ABSTRACT
Modern music production often uses pre-recorded pieces of audio, so-called samples, taken from a huge sample database. Consequently, there is an increasing demand to extensively adapt these samples to their intended new musical environment in a flexible way. Such an application, for instance, retroactively changes the key mode of audio recordings, e.g. from a major key to minor key, by a frequency selective transposition of pitch. Recently, the modulation vocoder (MODVOC) has been proposed to handle this task. In this paper, two enhancements to the MODVOC are presented and the subjective quality of its application to selective pitch transposition is assessed. Moreover, the proposed scheme is compared with results obtained by applying a commercial computer program, which became newly available on the market. The proposed method is clearly preferred in terms of the perceptual quality aspect «melody and chords transposition», while the commercial program is favored by the majority with regard to the aspect «timbre preservation».

Index Terms— Vocoder, Amplitude modulation, Frequency modulation, Transposition

1. INTRODUCTION

At present, audio effects are increasingly enhanced by incorporating knowledge about the (local) semantic content of the music signal to be processed. This implies the application of a certain effect in a context adaptive and selective manner. Such an application, for instance, retroactively changes the key mode of audio recordings, e.g. from a major key to minor key by a frequency selective transposition of pitch. While selective pitch transposing and time scaling audio effects are already commercially available and well established for the processing of monophonic content, polyphonic content still poses great challenges to modern signal processing and thus is subject of current scientific investigation activities.

Essentially, source separation could be used to decompose the polyphonic content into monophonic streams, which are then separately processed [1][2]. This includes an initial multiple fundamental frequency (f0) estimation step [3] and a subsequent grouping of spectral components into several estimated source objects. A precise grouping into source objects is, however, a tedious and error prone method, especially if the success of these methods strongly depends on the musical content of the individual item and on the reliability of the estimation and classification steps. In contrast to the aforementioned source separation method, the modulation vocoder (MODVOC) [4][5][6] follows an approach based on perceptual properties.

The paper is structured as follows: in Section 2 the MODVOC will be briefly introduced. In Section 3, two enhancements are proposed for selective pitch transposition using the MODVOC. Section 4 evaluates the overall subjective audio quality of the MODVOC for frequency selective pitch transposition application, and the merit of the proposed enhancements to the basic MODVOC principle. Additionally, the MODVOC is compared to a commercial audio software for polyphonic audio manipulation, «Melodyne editor» by Celemony [7]. Two main perceptual aspects, melody and chords transposition and timbre preservation, of the total subjective quality rating are assessed separately in detail, using preference tests for both synthetic and natural sound recordings.

2. MODVOC SELECTIVE PITCH TRANSPOSITION

In human auditory perception, the different sound contributions contained in a certain spectral region of a polyphonic mix are fused into a single joint sonic impression given a sufficiently narrow spectral distance of these contributions. The fundamental idea of the MODVOC is to jointly process signal components which are also perceived by humans as a sonic entity. Consequently, it is proposed to decompose polyphonic audio material into signal adaptive multiband components prior to a modulation analysis on each component. Most importantly, these multiband components must be aligned with spectral local centers of gravity (COG). In this way, the modulation parameters obtained by further analysis can be closely related to perceptual parameters. More precisely, the audio signal is decomposed into a set of signal adaptive carrier frequencies and their associated amplitude modulation (AM) and frequency modulation (FM). The carrier frequency, defined to be the local COG, corresponds to the mean frequency that is perceived by a listener due to the spectral contributions in that frequency region.

A conventional, global transposition changes the original key of a music signal towards a target key (e.g. from C major to G major) while preserving the original tempo. Due to its signal adaptive nature, the MODVOC can go beyond this task. 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 [4]. 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 MIDI pitches are mapped towards new target values, the overall musical character that is determined by the key mode is manipulated.

3. MODVOC ENHANCEMENTS

3.1. Envelope Shaping

Figure 1 shows the basic MODVOC analysis, processing and synthesis chain (black color). Firstly, the audio signal is transformed into frequency domain by the Discrete Fourier Transform (DFT). Given the spectral representation, a set of adaptive spectral bandpass weighting functions that is aligned with local COG positions is calculated. After applying the bandpass weighting to the spectrum, the signal is transferred into the time domain and the analytic signal is derived using Hilbert transform. These two processing steps can be efficiently combined by calculating a single-sided IDFT on each bandpass signal. Given the discrete time bandpass signal, the estimation of AM and IF is implemented by taking the magnitude and by phase differencing of the analytic signal, respectively. The MODVOC synthesis renders the output signal on an additive basis from all bandpass components. Each component is re-synthesized by modulating its carrier frequency by the associated AM and FM.

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 drum beats or castanets, the preservation of global coherence by temporal envelope shaping (ES) can greatly improve the reproduction quality of these signals [6]. The preservation of global coherence can be addressed by linear prediction in the spectral domain. Similar approaches have been utilized in audio codecs, for instance by the temporal noise shaping (TNS) tool [8][9] in MPEG 2/4 advanced audio coding (AAC). Figure 1 outlines the integration of this technique into the MODVOC processing scheme (in red color). In the analysis, subsequent to the initial DFT of the input signal $x$, linear prediction coefficients (LPC) of a forward predictor along frequency direction having the impulse response $h(\omega)$ are derived. Subsequently, this 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 $H(e^{j\omega})$, is re-stored by a multiplicative application of the same to the sum signal yielding the envelope shaped output signal $y$. The engagement of envelope shaping can be switched on or off signal adaptively depending on the prediction gain.

