AN AMPLITUDE- AND FREQUENCY-MODULATION VOCODER FOR AUDIO SIGNAL PROCESSING

Sascha Disch

Laboratorium für Informationstechnologie,
Leibniz Universität Hannover, LFI

Hannover, Germany
disch@tnt.uni-hannover.de

Bernd Edler

Laboratorium für Informationstechnologie,
Leibniz Universität Hannover, LFI

Hannover, Germany
edler@tnt.uni-hannover.de

ABSTRACT

The decomposition of audio signals into perceptually meaningful modulation components is highly desirable for the development of new audio effects on the one hand and as a building block for future efficient audio compression algorithms on the other hand. In the past, there has always been a distinction between parametric coding methods and waveform coding: While waveform coding methods scale easily up to transparency (provided the necessary bit rate is available), parametric coding schemes are subjected to the limitations of the underlying source models. Otherwise, parametric methods usually offer a wealth of manipulation possibilities which can be exploited for application of audio effects, while waveform coding is strictly limited to the best as possible reproduction of the original signal. The analysis/synthesis approach presented in this paper is an attempt to show a way to bridge this gap by enabling a seamless transition between both approaches.

1. INTRODUCTION

Modulation analysis/synthesis systems that decompose a wideband signal into a set of components each comprising carrier, amplitude modulation, and frequency modulation information have many degrees of freedom since in general this task is an ill-posed problem. In order to define a useful representation one has to fix some basic conditions. Our approach is to satisfy the condition that the extracted information is perceptually meaningful and interpretable 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. For example methods that modify sub band magnitude envelopes of complex audio spectra and subsequently recombine them with their unmodified phases [1] for re-synthesis do not satisfy this constraint. Effectively, this leads to the design goal, that the extracted carrier information alone should allow for a coarse but perceptually pleasant and representative 'sketch' reconstruction of the audio signal and any successive application of AM and FM related information should refine this representation towards full detail and transparency.

The paper is structured as follows: First we briefly describe related technology in the field of the vocoder and audio modulation decomposition. After this we motivate our approach to modulation decomposition followed by a description of a modulation analysis/synthesis system. We then outline some ideas of modulation processing and present results obtained by application of these methods. These results are further accompanied by a listening test. Finally we suggest fields of application for this technology and conclude with a summary.

2. RELATED TECHNOLOGY

The vocoder (or 'VODER') was invented by Dudley as a manually operated synthesizer device for generating human speech [2]. Some considerable time later the principle of its operation was extended towards the so-called phase vocoder [3][4]. The phase vocoder operates on overlapping short time DFT spectra and hence on a set of sub band filters with fixed centre frequencies. The vocoder has found wide acceptance as an underlying principle for manipulating audio files. For instance, audio effects like time-stretching and pitch transposing are easily accomplished by a vocoder [5]. Since then, a lot of modifications and improvements to this technology have been published. Specifically the constraints of having fixed frequency analysis filters was dropped by adding a fundamental frequency (*'f0'*) derived mapping, for example in the 'STRAIGHT' vocoder [6]. Still, the prevalent use case remained to be speech coding/processing.

Another area of interest for the audio processing community has been the decomposition of speech signals into modulated components. Each component consists of a carrier, an amplitude modulation (AM) and a frequency modulation (FM) part of some sort. A signal adaptive way of such a decomposition was published e.g. in [7] suggesting the use of a set of signal adaptive band pass filters. In [8] an approach that utilizes AM information in combination with a *'sinusoids plus noise'* parametric coder was presented. Another decomposition method was published in [9] using the so-called 'FAME' strategy: here, speech signals have been decomposed into four bands using band pass filters in order to subsequently extract their AM and FM content. Most recent publications also aim at reproducing audio signals from AM information (sub band envelopes) alone and suggest iterative methods for recovery of the associated phase information which predominantly contains the FM [10].

Our approach presented herein is targeting at the processing of general audio signals hence also including music. It is similar to a phase vocoder but modified in order to perform a signal dependent perceptually motivated sub band decomposition into a set of sub band carrier frequencies with associated AM and FM signals each. We like to point out that this decomposition is

perceptually meaningful and that its elements are interpretable in a straight forward way, so that all kinds of modulation processing on the components of the decomposition become feasible.

