MULTIBAND PERCEPTUAL MODULATION ANALYSIS, PROCESSING AND SYNTHESIS OF AUDIO SIGNALS

Sascha Disch, Bernd Edler

Leibniz Universität Hannover, Laboratorium für Informationstechnologie (LFI)
Schneiderberg 32, 30167 Hannover, Germany

Index Terms— Audio coding, Amplitude modulation, Frequency modulation, Time-frequency analysis

ABSTRACT

The decomposition of audio signals into perceptually meaningful multiband modulation components opens up new possibilities for advanced signal processing. The signal adaptive analysis approach proposed in this paper will be shown to provide a powerful handle on the signal's perceptual properties: pitch, timbre or roughness can be manipulated straight forward. Additionally a synthesis method is specified providing high subjective perceptual quality. Furthermore, as an application example, a novel audio processing technique is proposed which changes the key mode of a given piece of music e.g. from major to minor key or vice versa.

1. INTRODUCTION

The task of decomposing a wide-band audio signal into a set of components each comprising carrier, amplitude modulation, and frequency modulation is an ill-defined problem with an infinite number of solutions. Thus we pose the constraint of the decomposition being straight forward interpretable and perceptually meaningful in a sense that modulation processing applied on the modulation information should produce perceptually smooth results avoiding undesired artefacts introduced by the limitations of the modulation representation itself. This leads to the design goal, that the extracted carrier information alone should already allow for a coarse but representative 'sketch' reconstruction of the audio signal and any successive application of modulation related information should refine this representation towards full detail. This motivates the approach of partitioning the fullband signal by an adaptive set of band pass filters having local spectral centers of gravity as center frequencies rather than utilizing a static filter bank. Each band pass signal is further decomposed into amplitude modulation (AM) and frequency modulation (FM) providing the desired decomposition.

The paper is structured as follows: First we relate our work to other publications in the field. Then we motivate our approach to modulation decomposition followed by a description of our modulation analysis/synthesis system. We show a novel application ('key mode change') based on this method in the field of post-production/audio effects, that has not been considered possible so far.

2. BACKGROUND

The following analysis/synthesis strategy was initially proposed in [1]. However, we see some relation to previous work published by A. Master [2]. Moreover there are touch points with recent work of A. Röbel focused on solely modeling sinusoidal components [3]. Another recent publication of a similar decomposition is restricted to speech signals only and relies on additional a-priori side information on the fundamental frequency and the number of carriers to be modeled. In contrast to our method, psychoacoustic criteria are not part of the model [4]. The basic idea of using signal adaptive band pass filters dates back to a publication by A. Rao et al [5]. However there are also methods which use fixed band pass filters [6][7].

In our proposal the audio signal is decomposed into a signal adaptive set of (analytical) band pass signals, each of which is further divided into a sinusoidal carrier and its AM and FM. The set of band pass filters is computed such that on one hand it seamlessly covers the full-band spectrum and on the other hand the single filters are centered at local centers of gravity (COG) each. Additionally the human perception is accounted for by choosing the filter pass-band to follow a perceptual scale e.g. the ERB scale [8]. The local COG corresponds to the 'mean' frequency that is perceived by a human listener due to the spectral contribution in that frequency region. To see this relationship, note the equivalence of COG and 'intensity weighted average instantaneous frequency' (IWAIF) as derived in [9]. Moreover the use of band pass signals centered at local COG positions correspond to the 'regions of influence' based phase locking of traditional phase vocoders [10][11][12][13]. Our band pass signal envelope representation and the traditional region-of-influence phase locking both preserve the temporal envelope of a band pass signal: Our one intrinsically and the latter one by ensuring local spectral phase coherence during synthesis. AM and FM with respect to a sinusoidal carrier of a frequency corresponding to the local COG are captured in the amplitude

envelope and in the heterodyned phase of the analytical band pass signals, respectively. A synthesis method renders the output signal from the (processed) parameters, being carrier frequency, AM and FM.

