AN ITERATIVE SEGMENTATION ALGORITHM FOR AUDIO SIGNAL SPECTRA DEPENDING ON ESTIMATED LOCAL CENTERS OF GRAVITY

Sascha Disch,[*]
Laboratorium für Informationstechnologie (LFI)
Leibniz Universität Hannover
Schneiderberg 32, 30167 Hannover, Germany
disch@tnt.uni-hannover.de

Bernd Edler
Laboratorium für Informationstechnologie (LFI)
Leibniz Universität Hannover
Schneiderberg 32, 30167 Hannover, Germany
edler@tnt.uni-hannover.de

ABSTRACT

Modern music production and sound generation often relies on manipulation of pre-recorded pieces of audio, so-called *samples*, taken from a huge database. Consequently, there is a increasing request to extensively adapt these samples to any new musical context in a flexible way. For this purpose, advanced digital signal processing is needed in order to realize *audio effects* like *pitch shifting*, *time stretching* or *harmonization*. Often, a key part of these processing methods is a signal adaptive, block based spectral segmentation operation. Hence, we propose a novel algorithm for such a spectral segmentation based on local *centers of gravity* (COG). The method was originally developed as part of a multiband modulation decomposition for audio signals. Nevertheless, this algorithm can also be used in the more general context of improved vocoder related applications.

1. INTRODUCTION

There is an increasing demand for digital signal processing techniques that address the need for extreme signal manipulations in order to fit pre-recorded audio signals, e.g. taken from a database, into a new musical context. In order to do so, high level semantic signal properties like pitch, musical key and scale mode are needed to be adapted. All these manipulations have in common that they aim at substantially altering the musical properties of the original audio material while preserving subjective sound quality as good as possible. In other words, these edits strongly change the audio material musical content but, nevertheless, are required to preserve the *naturalness* of the processed audio sample and thus maintain *believability*. This ideally requires signal processing methods that are broadly applicable to different classes of signals including polyphonic mixed music content.

Therefore, a method for analysis, manipulation and synthesis of audio signals based on multiband modulation components has been proposed lately [1][2]. The fundamental idea of this approach is to decompose polyphonic mixtures into components that are perceived as sonic entities anyway, and to further manipulate all signal elements that are contained in one component in a joint fashion. Additionally, a synthesis method has been introduced that renders a smooth and perceptually pleasant yet - depending on the type of manipulation applied - drastically modified output signal. If no manipulation whatsoever is applied to the components the method has been shown to provide transparent or near-transparent subjective audio quality [1] for many test signals.

[*] This work was supported by Fraunhofer IIS, Erlangen, Germany.

An important step for our block based polyphonic music manipulation, e.g. the multiband modulation decomposition, is the estimation of local *centers of gravity* (COG) [3][4] in successive spectra over time. This paper amends the detailed description of an iterative algorithm, that can be used to determine a signal adaptive spectral decomposition that is aligned with local COG of the signal.

The COG approach may be reminiscent of the classic *time-frequency reassignment* (t-f reassignment) method. For an extensive overview on this technique the reader is referred to [5]. Basically, t-f reassignment alters the regular time-frequency grid of a conventional *Short Time Fourier Transform* (STFT) towards a time-corrected instantaneous frequency spectrogram, thereby revealing temporal and spectral accumulations of energy that are better localized than implicated by the t-f resolution compromise inherent in the STFT spectrogram. Often, reassignment is used as an enhanced front-end for subsequent partial tracking [6]. In contrast to that, our algorithm directly performs a spectral segmentation on a perceptually adapted scale, while t-f reassignment solely provides for a better localized spectrogram and leaves the segmentation problem to later stages, e.g. partial tracking.

Other related publications aim at the estimation of multiple fundamental frequencies [7][8] by grouping spectral peaks which exhibit certain harmonic relations into separate sources. However, for complex music composed of many sources (like orchestral music), this approach has no reasonable chance. In contrast, the approach presented in this paper does not attempt to decompose the signal into its sources, but rather segments spectra into perceptual units which can be further manipulated conjointly.

