A DATABASE OF PARTIAL TRACKS FOR EVALUATION OF SINUSOIDAL MODELS

Leonardo O. Nunes and Luiz W. P. Biscainho[*]
PEE/COPPE, Federal Univ. of Rio de Janeiro
Rio de Janeiro, Brazil
{lonnes,wagner}@lps.ufrj.br

Paulo A. A. Esquef[*]
CSC, National Lab. for Scientific Computing
Petrópolis, Brazil
pesquef@lncc.br

ABSTRACT

This paper presents a database of partial tracks extracted from synthetic as well as pre-recorded musical signals, designed to serve as an ancillary tool for evaluation of sinusoidal analysis algorithms. In order to accomplish this goal, the database requirements have been carefully specified. A semi-automatic analysis methodology to ensure the track parameters are precisely estimated has been employed. The overall methodology is validated via the application of performance tests over the synthetic source-signals.

1. INTRODUCTION

The modeling of the resonant part of musical signals as a sum of amplitude- and frequency-modulated sinusoids, the so-called sinusoidal modeling (SM) [1, 2], has found many applications [3] in digital audio processing. The model's popularity lies in its ability to represent well the time evolution of resonant signals as a set of separate spectrally salient components, called sinusoidal tracks. This paper addresses the difficulty of comparing the performance of different analysis algorithms by objective means.

Performance comparison of sinusoidal analysis algorithms when using natural signals is difficult, mainly due to the lack of referential data against which to confront the attained results. Thus, indirect means to performance assessment are usually resorted to. For example, a solution proposed in [3] to compare several sinusoidal track estimators was to first reconstruct new versions of the test signals from their estimated tracks. Then, pre-defined criteria on the results of comparative listening tests with the original and synthesized versions of the test signals were used to evaluate system performance. However, that kind of evaluation does not explicitly provide quantitative data that can be used to choose one system in detriment of another. Moreover, the information provided by subjective tests might not be insightful, depending on the final audio application in which the estimated tracks will be used.

Objective performance evaluation of sinusoidal analysis methods is rendered feasible when reference sinusoidal tracks that represent well the signals under test are available. This way, objective measures such as the ones presented in [4] can be calculated. Such set of reference tracks can be easily created for synthetic signals. However, for natural signals, the reference tracks need themselves to be extracted by a sinusoidal analysis system, the choice of which is likely to bias further comparisons.

In this paper, the development of a database containing audio signals along with their reference sinusoidal tracks is described. Such *corpus*, called TDB (Sinusoidal Track Database) would provide ancillary data for objective benchmarking of sinusoidal analysis algorithms. In order to ensure the reliability of the tracks stored in the TDB, it is necessary to judiciously define the sinusoidal analysis algorithms used in its creation and resort to manual intervention when judged necessary during the analysis procedure.

The paper is organized as follows. In Section 2 a brief overview of sinusoidal modeling is given. The signals whose tracks are to be part of the TDB (source-signals) are described in Section 3. In Section 4, the desired characteristics of TDB are specified. In Sections 5 and 6, the analysis methods used for obtaining the reference tracks and the methodology used in their extraction are respectively detailed. In Section 7 tests are performed to validate the proposed methodology. Conclusions are drawn in Section 8.

2. SINUSOIDAL ANALYSIS OVERVIEW

Sinusoidal modeling [1] describes an audio signal $x(t)$ as a sum of L sinusoids, i.e.,

$$x(t) = \sum_{l=1}^{L} A_l(t) \sin(\Psi_l(t)), \tag{1}$$

where $A_l(t)$ and $\Psi_l(t)$ are, respectively, the amplitude and the phase modulations of partial l. Usually, Eq. (1) is replaced by a discrete-time model,

$$x[n] = \sum_{l=1}^{L} A_l[n] \sin\left(\Psi_l[n]\right). \tag{2}$$

For a given partial l, the approximations $A_l[n] \approx A_l$ and $\Psi_l[n] \approx \Omega_l n + \Psi_l[0]$, where A_l and Ω_l are constant values, hold true within a sufficiently short N-sample frame.

The main objective of a sinusoidal analysis algorithm consists in estimating A_l and Ω_l across frames. The typical stages [2] in the analysis portion of an SM system are illustrated in Fig. 1.

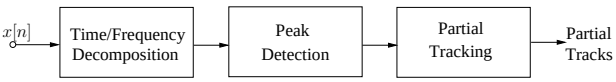

Figure 1: Processing stages of a sinusoidal analysis system.

The 'time / frequency decomposition' stage usually involves the discrete-time Short-Time Fourier Transform (STFT) of the audio signal $x[n]$, i.e.,

$$X[m,k] = \frac{1}{N} \sum_{n=0}^{N-1} w[n] x[n + mH] e^{-jk\frac{2\pi}{N}n}, \tag{3}$$

where $w[n]$ is a window of length N, e.g. the Hamming window, k is the frequency bin index, m is the frame index, and H is the

[*] The authors wish to thank CNPq, CAPES, and FAPERJ for financial support via grants 306607/2009-3, 306640/2009-0, and E26/103.098/2008.

frame hop (in samples) along time. Out of $|X[m,k]|$, the 'peak detection' stage is supposed to select only those peaks in frame m that correspond to genuine stationary sinusoidal components. Finally, the 'partial tracking' is responsible for coherently grouping peaks across consecutive frames into sinusoidal tracks.

