A METHOD FOR THE MODIFICATION OF ACOUSTIC INSTRUMENT TONE DYNAMICS

Marco Fabiani, [*]

Dept. of Speech, Music and Hearing (CSC-TMH),
Royal Institute of Technology (KTH)
Stockholm, Sweden
himork@kth.se

ABSTRACT

A method is described for making natural sounding modifications of the dynamic level of tones produced by acoustic instruments. Each tone is first analyzed in the frequency domain and divided into a harmonic and a noise component. The two components are modified separately using filters based on spectral envelopes extracted from recordings of isolated tones played at different dynamic levels. When transforming from low to high dynamics, additional high frequency partials are added to the spectrum to enhance the brightness of the sound. Finally, the two modified components are summed and a time domain signal is synthesized.

1. INTRODUCTION

It is a well known fact that spectral characteristics of tones produced by acoustic instruments co-vary with sound level, when played at different dynamic levels [1]. In general, the timbre of louder tones is bright whereas that of softer tones is more dull. However, a more detailed analysis reveals that the exact timbral changes vary in a complex way depending on both instrument and register.

In [2] we proposed a prototype system aimed at a rule-based expressive modification of audio music recordings. We pointed out that in order to obtain a realistic performance, three main parameters must be controlled: tempo, articulation and dynamics. We presented a model for tempo modifications in [3]. During the development of the system it became clear that articulation and dynamics modifications must be applied on a note-by-note basis. For this reason, an analysis/synthesis approach was considered the best choice: with the assistance of a score file, each tone is identified and extracted as a group of partials tracks, following a sinusoidal plus residual (noise) model (see [4] for an overview). At this point, it is possible to modify the dynamic level of each tone in a realistic way by changing both its amplitude and its spectrum.

It has been shown in many studies (e.g. [5]) that the spectral centroid is a good estimator of a tone's brightness. It is usually defined as

$$f_{SC} = \frac{\sum_{n=0}^{N-1} f(n)H(n)}{\sum_{n=0}^{N-1} H(n)} \qquad (1)$$

where $f(n)$ and $H(n)$ are the center frequency and the magnitude of the nth bin of the signal's Fourier transform, respectively. A high spectral centroid indicates more energy in the higher frequencies of the spectrum, and thus a brighter timbre and, likely, a louder tone. A shelf-filter with varying cut-off frequency and slope or simple high- and low-pass filters are simple solutions to move the spectral centroid in a controlled way (e.g. [6, 7, 8]). However, we found two limitations when following this method. First, using a single filter on the complete signal affects not only the harmonic component of the signal but also the noise component. We have seen that, especially for certain instruments such as the flute, the harmonic and noise components change in different ways, and thus a better result can be obtained by handling them separately. Second, there is an inherent limit to how much the spectral centroid can be shifted upward. Consider the opposite case, a downward shift. Assume we have already separated the harmonic part of the signal. By using a low-pass filter, it is possible to lower the spectral centroid by reducing the energy of the higher partials, eventually until they fall under the noise floor. On the other hand, by using a high-pass filter we can only enhance those partials that were above the noise floor, and thus detected, in the first place. Consequently, turning a *piano* tone into a *forte* one can be very difficult, since the required high frequency partials are missing. In the worst case, applying a high-pass filter to the complete signal (harmonic and noise parts together), results solely in the amplification of the high frequency noise. Furthermore, simple shelve filters are usually too generic to reflect the differences between instruments or instrument's registers.

A synthesis technique employing a sinusoidal model and dynamic filters based on spectral envelopes extracted from instrument recordings is described in [9]. The authors' aim was not to modify existing tones, but to synthesize natural sounding ones. Nevertheless, their approach is similar to the one proposed in the present paper in that different filters are matched towards average spectral envelopes extracted from a number of recordings of isolated tones. These filters are then applied to sounds from a wavetable to generate different dynamic levels. In order to avoid the "problem of the missing partials", the wavetable is created using only the tones with the highest spectral centroid, i.e. the largest number of partials. Thus, only low-pass filters are actually used. In our case, since we want to modify existing tones, this approach is impracticable because we would have to replace the entire tone with a synthetic one. The primary advantage of modifying an existing tone is that a lot of the nuances that make it sound natural are already present. We want to preserve as much of the original sound as possible.