Two basic kinds of parameter processing are conceivable [1]: change of carrier frequency while preserving modulation and change of the degree of modulation detail while keeping carriers unaffected. The first option is targeting at applications like pitch shift/correction or musical key transposition the latter at timbre alterations for controlling auditory roughness [14][15].

3. MODULATION ANALYSIS/SYNTHESIS

3.1. Analysis

The signal decomposition into carrier signals and their associated modulation components is depicted in Figure 1. In the picture, the theoretical 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% overlap, roughly corresponding to a time interval of 340 ms and a stride of 85 ms using a DFT on a windowed signal block. The window is a 'flat top' window according to Equation (1). This ensures that the centered N/2 samples used for 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.

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

Having the spectrum, next a set of signal adaptive spectral weighting functions (having band pass characteristic) that are each centered at a local COG positions has to be calculated. Care has to be taken that the resulting set of filters on the one hand covers the spectrum seamlessly and on the other hand adjacent filters do not overlap too much since this will result in undesired beating effects after the synthesis of (modified) components. This involves some compromise with respect to the bandwidth of the filters which should follow a perceptual scale but, at the same time, have to provide seamless spectral coverage and alignment with the local COG positions [1].

The resulting band pass weighting functions are applied to the DFT spectrum each. A single sided iDFT on every band pass spectrum yields the desired time domain analytical band pass signals. An example of such a segmentation is shown in Figure 2.

Subsequently, each analytic signal is heterodyned by its estimated carrier frequency. Finally, the signal is further decomposed into its amplitude envelope and its instantaneous

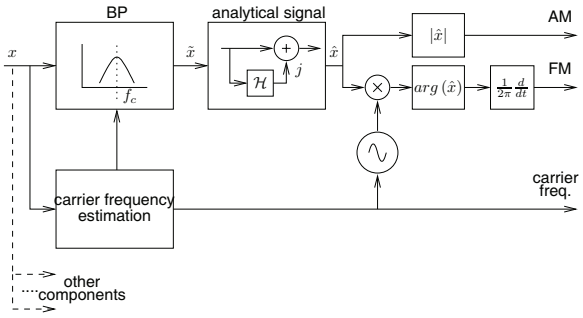

Fig. 1. Analysis

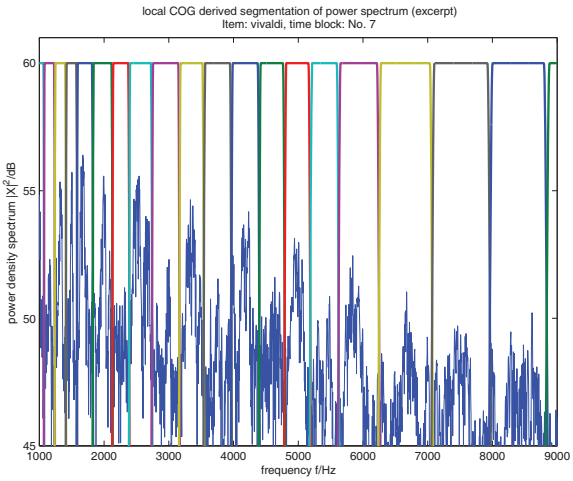

Fig. 2. Spectral Segmentation

frequency (IF) track yielding the desired AM and FM signals. If a separate processing of AM and FM parameters is intended, a subsequent smoothing of the FM, e.g. by constrained polynomial fitting, is adviseable in order to decouple AM and FM and thereby minimize artifacts[1].

It should be stressed at this point that the perceptually correct spectral segmentation of the signal is of paramount importance for a convincing result of any modulation parameter processing.

3.2. Efficient spectral segmentation

To facilitate the task of segmenting the spectrum into perceptually adapted non-uniform and, at the same time, COG centered bands, a mapping of the magnitude spectrum is performed onto a perceptually motivated scale prior to COG calculation and segmentation. Now the problem is simplified to an alignment of a set of approximately uniform segments with respect to the signal's local COG positions.

