AN ENHANCED MODULATION VOCODER FOR SELECTIVE TRANSPOSITION OF PITCH

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

In previous papers, the concept of the modulation vocoder (MODVOC) has been introduced and its general capability to perform a selective transposition on polyphonic music content has been pointed out. This renders applications possible which aim at changing the key mode of pre-recorded PCM music samples. In this paper, two enhancement techniques for selective pitch transposition by the MODVOC are proposed. The performance of the selective transposition application and the merit of these techniques are benchmarked by results obtained from a specially designed listening test methodology which is capable to govern extreme changes in terms of pitch with respect to the original audio stimuli. Results of this subjective perceptual quality assessment are presented for items that have been converted between minor and major key mode by the MODVOC and, additionally, by the first commercially available software which is also capable of handling this task.

1. INTRODUCTION

In previous papers [1][2][3], the concept of the modulation vocoder (MODVOC) has been introduced and its general capability to perform a meaningful selective transposition on polyphonic music content has been pointed out. This renders applications possible which aim at changing the key mode of pre-recorded PCM music samples [2]. In this paper, two enhancement techniques for selective pitch transposition by the MODVOC are proposed, specifically being *envelope shaping* and *harmonic locking.* The performance of the application and the merit of these techniques are demonstrated by results obtained from a specially designed listening test methodology which is capable to govern extreme changes in terms of pitch with respect to the original audio stimuli. Results of this subjective perceptual quality assessment are presented for items that have been converted between minor and major key mode by the MODVOC and, additionally, by the first commercially available software which can handle such a polyphonic manipulation task (*Melodyne editor* by *Celemony*). The software implements a technology which has been branded and marketed by the term *direct note access* (DNA). Up to best knownledge, there have been no scientific publications by Celemony related to the underlying technology of DNA processing. However, a patent has been published lately, presumably covering and thus disclosing the essential functionality of DNA [4]. It is worthwhile to note that while Melodyne editor initially performs an automatic analysis of the entire audio file before allowing for any manipulations the MODVOC operates on a block-by block basis thus potentially allowing for real-time operation.

The paper is organized as follows. First, in section 2 a recap on the underlying principle of modulation vocoder analysis and synthesis is given, followed by a description in section 3 how selective pitch transposition can be performed by the MODVOC. Next, in section 4 two techniques are introduced which have the potential to enhance the MODVOC in terms of perceptual quality, specifically *envelope shaping* and *harmonic locking*. In section 5, a listening test is described which assesses the perceptual quality of the MODVOC for selective pitch transposition. The test results are given in section 6 and, finally, we end with some conclusions.

2. BACKGROUND

2.1. Modulation vocoder (MODVOC)

The multiband modulation decomposition [2] dissects the audio signal into a signal adaptive set of (analytic) 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 *centers of gravity* (COGs). 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 [5].

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 [6][7]. 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 analytic 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

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

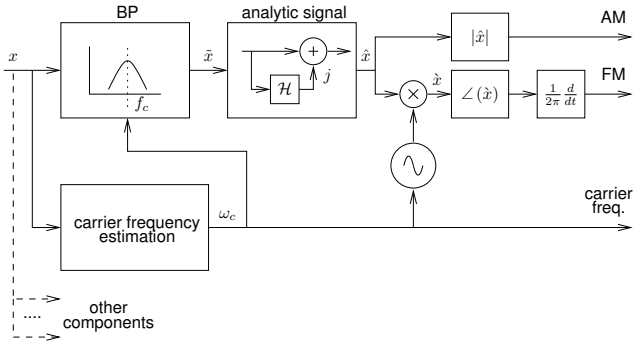

Figure 1: Modulation analysis.

of the multiband components is shown. All other components are obtained in a similar fashion. First, a broadband input signal x is fed into a bandpass filter that has been designed signal adaptively yielding an output signal $\widetilde{x}$. Next, the analytic signal is derived by the Hilbert transform according to Equation (1).

$$\widehat{x}\left(t\right) = \tilde{x}\left(t\right) + j\mathcal{H}\left(\tilde{x}\left(t\right)\right) \tag{1}$$

The AM is given by the amplitude envelope of $\widehat{x}$

$$AM\left(t\right) = \left|\widehat{x}\left(t\right)\right| \tag{2}$$

while the FM is obtained by the phase derivative of the analytic signal heterodyned by a stationary sinusoidal carrier with angular frequency ω_c. The carrier frequency is determined to be an estimate of the local COG. Hence the FM can be interpreted as the IF variation at the carrier frequency f_c.

$$\begin{aligned} \dot{x}\left(t\right) &= \widehat{x}\left(t\right) \cdot \exp\left(-j\omega_c t\right) \\ FM\left(t\right) &= \frac{1}{2\pi} \cdot \frac{d}{dt} \angle\left(\dot{x}\left(t\right)\right) \end{aligned} \tag{3}$$

The estimation of local COG and the signal adaptive design of the front-end filterbank is one of the key parts of the modulation analysis and has been described in a dedicated publication [3].