3. BASIC PRINCIPLE

To achieve the goal stated above, we rely on the observation that perceptually similar signals exist. A sufficiently narrow-band tonal band pass signal is perceptually well represented by a sinusoidal carrier at its spectral *'centre of gravity'* (COG) position and its Hilbert envelope. This is rooted in the fact that both signals approximately evoke the same movement of the basilar membrane in the human ear [11]. A simple example to illustrate this is the two-tone complex (1) with frequencies f_1 and f_2 sufficiently close to each other so that they perceptually fuse into one (over-) modulated component

$$s_t(t) = \sin(2\pi f_1 t) + \sin(2\pi f_2 t) \tag{1}$$

A signal consisting of a sinusoidal carrier at a frequency equal to the spectral COG of s_t and having the same absolute amplitude envelope as s_t is s_m according to (2)

$$s_m(t) = 2\sin\left(2\pi \frac{f_1 + f_2}{2} t\right) \cdot \left|\cos\left(2\pi \frac{|f_1 - f_2|}{2} t\right)\right| \tag{2}$$

In Figure 1 (top and middle plot) the time signal and the Hilbert envelope of both signals are depicted. Note the phase jump of π in the first signal at zeros of the envelope as opposed to the second signal. Figure 2 displays the power spectral density plots of the two signals (top and middle plot).

Although these signals are considerably different in their spectral content their predominant perceptual cues – the 'mean' frequency represented by the COG, and the amplitude envelope - are similar. This makes them perceptually mutual substitutes with respect to a band-limited spectral region centred at the COG as depicted in Figure 1 and Figure 2 (bottom plots). The same principle still holds true approximately for more complicated signals.

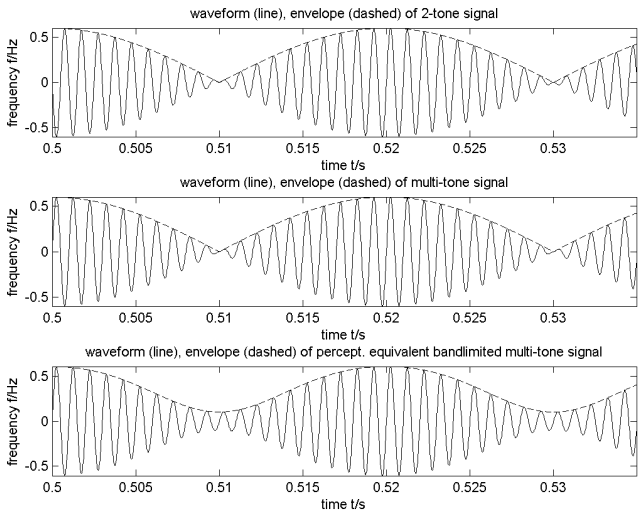

Figure 1: *Waveform and envelope of two-tone signal, multi-tone signal and appropriately band-limited multi-tone signal.*

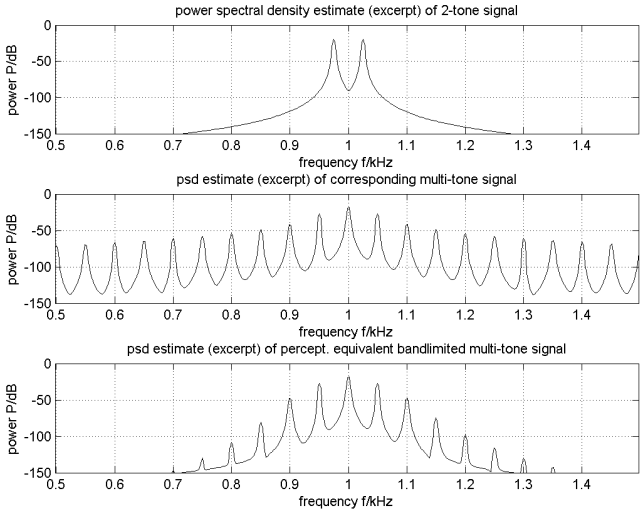

Figure 2: *Power spectral density of two-tone signal, multi-tone signal and appropriately band-limited multi-tone signal.*

4. THE PROPOSED SYSTEM

4.1. System Design Considerations

The proposed system consists of a modulation analysis/decomposition part, a modulation synthesis part and, if desired, a modulation processing unit. All units have been designed to support block based real-time computation. The processing of a certain time block is only dependent on parameters of previous blocks; no look ahead is required in order to keep the overall processing delay as low as possible.

