Audio Engineering Society

Convention Paper

Presented at the 125th Convention
2008 October 2–5 San Francisco, CA, USA

Parametric Interpolation of Gaps in Audio Signals

Alexey Lukin[1], and Jeremy Todd[2]

[1] *Dept. of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia*

[2] *iZotope Inc., Cambridge, MA 02139, US*

Correspondence should be addressed to Alexey Lukin (lukin@graphics.cs.msu.ru)

ABSTRACT

The problem of interpolation of gaps in audio signals is important for the restoration of degraded recordings. Following the parametric approach over a sinusoidal model recently suggested in JAES by Lagrange et al., this paper proposes an extension to this interpolation algorithm by considering the interpolation of a noisy component in a "sinusoidal+noise" signal model. Additionally, a new interpolator for sinusoidal components is presented and evaluated. The new interpolation algorithm is suitable for a wider range of audio recordings than just the interpolation of a sinusoidal signal component.

1. INTRODUCTION

The problem of interpolation of time-frequency regions is important for the restoration of audio data degraded by interfering events of significant duration. Typical time-domain interpolation methods use linear predictive coding (LPC) to predict the signal "into the gap" from each side of the gap [1], [2]. Two (forward and backward) extrapolated predictions are crossfaded to form a smooth transition. LPC extrapolation is able to model a mixture of sinusoids with fixed frequencies on each side of the gap. However no linking of such sinusoids is attempted. Also, LPC extrapolation is not using joint information from both sides of the gap to recover a realistic amplitude envelope of the interpolated signal.

An improved time-domain method called LSAR (least-squares autoregressive interpolation) is proposed in [3]. It is based on LPC, but uses joint information from both sides of the gap, and also contains the Expectation Maximization algorithm to jointly optimize AR coefficients and the interpolated audio signal.

A well-known problem with algorithms based on LPC is over-smoothing of the interpolated signal due to lack of proper excitation [4]. This happens because LPC only models a stationary tonal part of the signal. Some solutions are suggested in [4] and [5]: the excitation signal in the gap is either synthesized as random noise [5] or copied from the surrounding section of the audio signal [4]. These improvements allow for very realistic synthesis of both tonal and noisy signal components — but only when stationarity assumption holds.

However when the signal departs from stationarity, most time-domain interpolation methods are facing difficulties, especially when the length of the interpolated gap is above 20 ms [3]. More successful are the algorithms considering a 2-dimensional time-frequency structure of a signal (Fig. 1). They are referred to as spectral interpolation algorithms.

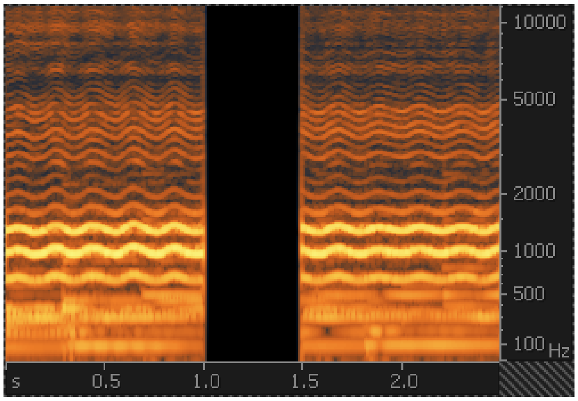

Fig. 1: 500-ms gap in the sustained vocal note.

Simple algorithms for spectral interpolation include multiband interpolation of subsampled spectrogram data. In [3], an algorithm is presented that works on the STFT (short-time Fourier transform) data and independently interpolates each frequency channel of spectral coefficients using a LSAR interpolator. Such simple algorithms work reasonably well for stationary tonal signals, but fail to correctly resolve pitch changes or reconstruct noisy sections of audio data in the interpolated segment (Fig. 2).