Practically, in a discrete time system, the component extraction is carried out jointly for all components as illustrated in Figure 2. The proposed processing scheme supports real-time computation. The processing of a certain time block is only dependent on parameters of previous blocks. Hence, no look-ahead is required in order to keep the overall processing delay as low as possible. The processing is computed on a block-by-block basis using e.g. 75 % analysis block overlap and application of a *discrete fourier transform* (DFT) on each windowed signal block. The window is a *flat top* window according to Equation (4). This ensures that the centered $N/2$ samples that are passed on for the subsequent modulation synthesis utilizing 50 % overlap are unaffected by the skirts of the analysis window. A higher degree of overlap may be used for improved accuracy at the cost of increased computational complexity.

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

Given the spectral representation, next a set of signal adaptive spectral bandpass weighting functions that is aligned with local

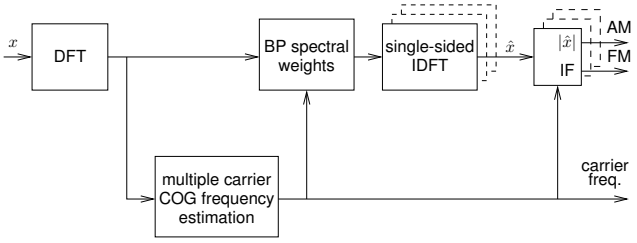

Figure 2: Implementation - Modulation analysis.

COG positions is calculated. After application of the bandpass weighting to the spectrum, the signal is transferred into 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. Given the discrete time bandpass signal, the estimation of the IF (3) is implemented by phase differencing as defined in Equation (5) where * denotes the complex conjugate. This expression is conveniently used since it avoids phase ambiguities and hence the need for phase unwrapping.

$$FM\left(n\right) = \angle\left(\dot{x}\left(n\right)\dot{x}\left(n-1\right)^\star\right) \tag{5}$$

2.3. Modulation synthesis

The signal is synthesized on an additive basis of all components. Successive blocks are blended by *overlap-add* (OLA) which is controlled by the bonding mechanism. The component bonding ensures a smooth transition between the borders of adjacent blocks even if the components are substantially altered by a modulation domain processing. The bonding does only take the previous block into account thus potentially allowing for real-time processing. It essentially performs a pair-wise match of the components of the actual block to their predecessors in the previous block. Additionally, the bonding aligns the absolute component phases of the actual block to the ones of the previous block.

For one component the processing chain is shown in Figure 3. In detail, first the FM signal is added to the stationary carrier frequency and the resulting signal is passed on to an OLA stage, the output of which is temporally integrated subsequently. A sinusoidal oscillator is fed by the resulting phase signal. The AM signal is processed by a second OLA stage. Next, the output of the oscillator is modulated in its amplitude by the AM signal to obtain the additive contribution of the component to the output signal. In a final step, the contributions of all components are summed to obtain the output signal y.

3. SELECTIVE PITCH TRANSPOSITION

A well-known application in the realm of audio effects is the global *transposition* of an audio signal. The processing required for this audio effect is a simple multiplication of the carriers with a constant transposition factor. By also multiplying the FM with the same factor it is ensured that, for each component, the relative FM modulation depth is preserved. Since the temporal structure of the input signal is solely captured by the AM signals it is unaffected by the processing. Global transposition changes the original key

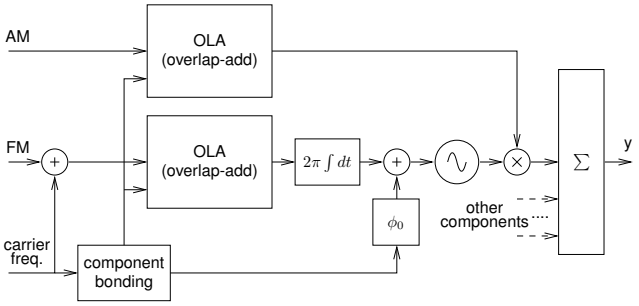

Figure 3: Modulation synthesis.

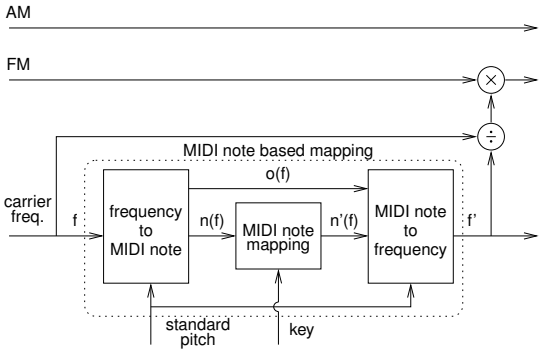

Figure 4: Selective transposition on MODVOC components. Carrier frequencies are quantized to MIDI notes which are mapped onto appropriate corresponding MIDI notes. Preservation of relative FM modulation depth by multiplication of the mapped components by the ratio of original and modified carrier frequency.

of a music signal towards a target key (e.g. from C major to G major) while preserving the original tempo.