Each track models an amplitude- and frequency-modulated sinusoid. The database described in this paper consists of sets of such sinusoidal tracks, each set ideally capturing completely the resonant information of a chosen musical signal.

3. SOURCE-SIGNALS

The source-signals are specified according to increasing levels of content complexity, enabling the performance assessment of a given analysis method under different levels of difficulty. For that, more challenging analysis scenarios , e.g. *vibrato* playing and polyphony, are introduced from one level of complexity to the next.

3.1. Categorization and Specification of Source-Signals

Following the division proposed in [5], four levels are specified:

Level 0 – Containing synthetic musical signals.

Level 1 – Containing recordings of single musical notes played on acoustic instruments, one per signal.

Level 2 – Containing recordings of musical excerpts played on polyphonic acoustic instruments, one per signal.

Level 3 – Containing recordings of musical excerpts played on more than one acoustic instrument per signal.

Level 0 is included so to provide referential data upon which to verify if the analysis methods used to extract the sinusoidal tracks meet the requirements of the TDB (see Section 4). Level 1 contains the simplest recorded musical signals of the TDB. Level 2 allows the performance evaluation of sinusoidal analysis algorithms for note emissions with *vibrato* and *tremolo*, as well as for note transitions. At last, Level 3 allows the evaluation of sinusoidal analysis algorithms for different degrees of polyphony, including the case of crossing frequency trajectories.

The duration of signals used in Levels 1 to 3 was kept between 2 and 15 s to allow their use in formal listening tests, if needed.

All source-signals must be monaural, stored in PCM format with sampling rate of 44.1 kHz and 16-bit precision, amplitude-normalized so to have maximum absolute value equal to 1. Other source-signals sharing this format can be used to expand the TDB.

3.2. Chosen Source-Signals

The signals chosen for Levels 1 to 3 are excerpts of recordings taken from the Real World Computing Database (RWC) [6]. More information on those signals, including their localization within the RWC, can be found in [7].

Next, the signals selected for each level are briefly described.

3.2.1. Level 0

The signals associated to Level 0 were created specifically for the TDB. For this, a trumpet sound model generated by the additive synthesizer described in [8] was used. Since trumpet sounds typically have a high number of partials, a large number of sinusoidal tracks are required for their proper modeling.

Three signals, described in Table 1, were created: (1) a signal with *vibrato*; (2) a signal with *tremolo*; and (3) a three-note sequence, contaminated with additive noise.

Once synthesized, the signals were stored in WAVE files according to the specifications defined in Section 3.2.

3.2.2. Level 1

The signals chosen for Level 1 are excerpts of the so-called 'musical instrument' database of the RWC [9]. Table 2 summarizes their characteristics.

Recordings of musical instruments exhibiting different characteristics and from different families were chosen. Nevertheless, overall the selected signals are tonal and conceivably well-modeled by Eq. (2).

3.2.3. Level 2

The selected signals for Level 2 were taken as excerpts from the so-called 'classical music' and 'jazz music' recording databases of the RWC[6]. Table 3 gives composer and title of the musical piece played in each chosen signal.

The Organ signal has low-register notes, including one note with f0 = 62 Hz. The Violoncello signal exhibits rapid transitions between notes, which hinder the detection of track onsets and offsets. On the other hand, the Violin signal contains notes with both *vibrato* and *tremolo*. At last, the Piano signal contains chords, being the first case of polyphony in the TDB.

3.2.4. Level 3

Level 3 is composed of two signals:

1. An excerpt of the recording *Crescent Serenade* from H. Kobayashi, taken from the 'jazz music' recording database of the RWC [6], including piano, bass, drums, and tenor saxophone. Two factors in this signal can stress a sinusoidal analysis system: polyphony and the presence of an unpitched percussive instrument that cannot be well-modeled by Eq. (2).
2. Two simultaneous trombone *glissandi* in contrary motion, with crossing partials, edited from two tracks of the 'musical instrument' database of the RWC.

4. SPECIFICATION OF THE TDB

In this section, specifications with the objective of guiding the construction of the TDB are defined. The underlying strategy behind the analysis methodology is to set stringent specifications for the methods involved in track estimation. In practice, conformance of the analysis tools with the defined specifications can only be verified via synthetic source-signals. Nevertheless, the bet is that a similar level of accuracy in model representation be also attained when dealing with recorded signals. Any further TDB extensions through analysis of new source-signals are expected to follow the specifications stated in this section.

4.1. Track Specification

A sinusoidal track should be composed of three vectors, containing samples of the temporal trajectory of the following parameters:

Table 1: Main characteristics of the signals in Level 0 of the TDB.

Name	f0 (Hz)	Length (s)	No. Partials	Characteristics
Vibrato	156	2	26	*vibrato* rate = ± 4 Hz *vibrato* depth = ± 1 Hz
Tremolo	156	2	26	*tremolo* rate = ± 8 Hz *tremolo* extension = ± 50 %
Note Sequence	note 1 = 156	0.5	26	0.2 s-long pause between notes 2 and 3 additive pink noise 60 dB SNR
	note 2 = 262	1	21	
	note 3 = 622	0.2	10	

Table 2: Characteristics of the signals chosen in Level 1 of the TDB.

Name	f0 (Hz)	Length (s)	Characteristics
Violin	247	3.1	*Vibrato* playing, with *mezzo-forte* dynamics.
Piano	207	2.2	Conventional execution, with no use of pedal, and *forte* dynamics.
Saxophone	148	4.2	Conventional execution, with *forte* dynamics.
Accordion	312	2.5	Conventional execution, with *forte* dynamics.