A parametric approach has been suggested by Maher in [6]. It uses a McAulay–Quatieri sinusoidal model [7] to identify tonal components in the signal. After tonal components on both sides of the gap are matched, a polynomial extrapolation[1] of their trajectories inside the gap is performed.

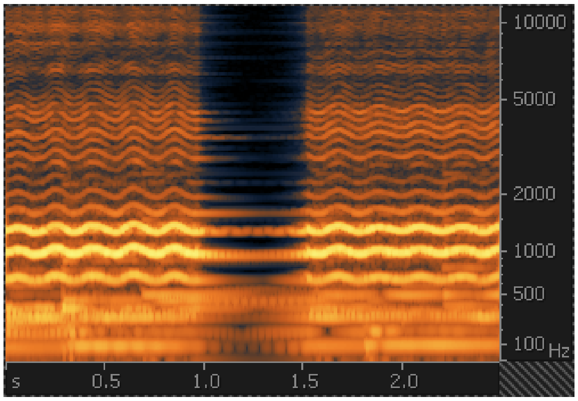

Fig. 2: Interpolation by a multiband LSAR algorithm.

Recently, an improvement of Maher's method has been published by Lagrange et al. in [8]. This new method is able to correctly interpolate pitch variations in the audio signal, such as vibrato or pitch slide effects, by using LPC (linear predictive) extrapolation of sinusoidal trajectories (Fig. 3). While being a significant improvement, this approach still only models the tonal part of the signal and does not address the problem of interpolation of noisy segments.

In this paper, an extension to the Lagrange's method is proposed, adding the interpolation of the noisy residual signal. A new interpolator for partials trajectories is also suggested to replace the LPC method of [8].

The proposed improvements make the interpolation algorithm suitable for a wider range of audio recordings than just the interpolation of a sinusoidal signal component, because the improved algorithm is able to interpolate noisy and mixed-content signals.

However it should be mentioned that the presented algorithm still relies on smooth changes of signal parameters inside the interpolated interval. It is unable

[1] Throughout this paper, "interpolation" means filling in missing data using information from the left and right sides of the gap. Similarly, "extrapolation" means filling in missing data using information from only one side of the gap.

to recover a transient that is completely missing, nor is it able to interpolate a signal which behaves unpredictably. A different class of gap filling algorithms, based on a waveform substitution, is able to recover repeating transients using the search of similar examples in the audio recording. However such algorithms assume the repeatable nature of audio signals and may have problems with smooth blending of a recovered section into the gap.

2. SINUSOIDAL INTERPOLATION

Our interpolation of a sinusoidal signal component closely follows Lagrange's method [8].

The first step is the identification of tonal signal components at both sides of the region (gap) to be interpolated. The signal is modeled as a sum of time-varying sinusoidal components ("partials"), as described by McAulay–Quatieri in [7].

$$s(t) = \sum_{k=1}^{K} P_k(t) = \sum_{k=1}^{K} A_k(t) \cos(\phi_k(t)) \qquad (1)$$

$$\phi_k(t) = \phi_k(0) + 2\pi \int_0^t f_k(u)\, du \qquad (2)$$

Each partial $P_k(n)$ detected from the STFT spectrum is represented by its amplitude envelope $A_k(n)$, frequency $f_k(n)$ and phase $\phi_k(n)$ functions in time:

$$P_k(n) = \{A_k(n), f_k(n), \phi_k(n)\} \qquad (3)$$

Here k is the index of the partial, and n is the discrete time index.

2.1. LPC Method

When the partials are identified on each side of the gap and their parameters (trajectories) are detected, the second step comprises extrapolation (prediction) of the trajectory of each partial inside the gap. In [8], the Burg method is used for estimation of LPC model parameters from a limited number of samples. When LPC parameters are evaluated, linear predictive modeling of the amplitude envelope and frequency trajectory is carried out inside the gap for all partials from the left and right sides of the gap.

