5

COGNITION-INFORMED PITCH DESCRIPTION

5.1 INTRODUCTION

In empirical, corpus-based music research, we may want to be able to describe high-level, cognition-related qualities of music, such as its complexity, expectedness and repetitiveness, from raw audio data. The features we would need to do this have not gotten the attention they deserve in MIR's audio community, perhaps due to the 'success' of low-level features when perceptual and cognitive validity is not a concern (i.e., most of the time—see chapter 3).

In this chapter, we propose a set of novel cognition-informed and interpretable descriptors for use in mid- to high-level analysis of pitch use, and applications in content-based retrieval. They are based on symbolic representations that were originally inspired by results in music cognition, and have been since been shown to work well in symbolic music analysis. We focus on features that describe the use of *pitch*, which we use here not just in its perceptual definition (a psycho-acoustic dimension related to the frequency of a sound event), but in a wider sense that includes both *harmony* and *melody*. In the long run, we believe, better mid-level and high-level pitch descriptors will provide insight into the building blocks of music, including riffs, motives and choruses.

In the second part of this chapter, we test the new descriptors in a cover song detection experiment. At the end of this section, we motivate why this type of application can serve as a good test case. Sections 5.3.1 – 5.3.3 present the data, methods, and results.

5.1.1 *Improving Pitch Description*

Symbolic Music Description and Music Cognition

In chapter 1.2.4, we discussed the theoretical arguments for symbolic vs. audio-based music representations. Both representations also have *practical* advantages. Symbolic music representations encode music in terms of discrete events. Discrete music events such as notes can be *counted*, making statistical modeling more straightforward (see many of the systems reviewed by Burgoyne in [21]). Symbolic representations also allow more easily for models that acknowledge the order of events in time or look for hierarchical structure on the music (see de Haas [61] and chapter 3). None of these abstractions are easily accessed through currently available audio features. This may explain why symbolic music has been the representation of choice in all of the corpus-based music cognition research reviewed in chapter 3. Since we aim to develop new audio representations that get a step closer to describing cognition-level qualities of music, we may be able to take some inspiration from the technologies that exist in symbolic music description.

Rhythm description is not included in this chapter. The state of the art in rhythm description, e.g., measuring inter onset intervals or syncopation, requires a reliable method of estimating and characterizing streams of salient onsets. Robust rhythm description thus involves two of the most difficult remaining challenges in MIR right now: note segmentation in a polyphonic mix, and separation of notes into streams (see section 2.1.4 on melody extraction). This makes rhythm description on any level beyond tempo and meter very difficult with the current state of the art in the above tasks, similarly to how complete music transcription isn't robust enough to be useful at present time. Timbre is also not considered in this chapter, as it is not typically described in symbolic terms as much as melody, harmony and rhythm. We come back to this point in the conclusions in chapter 9.

We focus on harmony and melody, or 'pitch'. Melodic pitch estimation, discussed in section 2.1.4, involves fewer steps than music transcription, making it quite reliable in comparison. To ensure robustness, however, we will make two further restrictions on the kind of representations we pursue.

1. melody description should not require note segmentation
2. harmony description should not involve chord estimation

The state-of-the-art in chord estimation is considerably more successful then rhythm transcription, yet any transcription in general runs the risk of introducing unknown biases due to the assumptions of the transcription systems (e.g., for many chord transcriptions: datasets on which it was trained or evaluated). See also section 3.2.3 for a more elaborate discussion of this issue.

Finally, we add two more restrictions.

3. descriptors should be invariant to non-pitch-related facets such as timbre, rhythm and tempo
4. descriptors should have a fixed size

Chroma features or pitch class profiles are a proven and relatively robust representation of harmony but, like most feature time series, vary in length with the audio from which they have been extracted, and are not tempo- and translation-invariant. An adequate fixed-size descriptor should capture more detail than a simple chroma pitch histogram, while preserving tempo and translation invariance.

5.1.2 *Audio Description and Cover Song Detection*

In the second part of this chapter, three new descriptors will be evaluated. We now argue that the task of *scalable cover song retrieval* is very suitable for developing descriptors that effectively capture mid- to high-level musical structures, such as chords, riffs and hooks.

Cover detection systems, as explained in section 2.2.3, take a query song and a database and aim to find other versions of the query song.

Most successful cover detection algorithms are built around a two-stage architecture. In the first stage, the system computes a time series representation of the harmony or pitch for each of the songs in a database. In the second stage, the time series representing the query is compared to each of these representations, typically through some kind of alignment, i.e., computing the locations of maximum local correspondence between the two documents being compared. Such alignment methods are very effective, but computationally expensive.

Consider an archivist or musicologist, who aims to exhaustively search for musical relations between documents in a large audio corpus, e.g., an archive of folk music recordings. Archives like this may contain large numbers of closely related documents, such as exact duplicates of a recording, different renditions of a song, or variations on a common theme. The particularities of such variations are of great interest in the study of music genealogies, oral transmission of music, and other aspects of music studies [196].

