A Novel Chroma Representation of Polyphonic Music based on Multiple Pitch Tracking Techniques

Matthias Varewyck
matthias.varewyck@
ugent.be

Johan Pauwels
johan.pauwels@
ugent.be

Jean-Pierre Martens
martens@
elis.ugent.be

Department of Electronics and Information Systems
Ghent, Belgium

ABSTRACT

It is common practice to map the frequency content of music onto a chroma representation, but there exist many different schemes for constructing such a representation. In this paper, a new scheme is proposed. It comprises a detection of salient frequencies, a conversion of salient frequencies to notes, a psychophysically motivated weighting of harmonics in support of a note, a restriction of harmonic relations between different notes and a restriction of the deviations from a predefined pitch scale (e.g. the equally tempered western scale). A large-scale experimental evaluation has confirmed that the novel chroma representation more closely matches manual chord labels than the representations generated by six other tested schemes. Therefore, the new chroma representation is expected to improve applications such as song similarity matching and chord detection and labeling.

Categories and Subject Descriptors

I.5.4 [**Pattern recognition**]: applications—*Waveform analysis, signal processing*

General Terms

Algorithms, theory

1. INTRODUCTION

The harmonic structure of a song is mainly determined by chord progressions. A chord is defined as the simultaneous sounding of multiple musical notes. Each note is characterized by a fundamental frequency, hereafter called a pitch, and by harmonics thereof. On an equally tempered scale, the eligible notes are equidistantly spaced on a logarithmic frequency axis. Moreover, the pitch of each note can be multiplied/divided by a power of two until a value in a predetermined (but arbitrarily chosen) reference octave is obtained. This way, all eligible notes can be projected onto 12 chromas or pitch classes, denoted by musical symbols A, B, etc. This 12-dimensional representation can reveal the evidences of the chromas being present in the audio signal. Such a description is called a pitch class or chroma profile [5]. In this paper, the latter is used.

A chroma extractor is an algorithm that usually slices the signal into frames so as to compute a chroma profile for each frame. Per frame, the computation usually breaks down into (1) the construction of a log-frequency representation of the spectral content of the audio and (2) the conversion of that representation into a chroma profile. Although this general outline is shared by many chroma extractors, there are many extractors differing in the detailed implementation of this outline. Some algorithms first produce a short-term amplitude spectrum that is subsequently resampled along the log-frequency scale, whereas others apply a constant-Q transform [1] that immediately generates a log-frequency amplitude spectrum. Some algorithms fold each sample of the spectrum to a chroma [2] whereas others [6] just fold the samples corresponding to salient peaks in the spectrum. Some extractors make a distinction between frequencies that are or that are not supported by higher harmonics [9]. Some extractors also embed schemes for correcting the detected pitches whenever they appear to be slightly out-of-tune [4].

The aim of our research was to revisit some formerly proposed algorithms and to conceive a novel algorithm that produces chroma profiles possessing the potential to enhance tonality detection applications.

In the rest of this paper we first provide a detailed description of the novel chroma extractor. Then we describe an experimental evaluation measuring the similarity of the novel chroma profiles and the profiles emerging from previously published extractors with the 'ideal' profiles retrieved from manually extracted chord segments and labels.

2. PROPOSED CHROMA EXTRACTOR

The proposed chroma extractor works on a frame-by-frame basis. Per frame, it performs the following operations: (1) a constant bandwidth spectral analysis, (2) a determination of salient frequencies, (3) an extraction of multiple pitches, and (4) a transformation of the extracted pitches to a 12-dimensional chroma profile. Figure 1 illustrates the outputs of the different steps for a realistic frame. The introduction of a salient frequency extractor constitutes an important change in strategy with respect to our former chroma extractor [2]. However, other changes, described in the subsequent sections, have equally contributed to the experimentally established quality improvements.

