Mitigation of Long Gaps in Music Using Hybrid Sinusoidal+Noise Model with Context Adaptation

Maciej Bartkowiak
Chair of Multimedia Telecommunications
and Microelectronics, Poznań University of Technology
Polanka 3, 60-965 Poznań, Poland,
e-mail: mbartkow@multimedia.edu.pl

Bernard Latanowicz
Poznań Supercomputing and Networking Center,
Network Department,
Dąbrowskiego 79a, 60-529 Poznań, Poland,
e-mail: blatan@man.poznan.pl

***Abstract*—The paper describes a technique allowing to conceal or mitigate long gaps (up to 0.5 second) in music programs. Missing content is replaced by signal synthesized from spectral models using data surrounding the gap. Tonal components are synthesized using a sinusoidal model. A heuristic adaptive algorithm is employed to link model parameters across the gap. Prior to linking, sinusoidal partials are categorized as stable or variable, allowing to properly dealing with vibrato or glissando notes in music. The noise part is synthesized using a warped LPC model. Results of blind listening tests are reported.**

I. Introduction

Long gaps may occur by a transmission or storage of digital audio due to lost data packets, lost synchronization or physical damages of the storage medium. Also, a signal level overload during a music recording session usually leads to severely distorted data. These segments of damaged signal may be as long as tens or even hundreds of milliseconds (thousands of samples) and are very disturbing for the listener [1,2]. In general, reconstruction of such missing data through interpolation on a waveform level is an unrealistic challenge. Luckily, human perception of sounds relies mainly on the modulations carried by audio waveforms. Hence, for the gap to remain unnoticeable, perceptual signal properties should be reconstructed rather than the exact waveform. However, even this goal is quite elusive due to the fundamental non-stationarity of audio signals such as speech or music. The longer the gap the more of important semantic content is lost. Therefore, in cases of long gaps, only a concealment technique may be offered in order to decrease the level of listener annoyance. Usually, an artificial audio data segment is synthesized that matches the content preceding and following the gap in a perceptually seamless way.

The problem of missing audio data concealment has been addressed by many researchers, however most papers deal with quite short gaps and/or speech signals. The solution may have a form of extrapolation or interpolation, depending on the time constraints in particular application. For long gaps, the only realistic approach is interpolation. The general idea is typically to build two signal models using data segments surrounding the gap, find a correspondence between model parameters, form a time-varying model and finally synthesize a required number of samples to fit the gap. The existing approaches may be categorized into time-domain and frequency-domain.

In time domain methods, usually two separate high-order auto-regressive (LPC-like) models are built upon the available data segments [2,3,4]. Subsequently, both segments are extrapolated for a number of missing samples by recursively feeding the predictor with already predicted samples (a reversed time processing is applied to the segment after the gap). Finally, left and right-hand predictor outputs are merged with a simple cross-fade window in order to ensure smooth transition. Such approach allows to seamlessly reconstructing gaps of 25-50ms, however it works well only for stationary and strongly tonal sounds.

Frequency domain methods rely on frame-based short-time spectrum being slowly evolving compared to the signal in time domain [2]. This allows applying a simpler, low-order interpolation of the spectral content after calculating the FFT from short segments preceding and following the gap. The interpolated STFT frames replace those representing missing samples, and the signal is reconstructed using an overlap-add (OLA) procedure [2,5].

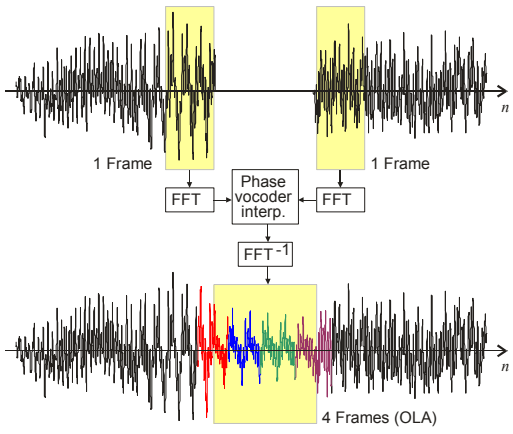

Fig. 1. Reconstruction of missing samples using spectral interpolation

For short gaps (below 25ms), a simple spectral interpolation according to the phase vocoder (PV) principle [6] may