Cover song detection can be helpful here, but alignment-based techniques are no longer an option: a full pair-wise comparison of 10,000 documents could easily take months.[1] This is why some researchers have been developing the more scalable cover song detection techniques reviewed in section 2.2.3. Scalable strategies are often inspired by audio fingerprinting and involve the computation of an indexable digest of (a set of) potentially stable landmarks in the time series, which can be stored and matched through just a few inexpensive look-ups.

The challenges laid out above make cover song detection an ideal test case to evaluate a special class of descriptors: harmony, melody and rhythm descriptors, global or local, which have a fixed dimensionality and some tolerance to deviations in key, tempo and global structure. If a collection of descriptors can be designed that accurately describes a song's melody, harmony and rhythm in a way that is ro-

1 MIREX 2008 (the last to report run times) saw times of around 1.2–3.2 seconds per comparison. These algorithms would take 1.8–6 years to compute the $\frac{1}{2}10^8$ comparisons required in the above scenario, or 6–20 weeks at current processor speeds (assuming eight years of Moore's law—processor speeds doubling every 2 years).

bust to the song's precise structure, tempo and key, we should have a way to determine similarity between the musical 'gist' of two songs and assess if the underlying composition is likely to be the same.

Note that the interest, in this case, is not necessarily in the large-scale performance or efficiency of the system, but in the evaluation of a fixed-sized descriptor in the context of performance variations.

5.2 COGNITION-INSPIRED PITCH DESCRIPTION

Symbolic Music Description and Expectation

The possibility of computing statistics on discrete musical events, and the possibility of representing time, have inspired some researchers to use music data to test models of musical expectation. There is an increasing amount of evidence that the primary mechanism governing musical expectations is statistical learning [78, 153]. On a general level, this implies that the relative frequencies of musical events play a large role in their cognitive processing. Expectations resulting from the exposure to statistical patterns have been shown to affect the perception of melodic complexity and familiarity, preference, and recall [78], making them a particularly interesting for some of the applications in this thesis.

Statistical distributions of musical events are often modeled using the notion of *bigrams*. Bigrams are, simply put, ordered pairs of observations. Word bigrams and letter bigrams, for example, are much-used representations in natural language processing and computational linguistics [116]. Figure 21 shows the set of word bigrams extracted from the phrase "to be or not to be".

In *melody* description, numerous authors have proposed representations based on pitch bigrams, most of them from the domain of cognitive science [110, 134, 166]. This comes as no surprise: distributions of bigrams effectively encode two types of probabilities that influence expectation: the prior probability of pairs of pitches, and, if we condition on the first pitch in each pair, the conditional frequency of a pitch given the one before.

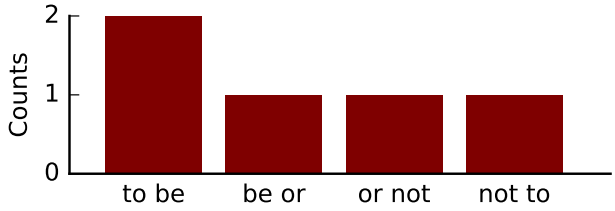


Figure 21.: A bigram count for the phrase "to be or not to be".

In the description of *harmony*, bigram and other n-gram representations have been used as well. In [167], for example, Rohrmeier and Graepel find that chord bigram-based models predict chord sequences better than plain HMM on a subset of the *Band-in-a-box* corpus of Jazz chord sequences, though not as good as trigrams, higher-order n-grams and autoregressive HMM.

When symbolic data are not available, bigrams in the strict sense are difficult to compute. For this reason, the features we propose are approximations of the notion of bigrams that can also be interpreted as probability distributions, encode a notion of order in time, and can be computed from audio. The descriptors will be named *audio bigrams*.

5.2.1 *Pitch-based Audio Bigrams*

First, we propose three pitch bigram-like representations.

The Pitch Bihistogram

The first new feature is a melody descriptor. It essentially captures how often two pitches p_1 and p_2 occur less than a distance d apart.

Consider a 12-dimensional melody time series $M(t, p)$. As in chroma, M contains pitch activations, quantized to semitones and folded to one octave. If a pitch histogram is defined as:

$$h(p) = \sum_t M(t, p), \tag{38}$$

with $p \in \{1, 2, \ldots 12\}$, the proposed feature is then defined:

$$PB(p_1, p_2) = \sum_t M(t, p_1) \max_\tau (M(t + \tau, p_2)) \tag{39}$$

with $\tau = 1, 2 \ldots \Delta t$. This will be referred to as the *pitch bihistogram* (*PB*), a bigram representation that can be computed from any melodic pitch time series, up to arbitrarily high frame rates. Note that the use of pitch classes rather than pitch creates an inherent robustness to octave errors in the melody estimation step, making the feature insensitive to one of the most common errors encountered in pitch extraction.

