A SYSTEM BASED ON SINUSOIDAL ANALYSIS FOR THE ESTIMATION AND COMPENSATION OF PITCH VARIATIONS IN MUSICAL RECORDINGS

Luís F. V. de Carvalho *

Electrical Eng. Program - COPPE/SMT,
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
luis.carvalho@smt.ufrj.br

Hugo T. de Carvalho

Institute of Mathematics
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
hugo.carvalho@smt.ufrj.br

ABSTRACT

This paper presents a computationally efficient and easily interactive system for the estimation and compensation of speed variations in musical recordings. This class of degradation can be encountered in all types of analog recordings and is characterized by undesired pitch variations during the playback of the recording. We propose to estimate such variations in the digital counterpart of the analog recording by means of sinusoidal analysis, and these variations are corrected via non-uniform resampling. The system is evaluated for both artificially degraded and real audio recordings.

1. INTRODUCTION

The problem of speed variations in old recordings is quite ubiquitous: for example, the puncture of vinyl and gramophone disks could be not well centered, and this kind of media, when subject to high temperature, could be bent; also, poorly stored magnetic tapes can be stretched. In both cases, when the degraded media is reproduced the playback speed will not be constant, causing an effect that is perceived as a pitch variation along the signal. Because of this audible effect, this defect is also known as "wow" in the literature. When considering historical collections, where it is very common to have only one copy of the recording available, it is then important to develop methods to identify and remove this effect of the degraded recording.

The study of quantification of "wow" dates back to the 40's [1, 2]. Depending on the cause of the degradation, mechanical methods can be used to restore such recordings: for example, correctly centering the puncture on a disk is a quite efficient way of undoing the degradation, but it only works in this particular case. For more general causes of this degradation, more sophisticated methods are required, that enable the use of digital signal processing, since the basic idea behind all the proposed restoration methods is to re-sample the degraded signal in a non uniform way such that the speed variation is compensated [3]. Therefore, it is then necessary to firstly estimate the so-called *pitch variation curve* (PVC) from the degraded signal. Then, a time-varying resampling algorithm is applied on the PVC. One of the first proposed methods following this guideline is [4, 5], where the curve is estimated via a statistical procedure from the spectrogram of the degraded signal and then used to resample it. A drawback of this method is that it is quite computationally intensive. This same idea was also explored in [6, 7], where an improved method for estimating the peaks in the spectrogram is proposed, as well as a different modeling for the pitch variation curve is employed. Since this modeling is parametric and sinusoidally-based, it can fail to describe more general curves. Other methods for determining the distortion are proposed in [8], although they are difficult to implement. Also in [8] an extensive discussion and comparison of several estimation methods is presented. Lastly, commercial tools are also available, for instance Capstan[1].

In this paper we propose a computationally efficient and non-parametric method which requires low amount of user interaction for determining the PVC based on a sinusoidal analysis of the degraded signals, as well as a time-varying resampling scheme that uses the estimated curve to restore the degraded signal.

The paper is organized as follows: in Section 2 an outline of the proposed solution is presented, and the next sections describe each step in more details; Section 3 presents the peak detection method employed in the sinusoidal analysis, followed by the peak tracking algorithm in Section 4; in Section 5 it is shown how the PVC is obtained from the estimated tracks, and Section 6 describes how the PVC is used in the time-varying resampling algorithm; results are presented and conclusions are drawn in Sections 7 and 8, respectively.

2. OUTLINE OF THE PROPOSED SOLUTION

The proposed solution has essentially three steps, as shown in Figure 1: the degraded signal is given as input to a sinusoidal analysis algorithm, whose estimated tracks are used to obtain the PVC, subsequently used in the time-varying resampling algorithm in order to restore the degraded signal. In this Section we briefly recall some aspects of the sinusoidal analysis and outline how these aforementioned steps are interconnected.

Sinusoidal analysis is a well-known multi-purpose technique in audio processing where small excerpts of a digital audio signal $x[n]$ are described as a sum of sinusoidal components [9]:

$$x[n] = \sum_j A_j[n] \cos(\Psi_j[n]), \tag{1}$$

where $A_j[n]$ and $\Psi_j[n]$ represent the time-varying envelope and the phase modulations of each component j, respectively.