However, due to the signal adaptive nature of the proposed modulation analysis, the modulation vocoder has the potential to go beyond this task. Now, even the transposition of *selected components* of polyphonic music becomes feasible, enabling applications which e.g. alter the key mode (e.g. from C major to C minor) of a given music signal [2]. This is possible due to the fact that each component carrier closely corresponds to the perceived pitch in its spectral region. If only carriers that relate to certain original pitches are mapped towards new target values, the overall musical character that is determined by the key mode is manipulated.

The necessary processing on the MODVOC components is depicted in Figure 4. Within the MODVOC decomposition domain, the carrier frequencies are quantized to MIDI notes which are subsequently mapped onto appropriate corresponding MIDI notes. For a meaningful reassignment of midi pitches and note names, a-priori knowledge of mode and key of the original music item is required. The AM of all components is not acted upon at all since these contain no pitch information.

Specifically, the component carrier frequencies f, which represent the component pitch, are converted to MIDI pitch values m according to Equation 6, where f_{std} denotes the standard pitch which corresponds to MIDI pitch 69, the note «A0».

Original note	Target note
C	C
D	D
E	Eb
F	F
G	G
A	Ab
B	Bb

Table 1: MIDI note mapping table for a mode transformation from C major to C natural minor. The mapping applies for the notes of all octaves.

$$\begin{aligned} m(f) &= 69 + 12 \cdot \log_2 \frac{|f|}{f_{std}} \\ n(f) &= \text{round}(m(f)) \\ o(f) &= m(f) - n(f) \end{aligned} \tag{6}$$

$$\begin{aligned} n &\rightarrow n' \\ f' &= f_{std} \cdot 2^{\left(n' + o(f) - 69\right)/12} \end{aligned} \tag{7}$$

Subsequently MIDI pitches are quantized to MIDI notes $n(f)$ and, additionally, the pitch offset $o(f)$ of each note is determined. By utilization of a MIDI note mapping table which is dependent on key, original mode and target mode, these MIDI notes are transformed to appropriate target values n'. In Table 1, an exemplary mapping is given for key of C from major to natural minor. Lastly, the mapped MIDI notes including their pitch offsets are converted back to frequency f' in order to obtain the modified carrier frequencies that are used for synthesis (Equation 7). Additionally, in order to preserve the relative FM modulation depth, the FM of a mapped component is multiplied by the individual pitch transposition factor which is obtained as the ratio of original and modified carrier frequency. A dedicated MIDI note onset/offset detection is not required since the temporal characteristics are predominantly represented by the unmodified AM and thus are preserved.

4. MODVOC ENHANCEMENTS

4.1. Envelope shaping

It has been stated in subsection 2.1 that the MODVOC processing preserves spectral coherence in the passband area surrounding the carrier locations. However, the broadband global spectral coherence is not preserved. For quasi-stationary signals this has only minor impact on the perceptual quality of the synthesized signal. If the signal contains prominent transients like e.g. drum beats or castanets, the preservation of global coherence can greatly improve the reproduction quality of these signals.

The preservation of global coherence can be improved by linear prediction in the spectral domain. Since long, similar approaches are utilized in audio codecs, for instance by the *temporal noise shaping* (TNS) tool [8] in MPEG 2/4 *advanced audio coding* (AAC). In [9], the combination of a high resolution time-frequency transform and spectral prediction is shown to essentially correspond to a signal adaptive transform.

Figure 5 outlines the integration of this technique into the MODVOC processing scheme. In the analysis, subsequent to the initial DFT of the input signal x, *linear prediction coefficients* (LPC) of

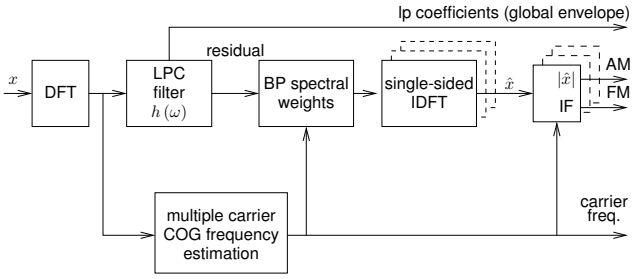

Figure 5: Modulation analysis with *envelope shaping*.