Alternatively, scale degrees can be used instead of absolute pitch class. In this scenario, the melody time series $M(t, p)$ must first be aligned to an estimate of the piece's overall tonal center. As a tonal center, the tonic can be used. However, for extra robustness to mis-estimation of the tonic, we suggest to use the tonic for major keys and the minor third for minor keys, so that mistaking a key for its relative major or minor has no effect.

Chroma Correlation Coefficients

The second feature representation we propose looks at vertical rather than horizontal pitch relations. It encodes which pitches appear simultaneously in a 12-dimensional chroma time series $H(t, p)$. From $H(t, p)$ we compute the correlation coefficients between each pair of chroma dimensions to obtain a 12×12 matrix of *chroma correlation* coefficients $CC(p_1, p_2)$:

$$CC(p_1, p_2) = \sum_t H^\star(t, p_1)\ H^\star(t, p_2), \tag{40}$$

in which $H^\star(t, p)$ is $H(t, p)$ after column-wise normalization, i.e., subtracting, for every p, the mean and dividing by the standard deviation:

$$H^\star = \frac{H(t, p) - \mu(p)}{\sigma(p)} \tag{41}$$

$$\begin{aligned} \mu(p) &= \tfrac{1}{n} \textstyle\sum_t H(t, p) \\ \sigma^2(p) &= \tfrac{1}{n-1} \textstyle\sum_t (H(t, p) - \mu(p))^2 \end{aligned} \tag{42}$$

This descriptor is similar to the chroma covariance feature proposed by Kim in [90]. Like the pitch bihistogram, the chroma features can be transposed to the same tonal center (tonic or third) based on an estimate of the overall or local key.

Harmonization
Finally, the harmonization feature (*HA*) is a set of histograms of the harmonic pitches $p_h \in \{0, \ldots, 11\}$ as they accompany each melodic pitch $p_m \in \{0, \ldots, 11\}$. It is computed from the pitch contour $P(t)$ and a chroma time series $H(t, p_h)$, which should be adjusted to have the same sampling rate.

$$HA(p_m, p_h) = \sum_t M(t, p_m)\ H(t, p_h). \tag{43}$$

Musically, the harmonization feature summarises how each note of the pitch tends to be harmonised.

From a memory and statistical learning perspective, the chroma correlation coefficients and harmonization feature may be used to approximate expectations that include: the expected consonant pitches given a chord note, the expected harmony given a melodic pitch, and the expected melodic pitch given a chord note. Apart from [90], where a feature resembling the chroma correlation coefficients is proposed, information of this kind has yet to be exploited in a functioning (audio) MIR system. Like the pitch bihistogram and the chroma correlation coefficients, the harmonization feature has a dimensionality of 12×12.

5.2.2 *Pitch Interval-based Audio Bigrams*

The next three descriptors extend the pitch-based audio bigrams above to interval representations. Whereas pitch bigram profiles are expected to strongly correlate with the key of an audio fragment, interval bigrams are key-invariant, which allows them to be compared across songs.

Melodic Interval Bigrams

The melodic interval bigrams (*MIB*) descriptor is a two-dimensional matrix that measures which pairs of pitch intervals follow each other in the melody It is based on *pitch trigrams*, an extension of the two-dimensional bihistogram in equation 39:

$$\text{trigrams}(p_1, p_2, p_3) = \sum_t \max_\tau(M(t-\tau, p_1))\ m(t, p_2) \max_\tau(M(t+\tau, p_3)), \tag{44}$$

with again $\tau = 1 \ldots \Delta t$ and M the melody matrix, the binary chroma-like matrix containing the melodic pitch class activations. The result is a three-dimensional matrix indicating how often triplets of melodic pitches (p_1, p_2, p_3) occur less than Δt seconds apart.

The pitch class triplets in this feature can be converted to interval pairs using the function:

$$\text{intervals}(i_1, i_2) = \sum_{p=0}^{11} X((p - i_1) \bmod 12,\ i,\ (p + i_2) \bmod 12). \tag{45}$$

This maps each trigram (p_1, p_2, p_3) to a pair of intervals $(i_2 - i_1, i_3 - i_2)$. A broken major chord $(0, 4, 7)$ would be converted to $(4, 3)$, or a major third followed by a minor third. Applied to the pitch trigrams, the intervals function yields the *melodic interval bigrams* descriptor:

$$MIB(i_1, i_2) = \text{intervals}(\text{trigrams}(M(t, p))) \tag{46}$$

Harmonic Interval Co-occurrence

The harmonic interval co-occurrence descriptor measures the distribution of triads in an audio segment, represented by their interval representation. It is based on the *triad profile*, which is defined as the three-dimensional co-occurrence matrix of three identical copies of the chroma time series $H(t, p)$ (t is time, p is pitch class):

$$\text{triads}(p_1, p_2, p_3) = \sum_t H(t, p_1)\ H(t, p_2)\ H(t, p_3). \tag{47}$$

The triad profile can be made independent of absolute pitch by applying the intervals function (equation 45). This yields the *harmonic interval co-occurrence* matrix:

$$HIC(i_1, i_2) = \text{intervals}(\text{triads}(H(t, p))) \tag{48}$$

As an example, a piece of music with only minor chords will have a strong activation of HIC$(3, 4)$, while a piece with a lot of tritones will have activations in HIC$(0, 6)$ and HIC$(6, 0)$.