The main goal is to estimate the parameters $A_j[n]$ and $\Psi_j[n]$ from the respective audio excerpt. With this set of parameters in hand, several tasks could be performed, for example, feature extraction or some modification and posterior re-synthesis of the signal [9]. In our case, this framework is used to estimate the PVC, based on the idea that all the frequencies present in an audio signal

* The authors thank CAPES and CNPq Brazilian agencies for funding this work.

[1] http://www.celemony.com/en/capstan

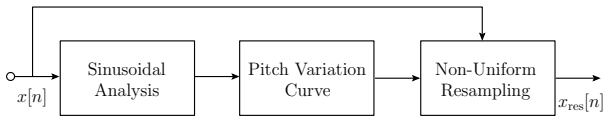

Figure 1: *Main steps of the proposed method.*

degraded by speed variation must contain some degree of deviation, more discussed in Section 5. Therefore, in this work it is more important to estimate the frequencies and amplitudes present within short excerpts of the audio signal than other quantities, and this estimation is performed as follows (see [10] for more details):

1. The whole signal $x[n]$ is segmented, with each segment being multiplied by a window function $w[n]$ (here the Hann window was employed) of length N_{W}, and contiguous segments have an overlap of N_{H} samples. Denote the n-th sample of the b-th block as $x_b[n]$, for $n = 1, \ldots, N_{\text{W}}$;
2. The N_{FFT}-point DFT of each segment is computed, its result being denoted by $X(k, b)$, representing the coefficient of the k-th frequency bin in the b-th block.
3. The most prominent peaks of each block that are more likely to be related to tonal components of the underlying signal are estimated via a procedure described in Section 3, based on [11];
4. Finally, in order to form tracks with the peaks estimated in the previous step, a simple algorithm is applied, where essentially a peak in some particular block is associated with the closest peak in the following block. This is a modification of the MQ algorithm [12], and is explained in more details in Section 4. The frequency of the i-th track in the b-th block of signal is denoted by $f_i[b]$.

Now, using an weighted average of the previously obtained tracks, the PVC is estimated via a simple procedure described in detail in Section 5. Finally, the PVC is given, together with the degraded signal, as input to a non-uniform resampling algorithm, detailed in Section 6. In the next Sections, the aforementioned steps are thoroughly discussed.

3. PEAK DETECTION

Since the presented method for estimating the PVC is based on sinusoidal analysis, it is important to employ a peak detection algorithm that rejects those noise-induced. Several methodologies have been proposed to detect tonal peaks in audio signals, including threshold-based methods [9, 13] and statistical analysis [14].

To separate genuine peaks from spurious ones, we adopt the *Tonalness Spectrum*, introduced in [11], since it is an easily extensible and flexible framework for sinusoidal analysis. This is a non-binary representation which indicates the likelihood of a spectral bin to be a tonal or non-tonal component. In this metric, a set of spectral features $\mathbb{V} = \{v_1, v_2, ..., v_{\text{V}}\}$ is computed from the signal spectrum and combined to produce the overall tonalness spectrum:

$$\mathcal{T}(k, b) = \left(\prod_{i=1}^{V} t_i(k, b) \right)^{1/\eta}, \tag{2}$$

where $t_i(k, b) \in [0, ..., 1]$, which is referred to as the *specific tonal score*, is calculated for each extracted feature v_i according to:

$$t_i(k, b) = \exp\left\{ -[\epsilon_i \cdot v_i(k, b)]^2 \right\}. \tag{3}$$

This measure can be explained as the probability of the given feature present a tonal component in the bin k of block b. The factor ϵ_i is a normalization constant which ensures that all specific features will equally contribute to the tonalness spectrum when combining them, and it is obtained by setting in Eq. 3 the specific tonal score of the median of the feature in each block to 0.5. This yields the following expression for the normalization constant:

$$\epsilon_i = \frac{\sqrt{\log(2)}}{\overline{m_{v_i}}}, \tag{4}$$

where $m_{v_i}(b)$ is the median value of feature i in the block b and $\overline{m_{v_i}}$ is the mean over all blocks of all files (in the case that a dataset is being analyzed).