An alternative solution to the problem of the missing high frequency partials is the use of techniques known as "bandwidth extension". They are used in audio coding algorithms to enhance the quality of low bitrate encoded audio (e.g. mp3PRO, AAC+, see [10, 11, 12]). Our approach is based on the Accurate Spectral Replacement (ASR, [13]) technique. The spectrum is divided into its harmonic (sinusoidal) and noise components, and a smooth spec-

[*] This work was supported by BrainTuning (FP6-2004-NEST-PATH-028570) and SAME (FP7-ICT-STREP-215749) projects

tral envelope is computed. Only the lower frequency partials are preserved and directly transmitted to the decoder, while the higher frequency partials are estimated at the decoder side from a spectral envelope and the frequency of the lower partials.

In this paper, we present an algorithm for the modification of the dynamic level of different acoustic instruments (currently trumpet, clarinet, flute and violin, but it can be extended to other instruments, provided we record suitable samples). The signal is first decomposed into its harmonic and noise components. Different filters, matched to spectral envelopes extracted from recordings of isolated tones, are applied to the two parts of the signal. In case of a transformation from a low to a high dynamic level, the missing high frequency partials are synthesized and added to the spectrum. Finally, the modified harmonic and noise components are summed and a new time-domain signal is synthesized.

2. FILTER ESTIMATION

2.1. Recording of instrument samples

To extract the spectral envelopes needed to create the filters, some available sound libraries were first tested (e.g. MUMS[1], MIS[2]). These recordings proved unsuitable for our purposes since they did not provide the calibration data needed in order to compare different players of the same instrument, or different exemplars of the same instrument. Furthermore, we wanted to have a wide range of dynamic levels as well as a wide range of pitches. Thus, we decided to record our own samples.

For each instrument, at least two expert musicians were asked to perform a number of tones at different dynamic levels on their own instrument. They were asked to play each combination of pitch and dynamic level twice, to reduce the risk of unusable samples. A summary of the dynamic level and pitch ranges for each instrument is presented in Table 1. The musicians were instructed to play the *pp* as the softest level they could produce, the *mf* as a “standard” level and the *ff* as the loudest, yet still comfortable, level. They were also asked to play each tone for 2-3 seconds, and to maintain a steady level (sustain with no vibrato) for at least 1 second. The recordings were performed in a semi-anechoic room using a Bruel&Kjaer type 4003 microphone at a sampling rate of 96kHz. A calibration signal (1kHz sine wave, 94dB at 1 cm distance) generated using a Extech Sound Level Calibrator 447766 was recorded before each session. The microphone was placed at a distance of approximately 0.5 m from the instrument.

Table 1: *Summary of dynamic range and tone range of the recorded samples.*

Instrument	Dynamic levels	Tone range (MIDI)
Violin	*pp, p, mf, f, ff*	G3-E6
Trumpet	*pp, p, mp, mf, f, ff*	E3-A#5
Clarinet	*pp, p, mp, mf, f, ff*	D3-A#5
Flute	*pp, p, mf, f, ff*	C4-C7

2.2. Estimation of loudness

For the purpose of this study, it was not important to obtain absolute values of sound level for each tone. Instead, we were more interested in the perceptual differences in loudness and timbre between different dynamic levels. Thus, we decided to estimate the loudness of each sample using a simple loudness model, the ITU-R standard BS.1770-1 [14]. According to the specifications, in the first stage, the signal is pre-filtered to account for the acoustic effects of the head; in the second stage, an RLB (Revised Low-frequency B) weighting curve modeled as a simple high-pass filter is applied; finally, the mean-square energy $\overline{z}$ is measured. The loudness (mono signal) is then obtained as

$$L = -0.691 + 10\log_{10}(\overline{z}/\overline{z}_{ref}) \tag{2}$$

where $\overline{z}_{ref}$ is the mean-square energy of the calibration signal.

The loudness estimation was performed in the frequency domain. We first computed $F(n)$, the Short Time Fourier Transform, using the Odd-DFT transform (frame length ≈ 9 ms, overlap 75%), where n is the analysis frame index. The Odd-DFT, which is briefly explained in [15], was used to be fully compliant with the ASR method described in [13], as mentiod in section 1. For each frame, the impulse response of the two filters of the loudness model (i.e. head response and RLB) was multiplied by the magnitude of $F(n)$. The squared amplitudes of the resulting spectrum bins were summed, and the result divided by the number of frequency bins to obtain the mean-square energy $z(n)$ for each frame. This resulted in an amplitude curve that was used to detect the onset and offset of the tone and the boundaries of the sustain part.