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

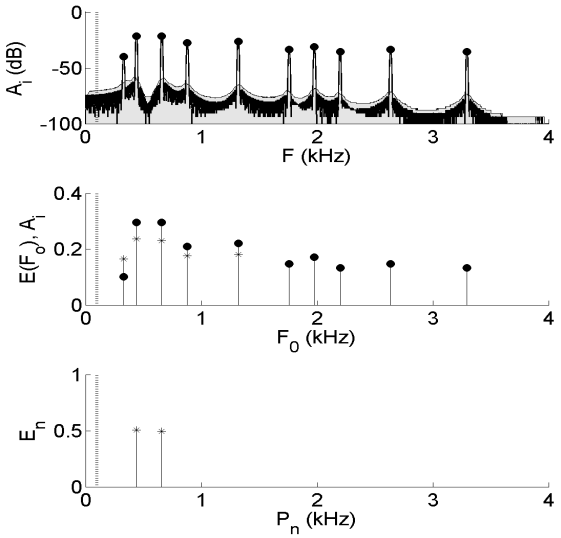

Figure 1: Illustration of the chroma profile extraction. The frame is a combination of three notes with pitches of 330 Hz (no harmonics), 440 Hz and 660 Hz (both having 5 harmonics). The top panel depicts the amplitude spectrum, the background spectrum (grey area) and the detected peaks (dots). The central panel depicts the detected peaks (dots) together with the corresponding pitches and their evidences (stars). The bottom panel shows the retained pitches and their normalized evidences.

2.1 Spectral analysis

The spectral analysis is a standard frame-based constant bandwidth analysis performed on audio that was resampled to 8 kHz. The analysis is conceived in view of our aim to provide chroma profiles that are suitable for chord classification. We therefore wanted to achieve that stable pure tones differing by one semitone are resolvable in a large part of the frequency range. We also wanted to prevent that common tone changes cause too much broadening of the spectral peak caused by such a tone. We argue that a frame size of 150 ms yields a good compromise. It corresponds to a 3-dB bandwidth of 7 Hz, which is smaller than a semitone for all frequencies above 130 Hz. In order to guarantee that the position and strength of a salient pure tone component can be estimated sufficiently accurately, we apply the DFT on a zero padded frame (8192 samples). This then yields a discrete amplitude spectrum with a bin of about 1 Hz.

2.2 Finding salient frequencies

Once available, the discrete log-amplitude spectrum is scanned from low to high frequencies to detect salient maxima. A maximum is called salient if it is (1) surrounded by two zones of at least 7 Hz (the bandwidth of the spectral analysis) wide that stay more than H dB below the value of the maximum, and (2) standing out more than L dB above the background spectrum. Since the first priority is to detect all peaks that correspond to relevant frequencies, the parameter H is set just to achieve that most random peaks emerging from a non-periodic sound do not result in salient frequencies. From informal tests we derived that $H = 3$dB is a very reasonable value.

The background spectrum is computed by filtering the amplitude spectrum with a median filter. The window size of the median filter should be large enough to make the filter 'blind' for a spectral peak emerging from a genuine tone, and short enough to preserve enough spectral detail in the estimated background spectrum. Since for our settings of the spectral analysis, the peak caused by a tone is about 7 Hz wide, we selected a window size of 80 bins. For the choice of L we observe that if salient peaks are detected with a margin of H dB, the margin L should be larger than H. Again, informal tests showed that $L = 6$ dB is an appropriate choice.

At this point in the processing of a frame we obtain a set $\{F_i, A_i\}$ of salient frequencies F_i and their amplitudes A_i. From these we will now derive a set of pitches.

2.3 Finding salient pitches

The term pitch refers to a virtual tone evoked by a set of physical frequencies that are in a harmonic relation to that tone, and a note is used to indicate the signal composed of these harmonics. Based on reasonable assumptions and arguments, we have conceived a novel multiple pitch extractor for determining the most salient pitches evoked by a signal.

2.3.1 Proposing pitch candidates

First we assume that only notes with a sufficiently high pitch are interesting candidates for building a chord. Therefore, notes with a pitch smaller than some lower bound F_L, which was set to 100 Hz, will be presumed to represent bass notes being generally not essential for determining a chord label. The same lower bound of 100 Hz was also introduced in e.g. [4]. In that paper there is even an upper limit on an eligible pitch. Next we assume that a salient pitch of frequency F_o must be marked by a salient frequency F_o, an assumption that has also been made in other extractors (e.g. [8]). Following these two assumptions, the set of pitch candidates can be limited to the set of salient frequencies F_i that are larger than F_L.

