THE TONALNESS SPECTRUM: FEATURE-BASED ESTIMATION OF TONAL COMPONENTS

Sebastian Kraft
Helmut-Schmidt-University
Department of Signal Processing
Hamburg, Germany
skraft@hsu-hh.de

Alexander Lerch
zplane.development
Berlin, Germany
lerch@zplane.de

Udo Zölzer
Helmut-Schmidt-University
Department of Signal Processing
Hamburg, Germany

ABSTRACT

The tonalness spectrum shows the likelihood of a spectral bin being part of a tonal or non-tonal component. It is a non-binary measure based on a set of established spectral features. An easily extensible framework for the computation, selection, and combination of features is introduced. The results are evaluated and compared in two ways. First with a data set of synthetically generated signals but also with real music signals in the context of a typical MIR application.

1. INTRODUCTION

A multitude of algorithms in the area of audio signal processing focus only on sinusoidal components of a signal because noisy or non-sinusoidal components may either have a negative impact on the algorithm's performance or they need to be processed separately. Examples of such algorithms are

- analysis/synthesis systems based on sinusoidal signal models such as phase vocoders and audio codecs, for which the audio quality directly depends on correct sinusoidal identification,
- systems for audio restoration, and
- audio analysis systems, especially pitch-based systems for Music Information Retrieval (MIR), such as key detection, chord detection, music transcription and source separation, which all may benefit by suppressing noisy components.

When considering music signals it is a valid assumption that sinusoidal components are evoked by tones and thus we will refer to them as tonal components. Since the term "tonality" is commonly used to describe a harmonic or key context we use the term *tonalness* (as an antonym of noisiness) for the amount of sinusoidality. The tonalness is a likelihood or a continuous score as opposed to the commonly used binary classification of components. It is the authors' believe that hard thresholding and the resultant reduction of information should in general be avoided in the early processing stages.

The detection of tonalness or the identification of sinusoidal components is a common pre-processing step which might have a major impact on a system's overall performance. Nevertheless, the systematic evaluation of this pre-processing step is frequently missing in most publications. In those publications that deal with the evaluation of tonalness measures, the evaluation is mostly done with synthetic signals, raising the question if these results also apply to real-world signals and can be assumed to be application-independent. Therefore, in this paper we will evaluate with synthetic signals but set the focus on a real application and the processing of music recordings.

The paper describes a formal way to develop, combine, select and evaluate spectral features for the detection of tonalness in a spectrum. After a short overview of the related work in the following section we will define a generic feature framework in Sect. 3 and describe an exemplary set of features in Sect. 4. The evaluation is performed with synthetic test data as well as with a key detection algorithm — a typical MIR task — on real-world data sets in Sect. 5.

2. RELATED WORK

As the variety of applications benefiting from a sinusoidal detection suggest, there have been numerous publications in this area, of which only a subset can be presented in this paper. Many of the features in Sect. 4 are partly based on such established methods.

Charpentier detected harmonic components based on the phase spectrum. He evaluated the difference between the reassigned frequency and the bin frequency and also expected the neighboring bins of a peak to have the same phase as the peak itself [1]. Roebel et al. used a similar phase based feature but they also employed the peak's energy location according to its group delay as well as the bandwidth of a spectral peak for a classification into sinusoidal vs. non-sinusoidal peaks [2].

Peeters and Rodet, as well as later Lagrange, presented an amplitude-based measure computing the correlation between the magnitude spectrum and the shifted spectrum of the employed window function together with a phase-derived measure that com-

pares the frequency of a peak with its reassigned (instantaneous) frequency [3, 4].

A more simple amplitude based measure — which can be found in nearly every publication in this area — is the search for local maxima in the magnitude spectrum. For example, it has been used in the context of speech separation by Parsons [5]. He also used the peak's symmetry, its proximity to the next peak as well as the continuity of the frequency bin's phase for the detection of "peak overlaps". Terhardt extended the concept of local maximum by taking into account more distant bins, more specifically bins with a distance of 2 and 3, to be a certain level lower than the maximum itself [6].

Serra proposed to use a measure of "peakiness" of local maxima by comparing the bin magnitude with the surrounding local minima; he also defined a frequency and magnitude range for detecting peaks [7].

All of the publications listed above make a binary decision whether a spectral bin is considered to be tonal or not. Kulesza and Czyzewski presented an algorithm that tries to estimate the likelihood of a bin being tonal [8] and referred to this as a scoring classifier. This non-binary decision makes this algorithm probably most similar to the one proposed here. Their approach combines several features and uses a combination of heuristics and both binary and non-binary features to compute the resulting likelihood.

3. FRAMEWORK

The input samples x are split into frames of length N_W and are weighted by a window function w. The windowed signal is zero padded to a length $N_\mathrm{FFT} \geq N_\mathrm{W}$ and the DFT yields the spectra $X(k,n)$ with frequencies k and frame indices n. The hop size between two frames is N_H.

A set of spectral features $\mathbb{V} = \{v_1, v_2, \ldots, v_\mathrm{V}\}$ is extracted from the spectrum. The design of the features is based on the following assumptions for the input signal:

- it is a time-varying mixture of tonal and non-tonal components
- it has an undefined number of voices (polyphony)
- the spectral envelope of both tonal and non-tonal components is unknown
- it is stationary for at least a minimum length of time
- its tonal components are deterministic, i.e. their phase will not change erratically between the points of observation

Each feature by itself should be simple to compute as well as simple to understand and should focus on one individual property or aspect of a tonal component.