4.2. Modulation Analysis

The decomposition into carrier signals and their associated modulation components is depicted in Figure 3.

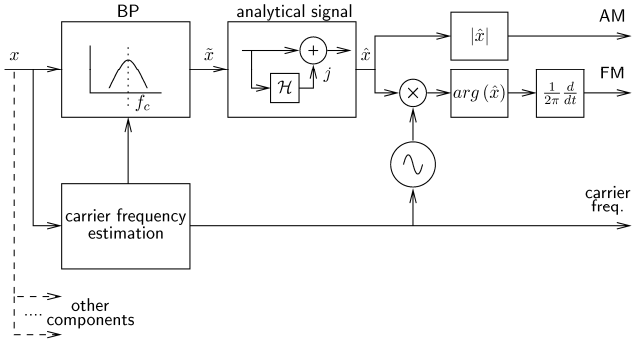

Figure 3: *Overview of the signal adaptive modulation decomposition scheme.*

In the picture the signal flow for the extraction of one component is shown. All other components are obtained in a similar fashion. The extraction is carried out on a block-by-block basis using a block size of N = 2^{14} at 48 kHz sampling frequency and ¾ over-

lap, roughly corresponding to a time interval of 340 ms and a stride of 85 ms. It consists of a signal adaptive band pass filter that is centred at a local COG [12] in the signal's DFT spectrum.

The local COG candidates are estimated by searching positive-to-negative transitions in the *CogPos* function defined in (3). A post-selection procedure ensures that the final estimated COG positions are approximately equidistant on a perceptual scale.

$$
\begin{aligned}
CogPos(k,m) &= \frac{nom(k,m)}{denom(k,m)} \\
nom(k,m) &= \alpha \sum_{i=-B(k)/2}^{+B(k)/2} \left(iw(i) \left| X(k+i,m) \right|^2 \right) \\
&\quad + (1-\alpha)\, nom(k,m-1) \\
denom(k,m) &= \alpha \sum_{i=-B(k)/2}^{+B(k)/2} \left(w(i) \left| X(k+i,m) \right|^2 \right) \\
&\quad + (1-\alpha)\, denom(k,m-1) \\
\alpha &= \frac{1}{\tau F_s}
\end{aligned}
\qquad (3)
$$

For every spectral coefficient index *k* it yields the relative offset towards the local centre of gravity in the spectral region that is covered by a smooth sliding window *w*. The width *B(k)* of the window follows a perceptual scale, e.g. the Bark scale. *X(k,m)* is the spectral coefficient *k* in time block *m*. Additionally, a first order recursive temporal smoothing with time constant τ is done.

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 [12]. The COG estimation window and the transition bandwidth of the resulting filter are chosen with regard to resolution of the human ear (*'critical bands'*). Here, a bandwidth of approx. 0.5 Bark was found empirically to be a good value for all kinds of test items (speech, music, ambience). Additionally, this choice is supported by the literature [13].

Subsequently, the analytic signal is obtained using the Hilbert transform of the band pass filtered signal and heterodyned by the estimated COG frequency. Finally the signal is further decomposed into its amplitude envelope and its instantaneous frequency (IF) track yielding the desired AM and FM signals. Note that the use of band pass signals centred at local COG positions correspond to the *'regions of influence'* paradigm of a traditional phase vocoder. Both methods preserve the temporal envelope of a band pass signal: The first one intrinsically and the latter one by ensuring local spectral phase coherence.

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 compromises with respect to the bandwidth of the filters which follow a perceptual scale but, at the same time, have to provide seamless spectral coverage. So the carrier frequency estimation and signal adaptive filter design turn out to be the crucial parts for the perceptual significance of the decomposition components and thus have strong influence on the quality of the re-synthesized signal. An example of such a compensative segmentation is shown in Figure 4.