In this paper, we start with a brief review of the aforementioned modulation analysis/synthesis system. In the following, we focus on the details of a novel multiple local COG estimation algorithm followed by the derivation of a set of bandpass filters aligned with the estimated COG positions. Some exemplary result data of the COG estimation and its associated set of of bandpass filters is presented and discussed.

2. MODULATION DECOMPOSITION

2.1. Background

The multiband modulation decomposition dissects the audio signal into a signal adaptive set of (analytical) bandpass signals, each of which is further divided into a sinusoidal carrier and its *amplitude modulation* (AM) and *frequency modulation* (FM). The set of bandpass filters is computed such that on the one hand the fullband spectrum is covered seamlessly and on the other hand the

filters are aligned with local COGs each. Additionally, the human auditory perception is accounted for by choosing the bandwidth of the filters to match a perceptual scale e.g. the ERB scale [9].

The local COG corresponds to the mean frequency that is perceived by a listener due to the spectral contributions in that frequency region. Moreover, the bands centered at local COG positions correspond to *regions of influence* based phase locking of classic phase vocoders [10][11][12][13]. The bandpass signal envelope representation and the traditional region of influence phase locking both preserve the temporal envelope of a bandpass signal: either intrinsically or, in the latter case, by ensuring local spectral phase coherence during synthesis. With respect to a sinusoidal carrier of a frequency corresponding to the estimated local COG, both AM and FM are captured in the amplitude envelope and the heterodyned phase of the analytical bandpass signals, respectively. A dedicated synthesis method renders the output signal from the carrier frequencies, AM and FM.

2.2. Modulation analysis

A block diagram of the signal decomposition into carrier signals and their associated modulation components is depicted in Figure 1. In the picture, the schematic signal flow for the extraction of one component is shown. All other components are obtained in a similar fashion. Practically, the extraction is carried out jointly for all components on a block-by-block basis using e.g. a block size of $N = 2^{14}$ at 48 kHz sampling frequency and 75% analysis overlap - roughly corresponding to a time interval of 340 ms and a stride of 85 ms - by application of a *discrete fourier transform* (DFT) on each windowed signal block. The window is a 'flat top' window according to Equation (1). This ensures that the centered N/2 samples that are passed on for the subsequent modulation synthesis are unaffected by the slopes of the analysis window. A higher degree of overlap may be used for improved accuracy at the cost of increased computational complexity.

$$window\,(i)_{analysis} = \begin{cases} \sin^2(\frac{2i\pi}{N}) & 0 < i < \frac{N}{4} \\ 1 & \frac{N}{4} \leq i < \frac{3N}{4} \\ \sin^2(\frac{2i\pi}{N}) & \frac{3N}{4} \leq i < N \end{cases} \quad (1)$$

Given the spectral representation, next a set of signal adaptive spectral weighting functions (having bandpass characteristic) that is aligned with local COG positions is calculated.

After application of the bandpass weighting to the spectrum, the signal is transformed to the time domain and the analytic signal is derived by Hilbert transform. These two processing steps can be efficiently combined by calculation of a single-sided IDFT on each bandpass signal.

Subsequently, each analytic signal is heterodyned by its estimated carrier frequency. Finally, the signal is further decomposed into its amplitude envelope and its *instantaneous frequency* (IF) track, obtained by computing the phase derivative, yielding the desired AM and FM signals [1].

2.3. Modulation synthesis

The signal is synthesized on an additive basis of all components. For one component the processing chain is shown in Figure 2. Like the analysis, the synthesis is performed on a block-by-block basis. Since only the centered N/2 portion of each analysis block is evaluated for synthesis, a synthesis overlap factor of 50% results.