Tone onsets and offsets were detected using a threshold set at

$$A_{thr} = \frac{max(z(n)) - min(z(n))}{2} \tag{3}$$

The sustain was defined as the longest part of the signal (not exceeding 1 sec) where the derivative of the amplitude curve $z(n)$ did not exceed ± 0.5.

The average loudness as defined in equation 2 was computed for both the entire tone and the sustain part, where $\overline{z}$ is the average of $z(n)$ over the frames comprised between the tone onset and offset, and between the sustain limits, respectively. The sustain loudness values were then used to normalize the samples to the same level, as well as to compute the instrument's dynamic range. Figure 1 shows the average loudness of the sustain part for one of the recorded trumpet players.

2.3. Partials extraction

Partials detection and extraction was not fully automatic, but based on the *a priori* knowledge of the approximate f_0, and on the assumptions that the pitch of the tone was held constant over the entire sample and that the spectrum was perfectly harmonic.

For each analysis frame, a peak detection algorithm was used to extract all possible partials from the magnitude of $F(n)$. For each detected peak, its frequency was estimated as described in [15]. The method interpolates the real frequency of a peak in the spectrum from the amplitude of the peak's frequency bin and the two nearest bins, and by also considering the shape of the analysis window. For each frame, the value of f_0 was corrected using the estimated frequency of the peak closest to the expected fundamental frequency. If no peaks in a range of ±35 *cents* around the

[1] McGill University Master Samples, http://www.music.mcgill.ca/resources/mums/html/

[2] University of Iowa Instrument Samples, http://theremin.music.uiowa.edu/MIS.html/

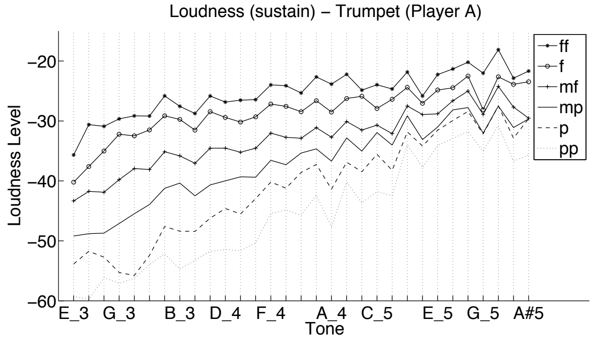


Figure 1: *Loudness values for each tone played by trumpet player A (average of the two collected samples).*

expected f_0 were found, the frame was considered to be containing only noise.

In the first phase, all the peaks that were not harmonically related to the extracted f_0, with an error margin of ±25 *cents*, were eliminated. A noise floor was then computed by taking the spectral envelope of the signal minus the detected partials. In the second phase all the peaks deemed to be harmonic, but that had an amplitude below the estimated noise floor, were eliminated. The noise component of the signal was finally obtained by subtracting the harmonic components from the original signal.

2.4. Spectral envelope estimation

The spectral envelopes used to define the spectrum modification filters were computed for both the harmonic and the noise components using only the sustain part of the tone. The algorithms and the `Matlab` code described in [16], and in particular the LS method, were used because they facilitate the extraction of envelopes based only on specific frequency bins. The algorithm uses a cepstral parametrization of the envelope. The parameters are estimated iteratively by minimizing an error function. For each recorded tone (i.e. two samples for each pitch-dynamics combination), we computed the average spectral envelope of both components of the signal.

As expected, the spectrum of different tones played at the same level had sometimes very different characteristics, especially when the pitch difference increased. For the trumpet, for instance, a sudden change in the spectrum could be seen at the register change. For the violin, each string had different spectral characteristics. We thus decided to group the samples and produce several spectral envelopes for each instrument. The trumpet and clarinet samples were grouped into five half octaves, while the violin samples were grouped according to the string they had been played on. The spectral envelopes for all the samples of one group were then averaged. Finally, a differential spectral envelope was computed by subtracting from all the envelopes the envelope of the *ff* samples (as in [9]). Figure 2 shows the extracted average spectral envelopes (harmonic components only) for the first four semi-octaves of a trumpet, while figure 3 shows the average spectral envelopes for the tones played on the four strings of a violin.