Table 3: Composition and composer information for the signals used in Level 2 of the TDB.

Name	Length (s)	Composer	Composition	f0 Range (Hz)
Organ	10.9	Johann S. Bach	Passacaglia and Fugue in C minor	from 62 to 104
Violin	3.4	Johann S. Bach	Partita no. 6 for Violin Solo	from 210 to 427
Piano	10.2	Not identified	Identified as *Jive*	from 193 to 603
Violoncello	10.2	Johann S. Bach	Suite no. 1 for Violoncello Solo in G major	from 97 to 143

amplitude (in linear scale), frequency (in Hz), and phase (in radians). Besides these vectors, the track should store the time instant (in seconds) when it was first detected.

In the case of an analysis system that uses a frame-by-frame analysis of the signal, the parameters' sampling rate is usually defined by the hop between consecutive frames. Ideally, the spectral content of the trajectories is limited to 20 Hz [10], which would require a sampling rate greater than 40 Hz. However, oversampling is commonly employed by sinusoidal analysis systems to facilitate obtention of the sinusoidal tracks. Considering that rates as high as 200 Hz were reported in [11], a sampling rate $F_{\text{sp}} = 500$ Hz was considered adequate for the track parameters. Moreover, the parameters should be stored in double-precision floating-point format.

In the TDB, a track from a source-signal should start and end at times coherent with the perception of the event that induced the track. In *glissando* or *legato* passages, the tracks should capture these effects, if they are perceived as continuous events (i.e. without the perception of onsets). As regards the accuracy of onset and offset time estimates, the tolerance adopted by onset detection algorithms is usually equal to 50 ms [12]. A smaller maximum error of 20 ms is adopted in the TDB for the estimation of track onsets and offsets.

In order to ease the detection of genuine spectral peaks, an algorithm for noise floor estimation [13] can be used. Only those peaks whose magnitude is above this floor are to be selected.

If a given spectral peak is not detected in a certain time interval, causing an improper track interruption, then an interpolation algorithm can be used to find the missing track parameters. A segment of up to 8 ms (equivalent to 4 consecutive parameter samples) can be interpolated. This choice was inspired by the duration limit below which humans cannot hear a mute in a sinusoid [14]. Track gaps longer than 8 ms are left intact.

4.2. Specification of Estimation Errors

In this section, the maximum tolerable errors for the estimated parameters (frequency and amplitude) of a track are specified. The objective is to provide a confidence interval around the estimated parameters. Hence, performance evaluations that use the TDB can take into account this uncertainty about the reference parameters.

In order to specify the errors, psychoacoustic criteria will be defined. The main idea is to use the so-called JND (Just Noticeable Difference), that provides, for a given physical quantity, the least difference in level that is perceptible by humans. Appropriate JND in frequency and intensity (amplitude) reported in the literature for pure tone experiments are going to be explored: the estimation error for each parameter will be considered imperceptible whenever smaller than the JND in that quantity.

4.2.1. Frequency

For short-duration sinusoids, it was experimentally observed that the JND in frequency δf is a function of frequency and is reduced as stimulus intensity grows [15]. The dependency

$$\log_{10} \delta f = a\sqrt{f} + b \tag{4}$$

is reported, where the parameters a and b depend on the intensity and are chosen to approximate experimental data.

The maximum acceptable relative error for the frequency estimate f is defined as

$$e^{\text{freq}}(f) = \frac{\delta f}{f}, \tag{5}$$

where δf is obtained through Eq. (4).

The intensity with which a given track will be reproduced (after synthesis) is unknown. Thus, aiming at the smallest JND in frequency, the values $a = 0.028$ and $b = -0.696$ were chosen,

since they were obtained for the largest intensity reported in [15], i.e. 80 dB SL[1].

Evaluation of Eq. (5) with 20 Hz $\leq f \leq$ 10 kHz reveals a minimum at about $f = 962$ Hz, where $e^{\text{freq}}(f) = 0.15$ %. Eq. (5) will be used to provide frequency-dependent tolerances tolerances for the estimation error.

4.2.2. Amplitude

The dependence of the smallest perceptible increment δI in relation to the intensity I of a pure sinusoid was found in [17] to be

$$\frac{\delta I}{I} = 0.463 I^{-0.072}. \quad (6)$$

Moreover, a negligible influence of the sinusoid frequency on δI was also reported.

The maximum relative error for amplitude estimation was chosen as 20 %, a tolerance that is always smaller than $\frac{\delta I}{I}$ for intensities as high as 100 dB SL. This stringent choice was made in order to overcome the uncertainty regarding the intensity level at which a given track will be reproduced after synthesis.

4.2.3. Phase

Absolute and relative phase differences in tonal sounds are hardly perceived by humans [16]. However, phase estimation can be important for synthesis and modification algorithms of sinusoidal tracks [1]. For this reason, the TDB includes phase estimates for each track at each analysis frame. The specification of a maximum error, however, was not considered necessary, given that phase values vary little around the frequencies of resonant components [10].

5. ANALYSIS METHODS

In this section, the sinusoidal analysis methods used to generate the TDB are described. The following criteria were used to their choice: (1) Be able to extract tracks that meet the specifications defined in Section 4; (2) Allow manual adjustment of their processing parameters for each signal to be analyzed. Different methods than the ones presented in this section can be used in future extensions of the TDB, provided they are able to meet the specifications given in Section 4.