be applied (fig. 1). This involves separate processing of amplitude and phase of corresponding spectral bins. Magnitudes are linearly interpolated between known values of border frames, while phase is linearly extrapolated from last frame, taking into account particular bin frequency and time advance of consecutive frames. This approach does not offer satisfactory quality for longer gaps due to the already mentioned non-stationarity of music (fig. 2). Furthermore, phase continuity principle may be applied only for tonal spectral components, because noise is usually characterized by an incoherent phase spectrum. Since PV does not distinguish tonal partials from noise, longer interpolated segments are often rendered as unnatural pseudo-periodic "buzzy" sound.

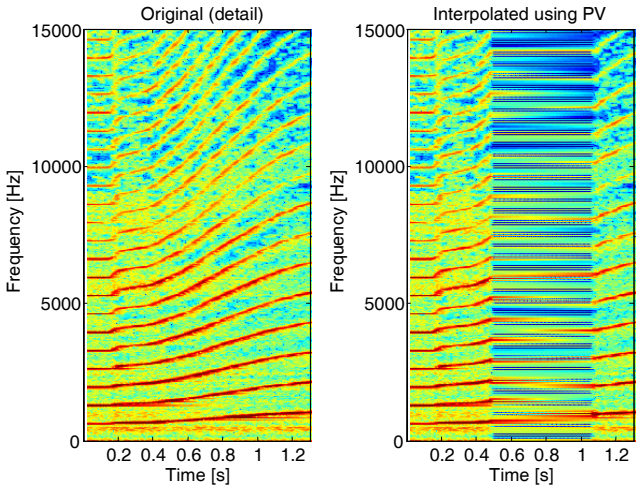

Fig. 2. Reconstruction problems due to non-stationary of music signal and inappropriate dealing with phase for noise component. Note that wrong harmonic partials are merged in a glissando note

A more practical approach is to process only spectral peaks corresponding to sinusoidal-like partials. Due to the variability of sounds, such peaks in data segments surrounding the gap usually do not correspond 1:1, therefore a peak matching procedure is required. The natural candidate method for detection and matching of peaks is the sinusoidal model (SM) [7,8]. Within SM, sinusoidal-like components are detected in the short-time magnitude spectra of consecutive signal frames. Subsequently, corresponding parameters (local average frequency, amplitude and phase) are estimated and linked by a tracking algorithm that takes into account parameter difference in consecutive frames. The sinusoidal trajectories obtained in this way allow resynthesizing the signal (1), with accuracy depending on estimation errors and successful tracking,

$$\hat{x}(t) = \sum_{n=1}^{N} A_n(t) \sin\left(2\pi \int_0^t f_n(\tau)\, d\tau + \varphi_n(0)\right), \qquad (1)$$

where n is a sinusoidal partial number, $A_n(t)$ and $f_n(t)$ are its interpolated parameters in current frame, $\varphi_n(0)$ is the phase of partial at the end of previous frame, and N is the number of partials.

II. Reconstruction Based on Sinusoidal Modeling

Early applications of SM to reconstruction of missing audio data employed sinusoidal tracking in a close neighborhood of the gap for simple prediction of the evolution of spectral peaks. For example, a group of methods proposed in [9] used a straightforward extrapolation of sinusoidal trajectories (fig. 3) followed by linking to the trajectories beginning after the gap that are closest to the predicted endpoints, and interpolation of the linked segment.