The feature set comprises simple and established features, including a few that are purely based on information from the current magnitude spectrum of a block, such as frequency deviation, peakiness and amplitude threshold; some that are based on spectral changes over time, such as amplitude and frequency continuity; and one feature that is based on the phase and amplitude of the signal, which is the time window center of gravity.

Although the proposed method for estimating speed variations is based on a time-frequency representation, the peak detection stage is characterized by analyzing each frame individually. Therefore we take into account only those features which extract information from a single block.

The evaluation of results in [11] showed that, although the combination of features intuitively and empirically performs better than individual scores, combinations of more than three features did not achieve better representations. Moreover, it was also reported that combination with a simple product, that is, setting η to 1 in Eq. 2, yielded better results than with the distorted geometric mean.

Taking these information into account, our tonalness spectrum is formed by the combination of the *amplitude threshold* and *peakiness* features, since in [15] the combination of these features achieved good results on the detection of fundamental frequencies and their harmonics in music signals. Therefore, in our case, the expression in Eq. 2 can be simplified to:

$$\mathcal{T}(k, b) = t_{\text{PK}}(k, b) \cdot t_{\text{AT}}(k, b), \tag{5}$$

with $t_{\text{PK}}(k, b)$ and $t_{\text{AT}}(k, b)$ being the specific tonal scores of the peakiness and amplitude threshold, respectively.

The peakiness feature measures the height of a spectral sample in relation to its neighboring bins, and it is defined as:

$$v_{\text{PK}}(k, b) = \frac{|X(k + p, b)| + |X(k - p, b)|}{|X(k, b)|}, \tag{6}$$

where the distance p to the central sample should approximately correspond to the spectral main lobe width of the adopted window function when segmenting the signal, so side lobe peaks can be avoided.

The amplitude threshold feature measures the relation of the magnitude spectrum by an adaptive magnitude threshold, and it is defined as:

$$v_{\text{AT}}(k, b) = \frac{r_{\text{TH}}(k, b)}{|X(k, b)|}, \tag{7}$$

where $r_{\mathrm{TH}}(k,b)$ is a recursively smoothed version of the magnitude spectrum:

$$r_{\mathrm{TH}}(k,b) = \beta \cdot r_{\mathrm{TH}}(k-1,b) + (1-\beta)\cdot|X(k,b)|, \quad (8)$$

with this filter being applied in both forward and backward direction in order to adjust the group delay, and $\beta \in [0,1]$ being a factor that is empirically tweaked.

3.1. Peak Selection

After computing the tonalness spectrum, all peaks k_i are selected in each block, and those which do not fulfill the following criterion are discarded:

$$\mathcal{T}(k_i,b) \geq \mathcal{T}_{\mathrm{TH}}, \quad (9)$$

where $\mathcal{T}_{\mathrm{TH}} \in [0,1]$ is an empirically adjusted likelihood threshold.

Moreover, since our tonalness spectrum measurement evaluates peaky components and their surroundings, independently of their absolute magnitude amplitudes, some small and insignificant peaks could present a high tonalness likelihood and thus be selected. Hence the following criterion is employed in each block to discard such irrelevant peaks:

$$|X(k_i,b)| \geq \gamma \cdot \max|X(k,b)|, \quad (10)$$

where γ is a percentage factor, and satisfactory peak selections were obtained by setting this factor to around 1%.

Figure 2 illustrates the peak selection stage in a block of an audio signal containing a single note being played by a flute. The criteria in Eq. 9 and 10 were set to 0.8 and 1%, respectively. It can be seen that the tonalness spectrum is a powerful representation of tonal components, when comparing it with the original magnitude spectrum.