Harmonization Intervals
Finally, the harmonization feature can be extended to obtain the harmonization intervals (HI) feature, defined as:

$$HI(i) = \sum_t \sum_{p=0}^{12} M(t, p)\ H(t, (p + i) \bmod 12) \tag{49}$$

Unlike the 12×12 *MIB* and *HIC*, the *HI* is 12-dimensional, and measures the distribution of intervals between the melody and harmony.

5.2.3 *Summary*

We have proposed three 12×12 melody and harmony descriptors based on pitch, and two 12×12 and one 12-dimensional descriptor based on pitch intervals. The first three will now be used in an experiment to test how much harmonic and melodic information they encode. The last three features will be used in part iii of this thesis, where a key-invariant descriptor is needed.

Finally, we note that the term 'audio bigrams' assumes a necessarily loose interpretation of the term bigrams. It is a loose interpretation in that not all pitch pairs in the pitch bihistogram follow each other immediately—some other pitch content might be present in between—and pitch pairs in the chroma correlation feature are simultaneous rather than adjacent.[2] However, this loose interpretation is necessary if we want to apply the idea of bigrams to audio at all. Because of the

2 similar to 'skipgrams' in natural language processing.

continuous nature of audio signals, there is no notion of 'adjacency' without resorting to some arbitrary discretization of time—we can only measure what is 'close together'.[3] The use of the word bigrams will also become more clear as we define the concept of audio bigrams more formally in the next chapter.

5.3 EXPERIMENTS

To test whether the features we propose capture useful information, we perform a number of cover song detection experiments.

5.3.1 *Data*

Two datasets are used: *covers80*, a standard dataset often used as as a benchmark, and the *translations* dataset. The *covers80* dataset is a collection of 80 cover song pairs, divided into a fixed list of 80 queries and 80 candidates. Results for this dataset have been reported by at least four authors [174], and its associated audio data are freely available. It is not as big as the much larger *Second Hand Song* dataset.[4] The problem with the Second Hand Song dataset, however, is that it is distributed only in the form of standard Echo Nest features. These features do not include any melody description, which is the basis for two of the descriptors proposed in this chapter.

The *translations* dataset is a set of 150 recordings digitized especially for this study: 100 45-rpm records from the 50's and 60's, and 50 78-rpm records, most of them from before 1950. They have been selected from the collections of the Netherlands Institute for Sound and Vision, who own a 'popular music heritage' collection that was, until recently,

3 Of course, frequency-domain representations of signals come in discrete frames by construction, but to choose this same discretization for measuring pitch transitions would restrict the scope of our features to pitch patterns on very short time scales. A discretization based on onsets or beats may be more meaningful. Here, however, we argue that the feature should be meaningful both in the presence and absence of rhythm.

4 `http://labrosa.ee.columbia.edu/millionsong/secondhand`

only accessible in the form of manually transcribed metadata (such as titles, artists, original title, composer).

Amongst these records are 50 pairs of songs that correspond to the same composition. All of these tunes are translated covers or re-interpretations of melodies with a different text—in the early decades of music recording, it was very common for successful singles to be re-recorded and released by artists across Europe, in their own language.[5] Such songs are especially interesting since they guarantee a range of deviations from the source, which is desirable when models of music similarity are tested. Some pairs are re-recordings by the original artists, and thus very similar, other song pairs need a very careful ear to be identified as 'the same'.

5.3.2 *Methods*

Features

Four experiments were carried out, each following the same retrieval paradigm. The features that will be used are the three audio pitch bigram representations proposed in section 5.2.1: the pitch bihistogram, chroma correlations coefficients and the harmonization feature.[6]

Similarity

A query song is taken out of the collection and its feature representation is computed. The representations for all candidate songs are then ranked by similarity, using the *cosine similarity* $s_{\cos}$,

$$s_{\cos}(x, y) = \cos(\alpha_{xy}) = \frac{x \cdot y}{\|x\| \, \|y\|}, \tag{50}$$

5 from correspondence with the curators at the Netherlands Institute for Sound and Vision and Meertens Institute

6 The other, pitch interval-based audio bigrams had not been formalized yet at the time when this experiment was carried out, but close variants of theses features were found to perform no better than the pitch-based audio bigrams in initial tests.

where α_{xy} is the angle between the vectors x and y. Note that, even though no indexing is used, cosine similarities can be computed much faster than alignment-based distances.

