TIME-FREQUENCY REASSIGNED FEATURES FOR AUTOMATIC CHORD RECOGNITION

Maksim Khadkevich, Maurizio Omologo

Fondazione Bruno Kessler-irst, Via Sommarive, 18 - Povo - 38050 Trento, Italy

ABSTRACT

This paper addresses feature extraction for automatic chord recognition systems. Most chord recognition systems use chroma features as a front-end and some kind of classifier (HMM, SVM or template matching). The vast majority of feature extraction approaches are based on mapping frequency bins from spectrum or constant-Q spectrum to chroma bins. In this work a set of new chroma features that are based on the time-frequency reassignment (TFR) technique is investigated. The proposed feature set was evaluated on the commonly used Beatles dataset and proved to be efficient for the chord recognition task, outperforming standard chroma.

Index Terms— chord recognition, chroma, time-frequency reassignment

1. INTRODUCTION

Automatic chord recognition has always been of a great interest to Music Information Retrieval (MIR) community. Chord sequence can serve as a robust mid-level representation for a variety of MIR tasks. During the past few decades several approaches were developed. The majority of the proposed approaches can be decomposed into the following structural parts: feature extraction, pre-filtering, classification and post-filtering.

Chroma feature that was introduced by Fujishima [1] has proven to be an effective tool for capturing harmonic structure. The most common way of calculating chromagram is to transform the signal from the time domain to the frequency domain with the help of short-time Fourier transform (STFT) or constant-Q transform and subsequent energy mapping of spectral bins to chroma bins [2, 3].

When performing chroma extraction, the signal in a given analysis frame is assumed to be stationary and it is also assumed that no note transitions occur inside it. Transients and noise may cause energy assignment to some frequencies that do not occur in the signal. At the same time frame size should be long enough to provide reasonable frequency resolution. A trade-off between frequency resolution and stationarity should be made for a particular task. The most widely used frame sizes for capturing spectral content to form chroma vectors are 192ms - 360ms. As a rule, to provide smoothed feature sequence a high overlap ratio (50% – 90%) with subsequent median filtering or averaging is applied. However, using such window lengths introduces inaccuracies with rapidly changing notes. On the other hand, short window lengths does not provide reasonable frequency resolution.

In this paper we introduce two alternative chroma features and provide their comparative characteristics. The structure of the paper is as follows: in section 2 the formulation of the time-frequency reassignment technique is introduced. Sections 3 and 4 describe the chord recognition system and evaluation metrics. Experimental results and conclusions are then given in sections 5 and 6 respectively.

2. REASSIGNED SPECTRUM FOR CHROMAGRAM CALCULATION

In the past few years a lot of different techniques for accurate and relevant feature extraction in automatic chord recognition have been proposed. In this section we examine the performance of the chromagram that is based on the reassigned spectrum.

Feature extraction process is aimed at transforming a given waveform into a representation that captures desirable properties of an analyzed signal. A great deal of acoustic features is derived from some kind of time-frequency representations, which can be obtained by mapping audio signal from one-dimensional time domain into two-dimensional domain of time and frequency.

Spectrogram is one of the most widely spread time-frequency representations that has been successfully used in a variety of applications, where spectral energy distribution changes over time. However, spectrogram possesses several drawbacks, such as unavoidability of a compromise between time and frequency resolutions.

Time-frequency reassignment technique was initially proposed by Kodera et al. [4]. The main idea behind TFR technique is to remap spectral energy of each spectrogram cell into another cell that is the closest to the true region of support of the analyzed signal. As a result, "blurred" spectral representation becomes "sharper" that allows one to derive spectral features from reassigned spectrogram with much higher time and frequency resolution. Some papers have already investigated the usage of reassigned spectrogram in different tasks, such as sinusoidal synthesis [5], cover song identification [6] and many others.

Now some mathematical foundations for the TFR technique are provided. Let $x(n)$ be a discrete signal in the time domain sampled at a sampling frequency F_s. At a given time instant t, STFT is performed on the signal weighted by a window function $w(n)$ as follows:

$$X(t,k) = \sum_{n=0}^{M-1} w(n)x(n+t)e^{-2\pi jnk/M}, \quad (1)$$

where k and M denote a bin number and the window size respectively. Spectrogram is derived from (1) as shown in (2).

$$S(t,k) = |X(t,k)|^2 \quad (2)$$

The majority of chromagram extraction techniques uses this representation for mapping spectral energies to chroma bins, ignoring phase information as shown in (3).

$$n(f_k) = 12\log_2\left(\frac{f_k}{f_{ref}}\right) + 69, n \in \Re^+, \quad (3)$$

where f_{ref} denotes the reference frequency of "A4" tone, while f_k and n are the frequencies of Fourier transform and the semitone bin scale index, respectively.