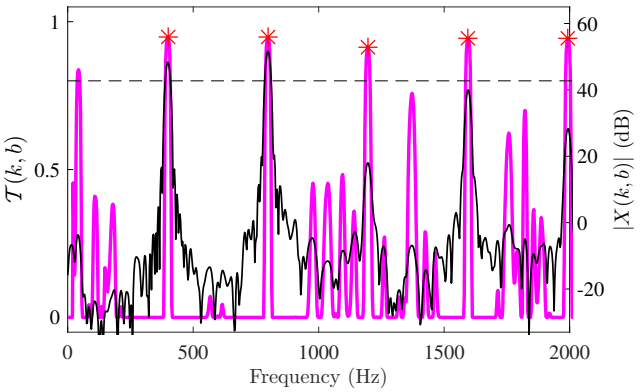

Figure 2: *Illustration of the peak detection stage. The thicker and brighter line represents the tonalness spectrum of a block from an audio signal, whose magnitude spectrum is indicated on the thinner and darker line. The likelihood threshold is represented by the horizontal dashed line and the selected peaks are marked with asterisks.*

4. PARTIAL TRACKING

To conclude the sinusoidal analysis stage, the spectral peaks selected in each frame are grouped into time-changing trajectories in which both frequency and amplitude can vary. This process is referred to as *partial tracking*.

Several methods have been proposed to track spectral peaks. Relevant works include the classical McAuley & Quatieri (MQ) algorithm [16], its extended version using linear prediction [17], and a solution via recursive least-squares (RLS) estimation [18].

In this work, a modified version of the MQ algorithm is employed, which is described in [12], since this is a computationally efficient technique that is easily implementable and achieves good representation of sinusoidal tracks.

In this algorithm, a track can be marked with three different labels: it *emerges* when a peak is not associated with any existing track, *remains* active while it is associated with peaks, and *vanishes* when it finds no compatible peak to incorporate. Defining $f_{i,b}$ and $A_{i,b}$ the frequency and magnitude amplitude of the ith detected peak in frame b, the algorithm can be explained as follows:

1. For each peak $f_{j,b+1}$ a search is performed to find a peak $f_{i,b}$ from a track which had remained active until the frame b, satisfying the condition $|f_{i,b} - f_{j,b+1}| < \Delta f_{i,b}$. The parameter $\Delta f_{i,b}$ controls the maximum frequency variation, and is set to a quarter tone around $f_{i,b}$.
2. If the peak $f_{j,b+1}$ finds a corresponding track in the previous frame satisfying the condition described in step 1, it associates with this track, which remains active. If two or more peaks satisfy the condition, the peak that minimizes

$$J = (1-\kappa)\frac{|f_{i,b} - f_{j,b+1}|}{f_{i,b}} + \kappa\frac{|A_{i,b} - A_{j,b+1}|}{A_{i,b}} \quad (11)$$

   is selected, where $\kappa \in [0,1]$ is a weighting parameter that controls the influence of the relative frequency and amplitude differences in the cost function in Eq. 11.
3. When the peak of a track in b is not associated with any peak in $b+1$ satisfying the condition, it is marked as *vanishing* and a virtual peak with its same frequency and amplitude is created in $b+1$. When the track reaches D consecutive virtual peaks, it is then terminated.

Except for the first frame, where all the peaks invariably start new tracks, these steps are performed in all frames, until all peaks are labeled. Lastly, short tracks whose length is less than a number E of frames are removed.

5. GLOBAL PITCH VARIATION CURVE

A consequence of speed variations in musical signals is the deviation of all their frequencies by the same percentage factor, that is, a pitch modification. Therefore, a curve which combines the variations of the main frequency components of the signal is a suitable metric for estimating the defect.

In [4, 5] a Bayesian procedure is proposed to estimate the pitch variation curve, but it is quite computationally expensive. An alternative is to determine the distortion using only the most prominent spectral component of the signal, a technique proposed in [8]. However, this method is not practical, and such spectral component may contain frequency modulations not corresponding to speed variations, interfering then with the correct curve estimation.

Our proposed approach consists of calculating a weighted average curve from the computed tracks, thus exhibiting an overall

behavior of how the tracks vary with time. If a global frequency variation in a part of the signal is detected, this might be a promising evidence of a speed variation in that part.