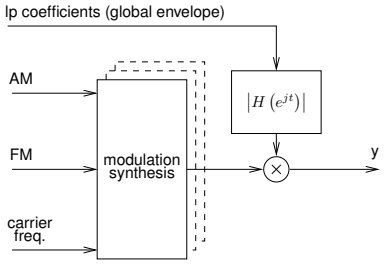

Figure 6: Modulation synthesis with *envelope shaping*.

a forward predictor along frequency having the impulse response $h(\omega)$ are derived by e.g. the autocorrelation method minimizing the prediction error in a least squares sense. Subsequently, the filter is applied to the spectral values and the residual signal is further processed by the MODVOC algorithm. The filter coefficients, representing the global envelope, are conveyed to the synthesis stage. In the synthesis, the global envelope, derived by evaluation of the prediction filter on the unit circle $\left|H\left(e^{jt}\right)\right|$, is restored by a multiplicative application of the same to the sum signal yielding the output signal y as illustrated in Figure 6.

4.2. Harmonic locking

Most instruments excite harmonic sounds consisting of a fundamental frequency part and its harmonics being approximately integer multiples of the fundamental frequency. Since musical intervals obey a logarithmic scale, each harmonic overtone resembles a different musical interval with respect to the fundamental (and its octaves). Table 2 lists the correspondence of harmonic numbers and musical intervals for the first seven harmonics.

Thus, in the task of selective transposition of polyphonic music content, there exists an inherent ambiguity with respect to the musical function of a MODVOC component. If the component originates from a fundamental is has to be transposed according to the desired scale mapping, if it is dominated by a harmonic to be attributed to a fundamental it has to be transposed together with this fundamental in order to best preserve the original timbre of the tone. From this there emerges the need for an assignment of each MODVOC component in order to select the most appropriate transposition factor.

To achieve this, the simple processing scheme introduced in section 3 has to be extended by a harmonic locking functionality. The harmonic locking examines all MODVOC components prior to transposition whether a component is to be attributed to a fundamental or is to be regarded as an independent entity. This is performed by an iterative algorithm. The flowchart of this algorithm is depicted in Figure 7. The algorithm evaluates frequency ratios, energy ratios and envelope cross correlations of a test component t with respect to all other components indexed by $i \in [0...I-1] \setminus t$ with I denoting the total number of components. The succession of test components during the iteration is determined by their A-weighted energy such that the evaluation order is in sequence of decreasing energy. The A-weighting [10][11] is applied to model the perceptual prominence of each component in terms of its loudness [12].

Harmonic number			Interval name
1	2	4	perfect unison (P1)
			minor second (m2)
		9	major second (M2)
			minor third (m3)
	5		major third (M3)
			perfect fourth (P4)
			tritone
	3	6	perfect fifth (P5)
			minor sixth (m6)
			major sixth (M6)
		7	minor seventh (m7)
			major seventh (M7)

Table 2: Harmonic numbers and related musical intervals with respect to the fundamental and its octaves.

The following features are examined by thresholding

- Harmonic carrier frequency match
- Harmonic carrier frequency missmatch
- Component energy
- Normalized amplitude envelope correlation at zero-lag

The frequency match and missmatch are definded according to Equation 8 with f_t being the test component carrier frequency and f_i being the component with index i. For the frequency match, all multiples greater than 1 are potential harmonics. A suitable threshold value for the frequency missmatch allowable for a potential harmonic is e.g. 22 Hz.

$$\begin{aligned} match_i &= \text{round}\left(\frac{f_i}{f_t}\right) \\ missmatch_i &= |f_t - (match_i \cdot f_t)| \end{aligned} \tag{8}$$

The a-weighted component energy ratio (Equation 9) of harmonics versus fundamental is required to be smaller than a predefined threshold reflecting the fact that for the vast majority of instruments the harmonics exhibit lower energy than the fundamental. A suitable threshold value, for instance, is the ratio of 0.6.

$$nrgRatio_i = \frac{nrg_i}{nrg_t} \tag{9}$$

The normalized zero-lag cross correlation of the envelope of the test component env_t and the envelope env_i of the component with index i is definded by Equation 10. This measure exploits the fact that a fundamental and its harmonics share a rather similar temporal envelope within the block length M. A suitable threshold value was determined to be 0.4 by informal experiments.