The third step is to match partials extrapolated from left and right sides of the gap. This matching is based on evaluation of the difference in predicted frequency and amplitude trajectories of each pair of "left" and "right" partials. Some partials with large differences are allowed to remain unmatched (for example, some partials could have completely faded out during the gap, and they do not have a match on the right side of the gap).

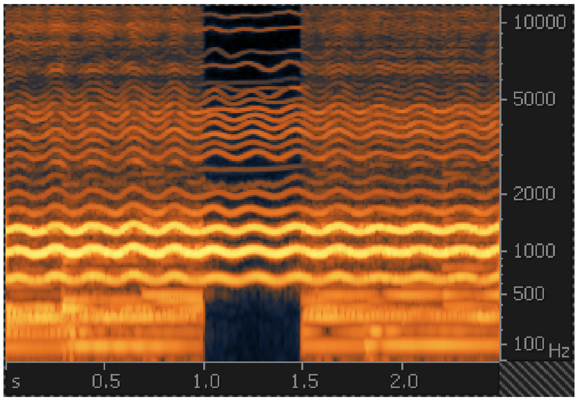

Fig. 3: Interpolation of partials by the Lagrange method.

Once the pairs of matching partials are identified, their trajectories are interpolated inside the gap by means of a crossfaded forward/backward LPC extrapolation, and the final synthesis of partials is performed from these interpolated trajectories (Fig. 3).

Our contribution to the method is the consideration of other, more powerful methods of extrapolation/interpolation of partials trajectories. While the crossfaded forward and backward LPC prediction suggested in [8] produces reasonably good results, it considers the left and right sides of every partial separately. As a result, the extrapolation of a partial from one side of the gap does not meet the known trajectory of that partial on the other side of the gap. This problem has been alleviated by crossfading in [8].

2.2. LSAR Method

A method using a LPC model and utilizing information from both sides of the gap is described in [3]; it is known as LSAR (least-squares autoregressive) interpolation. This method not only minimizes the LPC prediction error inside the gap, but also ensures that the extrapolated trajectory meets the corresponding matching trajectory at the other side of the gap.

This is done by a joint minimization of LPC prediction error inside the gap and on both sides of the gap.

This method produces more consistent interpolation of trajectories, but it is still sensitive to inaccuracies in estimated partials trajectories.

2.3. DFT Method

Another interpolation method recently proposed by Meisinger et al. in [9], simplifies the LPC model by removing exponential envelope terms from the model, which yields a "sum of constant-amplitude sinusoids" model. An effective iterative algorithm exists for fitting this model to the known trajectory of a partial on both sides of the gap simultaneously. Here is a brief review of the method.

The known signal samples $f(n)$ are approximated by the parametric model:

$$g(n) = \sum_{i \in \Omega} c_i \psi_i(n) \qquad (4)$$

Here Ω describes the set of used basis functions $\psi_i(n)$, and c_i are expansion coefficients. The support area $\mathcal{L}$ for basis functions consists of two areas: $\mathcal{A}$ – where the signal samples are known, and $\mathcal{B}$ – where signal samples are to be estimated (gap); $\mathcal{L} = \mathcal{A} \cup \mathcal{B}$.

To determine optimal expansion coefficients c_i, the weighted error energy between $f(n)$ and $g(n)$ is evaluated at the support area $\mathcal{A}$:

$$E_{\mathcal{A}} = \sum_{n \in \mathcal{L}} w(n) \cdot (f(n) - g(n))^2 \qquad (5)$$

Here the weighting function $w(n)$ is non-zero only in $\mathcal{A}$, and it tries to emphasize the samples that are more important for extrapolation — usually those close to the edges of area $\mathcal{A}$, as given by some proximity metric ρ.

$$w(n) = \begin{cases} \rho(n), & n \in \mathcal{A} \\ 0, & n \in \mathcal{B} \end{cases} \qquad (6)$$

The iterative process of finding expansion coefficients c_i is derived in [9]. On each iteration, it modifies only one expansion coefficient — the one that minimizes the weighted error energy $E_{\mathcal{A}}$.