The weights in this average are the magnitude amplitudes of the detected peaks. The motivation for weighting the curve is that the most important tracks should contribute more to the PVC than the less prominent ones. This metric can be interpreted as an extended version of the method proposed in [8]. Since this metric takes into account more frequency components and it is refined by their magnitude amplitudes, it is expected that this overall pitch variation curve can be more reliable than taking only one component.

For aesthetic reasons, it is quite common the presence of notes played with ornaments in music recordings, such as vibrato. This is often seen in bowed string or woodwind instruments, and in singing voice as well. Thus, when tracking an instrument recording with such effect, parts where vibrato occurred could be erroneously detected as speed variations. However, if the instrument is in a recording with other musical instruments, which would not necessarily be synchronized with its vibrato, that effect would be attenuated by computing an average of the tracks. Nevertheless, the method presented in this paper can be implemented in a way that the user may select the specific parts of the audio signal to be restored.

5.1. Extraction of the weighted global average of the tracks

In the first step of the global pitch variation curve extraction, the mean of all tracks are shifted to zero, and then each of them are normalized by its respective frequency average in time, this way obtaining the percentage average of each sinusoidal trajectory. For a track i, this stage is mathematically described as:

$$\tilde{f}_i[b] = \frac{f_i[b] - \bar{f}_i}{\bar{f}_i}, \tag{12}$$

with $f_i[b]$ being the frequency of the ith track in the frame b, and $\bar{f}_i$ being the arithmetic mean of all frequencies in the ith track.

Following that, the weighted average of each frame b is calculated:

$$\tilde{f}'[b] = \frac{\sum_i A_i[b]^\alpha \tilde{f}_i[b]}{\sum A_i[b]^\alpha}, \tag{13}$$

where $A_i[b]$ represents the magnitude amplitude of the ith track in the frame b and $\alpha \in [0, 1]$ is a parameter that controls the influence of large amplitudes over smaller ones, which we will from now on refer to as the *weighting factor*. The motivation for this parameter is that the magnitude amplitude of the harmonic components of tonal peaks drop sharply as the harmonic order increases, and one may desire to increase the participation of such medium and small amplitude harmonics in the computation of the pitch variation curve.

It can be noticed that when $\alpha = 0$, the Eq. 13 happens to be the arithmetic mean of the tracks, indicating that all tracks will equally contribute to the PVC; analogously, when $\alpha = 1$, the PVC is set so the small-amplitude peaks influence less in its estimation. To minimize the effects of false tracks, the weighting factor would naturally be set to a value close to 1, and empirical tests indicate that setting α between 0.7 and 1 achieve satisfying results.

The curve $\tilde{f}'[b]$ may present undesired small irregularities, which are caused by frequency inaccuracies and the tracking of non-tonal peaks which do not correspond to frequency components. Hence, this curved is smoothed by a moving average filter,

$$\tilde{f}[b] = \frac{1}{N_{\mathrm{MA}}} \sum_{i=0}^{N_{\mathrm{MA}}-1} \tilde{f}'[b+i], \tag{14}$$

where N_{MA} is the order of the filter, and the final PVC is then the vector $\tilde{\mathbf{f}}$, whose components are $\tilde{f}[b]$, for $m = 1, \ldots, N_{\mathrm{B}}$, with the latter term being the total number of blocks.

It is then expected that the pitch variation curve of a signal exhibits values around zero. For a better interpretation of this curve, it is interesting to normalize it, which is realized by shifting its average to 1. This is a better notation when it is necessary to adopt a frequency reference. For example, curves related to parts in which there was no speed variations now exhibit values around 1, indicating that the frequencies of their tracks have all been multiplied by 1. If in a frame transition there is a deviation of 0.6%, this now is represented in the curve as 1.006, that is, all frequencies in this transition on average were multiplied by 1.006.

6. NON-UNIFORM RESAMPLING

In this section, it is described how non-uniform sampling rate conversion can be employed from the PVC to compensate speed variations in digital audio signals.