The following notation will be used hereafter: the frequency trajectory estimate of track i at frame m will be denoted as $\hat{f}_{i,m}$; similarly, the amplitude trajectory estimate will be denoted as $\hat{A}_{i,m}$. The amplitude and frequency estimates of the p-th peak detected at frame m will be denoted as $\overline{A}_{p,m}$ and $\overline{f}_{p,m}$, respectively.

The remainder of this section presents the signal processing stages in the order they are applied to the input signal, and defines the analysis parameters that are not manually adjusted.

5.1. Sub-band Division

The first processing stage consists of the division of the signal in R frequency sub-bands. This strategy is adopted so to allow a finer control of the analysis parameters and a differentiated treatment of the estimation errors for each frequency band. The sub-band decomposition is obtained through the method presented in [18], in which the iterative application of the structure exhibited in Fig. 2 to the signal $x_l[n]$ is used. The filter $h[n]$ is a linear-phase low-pass filter with order 256. Its pass-band goes up to 0.4π rad and its stop-band starts at 0.45π rad (normalized frequency) with a target attenuation of 100 dB.

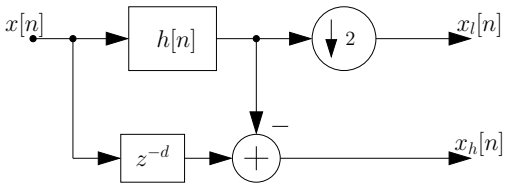

Figure 2: Basic element of the so-called Laplacian Pyramid iteratively used to decompose the spectrum into R octave sub-bands. The signal $x[n]$ is divided into two sub-bands: $x_h[n]$ and $x_l[n]$ that contain, respectively, the high- and low-frequency content. Octave subdivision of the lowest-frequency sub-band can be obtained through the re-application of this structure to the signal $x_l[n]$.

In Fig. 3 the magnitude response of the equivalent sub-band filters for $R = 4$ is shown. It can be noticed that a superposition between adjacent bands occurs. The sub-bands are indexed in decreasing order of pass-band centroid, with sub-band index $r = 1$ indicating the octave sub-band with the widest bandwidth.

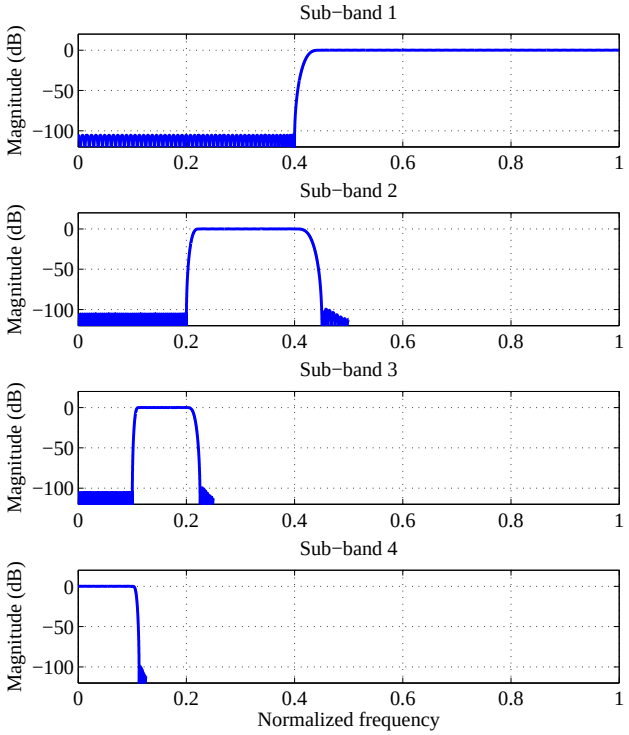

Figure 3: Magnitude of the frequency response of the 4 (equivalent) sub-band filters obtained from the structure in Fig. 2. The sub-bands are exhibited from DC to half the normalized sampling frequency of the input signal.

Each of the processing stages described from Section 5.2 to Section 5.6 will be executed individually for each sub-band.

[1] Above the loudness threshold [16] (Sensation Level) for a given frequency.

5.2. Time-Frequency Mapping

The time-frequency mapping of each sub-band signal is obtained through an STFT (see Eq. (3)). For this, the sub-band signal is segmented into frames of length N samples, using a Hann window. For the r-th sub-band, the hop size can be obtained through

$$H^r = \left\lfloor \frac{1}{F_{\mathrm{s_p}}} \times 2^{-(r-1)} F_{\mathrm{s}} \right\rfloor, \tag{7}$$

where $\lfloor b \rfloor$ denotes the floor operation, F_{s} (in Hz) is the sampling rate of the source-signal, and $2^{-(r-1)}F_{\mathrm{s}}$ is the sampling rate of the signal in sub-band r. Eq. 7 guarantees a parameter sampling rate at least equal to $F_{\mathrm{s_p}}$, thus respecting the specifications defined in Section 4. A disadvantage of this method is that the generated track parameters will possess a different sampling rate in each sub-band. A solution to this issue is given in Section 5.7.

Once the hop is defined, the spectrum of a frame can be obtained through a DFT with length $Z_p N$ (N in samples) for each block, where Z_p is the zero-padding factor. In Section 5.4, the values for Z_p are defined. The window length N is chosen differently for each sub-band and signal under analysis.