Figure 1 gives an overview of the feature computation and combination process which shows some similarity to a simplified Radial Basis Function Network [9]. First, each feature v_i is the input of an exponential function $\varphi(\cdot)$. Its output

$$\mathfrak{t}_i(k,n) = \varphi\left(v_i(k,n)\right) = \exp\left(-\epsilon_i \cdot v_i(k,n)^2\right) \qquad (1)$$

will be referred to as the *specific tonal score* $\mathfrak{t}_i(k,n) \in [0,\ldots,1]$. This score can be interpreted as a measure of likelihood of bin k in frame n representing a tonal component with respect to feature i. The normalization constant ϵ_i will be explained later in Sec. 4.2.

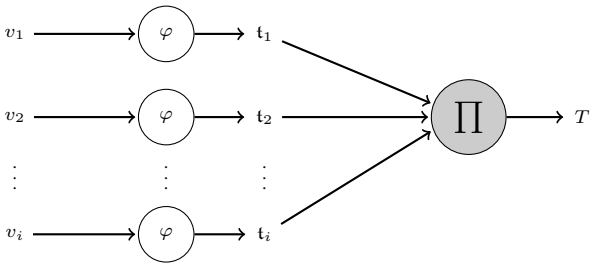


Figure 1: *Processing of individual feature outputs v_i by exponential function φ finally leading to the combined tonalness T.*

Finally, these specific tonal scores are combined to yield the overall tonalness

$$T(k,n) = \left(\prod_{i=1}^{V} \mathfrak{t}_i(k,n)\right)^{\frac{1}{\eta}}, \qquad \eta \in [1,\ldots,V]. \qquad (2)$$

The exponent η can be chosen in the range $[1,\ldots,V]$ to continuously adjust the mean between a simple product and a geometric mean, respectively.

4. FEATURE SET

The feature set described below mainly comprises various established features from several publications and does not claim to be exhaustive. It has an exemplary character and it is easy to extend the feature set with additional features or to modify the presented ones.

4.1. Detailed Feature Description

Since the feature output $v_i(k,n)$ is in turn the input of the exponential function in Eq. (1), the direct feature output is zero for tonal and maximum for non-tonal bins. Example plots of some of the resulting tonal scores are shown in Fig. 2.

4.1.1. Amplitude Continuity

The amplitude of a tonal bin is expected to be constant for several time frames. Thus, the amplitude change at a bin is a measure of tonalness and the feature

$$v_\mathrm{ACT}(k,n) = \frac{\big|\,|X(k,n)| - |X(k,n-1)|\,\big|}{|X(k,n-1)|} \qquad (3)$$

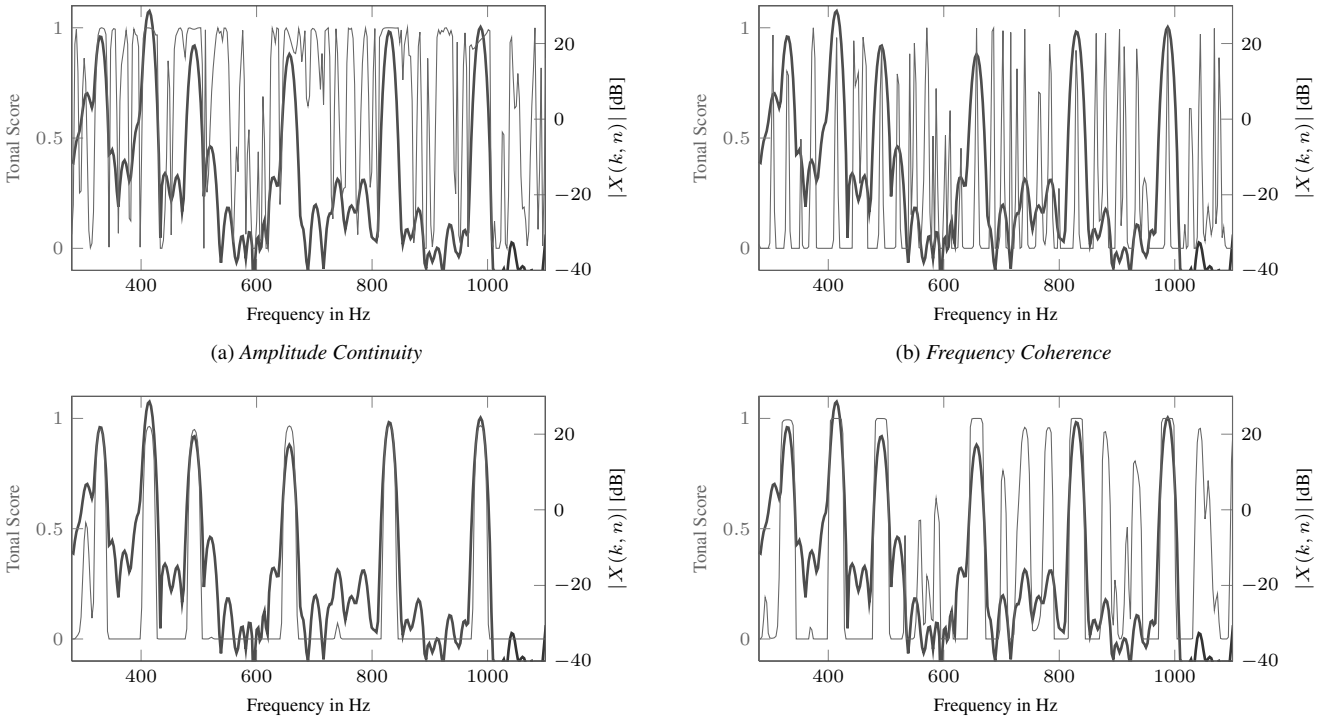
is defined as the relative bin amplitude difference between two neighbouring magnitude spectra.