These methods allowed successfully combining the trajectories and mitigating missing audio content as long as 30-50ms. However, for longer gaps the evolution of sinusoidal partial frequencies and amplitudes often cause significant mismatch between trajectories, so it is very difficult to establish a reliable link between both sides. Also, the simple model of partial variations does not provide a realistic interpolation of parameters; hence the resulting evolutions are often unnaturally smooth. It is partially due to the prediction algorithm that was not able to adapt to modulations present in music in a form of vibrato and tremolo. Also, the tracking algorithm of [7] did not adapt to the musical context.

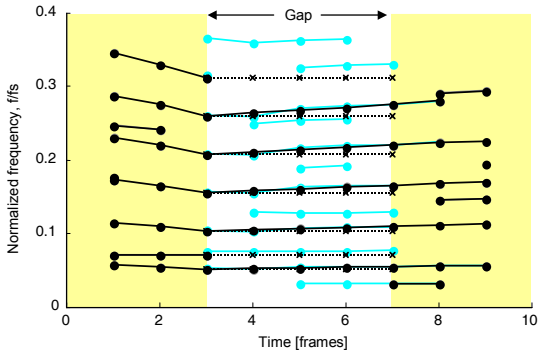

Fig. 3. Application of simple linking algorithm for sinusoidal trajectories of data surrounding a gap. Sinusoidal trajectories shown in blue correspond to actual data in original signal. Dotted lines represent extrapolation results. Solid black trajectories in the white area represent the interpolated data.

An important progress in tracking adaptability was achieved by the application of Hybrid Markov Model and Viterbi algorithm [10] as well as linear prediction (LP) [11]. The latter modeled the evolution of partial frequencies and amplitudes as auto-regressive (AR) processes. It is of a great advantage considering that pitch and intensity variations in many natural sounds are generated by the motion of player's hand that in turn is governed by simple kinetics. An LP-based tracking algorithm is capable of learning the character of typical vibrato and tremolo from the beginning of the note and accurately predicts its further evolution. This possibility has been exploited in [12] for interpolation of long gaps in audio.

The offline technique proposed in [12] applied Burg method of linear prediction to track sinusoidal partials in the data preceding and following the gap. The evolutions of partial frequencies and amplitudes within gap were predicted (extrapolated in two directions) using the known trajectories ending at both gap boundaries. The final signal was obtained by taking a weighted average of both prediction outputs and resynthesizing the signal from interpolated trajectory parameters. According to the authors, this method allowed for successful mitigation of gaps in music recordings as long as 320-820ms, however the quality strongly depended on the character of music, its degree of tonality and depth of modulations.

In our experience this approach still exhibits several drawbacks. First of all, the LP-based tracking has a tendency to connect spectral peaks of different partials into one trajectory, thus producing excessive modulations of the synthesized content due to the prediction problem being ill

posed. For example, in a piano music with high polyphony, there are a high number of spectral peaks, and some may be missed during sinusoidal analysis. Other peaks are sometimes linked into a wrong trajectory that misleads the frequency predictor in the gap segment to produce a sound with vibrato, while there must not be a vibrato in piano sounds.

Secondly, only tonal components of the signal are synthesized and in case of a more percussive music the result is quite annoying, because it consists of sinusoids instead of noise. In the following, we address both issues.

III. Context Adaptation for the SM-based Prediction and Reconstruction

The simple idea of context adaptation is to observe a longer period of signal surrounding the gap in order to learn the character of music that may contain mixed instruments of constant and variable pitch. In our approach this adaptation is realized by a tracking algorithm that exploits various matching criteria for connecting spectral peaks into sinusoidal trajectories. The algorithm first attempts to connect as many spectral peaks exhibiting small frequency deviation within a long sequence of STFT frames as possible. The objective is to maximize the likelihood function,

$$L(f_n^k, f_m^{k+1}, A_n^k, A_m^{k+1}) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(f_n^k - f_m^{k+1})^2}{2\sigma_f^2} - \frac{(A_n^k - A_m^{k+1})^2}{2\sigma_A^2}}, \qquad (2)$$

where $f_n^k, f_m^{k+1}, A_n^k, A_m^{k+1}$ denote respectively the frequencies (in ERB scale) and amplitudes (in dB) of spectral peaks in the current and the next frame, and σ_f , σ_A are model parameters that are learned for a period of several seconds in the neighborhood of the gap. A maximum absolute frequency difference is additionally constrained at this tracking stage.