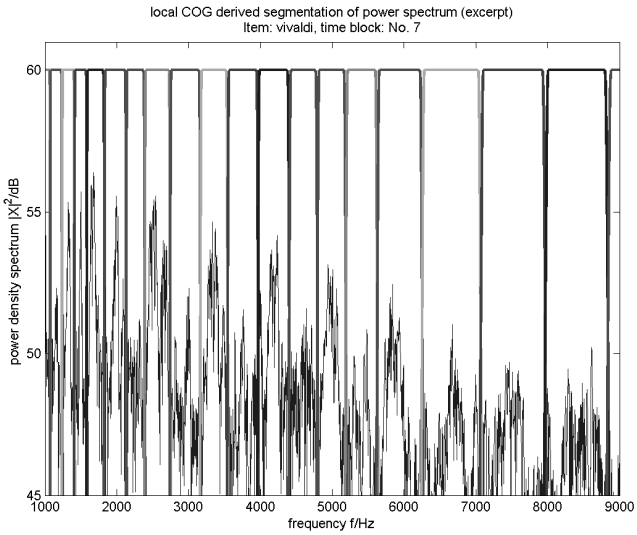

Figure 4: *Signal adaptive spectral segmentation.*

4.3. Modulation Synthesis

The signal is synthesized on an additive basis of all components. For one component the processing chain is shown in
Figure 5. Like the analysis, the synthesis is performed on a block-by-block basis. Since only the centred N/2 portion of each analysis block is used for synthesis, an overlap factor of ½ results. A component bonding mechanism is utilized to blend AM and FM and align absolute phase for components in spectral vicinity of their predecessors in a previous block. Spectral vicinity is also calculated on a bark scale basis to again reflect the sensitivity of the human ear with respect to pitch perception.

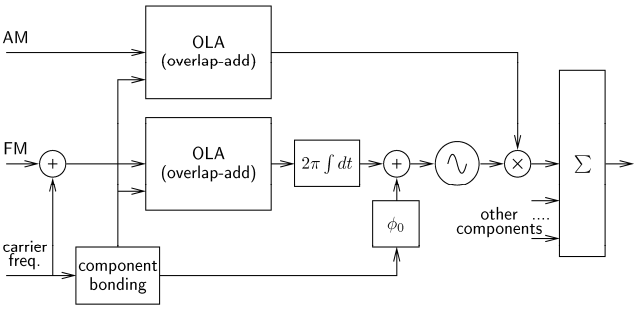

Figure 5: *Overview of the synthesis scheme.*

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

4.4. Modulation Processing

Having the modulation components at hand, new and interesting processing methods become feasible. A great advantage of the modulation decomposition presented herein is that the proposed analysis/synthesis method implicitly assures that the result of any modulation processing - independent to a large extent from the exact nature of the processing - will be perceptually smooth (free from clicks, transient repetitions etc.). A few examples of modulation processing are subsumed in Figure 6.

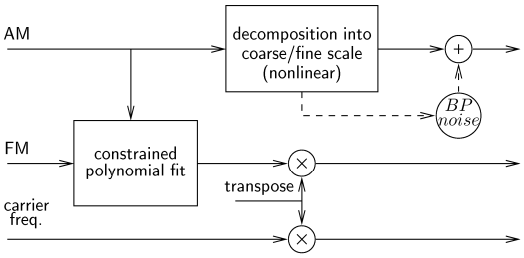

Figure 6: *Processing of the modulation components.*

For sure a prominent application is the *'transposing'* of an audio signal while maintaining original playback speed: This is easily achieved 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.

If only a subset of carriers corresponding to certain predefined frequency intervals is mapped to suitable new values, the key mode of a piece of music can be changed from e.g. minor to major or vice versa. To achieve this, the carrier frequencies are quantized to MIDI numbers which are subsequently mapped onto appropriate new MIDI numbers (using a-priori knowledge of mode and key of the music item to be processed). Lastly, the mapped MIDI numbers are converted back in order to obtain the modified carrier frequencies that are used for synthesis. Again, a dedicated MIDI note onset/offset detection is not required since the temporal characteristics are predominantly represented by the unmodified AM and thus preserved.