When an audio signal is played back with a different rate from that which was originally sampled, the perceived pitch is modified, as well as its time duration is distorted. It turns out, as explained in Section 1, that speed variations yield exactly pitch and time variations. Therefore sampling rate conversion is a suitable technique for compensating such defects in digital versions of degraded recordings.

Furthermore, since speed variations are not constant in time, this sampling rate conversion must be non-uniform, which can also be referred to as *time-varying resampling*. In [3] different methods for "wow" reduction are compared, and the sinc-based interpolation [19] achieved less distortions in reconstructed signals. Therefore, an algorithm that receives as input the digital version of the degraded signal and its PVC, and its output the reconstructed signal using the sinc-based time-varying resampling was implemented.

The sinc-based interpolation is closely related to the analog interpretation of resampling, which is the reconstruction of the continuous version from the discrete signal, followed by resampling it with the new desired rate. The expression for the converted signal $x_{\mathrm{rec}}[n']$ with new rate F_{s}' using the sinc-based resampling is given by [19]:

$$x_{\mathrm{rec}}[n'] = \sum_{n=-N_{\mathrm{T}}}^{N_{\mathrm{T}}} x[n] \,\mathrm{sinc}\left(\pi \left(\frac{n'}{f_{\mathrm{r}}} - n\right)\right), \tag{15}$$

where $f_{\mathrm{r}} = F_{\mathrm{s}}'/F_{\mathrm{s}}$ is the resampling factor, i.e. the ratio between the desired and original sampling rates, and $2N_{\mathrm{T}} + 1$ is the number of samples used in the interpolation. The reconstruction of a sample is illustrated in the Figure 3.

6.1. Implementation

Since the pitch variation curve $\tilde{\mathbf{f}}$ from Eq. 14 is normalized to 1, each of its elements correspond exactly to the term f_{r} in Eq. 15. However, each element is associated to a block of the signal, which was segmented during the STFT procedure. Hence the transition between blocks must be modeled sample by sample.

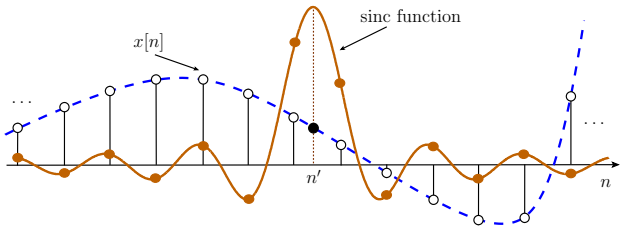

Figure 3: *Illustration of the sinc-based interpolation method.*

The following steps summarize the non-uniform resampling procedure that was implemented in this work, given a digital audio signal and its pitch variation curve:

1. Firstly, each element $\tilde{f}[b]$ of the vector $\tilde{\mathbf{f}}$ is associated to the central sample of its respective block;
2. Secondly, the number of samples which will be reconstructed between each pair of consecutive central samples of $\tilde{f}[b]$ and $\tilde{f}[b+1]$ is calculated, so the transition between these central samples be smooth. This is obtained by realizing a linear interpolation to obtain the positions of each sample to be reconstructed (which can generically be represented by the sample in n' in Figure 3);
   (a) This process inevitably involves numerical approximations, which in this context are translated into phase deviations. Such approximations must be taken into account, and therefore compensated in order to eliminate audible artifacts.
3. Finally, the expression in Eq. 15 is applied for each new sample.

It is also worth mentioning that the correction of speed variations can be performed offline, so N_{T} can be chosen large enough in order not to compromise the quality of the reconstructed signal. Experiments in [3] showed that using $N_{\mathrm{T}} = 100$ achieved inaudible distortions. Moreover, our experiments with such number of samples or even less also did not degraded the signal.

7. RESULTS

This section presents the evaluation of the proposed algorithm. Several degraded signals were investigated, and two of them are shown in this paper, one from an artificially degraded recording, and one from a real recording. The audio signals presented here as well as all implemented codes in this work are available in [20]. All tests were realized with the same set of parameters, which are summarized in Table 1.