In a subsequent step, the remaining unlinked spectral peaks are connected using the Burg predictor method similar to [11] in order to form trajectories representing partials of varying frequency. The algorithm also attempts to combine peak sequences evolving in frequency in a similar way into harmonic groups. Tracking of such groups is much more reliable. Finally, the trajectories are partitioned into two sets (stable and variable), according to what rule has been used for linking them.

For predicting the evolution of sinusoidal trajectories in the gap region, different strategies are employed. Trajectories from the stable set are not predicted by the LP method, but rather a constant extrapolator is used that takes a mean frequency of the spectral peaks within each trajectory. In this way we avoid the risk of introducing a false pitch modulation to sounds that should exhibit constant pitch. Matching of these is simply based on the frequency difference,

$$d_f(\vec{f}_{k_1}, \overleftarrow{f}_{k_2}) = \left| \vec{f}_{k_1} - \overleftarrow{f}_{k_2} \right| < \varepsilon_f \qquad (3)$$

where the arrow-marked terms represent predicted parameter values, and k_1 to k_2 denote the frame numbers corresponding to the gap.

For variable partials, left and right-hand predicted trajectories are matched across the gap using normalized L_1 distance criteria,

$$d_f(\underrightarrow{f}, \underleftarrow{f}) = \frac{1}{1+k_2-k_1} \sum_{k=k_1}^{k_2} \left| \vec{f}_k - \overleftarrow{f}_k \right| < \varepsilon_f \qquad (4)$$

$$d_A(\underrightarrow{A}, \underleftarrow{A}) = \frac{1}{1+k_2-k_1} \sum_{k=k_1}^{k_2} \left| \vec{A}_k - \overleftarrow{A}_k \right| < \varepsilon_A, \qquad (5)$$

The optimal thresholds of ε_f and ε_A are determined empirically. The matched predicted ends are combined by using a weighted average (fig. 4). Unmatched trajectories are not removed from the model, but rather smoothly faded to zero. The length of this fading is determined by the length of the trajectory. For trajectories shorter than the length of the gap, the extrapolated segment is also appropriately shortened.

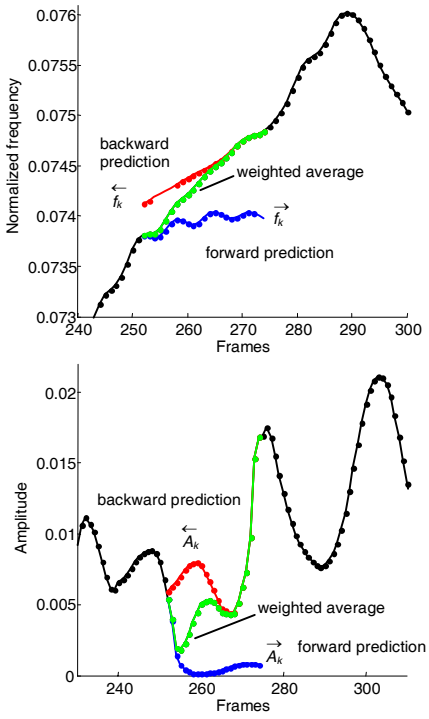

Fig. 4. An example of combining matched trajectories of frequency and amplitude for a single partial of the variable class

After successful matching, a tonal part of the signal is synthesized according to (1). Care is taken of exact matching of the phase of partials to the known phase at frame boundaries [7] which allows for seamlessly combining the synthetic signal with the available original.

IV. Noise Interpolation

The important disadvantage of SM is that it is not able to realistically synthesize the background noise, especially in the context of spectrally interpolated signals. To cope with this problem we perform an additional modeling step using LPC. For this purpose, two additional short segments of signal are synthesized from sinusoidal trajectories using (1) to match known signal segments preceding and following the gap. These synthetic signals are subtracted from the original signal, so that two segments of residual are obtained that contain mostly the background noise.