A more advanced processing is targeting at the modification of a signal's modulation properties: For instance it can be desirable to modify a signal's *'roughness'* [14][15] by modulation filtering. In the AM signal there is coarse structure related to on- and offset of musical events etc. and fine structure related to faster modulation frequencies (~30-300 Hz). Since this fine structure is representing the roughness properties of an audio signal (for carriers up to 2 kHz) [15][16], auditory roughness can be modified by removing the fine structure and maintaining the coarse structure.

To decompose the envelope into coarse and fine structure, nonlinear methods can be utilized. For example, to capture the coarse AM one can apply a piecewise fit of a (low order) polynomial. The fine structure (residual) is obtained as the difference of original and coarse envelope. The loss of AM fine structure can be perceptually compensated for - if desired - by adding band limited 'grace' noise scaled by the energy of the residual and temporally shaped by the coarse AM envelope.

Note that if any modifications are applied to the AM signal it is advisable to restrict the FM signal to be slowly varying only, since the unprocessed FM may contain sudden peaks due to beating effects inside one band pass region [17][18]. These peaks appear in the proximity of zero [19] of the AM signal and are perceptually negligible. An example of such a peak in IF can be seen in the signal according to formula (1) in Figure 1 in form of a phase jump of pi at zero locations of the Hilbert envelope. The undesired peaks can be removed by e.g. constrained polynomial fitting on the FM where the original AM signal acts as weights for the desired goodness of the fit. Thus spikes in the FM can be removed without introducing an undesired bias.

Another application would be to remove FM from the signal. Here one could simply set the FM to zero. Since the carrier signals are centred at local COGs they represent the perceptually correct local mean frequency.

5. RESULTS

5.1. Spectrogram Plots

In the following, some spectrograms are presented that demonstrate the properties of the proposed modulation processing schemes. Figure 7 shows the original log spectrogram of an excerpt of an orchestral classical music item (Vivaldi).

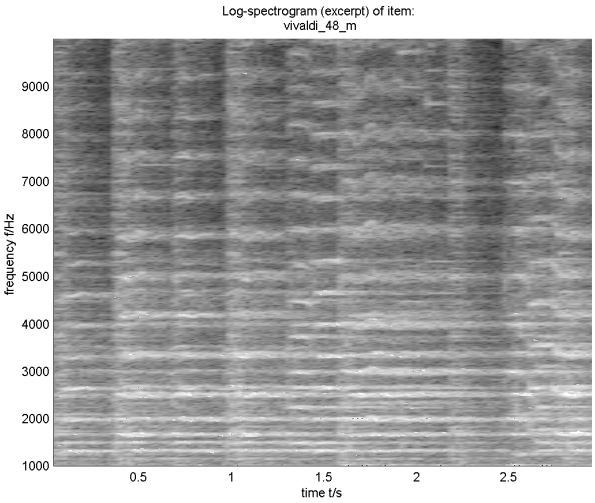

Figure 7: *Spectrogram of the original classical music item.*

Figure 8 to Figure 11 show the corresponding spectrograms after various methods of modulation processing in order of increasingly restored modulation detail. Figure 8 illustrates the signal reconstruction solely from the carriers. The white regions correspond to high spectral energy and coincide with the local energy concentration in the spectrogram of the original signal in Figure 7. Figure 9 depicts the same carriers but refined by non-linearly smoothed AM and FM. The addition of detail is clearly visible. In Figure 10 additionally the loss of AM detail is compensated for by addition of envelope shaped 'grace' noise which again adds more detail to the signal. Finally the spectrogram of the synthesized signal from the unmodified modulation components is shown in Figure 11. Comparing the spectrogram in Figure 11 to the spectrogram of the original signal in Figure 7 illustrates the very good reproduction of the full details.