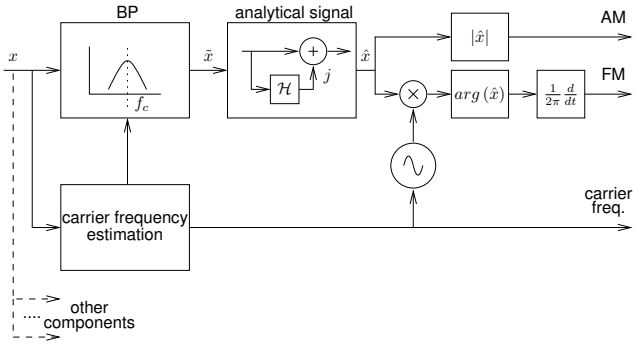

Figure 1: Modulation analysis

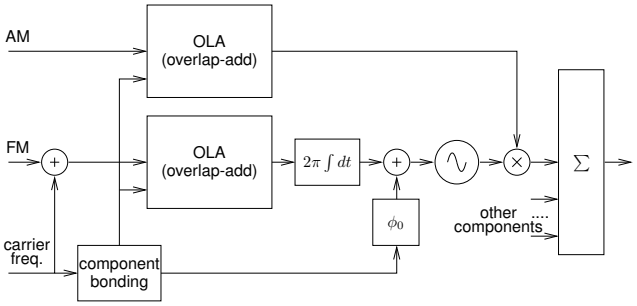

Figure 2: Modulation synthesis

Overlap-add (OLA) is applied in the parameter domain rather than on the readily synthesized signal in order to avoid beating effects between adjacent time blocks. The OLA is controlled by a component bonding mechanism, that, steered by spectral vicinity (measured on an ERB scale), performs a pair-wise match of components of the actual block to their predecessors in the previous block. Additionally, the bonding alignes the absolute component phases of the actual block to the ones of the previous block.

In detail, firstly the FM signal is added to the carrier frequency and the result is passed on to the OLA stage, the output of which is integrated subsequently. A sinusoidal oscillator is fed by the resulting phase signal. The AM signal is processed by a second OLA stage. Finally, the output of the oscillator is modulated in its amplitude by the resulting AM signal to obtain the additive contribution of the component to the output signal.

It should be emphasized that an appropriate spectral segmentation of the signal within the modulation analysis is of paramount importance for a convincing result of any further modulation parameter processing. Therefore, in this paper, a novel suitable segmentation algorithm is presented.

3. ITERATIVE SEGMENTATION ALGORITHM

3.1. Principle

The segmentation algorithm proposed herein consists of an initial COG spectral position candidate list that is iteratively updated by refined estimates. In the process of refinement, addition, deletion or fusion of candidates is incorporated, thus the method does not require a-priori knowledge of the total number of final COG estimates. The iteration is implemented by two loops. All necessary

operations are performed on a spectral representation of the signal. The details are outlined in the following.

3.2. Pre-processing

For each signal block, a *power spectral densitiy* (psd) estimate is obtained by computing the DFT spectral energy. Subsequently, in order to remove the global trend, the psd is normalized on a smoothed psd that is calculated by linear regression. Prior to division, both quantities are temporally smoothed by a first order IIR filter with time constant of approx. 200 ms.

Next, a mapping of the psd is performed onto a perceptual scale prior to COG calculation and segmentation in order to facilitate the task of segmenting a spectrum into perceptually adapted non-uniform and, at the same time, COG centered bands. Thereby the problem is simplified to the task of an alignment of a set of approximately uniform segments with the estimated local COG positions of the signal.

As a perceptual scale the ERB scale [9] is applied which provides better spectral resolution at lower frequencies than e.g. the BARK scale. The mapped spectrum is calculated by interpolation of the uniformly sampled spectrum towards spectral samples that are spaced following the ERB scale (2).

$$ERB\left(f\right) = 21.4\log_{10}\left(0.00437f + 1\right) \tag{2}$$

3.3. Iterative center of gravity estimation

The iterative COG estimation flowchart is depicted in Figure 3. For each time block k, a sorted position candidate list c is initialized with a uniformly spaced grid of N candidate positions $c(n)$ having a spacing S. Most important, the parameter S sets the spectral resolution of the estimates obtained in the course of the iteration process. Phrased differently, the parameter S determines what is considered to be the local scope of the COG estimation.