Once the pitch candidate F_o is identified, its evidence is computed. Therefore, consider all $F_i > F_o$ for which an integer $1 < k_i \leq K$ exists that satisfies

$$|F_i/k_i - F_o| < \epsilon\, F_o, \;\; k_i = \text{round}\,(F_i/F_o)\,, \qquad (1)$$

We opted for a tolerance ϵ of 3%, corresponding to half a semitone. K is considered one of the free parameters of our algorithm. If I_o represents the set of frequencies F_i meeting condition (1), then the evidence for pitch F_o is defined as a product of two factors

$$E(F_o) = \left(A_o^{0.5}\right)^{\alpha} \cdot \left(\frac{\sum\limits_{i \in I_o} A_i^{0.5} \gamma^{k_i}}{\sum\limits_{i \in I_o} \gamma^{k_i}} \right)^{1-\alpha} \qquad (2)$$

The first factor complies with our assumption that no pitch can be hypothesized unless it is also identified as a salient frequency. The second factor is the classical term following from sub-harmonic summation [7], but with two modifications: the use of square roots of the amplitudes and the introduction of a normalization factor.

The amplitude weighting is psychophysically motivated by the fact that the perceived loudness of a tone is roughly

proportional to the square root of its amplitude. The parameter α is introduced to have control over the balance between the importance of the fundamental frequency and that of its harmonics to the pitch evidence. It will be optimized on the basis of experiments. The parameter γ in the second factor must suppress the importance of the higher order harmonics. We used $\gamma = 0.75$ which was formerly found optimal in a pitch extractor working on vocal queries [3].

Note that since the sampling frequency is 8 kHz and since I_o consists of the harmonics of F_o only, the above process implies that no pitch larger than 2kHz can be hypothesized.

At this point we obtain a set $\{F_{oj}, E_{oj}\}$ of pitches and their evidences. In the next section we describe a method that will further reduce this pitch set.

2.3.2 Eliminating pitch candidates

We argue that if multiple salient frequencies F_i are all harmonics of the same fundamental frequency F_o, they should give rise to only one pitch. Consequently, pitches F_{oj} that are in a harmonic relation should be replaced by a single pitch. Nevertheless, to comply with the assumption that high order harmonics do not contribute much to the evidence of a pitch, we accept a high pitch that is in a harmonic relation to a pitch that is more than K times lower.

The pitch elimination process starts by treating the pitch candidates emerging from the previous step in a decreasing order of evidence. The pitch with highest evidence is always retained. The remaining pitches F_{oj} are rejected if they are either a harmonic or subharmonic of a formerly retained pitch P. The rejection of F_{oj} is carried out as follows:

$$\text{Reject } F_{oj} > P \quad \text{if} \quad \exists k \leq K : |F_{oj}/k - P| < \epsilon\, P$$
$$\text{Reject } F_{oj} < P \quad \text{if} \quad \exists k \leq K : |k\, F_{oj} - P| < \epsilon\, P$$

This process is repeated for all candidate pitches. The value of ϵ was obviously copied from Equation (1). For each surviving pitch P_n, we then convert the pitch evidences $E(P_n)$ emerging from the previous step to a pitch salience E_n

$$E_n = \frac{E(P_n)}{\sum_n E(P_n)} \tag{3}$$

representing the relative importance of that pitch in the ensemble of the retained pitches. I.e., E_n is supposed to be a kind of model of the probability that one can hear out pitch P_n when listening to the sound.

2.4 Pitch-to-chroma conversion

In the final stage each P_n is assigned to the closest note frequency F on the equally tempered western scale. Then the corresponding element of the chroma profile is incremented by E_n times the value of a triangular function with its peak in F and its zeros in $0.97F$ and $1.03F$. Pitches diverging from the western scale are thus automatically suppressed in the formation of chroma evidences. In fact, a pitch located halfway between two eligible notes will not contribute at all. This strategy differs from the one advocated by [8] and others: map an out-of-tune pitch to the closest note frequency on the scale and add it with full weight to the chroma profile.