5.3. Peak Detection

A scheme that uses two thresholds, one absolute and the other local, was employed in the spectral peak detection task. The overall strategy enforces that only peaks well above an estimated noise floor are detected, as specified in Section 4).

Both thresholds are obtained through the application of the method described in [13] to the magnitude spectrum of the signal. For each frame m, the method yields a noise floor, which is then adopted as the local threshold $T_{\mathrm{l}}[m,k]$. The global threshold $T_{\mathrm{a}}[k]$ is the smoothed noise floor obtained by averaging the local threshold over noise-only frames.

The sinusoidal peaks are then obtained through the selection of the points in the magnitude spectrum that satisfy

$$\begin{aligned} |X[m,k]| &> |X[m,k-1]| \\ |X[m,k]| &> |X[m,k+1]| \\ |X[m,k]| &> 3T_{\mathrm{a}}[k] \\ |X[m,k]| &> d[m,k]T_{\mathrm{l}}[m,k]. \end{aligned} \tag{8}$$

The first two conditions ensure that only spectral peaks are selected. The third condition restricts selection to peaks that are three times above the level of the estimated noise floor. The fourth condition is similar to the third, but more flexible, since the multiplier $d[m,k]$ allows to manually adjust the local threshold.

Optionally, the number of peaks that satisfy the criteria in (8) can be forcefully limited to an arbitrary maximum. If so, only those peaks with largest magnitudes are selected in a given frame.

5.4. Frequency Estimation

A frequency estimator based on quadratic interpolation [19] along spectral peaks was selected, since its estimation error can be controlled through the zero-padding factor Z_p. Thus, peak frequency errors can be forced to fall below a margin specified by the minimum of Eq. (5) in the frequency range of the sub-band in question , as was shown in [19].

5.5. Magnitude and Phase Estimation

Once the frequency is estimated, the magnitude and phase values for the p-th peak can be computed according to [20]

$$\overline{S}_{p,m} = \frac{1}{N} \sum_{n=0}^{N-1} s_m^r[n] e^{-jn\overline{f}_{p,m}}, \tag{9}$$

where $\overline{f}_{p,m}$ is in radians per sample and $s_m^r[n]$ is the source-signal at frame m and sub-band r. The procedure used to divide the signal into sub-bands can distort the estimated amplitude and phase values, specially for frequencies close to the transition bands. To compensate for such distortion it suffices to compute

$$\overline{S}_{p,m}^{\mathrm{mod}} = \frac{\overline{S}_{p,m}}{\overline{H}(e^{j\overline{f}_{p,m}})}, \tag{10}$$

where $\overline{H}(.)$ is the complex frequency response of the equivalent sub-band filter. In practice, it is sufficient to apply the compensation of Eq. (10) only for the final decomposition branch of a given sub-band, i.e. the filter $h[n]$ decimated by 2.

The amplitude of the p-th peak at frame m is estimated as $\overline{A}_{p,m} = |\overline{S}_{p,m}^{\mathrm{mod}}|$, and the phase as $\overline{\Theta}_{p,m} = \angle \overline{S}_{p,m}^{\mathrm{mod}}$.

5.6. Partial Tracking

In the partial tracking stage, a modified version of the McAulay & Quatieri [1] (MQ) algorithm was used. It operates on a frame-by-frame basis.

The following steps summarize the operation of the algorithm for a given frame m. If $m = 1$ or there was no active track in frame $m-1$, create a new track starting at each peak detected in the current frame. Otherwise:

1. Select the track i with the largest amplitude $\hat{A}_{i,m-1}$ in frame $m-1$, provided it has not been selected in this frame yet.
2. Among the peaks in the current frame that have not been assigned to any track, look for the ones that satisfy the inequality $|\hat{f}_{i,m-1} - \overline{f}_{p,m}| \leq \Delta_f$.
   (a) If there are none, update track i according to $\hat{f}_{i,m} = \hat{f}_{i,m-1}$ and $\hat{A}_{i,m} = \hat{A}_{i,m-1}$ and tag this frame. If the track has D consecutive tagged frames, then it is considered finished, and its final trajectories are trimmed to exclude the tagged frames at its end. If a finished track has a length inferior to E frames, it is discarded.
   (b) Otherwise, select as a track continuation the peak among those found that minimizes

$$J = (1-\kappa)\frac{|\hat{f}_{i,m-1} - \overline{f}_{p,m}|}{\hat{f}_{i,m}} + \kappa \frac{|\hat{A}_{i,m-1} - \overline{A}_{p,m}|}{\hat{A}_{i,m}}, \tag{11}$$

   where $\kappa \in [0,1]$ is a parameter that controls the weight of the relative frequency and amplitude differences in the cost function. Update the track parameters according to $\hat{f}_{i,m} = \overline{f}_{\hat{p},m}$ and $\hat{A}_{i,m} = \overline{A}_{\hat{p},m}$.
3. Return to step 1, until all tracks have been selected.
4. Start a new track from every peak in the current frame not assigned to any track.

The track parameters at the tagged frames are replaced by suitable values through an interpolation algorithm described in Section 5.7.1. The parameter D has been chosen as 4 frames to meet the specification of the maximum number of frames that can be interpolated (see Section 4).

5.7. Post-processing

After obtained, the set of tracks in each sub-band can be submitted to the following modifications:

1. Interpolation across missing values for each track.
2. Manual selection of tracks by visual inspection (in comparison with a high resolution spectrogram) followed by listening to the synthesized signal.
3. Reduction of the sampling rate of the parameters to F_{s_p}.