4.1.2. Frequency Continuity

A similar constraint can be applied to the change of the instantaneous frequency f_I of a bin over time:

$$v_\mathrm{FCT}(k,n) = \left|f_\mathrm{I}(k,n) - f_\mathrm{I}(k,n-1)\right|. \qquad (4)$$

We choose the frequency reassignment operator introduced by Auger and Flandrin [10] for our implementation because its accuracy is



(a) *Amplitude Continuity*

(b) *Frequency Coherence*

(c) *Amplitude Threshold*

(d) *Peakiness*

Figure 2: *Sample plots of some specific tonal scores when processing a recorded piano chord. The magnitude spectra are plotted in black and the tonal scores are overlaid with a light grey line.*

hop size independent. The reassigned frequency f_{I} is the bin frequency $2\pi k$ minus a frequency offset Δ_ω:

$$\Delta_\omega = \Im\left\{\frac{X_{\mathrm{D}}(k,n)X^*(k,n)}{|X(k,n)|)}\right\} \quad (5)$$

$$f_{\mathrm{I}}(k,n) = 2\pi k - \Delta_\omega \cdot N_{\mathrm{W}}. \quad (6)$$

$X_{\mathrm{D}}(k,n)$ is the spectrum of the current time frame weighted by the time derivative of the window function and $X^*(k,n)$ is the complex conjugate of $X(k,n)$.

Note that both the *Amplitude Continuity* and the *Frequency Continuity* will fail to work reliably in the case of strongly modulated input signals. A typical example for such a modulation is *vibrato*. In order to improve results with such signals, tracking of the sinusoidal trajectories would be necessary.

4.1.3. Frequency Deviation

Due to spectral leakage, bins close to a tonal component should have the same phase and thus the same instantaneous frequency as the tonal bin itself. Therefore, a feature

$$v_{\mathrm{FD}}(k,n) = \Big|f_{\mathrm{I}}(k,n) - f_{\mathrm{I}}(k-\gamma,n) + f_{\mathrm{I}}(k,n) - f_{\mathrm{I}}(k+\gamma,n)\Big| \quad (7)$$

is defined measuring the difference of instantaneous frequencies between a center and two neighbouring bins. The distance to the surrounding bins is given by the factor $\gamma = {}^{N_{\mathrm{FFT}}}/_{N_{\mathrm{W}}}$, taking into account the degree of spectral interpolation.

4.1.4. Frequency Coherence

Another way to utilize the reassigned frequency is to derive a tonalness criterion directly from the frequency offset Δ_ω between the bin frequency and the instantaneous frequency:

$$v_{\mathrm{FC}}(k,n) = |\Delta_\omega \cdot N_{\mathrm{W}}|\,. \quad (8)$$

4.1.5. Amplitude Threshold

Since the tonal components are expected to be salient and to have more energy than noisy parts, a magnitude threshold can be applied to increase the likelihood of components above the threshold and decrease the likelihood of other components accordingly. This is achieved by the ratio

$$v_{\mathrm{AT}}(k,n) = \frac{r_{\mathrm{TH}}(k,n)}{|X(k,n)|} \quad (9)$$

in which the smoothed magnitude spectrum

$$r_{\mathrm{TH}}(k,n) = \alpha \cdot r_{\mathrm{TH}}(k-1,n) + (1-\alpha)\cdot|X(k,n)|$$

serves as an adaptive threshold and is computed with a single pole low pass filter. The filter is applied over the frequency in both the forward and the backward direction to compensate for group delay. The filter coefficient α has been adjusted empirically.

4.1.6. Peakiness

Declaring the local maxima in the magnitude spectrum to be candidates for being tonal is a rather self-evident step in spectral analysis. A non-binary implementation of a local maximum feature is a measure of peakiness

$$v_{\mathrm{PK}}(k,n) = \frac{|X\,(k-m,n)\,| + |X\,(k+m,n)\,|}{|X(k,n)|}\,, \qquad (10)$$

which is the ratio between the sum of two surrounding bins and the center. The distance m should roughly correspond to the spectral main lobe width of the window function.

