Summary of Presentation on Audio-based Music Similarity Analysis

This presentation is about techniques and applications on audio-based music similarity. The investigation is largely based on Jonathan Foote's work (Foote 1997; Foote et al. 2002) and Beth Logan's work (Logan and Saloman 2001).

According to the literatures that have been reviewed, because of the soaring of music industry and internet audio-based music resources, which brings the world into a song-corpus-based era from the old album-based era, there are several types of promising applications based on audio music similarity such as, query music by similarity, automatic play-list generation, automatic Dick Jockey and automatic summarizing and categorization.

The approaches of the reviewed literature can be summarized to have common properties as follows: First, they all try to find a hidden metaphor to relate acoustic similarity with statistical audio features, based on certain defined application domains. Second, distance between features from different audio samples is taken as the measure of similarity. Different mathematic distance measures are used without definite direction on their relationship with perceptual qualities. Finally, supervised or unsupervised approaches are both used.

The common steps of a typical audio-based similarity analysis approach are as follows: First, the time-domain signal of an audio sample is windowed and some low-level features (usually frequency-domain or cepstral features) are extracted. This is often called Audio Parameterization (Foote 1997). Second, a process of feature quantization is launched. Low-level features are transformed into some higher-level statistical features according to the metaphor, resulting a set of dimensionality-lowered features. Either supervised or unsupervised approach can be taken. In a supervised approach, the quantization is discriminative and involves off-line training. In an unsuperized approach, statistical clustering is often used. Finally, the distances between high-level features from different audio samples, as the similarity measure, are derived, based on certain type of measures. The results of measure query against the target corpus are then ranked by similarity in terms of this distance.

In the first paper mentioned above, Foote demonstrated an audio content-based retrieval technique that is purely data-driven and is not dependent on particular audio characteristics. The metaphor used for audio similarity is first to find a sample-specific template, which is a histogram of MFCC feature vectors partitioned from an primitive feature space, and then to compare the templates from different audio samples to calculate the distance between them as the similarity measure. A supervised tree-based vector quantizer is used for generation of the histogram. Different measures such as Euclidean and cosine distance measures are used. Experiments on simple sounds and music clips are performed and compared to the Muscle Fish system (Foote 1999), showing the effectiveness of the approach and its advantage on music-clips-based tasks. A web demo query system built on top of the technique is available (Demo: Music Retrieval by Content).

In the second mentioned Foote’s paper, a technique on audio retrieval by rhythmic similarity is introduced. This unsupervised technique is not restricted to certain musical genres or music with certain rhythmic features. The metaphor used for similarity is a Beat Spectrum based on the idea that the autocorrelation of spectral-related audio features, within a certain time range, can hint rhythmic information. A novel technique called Similarity Matrix is used to derive the Beat Spectrum and visualize the audio structure. Experiments of this technique has been made on both different-tempo versions of same music and small group of different music excerpts with different distance measure, showing cosine distance and a new measure called Fourier Beat Spectral Coefficients generated excellent precision of 96.7%. An automatic play-list generation application using this technique is introduced (Foote et al. 2002).

In the Logan’s paper, an unsupervised technique based on the idea of Foote’s histogram-like approach is introduced. Instead of trained tree-based quantization, which may risk too much emphasis on several local bins, a statistical clustering is used to form higher-level feature from MFCC-based low-level audio features. A probability-based distance measure Earth Mover’s Distance is used. Evaluations on both objective and subjective relevance are performed, which demonstrated the effectiveness of the technique. The experiment on the robustness of the technique to data corruption is also given, showing its excellent anti-corruption ability. A play-list system based on the technique is introduced (Logan and Saloman 2001).

The future effort towards application of audio-based music similarity includes employing common classification and machine-learning techniques, combining knowledge constraints and finding theoretical basis of the selection of distance measure in terms of perpetual properties of music (Logan and Saloman 2001).

Bibliography

Foote, J. 1997. Content-based retrieval of music and audio. *Multimedia Storage and Archiving Systems II, Proceeding of SPIE* 3229: 138–47.

Foote, J. 1999. An overview of audio information retrieval. *Multimedia Systems* 7(1): 2–11.

Foote, J., M. Cooper, and U. Nam. 2002. Audio retrieval by rhythmic similarity. *Proceedings of the 3rd International Symposium on Musical Information Retrieval*: 265–66.

Logan, B., and A. Saloman. 2001. A music similarity function based on signal

analysis. *IEEE International Conference on Multimedia and Expo.*
Demo: Music Retrieval by Content.
http://www.fxpal.com/people/foote/musicr/doc0.html (accessed 10 March 2005).