless MIDI- or audio-files accessible via the Internet.

There are a lot of MIR systems in various stages of development. These systems all have the same task- to enable users to search for music in a database. But there are very few systems that are actually publicly accessible and comparable, To date, there has been no formal analysis or quantitative comparison methodology (benchmark) of the available preliminary MIR systems. Some systems work only with MIDI representations, some with monophonic transcriptions, and some with scores. In addition, each system has a different set of files available in its database. To date, there is no online, publicly available system, that attempts to search for music based on polyphonic transcriptions. Thus, one goal of this work is to find ways by which these different systems can be compared. A benchmarking of MIR search engines will also provide an effective measure of the progress in the field.

The work presented here addresses the methodology and results of a benchmarking analysis of music information retrieval services. Benchmarking MIR search engines techniques based on a portfolio approach are discussed and presented showing actual retrieval results. Differences in search engines index constructions methods are also reviewed. An additional motivation is to inspire researchers to improve their work and incorporate additional tools.

2. MUSIC INFORMATION RETRIEVAL METHODS

Proposed Music Information Retrieval Systems may be divided into two categories; those that search symbolic representations of music, and those that search raw audio files. The symbolic representations typically consist of MIDI files or Common Music Notation (CMN). The CMN files may be actual digital reproductions (scanned images) of score sheets, or as file formats that allow for an appropriate electronic representation of CMN[1],[2]. The symbolic representations usually consist of a list of instructions as to how the piece should be played. These include the notes, when and for how long each is played, the dynamics and the instruments that should be used. Other symbolic representations that may be searched include piano rolls and Parsons notation[3]. A typical query may involve a search for files with a given sequence of notes, and might produce a list of MIDI files from a database. Such queries are pertinent to musicians and musicologists who have a knowledge of musical representations.

The raw audio files are typically WAV or mp3 file format. Essentially, they are digital representations of an actual recording. Thus they contain a level of complexity that is not found in the symbolic representations. The composition is contaminated by noise and incorporates slight variations in the timing and dynamics of the notes. By comparison, symbolic representations are ambiguous, since they often leave certain characteristics of the piece unspecified. Thus, two performances may have the same MIDI or CMN representations, but differ notably in their audio files.

A raw audio based MIR system would offer many possible options as queries, depending on the complexity of the system. Songs may be hummed, sung or whistled into a microphone[4], retrieved from an internet radio station (www.clango.com), played from the computer's CD drive, or selected from audio files on the hard drive. The retrieved documents may include files of various formats, possibly from a large remote digital library, from a small media file library on the user's hard drive (www.soundfisher.com), or across the World Wide Web.

MIR systems that operate on audio files have followed two approaches, feature extraction[5] and transcription[6]. Feature extraction involves finding certain features, such as the mean and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

variance that typifies the audio or a portion thereof. The query and all files in the database are classified in terms of these parameters. Retrieval systems then operate as multidimensional searches on these parameters. Fast search methods have been described for such systems.[7] No attempt is made to relate these features to the musical qualities they might represent, e.g., energy to loudness, frequency to pitch.

Transcription based raw audio MIR systems convert the query into a symbolic representation, and seek to match it against symbolic representations of the audio files in the database. Thus such a technique typically uses feature extraction as well, but then has an intermediate step attempting to relate these features to a description of the notes and instruments. This is an exceedingly difficult to task, and to date, no system achieves this effectively and accurately over a wide range of music.

2.1 The case for transcription

Transcription is greatly simplified when the music is monophonic as opposed to polyphonic. That is, at most one note is played at any given time. For polyphonic music, any note may begin before the previous note, or set of notes has finished. Monophonic music may be transcribed accurately and quickly. Thus, monophonic queries, such as humming, whistling and (in most cases) singing, could be easily used to retrieve accurate symbolic representations. Similarly, a large database of monophonic music could be quickly catalogued by their monophonic transcriptions. Furthermore, effective and efficient search methods, such as string matching have been implemented on monophonic transcriptions. [8]

Although polyphonic transcription may be an unsolvable problem (imagine trying to transcribe a performance of an unknown piece by an entire symphony orchestra note for note), it may still achieve partial success in limited situations. Monophonic transcription methods have been used on polyphonic music[9], and polyphonic transcription has been achieved with partial success on a single instrument (piano)[6].