If basis functions $\psi_i(n)$ are selected to be basis functions of the DFT (discrete Fourier transform) on the segment $\mathcal{L}$, then an efficient computational algorithm exists for the expansion. Such a set of $\psi_i(n)$ is also a good fit for our needs, because sinusoidal basis functions of the DFT can compactly represent periodic changes in partials trajectories (such as vibrato and tremolo). Due to the arbitrary shape of the allowed support area $\mathcal{A}$ for the known data, the method can be used both for extrapolation of single-ended partials into the gap as well as interpolation of matched pairs of partials inside the gap.

In our algorithm, we remove a linear trend (including DC offset) from the detected trajectories of partials before interpolation/extrapolation. After that, we are running the DFT-based algorithm with just 2 iterations to produce a good approximation to frequency and amplitude trajectories (Fig. 4). Using more than 2 iterations over-fits the model to possibly inaccurate and noisy estimated partials trajectories, which is undesirable.

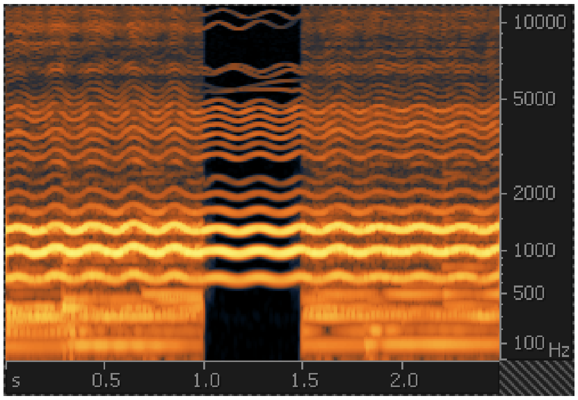

Fig. 4: Interpolation of partials by DFT method.

2.4. Joint Interpolation of Multiple Partials

Often partials belonging to the same musical instrument have similar pitch variations over time. When being interpolated independently, their trajectories can diverge quite significantly during the interpolated gap. Harmonic relationships of the partials are destroyed and the synthesized sound becomes unnatural (Fig. 4). So, when the signal is known to be mainly harmonic, a joint interpolation of frequency trajectories is proposed (Fig. 5). We do not develop any automatic algorithm here to detect such situations, leaving the decision on signal harmonicity up to the user.

It can be observed that frequencies of partials in harmonic relationship with each other are represented by parallel trajectories on a log-scale time-frequency plot. So, it is sufficient to only interpolate one partial in a log frequency scale and then its frequency-shifted trajectory would fit other partials. To improve robustness to possibly noisy and inaccurate trajectory data, we are averaging frequency trajectories of all reliably detected partials. Then the interpolation is carried out for the averaged frequency trajectory.

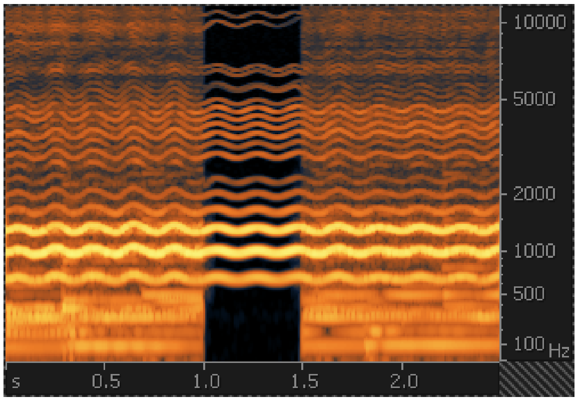

Fig. 5: Joint interpolation of partials by the proposed method.

Once the interpolated trajectory is calculated, we are shifting it on a log-frequency scale to fit other partials that took part in averaging. The interpolated averaged trajectory may not perfectly fit some of these individual partials due to non-strict harmonic relationships or noisy data. To improve smoothness at junctions of interpolated trajectories in zone $\mathcal{B}$ with original trajectories in zone $\mathcal{A}$, a short crossfade with LPC-predicted trajectories can be applied near the ends of the gap (Fig. 6), since LPC-predicted trajectories join the original data very smoothly. Thus, we are using both LPC and DFT interpolation algorithms for trajectories of partials.