On the other hand, the result of STFT $X(t,k)$ can be presented in the following form:

$$X(t,k) = M(t,k)e^{j\phi(t,k)}, \quad (4)$$

where $M(t,k)$ is the magnitude, and $\phi(t,k)$ the spectral phase of $X(t,k)$. As was shown in [7], reassigned time-frequency coordinates $(\hat{t},\hat{\omega})$ can be calculated as

$$\hat{t}(t,\omega) = -\frac{\partial\phi(t,\omega)}{\partial\omega} \tag{5}$$

$$\hat{\omega}(t,\omega) = \omega + \frac{\partial\phi(t,\omega)}{\partial t} \tag{6}$$

Efficient computation of $\hat{t}(t,\omega)$ and $\hat{\omega}(t,\omega)$ in the discrete-time domain was proposed by Auger and Flandrin [8] and takes the following form:

$$\hat{t}(t,\omega) = t - \Re\left\{\frac{X_{\mathcal{T}w}(t,\omega)\cdot X^*(t,\omega)}{|X(t,\omega)|^2}\right\} \tag{7}$$

$$\hat{\omega}(t,\omega) = \omega + \Im\left\{\frac{X_{\mathcal{D}w}(t,\omega)\cdot X^*(t,\omega)}{|X(t,\omega)|^2}\right\} \tag{8}$$

where $X_{\mathcal{D}w}$ is the STFT of the signal weighted by a frequency-weighed window function, $X_{\mathcal{T}w}$ is the STFT of the signal weighted by a time-weighed window function ([7]). Reallocating spectral energy from spectrogram coordinate (t,w) to $(\hat{t},\hat{\omega})$ concludes the reassignment operation. As a result more precise estimates of spectral energy distribution are obtained. However, reassigned spectrogram can be noisy. A random energy can be located in points where there are no obvious harmonic or impulsive components. The principle of the reassignment technique is to reallocate energy from the geometrical center of the analysis window to the "center of gravity" of the spectral component this energy belongs to. Meanwhile, in some spectral regions, where there are no dominant components, large energy reassignment both in time and frequency can be observed. In order to obtain a better spectral representation and to refine the spectrogram keeping the energy of harmonic components and deemphasizing that of noisy and impulsive components, the following condition should be met ([9])

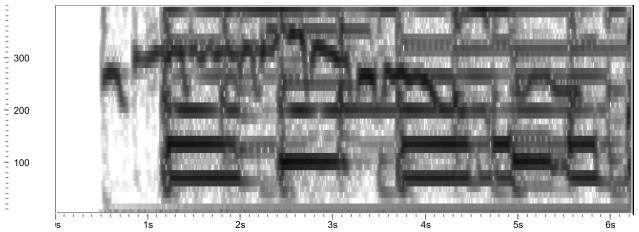
$$\left|\frac{\partial\phi^2(t,\omega)}{\partial t\partial\omega} + 1\right| < A \tag{9}$$

where A is the tolerance factor, which defines the maximum deviation of the acceptable spectral component from a pure sinusoid. The optimal value of A depends on a particular task and can be empirically determined. Fullop and Fitz reported in [10] that 0.2 is often a reasonable threshold for speech signals. Efficient computation of $\frac{\partial\phi^2(t,\omega)}{\partial t\partial\omega}$ is given in [7] and can be expressed as follows

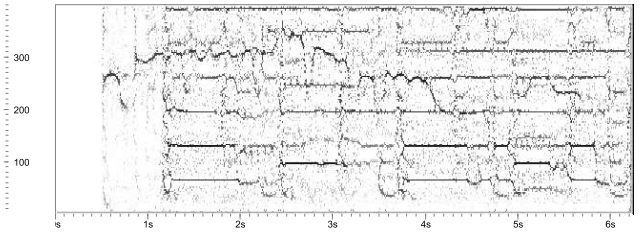
$$\begin{aligned}\frac{\partial\phi^2(t,\omega)}{\partial t\partial\omega} = &\Re\left\{\frac{X_{\mathcal{T}\mathcal{D}w}(t,\omega)X^*(t,\omega)}{|X(t,\omega)|^2}\right\} \\ &-\Re\left\{\frac{X_{\mathcal{T}w}(t,\omega)X_{\mathcal{D}w}(t,\omega)}{X^2(t,\omega)}\right\}\end{aligned} \tag{10}$$

where $X_{\mathcal{T}\mathcal{D}w}(t,\omega)$ is the STFT of the signal weighted by time-frequency-weighed window function ([7]).