A description of the algorithms employed in steps 1 and 3 follows.

5.7.1. Missing Data Interpolation

An interpolation algorithm is necessary to conceal the temporal gaps in the trajectories of the parameters. The adopted interpolation method [21] uses an AR model to represent the temporal evolution of the parameters in question. Hence, it is well suited for capturing the quasi-periodical nature of track trajectories, both in frequency (*vibrato* playing) and in magnitude (*tremolo* playing). The method formulation is presented here for interpolation of a frequency trajectory, but it can be directly replicated for the amplitude and phase trajectories. Given a vector $\hat{\mathbf{f}}_i$ containing the M samples of the (supposedly stationary) frequency trajectory of track i, its AR model of order q can be written in matrix form as

$$\mathbf{e} = \mathbf{P}\hat{\mathbf{f}}_i, \tag{12}$$

where $\mathbf{e}$ is a vector of length $(M-q)$ and

$$\mathbf{P} = \begin{bmatrix} -a_q & \cdots & -a_1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -a_q & \cdots & -a_1 & 1 \end{bmatrix} \tag{13}$$

an $(M-q)\times M$ matrix containing the models coefficients. Considering there are $l << (M-l)$ missing samples in vector $\hat{\mathbf{f}}_i$, it can be divided into two vectors: $\hat{\mathbf{f}}_i^k$ containing the $M-l$ known samples and $\hat{\mathbf{f}}_i^u$ containing the unknown l samples. Through an adequate partition of the matrix $\mathbf{P}$, one can write

$$\mathbf{e} = \mathbf{P}\hat{\mathbf{f}}_i = \mathbf{P}^k\hat{\mathbf{f}}_i^k + \mathbf{P}^u\hat{\mathbf{f}}_i^u, \tag{14}$$

where $\mathbf{P}^k$ and $\mathbf{P}^u$ are partitions of $\mathbf{P}$ corresponding to $\hat{\mathbf{f}}_i^k$ and $\hat{\mathbf{f}}_i^u$, respectively. The vector $\hat{\mathbf{f}}_i^u$ is obtained [21] via the minimization of the quadratic error $\mathbf{e}^T\mathbf{e}$ w.r.t $\hat{\mathbf{f}}_i^u$. For track trajectory interpolation up to 4 samples, an order $q=4$ was found to be sufficient to ensure satisfactory results.

5.7.2. Track Resampling

Recalling that the parameter trajectories have been obtained at a different sampling rate depending on the sub-band in which they were estimated (see Section 5.2), a simple resampling method was devised.

The sampling rate reduction of the parameter (amplitude, frequency, and unwrapped phase) streams assumes the underlying continuous trajectories as piece-wise linear. Once this approximation is made, the trajectories are resampled so to obtain samples spaced exactly 2 ms apart (equivalent to the 500 Hz sampling rate). In order to ensure the synchronism of the tracks, the starting time sample of each track is rounded towards the nearest integer multiple of 2 ms. This simple method can be adopted because of the high oversampling of track parameters (see Section 4).

6. OBTAINED TRACKS

In this section, information is presented regarding how the TDB was constructed using the analysis methods discussed in the previous section. The procedure used to adjust the parameters is also briefly discussed.

For each source-signal, the analysis parameters were initialized according to the standard values given in Table 4. With the exception of the number of sub-bands R, all other parameters are identically initialized for every sub-band.

Table 4: Initial values of the analysis parameters.

Stage	Parameter
Sub-band Division	$R=4$
STFT	N samples, equivalent to 40 ms
Peak Detection	$d[m,k]=3$
Tracking	$\Delta_f = 0.03$, $E=15$ frames, and $\kappa = 0.5$

The first parameter to be adjusted is R, according to the following steps: (1) The signal is divided into 4 sub-bands[2]; (2) An STFT of the signal in sub-band 4 is obtained; (3) If partials in the spectrogram are visually too clustered, R can be incremented.

Once R is chosen, the remaining analysis parameters are adjusted for each sub-band. The signal in the sub-band with highest index is analyzed first, and so on. The window length can be adjusted to balance the time and frequency resolutions of the obtained spectrogram.

The local threshold $d[m,k]$ in the peak detection stage, can be raised during abrupt changes in the source-signal, such as an *onset*. Additional information about the signal under analysis, such as harmonicity, can aid to set this threshold.

For the partial tracking stage, the analysis parameters are chosen in order to ensure the similarity between the tracks and the partials as observed in the spectrogram. In practice, the parameters were adjusted so that the tonal part of the signal was correctly modeled, even if that provoked the creation of spurious tracks which can be easily removed by the manual post-processing step.

For each signal, the adjustments performed during analysis were recorded in Matlab® scripts. An example of such script and more detailed information regarding the obtained tracks are available in [7]. The set of tracks associated with a given signal is made available in a .mat file.

7. VALIDATION OF THE ANALYSIS METHODS

In this section, the synthetic signals from Level 0 are used to check if the analysis methods described in Section 5 are able to meet the

[2] An exception was taken for signal 2 in level 3, which could be well-analyzed with $R=1$.

TDB specifications defined in Section 4.2. In particular, specification conformance of the maximum tolerable estimation errors for track frequency, amplitude, onset time and offset time has been verified. For that, the following test procedure is used: (1) The track parameters for the signal under test are estimated by means of the methods described in Section 5; (2) A point-wise comparison between estimated frequencies trajectories and reference track parameters is made.