$$\begin{aligned} c\left(n\right) &= nS \\ n &\in [1, 2..., N] \end{aligned} \tag{3}$$

The iteration process consists of two nested loops. The outer loop calculates the position offset ($posOff$) of the candidate position from the true local center of gravity by application of a negative-to-positive linear slope function of size $2S$, weighted by weights $g\left(i\right)$, to each candidate position n on the preprocessed psd estimate of a signal block (4).

$$\begin{aligned} posOff\left(n\right) &= \text{round}\left(\frac{\sum_i \left(w_i\left(n\right) \cdot idxOff\left(i\right)\right)}{\sum_i w_i\left(n\right)}\right) \\ w_i\left(n\right) &= \text{psd}\left(c\left(n\right) + idx\left(i\right)\right) \cdot g\left(i\right) \\ idxOff\left(i\right) &= i - S + 0.5 \\ idx\left(i\right) &= \text{round}\left(idxOff\left(i\right)\right) \\ i &\in [0, 1, 2..., 2S-1] \end{aligned} \tag{4}$$

In a next step (5), all candidate positions from the list are updated by their position offset.

$$c\left(n\right) := c\left(n\right) + posOff\left(n\right) \tag{5}$$

Each candidate position that violates the border limitations is removed from the list as indicated by (6) and the number of remaining candidate positions N is decremented by 1.

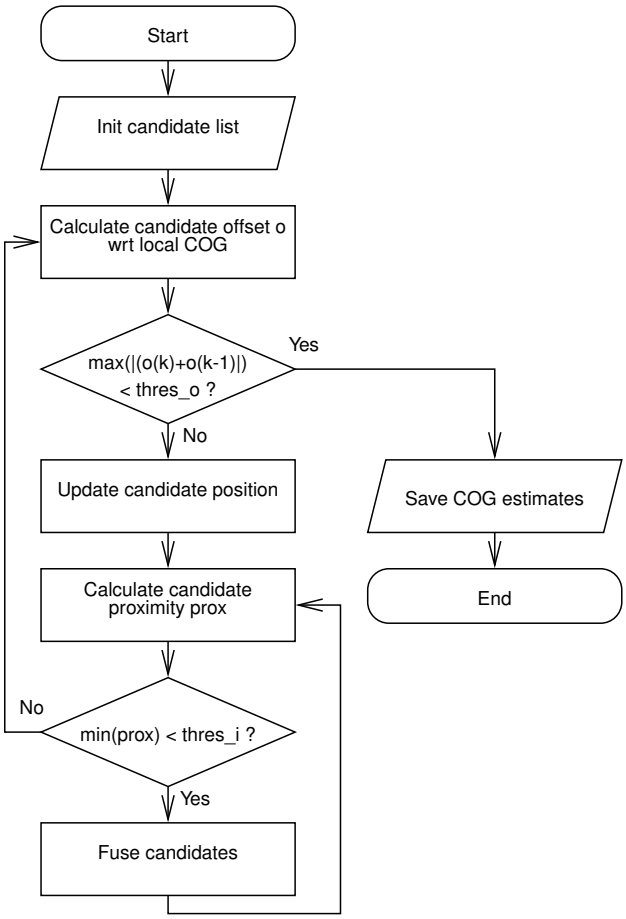

Figure 3: Flowchart of iterative COG estimation

$$\begin{aligned} &\text{if } \left(c\left(n\right) < S\right) \vee \left(c\left(n\right) > NS\right) \rightarrow \\ &c\left(x\right) := c\left(x+1\right) \ \forall x \in [n+1, ..., N-1] \\ &N := N - 1 \end{aligned} \tag{6}$$

If the absolute value of the sum of the actual and the previous position offset of all candidates is smaller than a predefined threshold the outer iteration loop is exited (7). Note that using this type of condition also terminates the iteration in case if the position offset toggles back and forth between two values.

$$\max\left(|posOff_k\left(n\right) + posOff_{k-1}\left(n\right)|\right) < thres\_o \tag{7}$$