Evaluation

To evaluate the results, three evaluation measures are used. 'Recall at 1' (R_1) is the fraction of queries for which the correct song is returned in first position. It will be used in the evaluation of the *covers80* results, as it is the measure most commonly used to compare results for this dataset. 'Recall at 5' is the fraction of queries for which the correct cover is returned in the top 5 ranked results. It is included to give an impression of the performance that could be gained if the current system were complemented with a good alignment-based approach to sort the top-ranking candidates, as proposed by [198], among others. Mean Average Precision (MAP), a more standard evaluation measure, will be used for the *translations* dataset.[7] To compute it, the candidates' ranks r are used to obtain the reciprocal rank r^{-1} for each relevant document returned. Since the datasets in these experiments only contain one relevant document to be retrieved for each query, precision and reciprocal rank are the same, and the mean Average Precision can simply be obtained by taken the mean of r^{-1} over all queries.

In the *translations* dataset, every song that is part of a cover pair is used as a query, and the candidate set always consists of all the other songs. In the *covers80* dataset, a fixed list of 80 queries and 80 candidates is maintained. A random baseline was established for this configuration at MAP = 0.036 for the *translations* dataset, with a standard deviation of 0.010 over 100 randomly generated distance matrices. In *covers80*, less songs are used as a retrieval candidate. Random baselines were found at R_1 = 0.012 (0.013), R_5 = 0.060 (0.026).

7 The difference is due to the a change in small change in implementation between the experiments and the availability of different baselines for each of the datasets.

Experiment 1: Global Fingerprints

The first experiment involves a straightforward evaluation of a few feature combinations using the *covers80* dataset. The three descriptors were extracted for all 160 complete songs. Pitch contours were computed using Melodia and chroma features using HPCP, with default settings [168].[8] For efficiency in computing the pitch bihistogram, the pitch contour was median-filtered and downsampled to $1/4$ of the default frame rate. The bihistogram was also slightly compressed by taking its point-wise square root.

As we observed that key detection was difficult in the present corpus, the simplest key handling strategy was followed: features for a query song in this experiment were not aligned to any tonal center. Instead, each query is transposed to all 12 possible tonics, and the minimum of the 12 distances to each other fingerprint is used to rank candidates.

All representations (*PB*, *CC* and *HA*) were then scaled to the same range by normalizing them for each fragment (subtracting the mean of their n dimensions, and dividing by their standard deviation; $n = 144$). In a last step of the extraction stage, the features were scaled with a set of dedicated weights $w = (w_{PB}, w_{CC}, w_{HA})$ and concatenated to 432-dimensional vectors. We refer to these vectors as the *global fingerprints*.

The main experiment parameters for the features described above are d, the look-ahead window of the bihistogram *PB*, and the weighting w of each of the three features when all are combined. Parameter d was found to be optimal around 0.500 s. Figure 22 shows the pitch bihistogram, chroma correlation, and harmonization descriptors in matrix form for an audio fragment from the *translations* dataset.

Experiment 2: Thumbnail Fingerprints

In a second experiment, the songs in the database were first segmented into structural sections using structure features, as described by Serrà [175]. This segmentation approach performed best at the

8 see section 2.1.4 and `mtg.upf.edu/technologies`

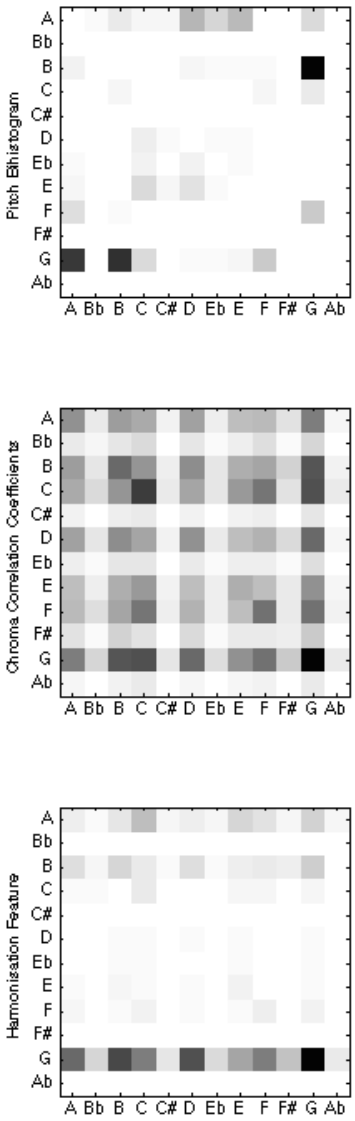


Figure 22.: An example of the pitch bihistogram, chroma correlation, and harmonization descriptors for an audio fragment from the *translations* dataset (in matrix form, higher values are darker). The pitch bihistogram at the top shows how pitch classes A and B appear closely after G, and G after B.