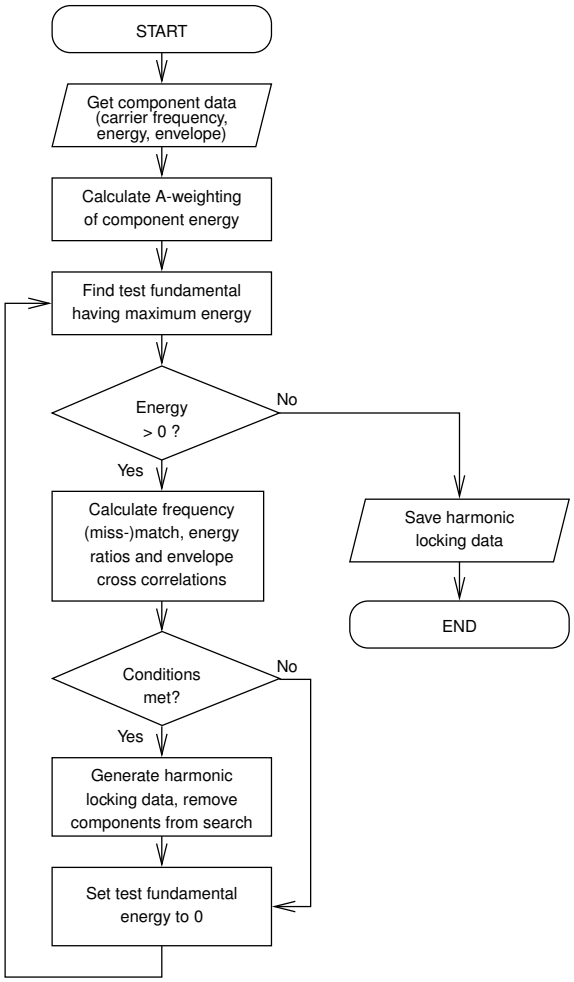

Figure 7: Flowchart of *harmonic locking*. Components that match the conditions of being harmonics of a test fundamental are iteratively labeled and removed from search space.

$$xcorr_i = \frac{\sum_{m=0}^{M-1} env_i(m) \cdot env_t(m)}{\sqrt{\sum_{m=0}^{M-1} env_i^2(m) \sum_{m=0}^{M-1} env_t^2(m)}} \quad (10)$$

After being examined, all components i that meet all of the threshold conditions are labeled as harmonics to be locked with respect to the test component and are subsequently removed from the search. Next, the test component is also excluded from further iterations by settings its energy to zero. The algorithm is repeated until all components have been assigned which is indicated by the maximum component energy being zero.

Figure 8 shows the enhanced processing scheme of selective transposition by the MODVOC incorporating harmonic locking. As opposed to Figure 4, only non-locked components enter the transposition stage while locked components are modified in a second stage by the same transposition factor that has been applied to their attributed fundamentals.

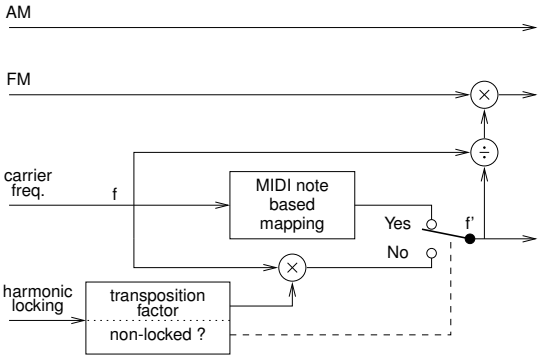

Figure 8: Enhanced selective transposition on MODVOC components using *harmonic locking*. Only non-locked carrier frequencies are quantized to MIDI notes which are mapped onto appropriate corresponding MIDI notes. Locked components are transposed by multiplication by the ratio of original and modified carrier frequency of their attributed fundamental.

5. LISTENING TEST

5.1. Scope

In order to evaluate the subjective audio quality of the modulation vocoder (MODVOC) for the application of selective pitch transposition and, moreover, the merit of the proposed enhancements to the basic MODVOC principle, a set of exemplary audio files has been assembled and processed accordingly. Additionally, the MODVOC technology is compared to the first commercially available audio software for polyphonic audio manipulation. *Melodyne editor* by *Celemony* which is on purchase since late 2009.

5.2. Methodology

Since the processing under test drastically alters the audio content of a signal, a direct comparison of original and processed signal - usually an inherent part in standard listening tests - is apparently not expedient in this case. In order to nonetheless measure the the subjective audio quality in a meaningful way, a special listening test procedure has been applied: the listening test set originates from symbolic MIDI data that is rendered into waveforms using a high quality MIDI expander. This approach enables a direct comparison of similarly altered audio files within the test and allows for an investigate into the effect of the selective pitch processing in isolation. The procedure of generating the test set is summarized in Figure 9. The original test signals are prepared in symbolic MIDI data representation (upper left). A second version of these signals is generated by a symbolic MIDI processing which resembles the target processing under test on the waveform rendered original audio (upper right). Subsequently, these signal pairs are rendered by a high quality MIDI expander into waveform (WAV) files (lower left and right). In the listening test, the waveform rendered from the processed MIDI file and several modulation vocoder (MODVOC) processed version of the rendered original MIDI file are compared (lower right). Additionally, the output of the MODVOC is compared to the output of Melodyne editor.