Next, the inner loop is executed. The inner loop iteratively fuses the closest (according to a certain proximity measure) two position candidates that violate a predefined proximity restriction due to the position update provided by the outer loop into one single new candidate, thereby accounting for perceptual fusion. The proximity measure is the spectral distance of the two candidates (8).

$$\begin{aligned} |c\left(n\right) - c\left(n+1\right)| &< thres\_i \\ thres\_i &:= S \end{aligned} \tag{8}$$

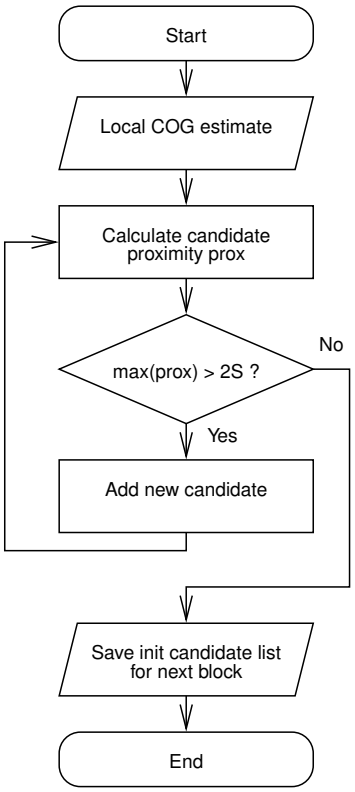

Figure 4: Flowchart of improved initialisation

Each newly calculated joint candidate is initialized to occupy the energy weighted mean position of the two former candidates (9).

$$\begin{aligned} c(n) &:= \text{round}\left(\frac{w(n)\,c(n) + w(n+1)\,c(n+1)}{w(n) + w(n+1)}\right) \\ w(n) &= \sum_i \left(\text{psd}\left(c(n) + idx(i)\right) \cdot g(i)\right) \\ c(x) &:= c(x+1)\ \forall x \in [n+1, ..., N-1] \\ N &:= N - 1 \end{aligned} \tag{9}$$

Both former candidates are deleted from the list and the new joint candidate is added to the list. Consequently, the number of remaining candidate positions N is decremented by 1. The inner loop iteration terminates if no more candidates violate the proximity restriction.

The final set of COG candidates constitutes the estimated local centers of gravity positions.

3.4. Improved initialization

In order to speed up the iteration process the initialization of each new block can advantageously be done using the COG position estimate of the previous block since it is already a fairly good estimate of the actual positions. This applies due to the block overlap in the analysis and hence the appropriate assumption of a limited change rate in temporal evolution of COG positions.

Still, care has to be taken to provide enough initial position estimates to also capture the possible emergence of new COG. Therefore, position candidate gaps in the estimate spanning a distance greater than $4S$ are filled by new COG position candidates (10) thus ensuring that potential new candidates are within the scope of the position update function. Figure 4 shows a flow chart of this extension to the algorithm.

The apposition of additional candidates to the list is accomplished with a loop that terminates if no more gaps larger than $4S$ are found.

$$\begin{aligned} &\text{if } \left(c(n+1) - c(n)\right) > 4S \rightarrow \\ &c(x+1) := c(x)\ \forall x \in [N, N-1, ..., n+1] \\ &c(n+1) := \text{round}\left(\frac{c(n) + c(n+1)}{2}\right) \\ &N := N + 1 \end{aligned} \tag{10}$$

3.5. Design of bandpass filter set

After having determined the COG estimates in the ERB adapted domain the COG positions are mapped back into the linear domain by solving (2) for f.

Next, a set of bandpass filters is calculated in the form of spectral weights, which are to be applied to the DFT spectrum of the broadband signal.

The bandpass filters are designed to have a pre-defined roll-off with sine-squared characteristic. To achieve the desired alignment with the estimated COG positions, the design procedure described in the following is applied. Firstly, intermediate positions between adjacent COG position estimates are calculated. Then, at this transition points, the roll-off parts of the spectral weights are centered such that the roll-off parts of neighboring filters sum up to one. The middle section of the bandpass weighting function is choosen to be flat-top equal to one.