2012 MIREX evaluation exchange in the task of 'music structure segmentation', both for boundary recovery and for frame pair clustering. (A slight simplification was made in the stage where sections are compared: no dynamic time warping was applied in our model, assuming constant tempo.) From this segmentation, two non-overlapping thumbnails are selected as follows:

1. Simplify the sequence of section labels (e.g. ababcabcc): merge groups of section labels that consistently appear together (resulting in AAcAcc for the example above).
2. Compute the total number of seconds covered by each of the labels A, B, C... and find the two section labels covering most of the song.
3. Return the boundaries of the first appearance of the selected labels.

The fingerprint as described above are computed for the full song as well as for the resulting thumbnails, yielding three different fingerprints: one global and two *thumbnail fingerprints*, stored separately. As in experiment 1, we transposed query thumbnails to all keys, resulting in a total of 36 fingerprints extracted per query, and 3 per candidate.

Experiment 3: Stability Model

In the last experiment on the *covers80* data, we want to show that the interpretability of the descriptors allows us to easily incorporate musical knowledge.

In [18], a collaboration with Dimitrios Bountouridis, a model of stability in cover song melodies was introduced. The model was derived independently of these experiments, through analysis of a separate dataset of transcribed melodies of cover songs variations, the *Cover Song Variation* dataset. The dataset transcriptions of 240 performances of 60 distinct song sections from 45 song. It includes four or more performances of each section, as described in [17].

Given the melody contour for a song section, the model estimates the stability for each note in the melody. Stability is defined as the

probability of the same pitch appearing in the same place in a performed variation of that melody. The empirical stability is based on multiple sequence alignment of melodies from the database.

The stability estimates produced by the model are based on three components that are found to correlate with stability: the duration of notes, the position of a note inside a section, and the pitch interval. The details of the model and its implementation are described in [18]. As an example, figures 24 and 25 show how stability relates to pitch interval and position in a section.

From these three findings, two were integrated in the cover detection system. The stability vs. position curve (with position scaled to the $[0, 1]$ range) was used as a weighting to emphasize parts of the melody before computing the thumbnails' pitch bihistogram. The stability per interval (compressed by taking its square root) was used to weigh the pitch bihistogram directly. (Note that each bin in the bihistogram matrix corresponds to one of 12 intervals.) The trend in duration is weak compared with the other effects, so is not used in the experiments in this study.

Experiment 4: The translations *Dataset*

In experiment 4, we apply the global fingerprints method to a real music heritage collection, the *translations* dataset. This experiment is performed to find out what range of accuracies can be obtained when a dataset is used that better represents the musicology scenario described in the beginning of this chapter (5.1.2).

Key handling is approached differently here: fingerprints are transposed to a common tonal center, as found using a simple key-finding algorithm. A global chroma feature is computed from a full chroma representation of the song. This global profile is then correlated with all 12 modulations of the standard diatonic profile to obtain the tonic. The binary form (ones in the 'white key' positions and zeros in the others) is used, as in figure 23.[9]

9 Note that this doesn't assume that a melody is in major: minor key melodies simply get aligned to their third scale degree, as suggested earlier.

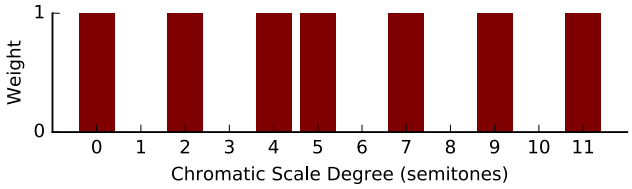


Figure 23.: Diatonic pitch profile used for key handling in the *translations* dataset.

5.3.3 *Results & Discussion*

Table 2 summarizes the results of the all four experiments.

Experiment 1

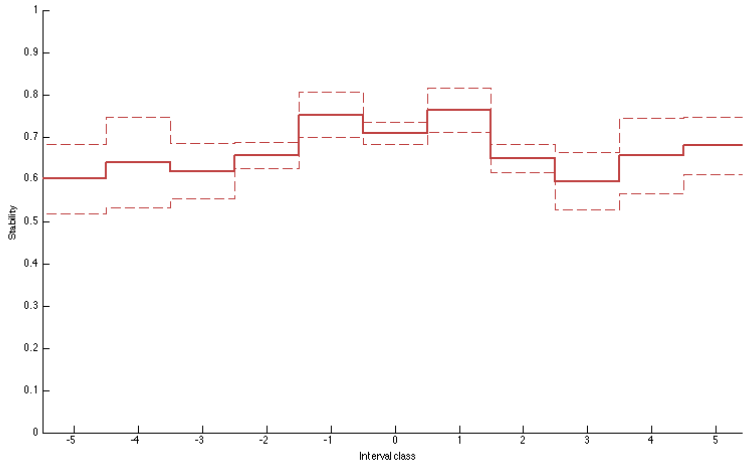
In Experiment 1, each descriptor was first tested individually (only one of w_B, w_C, w_H is non-zero). Results for h, a simple 12-dimensional melodic pitch histogram, and g, a harmonic pitch histogram (chroma summed across time), are added for comparison. They set a strong baseline of around R_1 = 16%–18%. From the newly proposed features, the harmony descriptors (chroma correlation coefficients) perform best, with an accuracy of over 30%, and when looking at the top 5, a recall of 53.8%.