For the purposes of information retrieval, it is not necessary that the transcription be exact or even produce an audibly close match to the original file. All that matters is that the transcription results in close matches in the retrieved documents. Thus, transcriptions that are prone to certain errors may still be very effective in MIR. Small databases containing quite different pieces of music would have transcriptions so different from each other that even a poor transcription method might retrieve the closest matched file.

Transcription based approaches have the advantage that they can be used in hybrid symbolic/raw audio MIR systems. For instance, both queries and retrieved documents may include MIDI files. One could envisage a situation where the query is a musical score, and the retrieved documents are all live performances of that score in the database. Since, transcription based methods incorporate symbolic representation MIR methods, they may be easily integrated into existing systems that search on CMN and MIDI files. This has the advantage that ongoing work in symbolic representation will improve the effectiveness of transcription retrieval techniques.

2.2 The case for feature extraction

Feature extraction and classification techniques have the simple advantage that they succeed where transcription fails. There are no preconditions regarding the complexity of the music, or that it fit into a notational format. Thus, the music may incorporate speech, noise and atonalities. Indeed, it need not even be music. Speech files may be incorporated into the database, and a classification technique should have few problems distinguishing them. Audio with similar noise levels

Feature extraction techniques are considerably faster than transcription based techniques. The added complexity of converting the features into a sequence of notes is where most of the computational time lies. In the case of polyphonic transcription the music is segmented into frames. Even once a note has been identified to exist in a frame, the preceding and following frames must also be analysed to identify that note's duration and to distinguish the existence of other notes. Such analysis must be rigorous and the total number of frames to analyse is not known.[10] No such analysis is necessary when the features are not used for transcription. Indeed, features may be chosen for their usefulness in distinguishing musical pieces and for their lack of computational complexity.

Feature extraction has uses related to MIR that extend beyond retrieval of documents from a database. It may be used to classify music by genre[11], to identify a song heard on the radio, to identify copyright infringement on shared music file systems and to generate playlists from CDs or stored files. Indeed, most of the commercial content-based identification software available from Audible Magic (www.audiblemagic.com) is based on feature extraction techniques.[12]

The answer, in an ideal situation, is to use a hybrid of transcription and nonmusical feature extraction. Thus, searches can be made on testbeds that include speech, noise, atonal and complex music. Both queries and retrieved files may include symbolic representations. Transcriptions may be generated where possible, and presented with the retrieved audio. It is expected that relevance includes both music with similar transcriptions and with similar production and other factors that may only be identified through feature extraction. Thus weighting of retrieved files should combine, where possible, transcription similarity and feature similarity.

3. ONLINE MIR SYSTEMS

For the purposes of this work, we considered five online MIR systems. The systems considered all have certain properties in common. They may all be used online via the World Wide Web. They all are used by entering a query concerning a piece of music, and all may return information about music that matches that query. However, these systems differ greatly in their features, goals and implementation. These differences are discussed in detail below.

3.1 CatFind

CatFind[13] allows one to search MIDI files using either a musical transcription or a melodic profile based on the Parson's Code. It has minimal features, and was intended primarily for demonstration. Although it seems unlikely that this system will be extended, it is still useful here as a system for comparison.

Table 2. A comparison of features between the online MIR Systems.

	MelDex	ThemeFinder	CatFind	MelodyHound	Music Retrieval Demo
Query By Humming	Y	N	N	Y	N
Query By Parsons Code	N	Y	Y	Y	N
Database Size	11,169	~21,500	270[1]	~10,210	>250
Extendable	N	Y[2]	N	Y	N
Display Score	Y	Y	N	Y[3]	N
Distance Information	Y	N	N	N	Y
Ranking	Y	N	N	Y[4]	Y
Partial Matching	Y	N	N	Y	Y
Multiple File Formats	Y[5]	Y	N	N	N
Multiple Search Options	Y	Y	Y	Y	N
Polyphonic	N	N	N	N	Y

[1] An alternative version of Catfind uses a database of 20,000 files but is currently not working.