In designing the roll-off characteristic, a trade-off has to be made with respect to spectral selectivity on the one hand and temporal resolution on the other hand. Also, allowing multiple filters to spectrally overlap may add an additional degree of freedom to the design restrictions. The trade-off may be chosen in a signal adaptive fashion for e.g. improving on the reproduction of transients.

4. RESULTS

Figures 5, 6, 7, 8 visualize results obtained by the proposed iterative local COG estimation algorithm of subsection 3.3 that has been applied to different test items.

The test items are two separate pure tones, two tones that beat with each other, plucked strings ('MPEG Test Set - sm03') and orchestral music ('Vivaldi - Four Seasons, Spring, Allegro'). In these figures, the perceptually mapped, smoothed and globally detrended spectrum is displayed (gray, line plot) along with the COG estimates (black, stem plot). The COG estimates are numbered in ascending order. While e.g. the estimates no.22, no.26 of Figure 5 and estimates no.18 and no.19 of Figure 7 correspond to sinusoidal signal components, estimate no.22 of Figure 6, estimates no.23 and no.25 of Figure 7 and most estimates of Figure 8 capture spectrally broadened or beating components, which are nevertheless detected and segmented well, thus grouping them into perceptual units.

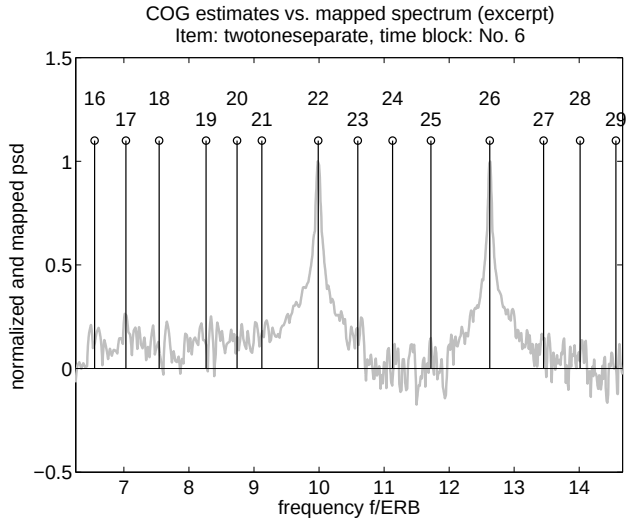

Figure 5: Two separate tones - Local centers of gravity (black, stem plot) vs. mapped spectrum (gray, line plot)

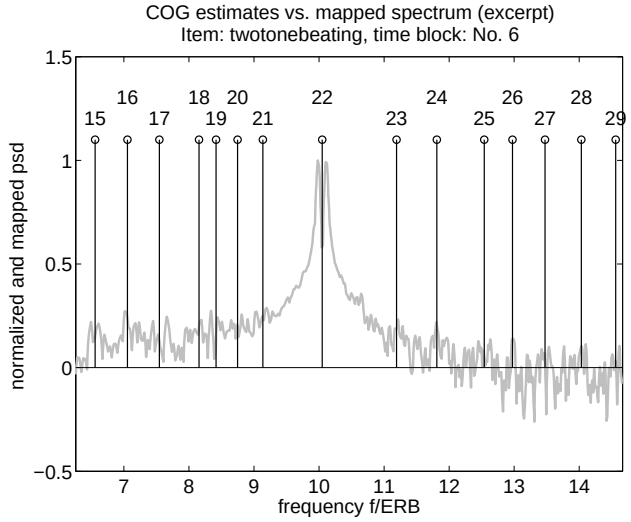

Figure 6: Two beating tones - Local centers of gravity (black, stem plot) vs. mapped spectrum (gray, line plot)