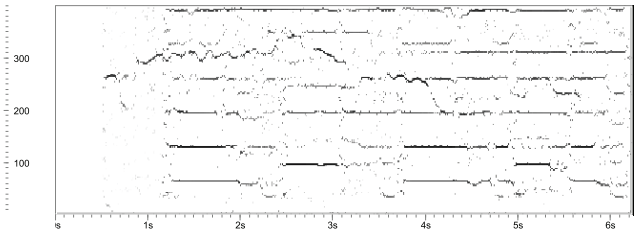
Comparison of spectrogram, reassigned spectrogram and "refined" reassigned spectrogram for an excerpt from "Girl", the Beatles is provided in Figure 1. All spectrograms are computed using Hanning window of 192 ms with 90% overlapping.



(a) Spectrogram



(b) Reassigned spectrogram



(c) Harmonic reassigned spectrogram with tolerance factor set to 0.4

Fig. 1: Time-Frequency representation of an excerpt from "Girl", the Beatles. All spectrograms are computed using Hanning window of 192 ms with 90% overlapping.

3. CHORD RECOGNITION SYSTEM

3.1. Front-End processing

In the chord recognition system under study, before extracting features the tuning procedure described in [11] is applied in order to find the mis-tuning rate and set the reference frequency f_{ref} for the "A4" tone. The necessity of tuning appears when audio was recorded from instruments that were not properly tuned in terms of semitone scale.

The feature extraction process starts with downsampling the signal to 11025 Hz and converting it to the frequency domain by a STFT applying Hanning window of N samples with 90% overlap ratio. Direct folding of spectral energy to chroma bins using formula (3) produces standard chroma (STD) feature. Applying time-frequency reassignment technique before chroma wrapping results in a "reassigned" chroma (RC). "Harmonic reassigned" chroma feature (HRC) calculation is based on the reassigned spectrum when fulfilling the condition introduced in (9). In the last stage semitone bins are mapped to pitch classes, which results in the sequence of 12-dimensional chroma vectors:

$$b(n) = \mathrm{mod}(n,12) \tag{11}$$

where $b(n)$ and n denote semitone bin indices in wrapped and unwrapped chroma respectively.

3.2. Statistical classifier

This section briefly introduces a statistical classifier used for evaluations. In the following, usage of hidden Markov models is pretty much similar to what was described in [11] and [12]. However, in this paper we investigate the usage of multi-stream observation layer, where two observation vector streams model lower (bass) and higher harmonic content. Similar technique was used in [13] and [14].

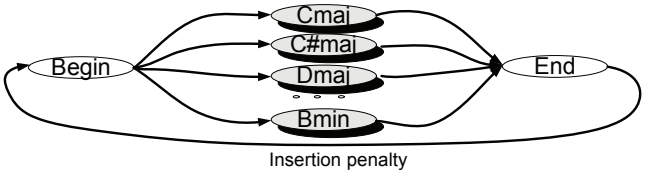


Fig. 2: Connection scheme of trained models for decoding.

In [14] a dynamic Bayesian network is configured to contain bass and treble observable layers.

In a multi-stream HMM observation probability distribution $b_j(o_t)$ representing the probability to emit observation symbol o_t at time instant t is defined as follows:

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_{js}} c_{jsm} \mathcal{N}\left(o_{st}; \mu_{jsm}, \Sigma_{jsm}\right) \right]^{\gamma_s}, \qquad (12)$$

where M_{js} denotes the number of mixture components in state j for stream s, c_{jsm} is the weight of the m-th component and $\mathcal{N}\left(o_{st}; \mu_{jsm}, \Sigma_{jsm}\right)$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ. Continuous density models are here used in which each observation probability distribution is represented by a mixture of multivariate Gaussians. Each Gaussian component $\mathcal{N}\left(o_{st}; \mu_{jsm}, \sum_{jsm}\right)$ can be expressed as

$$\mathcal{N}\left(o; \mu, \Sigma\right) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(o-\mu)'\Sigma^{-1}(o-\mu)\right) \qquad (13)$$

where n is the dimensionality of observation o. The term γ_s is a stream weight. Varying this parameter allows one to emphasize or deemphasize the contribution of a particular stream.

For each chord type a separate HMM is created. Each model consists of 1 – 3 emitting hidden states. Observation probability distributions are learned from data in the training stage. Feature vector components are assumed to be uncorrelated with one another, so the covariance matrix has a diagonal form.

