Weighted Scales

A High-Level Descriptor in the MPEG-7 Framework

Grégoire Carpentier
gregoire.carpentier@ircam.fr

Jérôme Barthélemy
Jerome.barthelemy@ircam.fr

May 9th, 2005

Abstract - In this paper we present the current development of the Weighted Scale Descriptor Scheme (WSDS), a tool for representing scales and temperaments used in music pieces, in order to extend query-by-content possibilities in music databases, allowing information retrieval on scales and modes similarity criterions. Our descriptor is an extension of the currently existing scale descriptor in the MPEG-7 MelodyDS.

WSDS was first introduced in March 2004 at the 68th MPEG meeting in Munich, joint with the 3rd Musicnetwork open workshop (see MPEG documents m10568 and n6454). WSDS was refined in October 2004 at the 70th MPEG meeting in Palma de Mallorca (see MPEG documents m11313 and n6811) and will be finalized at the 73rd MPEG meeting in Poznan.

1 General Purpose and usage

The development of WSDS was motivated by the need for a fine metadata representation of scales, modes and temperaments used in melodic segments. Scales, modes and temperaments have been recognized for centuries as the basics of musical harmony, and there is still no multimedia-integrated tool for representing them in an efficient manner, whereas descriptors have already been proposed for rhythm, meter, rhythmic patterns, timbre, melodies, etc… The MPEG-7 open standard, which is devoted to content description, and its extensible XML schema, provide a suitable framework for the development of WSDS.

Briefly speaking, a scale is a discrete set of pitches, each one having a precise ratio of frequencies with a “base pitch”. In addition to these ratios, each element might also be

characterized by is relative importance within the scale. In western tonal music for instance, the base pitch and the 8th element of the scale (the fifth) play are playing a prominent role.

A fine description of scales, modes and temperaments can be achieved by using the frequency ratios between the elements of the scale and the base pitch. To this aim the scale descriptor used in the MPEG-7 Melody Descriptor Scheme can be used as the core of WSDS. Taking into account the relative roles played by the different elements of the scale, we complete this descriptor with a complementary estimation of the weight of each element of the scale. In addition, a model for ensuring choice of the base pitch must be defined, in order to avoid eventual inconsistencies between similar scales using different base pitches chosen on an arbitrary basis. To this end, the base pitch should be chosen as "the most important" pitch of the scale.

The main application of our Weighted Scale descriptors consists in advanced query-by-content in musical databases. In terms of audio indexation, our descriptor should extend the current query-by-content possibilities to what one might call "query-by-genre", where "genre" is a given musical mode, scale or temperament. This topic actually addresses the following issues:

- Finding out in a music database of all music pieces which bear some significant resemblance (in terms of musical mode, scale or temperament) to a given input music sample.
- Inferring from music databases the existence of significant clusters, i.e. performing categorization on a mode, scale or temperament resemblance criterion.

These both tasks rely on the capability of performing the comparison between two sets of Weighted Scales descriptors, i.e. on the capability of defining a relevant similarity measure on the descriptors' space.

2 The WSDS framework

2.1 Scale values

Each element of the scale is expressed as a ratio of frequency in semitones from the base pitch and expressed modulo the Transposing Ratio (when it is specified – see below). For an equal-tempered chromatic scale (and when the Transposing Ratio is specified), the values are [1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0]. The following formula is applicable (informative):

$$Elementvalue[n] = \left(12 \log 2\left(\frac{F_0[n]}{F_0[0]}\right)\right) \bmod TR$$

where $F_0[0]$ is the base pitch of the scale, and TR the Transposing Ratio. When TR is not specified, the modulo function does not apply.

An example of a non-traditional scale is the Bohlen-Pierce scale, which contains 13 notes that repeat after an octave and a fifth (a perfect twelfth, or a frequency ratio of 1:3). This scale would then be: [1.3324 3.0185 4.3508 5.8251 … 17.6871]

2.2 Scale weights

Motivation for the design is based on the need to associate weight values to the existing scale descriptor defined in MPEG-7 Melody Description Scheme. The scale descriptor is completed by a set of weights for each value of the scale. These weights can be estimated by different methods. The weights proposed in our descriptor scheme are determined by computing the total duration and the total energy of each element of the scale within the melody to be described.

2.3 Base pitch

The base pitch must be chosen as the "most important" element of the scale. We use a statistic model to determine which pitch is to be the "most important". Obviously, the model should allow different statistic methods in order to cope

with different musical contexts. These models are discussed below.

2.4 Transposing ratio

The Transposing Ratio (TR) is defined as the ratio between two base pitches of two consecutive cycles of the scale (when scale is repeated in cycles), using the same units as for the scale values, and the same formula. For European scales, Transposing Ratio is generally 12. For the Bohlen-Pierce scale, the Transposing Ratio is 19.0196.