After performing a minimal grid search with weights in $\{1, 2, 3\}$, it is found that, when the three new features are used together, R_1 and R_5 improve slightly. The chroma correlation coefficients contribute most, before the pitch bihistograms and harmonization features.

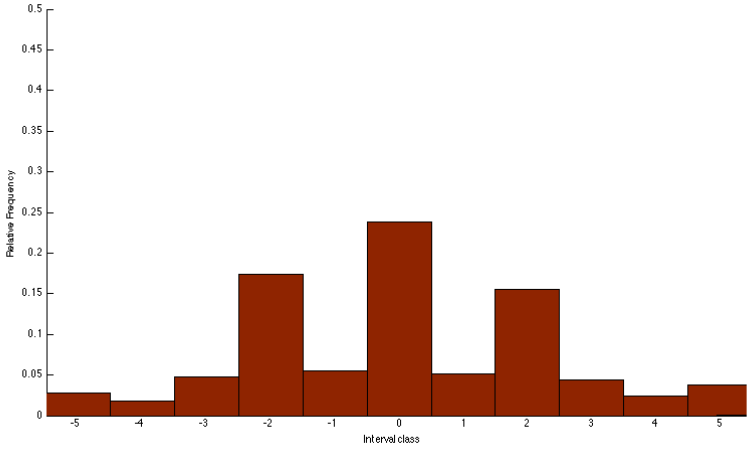
Experiment 2

Results for experiment 2 show that the global fingerprints outperform the thumbnail fingerprints (42.5% vs. 38.8%), and combining both types does not increase performance further.

In less optimal other configurations, it was observed that thumbnail fingerprints sometimes outperformed the global fingerprints, but

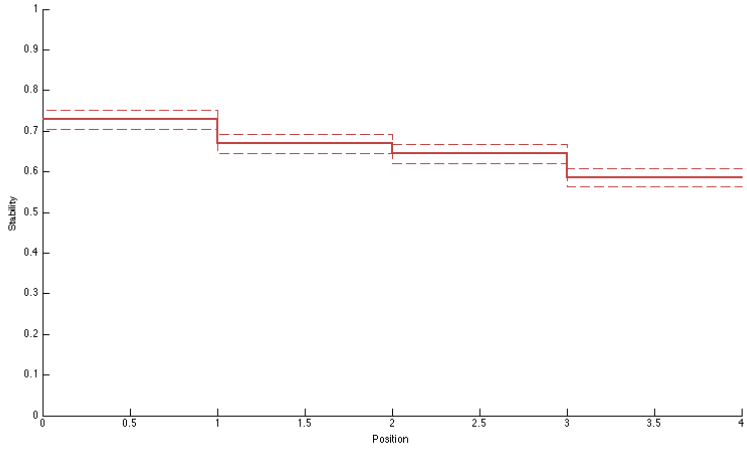


(a) Stability of notes by preceding pitch interval

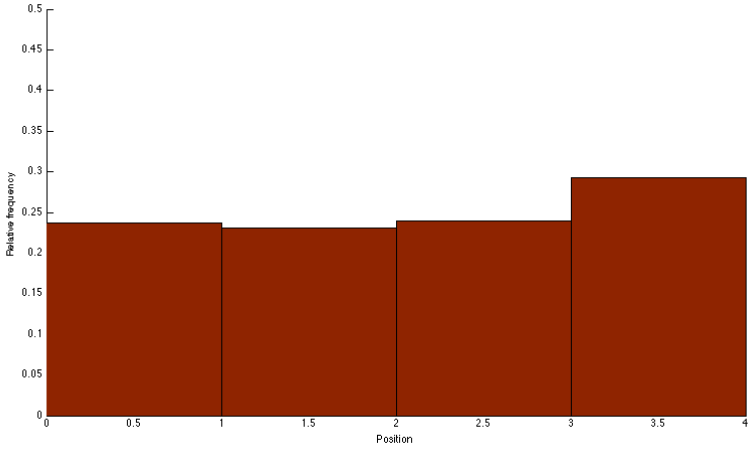


(b) Histogram of pitch interval

Figure 24.: Mean stability of melody notes as a function of pitch interval. Adapted from [18].