Two LPC models of moderate order (typically 20) are calculated from this data, using a frequency-warped LPC technique of [13]. The reason of using warped LPC is to obtain a perceptually uniform resolution in the low frequency range. Predictor coefficients of both are subsequently converted into Line spectral frequencies (LSF) representation [14] that allows for convenient interpolating between two models. A sequence of linearly interpolated LSFs is calculated for missing frames. At each frame, the

interpolated LSF values are converted back to the predictor coefficients form. These are used in an LPC synthesis filter configuration that shapes a segment of white noise. In this way, a sequence of noise segments is obtained such that its power spectral density morphs from that of the beginning of gap towards the noise after the gap. All these segments are combined using a standard OLA procedure. At the end, the interpolated noise is added to the synthetic tonal signal and replaces the missing samples within the gap.

V. Experimental results

We have performed a number of tests in order to verify the performance of the new technique with respect to an implementation of [12] and a simple PV-based reconstruction. Gaps of lengths varying from 100ms to 1s have been cut randomly from a selection of music recordings exhibiting various challenges and processed with all three algorithms. The reconstructed excerpts have been both inspected visually (using spectrograms – cf fig. 5,6) as well as subjectively evaluated [16] in a double blind critical listening environment according to the MUSHRA methodology [15]. The general conclusion is that the proposed enhanced technique always outperforms that of [12], however the audibility of the difference varies with the character of music. Certainly, the most substantial advantage in subjective quality appears in the case of percussive as well as densely polyphonic piano music (fig. 6).

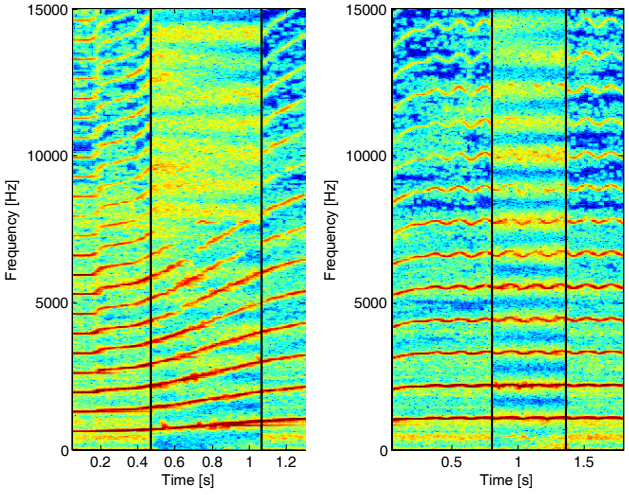

Fig. 5. Two reconstruction examples for a violin note with glissando and vibrato. Straight black lines represent gaps boundaries. Missing high frequency tonal partials are compensated by noise (cf fig. 2), a natural shape of vibrato modulation is well recreated.

VI. Conclusions

An enhanced mitigation technique for long gaps (up to 0.5s) in music signals has been presented. We show by experiments that an application of a simple musical context adaptation within a sinusoidal model together with separate modeling of the residual noise using an LSF-domain interpolated filter brings a substantial improvement in terms of seamlessly concealed gap. The applications of the proposed technique are in music data streaming and compression.

Acknowledgmnents

This work was supported by the public founds as a research project in years 2008-2010.

