Statistical Significance in Song-Spotting in Audio

Pedro Cano, Martin Kaltenbrunner, Oscar Mayor, Eloi Batlle
Music Technology Group
IUA-Pompeu Fabra University
{pedro.cano,modin,oscar.mayor,eloi.batlle}@iua.upf.es

ABSTRACT

We present some methods for improving the performance a system capable of automatically identifying audio titles by listening to broadcast radio. We outline how the techniques, placed in an identification system, allow us detect and isolate songs embedded in hours of unlabelled audio yielding over a 91% rate of recognition of the songs and no false alarms. The whole system is also able of working real-time in an off-the-shelf computer.

1. INTRODUCTION

A monitoring system able to automatically generate play lists of registered songs can be a valuable tool for copyright enforcement organizations and for companies reporting statistics on the music broadcasted. The difficulty inherent in the task is mainly due to the difference of quality of the original titles in the CD and the quality of the broadcasted ones. The song is transmitted partially, the speaker talks on top of different fragments, the piece is maybe playing faster and several manipulation effects are applied to increase the listener's psycho-acoustic impact (compressors, enhancers, equalization, bass-booster, etc…). An additional difficulty is that there are no markers in broadcasted radio informing when the songs start and end.

In this scenario, the article focus on the pattern matching techniques that, given a sequence of audio descriptors, are able to locate a song in a stream avoiding false alarms. Shortly the whole system works as follows, off-line and out of a collection of music representative of the type of songs to be identified, an alphabet of sounds that describe the music is derived. These audio units are modeled with Hidden Markov Models (HMM). The unlabelled audio and the set of songs are decomposed in these audio units. We end up then with a sequence of letters for the unlabelled audio and a database of sequences representing the original songs. By approximate string matching the song sequences that best resembles the audio the most similar song is obtained. We point out the importance of assessing statistical relevance on the best matching song found in order to avoid false positives. We end up explaining how these techniques can be applied to continuous stream of audio and commenting the results.

2. AUDIO PERCEPTUAL UNITS

From an acoustic point of view, music can be described as a sequence of acoustic events. To be able to identify titles it is relevant to extract information about the temporal structure of these sequences. The first step converts the acoustic signal into a sequence of abstract acoustic events. Speech events are described in terms of phones. In music modeling this is not so straightforward. Using, for instance notes would have disadvantages: Often notes are played simultaneously (accords, polyphonic music) and music samples contain additional voices or other sounds. The approach therefore followed is learning relevant acoustic events, that is, finding the set of "fundamental sounds" in which we can decompose audio and representing them with a letter. The alphabet of audio perceptual units is derived through unsupervised clustering using cooperative HMM from a database of several thousand titles [1].

3. SEQUENCE ALIGNMENT

Having derived HMM models for the audio perceptual units, we can decompose the songs into a symbolic representation. Instead of comparing raw audio, for identifying titles, we compare the sequence of letters of unknown audio against the sequences corresponding to all the songs to identify. The search for a sequence in a database similar to the query sequence is performed by approximate string pattern matching [2]. A measure of the difference between two sequences is the edit distance, defined as the minimum number of character insertions, deletions and substitutions needed to make them equal. An arbitrary weight can be associated with every edit operation, as well as with a match.

The dynamic programming algorithm is guarantied to find the best alignment between a pair of sequences given a particular choice of scoring matrix and gap penalties [3]. There are several variants of the dynamic programming algorithm that yield different kinds of alignments. The Neddleman and Wunsch is a global alignment, that is to say, it aligns the entire length of both sequences. For our particular case this is not suitable since it is typical that a song in the radio is broadcasted partially. The variant known as the Smith-Waterman algorithm yields a local alignment. It aligns the pair of regions within the sequencesIn our application, since the query audio sequence must be compared to several thousand titles, we run a heuristic approximation to the Smith-Waterman algorithm that allows us perform the matching much faster named FASTA[4].

3.1 The choice of substitution scores

The weighted scores for substitutions of the edit distance are calculated to account for bias in the replacement of symbols between the original and the broadcasted song sequences. A set of original CD and corresponding radio songs are selected and manually edited by cutting pieces so that the pieces of audio are synchronized. Then a similarity ratio, R_{ij} is computed for the symbols in the sequences

$$R_{ij} = \frac{q_{ij}}{p_i p_j}$$

where q_{ij} is the relative frequency with which the symbols i and j are observed to replace each other in the manually aligned sequences. p_i and p_j are the frequencies at which the symbols i and j occur in the set of songs in which the substitutions are observed. Their product, $p_i\ p_j$, is the frequency at which they would be expected replace each other if the replacements were random. If the observed replacement rate is equal to the theoretical replacement rate, then the ratio is one ($R_{ij} = q_{ij} / p_i p_j = 1.0$). If the replacements are favored with the manipulative effects above described the ratio will be greater than one and if there is selection against the replacement the ratio will be less than one. The similarity reported in the similarity score matrices S_{ij} is the logarithm to this ratio.