3. EXPERIMENTAL EVALUATION

In this section, the proposed chroma extractor is compared to six reference extractors: two in-house created baseline algorithms and four formerly published extractors of which two are found on the web, one is implemented by us and the last one is the extractor we formerly developed ourselves. Each extractor is run with its preferred parameter settings.

3.1 Reference chroma extractors

The first baseline algorithm (BL-STFT) creates a linear amplitude spectrum which is subsequently converted to a log-frequency spectrum by distributing each sample over the two closest bins on a log-frequency scale. Finally, the log-frequency bins are folded into one octave. The second baseline extractor (BL-constQ) embeds a constant-Q transform [1] which gets directly converted to a chroma profile.

Three of the four formerly published extractors were selected for their own singularities: the first one[1] from Ellis and Poliner [4] uses the mean phase derivative instead of the mean amplitude of the individual FFT bins. The second one[2] from Gomez [6] only permits peaks in the interpolated amplitude spectrum to contribute to the chroma profile. The third one from Lee [8] converts a constant-Q power spectrum to an harmonic power spectrum with pitch evidence being solely coming from even harmonics.

The last reference extractor, described in [2], incorporates the notions of a subharmonic summation and a background spectrum, but it does not extract salient frequencies. Furthermore, it extracts the background spectrum in the log-frequency domain, it estimates the pitch evidence differently and it allows pitches to exhibit a harmonic relationship.

3.2 The dataset

The data set consists of 161 30s-excerpts originating from different songs, covering different tempos (between 40.2 and 188.0 bpm) and genres: dance/techno (14), jazz/blues (12), classical (30), new age/ambient (14), pop (38), rap/hip hop (7), R&B/reggae (5), rock (12) and world music (29). The fragments were formerly enriched with meter information (see [10]) and were now manually annotated with chords. A chord was defined as a composition of three or more chromas and chord segments were forced to start and end on tatum grid points. Chord segments can be separated from each other by intervals without a chord. The 2760 labeled chord segments cover 81.8% of the total signal duration. A large part of the uncovered time originates from the fact that in 15 songs, the human expert was unable to identify any chord segment. The dataset is divided in a development set of 23 systematically selected songs (songs 1, 8, 15, etc.) and an evaluation set consisting of the remaining 138 songs.

3.3 Experimental procedure

The experimental evaluation aims at quantifying the resemblance between the computed chroma profiles and the annotated chord profiles in the frames belonging to annotated chord segments. We argue that the better this resemblance is, the more chance there is that the computed chroma profiles can give rise to an accurate chord detection and classification.

The computed chroma profiles of the frames in a chord segment are averaged to a 12-dimensional chroma vector $\mathbf{V_c}$ per segment. Similarly, the annotated chord label is converted to a 12-dimensional target chroma vector $\mathbf{V_t}$ consisting of ones and zeros indicating whether the corresponding

[1] http://labrosa.ee.columbia.edu/matlab/chroma-ansyn/

[2] http://users.jyu.fi/∼lartillo/mirtoolbox/

chroma coincides with that of a note of the annotated chord. The cosine similarity between $\mathbf{V_c}$ and $\mathbf{V_t}$ is then defined as

$$S(\mathbf{V_c}, \mathbf{V_t}) = \frac{\langle \mathbf{V_c}, \mathbf{V_t} \rangle}{\|\mathbf{V_c}\| \, \|\mathbf{V_t}\|}$$

and is a number between 0 and 1. If the computed chroma vector is zero, the similarity is automatically set to zero as well. The mean cosine similarity across all chord segments in the evaluation set is considered as a quality measure for the considered chroma extractor.

Additionally to the cosine similarity, we use the mean reciprocal rank (MRR) [4]. Therefore, for each chord segment, the extractors are ranked according to their mean cosine similarity. Subsequently, the MRR of an extractor is computed as the mean of the reciprocal of its ranks across all the individual segments.

Since different extractors work with different frame sizes, there may be different phase shifts between annotations and extractor outputs. The evaluation program uses the best phase shift compensation for each algorithm.