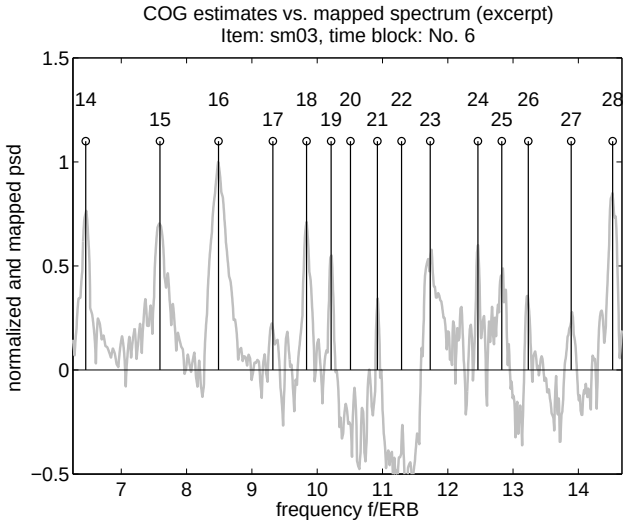

Figure 7: Plucked strings - Local centers of gravity (black, stem plot) vs. mapped spectrum (gray, line plot)

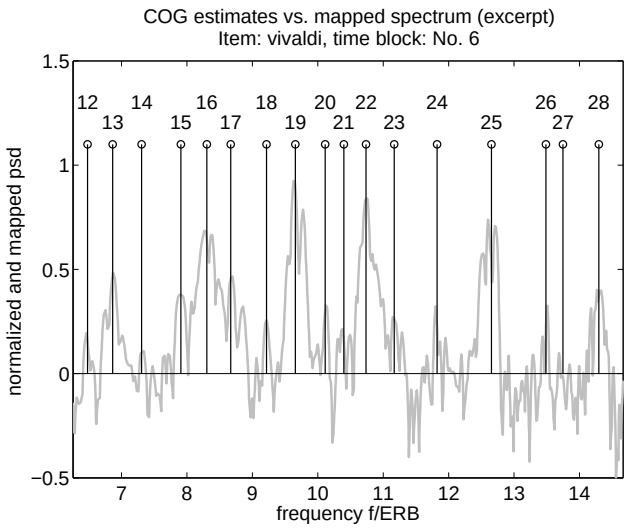

Figure 8: Orchestral music - Local centers of gravity (black, stem plot) vs. mapped spectrum (gray, line plot)

In Figures 9 and 10, the original - non pre-processed - psd of the signal block is depicted (gray) and a set of bandpass filters (black) is sketched, that has been designed as outlined in subsection 3.5. It is clearly visible, that each filter is aligned with a COG estimate and pairwise smoothly overlaps with its adjacent subband filters.

5. CONCLUSION

An important step in block based (polyphonic) music manipulation is the estimation of local *centers of gravity* (COG) in successive spectra over time. Motivated by the development of a signal adaptive multiband modulation decomposition, a detailed method and algorithm that estimates multiple local COG in the spectrum of an arbitrary audio signal has been proposed. Moreover, a design scheme for a set of resulting bandpass filters aligned to the estimated COG positions has been described. These filters may be utilized to subsequently separate the broadband signal into signal dependend perceptually adapted subband signals.

Exemplary results obtained by application of this method have been presented and discussed. However, a subjective audio quality assessment by listening tests evaluating applications that are based on the presented segmentation method are beyond the scope of this paper and will be the subject of future publications. Although developed in the context of a dedicated multiband modulation decomposition scheme, the proposed algorithm can potentially be used in the more general context of audio post-processing, audio effects and improved vocoder applications.