3. RESIDUAL INTERPOLATION

After the sinusoidal part of the audio is modeled and interpolated, there is a need to interpolate the rest of the audio data that was not represented by the sinusoidal model. Such a residual corresponds to signal noise and minor harmonics not resolved by the sinusoidal modeling. For most real-world music signals, this residual component carries a significant part of the total signal energy, and omitting it from the synthesized gap leads to a significant drop in the overall loudness and change in the overall timbre (Fig. 5).

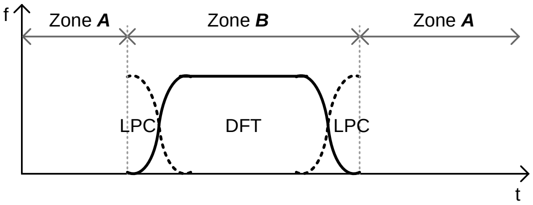

Fig. 6: Weighting windows for crossfading trajectories interpolated by LPC and DFT.

The first step is obtaining the noisy residual in zone $\mathcal{A}$, where the signal is known. This is achieved by subtracting the synthesized partials from the audio signal on both sides of the gap (Fig. 7).

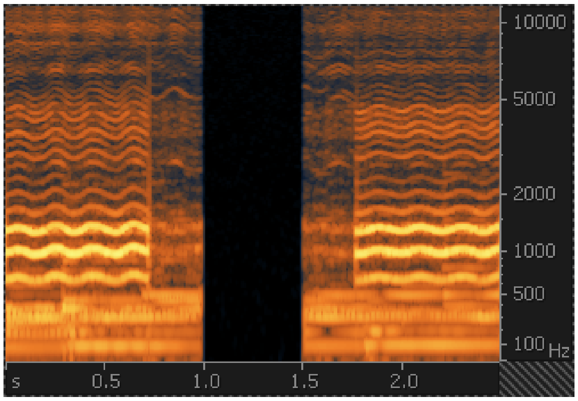

Fig. 7: Residual signal after subtraction of synthesized partials.

The second step is interpolation of the noisy residual into the gap. This is done by a multiband LSAR interpolator working on STFT coefficients of the residual signal.

Our filter bank is a Short-Time Fourier Transform with a 50-ms Hann window and a hop size being 1/8-th of the window size. Just as with a sinusoidal modeling, STFT window duration can be chosen higher for stationary audio material or lower for quickly changing material, but we are leaving this decision up to the user. The synthesis filter bank is quite similar: inverse STFT with 50-ms Hann windows.

Our implementation of a multiband LSAR method is working on STFT coefficients with LPC order 20, which covers approximately 125 ms of audio signal at each side of the gap (this is invariant of the sampling rate). Linear prediction coefficients are calculated using Levinson-Durbin recursion from a 38-point (240-ms) autocorrelation window. Since the autocorrelation window at least partially covers the gap, the resulting LPC coefficients are biased. A 3-iteration Expectation Maximization algorithm is run to reduce the bias and improve the quality of LSAR interpolation (see [3] for details).

In [5], a drawback of LSAR interpolation is described: in long interpolated gaps, the LSAR interpolator tends to underestimate the energy of the signal due to lack of proper excitation (Fig. 8).