Apart from the MODVOC processed conditions, the test includes a condition obtained by using Melodyne editor which is currently the only commercial application to address this type of

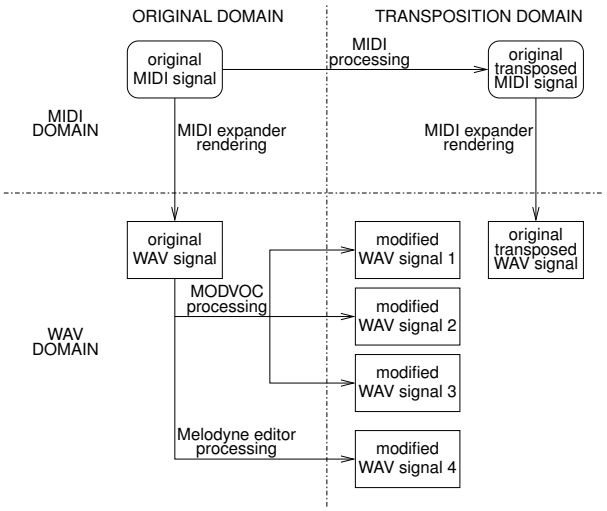

Figure 9: Procedure of generating the test set for evaluation of the subjective quality of MODVOC processing for the task of selective pitch transposition.

audio manipulation and thus can be seen as the industry standard. Melodyne editor initially performs an automatic analysis of the entire audio file. After the initialization phase, Melodyne suggests a decomposition of the audio file. By user interaction, this decomposition can be further refined. For the sake of a fair comparison to the MODVOC processing results, we chose to base our evaluation on the outcome of this automatic initial analysis since, apart from the a-priori knownledge of key and standard pitch, the MODVOC decomposition is fully automatic as well.

The listening test setup was based on a standard *MUltiple Stimuli with Hidden Reference and Anchor* (MUSHRA) test according to the ITU recommendation BS.1534 [13]. MUSHRA is a blind listening test. Only one person at a time is subjected to the test. For each item, the test presents all test conditions along with the hidden reference and a hidden lowpass filtered anchor to the listener in a time-aligned fashion. Hidden reference and lower anchor are include in order to check the listeners reliability. Switching between conditions while listening is permitted and so is setting a loop on arbitrarily selected partitions of the item as is suggested in the BS.1116-1 [14] and is applicable to MUSHRA tests as well. There is no limit of the number of repetitions the test subjects could listen to before rating the item and proceeding to the next test item, thus allowing for a very close comparison and thorough examination of the different conditions. The perceptual quality of the items is rated on a scale ranging from «excellent» (100 points) via «good» and «fair» up to «poor» (0 points). The sequence of test items is randomly ordered and moreover, the order of the conditions of each item is randomized as well.

5.3. Test items and conditions

The eight test items have been sourced from the MUTOPIA project[1], which provides free sheet music for public use. Suitable excerpts having an approximate duration of 20 seconds at maximum have been extracted from various pieces of classical music, containing both single instruments (e.g. G, E) and dense full orchestra parts (e.g. F). Also, dominant instrumental solo melodies accompanied by other instruments (for example C) are included in the test set. Besides the short-term quasi-stationary tonal parts, also percussive elements are contained in several items (onsets of plucked guitar in C and piano in G) which pose a special challenge on the transient response of the system under test. Table 3 lists all items of the set.

[1] http://www.mutopiaproject.org/

name	description	instruments	key mode
A	Violin Concerto, J. S. Bach, BWV1041	Orchestra	Amin
B	Eine kleine Nachtmusik, W. A. Mozart, KV525 Mv1	String Quartet	Gmaj
C	Berceuse, G. Fauré, Op56	Flute and Guitar	Emaj
D	Nocturno, F. Strauss, Op7	Horn and Piano	Dbmaj
E	Waltz, F. Carulli, Op241 No1	Guitar	Cmaj
F	Ein Musikalischer Spass, W. A. Mozart, KV522 Mv1	Horns, Violin, Viola, Cello	Fmaj
G	Ode an die Freude, L. V. Beethoven	Piano	Gmaj
H	Piano Trio, L. V. Beethoven, Op11 Mv3	Clarinet, Cello and Piano	Bbmaj

Table 3: Set of MIDI test items.

condition	name	description
1	*_reference	MIDI transposed original
2	*_3k5Hz_reference	3.5 kHz lowpass filtered original (anchor)
3	*_MODVOC	MODVOC
4	*_MODVOC_harm	MODVOC with harmonic locking
5	_MODVOC_harm_es	MODVOC with harmonic locking and envelope sharping
6	*_dna	Melodyne editor (DNA) fully automatic mode

Table 4: Test conditions.

The MIDI processing for obtaining the original transposed signals has been done in *Sonar8* manufactured by *Cakewalk*. The high quality waveforms rendering has been performed using *Bandstand* from *Native Instruments* in sound library version 1.0.1 R3. The MODVOC processing was evaluated in three different combinations with the two enhancement processing steps being *harmonic locking* (see subsection 4.2) and *envelope shaping* that has been introduced in section 4.1. For comparison to Melodyne editor, version 1.0.11 was utilized. All conditions are listed in Table 4.