4. STATISTICAL SIGNIFICANCE

Considering the possible uses of the system, a great concern in the similarity searching above described is a false-positive error. We would not like to include in a play list for a copyright enforcement association a song that has not been played. Any two sequences composed of letters from the same alphabet can be aligned to show some measure of similarity. Typically alignment scores of unrelated sequences are small, so that the occurrence of unusually large scores can be attributed to a match. However, even unrelated sequences can occasionally give large scores in the local alignment regime. Although these events are rare, they become important when one attempts a search of a big and expanding sequence database. How often will an event at least as extreme as the one just observed happen if these events are the result of a well defined, specific, random process? It is imperative to understand the statistics of the high-scoring events, in order to estimate the statistical significance of a high-scoring alignment.

In the case of gapless alignment, it is known rigorously [6] that the distribution of alignment scores of random sequences is the Gumbel or extreme value distribution (EVD), which has a much broader tail than that of the Gaussian distribution. For the case of gapped alignment, there is no theory available to predict the distribution of alignment scores for random sequences. It has been conjecture that the score distribution is still of the Gumbel form. Also our tests on sequence of descriptors extracted from audio seem to show a good fit to the Extreme Value Distribution. The EVD is of the form:

$$E = Kmne^{-\lambda S}$$

where E is the expected number of hits with score >=S, m is the size of the query sequence, n is the size of the database. λ and K the are Gumbel constants and must be estimated from a large scale comparison of random sequences. The FASTA or various implementation of the SW algorithm, produce optimal alignment scores for the comparison of the query sequence to sequences in the database. Most of these scores involve unrelated sequences, and therefore can be used to estimate λ and K.

5. ON-LINE SYSTEM

We have then a method for comparing fragments of audio against a database of songs for a best match and statistical method for assessing its goodness. Both the symbolic extraction and the matching against the database run fast on a normal machine. The approach for, having a continuous stream of broadcasted audio, identify songs consists in sending hypothesis to match against the database every few seconds. That is, the superstring resulting from the conversion of the raw audio to symbols is windowed with overlap. So every 10 seconds, a sequence corresponding to two and a half minutes of sound is compared to the database. As a result of each comparison a set of candidates is shown along with its expectation (E-value). A candidate with sufficiently low E-value suggests that the query is related to that candidate sequence and therefore can be added to the play list. Along with the candidate sequence, an alignment with the query is provided. With the timing associated to the query sequence an estimation of the beginning and ending time of the song broadcasted can be obtained and printed in the play list.

6. RESULTS

The system has been tested with 24 hours of radio recorded from 10 different stations against a database of around 2500 songs of commercial music. The radio data contains among music, jingles commercials... 147 songs registered in the system (its original version is in the database). The system yields a result of 133 (little over a 91%) songs recognized and no false positive. By lowering the threshold of acceptance of a candidate raises the results to 135 correctly identified but false positives appear as well. When working on-line, the delay between the moment a song starts sounding and it is added correctly to the play list is about one minute as average. The system runs in more than real-time in a Pentium III 500Mhz.

7. REFERENCES

[1] Batlle, E., Cano, P., Automatic Segmentation for Music Classification using Competitive Hidden Markov Models, Proceedings International Symposium on Music Information Retrieval (2000)

[2] Gusfield, D., Algorithms on Strings, Trees and Sequences, Cambridge University Press (1997)

[3] Smith, T.F. and Waterman, M.S., Identification of common molecular subsequences, Journal of Molecular Biology. (1981), 195-197.

[4] Pearson, W.R. and Lipman, D.J. Improved tools for Biological Sequence Comparison. Proc. Natl. Acad. Sci. (1988) 85 : 2444-2448.

[5] Nicholas, H.B., Deerfield D. W., Ropelewski, A.J. A Tutorial on Searching Sequences Databases and Sequence Scoring Methods (1997)

[6] Karlin, S. And Altschul, S.F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA 87 (1990), 2264-2268.