The first test was performed with an excerpt of orchestral music, containing long notes being played by bowed instruments. An artificial sinusoidal pitch variation was imposed into this signal and the proposed method was applied, so both true and estimated PVCs can be compared. Although the curves slightly differ in terms of their amplitudes, as can be seen in the first graph of Fig. 4, a satisfactory correspondence can be observed between them. As can be seen in the second graph of Fig. 4, the pitch variation curve of the resampled signal shows only slight variations, which can be explained by the amplitude difference mentioned before; however, informal listening tests reported no audible pitch variations. It is also worth mentioning that there are no precise objective measurements to compare the original signal with the restored one.

Table 1: *Parameters used to estimate the PVC.*

Parameter	Value
N_{W}	4096 samples
N_{H}	256 samples
N_{FFT}	16384 samples
β	$1500/N_{\mathrm{FFT}}$
$\mathcal{T}_{\mathrm{TH}}$	0.75
γ	1%
κ	0.6
D	5 blocks
E	10 blocks
α	0.8

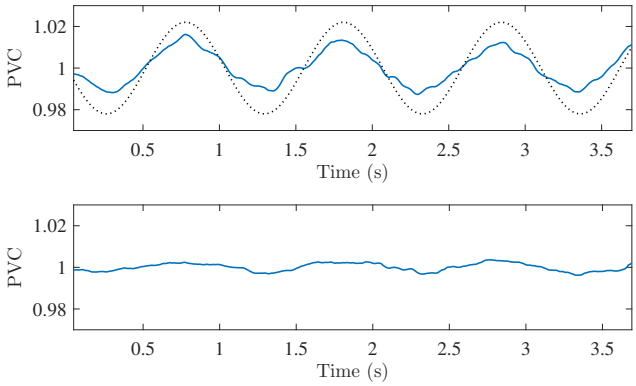

Figure 4: *Restoration of the signal 'orchestra'. In the first graph, the solid line and dotted line represent the estimated and true PVCs, respectively. The PVC of the restored signal is shown in the second graph.*

The second test was realized on a piano recording which presents genuine speed variations. Figure 5 shows the PVCs of the degraded and restored signals, respectively. What sorts out from the graphs is that the proposed method estimated the shape of the variations with considerable accuracy, and therefore the result was considered satisfying. However, the PVC of the restored signal still presents some slight variations, but this can be explained by the inaccuracies of the sinusoidal analysis stage, more precisely during the partial tracking stage.

8. CONCLUSION

This work has presented a computationally efficient system which requires minimum user interaction for the estimation and correction of speed variations in the playback of musical recordings. The stage of estimating the pitch variation curve was based purely on sinusoidal analysis, and the good results indicated that the proposed framework can serve, for example, as the core of a more sophisticated and robust professional tool for audio restoration.

It can be said that the partial tracking stage appears as the less robust stage in the system, and it may have the biggest influence in the inaccuracies reported in Figures 4 and 5. Therefore, future works include the development of a more robust algorithm for peak tracking, possibly operating simultaneously with a group

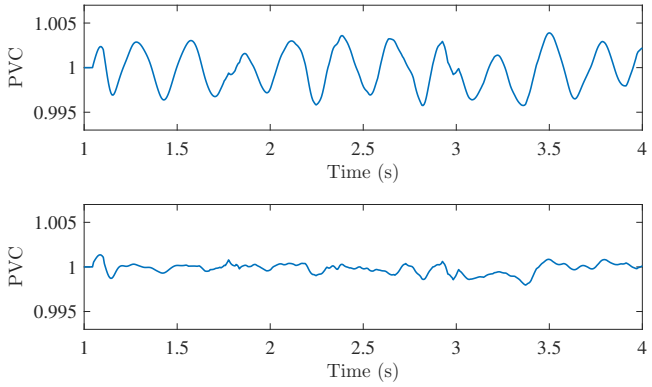

Figure 5: *Restoration of the signal 'piano'. The estimated PVC is depicted in the first graph, and the restored PCV is shown in the second one.*

of blocks, as proposed in [17]. Other potential improvements to be implemented include the development of a better method for converting the set of tracks into the pitch variation curve, and the assessment of results via subjective tests.