A suitable perceptual scale is the ERB scale [8] since it can be expressed fully analytical. The mapped spectrum is calculated by interpolation of the uniformly spaced support

points towards sample points according to the ERB scale described by Equation (2).

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

The local COG candidates are estimated on the mapped magnitude spectrum by searching positive-to-negative transitions in the CogPos function defined in Equation (3).

$$\begin{gathered} n(k,m) = \alpha \sum_{i=-B/2}^{B/2} i\,|X(k+i,m)|^2 + (1-\alpha) n(k,m-1) \\ d(k,m) = \alpha \sum_{i=-B/2}^{B/2} |X(k+i,m)|^2 + (1-\alpha) d(k,m-1) \\ CogPos(k,m) = \tfrac{n(k,m)}{d(k,m)} \\ \alpha = \tfrac{1}{\tau F_s};\ i \in \mathbb{N} \end{gathered} \tag{3}$$

At time block m, for every spectral coefficient index k it yields the relative offset towards the local center of gravity in the spectral region that is covered by a smooth sliding window w. The width B of the window is chosen in the range of 0.5-1 ERB. X(k,m) is the spectral coefficient k in time block m of the mapped magnitude spectrum. Additionally, a first order recursive temporal smoothing with time constant τ is done.

A post selection/processing procedure ensures that the final estimated COG positions are approximately equidistant on a perceptual scale.

Last, the set of weighting functions having band pass character is fitted to the final COG positions and mapped back to the linear domain by interpolation of the ERB spaced segment support points towards uniformly spaced sample points. The same applies for the COG positions which directly translate to carrier frequencies.

3.3. Synthesis

The signal is synthesized on an additive basis of all components. For one component the processing chain is shown in Figure 3. 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 used for synthesis, an overlap factor of 50% results. A component bonding mechanism is utilized to blend AM and FM and align absolute phase for components in spectral vicinity (measured on an ERB scale) of their predecessors in a previous block. Blending is done in the parameter domain rather than on the readily synthesized signal, thus beating effects between adjacent time blocks are avoided.

In detail firstly the FM signal is added to the carrier frequency and the result is passed on to the overlap-add (OLA) stage. The output is integrated to obtain the absolute phase of the component to be synthesized. A sinusoidal oscillator is fed by the resulting phase signal. The AM signal is processed by a parallel OLA stage. Finally the oscillator's output is modulated in its amplitude by the resulting AM signal

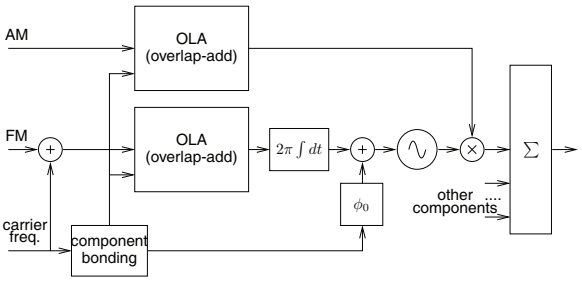

Fig. 3. Synthesis

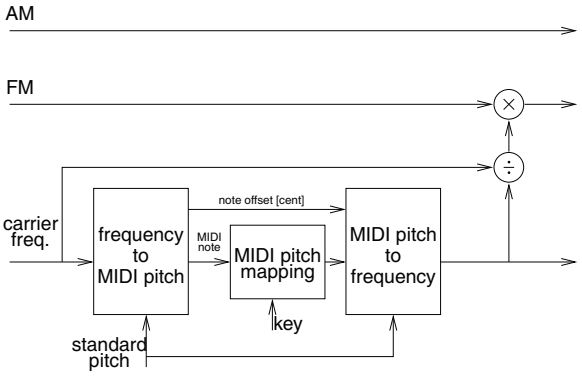

Fig. 4. Key mode change processing

to obtain the components' additive contribution to the output signal.