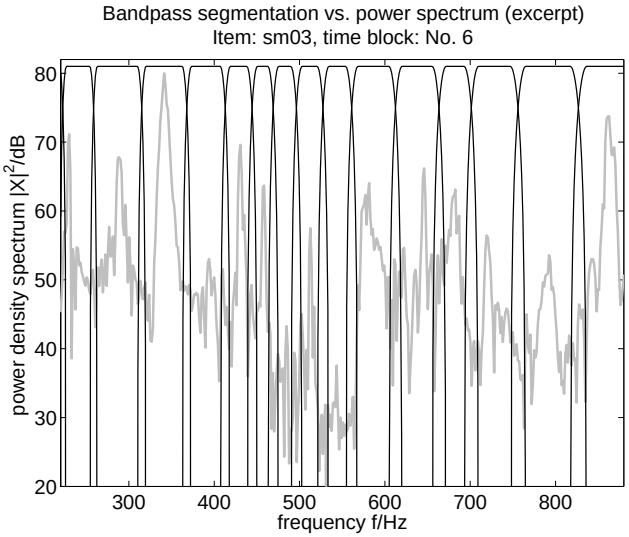

Figure 9: Plucked strings - Bandpass filters (black) aligned with local centers of gravity vs. power spectrum (gray)

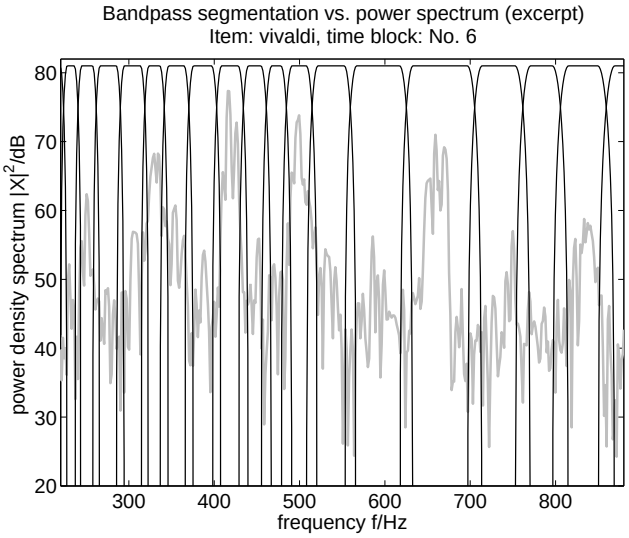

Figure 10: Orchestral music - Bandpass filters (black) aligned with local centers of gravity vs. power spectrum (gray)

6. REFERENCES

[1] S. Disch and B. Edler, "An amplitude- and frequency modulation vocoder for audio signal processing," *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, 2008.

[2] S. Disch and B. Edler, "Multiband perceptual modulation analysis, processing and synthesis of audio signals," *Proc. of the IEEE-ICASSP*, 2009.

[3] J. Anantharaman, A. Krishnamurthy, and L. Feth, "Intensity-weighted average of instantaneous frequency as a model for frequency discrimination.," *J. Acoust. Soc. Am.*, vol. 94, pp. 723–729, 1993.

[4] Q. Xu, L. L. Feth, J. N. Anantharaman, and A. K. Krishnamurthy, "Bandwidth of spectral resolution for the "c-o-g" effect in vowel-like complex sounds," *Acoustical Society of America Journal*, vol. 101, pp. 3149–+, May 1997.

[5] A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *Journal of the Acoustical Society of America*, vol. 119, pp. 360–371, 2006.

[6] K. Fitz and L. Haken, "On the use of time-frequency reassignment in additive sound modeling," *Journal of the Audio Engineering Society*, vol. 50(11), pp. 879–893, 2002.

[7] A Klapuri, *Signal Processing Methods For the Automatic Transcription of Music*, Ph.D. thesis, Tampere University of Technology, 2004.

[8] Chunghsin Yeh, *Multiple fundamental frequency estimation of polyphonic recordings*, Ph.D. thesis, École doctorale edité, Université de Paris, 2008.

[9] B. C. J. Moore and B. R. Glasberg, "A revision of zwicker's loudness model," *Acta Acustica*, vol. 82, pp. 335–345, 1996.

[10] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[11] Ch. Duxbury, M. Davies, and M. Sandler, "Improved time-scaling of musical audio using phase locking at transients," in *112th AES Convention*, 2002.

[12] A. Röbel, "A new approach to transient processing in the phase vocoder," *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, pp. 344–349, 2003.

[13] A. Röbel, "Transient detection and preservation in the phase vocoder," *Int. Computer Music Conference (ICMC'03)*, pp. 247–250, 2003.