3. SIGNAL MODIFICATION

Once all the spectral envelopes, and thus the response of the dynamic modification filters, had been obtained, we could proceed to the modification of the recorded samples. In an automatic modification system, such as the one described in [2], the dynamic level of each tone in the original signal need to be estimated. This is a relatively difficult task since the conditions in which the recordings were made are usually not known (distance from the microphone, calibration, possible post-production modifications) and thus the sound level alone is not a reliable estimator of the tone's dynamic level. For the purpose of this study, we will nevertheless assume that we have already obtained this information.

3.1. Filtering

Through the passages described in section 2.4, an "envelopes matrix" $H(d, p)$ was created, where each line d corresponded to a dynamic level and each column p to a group of pitches. Suppose we want to transform the tone T, belonging to pitch group p_T, from level d_1 to d_2: we will first select the two correct envelopes $H(d_1, p_T)$ and $H(d_2, p_T)$ and compute their difference (the envelopes being expressed in dB). Thus, the frequency response of the required modification filter is

$$H_{mod} = H(d_2, p_T) - H(d_1, p_T) \qquad (4)$$

It is important to point out that H_{mod} is not the difference between the tone to be modified and the target envelope: this would simply turn the tone into the target. Instead, H_{mod} is the difference between the target reference envelope and the reference envelope with the same dynamic level as the tone to be modified. Different exemplars of the same instrument have different spectral envelopes (because of e.g. different internal resonances, different materials, different construction techniques). Nevertheless, if we assume these specificities to be constant throughout the dynamic range of an instrument, by taking H_{mod} as in equation 4, we remove these specificities. This implies that, in theory, the modified tone does not inherit the characteristics of the reference samples, but maintains the character of the original instrument.

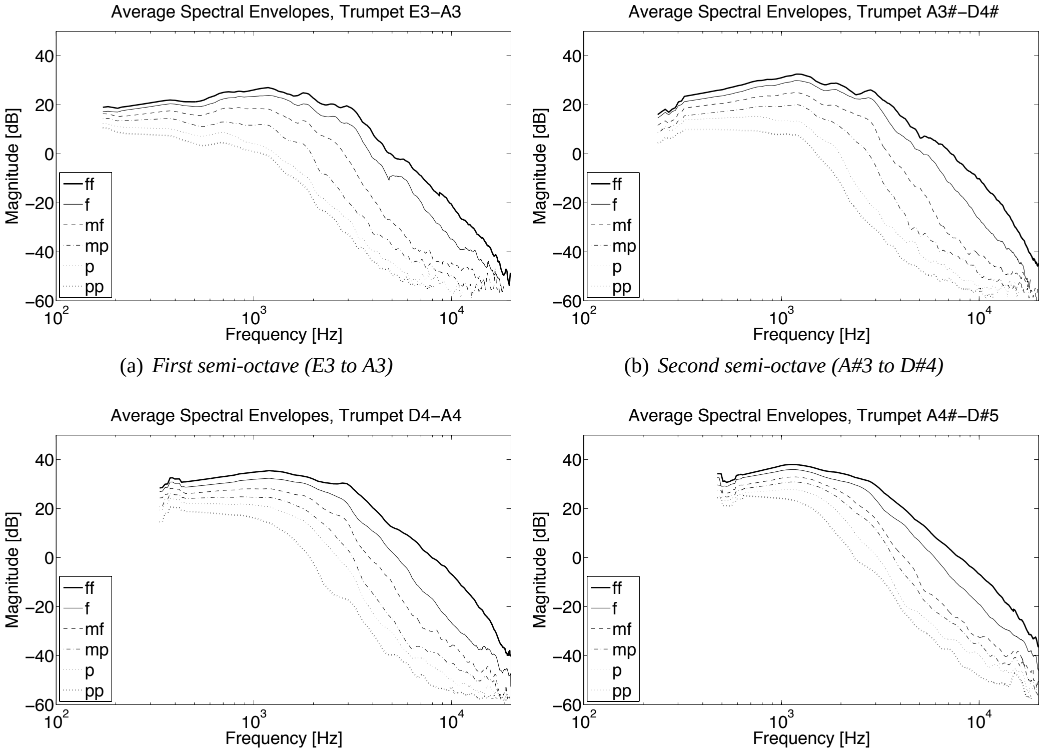
The magnitude of the modified spectrum is then