[2] The user can currently provide additional feedback and relevant hypertext links, but not extend the number of files in the database.

[3] MelodyHound offers links whereby the score and/or the music can be purchased online.

[4] MelodyHound ranks the results but provides no confidence measure.

[5] MelDex allows the user's query and its transcription to be given in various formats, but the returned results are only given as MIDI files.

3.2 MelDex

This allows searching of the New Zealand Digital Library. The MELody inDEX system[14, 15] is designed to retrieve melodies from a database on the basis of a few notes sung into a microphone. It accepts acoustic input from the user, transcribes it into common music notation, then searches a database for tunes that contain the sung pattern, or patterns similar to it. Thus the query is audio although the retrieved files are in symbolic representation. Retrieval is ranked according to the closeness of the match. A variety of different mechanisms are provided to control the search, depending on the precision of the input.

3.3 MelodyHound

This melody recognition system[16] was developed by Rainer Typke in 1997. It was originally known as "Tuneserver" and hosted by the university of Karlsruhe. It searches directly on the Parsons Code and was designed initially for Query By Whistling. That is, it will return the song in the database that most closely matches a whistled query.

3.4 ThemeFinder

Themefinder[17], created by David Huron, et. al.,[18] allows one to identify common themes in Western classical music, Folksongs, and latin Motets of the sixteenth century. Themefinder provides a web-based interface to the Humdrum thema command[19], which in turn allows searching of databases containing musical themes or incipits (opening note sequences).

Themes and incipits available through Themefinder are first encoded in the kern music data format. Groups of incipits are assembled into databases. Currently there are three databases: Classical Instrumental Music, European Folksongs, and Latin Motets from the sixteenth century. Matched themes are displayed on-screen in graphical notation.

3.5 Music Retrieval Demo

The Music Retrieval Demo[20] is notably different from the other MIR systems considered herein. The Music Retrieval Demo performs similarity searches on raw audio data (WAV files). No transcription of any kind is applied. It works by calculating the distance between the selected file and all other files in the database. The other files can then be displayed in a list ranked by their similarity, such that the more similar files are nearer the top.

Distances are computed between *templates*, which are representations of the audio files, not the audio itself. The waveform is Hamming-windowed into overlapping segments; each segment is processed into a spectral representation of *Mel-frequency cepstral coefficients*. This is a data-reducing transformation that replaces each 20ms window with 12 cepstral coefficients plus an energy term, yielding a 13-valued vector.

The next step is to quantize each vector using a specially-designed quantization tree. This recursively divides the vector space into *bins*, each of which corresponds to a leaf of the tree. Any MFCC vector will fall into one and only one bin. Given a segment of audio, the distribution of the vectors in the various bins characterize that audio. Counting how many vectors fall into each bin yields a histogram template that is used in the distance measure. For this demonstration, the distance between audio files is the simple Euclidean distance between their corresponding templates (or rather 1 minus the distance, so closer files have larger scores). Once scores have been computed for each audio clip, they are sorted by magnitude to produce a ranked list like other search engines.

4. COMPARISON OF MIR SYSTEMS

In Table 2, we present a comparison of the features of the various MIR systems under investigation. Note first that each of these systems was designed for a different purpose, and none of them can be considered a finished product.

This table allows one to get an overview of the state of the MIR systems available., the features that one may wish to include in an MIR system, and the areas where improvement is most necessary. It also highlights the need for a standardized testbed. Each of the MIR systems use a different database of files for audio retrieval. Both CatFind and the Music Retrieval Demo have databases with less than 500 files. Thus, any benchmarking estimates, such as retrieval times and efficiency, are rendered useless. MelDex, MelodyHound and ThemeFinder have databases containing over 10,000 files. This should be sufficient for estimating search efficiency and slalability.