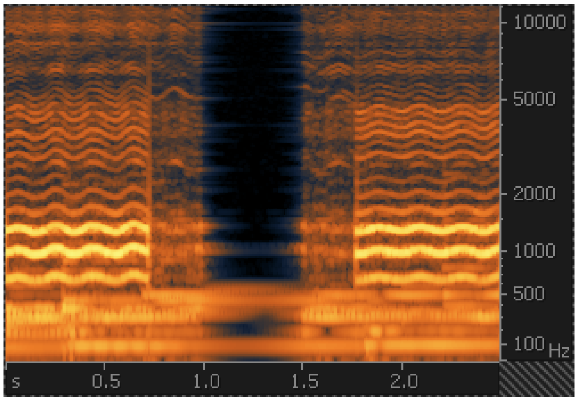

Fig. 8: Interpolation of the residual signal by a multiband LSAR method.

To address this problem, we suggest extraction of a linear trend (in time) from the magnitude envelope of the interpolated signal in every frequency band. Such a trend is calculated using signal magnitudes in zone $\mathcal{A}$ (see Fig. 10). Then, LSAR interpolation is only performed on a zero-mean residual signal after subtraction of a linear trend. Adding this trend back to the signal after LSAR interpolation ensures that the overall energy of the interpolated signal is preserved (compare Fig. 8 and 9).

Another problem with the LSAR algorithm is the performance of a multiband LSAR on noisy signals: the method fails to model inter-band correlations of STFT coefficients. Since STFT is a windowed transform, there exists a correlation (also known as "vertical phase coherence") between coefficients in neighboring frequency bins, both for noisy and sinusoidal signals. Interpolating frequency bands independently does not preserve this correlation, which results in loss of power and "phasiness" artifact for both noise and sinusoids.

To address the problem of vertical phase coherence, we suggest explicitly identifying residual frequency bands containing mostly noise, and substituting (properly correlated) STFT coefficients of a white noise in the synthesized signal at these frequency bands (Fig. 9). The power of noise is scaled according to the signal energy at the sides of the gap.

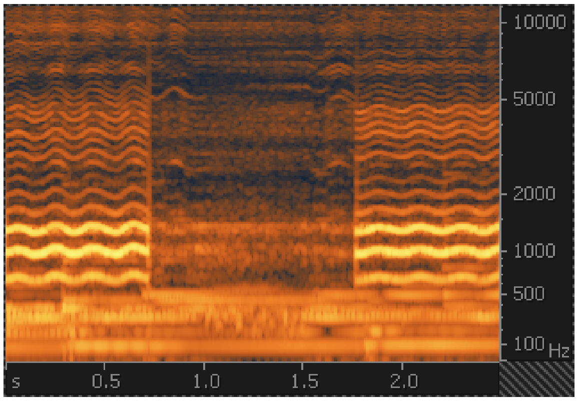

Fig. 9: Interpolation of the residual signal by the proposed method.

The proposed way for identifying regions of such noise substitution is based on comparing energy of LSAR interpolation (without linear trend extraction) to the level of the linear trend. Where the LSAR interpolation drops below the level of the linear trend, the STFT coefficients of white noise should be mixed in to compensate for the level drop (Fig. 10). This also improves vertical phase coherence, because phases of a substituted white noise are properly correlated.

After the residual signal is interpolated, it is mixed with interpolated sinusoidal components to produce the final result of the algorithm (Fig. 11)

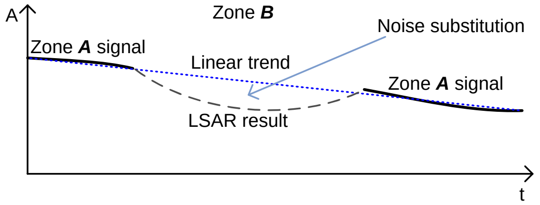

Fig. 10: Detection of regions for noise substitution.

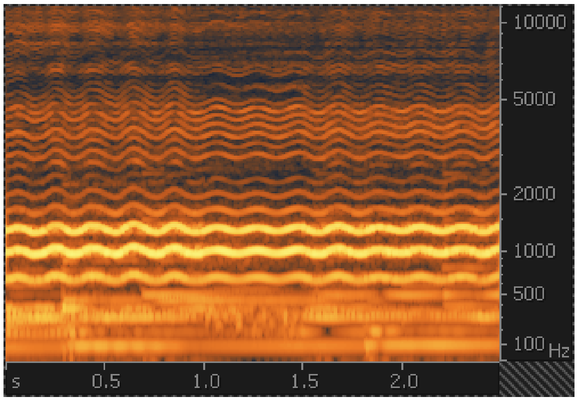

Fig. 11: Combining the interpolated residual with sinusoidal components.