$$|\hat{F}(n)| = a \cdot |F(n)| \cdot H_{mod} \qquad (5)$$

where a is a gain factor calculated as the difference between the loudness of the samples used to generate the spectral envelopes (see section 2.2). The gain is necessary because the spectral envelopes were computed from samples normalized for loudness. The gain factor can also be omitted, and the sound level modification left to the user in a later stage. Time-varying modifications (i.e. *crescendo* and *decrescendo*) are possible by interpolating envelopes between two levels over time.

As mentioned in section 1, for some instruments, such as the flute, the harmonic and the noise parts of the signal show different behaviors. For this reason, two sets of spectral envelopes were estimated, one from the harmonic and one from the noise components of the samples. To obtain more realistic modifications, equations 4 and 5 were used on the harmonic and the noise spectra separately. The results were then summed to obtain the complete modified spectrum.

The method described until now is very similar to that described in [9]. The main limitation we found by using this approach is that only a reduction of the dynamic level is effective,



(a) *First semi-octave (E3 to A3)*

(b) *Second semi-octave (A#3 to D#4)*

(c) *Third semi-octave (E4 to A4)*

(d) *Fourth semi-octave (A#4 to D#5)*

Figure 2: *Spectral envelopes (harmonic components only) for the tones played by trumpet player A, grouped in 4 semi-octaves*

while an increase suffers from a lack of high-frequency partials. Thus, we decided to use the concept of bandwidth extension, employed in audio signal coding, to extend the capabilities of our modification technique.

3.2. Bandwidth extension

Bandwidth extension was used only in those cases where an increase in dynamic level was required. During the partials extraction phase described in section 2.3, we had already estimated the f_0 in each analysis frame. It is thus fairly straightforward to identify the upper limit of the detected partials, and synthesize new ones above that limit, where needed. New peaks in the spectrum were added using the algorithm described in [13]. From the frequency and the amplitude of the partial, a three-point interpolated peak can be obtained, to be added to the magnitude $F(n)$, by inverting the formulas used for frequency estimation (see section 2.3). The amplitude of the partial was estimated from the continuation of the spectral envelope of the partials already present, after the multiplication with the filter's response. The spectral envelope was extended until it fell under the noise floor (see figure 4)

3.3. Signal resynthesis

The modified $F(n)$ could now be inverted back to the time domain. A problem at this point regards the phase response required to invert the transform. It is indeed possible, using the method from [13], to compute the phase response of the three-point synthesized peak. For this, though, the amplitude and frequency of the peak are not sufficient, but a phase value is also required. A solution to the problem is to arbitrarily set an initial phase for the synthetic partial track, and to compute the phase increment at each successive analysis frame based on its frequency. Considering the specific case of the application we proposed in [2], this approach seemed difficult to apply, since various other transformations to the spectrum are required apart from dynamics modifications. We