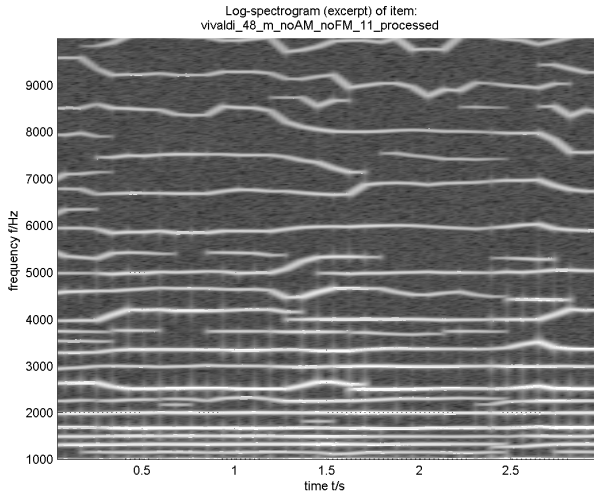

Figure 8: *Spectrogram of the synthesized carriers only.*

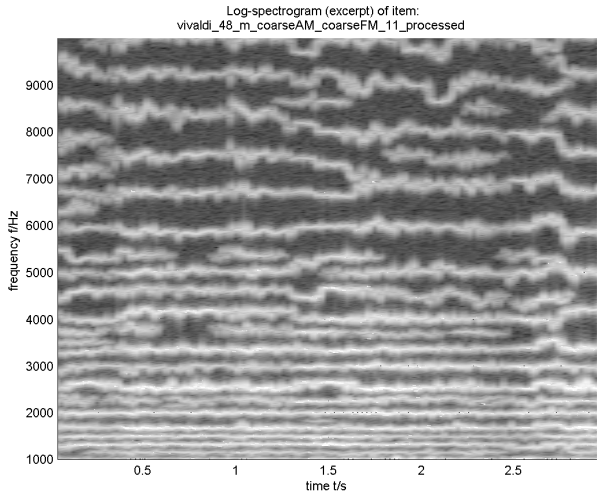

Figure 9: *Spectrogram of the carriers refined by coarse AM and FM.*

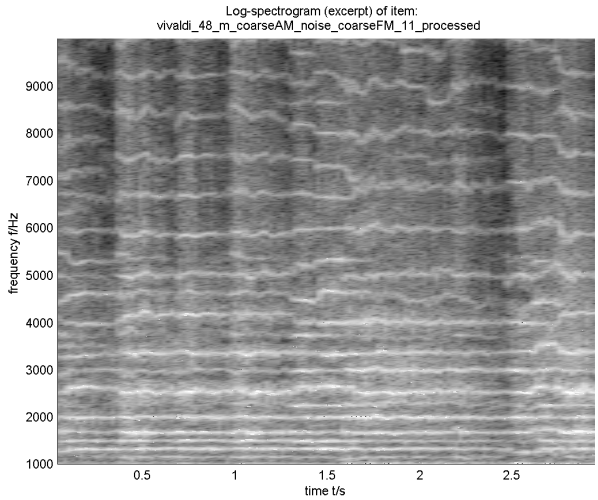

Figure 10: *Spectrogram of the carriers refined by coarse AM and FM, and added 'grace' noise.*

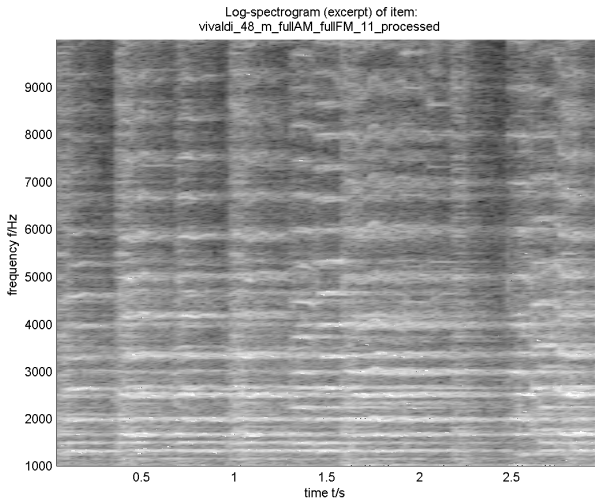

Figure 11: *Spectrogram of the carriers and unprocessed AM and FM.*

5.2. Listening test results

To evaluate the performance of the proposed method, a subjective listening test was conducted. The MUSHRA [21] type listening test was conducted using STAX high quality electrostatic headphones. A total number of 6 listeners participated in the test. All subjects can be considered as experienced listeners.

The test set consisted of the items listed in Table 1 and the configurations under test are subsumed in Table 2.