4. RESULTS

4.1. MUSHRA Test

Frequency selective pitch transposition drastically alters the audio content of a signal. A direct comparison of original and processed signal is apparently not expedient in this case. Thus, a special listening test procedure has been applied [6]: the listening test set originates from symbolic MIDI data that is manipulated in its key mode and rendered into waveforms using a high quality MIDI expander. Then, a ground truth obtained from manipulated and subsequently rendered MIDI files can be compared to waveform files which have been processed by the proposed methods in the waveform domain.

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 [10]. The perceptual quality of the items is rated on a scale ranging from «excellent» (100 points) via «good», «fair», «bad» and down to «poor» (0 points). The eight test items (A-H) listed in Table 1 have been sourced from the MUTOPIA project\(^1\). 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

\(^1\)http://www.mutopiaproject.org/
The MODVOC processing was evaluated in three different combinations with the two enhancement processing steps being harmonic locking (HL) and envelope shaping (ES). Additionally, the MODVOC is compared to a commercial audio software for polyphonic audio manipulation, Melodyne editor (DNA) by Celemony [7]. The subjective listening tests were conducted at the Fraunhofer IIS facility in an acoustically isolated listening lab equipped with STAX electrostatic headphones. In total, fifteen subjects contributed to the test result, where one listener had to be post-screened due to obviously failing to successfully identify the hidden original.

Figure 3 summarizes the results of the listening test. The perceptual quality for the items processed by selective pitch transposition ranges from «fair» to «good». The application of HL in the MODVOC is favored for every item clearly indicating its merit. In six out of eight cases the additional application of ES shows a tendency to improve 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. The score reflects overall quality judgment comprising aspects, like unnaturally sounding artifacts, such as the degradation of transients by pre- or post-echoes, pitch accuracy, correctness of melody and chords, and the 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 contributed to the overall score to a higher degree than the performance in terms of melody and chord transposition.

Figure 4 depicts the outcome based on score differences of the enhanced MODVOC variants (conditions 4 and 5), with respect to the plain MODVOC (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 only in item A and C.

### 4.2. Preference Tests on Quality Aspects

To investigate in depth which specific perceptual quality aspects may influence the overall rating of listeners, two main aspects denoted by «melody and chords transposition» and «timbre preservation» were further evaluated separately in a preference test. As opposed to Subsection 4.1, this test considers items which have already been transposed in the MIDI domain and subsequently rendered into waveforms. These waveforms were finally transposed back into their original key mode by both MODVOC and Melodyne editor. This modified procedure has been chosen to exclude any effect of listener failure to memorize melodies and chord progressions which otherwise would have sounded unfamiliar due to their changed key mode.

The four items of this test (1–4) are also listed in Table 1. The two alternative processing methods that were tested are MODVOC with HL and ES, and Melodyne editor (DNA). The original items were also available for supporting the actual decision process. Since synthetic items were tested, these non-transposed originals could be used as a ground truth with respect to both aspects, melody and chords transposition and timbre preservation. The listeners were asked to indicate their exclusive preference for one of the (randomized) conditions. Listeners were allowed to listen as often as desired to each of the test items. The listeners were instructed to focus on the following points. Quality of melody and chords transposition; do the melody and chords consistently sound as if originally played in that key throughout the item or are there «wrong notes» or bad intonations audible? Believability of timbre; is the timbre consistent
throughout the item and is it plausible for every instrument contained the mix? Twelve subjects participated in this test. All subjects were expert listeners and, at the same time, musicians capable of playing at least one instrument. Playout was from a Beyerdynamic DT-770 closed headphone. Figure 5 illustrates the listeners preference choice with respect to melody and chords transposition, and timbre preservation, respectively. For three out of four items, the MODVOC was preferred by the majority of listeners over Melodyne editor in terms of melody and chords transposition, for one item the preferences were on par. In contrast, in three out of four items Melodyne editor was preferred with regard to timbre preservation and for one item (solo piano), the MODVOC was clearly preferred.

4.3. Natural Recording Signals

Also, the perceptual quality of MODVOC and Melodyne editor for frequency selective pitch transposition application has been investigated for natural audio recordings in preference tests. The aim of this test is to compare the stability and performance of both algorithms with real-world signals that may exhibit natural inaccuracies in tuning or intonation due to the «human factor», and disturbing components originating from room reverb, ambience or tape hiss. Since no direct ground truth is available in this test, the listeners were instructed to consider the originals as an informal reference only for the timbre preservation property. Table 2 lists the set of test items. The same twelve persons that participated in the previously described synthetic items test also attended the natural items test. The results are depicted in Figure 6. In terms of melody and chords transposition, three of the MODVOC processed items are preferred, while item no.2 (Guitar/Orchestra) is equally preferred in both processing versions. The timbre preservation properties are consistently preferred in the Melodyne editor processed items.

5. CONCLUSION

In this paper, the application of the modulation vocoder (MODVOC) for frequency selective pitch transposition is proposed and, two enhancement techniques, envelope shaping (ES) and harmonic locking (HL), are presented. The subjective quality of the processing scheme is evaluated through formal listening tests. The test results lead to the conclusion that the plain MODVOC is truly enhanced by ES and HL. The overall audio quality of selective pitch transposition is scored by the listeners in a range spanning from «satisfactory» to «good». A comparison with frequency selective transposition results achieved by a commercially newly available software revealed that for the majority of items the commercial processing is rated with higher overall scores than the MODVOC processing. However, additional preference tests on the detailed quality aspects, «melody and chords transposition», and «timbre preservation» showed that for the majority of test items the MODVOC was preferred in terms of melody and chords transposition whereas Melodyne editor was chosen most often as the preference in terms of timbre preservation. It can be further speculated that the quality of melody and chords transposition, and the accuracy of timbre preservation possibly constitute opposing quality tradeoff aspects of any selective transposition scheme, at least for audio material containing complex polyphonic mixtures.

6. REFERENCES