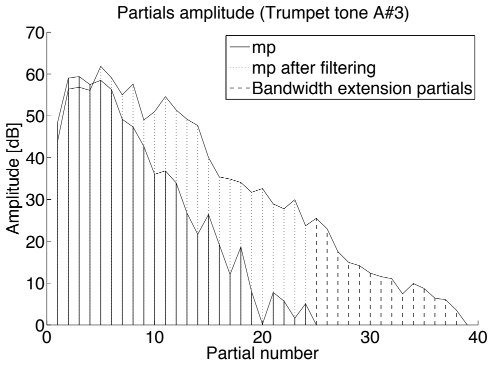
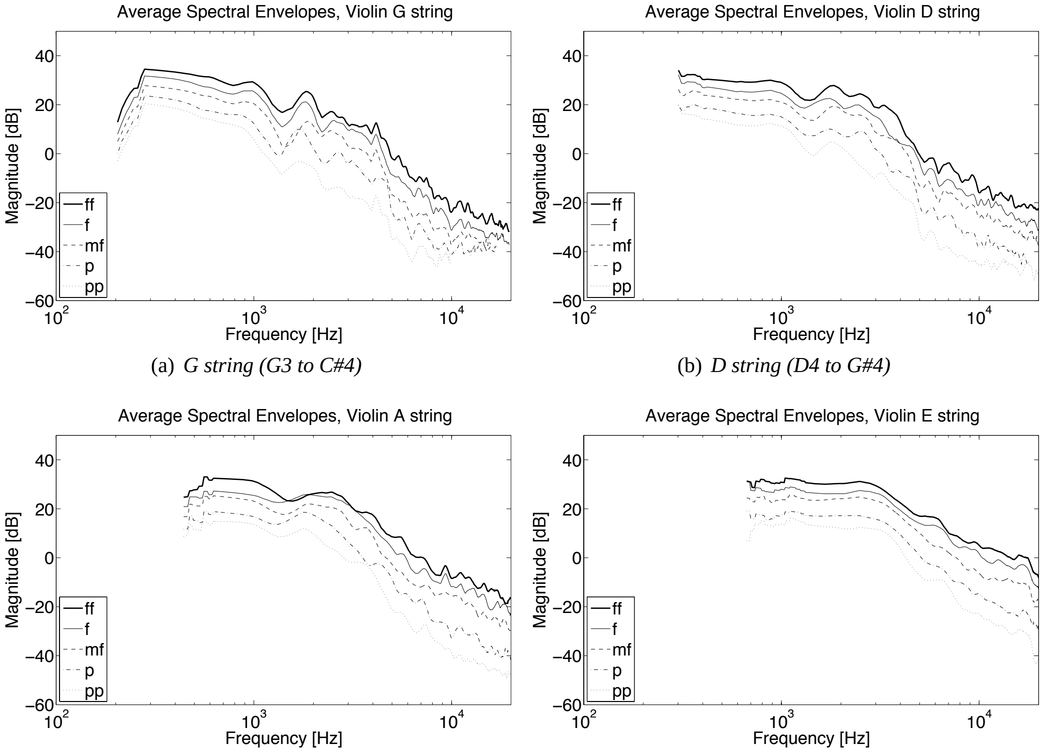


Figure 4: *Partials amplitudes of a* mp *trumpet tone (solid line), the same partials after filtering (dotted line), and the additional partials added by the bandwith extension technique (dashed line).*



(a) *G string (G3 to C#4)*

(b) *D string (D4 to G#4)*

(c) *A string (A4 to D#5)*

(d) *E string (E5 to E6)*

Figure 3: *Spectral envelopes (harmonic components only) for the tones played on the four strings of the violin by player A*

decided instead to completely discard the phase response, and construct a new one based solely on the magnitude of $F(n)$. For this purpose, a modified version of the Real Time Iterative Spectrum Inversion with Look-ahead (RTISI-LA) algorithm described in [17], was chosen. The original algorithm was extend to preserve transients in case of time-scale modifications, as described in [3].

4. CONCLUSIONS

A method is described for a natural sounding modification of the dynamic level of acoustic instrument sounds. It is based on previous methods, such as [9], that use filters matched to spectral envelopes extracted from instrument samples, to change the brightness of the sound. Two main limitations to this approach were found in the different behavior of the harmonic and noise components of the signal, and in the limited effect of enhancing the sound's brightness because of a lack of high frequency partials. The first problem was addressed by creating two different sets of filters for the harmonic and the noise components, and applying them independently. The second problem was addressed by adding high frequency partials to the signal, when needed.

The method described in this paper is a general approach to dynamics modification, our implementation being only one of the possible alternatives. We decided to use the ASR technique because it incorporates three of the features required in our system: accurate frequency estimation, analysis/synthesis and bandwidth extension. Nevertheless, the method can be implemented using other analysis/synthesis techniques (e.g. Serra's SMS [18]).

To be able to use the proposed model in a wider context, such as the rule-based performance modification system we proposed in [2], other problems need to be addressed. The estimation of the original level of a tone needs to be solved in order to choose the correct filters for the modification. We are currently working on a system using Support Vector Machine classifiers based only on spectral characteristics, ignoring sound level differences. Furthermore, the algorithm used for the resynthesis in the case of a two-channel recording can not preserve the stereo image, since the phase response of the two channels are computed independently. Finally, the method described in this paper only takes into account the sustain part of the signal. The attack plays also a very important role in defining the dynamic level of a tone, and its modification needs clearly to be addressed in the future to obtain a more effective result.

