ARTHUR: Retrieving Orchestral Music by Long-Term Structure

Jonathan Foote
FX Palo Alto Laboratory, Inc.
3400 Hillview Avenue
Palo Alto, CA 94304
foote@pal.xerox.com

The structure of most music is sufficient to characterize the work. Arthur G. Lintgen of Philadelphia proved able to identify unlabeled classical phonograph recordings by the softer and louder passages visible in the LP grooves. His example indicates that the long-term musical structure can be used for identification and retrieval. This paper presents a automatic music retrieval system inspired by Mr. Lintgen's approach, and is thus named ARTHUR in his honor. Like its namesake, ARTHUR retrieves audio on the basis of long-term structure, specifically the variation of soft and louder passages. Unfortunately, this technique is not robust for much popular music, a shortcoming shared with Mr. Lintgen. ARTHUR retrieves audio on the basis of long-term structure, specifically the variation of soft and louder passages. The long-term structure is determined from envelope of audio energy versus time in one or more frequency bands. Similarity between energy profiles is calculated using dynamic programming. Given an example audio document, other documents in a collection can be ranked by similarity of their energy profiles. Experiments are presented for a modest corpus that demonstrate excellent results in retrieving different performances of the same orchestral work, given an example performance or short excerpt as a query.

The Algorithm

The retrieval algorithm is relatively straightforward. First, an "energy profile" is computed for every audio document in the collection. The energy profile is a representation of the average acoustic energy versus time. overall energy structure of the documents is quite similar. This property is exploited by the ARTHUR system. Once the energy profile is computed, it can be compared with other profiles. The similarity between them is calculated using dynamic programming (DP), which calculates the cost of the best alignment. This is a good measure of signal similarity: identical signals will have a cost of zero, while increasing differences will increase the matching cost. For retrieval, the cost is used to rank corpus documents by similarity to the query. The DP algorithm is especially well suited to matching energy profiles. Unlike other methods, the DP algorithm accounts for differences in both the features and the relative timing. In other words, features need not match exactly, nor are they required to occur exactly at the same relative times. Thus the DP algorithm smoothly matches performances with variable dynamics, tempos, and tempo changes. In addition, DP easily handles the case when query and corpus files do not start and end at exactly the same time. This is particularly useful for IR, because it allows queries to be any shorter fragment of longer works.

Figure 1. Arthur G. Lintgen identifying a phonograph record from the grooves

Experiments

This paper presents results using an extremely modest corpus of less than 100 documents. The corpus for the first experiment contained 58 audio documents, including movements from three versions of Brahms' *Symphony No. 3*, For evaluation, different performances of the same movement were considered relevant, while different movements or works were not.

For the actual experimental evaluation, each of the three performances of the four movement of Brahms' Third Symphony was used as a query. Each of the 58 corpus documents was then ranked by similarity to each of the 12 queries. For every query, the other two performances of the same movement ranked higher than any other document, thus yielding recall and precision rates of 100% on this corpus. Investigation revealed that piano music was not retrieved nearly as well as purely orchestral music because of the high variability of the piano's acoustic energy. For a more challenging retrieval task, the corpus of Experiment I was augmented with four performances of the three movements of the Beethoven *Piano Concerto No. 2* and the Chopin *Concerto No. 2*. These additional documents increased the overall corpus size to 82. From the 72 documents

