The Intervalgram: An Audio Feature for Large-Scale Cover-Song Recognition

Thomas C. Walters, David A. Ross, and Richard F. Lyon
Google, Brandschenkestrasse 110, 8002 Zurich, Switzerland
tomwalters@google.com

Abstract. We present a system for representing the musical content of short pieces of audio using a novel chroma-based representation known as the ‘intervalgram’, which is a summary of the local pattern of musical intervals in a segment of music. The intervalgram is based on a chroma representation derived from the temporal profile of the stabilized auditory image [10] and is made locally pitch invariant by means of a ‘soft’ pitch transposition to a local reference. Intervalgrams are generated for a piece of music using multiple overlapping windows. These sets of intervalgrams are used as the basis of a system for detection of identical melodic and harmonic progressions in a database of music. Using a dynamic-programming approach for comparisons between a reference and the song database, performance is evaluated on the ‘covers80’ dataset [4]. A first test of an intervalgram-based system on this dataset yields a precision at top-1 of 53.8%, with an ROC curve that shows very high precision up to moderate recall, suggesting that the intervalgram is adept at identifying the easier-to-match cover songs in the dataset with high robustness. The intervalgram is designed to support locality-sensitive hashing, such that an index lookup from each single intervalgram feature has a moderate probability of retrieving a match, with few false matches. With this indexing approach, a large reference database can be quickly pruned before more detailed matching, as in previous content-identification systems.

Keywords: Cover Song Recognition, Auditory Image Model, Machine Hearing.

1 Introduction

We are interested in solving the problem of cover song detection at very large scale. In particular, given a piece of audio, we wish to identify another piece of audio representing the same underlying composition, from a potentially very large reference set. Though our approach aims at the large-scale problem, the representation developed is compared in this paper on a small-scale problem for which other results are available.

There can be many differences between performances with identical melodies. The performer may sing or play at a different speed, in a different key or on a different instrument. However, these changes in performance do not, in general, prevent a human from identifying the same melody, or pattern of notes. Thus,
given a performance of a piece of music, we wish to find a representation that is to the largest extent possible invariant to such changes in instrumentation, key, and tempo.

Serra [12] gives a thorough overview of the existing work in the field of melody identification, and breaks down the problem of creating a system for identifying versions of a musical composition into a number of discrete steps. To go from audio signals for pieces of music to a similarity measure, the proposed process is:

– Feature extraction
– Key invariance (invariance to transposition)
– Tempo invariance (invariance to a faster or slower performance)
– Structure invariance (invariance to changes in long-term structure of a piece of music)
– Similarity computation

In this study, we concentrate on the first three of these steps: the extraction of an audio feature for a signal, the problem of invariance to pitch shift (both locally and globally) and the problem of invariance to changes in tempo between performances of a piece of music. For the first stage, we present a system for generating a pitch representation from an audio signal, using the stabilized auditory image (SAI) [10] as an alternative to standard spectrogram-based approaches. Key invariance is achieved locally (per feature), rather than globally (per song). Individual intervalgrams are key normalized relative to a reference chroma vector, but no guarantees are made that the reference chroma vector will be identical across consecutive features. This local pitch invariance allows for a feature that can track poor-quality performances in which, for example, a singer changes key gradually over the course of a song. It also allows the feature to be calculated in a streaming fashion, without having to wait to process all the audio for a song before making a decision on transposition. Other approaches to this problem have included shift-invariant transforms [9], the use of all possible transpositions [5] or finding the best transposition as a function of time in a symbolic system [13]. Finally, tempo invariance is achieved by the use of variable-length time bins to summarize both local and longer-term structure. This approach is in contrast to other systems [5,9] which use explicit beat tracking to achieve tempo invariance.

While the features are designed for use in a large-scale retrieval system when coupled with a hashing technique [1], in this study we test the baseline performance of the features by using a Euclidean distance measure. A dynamic-programming alignment is performed to find the smallest-cost path through the map of distances between a probe song and a reference song; partial costs, averaged over good paths of reasonable duration, are used to compute a similarity score for each probe-reference pair.

We evaluate performance of the intervalgam (using both SAI-based chroma and spectrogram-based chroma) using the ‘covers80’ dataset [4]. This is a set of 160 songs, in 80 pairs that share an underlying composition. There is no explicit notion of a ‘cover’ versus an ‘original’ in this set, just an ‘A’ version and