MODULATION DECOMPOSITIONS FOR THE INTERPOLATION OF LONG GAPS IN ACOUSTIC SIGNALS

Pascal Clark and Les Atlas

Department of Electrical Engineering, University of Washington,
Seattle WA, 98195-2500, USA

ABSTRACT

This paper presents a modulation-based reconstruction method for audio signals across long gaps of missing samples. We use LTI filterbanks followed by a multiplicative model that decomposes subbands into constituent modulators and carriers. This processing separates slowly-varying envelopes, or modulators, from the high-frequency fine structure of the carriers. Since modulators can be downsampled, this decomposition allows faster gap reconstruction when applying standard interpolation algorithms on the downsampled modulators, particularly when interpolation requires matrix inversion.

Index Terms— modulation, signal reconstruction, acoustic signal processing, music, speech processing

1. INTRODUCTION

Gap interpolation is the process of reconstructing a contiguous segment of missing information in a digital signal. Such a need arises when, for example, an audio signal contains impulsive noise, or when a transmitted signal loses packets due to channel effects. Long gaps are difficult to repair in nonstationary audio signals, especially when the gap length is on the order of, or greater than, the interval of stationarity of the signal. The goal of this paper is to study the effects of using narrowband modulation decompositions as a preprocessor stage for the interpolation of gaps ranging from 10 ms to 500 ms.

Previous work in gap interpolation has included optimal least-squares solutions based on autoregressive modeling [1, 2]. These methods can give satisfactory results for small gaps less than 20 ms long. Recent attempts to extend the range of effective interpolation have used subband decompositions, in conjunction with non-optimal AR prediction [3] and with neural network nonlinear prediction [4].

Subband interpolation raises the possibility of using other classes of signal decomposition, such as demodulation. Many nonstationary signals, natural or man-made, can be represented as sums of slowly-varying narrowband processes that modulate high-frequency carriers. Modulation analysis has been used in source separation of music [5] and in speech perception studies [6, 7]. Due to their slowly varying nature, modulators are hypothetically easier to interpolate over long gaps.

The rest of this paper is summarized as follows. Section 2 discusses a modulation signal model in terms of a uniform filterbank. Section 3 then describes the process of demodulating a signal containing a gap and individually repairing each modulator and carrier. Section 4 presents experimental results, followed by a discussion and concluding remarks in Sections 5 and 6.

2. MODULATION SIGNAL MODEL

For natural signals with time-varying statistics, we assume a signal model of the following form:

$$x(t) = \sum_{k=0}^{K-1} s_k(t) = \sum_{k=0}^{K-1} m_k(t) c_k(t), \tag{1}$$

which expresses $x(t)$ as a sum of products between carriers $c_k(t)$ and modulators $m_k(t)$. The only constraints on this model are that each individual product $s_k(t)$ be bandlimited, and that the carriers be unimodular. Carriers thus contain fine-structure information, whereas modulators form corresponding amplitude envelopes.

As with any product model, there is no unique decomposition that satisfies (1) for a given signal $x(t)$. Adopting the methodology used by Atlas and Janssen [5], we first use a filterbank to separate $x(t)$ into subbands $s_k(t)$. Using the short-time Fourier transform (STFT), the subbands of the sampled signal $x[n]$ are:

$$s_k[n] = \sum_{m=-\infty}^{\infty} x[n+m] w[-m] e^{-j2\pi km/K}, \tag{2}$$

where $w[n]$ is a lowpass window of length K.

After the STFT filterbank, a carrier detection algorithm determines, for each $s_k[n]$,

$$c_k[n] = e^{j\phi_k[n]}. \tag{3}$$

This research was partially supported by AFOSR Grant FA95500610191.

Consequently,

$$m_k[n] = s_k[n] \cdot c_k^*[n], \tag{4}$$

where * denotes complex conjugation. The purpose of demodulation, then, is to determine $\phi_k[n]$ for a subband $s_k[n]$. Three such methods are considered in this paper, each of which is discussed in the following sub-sections.

2.1. Constant Carrier Demodulation

The first demodulation method defines the carriers of an arbitrary, sampled signal $x[n]$ to be constant-frequency tones spaced evenly around the unit circle. That is, for $k = \{0, ..., K-1\}$,

$$c_k[n] = \exp\left\{\frac{j2\pi kn}{K}\right\}. \tag{5}$$

This demodulation shifts each STFT subband of $x[n]$ to the baseband, essentially removing the fast angular rotation of the phasors defined in (5). The drawback of this decomposition is that, although reducing each subband to slowly-varying envelope-like modulators, it relies on a set of carriers that have no physical connection to the original signal.

2.2. Hilbert Demodulation

The second form of demodulation makes use of the analytic subband, which is defined as the sum of a real subband and its Hilbert transform. By virtue of the complex exponential in (2), the STFT subbands $s_k[n]$ are approximately analytic versions of real bandpass signals, with the exception being the few subbands that span both negative and positive frequencies near zero and the Nyquist rate. Ignoring these exceptions, Hilbert demodulation is as follows:

$$m_k[n] = |s_k[n]|, \quad \phi_k[n] = \angle s_k[n], \tag{6}$$

where $m_k[n]$ is called the Hilbert envelope and $\phi_k[n]$ the Hilbert instantaneous phase. In implementation, the relatively few non-analytic subbands are either discarded completely or left unmodified before and after interpolation.