4. EVALUATION AND CONCLUSION

We have evaluated the suggested algorithm on different synthetic and real-world samples.

It has been found that for gaps longer than 10–20 ms, the time-domain LSAR method usually cannot properly interpolate the signal and produces a drop in level [10].

A multiband version of LSAR works for gaps up to 100 ms if the signals are mostly tonal and stationary. If the signal contains significant noisy components, a multiband LSAR results in a level drop as well.

The sinusoidal interpolation algorithm of Lagrange [8] works well in connecting sinusoidal components for longer gaps, depending on the complexity of the harmonic structure. However it is only sufficient for instruments where the tonal part dominates and no significant noise is present.

The algorithm proposed in this paper provides a higher quality of interpolation on most samples and gap sizes because of its improved robustness of partials connection and synthesis of a noisy residual. In most cases, the version with joint interpolation of harmonics has been preferred to independent interpolation of harmonics, because it produces less spurious "detuning" of harmonics.

Additional examples of interpolation can be seen below in Fig. 12–17. More examples of interpolated spectrograms, audio samples and demo software can be obtained in [10].

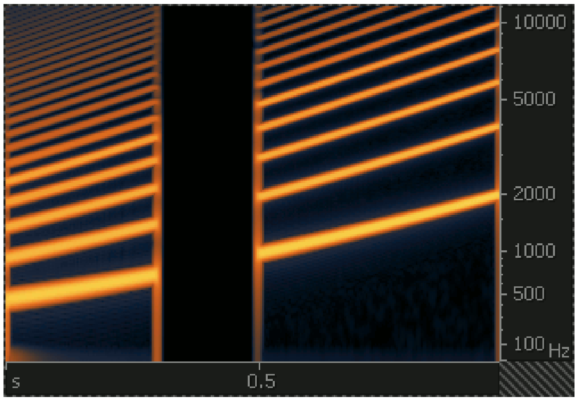

Fig. 12: 200-ms gap in the synthetic sliding tone.

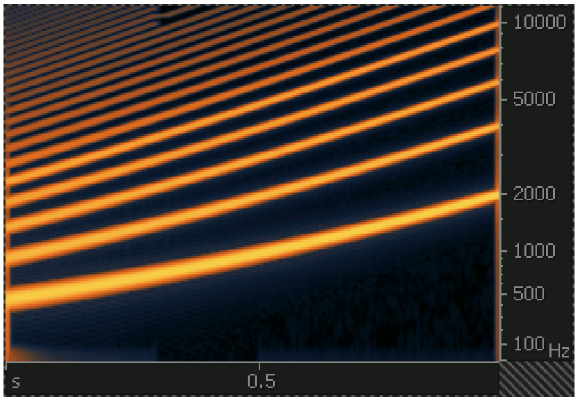

Fig. 13: Interpolated partials.

5. REFERENCES

[1] I. Kauppinen, J. Kauppinen, P. Saarinen, "A Method for Long Extrapolation of Audio Signals," Journal of the Audio Engineering Society, vol. 49, no. 12, pp. 1167–1180, Dec. 2001.

[2] P. Esquef, V Vealimeaki, K. Roth, I. Kauppinen, "Interpolation of Long Gaps in Audio

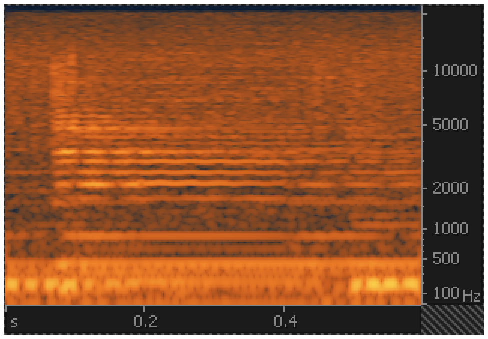

Fig. 14: Original noisy guitar chord.