Table 1: *Test items.*

Item	*Description*
vivaldi_48_m	Classical orchestral music
brahms_48_m	Classical orchestral music
si01_48_m	Harpsichord/MPEG
si03_48_m	Pitch pipe/MPEG
sm01_48_m	Bagpipe/MPEG
sm02_48_m	Glockenspiel/MPEG
sm03_48_m	Plucked string/MPEG

Table 2: *Configurations under test.*

File name extension	*Processing method*
hidden_reference	Hidden reference
3k5Hz	Lower anchor
fullAM_fullFM	No modulation processing
fullAM_coarseFM	Coarse FM information
coarseAM_noise_coarseFM	Coarse FM and AM information with added 'grace' noise

The chart plot in Figure 12 displays the outcome. Shown are the mean results with 95% confidence intervals for each item. The plots show the results after statistical analysis of the test results for all listeners. The X-axis shows the processing type and the Y-axis represents the score according to the 100-point MUSHRA scale ranging from 0 (bad) to 100 (transparent).

From the results it can be seen that the two versions having full AM and full or coarse FM detail score best at approx. 80 points in the mean, but are still distinguishable from the original. Since the confidence intervals of both versions largely overlap, one can conclude that the loss of FM fine detail is indeed perceptually negligible as stated in section 4.4. The version with coarse AM and FM and added 'grace' noise scores considerably lower but in the mean still at 60 points: this reflects the graceful degradation property of the proposed method with increasing omission of fine AM detail information.

Most degradation is perceived for items having strong transient content like glockenspiel and harpsichord. This is due to the loss of the original phase relations between the different components across the spectrum. However, this problem might be overcome in future versions of the proposed synthesis method by adjusting the carrier phase at temporal centres of gravity of the AM envelope jointly for all components.

For the classical music items in the test set the observed degradation is statistically insignificant.

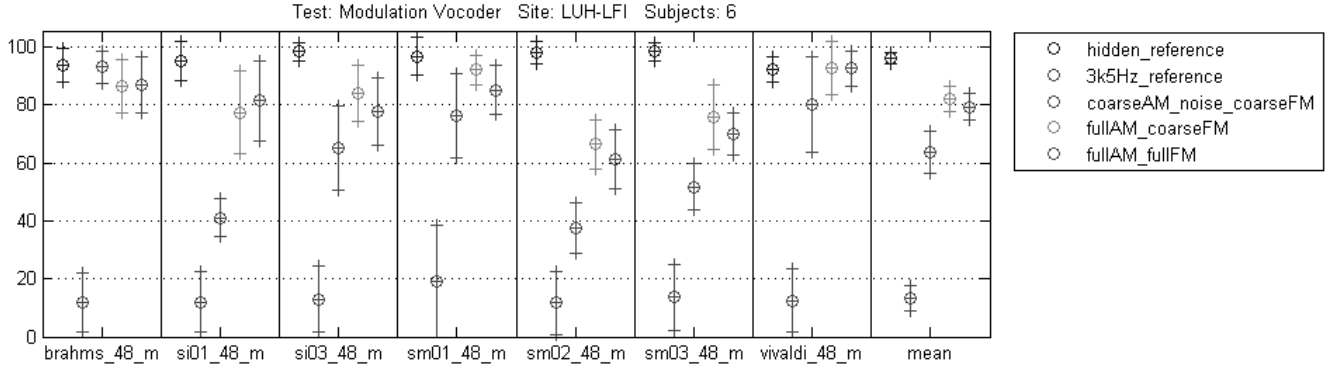

Figure 12: *Subjective audio quality test results (MUSHRA).*

6. APPLICATION

The analysis/synthesis method presented could be of use in different application scenarios: For audio coding it could serve as a building block of an enhanced perceptually correct fine grain scalable audio coder the basic principle of which has been published in [1]. With decreasing bit rate less detail might be conveyed to the receiver side by e.g. replacing the full AM envelope by a coarse one and added 'grace' noise.

Furthermore new concepts of audio bandwidth extension [20] are conceivable which e.g. use shifted and altered baseband components to form the high bands.

Improved experiments on human auditory properties become feasible e.g. improved creation of chimeric sounds in order to further evaluate the human perception of modulation structure [11].

Last not least new and exciting artistic audio effects for music production are within reach: either scale and key mode of a music item can be altered by suitable processing of the carrier signals or the psycho acoustical property of roughness sensation can be accessed by manipulation on the AM components.