5.4. Test setup

The subjective listening tests were conducted at the Fraunhofer IIS facility in an acoustically isolated listening lab that is designed to permit high-quality listening tests in an environment similar to an «ideal» living room. The listeners were equipped with *STAX* elec-

trostatic headphones that were driven from an *Edirol* USB sound interface connected to an *Apple MAC mini*. The listening test software was *wavswitch* by Fraunhofer IIS, operated in MUSHRA mode, providing a simple GUI to support the listener in performing the test. The listeners can switch between the reference (1) and the different conditions (2-7) during playout. Each listener can decide individually how long to listen to each item and condition. During the actual switching, the sound playout is muted. In the GUI, vertical bars visualize the rating attributed to each condition. We chose experienced listeners that are familiar with audio coding but as well have a musical background in order to get, on the one hand, an educated judgment on typical signal processing artifacts like pre- and post-echoes or dispersion of transients and on the other hand musical parameters such as spectral pitch, melody and timbre. In addition, the listeners were asked to provide their informal observations and impressions.

6. RESULTS

6.1. Absolute scores

15 subjects in total contributed to the test result, whereas one listener had to be post-screened due to failing to successfully identify the hidden original.

Figure 10 summarizes the results of listening test. The general quality level that, at this time, can be reached in selective transposition of pitch with either processing variant stretches from «fair» to «good» corresponding to a mean of approx. 60 MUSHRA points.

The application of harmonic locking in the MODVOC is favored for every item clearly indicating its merit. In the mean over all items the MOVOC featuring harmonic locking is rated significantly better than without. In six out of eight cases the additional application of envelope shaping shows a tendency of improving perceptual quality, which is also visible in the mean over all items, albeit no significance can be shown.

Furthermore, a tendency can be seen that the condition processed by Melodyne editor is mostly preferred by the listeners over the best rated MODVOC based processed condition except for items C and F where the MODVOC was rated better. However, none of this preferences is significant in the 95% confidence interval sense. Absolute score are widely accepted in the audio community to compare different conditions within a listening test. Nevertheless, to closer examine the above findings, in the following additionally score differences are considered.

6.2. Difference scores comparing plain MODVOC to enhanced MODVOC

Figure 11depicts the outcome based on score differences of the enhanced MODOVOC variants (conditions 4 and 5) with respect to the plain MODOVC (condition 3) results. Here, all enhanced MODVOC variants score considerably better than the plain MODVOC processing (all scores are well located above zero). There is significance in the 95% confidence sense for all items and conditions except for the application of harmonic locking alone in item A and C. Over all items, there is an significant improvement of ~8 MUSHRA points for harmonic locking and 10 points for the combination of harmonic locking and envelope shaping.

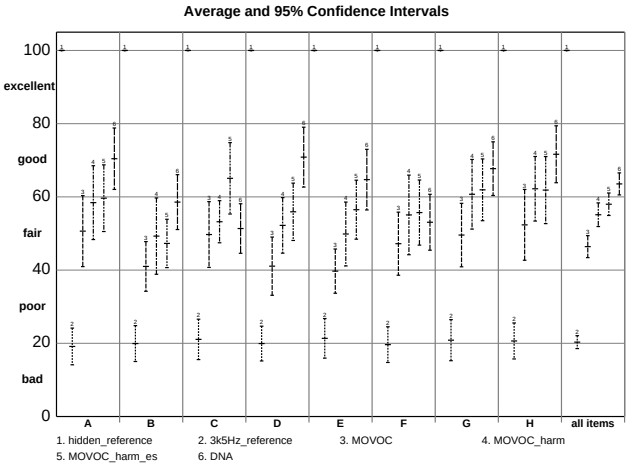

Figure 10: Absolute MUSHRA scores and 95% confidence intervals of listening test addressing selective pitch transposition. Items are according to table 3, test conditions are listed in table 4.

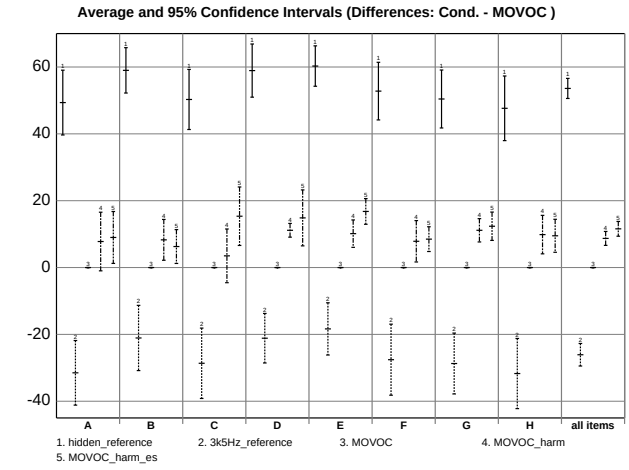

Figure 11: Difference MUSHRA scores with respect to condition 3 (MODVOC) and 95% confidence intervals of listening test addressing selective pitch transposition.