4. APPLICATION: POLYPHONIC KEY MODE CHANGE

Transposing of an audio signal while maintaining original playback speed is a challenging task. Using the proposed system, this is achieved straight forward by multiplication of all carrier components with a constant factor. Since the temporal structure of the input signal is solely captured by the AM signals it is unaffected by the stretching of the carrier's spectral spacing.

An even more demanding effect can be obtained by selective processing: the key mode of a piece of music can be changed from e.g. minor to major or vice versa. Therefor, only a subset of carriers corresponding to certain predefined frequency intervals is mapped to suitable new values. To achieve this, the carrier frequencies are quantized to MIDI notes which are subsequently mapped onto appropriate new MIDI notes (using a-priori knowledge of mode and key of the music item to be processed). The necessary processing is depicted in Figure 4.

The MIDI notes/tones to be mapped can be derived from the circle of fifth as depicted in Figure 5. Major to minor conversion is obtained by a leap of three steps counterclockwise, minor to major change by three steps clockwise. Lastly,

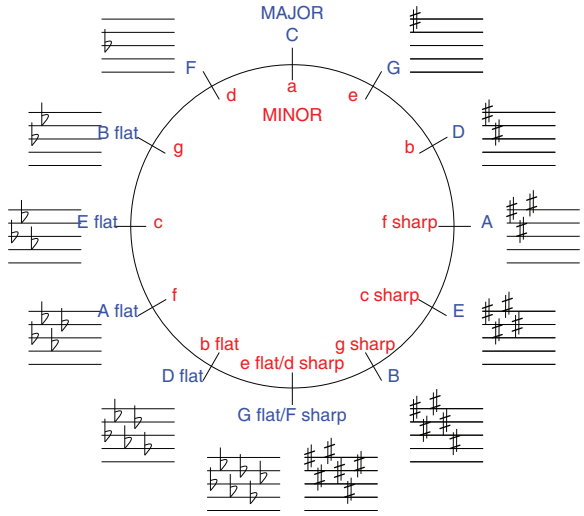

Fig. 5. Circle of fifth

the mapped MIDI notes are converted back in order to obtain the modified carrier frequencies that are used for synthesis. A dedicated MIDI note onset/offset detection is not required since the temporal characteristics are predominantly represented by the unmodified AM and thus preserved.

5. CONCLUSION

A novel method of multiband perceptual modulation analysis, processing and synthesis of audio signals has been proposed. Specifically, the analysis consists of a signal adaptive decomposition of the audio signal into sets of carriers, amplitude modulation (AM) and frequency modulation (FM). An efficient realization of the scheme using spectral mapping onto a perceptual scale has been given. Furthermore a suitable synthesis scheme for high quality audio rendering from modulation parameters has been proposed. The application of this method to the task of a key mode change of polyphonic pieces of music has been demonstrated.

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] A. Master, "Sinusoidal modeling parameter estimation via a dynamic channel vocoder model," *Proc. of the IEEE-ICASSP*, vol. 2, pp. 1857–1860, 2002.

[3] A. Röbel, "Adaptive additive modeling with continuous parameter trajectories," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 14, no. 4, pp. 1440–1453, 2006.

[4] Q. Li and L. Atlas, "Coherent modulation filtering for speech," *Proc. of the IEEE-ICASSP*, pp. 4481–4484, 2008.

[5] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 240–254, May 2000.

[6] S. Schimmel and L. Atlas, "Coherent envelope detection for modulation filtering of speech," *Proc. of the IEEE-ICASSP*, 2005.

[7] S. Schimmel, L. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," *Proc. of the IEEE-ICASSP*, vol. 4, pp. 605–608, 2007.

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

[9] 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.

[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.

[14] P. Daniel and R. Weber, "Psychoacoustical roughness: Implementation of an optimized model," *Acustica*, vol. 83, pp. 113–123, 1997.

[15] E. Zwicker and H. Fastl, *Psychoacoustics - Facts and Models*, Springer, Berlin - Heidelberg, 2. edition, 1999.