7. SUMMARY

A proposal of a system for decomposing an arbitrary audio signal into perceptually meaningful carrier and AM/FM components has been presented, which allows for fine grain scalability of modulation detail modification. An appropriate re-synthesis method has been given. Some examples of modulation processing principles have been outlined and the resulting spectrograms of an example audio file have been presented. A listening test has been conducted to verify the perceptual quality of different types of modulation processing and subsequent re-synthesis. Future application scenarios for this promising new analysis/synthesis method have been identified. Our results demonstrate that the proposed method provides appropriate means to bridge the gap between parametric and waveform audio processing and moreover renders new fascinating audio effects possible.

8. ACKNOWLEDGMENTS

This research has been funded by Fraunhofer IIS, Erlangen/Germany.

9. REFERENCES

[1] M. Vinton and L. Atlas, "A Scalable And Progressive Audio Codec," in *Proc. of ICASSP 2001*, pp. 3277-3280, 2001

[2] H. Dudley, "The vocoder," in *Bell Labs Record*, vol. 17, pp. 122-126, 1939

[3] J. L. Flanagan and R. M. Golden, "Phase Vocoder," in *Bell System Technical Journal*, vol. 45, pp. 1493-1509, 1966

[4] J. L. Flanagan, "Parametric coding of speech spectra," *J. Acoust. Soc. Am.*, vol. 68 (2), pp. 412-419, 1980

[5] U. Zoelzer, *DAFX: Digital Audio Effects*, Wiley & Sons, pp. 201-298, 2002

[6] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. of ICASSP 1997*, vol. 2, pp. 1303-1306, 1997

[7] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," in *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 240-254, 2000

[8] M. Christensen et al., "Multiband amplitude modulated sinusoidal audio modelling," in *IEEE Proc. of ICASSP 2004*, vol. 4, pp. 169-172, 2004

[9] K. Nie and F. Zeng, "A perception-based processing strategy for cochlear implants and speech coding," in *Proc. of the 26th IEEE-EMBS*, vol. 6, pp. 4205-4208, 2004

[10] J. Thiemann and P. Kabal, "Reconstructing Audio Signals from Modified Non-Coherent Hilbert Envelopes," in *Proc. Interspeech (Antwerp, Belgium)*, pp. 534-537, 2007

[11] Z. M. Smith and B. Delgutte and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," in *Nature*, vol. 416, pp. 87-90, 2002

[12] J. N. Anantharaman and A.K. Krishnamurthy, L.L Feth, "Intensity weighted average of instantaneous frequency as a model for frequency discrimination," in *J. Acoust. Soc. Am.*, vol. 94 (2), pp. 723-729, 1993

[13] O. Ghitza, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," in *J. Acoust. Soc. Amer.*, vol. 110(3), pp. 1628-1640, 2001

[14] E. Zwicker and H. Fastl, *Psychoacoustics - Facts and Models*, Springer, 1999

[15] E. Terhardt, "On the perception of periodic sound fluctuations (roughness)," in *Acustica*, vol. 30, pp. 201-213, 1974

[16] P. Daniel and R. Weber, "Psychoacoustical Roughness: Implementation of an Optimized Model," in *Acustica*, vol. 83, pp. 113-123, 1997

[17] P. Loughlin and B. Tacer, "Comments on the interpretation of instantaneous frequency," in *IEEE Signal Processing Lett.,* vol. 4, pp. 123-125, 1997.

[18] D. Wei and A. Bovik, "On the instantaneous frequencies of multicomponent AM-FM signals," in *IEEE Signal Processing Lett.*, vol. 5, pp. 84-86, 1998.

[19] Q. Li and L. Atlas, "Over-modulated AM-FM decomposition," in *Proceedings of the SPIE*, vol. 5559, pp. 172-183, 2004

[20] M. Dietz, L. Liljeryd, K. Kjörling and O. Kunz, "Spectral Band Replication, a novel approach in audio coding," in *112th AES Convention*, Munich, May 2002.

[21] ITU-R Recommendation BS.1534-1, "Method for the subjective assessment of intermediate sound quality (MUSHRA)," *International Telecommunications Union*, Geneva, Switzerland, 2001.