5. EVALUATION ISSUES

Table 1 listed and compared the features available in existing online MIR systems. However, this is not sufficient for effective benchmarking and evaluation of possible music information retrieval systems that may appear in the near future and be used with large file collection. The question of what features to evaluate is determined by what we can measure that will reflect the ability of the system to satisfy the user. In a landmark paper, Cleverdon[21] listed six main measurable quantities. This has become known as the Cranfield model of information retrieval evaluation. Here, those properties are listed and modified as applicable for MIR.

1. The coverage of the collection, that is, the extent to which the system includes relevant matter.
2. The time lag, that is, the average interval between the time the search request is made and the time an answer is given. Consideration should also be made of worst case or close to worst case scenarios. It may be that certain genres or formats of music, as well as certain types of queries, e. g., query and retrieval of polyphonic transcription based audio may require far more time than other queries. Furthermore, if the testbed is particularly large, dispersed or unindexed, such as with peer-to-peer based internet, then bandwidth limitations and scalability may greatly reduce efficiency while maximizing the collection size.
3. The form of presentation of the output. For MIR systems this not only means having the option of retrieving various formats, symbolic and audio, but it also implies identifying multiple performances of the same composition.
4. The effort involved on the part of the user in obtaining answers to his search requests. So far, MIR research has been dominated by audio engineers, computer scientists, musicologists and librarians. As the field expands to include developers and user interface experts this issue will acquire more significance.
5. The recall of the system, that is, the proportion of relevant material actually retrieved in answer to a search request;
6. The precision of the system, that is, the proportion of retrieved material that is actually relevant.

It is claimed that (1)-(4) are readily assessed. (5) and (6) together measure the ability of the system to retrieve relevant documents while at the same time holding back non-relevant ones. It is assumed that the more effective the system the more it will satisfy the user. In general, it is assumed that precision and recall are sufficient for the measurement of effectiveness. However, due to the complexity of MIR systems, and multimedia information retrieval systems in general, coverage, time lag, form and effort cannot be ignored. Thus the authors recommend that benchmarking procedures take all six factors into account, at least until the field has matured to such an extent that MIR systems share similar qualities in regard to these factors.

What is meant by relevance is unclear. Different users may differ about the relevance or non-relevance of particular documents to given questions. Because it is desired that MIR systems perform well in conditions where the users and their criteria is not known a priori, it is suggested that relevance be determined by subjective double-blind testing of the systems. This can be used to judge the relevance of retrieved results (6), but it does not guarantee that as many as possible relevant documents have been retrieved. That requires more objective testing and knowledge of the testbed collection(5). Specifically, it requires that the testbed include many possible relevant and nonrelevant documents that may 'trick' the retrieval system A situation is created where a number of questions exist for which the preferred responses are known.

6. SUBJECTIVE AND OBJECTIVE TESTING

Lets consider an implementation of a true polyphonic audio music retrieval system. None of the five systems under study have the ability to do this, and at this stage, there is no published system, online or elsewhere, that does successful polyphonic MIR.

Determination of success in such a system is far more complicated than benchmarking of text based MIR systems. One primary reason for this is the relative ease by which one can measure the accuracy of retrieval results. A simple ranking system may be used to compare retrieved text documents to the query. This is well-established and there is a great deal of knowledge of how irrelevant documents, e.g., very large documents that contain all words in the query, might be retrieved, and how to omit them from the search. In addition, most users can easily compare a document to a query and determine the success of the match. Thus for text based systems, both subjective and objective testing of the matching of retrieval to query is straightforward.

However, music information retrieval benchmarking is a far more complex task, both subjectively and objectively. First, many different measures can be used to objectively compare the similarity of audio files, such as the methods used by the MIR systems described above. But it is quite possible for systems to be very similar, but the measure still fails to show this. One simple example is time stretching or pitch shifting. A time-stretched file still shares a lot in common with the original, but many time domain measures of similarity will show little similarity. Similarly, pitch shifting will often destroy frequency domain based measures.

Thus, the authors propose the following for the creation of subjective and objective retrieval accuracy measures that can be used to compare MIR systems, and improve and fine-tune an individual MIR system.