6.3. Difference scores comparing MODVOC to Melodyne editor

Figure 12 displays the test scores as score differences with respect to condition 6 (Melodyne editor). For item C, the MODVOC in condition 5 scores significantly better than Melodyne editor while condition 4, albeit being slightly positive, and condition 3 are inconclusive in a 95% confidence interval sense (confidence intervals overlap with 0).

For items B (condition 2), F, G (condition 5) also no significant conclusion can be drawn, but a tendency for better performance of the modulation vocoder can be seen also for item C in condition 4 and item F in conditions 4 and 5. In all other cases the modulation vocoder scores significantly worse than Melodyne editor.

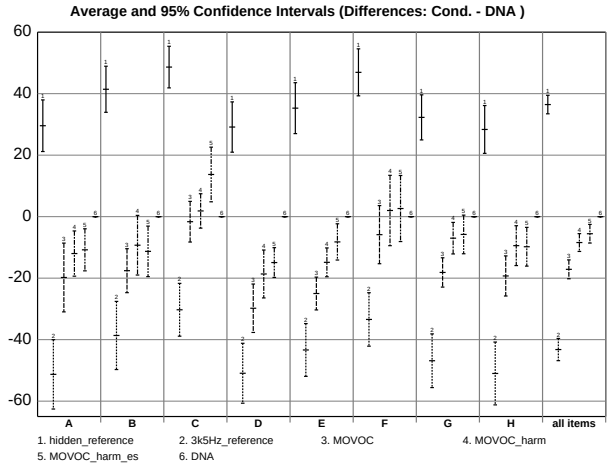

Figure 12: Difference MUSHRA scores with respect to condition 6 (DNA) and 95% confidence intervals of listening test addressing selective pitch transposition.

6.4. Discussion

The score reflects an overall quality judgment comprising aspects like unnatural sounding artifacts like degradation of transients by pre- or post-echos, pitch accuracy, correctness of melody and preservation of timbre. In order to interpret the results in more detail, the listeners were asked to note their informal observations alongside with noting the actual score. From these observations it can be concluded that the preservation of the timbre and absence of unnatural sounding artifacts were represented in the overall score to a higher degree than e.g. the goodness of melody preservation. Moreover, if a certain melody is unknown to the listener it seems that the test persons were not able to memorize the reference melody on short notice during the test and thus were unsure about the true melody. This can be an explanation of the higher overall rating of the Melodyne editor processed items, that have a higher fidelity with respect to preservation of timbre, especially of sounds originating from single instruments. However this comes at the prize of accidentally occurring severe melody errors that can happen presumably due to missclassification. The MODVOC is more robust in that respect since it does not predominantly rely on feature based classification techniques.

7. CONCLUSION

In this paper we have proposed two enhancement techniques for the modulation vocoder (MODVOC) for selective transposition of pitch. From the listening test results obtained for test signals rendered from MIDI it can be concluded that the perceptual quality of the plain MODVOC is indeed enhanced by *harmonic locking* and e*nvelope shaping*. Over all items, a increase of up to 10 MUSHRA points can be expected. In all cases, the main share of the improvement stems from the harmonic locking.

Moreover, the comparison of the MODVOC with a commercially newly available software (*Melodyne editor*) revealed that the general quality level that can be reached in selective pitch transposition, at this point of time, is located between «fair» and «good». For the majority of items the processing by Melodyne editor is preferred over the MODVOC processing, presumably due to a better preservation of timbre. Nonetheless, the MODVOC is more robust to missinterpretation of melody since it essentially does not mainly rely on classification decisions. However, this fact was only reflected in the informal comments by the listeners and was obviously not the major aspect of the overall perceptual quality ratings.

As opposed to the multi-pass analysis performed by Melodyne editor on the entire audio file prior to manipulation, the MODVOC is solely based on a single-pass blockwise processing potentially allowing for streaming or realtime operation scenarios. The restriction posed on the MODOVC hereby can be seen as another reason for the quality difference of both methods.

8. 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] S. Disch and B. Edler, "An iterative segmentation algorithm for audio signal spectra depending on estimated local centers of gravity," *12th International Conference on Digital Audio Effects (DAFx-09)*, 2009.

[4] P. Neubäcker, "Method for acoustic object-oriented analysis and note object-oriented processing of polyphonic sound recordings (ep2099024)," September 2009.

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

[6] 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.

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

[8] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (tns)," *101st AES convention, Los Angeles*, , no. Preprint 4384, 1996.

[9] J. Herre and J. D. Johnston, "A continuously signal-adaptive filterbank for high-quality perceptual audio coding," *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk*, 1997.

[10] ANSI, "Ansi standard s1.4-1983," 1983.

[11] ANSI, "Ansi standard s1.42-2001," 2001.

[12] H. Fletcher and W.A. Munson, "Loudness, its definition, measurement and calculation," *J. Acoust Soc Amer.*, vol. 5, pp. 82–108, 1933.

[13] ITU-R, "Method for the subjective assessment of intermediate sound quality (mushra)," 2001.

[14] ITU-R, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1994-1997.