3.4 Results

First we optimized the free parameters F_L, K and α of the proposed extractor on the development set. To that end we conducted a small grid search: $F_L = 80$, 100 or 120 Hz; $K = 5$ or 10 and $\alpha = 0.3$, 0.5 or 0.7. The mean cosine similarity changed only slightly between 0.756 and 0.801, and F_L turned out to be the most critical parameter, followed by K. For all grid points with an $F_L \geq 100$ Hz, the similarity was larger than 0.791. We finally selected $F_L = 100$ Hz, $K = 5$ and $\alpha = 0.5$.

In a second test we applied all chroma extractors on the evaluation set. The mean cosine similarities per extractor are shown on the left side of Figure 2, the MRRs on the right side. The proposed extractor achieves a mean similarity of

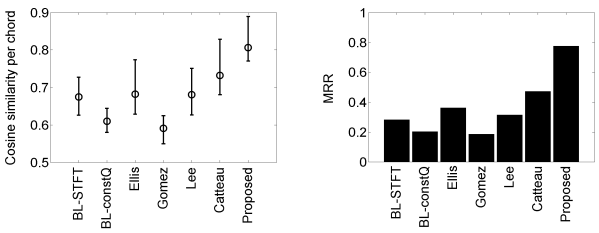

Figure 2: Left panel: the mean cosine similarities for the different chroma extractors. Also indicated is the interval between the 30-th and 70-th percentile. Right panel: the corresponding mean reciprocal ranks (MRR).

0.807 and a MRR of 0.775. The second best one reaches a mean similarity of 0.732 and an MRR of 0.473. A matched pairs signed rank Wilcoxon test on the cosine similarities shows that the proposed extractor outperforms the second best one with 0.06 at a significance level of 0.05.

4. CONCLUSION AND FUTURE WORK

We have proposed a new chroma extractor comprising a novel multi-pitch tracker. The latter puts specific restrictions on the eligible harmonic relations among pitches and it uses a psychophysically motivated measure of the harmonic tone importances while producing the pitch evidences.

We have experimentally verified on a sufficiently large (138 song excerpts) and diverse corpus that the new extractor produces chroma profiles that well resemble target chroma profiles which were retrieved from manual chord annotations. The mean cosine similarity (correlation) between the computed and the target chroma profiles is equal to 0.80 and significantly larger than that of six other algorithms.

The next step will be to investigate whether the seemingly improved quality of the chroma profiles can also be translated into an improved accuracy of existing chord detectors and labelers. This is not as trivial as it seems because these systems are either consciously or unconsciously optimized with respect to the inputs they are normally supplied with. Changing the inputs may require an adaptation of the system parameters and even of the system strategies.

5. ACKNOWLEDGEMENTS

This work was done in the context of the SEMA project, funded by 'Bijzonder Onderzoeksfonds Universiteit Gent' under contract GOA-1250604.

6. REFERENCES

[1] J. Brown. Calculation of a constant q spectral transform. *Journal of the Acoustical Society of America*, 89:425–434, 1991.

[2] B. Catteau, J.-P. Martens, and M. Leman. A probabilistic framework for audio-based tonal key and chord recognition. In *Advances in Data Analysis*, pages 637–644, 2007.

[3] T. De Mulder, J.-P. Martens, M. Lesaffre, M. Leman, B. De Baets, and H. De Meyer. Recent improvements of an auditory model based front-end for the transcription of vocal queries. In *Proceedings ICASSP*, pages 257–260, 2004.

[4] D. Ellis and G. Poliner. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *Proceedings ICASSP*, pages 1429– 1432, 2007.

[5] T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. In *Proceedings International Computer Music Association*, pages 464–467, 1999.

[6] E. Gómez. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, 18:294–304, 2006.

[7] D. Hermes. Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America*, 83:257–264, 1988.

[8] K. Lee. Automatic chord recognition from audio using enhanced pitch class profile. In *Proceedings International Computer Music Conference*, pages 306–313, 2006.

[9] S. Pauws. Musical key extraction from audio. In *Proceedings 5th International Symposium on Music Information Retrieval*, pages 96–99, 2004.

[10] M. Varewyck and J.-P. Martens. Assessment of state-of-the-art meter analysis systems with an extended meter description model. In *Proceedings 8th International Symposium on Music Information Retrieval*, pages 311–314, 2007.