Consider a polyphonic MIR system, where a sample of a piece of music has been entered as the query. The polyphonic ranking might typically operate by devising a polyphonic transcription of

the query, and attempting to match that transcription to queries in the database. Ranked results are returned that regarded as close to the query, but the ranking may not be a sufficiently good measure of proximity that includes all the similarities that would be noticed by the listener. Essentially this is because the scalar rank does not necessarily capture the multi-dimensional nature of human recognition capabilities. What is really needed is for the user to interact with the returned results, by listening to them. Thus we propose that any successful MIR system needs to be able to play each match back, in much the same way that IR systems like Google let the user read all the documents it suggests.

But we can extend the use of audio playback within MIR systems, to their design and evaluation. We base the following proposals on widely used Mean Opinion Score methodology as used in assessing audio codecs and reproduction systems.[22]

Assume we are comparing MIR systems A and B. We have a query which we call X. Each system returns r results. Thus we have two ordered lists of suggested retrievals/matches, A[r] and B[r]. We also have a number of human test subjects, each of whom listen to triplets of sound snippets, using the ABX methodology[23-25], as follows.

First, the original query X, is played. Then, in random order, A[1] or B[1], then the remaining of that pair. Then the test is repeated with A[2] and B[2] and so on down the returned list. For each such triplet of sounds, the listeners rank whether the first or the second sound they heard is closer to the query.

The results over one query will not provide enough evidence for a comparison, so the above process must be repeated over a set of queries. For now let us assume these are of similar genre[1]. Then we can determine, on average, which MIR system returns better, perceptually relevant, matches.

Clearly, similar testing procedures could be devised as part of the process of developing an MIR system, where we would compare a system before and after a modification. The challenge is to find a collective objective test which yields results that match a subjective test.

However, subjective testing will highlight additional factors, such as noisiness and timbre, which do not show up in any simple measure of transcription similarities. Also similarity of a transcription is subject to interpretation. Therefore, subjective measures may be used to modify and improve transcription based ranking, and to gauge the importance of other differences between audio files. Subjective and objective testing therefore work in conjunction.

In order to ensure that the subjective tests measure as many similarities as possible, the testbed should include both very diverse and very similar files. The diversity should include songs of varying length from various genres. Similar files should be included, with the intention that a retrieval system should both identify similarities, and distinguish between them. These should include, where possible

1. Multiple live and studio recordings
2. Recordings of the same composition by different artists, including recordings with different instruments
3. Variations on a theme, including variations far enough removed from the original that they are no longer relevant
4. Recordings in various genres
5. Recordings of widely varying length
6. Recordings of the same music in different file formats

As a final test, the queries should be varied in all the same ways as the testbed is varied. Reduced versions of the testbed should be searched as well, in order to identify how the MIR system scales with data set size, and how relevance changes with data scarcity.

[1] Future MIR systems may differ in their capabilities over different genres. Performance testing should accommodate for this.

7. CONCLUSION

In this work, we have laid down a framework for benchmarking of future MIR systems. At the moment, this field is in its infancy. There are only a handful of MIR systems available online, each of which is quite limited in scope. Still, these benchmarking techniques were applied to five online systems. Proposals were made concerning future benchmarking of full online audio retrieval systems. It is hoped that these recommendations will be considered and expanded upon as such systems become available.

8. ACKNOWLEDGMENTS

The authors would like to thank J. Stephen Downie for his insightful comments and criticisms concerning this work. Funding for this work was partially provided by the EPSRC.

9. REFERENCES

[1] Cahill, M. Using XML for Score Representation. in COST G-6 Conference on Digital Audio Effects (DAFx-01). 2001. Verona, Italy.

[2] Crawford, T., D. Byrd, and J. Gibson, The Nightingale Notelist, in Beyond MIDI: The handbook of musical codes, E. Selfridge-Field, Editor. 1997, MIT Press.

[3] Parsons, D., The Directory of Tunes and Musical Themes, ed. S. Brown. 1975: Cambridge.