We also plan to perform an objective comparison of different variations of the model (simple filter, bandwidth extension, different levels of grouping to extract spectral envelopes), as well as a perceptual evaluation of the quality of the resulting modified samples. Nevertheless, informal listening tests suggest an improvement in comparison to the simple filter approach when enhancing the brightness of a tone. Sound examples are available at: `http://www.speech.kth.se/~himork/dafx09/`

5. REFERENCES

[1] David A. Luce, "Dynamic spectrum changes of orchestral instruments," *Journal of the Audio Engineering Society*, vol. 23, no. 7, pp. 565–568, 1975.

[2] Marco Fabiani and Anders Friberg, "A prototype system for rule-based expressive modifications of audio recordings," in *Proceedings of ISPS 2007 (Int. Symp. of Performance Science 2007)*, Aaron Williamon and Daniela Coimbra, Eds., Porto, Portugal, November 2007, pp. 355–360, AEC (European Conservatories Association).

[3] Marco Fabiani and Anders Friberg, "Rule-based expressive modifications of tempo in polyphonic audio recordings," in *Computer Music Modeling and Retrieval. Sense of Sounds*, Berlin, July 2008, Lecture Notes in Computer Science, pp. 288–302, Springer Berlin.

[4] M. Wright, J. Beauchamp, K. Fitz, X. Rodet, A. Robel, X. Sierra, and G. Wakefield, "Analysis/synthesis comparison," *Organized Sound*, vol. 5(3), pp. 173–189, 2000.

[5] Anne Caclin, Stephen McAdams, Bennett K. Smith, and Suzanne Winsberg, "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 471–482, 2005.

[6] Minna Ilmoniemi, Vesa Valimaki, and Minna Huotilainen, "Subjective evaluation of musical instruments timbre modifications," in *Proc. of the Joint Baltic-Nordic Acoustics Meeting 2004*, Mariehamn, Aland, June 2004.

[7] Alfonso Perez, Jordi Bonada, Esteban Maestre, Enric Guaus, and Merlijn Blaauw, "Score level timbre transformations of violin sounds," in *Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, September 2008.

[8] Tae Hong Park, Jonathan Biguenet, Zhiye Li, Conner Richardson, and Travis Scharr, "Feature modulation synthesis," in *Proc. of the 2007 International Computer Music Conference (ICMC07)*, Copenhagen, Denmark, August 2007, International Computer Music Association, vol. 2, pp. 368–372.

[9] Andrew Horner and James Beauchamp, "Synthesis of trumpet tones using a wavetable and a dynamic filter," *Journal of the Audio Engineering Society*, vol. 43, no. 10, pp. 799–812, 1995.

[10] Erik Larsen, Ronald M. Aarts, and Michael Danessis, "Efficient high-frequency bandwidth extension of music and speech," in *Proc. of the 112th Convention of the Audio Engineering Society*, Munich, Germany, May 2002.

[11] Arttu Laaksonen, "Bandwidth extension in high-quality audio coding," M.S. thesis, Helsinki University of Technology, 2005.

[12] Chi-Min Liu, Wen-Chieh Lee, and Han-Wen Hsu, "High frequency reconstruction for band-limited audio signals," in *Proc. of the 6th International Conference on Digital Audio Effects (DAFx-03)*, London, UK, September 2003.

[13] Anibal J.S. Ferreira and Deepen Sinha, "Accurate spectral replacement," in *Proc. of the 118th Convention of the Audio Engineering Society*, Barcelona, Spain, May 2005.

[14] ITU-R, "Recommendation ITU-R BS.1770-1 algorithms to measure audio programme loudness and true-peak audio level," ITU-R recommendation, ITU-R, Switzerland, 2007.

[15] Anibal J.S. Ferreira and Deepen Sinha, "Accurate and robust frequency estimation in the odft domain," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz (NY), USA, October 2005.

[16] Marine Campedel-Oudot, Olivier Cappe, and Eric Moulines, "Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 469–481, July 2001.

[17] Xinglei Zhu, Gerald T. Beauregard, and Lonce Wyse, "Real-time iterative spectrum inversion with look-ahead," in *Proc. of the 2006 IEEE Internationl Conference on Multimedia and Expo (ICME 2006)*, Toronto, Canada, July 2006.

[18] Xavier Serra and Julius O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, pp. 12–24, 1990.