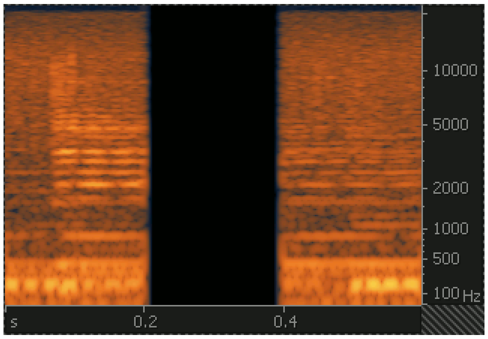

Fig. 15: 200-ms gap in the guitar chord.

Signals Using the Warped Burg's Method," Proceedings of the 6-th International Conference on Digital Audio Effects (DAFx-03), London, UK, Sept. 2003.

[3] S. Vaseghi, "Advanced Digital Signal Processing and Noise Reduction," Wiley, 2000, ISBN 0471626929.

[4] M. Niedzwiecki, K. Cisowski, "Smart Copying — A New Approach to Reconstruction of Audio Signals," IEEE Transactions on Signal Processing, vol. 49, no. 10, pp. 2272–2282, Oct. 2001.

[5] S. Godsill, P. Rayner, "Digital Audio Restoration — A Statistical Model-Based Approach," Springer-Verlag London Limited 1998, ISBN 3 540 76222 1.

[6] R. Maher, "A Method for Extrapolation of Missing Digital Audio Data," Journal of the Audio Engineering Society, 1994, vol. 42, no. 5, pp. 350–357.

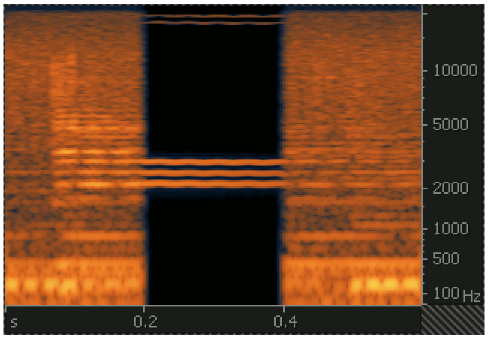

Fig. 16: Interpolated partials.

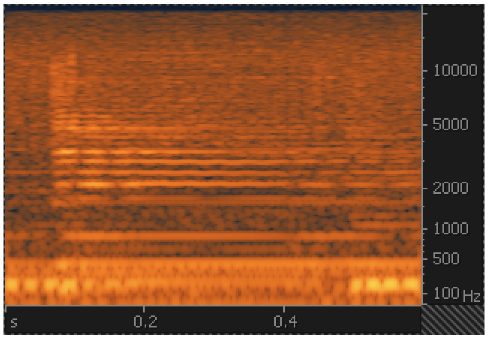

Fig. 17: Full interpolation (partials + residual).

[7] R. McAulay, T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," IEEE Transactions on Acoustics, Speech, Signal Processing, 1986, vol. 34, pp. 744–754.

[8] M. Lagrange, S. Marchand, J.-B. Rault, "Long Interpolation of Audio Signals Using Linear Prediction in Sinusoidal Modeling," Journal of the Audio Engineering Society, 2005, vol. 53, no. 10, pp. 891–905.

[9] K. Meisinger, A. Kaup, "Minimizing a Weighted Error Criterion for Spatial Error Concealment of Missing Image Data," Proceedings of IEEE International Conference on Image Processing (ICIP), Singapore, pp. 813–816, Oct. 2004.

[10] Demo web-page with spectrograms, audio files and software
http://www.izotope.com/tech/aes_interp