[4] Ghias, A., et al. Query By Humming. in ACM Multimedia (ACMMM). 1995. San Francisco, USA.

[5] Blum, T., et al., Content Based Classification, Search and Retrieval of Audio. IEEE Multimedia, 1996. 3(3): p. 27-36.

[6] Pickens, J., et al. Polyphonic Score Retrieval. in Proceedings of the 3rd International Symposium on Music Information Retrieval (to appear). 2002. Paris, France.

[7] Reiss, J.D., J.-J. Aucouturier, and M.B. Sandler. Efficient Multidimensional Searching Routines for Music Information Retrieval. in 2nd Annual International Symposium on Music Information Retrieval. 2001. Bloomington, Indiana, USA.

[8] Lemström, K. and S. Perttu. SEMEX - An Efficient Music Retrieval Prototype. in First International Symposium on Music Information Retrieval (ISMIR). 2000. Plymouth, Massachusetts.

[9] Lemström, K. and J. T arhio. Searching monophonic patterns within polyphonic sources. in The RIAO Conference. 2000. Paris, France.

[10] Bello, J.P. and M. Sandler. Blackboard system and top-down processing for the transcription of simple polyphonic music. in COST G-6 Conference on Digital Audio Effects (DAFx-01). 2001. Verona, Italy.

[11] Tzanetakis, G. and G. Essl. Automatic Musical Genre Classification Of Audio Signals. in Int. Symposium on Music Inform. Retriev. (ISMIR). 2001. Bloomington, IN, USA.

[12] Blum, T., et al., Method and Article of Manufacture for Content-Based Analysis, Storage, Retrieval and Segmentation of Audio Information, U. S. A. Patent No. 5,918,223 (June 29 1999), Muscle Fish (acquired by Audible Magic).

[13] Lap, Y.C., CatFind. 1999, University of Hong Kong: Hong Kong. **http://zodiac.csis.hku.hk:8192/catfind/Music/ContentSearch.html**

[14] Bainbridge, D., MELDEX: A Web-based Melodic Locator Service. Computing in Musicology, 1998. 11: p. 223-229.

[15] McNab, R.J., et al., The New Zealand Digital Library MELody inDEX. D-Lib Magazine, 1997. May. **http://www.cs.waikato.ac.nz/~nzdl/meldex/demo.html**

[16] Prechelt, L. and R. Typke, An interface for melody input. ACM Transactions on Computer-Human Interaction, 2001. 8(2): p. 133-149. **http://name-this-tune.com/**

[17] Kornstädt, A., Themefinder: A web-based melodic search tool. Computing in Musicology, 1997-98. 11: p. 231-236.

[18] Huron, D., C.S. Sapp, and B. Aarden, Themefinder. 2000. **http://www.themefinder.org**

[19] Wild, J., A Review of the Humdrum Toolkit: UNIX Tools for Musical Research, created by David Huron. Music Theory Online, 1996. 2(7).

[20] Foote, J.T. Content-Based Retrieval of Music and Audio. in Multimedia Storage and Archiving Systems II. 1997. Dallas, Texas: SPIE. **http://www.fxpal.com/people/foote/musicr/doc0.html**

[21] Cleverdon, C.W., J. Mills, and M. Keen, Factors Determining the Performance of Indexing Systems, Volume I - Design, Volume II - Test Results, ASLIB Cranfield Project, in Readings in Information Retrieval, K. Sparck Jones and P. Willett, Editors. 1997, Morgan Kaufmann: San Francisco.

[22] ITU-R Recommendation BS. 1116. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. 1999, International Telecommunications Union.

[23] Grusec, T., L. Thibault, and R. Beaton. Sensitive Methodolgies for the Subjective Evaluation of High Quality Audio Coding Systems. in Audio Engineering Society UK DSP Conference. 1992. London.

[24] Burstein, H., Approximation Formulas for Error Risk and Sample Size in ABX Testing. Journal of the Audio Engineering Society, 1988. 36: p. 879.

[25] Clark, D.L., A/B/Xing DCC. Audio, 1992. 76(4): p. 32.