Trained HMMs are connected as shown in figure 2. Such parameter as insertion penalty is introduced, which influences the transition probability between chords. Varying insertion penalty allows for obtaining labels with different degrees of fragmentation, as typically done in speech recognition tasks. As was shown in [12], insertion penalty (or self-transition probability in [15]) can have a significant impact on the overall performance.

In the experimental part two different HMM configurations are evaluated – baseline and multi-stream one. The former configuration includes one observation stream, where emitted symbols are chroma vectors. In the latter case an additional observation stream is added with bass chroma vectors served as observed sequence.

Songs from the training set are segmented according to the ground-truth labels so that each segment represents one chord. Chromagrams extracted from these segments are used for training, which is based on the application of the Baum-Welch algorithm. The recognition process is performed by running the Viterbi decoder.

4. EVALUATION METRICS

Three different estimates are used to evaluate the quality of a chroma vector. The first two that are *ratio (R)* and *cosine measure (CM)* are computed as proposed in [16]; the third one, which is *recognition rate (RR)* is explained below.

Let $c(n)$ be an unwrapped chroma vector extracted from a chord sample that was generated from a set of notes s. The *R* estimate is the ratio of the power in the expected semitone bins, over the total power. The expected semitone bins include the bins of the fundamentals and 3 partials for every note from set s.

For *CM* estimate a chroma template $y(n)$ is built so that its values are set to 1 in the chroma bins that correspond to the fundamentals and to 0.33 in the chroma bins that correspond to the first 3 partials. The CM estimate is then computed as $CM = \frac{\langle y \cdot c \rangle}{\|y\| \|c\|}$, where $\langle \cdot \rangle$ is the inner product and $\|\cdot\|$ is the L^2 norm.

RR measure is obtained by running the chord recognition task. RR is computed as the total duration of correctly classified chords divided by the total duration of the test material.

5. EXPERIMENTAL RESULTS

5.1. Chroma quality evaluation

For the first set of evaluations we used the University of Iowa [1] database of individual note recordings. The samples of the constituent notes for a given chord were mixed together, producing a waveform of 2 seconds duration. For the RR measure half of the generated material was used as training set, the other half was used for the test purposes. Chroma features were extracted with 192 ms window lengths, 0.9 overlap, Hanning windowing.

The evaluation results for three different chroma features are given in Figure 3.

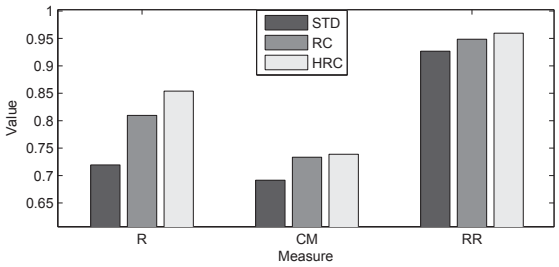


Fig. 3: Chroma quality estimates

In all the cases *HRC* and *RC* significantly outperform *STD* feature. The *ratio* estimate values proved the ability of the *HRC* to deemphasize noise and impulsive components, which frequently occur during the note onsets.

5.2. Chord recognition system evaluation

In order to show the advantages of the proposed feature set for the chord recognition task, a 3-fold cross validation was accomplished on the ubiquitous Beatles data set. All the songs were randomly divided into 3 folds.

Figure 4 depicts recognition rates for different number of Gaussians for STD, RC and HRC features. For each configuration the best insertion penalty is assumed. The results obtained indicate the optimal number of Gaussians for the given training/test set equal to 2048, since higher values do not bring significant improvement while increasing computational load drastically.

For the HRC feature an impact of tolerance factor A introduced in (9) on the recognition rate was investigated. The corresponding graph is shown in Figure 5. The optimal value of A for the chord

[1] http://theremin.music.uiowa.edu/MIS.html

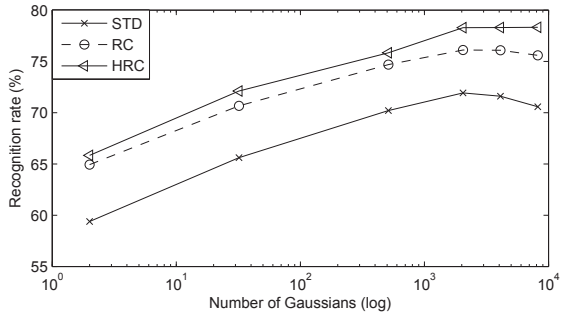


Fig. 4: Recognition rate as a function of the number of Gaussians

recognition task turned out to be 0.4 with the recognition rate of 78.28%, although small deviations on this parameter seem to have a minor impact in terms of loss of performance.