7.1. Reference Tracks

The reference tracks used to validate the TDB are obtained from the control data of the synthesizer used to generate the signals at Level 0 (see Section 3.2.1). The control signals are originally available with the same sampling rate as the synthesized signals, i.e. 44.1 kHz. Sample rate reduction to 500 Hz to obtain the reference tracks is carried out via the scheme described in Section 5.7.

After the sampling rate modification, the onset (offset) of each track is determined by the first instant in which the amplitude trajectory of the track is greater (smaller) than a pre-defined threshold. In the case of the Vibrato and Tremolo signals, the threshold is defined by the quantization noise due to their fixed-point binary representation. As both signals are quantized with 16 bits, the quantization level is approximately -96 dB (Full Scale). For the Note Sequence signal, the threshold varies with track frequency, according to the noise spectrum shape.

7.2. Frequency Estimation

In order to verify if the estimation error is smaller than the tolerance specified in Section 4.2.1, a test using the Vibrato signal is performed. Let $f_{i,m}$ be the i-th frequency trajectory of the synthesized signal at frame m. The relative estimation error can be obtained by

$$\frac{|\hat{f}_{i,m} - f_{i,m}|}{f_{i,m}}. \tag{15}$$

The amount by which the relative error surpasses the maximum tolerance $e^{\text{freq}}(f_{i,m})$ (Eq. (5)) is calculated, for the frequency $\hat{f}_{i,m}$. Then, the following metric was adopted

$$\Gamma(f_{i,m}, \hat{f}_{i,m}) = \frac{|\hat{f}_{i,m} - f_{i,m}|}{f_{i,m}} - e^{\text{freq}}(f_{i,m}). \tag{16}$$

In Fig. 4, the distribution of Γ for all points of the complete set of frequency trajectories of the signal under test can be seen. Note that only negative samples of Γ were obtained, indicating that the estimation error was smaller than the tolerance for the signal under test. Hence, it can be concluded that the analysis methodology is capable of providing frequency estimates for the tracks of the test signal within the specifications defined for the TDB.

7.3. Amplitude Estimation

The estimated amplitude trajectories for the Tremolo signal were compared with the corresponding reference trajectories. Being $A_{i,m}$ the i-th amplitude trajectory of the synthesized signal at frame m, the relative amplitude error is obtained as

$$\frac{|\hat{A}_{i,m} - A_{i,m}|}{A_{i,m}}. \tag{17}$$

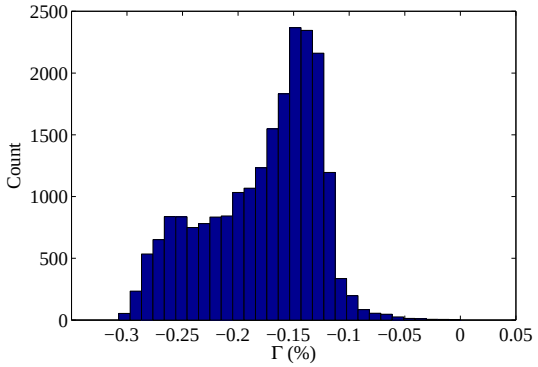

Figure 4: Distribution of the metric $\Gamma(f_{i,m}, \hat{f}_{i,m})$ in percentage for the tracks of the Vibrato signal.

In Fig. 5, a histogram of the relative amplitude error for all measured amplitudes over all identified tracks for the signal under test can be seen. Note that, in this example, the relative error does not exceed the maximum tolerance of 20 % defined in Section 4.2.2. Hence, the analysis methodology is capable of providing amplitude estimates for the tracks of the test signal within the specifications defined for the TDB.

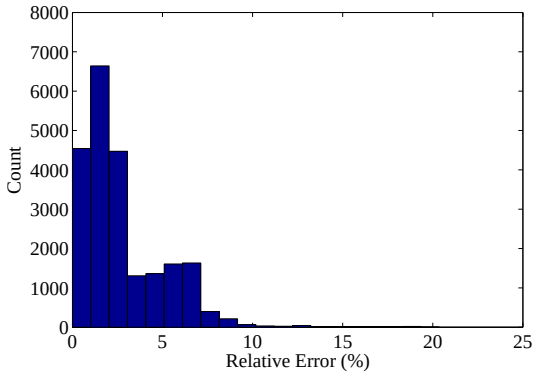

Figure 5: Distribution of the relative estimation error for the Tremolo signal.

7.4. Onset and Offset Time Estimation

The error in track onset/offset time estimation was validated using the Note Sequence signal. For this, the absolute difference between the onset time instants of the estimated tracks and those of the reference tracks is computed. A similar calculation was performed for the offsets.

In Fig. 6, the histogram of the measured absolute errors, calculated for each onset and offset of all identified tracks in the signal under test is exhibited. The largest error found was of 8 ms, duration that is shorter than the tolerable maximum error of 20 ms. A mean time support error of 1.2 ms was found for the tracks of the test signal. Hence, it can be concluded that the analysis methodology is capable of providing onset and offset estimates within the specifications defined for the TDB.