4.1.7. Extended Peakiness

An extended peakiness measure also includes more distant bins:

$$v_{\mathrm{EPK}}(k,n) = \frac{\sum_{s=1}^{3} |X(k-2\gamma s,n)| + |X\,(k+2\gamma s,n)\,|}{|X(k,n)|}. \qquad (11)$$

It is the relation of the sum of the magnitudes of the three surrounding bins to the magnitude of the center bin. The distances are multiples of $2\gamma = {}^{2N_{\mathrm{FFT}}}/_{N_{\mathrm{W}}}$ to make this feature independent of the degree of spectral interpolation.

4.1.8. Time Window Center of Gravity

Similar to the frequency reassignment operator from Eq. (5), Auger and Flandrin also defined a time reassignment operator [10]

$$\Delta_t \quad = \quad \Re\left\{\frac{X_{\mathrm{T}}(k,n)X^*(k,n)}{|X(k,n)|}\right\}, \qquad (12)$$

which gives the time offset of a certain frequency relative to the center of the current time frame. In this case $X_{\mathrm{T}}(k,n)$ is the spectrum retrieved from the time samples weighted by a time weighted window.

The actual feature

$$v_{\mathrm{TCG}}(k,n) \quad = \quad \left|\Delta_t \cdot \frac{1}{N_{\mathrm{W}}}\right| \qquad (13)$$

is the time deviation weighted by the window size. This feature can also be interpreted as a transient detection which categorizes transient events as not being sinusoidal and has already been used as in this area, e.g. by Röbel [11].

4.1.9. Random Feature

In order to have a base line for the evaluation and to verify that the presented features will result in a gain of information the random feature v_{RND} is introduced. The feature output is Rayleigh distributed and normalized in the same way as the other features. We expect the resulting random tonal score to perform worse than all other tonal scores in the following evaluation.

4.2. Feature Normalization

The individual features have different scaling and different cumulative distribution functions. This requires normalization of all features with the normalization constant ϵ_i, compare Eq. (1), to avoid favouring specific features when combining them.

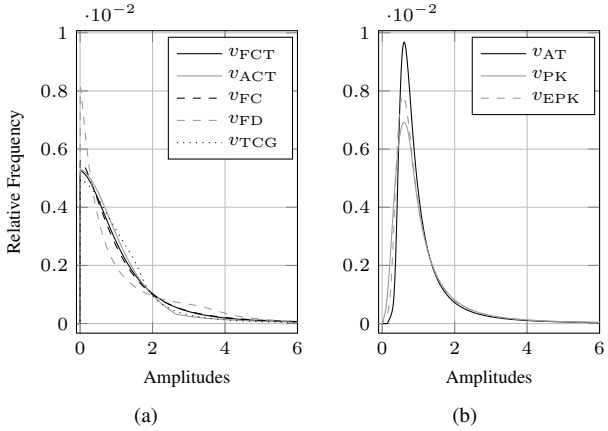


Figure 3: *Relative frequency distributions after normalization.*

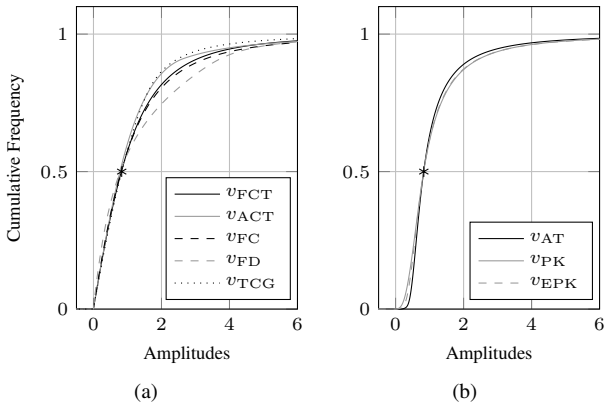


Figure 4: *Cumulative distribution functions after normalization.*

We choose to normalize the extracted features in a way that the median of the feature output yields a specific tonal score of 0.5. To compute ϵ_i, the individual feature outputs are analyzed while processing a database of music files (the ALERC data set is later described in more detail in Sect. 5.2.1). For each feature the median value $m_{v_i}(n)$ is calculated per frame and then the mean $\overline{m_{v_i}}$ over all frames and files is taken. Rearranging Eq. (1) and setting the target tonal score to 0.5 gives

$$\epsilon_i = \frac{\sqrt{\log\left({}^{1}/_{0.5}\right)}}{\overline{m_{v_i}}} \qquad (14)$$

to normalize the feature output. After normalization all features have similar relative frequency distributions as shown in Fig. 3. Note that the cumulative distribution functions all intersect the same $0.83/0.5$ point in Fig. 4.

5. EVALUATION

The evaluation of algorithms for the detection of tonal components is problematic due to the difficulties of finding a proper ground truth. It is not possible to annotate a large data set of real-world signals with appropriate labels when they are non-trivial complex mixtures of multiple sources. Therefore, practically all approaches

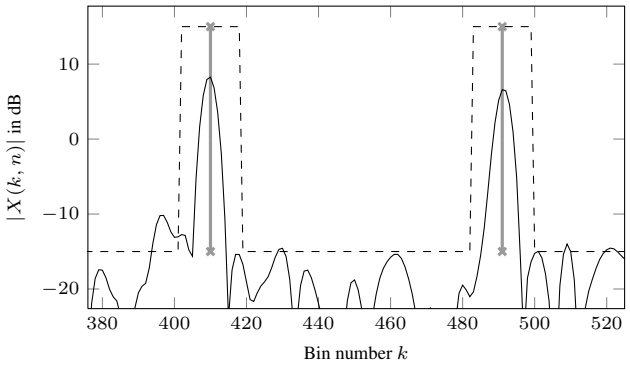


Figure 5: *Magnitude spectrum of the synthetic test signal with additive white noise and the overlayed mask (dashed line). The grey bars mark the expected sinusoidal peak centers.*