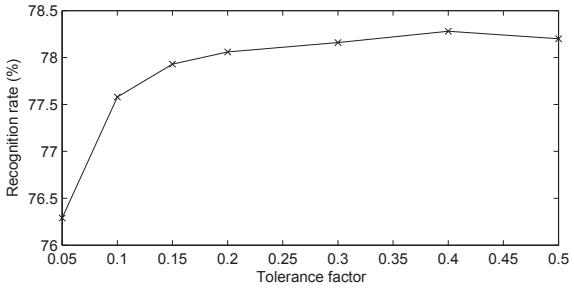


Fig. 5: Recognition rate for HRC as a function of the tolerance factor

The next set of experiments involved the technique of splitting frequency range used for chroma calculation into 2 parts: chroma and bass chroma. For computing bass-chroma, frequencies that correspond to the MIDI range between 24 (32.7 Hz) and 54 (185 Hz) notes are used. For chroma feature extraction frequency interval between 54 (185 Hz) and 96 (2093 Hz) MIDI notes is employed.

The summary on the recognition rates for different feature set configurations is given in Table 1. The experimental results showed an evident advantage of *HRC* and *RC* features over standard chroma. Having an additional observation stream that models bass content proved to be effective. The best system configuration based on the *HRC* feature and 2-stream observation layer in HMM achieved the highest result of 80.67%.

	STD	RC	HRC(A=0.4)
nobass	71.93	76.89	78.28
bass	74.29	80.19	80.67

Table 1: Recognition rates (%) for different feature set configurations

6. CONCLUSION

In this paper we investigated influence of different parameters on the performance of chord recognition system. Different frond-end configurations have been proposed. More sophisticated chromagram that is based on the time-frequency reassigned spectrogram proved to outperform the traditional one. Tolerance factor with the *HRC* features has been addressed and an optimal choice has been individuated. As for the classification component, a multi-stream HMM structure has been proposed, where the two observable layers represent harmonic content of the two frequency regions.

7. REFERENCES

[1] Takuya Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," in *Proceedings of the International Computer Music Conference*, Beijing, 1999.

[2] A. Sheh and D. P. Ellis, "Chord segmentation and recognition using em-trained hidden markov models," in *Proc. 4th International Conference on Music Information Retrieval*, 2003.

[3] H. Papadopoulos and G. Peeters, "Simultaneous estimation of chord progression and downbeats from an audio file," in *Proc. ICASSP*, 2008.

[4] K. Kodera, R. Gendrin, and C. Villedary, "Analysis of time-varying signals with small bt values," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 64 – 76, feb. 1978.

[5] Toshihiko Abe and Masaaki Honda, "Sinusoidal model based on instantaneous frequency attractors," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1292–1300, 2006.

[6] D. P. W. Ellis and G. E. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, April 2007, vol. 4.

[7] Kelly R. Fitz and Sean A. Fulop, "A unified theory of time-frequency reassignment," *CoRR*, vol. abs/0903.3080, 2009.

[8] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time- scale representations by the reassignment method," *IEEE Trans. Speech Audio Process*, vol. 5, no. 43, pp. 1068–1089, 1995.

[9] Douglas J. Nelson, "Instantaneous higher order phase derivatives," *Digital Signal Processing*, vol. 12, no. 2-3, pp. 416 – 428, 2002.

[10] S. A. Fulop and K. Fitz, "Separation of components from impulses in reassigned spectrograms," in *J. Acoust. Soc. Am. 121*, 2007, pp. 1510–1518.

[11] M. Khadkevich and M. Omologo, "Phase-change based tuning for automatic chord recognition," in *Proceedings of DAFX*, Como, Italy, 2009.

[12] M. Khadkevich and M. Omologo, "Use of hidden markov models and factored language models for automatic chord recognition," in *Proceedings of the 2009 ISMIR Conference*, Kobe, Japan, 2009.

[13] D. P. W. Ellis and Adrian Weller, "The 2010 labrosa chord recognition system," `http://www.ee.columbia.edu/~dpwe/pubs/Ellis10-chords`, 2010.

[14] Matthias Mauch and Simon Dixon, "Simultaneous estimation of chords and musical context from audio," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[15] T. Cho, R. J. Weiss, and J. P. Bello, "Exploring Common Variations in State of the Art Chord Recognition Systems," in *Proc. Sound and Music Computing Conference (SMC)*, Barcelona, Spain, jul 2010, pp. 1–8.

[16] C. Joder, S. Essid, and G. Richard, "A comparative study of tonal acoustic features for a symbolic level music-to-score alignment," mar. 2010, pp. 409 –412.