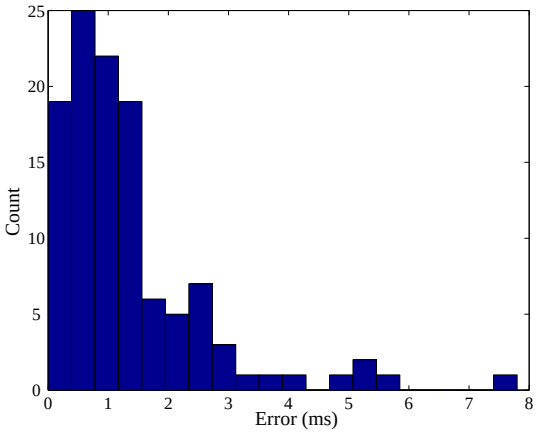

Figure 6: Distribution of the absolute error for the detection of onset and offset of each track for the Note Sequence signal.

8. CONCLUSION

This paper described a semi-supervised methodology to produce a database (called TDB) containing sinusoidal tracks extracted from both synthetic and pre-recorded musical signals. The TDB was designed to serve as a tool for comparison of different sinusoidal analysis systems. For this, prior to the obtention of the tracks, their desired attributes were carefully specified. Upper bounds for the estimation errors in the amplitude and frequency trajectories of each track were also defined. The employed estimators were validated by means of the synthetic signals in TDB, for which the referential trajectories were previously known. The estimators may conceivably perform similarly for the remaining signals of the database.

The applicability scope of the TDB is determined by its specifications: of course, the database is not suitable to assess sinusoidal analysis systems expected to surpass the accuracy dictated by the psychoacoustic criteria used in the construction of the TDB itself. Nevertheless, if judiciously used, the TDB is a new tool for fair evaluation of sinusoidal modeling algorithms.

9. REFERENCES

[1] R. McAulay and T. Quatieri, "Speech analsysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech Sig. Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[2] X. Serra and J. Smith III, "Spectral modeling synthesis: A sound analysis/synthesis system based on deterministic plus stochastic decomposition," *Comput. Music J.*, vol. 14, no. 4, pp. 12–24, Winter 1990.

[3] M. Wright, J. Beauchamp, K. Fitz, X. Rodet, A. Röbel, X. Serra, and G. Wakefield, "Analysis/synthesis comparison," *Organized Sound*, vol. 5, no. 3, pp. 173–189, Dec. 2000.

[4] M. Lagrange and S. Marchand, "Assessing the quality of the extraction and tracking of sinusoidal components: Towards an evaluation methodology," in *Proc. of the 9th Int. Conf. on Digital Audio Effects*, Montreal, Canada, Sept. 2006.

[5] P. Herrera-Boyer, "Setting up an audio database for music information retrieval benchmarking," in *The MIR/MDL Evaluation Project White Paper Collection*, J. S. Downie, Ed., pp. 53–55. 3rd edition, 2005.

[6] M. Goto, "Development of the RWC music database," in *Proc. of the 18th Int. Congress on Acoust.*, Kyoto, Japan, Apr. 2004, vol. I, pp. 553–556.

[7] TDB, "TDB companion webpage," 2010, available at: `http://www.gpa.lps.ufrj.br/tdb.html`.

[8] A. Horner and L. Ayers, "Modeling acoustic wind instruments with contiguous group synthesis," *J. Audio Eng. Soc.*, vol. 46, no. 10, pp. 868–879, Oct. 1998.

[9] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. of the 4th Int. Conf. on Music Information Retrieval*, Baltimore, USA, 2003, pp. 229–230.

[10] D. Ellis, "A perceptual representation of audio," M.Sc. Thesis, Massachusetts Institute of Technology, Feb. 1992.

[11] M. Lagrange, S. Marchand, and J.-B. Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 5, pp. 1625–1634, July 2007.

[12] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sept. 2005.

[13] N. Laurenti and G. De Poli, "A nonlinear method for stochastic spectrum estimation in the modeling of musical sounds," *IEEE Trans. Acoust. Speech Sig. Process.*, vol. 2, no. 15, pp. 531–541, Feb. 2007.

[14] B. Moore, R. Peters, and B. Glasberg, "Detection of temporal gaps in sinusoids: Effects of frequency and level," *J. Acoust. Soc. Am.*, vol. 93, no. 3, pp. 1563–1570, Mar. 1993.

[15] C. Wier, W. Jesteadt, and D. Green, "Frequency discrimination as a function of frequency and sensation level," *J. Acoust. Soc. Am.*, vol. 61, no. 1, pp. 178–184, Jan. 1977.

[16] B. Moore, *An Introduction to the Psychology of Hearing*, Elsevier, 5th edition, 2004.

[17] W. Jesteadt, C. Wier, and D. Green, "Intensity discrimination as a function of frequency and sensation level," *J. Acoust. Soc. Am.*, vol. 61, no. 1, pp. 169–177, Jan. 1977.

[18] S. Levine, T. Verma, and J. Smith III, "Multiresolution sinusoidal modeling for wideband audio with modifications," in *Proc. of the 1998 IEEE Conf. Acoust., Speech, and Sig. Process.*, Washington, USA, May 1998, vol. 6, pp. 3585–3588.

[19] M. Abe and J. Smith III, "Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks," in *Pres. at the 117th Conv. of AES*, San Francisco, USA, Oct. 2004, vol. 58, pp. 104–117.

[20] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results, and an application," *IEEE Trans. Sig. Process.*, vol. 48, no. 2, pp. 338–352, Feb. 2000.

[21] S. Godsill and P. Rayner, *Digital Audio Restoration*, Springer, 1998.