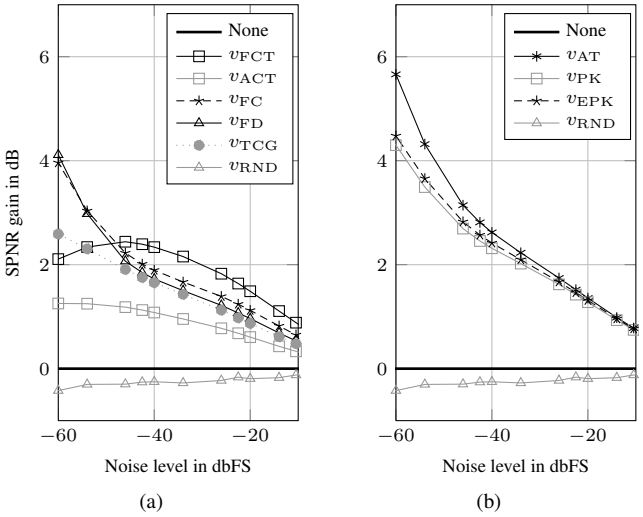


Figure 6: *Gain of the Sinusoidal Peaks to Noise Ratio (SPNR) with the synthetic test signal in dependency of the noise level.*

mentioned in Sect. 2 have been evaluated with synthetic test signals, usually composed of sinusoids and noise. Using these constructed signals allows to annotate tonal and non-tonal components accurately and thus enables detailed evaluation, but it is unclear whether the results can be generalized to real-world signals.

Therefore, we evaluate the features in two variants: first in the established way with synthetic test signals and second by measuring the performance of an MIR algorithm with and without a tonalness pre-processing step.

The sampling rate was 44.1 kHz for all test cases and the following STFT parameters were chosen:
$N_{\mathrm{W}} = 8192$, $N_{\mathrm{FFT}} = 2N_{\mathrm{W}} = 16384$, $N_{\mathrm{H}} = {}^{N_{\mathrm{W}}}/_{8} = 1024$.

5.1. Synthetic test cases

A test signal has been generated containing various combinations of tones with different amplitudes. Each tone consists of a fundamental frequency and 32 harmonics with an exponentially decaying spectral envelope. All tones or chords have zero attack and exponential decay times followed by a short gap of 0.2 s. As the synthesis was done by a simple sum of sinusoidals it was easy to build a spectral mask for all active frequencies per frame. The width of the spectral mask should roughly correspond to the main lobe width of the window function and was determined empirically. An example mask is shown in Fig. 5.

The assumption is that the tonalness should be close to zero outside of the mask and close to one inside. Furthermore, it is assumed that it reaches its maximum value at the center of the masked regions. When $\hat{X}(k,n) = X(k,n) \cdot T(k,n)$ is the magnitude spectrum weighted by the tonalness, we expect the non-tonal components to be attenuated while the tonal components remain untouched.

In order to evaluate the tonalness we define the Sinusoidal Peaks to Noise Ratio (SPNR)

$$\mathrm{SPNR} = 10 \log_{10} \left(\frac{\sum_{k \in \mathbb{P}} |\hat{X}(k,n)|^2}{\sum_{k \in \mathbb{N}} |\hat{X}(k,n)|^2} \right), \qquad (15)$$

with $\mathbb{P}$ denoting the set of indices of center bins of a masked region and $\mathbb{N}$ denoting the indices of all bins outside of the mask. In other words, the SPNR gives the ratio between the energy at the peak locations and the energy outside of the sinusoidal main lobes. The weighted magnitude spectrum should thus have a higher SPNR than the non-weighted magnitude spectrum.

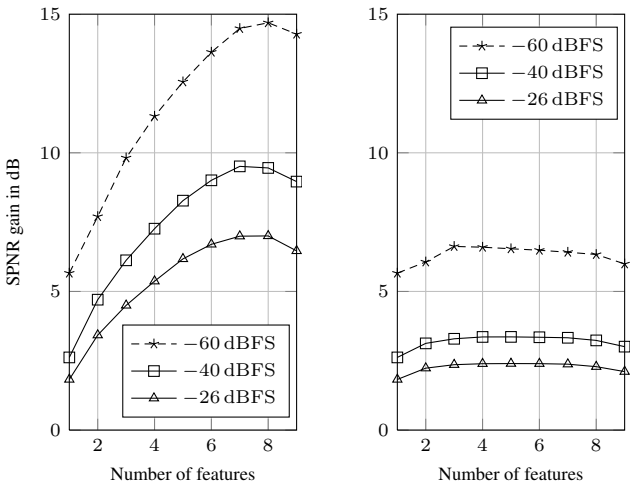
White noise with different levels was added to the described test signal and the SPNR was measured before and after weighting the spectra with the specific tonal scores. The resulting SPNR gains are plotted in Fig. 6. Using the specific tonal scores the SPNR is increased for all features. Only the random feature slightly degrades the SPNR as expected.