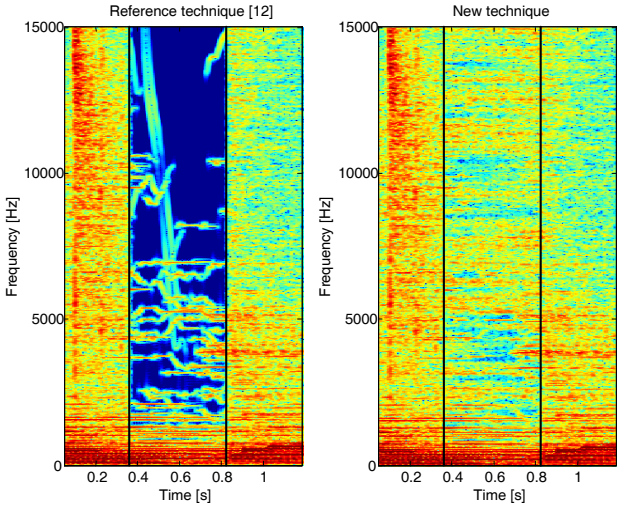

Fig. 6. A comparison of the reconstruction results of polyphonic jazz music with piano and percussion. Partials of piano sound are erroneously tracked with no context adaptation, and important part of noise energy is missing in the reference technique [12] (left). The new proposed technique (right) offers much more seamless mitigation of the missing contents.

References

[1] R. Veldhuis, *Restoration of Lost Samples in Digital Signals*, Englewood Cliffs, NJ: Prentice-Hall, 1990.

[2] S. Godsill, P. Rayner, O. Cappé, "Digital Audio Restoration" in *Applications of Digital Signal Processing to Audio and Acoustics*, (Kahrs, M., Brandenburg, K., Eds.), Kluwer Academic Publishers, Massachusetts, 2001.

[3] I. Kauppinen, J. Kauppinen, "Reconstruction Method for Missing or Damaged Long Portions in Audio Signal", *J. Audio Eng. Soc.*, vol. 50, pp. 594–602, July/Aug. 2002.

[4] P. A. Esquef, V. Välimäki, K. Roth, I. Kauppinen, "Interpolation of long gaps in audio signals using the warped Burg's method", *6th Int. Conf. Digital Audio Effects (DAFx-03)*, London, UK, 2003.

[5] D.W. Griffin, J.S. Lim, "Signal estimation from modified short-time fourier transform", *IEEE Trans. Acoust., Speech, Sig. Proc.*, ASSP-32(2):236-243, Apr 1984.

[6] M. R. Portnoff, "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform", *IEEE Trans. Acous., Speech, Sig. Proc.*, 24(3), June 1976.

[7] R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation", *IEEE Trans. Acous., Speech, Sig. Proc.*, 34(4), Aug. 1986

[8] J.O. Smith, X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation", *Int. Computer Music Conf.*, ICMC, University of Illinois, USA, 1987.

[9] R. C. Maher, "A method for extrapolation of missing digital audio data", *J. Audio Eng. Soc.*, vol. 42, no. 5, pp. 350–357, May 1994.

[10] P. Depalle, G. Garcia, X. Rodet, "Tracking of Partials for Additive Sound Synthesis using Hidden Markov Model", *IEEE Int. Conf. on Acoust., Speech and Sig. Proc. (ICASSP)*, vol. 1, April 1993, pp. 225–228.

[11] M. Lagrange, S. Marchand, M. Raspaud, J. B. Rault, "Enhanced Partial Tracking Using Linear Prediction," *Proc. Digital Audio Effects (DAFx) Conf.*, London, Sept. 2003, pp. 141–146.

[12] M. Lagrange, S. Marchand, J.-B. Rault, "Long Interpolation of Audio Signals Using Linear Prediction in Sinusoidal Modeling", *J. Audio Eng. Soc.*, 53(10): 891-905, 2005.

[13] A. Härmä, et al, "Frequency-Warped Signal Processing for Audio Applications", *J. Audio Eng. Soc.*, 48(11):1011-1031, 2000.

[14] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals", *J. Acoust. Soc. Am.*, 1(35), 1975.

[15] ITU-R, "Method for the subjective assessment of intermediate quality level of coding systems", ITU-R, Tech. Rep. BS. 1534-1, Rec. 2003.

[16] B. Latanowicz, *Reconstruction of damaged segments in music recordings*, MSc Thesis (in Polish), Poznan University of Technology, 2009.