The Transposing Ratio is not necessarily required. This allows describing through the same framework transposing scales as well as non-transposing ones. By default our descriptor assumes octave-transposing scales. The default Transposing Ratio is therefore 12.

3 Current development

This section aims at the following objectives:

- Show the evidence that weighted scales descriptor can be automatically extracted from simple audio signals, and briefly introduce the extraction method.
- Prove the interest of such descriptors with simple categorization examples.

In order to restrict the early steps of the development to quite simple cases, we only take into account here common Western music modes and temperaments, i.e. we assume the musical context to be the equally tempered context. The scale used in the first phase of our Core Experiment development will therefore have only 12 degrees and the Transposing Ratio (see upper in section 2.3) will always be 12.

3.1 F0 extraction

For F0 extraction we use the time-domain, autocorrelation-based algorithm called YIN[i]. YIN was designed and developed at IRCAM in 2002 by Alain de Cheveigné and is Open-Source under Copyright (CNRS/IRCAM Copyright © 2002, Centre National de la Recherche Scientifique). YIN was successfully tested with a database of monophonic samples from commercial recordings, including various instruments and genres.

3.2 Segmentation & pitch calculation

The segmentation phase (i.e. attack detection) is directly performed on YIN outputs, i.e. not only on a vector of local instantaneous frequencies, but also on a vector of local energy and an estimation of the local rate of periodicity. It basically consists in the following steps:

- First, each segment of the input signal where periodicity falls under a given threshold is considered as noise and irrelevant to F0 estimation.
- For the remaining signal we compute the rate of F0 local variation within a fixed-width moving window. An attack is detected every time the variation rate exceeds a given threshold. This threshold should be low enough to cope with eventual glissandi and high enough a avoid splitting a vibrato into different notes. From our we found out that a sensibility value of 50 cents (a quarter-tone) was convenient for our purpose.
- In order to avoid short notes (i.e. short inter-onset intervals) for which the YIN precision is generally poor we discards all notes with lower duration than a given resolution. Our experiments lead us to satisfying results with a Resolution value ion a range from 80 ms to 100 ms (milliseconds).
- Last, we assign for each remaining segment (attack to release or attack to next attack) a unique pitch obtained from the weighted average the interval local pitches. The average is weighted as a proportional function of the local periodicity rate.

3.3 Quantization – Transposing ratio

The frequencies estimated by filtering the YIN algorithm output are then quantized and converted into a discrete scale. This scale should be finer than the common standard MIDI specification (where notes are expressed in whole numbers of semitones) if we want to capture non-Occidental temperaments. In regard to our F0 extraction algorithm precision it seems reasonable to set the quantization unit at 25 cents, i.e. a quarter of a semitone. In a range of one octave, this scale has therefore 48 degrees.

However, the current limitations exposed at the beginning of this section made us restrict our scales descriptors to 12 degrees and the Transposing Ratio always be 12.

3.4 Base pitch

We use statistical methods to determine which degree of the scale should be chosen as the "most important":

- With our equally tempered context first restriction we use the famous Krumhansl-Kessler key finding algorithm[ii]. The 24 Krumhansl-Kessler key profiles could eventually be extended to other equally tempered modes such as Messiaen's scales, Bartok's scale or the Hungarian minor scale.
- When moving to our finer 48 degrees per octave quantization we can either extend the Krumhansl-Kessler key profiles by upsampling them or simply choose the Base Pitch as the one with a frequency that reaches (modulo the Transposing Ratio) the highest value of a weight function, i.e. the most "used" pitch of the quantized melody in terms of duration and power, and modulo the Transposing Ratio.

3.5 Scale weights

Once the Base Pitch as been identified a 12-length (within the equally-tempered restriction) vector of scale weights is then computed. Weights are calculated as a function of duration and power. Each degree's weight is normalized by the Base Pitch weight and set to 1.00 if it's higher.

3.6 Clustering

This last phase is still under development. The main task to perform is gathering a collection of melodic samples in which significant clusters (in terms of scales, modes and temperaments) are likely to be discovered by our software. Today we only have a small database (26 sound files), which contains only major or minor melodies. This database should be extended within the end of June in order to show a relevant evidence of the Weighted Scales descriptors clustering potentialities.

The categorization gives satisfying results with our little database. The major and minor clusters are well recognized and each sample is found to belong to the appropriate category. We use here a revised version of the common hierarchical linkage algorithm in which the user has not to know in advance the relevant number of clusters the algorithm should discover.

4 Future work

The future development of WSDS should be driven by the need to cope with a larger set of scales, modes and temperaments, ensuring the same categorization performances. The tasks to be performed in the future should therefore consist in:

- Extending the melodies database to other scales, modes and temperaments.
- Check the clustering algorithm robustness to this new database.

---

[i] de Cheveigné, A., and Kawahara, H. (2002). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., 111, 1917-1930.

[ii] Krumhansl, C. L. (1990). Cognitive Foundations of Musical Pitch. New York: Oxford University Press.