5.1.1. Sequential-Forward Selection

All individual tonal scores potentially increase the SPNR and it is of interest to see if a combination of tonal scores would further improve the results. To find the best feature combinations we have implemented a Sequential-Forward Selection strategy [12]. This is an iterative process by which first the best performing single feature is selected and then combined with all other individual features. The feature pair which improves the SPNR most is selected and then combined with the other remaining features. This selection process is carried out with the simple product as well as with the geometric mean feature combination (see Eq. (2)).

Figure 7 plots the gain of the SPNR values depending on the number of features chosen with the forward selection strategy and Table 1 shows the corresponding selected feature names for an input noise level of $-40\,\mathrm{dBFS}$. In general, one can see that it is possible to clearly improve the performance of the best individual feature by a combination with another feature. Also utilizing a simple product is far more effective than the geometric mean. The latter only shows a slight increase in SPNR and stagnates after 3 features, whereas the product can increase the SPNR up to a combination of 7 or 8 features. During the selection process, a combination with the random feature always degraded the SPNR and therefore, it has never been selected.

The order of the selected features is unfortunately not fixed and depends on the input noise level as well as the combination type (geometric mean or simple product). But we can observe the *Amplitude Threshold* and *Frequency Continuity* being among the first three selected features most of the time.



(a) *Product* (b) *Geom. mean*

Figure 7: *SPNR gain depending on the amount of combined features for different input noise levels.*

Combination	SPNR gain
Prod.	
AT	2.6 dB
AT, FCT	4.7 dB
AT, FCT, PK	6.1 dB
AT, FCT, PK, EPK	7.3 dB
AT, FCT, PK, EPK, FC	8.3 dB
AT, FCT, PK, EPK, FC, TCG	9.0 dB
AT, FCT, PK, EPK, FC, TCG, FD	9.5 dB
AT, FCT, PK, EPK, FC, TCG, FD, ACT	9.5 dB
Geom. mean	
AT	2.6 dB
AT, FCT	3.1 dB
AT, FCT, FD	3.3 dB
AT, FCT, FD, TCG	3.6 dB
AT, FCT, FD, TCG, EPK	3.6 dB

Table 1: *SPNR gain for different feature combinations and an input noise level of* $-40\,\mathrm{dBFS}$.

5.2. Real-world application

The results above indicate that all features are able to emphasize the sinusoidal or tonal parts of a spectrum and, compared to the random feature, it is evident that the tonalness includes a gain of information. In the second evaluation it is investigated whether these results also apply to real-world music signals and real-world applications.

Due to the difficulties of annotating tonal components mentioned above, an indirect approach of evaluation by using a key detection algorithm is proposed that allows the usage of real music signals as input data. Since a simple key detection algorithm depends on tonal components and will perform better with suppressed non-tonal components, the key detection performance can be evaluated and compared with and without utilizing the tonalness. This has the advantage that on the one hand the annotation of ground truth data is comparably easy and leads to a simple but reliable measure, and on the other hand the tonalness estimation is evaluated in the context of an everyday MIR application.

5.2.1. Key Detection

The algorithm implemented for detecting the musical key is kept as simple as possible in order to avoid too many processing steps between the evaluation metric and the tonalness spectrum itself.

First, the pitch chroma vector per block is computed. This is a twelve-dimensional histogram-like octave-independent vector showing the "strength" of the 12 semitone classes (C, C♯, D, . . ., B). It is computed by converting the spectrum to semi-tone bands and summing the energy of all bands with the distance of an octave [13]. Second, the average chroma vector per file is computed and finally the (Euclidean) distance between the extracted chroma vector and the shifted key profiles (acc. to Krumhansl [14]) is determined. The minimum of the 24 computed distances (12 major and 12 minor keys) identifies the most likely key of this piece.

There are, of course, more refined approaches to key detection which lead to better accuracies, see e.g. [15], but in this case we are only interested in comparing key detection results for different inputs of the pitch chroma vector calculation.

Two data sets have been used to evaluate the key detection rate. The first one has been previously used for the evaluation of a key detection system [16]. We will refer to this data set as ALERC; it consists of 145 full length songs in the *pop*/*rock* category and 65 songs in the *jazz*/*folk* category; all pieces have been manually annotated.

The second data set, used for validation, is the well known genre classification data set that we will refer to as GTZAN [17]. It has been manually annotated with key labels for the purpose of this evaluation. This data set consists of 1000 song snippets with a length of 30 s and is divided in 10 genres with 100 songs per genre. The classical genre has not been annotated and of the remaining categories 63 songs were not labeled because they contained a key modulation or were particularly difficult to annotate for other reasons. This results in a test set of 837 song snippets from the genres *blues*, *country*, *disco*, *hiphop*, *jazz*, *metal*, *pop*, *rock*, and *reggae*.

5.2.2. Individual tonal scores

The key detection is run with the unweighted magnitude spectra and the spectra weighted by the tonalness prior to the chroma vector calculation. For the unweighted spectra, the detection accuracy was 55.7 % for the ALERC data set and 40.4 % for the GTZAN data set. The achieved detection accuracy gains for the weighted spectra are visualized in Fig. 8 for both data sets.