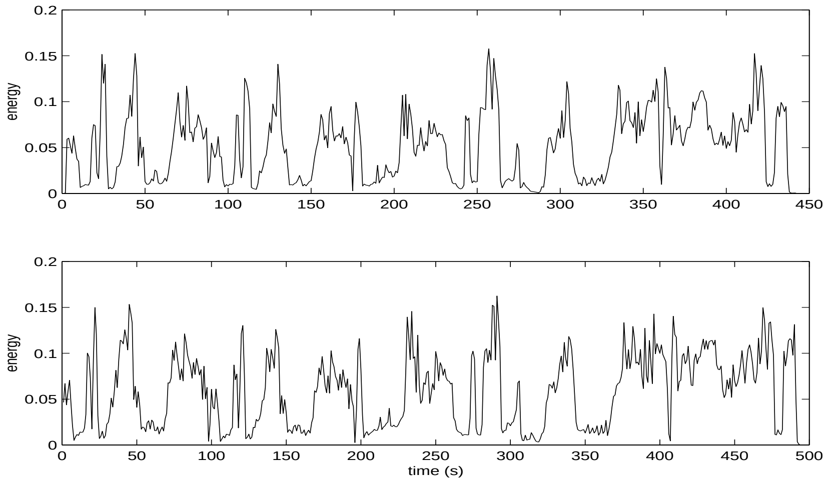

Figure 2. Energy profiles for two different performances of the first movement of Beethoven's *Fifth Symphony*
Top: Herbert von Karajan/Berliner Philharmoniker. Bottom: Eric Leinsdorf/Boston Symphony Orchestra.

retrieved at this cutoff of 3, 60 were relevant, giving a retrieval precision of 83% Retrieval performance for the original Brahms query set was not affected by the corpus expansion, and remained at 100%. We then attempted to improve retrieval by using features more informative than pure energy. For every audio document, a long-term spectral representation was computed using the Short-Time Fourier Transform. Using spectral features increased the retrieval performance from 83% to 96% on the piano query set. The retrieval performance on the original Brahms query set remained at 100%.

The last experiment investigates retrieval accuracy as a function of query length. The queries were variable-length fragments of the queries from the previous experiment. Once again, the DP algorithm can find the best match, regardless of where the clip starts or ends. Figure 3 shows the results of the experiment. As might be expected, longer queries perform better, and spectral features dramatically outperform purely energetic features. Queries needed to be truncated at 130 seconds so as not to exceed the length of the shortest query document; this is one reason the best results in this experiment do not approach the precision achieved when using the full-length query documents. The non-monotonic results are no doubt due to the small test corpus: experiments on larger corpora should yield smoother curves.

These experiments are primarily a proof of concept given the admittedly small corpus size. There is considerable scope for improving the retrieval performance yet further. Thanks to Steven Smoliar for discussions and providing much of the test corpus data.

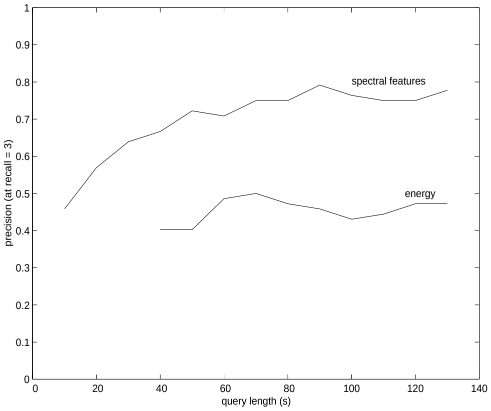

Figure 3. Retrieval performance versus query length, for piano query set (24 queries).

Suggested Reading

[1] Holland, Bernard. "A Man Who Sees What Others Hear." *The New York Times*. p. C28, 19 November 1981

[2] Foote, J. "Content-Based Retrieval of Music and Audio," in *Multimedia Storage and Archiving Systems II, Proc. SPIE*, Vol. 3229, Dallas, TX.

[3] Wold, E., Blum, T., Keislar, D., and Wheaton, J., "Classification, Search and Retrieval of Audio," in *Handbook of Multimedia Computing*, ed. B. Furht, pp. 207-225, CRC Press, 1999.

[4] Pye, D., "Content-based Methods for the Management of Digital Music," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2000,* vol. IV pp. 2437, IEEE

[5] J. Kruskal and D. Sankoff, "An Anthology of Algorithms and Concepts for Sequence Comparison," in *Time Warps, String Edits, and Macromolecules: the Theory and Practice of String Comparison*, eds. D. Sankoff and J. Kruskal, CSLI Publications, (Stanford) 1999

[6] Rabiner, L., and Juang, B.-H., *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ, 1993