(a) Stability of notes by position in the section (scaled to $[0, 1]$)



(b) Histogram of note positions

Figure 25.: Mean stability of melody notes as a function of position in a section. Adapted from [18].

this result didn't generalize to the more optimal configuration with weights $w = (2, 3, 1)$. It is difficult to tell, at this moment, whether the relatively poor performance of the thumbnailing strategy is due to an advantage in capturing all of the songs pitch patterns in one representation, or to poor segmentation.

Experiment 3

When the stability model is integrated in the thumbnail fingerprints, top 1 accuracy reaches 45.0%. This result can be situated between precisions reported using the first alignment-based strategies (Ellis, 42.5%) and a recent scalable system (Walters, 53.8%), see table [198].

This justifies the conclusion that the descriptors proposed in section 5.2.1 capture enough information to discriminate between individual compositions, which we set out to show.

The straightforward embedding of domain knowledge from external analyses further attests to the potential in optimising the proposed representations and fully adapt them to the scalable cover song detection problem.

Experiment 4

Using the *translations* dataset, the precision obtained using just pitch histograms (h) is again added for comparison, and the result is substantial, about 0.27.

However, using just the pitch bihistogram (PB) feature, a MAP of around 0.43 can be obtained, compared to a very competitive 0.42 for using just the chroma correlations (CC). When these features are combined, the MAP goes up to 0.53. In the latter configuration, 44 of the 100 queries retrieve their respective cover version in first place, or in other words, $R_1 = 0.44$, comparable to the accuracy in *covers80*.

The evaluation results for two existing cover detection system, evaluated by Ralf van der Ham for the *translations* dataset, are also included in the table [194]. Van der Ham evaluated Ellis' seminal algorithm based on cross-correlation and Serrà's alignment-based algorithm that is currently considered state-of-the-art [49, 178]. The pro-

posed descriptors are not only faster, but, as can be seen from Table 2, also more powerful than Ellis' cross-correlation method. Serrà's method is much slower, but far superior in performance.

In short, results for the *translations* dataset are not as good as state-of-the-art alignment-based methods, but fairly good for a scalable approach. Specifically, while the dataset poses some extra challenges to Ellis' method, performance for the proposed descriptors is on par with performance in the *covers80* dataset. This justifies the use of the proposed descriptors for the description of older popular music.

5.4 CONCLUSIONS

In this chapter, six new audio descriptors were proposed for the description of harmony and melody. Inspired by notion, from symbolic music analysis, of pitch and interval bigrams, we refer to them as 'audio bigrams', and distinguish (for now) between two kinds: pitch-based audio bigrams and pitch interval-based audio bigrams. Interpretations of the new pitch descriptors were discussed, and their descriptive power is tested in a cover song retrieval experiment.

Performance figures for the experiments, though not state-of-the-art, are a strong indication that the pitch bihistogram feature, the chroma correlation coefficients and the harmonization feature capture enough information to discriminate between individual compositions, proving that they are at the same time meaningful and informative, a scarce property in the MIR feature toolkit.

To illustrate the benefit of the features' simplicity and straightforward interpretation, an independent model of cover song stability has been successfully integrated into the system. Finally, the main findings were confirmed in a cover detection experiment on the *translations* dataset, a dataset of older popular music. In the next chapter, a generalized formulation of the audio bigram paradigm will be proposed.

covers80 dataset experiments	Descriptor	R_1	R_5
Random baseline		0.012	0.060
Ellis, 2006 [48]		0.425	
Ellis, 2007 [48]		0.675	
Walters, 2012 [198]		0.538	
Exp. 1: global fingerprints	h	0.188	0.288
	g	0.163	0.363
	$PB \Leftrightarrow w = (1,0,0)$	0.288	0.438
	$CC \Leftrightarrow w = (0,1,0)$	0.313	0.538
	$HA \Leftrightarrow w = (0,0,1)$	0.200	0.375
	$w = (2,3,1)$	0.425	0.575
Exp. 2: thumbnail fingerprints	$w = (2,3,1)$	0.388	0.513
Exp. 2: global + thumbnail fingerprints	$w = (2,3,1)$	0.425	0.538
Exp. 3: both fingerprints + stability model	$w = (2,3,1)$	0.450	0.563

translations dataset experiments	Descriptor	MAP
Random baseline		0.04
Ellis, 2006 [48]		0.40
Serrà, 2012 [178]		0.86
Exp. 4: global fingerprints only	h	0.27
	$PB \Leftrightarrow w = (1,0,0)$	0.43
	$CC \Leftrightarrow w = (0,1,0)$	0.42
	$HA \Leftrightarrow w = (0,0,1)$	0.30
	$w = (1,1,0)$	0.53

Note. Performance measures are *recall at 1* (R_1; proportion of covers retrieved 'top 1') and *recall at 5* (R_5; proportion of cover retrieved among the top 5) for the *covers80* dataset and MAP for the *translations* dataset. w are feature weights. h = pitch histogram, PB = pitch bihistogram, CC = chroma correlation coefficients, HA = harmonization.

Table 2.: Overview of cover song experiment results.