Most features significantly improve the key detection rate while the random score has nearly no impact on the results. Only the *Frequency Coherence* feature decreased the accuracy on both data sets and the *Frequency Deviation* yielded a relatively low gain. The best individual feature on the GTZAN data set is the *Extended Peakiness* with an improvement of 4.0 %. The simple *Peakiness*, *Amplitude Threshold* and *Time Window Center of Gravity* produce results in the same range and only *Frequency Continuity* and *Amplitude Continuity* did not perform as well. With the ALERC data set, a maximum gain of 7.6 % is found with the *Time Window Center of Gravity*; the remaining features are all a bit lower but still in a similar range. Interestingly, *Frequency* and *Amplitude Continuity* worked far better on this data set compared to the results from the GTZAN data set.

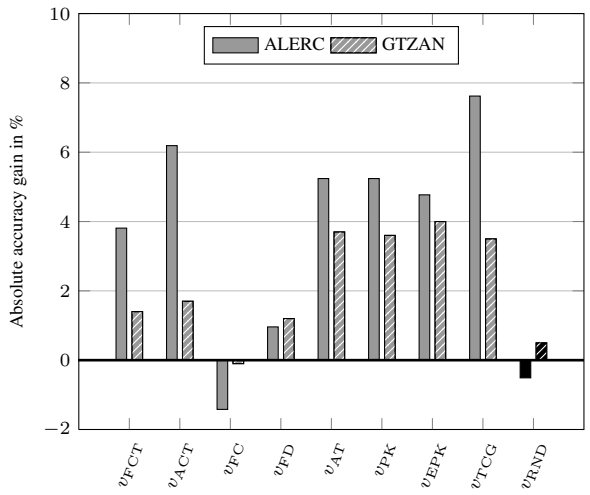


Figure 8: *Key detection accuracy gains by the use of tonalness weighted spectra for the GTZAN and ALERC data set.*

#	Combination	ALERC	GTZAN
	Product		
1	TCG	7.6 %	3.5 %
2	TCG, FCT	9.1 %	4.2 %
3	TCG, FCT, ACT	10.5 %	3.6 %
	Geom. mean		
1	TCG	7.6 %	3.5 %
2	TCG, FD	8.6 %	3.8 %
3	TCG, FD, PK	9.1 %	3.7 %

(a) *ALERC data set*

#	Combination	GTZAN	ALERC
	Product		
1	EPK	4.0 %	4.8 %
2	EPK, TCG	4.9 %	8.1 %
	Geom. mean		
1	EPK	4.0 %	4.8 %
2	EPK, ACT	4.5 %	7.2 %

(b) *GTZAN data set*

Table 2: *Achieved key detection accuracy gains by Sequential-Forward feature combination.*

It is noteworthy that the absolute detection accuracy on the GTZAN data set is lower than on the ALERC data set. By investigating the key results per genre, it was possible to identify the reason in a remarkably bad performance mainly in the blues, hiphop, and jazz genres. A possible explanation is the specific harmonic structure in the blues and jazz genres as well as the lesser amount of pitched content in the hiphop genre. Furthermore, the audio quality of the GTZAN dataset may influence the results as it includes, amongst others, radio recordings and files with quality degradations, while the ALERC data set exclusively consists of high quality CD recordings.

5.2.3. Sequential-Forward Selection

The same Forward Selection strategy as described in Sect. 5.1.1 has also been applied to the key detection evaluation and the best performing combinations are shown in Table 2. Again it is possible to improve the performance of the best individual feature by a combination with another one. However, combinations of more than 3 features did not result in further improvement and sometimes even slightly decreased the detection accuracy.

The product combination of *Time Window Center of Gravity*, *Frequency Continuity* and *Amplitude Continuity* yields an overall gain of 10.5 % on the ALERC data set; this is 2.9 % better than the best single feature on this data set. The best geometric mean combination is *Time Window Center of Gravity*, *Frequency Deviation* combined with *Peakiness* and improves the best individual tonal score by 1.5 %.

A similar behavior can be observed with the GTZAN data set, although the overall gain is lower and the maximum accuracy is already achieved by only using two features. A product combination of the *Extended Peakiness* and *Time Window Center of Gravity* results in a gain of 4.9 % and the geometric mean combination of *Extended Peakiness* and *Amplitude Continuity* is 4.5 % better than the unweighted spectra.

5.3. Discussion

Summarizing the evaluation is rather difficult because there is no best single feature or outstanding feature combination that excels in all evaluation tasks. From the results of the synthetic evaluation it is proven that all features are able to successfully detect tonal components of a signal. But these results could not be generalized to the key detection task where most of the features performed differently. The *Frequency Coherence*, for example, was en par with the other features for synthetic test signals but slightly decreased the key detection rate in the context of real-world signals.

When it comes to the combination of features the results from both the synthetic and the real-world task indicate that this can drastically improve the performance. Although we cannot pick a single superior combination, the results improve most if features dealing with diverse aspects of tonalness are combined. The *Peakiness*, for example, is solely based on information from the current magnitude spectrum while the *Frequency* and *Amplitude Continuity* take into account changes over time. Furthermore, the *Time Window Center of Gravity* is based on the complex spectrum considering the phase and the amplitudes of the signal.