As stated in [8], a fully automatic system, although tempting, is not feasible. Since the nature of pitch variations in old music recording is wide [5], this class of audio degradation requires a human operator for applying restoration algorithms. However, we believe that such systems can be minimally and friendly interactive, so non-professional users can restore their old domestic recordings by their own.

9. ACKNOWLEDGMENTS

The authors wish to express their sincere gratitude to Prof Luiz W. P. Biscainho for the motivation and ideas for this work. They also would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

10. REFERENCES

[1] U. R. Furst, "Periodic variations of pitch in sound reproduction by phonographs," *Proceedings of the IRE*, vol. 34, no. 11, pp. 887–895, Nov 1946.

[2] P. E. Axon and H. Davies, "A study of frequency fluctuations in sound recording and reproducing systems," *Proceedings of the IEE - Part III: Radio and Communication Engineering*, vol. 96, no. 39, pp. 65–75, Jan 1949.

[3] P. Maziewski, "Wow defect reduction based on interpolation techniques," in *Proceedings of the 4th National Electronics Conference*, Darlowko Wschodnie, Poland, Jun 2005, pp. 481–486.

[4] S. J. Godsill and P. J. W. Rayner, "The restoration of pitch variation defects in gramophone recordings," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 1993, pp. 148–151.

[5] S. J. Godsill, "Recursive restoration of pitch variation defects in musical recordings," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'94)*, Apr 1994, vol. 2, pp. 233–236.

[6] J. Nichols, "An interactive pitch defect correction system for archival audio," in *Proceedings of the AES 20th International Conference*, Budapest, Hungary, Oct 2001.

[7] J. Nichols, "A high-performance, low-cost wax cylinder transcription system," in *Proceedings of the AES 20th International Conference*, Budapest, Hungary, Oct 2001.

[8] A. Czyzewski, A. Ciarkowski, A. Kaczmarek, J. Kotus, M. Kulesza, and P. Maziewski, "DSP techniques for determining wow distortion," *Journal of the Audio Engineering Society*, vol. 55, no. 4, pp. 266–284, 2007.

[9] J. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *International Computer Music Conference*, Urbana, Illinois, USA, Aug 1987, pp. 290–297.

[10] P. A. A. Esquef and L. W. P. Biscainho, "Spectral-based analysis and synthesis of audio signals," in *Advances in Audio and Speech Signal Processing: Technologies and Applications*, H. M. Pérez-Meana, Ed., chapter 3. Idea Group, Hershey, 2007.

[11] S. Kraft, A. Lerch, and U. Zölzer, "The tonalness spectrum: feature-based estimation of tonal components," in *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx'14)*, Maynooth, Ireland, Sep 2014.

[12] L. Nunes, L. Biscainho, and P. Esquef, "A database of partial tracks for evaluation of sinusoidal models," in *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx'10)*, Graz, Austria, Sep 2010.

[13] N. Laurenti, G. De Poli, and D. Montagner, "A nonlinear method for stochastic spectrum estimation in the modeling of musical sounds," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 531–541, Feb 2007.

[14] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, Sep 1982.

[15] S. Kraft and U. Zölzer, "Polyphonic pitch detection by matching spectral and autocorrelation peaks," in *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO'15)*, Nice, France, Aug 2015, pp. 1301–1305.

[16] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, Aug 1986.

[17] M. Lagrange, S. Marchand, and J. B. Rault, "Using linear prediction to enhance the tracking of partials," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, May 2004, vol. 4, pp. 241–244.

[18] L. Nunes, R. Merched, and L. Biscainho, "Recursive least-squares estimation of the evolution of partials in sinusoidal analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, April 2007, vol. 1, pp. 253–256.

[19] F. Marvasti, *Nonuniform Sampling Theory and Practice*, Kluwer Academic Publishers, New York, USA, 2001.

[20] WOW, "WOW companion webpage," Available at `http://www.smt.ufrj.br/~luis.carvalho/WOW/`, 2017.