The combinations with a simple product always lead to better results than with the geometric mean. This seems reasonable in our evaluation environment as the geometric mean is just a nonlinear distortion of the simple product and increases the weight of bins with low tonalness. Thus the geometric mean will result in less damping of frequency bins with an uncertain tonalness.

A preliminary Principal Component Analysis (PCA) with the tonal scores retrieved from the ALERC data set revealed that the 8 features span a space of roughly 2-3 independent dimensions. This matches the results from the key detection task, where combinations of up to 3 features were able to improve the detection accuracy.

6. SUMMARY

In this paper, a framework and a set of spectral features to estimate the tonalness of spectral bins was presented. The tonalness indicates the likelihood of a spectral bin being tonal. We investigated a set of simple and established features to derive tonal scores and evaluated the tonalness estimation with individual features as well as with feature combinations. The evaluation was conducted with both, synthetic signals and real music in the context of a typical MIR application. The evaluation with the real-world data (key detection) shows a gain in accuracy of more than 10 % on the ALERC data set and approximately 5 % on the GTZAN data set for the best feature combinations. The differences in the evaluation results between the synthetic and the real-world data set indicate that the findings from theoretic evaluation procedures cannot be directly transferred to an application. Instead, it is necessary to perform the evaluation with real-world data and to carefully select from a diverse set of features specifically for the intended application.

The proposed tonalness spectrum is efficient to compute and the framework is easily extensible with more powerful and possibly application specific features. Examples include further features from [2] or more complex amplitude based measures like the correlation between magnitude spectrum and window function [3, 4] as well as high order auto-regressive modelling [18]. The application of post-processing options such as partial tracking [19, 20, 21] also can be expected to improve results.

7. REFERENCES

[1] F. J. Charpentier, "Pitch Detection Using the Short-Term Phase Spectrum," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Tokyo, 1986, pp. 113–116, IEEE.

[2] Axel Roebel, Miroslav Zivanovic, and Xavier Rodet, "Signal decomposition by means of classification of spectral peaks," in *Proceedings of the International Computer Music Conference (ICMC)*, Miami, 2004, ICMA.

[3] Geoffroy Peeters and Xavier Rodet, "Signal Characterization in terms of Sinusoidal and Non-Sinusoidal Components," in *Proceedings of the 1st Conference on Digital Audio Effects (DAFX)*, Barcelona, 1998.

[4] Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault, "Sinusoidal Parameter Extraction and Component Selection in a Non Stationary Model," in *Proceedings of the 5th International Conference on Digital Audio Effects (DAFX)*, Hamburg, 2002.

[5] Thomas W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *Journal of the Acoustical Society of America (JASA)*, vol. 60, no. 4, pp. 911, 1976.

[6] Ernst Terhardt, Gerhard Stoll, and Manfred Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *Journal of the Acoustical Society of America (JASA)*, vol. 71, no. 3, pp. 679, 1982.

[7] Xavier Serra, *A System for Sound Analysis / Transformation / Synthesis Based on a Deterministic plus Stochastic Decomposition*, Dissertation, Stanford University, 1989.

[8] Maciej Kulesza and Andrzej Czyzewski, "Frequency based criterion for distinguishing tonal and noisy spectral components," *International Journal of Computer Science and Security*, vol. 4, no. 1, pp. 1, Mar. 2010.

[9] Simon Haykin, *Neural Networks - A Comprehensive Foundation*, Prentice Hall Inc, Upper Saddle River, 1994.

[10] François Auger and Patrick Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[11] Axel Roebel, "Transient detection and preservation in the phase vocoder," in *Proceedings of the International Computer Music Conference (ICMC)*, Singapore, 2003, pp. 247–250.

[12] Alexander Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*, Wiley-IEEE Press, Hoboken, 2012.

[13] Mark A. Bartsch and Gregory H. Wakefield, "To Catch a Chorus: Using Chroma-Based Representations for Audio Thumbnailing," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, 2001, IEEE.

[14] Carol L. Krumhansl, *Cognitive Foundations of Musical Pitch*, Oxford University Press, New York, 1990.

[15] Özgür Izmirli, "Template based key finding from audio," in *Proceedings of the International Computer Music Conference (ICMC)*, Barcelona, Sept. 2005.

[16] Alexander Lerch, "Ein Ansatz zur automatischen Erkennung der Tonart in Musikdateien," in *Proceedings of the VDT International Audio Convention (23. Tonmeistertagung)*, Leipzig, Nov. 2004.

[17] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, July 2002.

[18] Toon van Waterschoot and Marc Moonen, "Comparison of Linear Prediction Models for Audio Signals," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, pp. 1–24, 2008.

[19] Thomas F. Quatieri and R. J. McAulay, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1449–1464, Dec. 1986.

[20] Phillipe Depalle, Guillermo Garcia, and Xavier Rodet, "Tracking of partials for additive sound synthesis using hidden Markov models," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Minneapolis, 1993, IEEE.

[21] Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault, "Enhancing the Tracking of Partials for the Sinusoidal Modeling of Polyphonic Sounds," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1625–1634, July 2007.