2.3. Coherent Demodulation

Loughlin and Tacer have observed that the Hilbert decomposition in (6) often yields physically meaningless quantities such as non-bandlimited carriers and negative instantaneous frequencies (IFs) [8]. They propose a more physically meaningful instantaneous frequency definition based on the time-varying spectral center-of-gravity (COG) of a signal, which forms the basis for coherent demodulation in this work.

Given the power spectral density estimate of a windowed segment from the kth subband centered at time t, the instantaneous frequency of that subband is defined as

$$\bar{\omega}_k(t) = \frac{\int_{B_1}^{B_2} \omega S_k(\omega, t) d\omega}{\int_{B_1}^{B_2} S_k(\omega, t) d\omega}, \tag{7}$$

where $[B_1, B_2]$ is the range of frequencies spanned by the mainlobe of the kth subband. The carrier phase is then found by integrating the IF, as in

$$\phi_k(t) = \int_{-\infty}^{t} \bar{\omega}_k(\tau) d\tau. \tag{8}$$

This carrier definition is similar to the monochromatic carrier derived by Papoulis [9] in the context of a bandpass wide-sense stationary process. In that case, the spectral COG is optimal in the sense of minimizing the variance of the derivative of the modulator, akin to a bandlimiting criterion.

3. INTERPOLATION

The three demodulation methods given in the previous section are now used as front-end decompositions for gap interpolation. Instead of applying an interpolation algorithm on samples of the fullband signal $x[n]$, the idea now is to reconstruct $m_k[n]$ and $c_k[n]$ over samples that correspond to the gap in $x[n]$.

Consider the signal $x[n]$ missing L samples in the contiguous interval $[n_1, n_2]$. Taking into account the effects of the subband transients at the gap edges, the filterbank extends the gap to $L + K - 1$ samples, corresponding to the interval $[n_1 - K/2, n_2 + K/2]$, where K is the number of subbands and also the length of the STFT subband impulse response. The modulators $m_k[n]$ and carriers $c_k[n]$ therefore also contain unknown samples in the interval $[n_1 - K/2, n_2 + K/2]$.

The modulators are then reconstructed via some conventional interpolation algorithm, using M samples on either side of the gap, yielding $\hat{m}_k[n]$. In the case of the Hilbert envelope, $\hat{m}_k[n]$ is full-wave rectified to comply with the definition of a non-negative magnitude value.

For constant-carrier modulation and coherent modulation, bandlimited carriers imply that the modulators are also bandlimited. Even the Hilbert modulators, although non-bandlimited, are heavily concentrated at low frequencies. It is therefore possible to downsample the modulators by a factor R depending on the bandwidth of the filterbank subbands. Consequently, the downsampled gap length is $(L+K-1)/R$, and M/R known samples are available on either side of the gap. Downsampling is desirable because of the reduced amount of computation needed for interpolation algorithms that involve matrix inversion.

Carrier interpolation is decidedly simpler than modulator interpolation. In the case of the constant-carriers method, the carriers are fixed and thus require no repair, so $\hat{c}_k[n] = c_k[n]$.

The Hilbert carriers, on the other hand, have non-bandlimited, and even discontinuous, instantaneous frequencies. Such behavior hinders elaborate interpolation efforts, so instead we fit a line to the first difference of the Hilbert phase $\phi_k[n]$, and then exponentiate the cumulative sum to form $\hat{c}_k[n]$. In the case of coherent demodulation, no explicit interpolation is performed; instead, the length of the carrier detection window is chosen to be longer than the gap, such that spectral COG estimates can be made within the gap based on the M samples on either side. In this study, we set the carrier detection window length to $2L$.

After interpolating the modulators and carriers, the final repaired signal $y[n]$ is synthesized by applying the inverse STFT algorithm on the remodulated subbands $\hat{s}_k[n] = \hat{m}_k[n] \cdot \hat{c}_k[n]$.

4. EXPERIMENT DESIGN AND RESULTS

The following experiment studies the results of modulation-based gap interpolation. Four decompositions are compared:

1. full-band, or no demodulation (FB),
2. coherent demodulation (CM),
3. Hilbert demodulation (HM), and
4. constant-carrier demodulation (CC).

For this study, we use the iterative least-squares autoregressive (AR) interpolator derived by Etter [2], which assumes an underlying linear prediction model with a single parameter (the AR order) and does not require additional assumptions or constraints (e.g., bandwidth constraints, specified excitation sequences). As described in Section 3, this interpolation algorithm is applied to each modulator as given by the demodulation methods listed above, except for FB, which applies the interpolator directly to the full-band signal.

In practice, the least-squares optimality of the chosen interpolator can be a drawback for long gaps. This is because the conventional interpolator minimizes the AR prediction error over the gap, which tends to reduce the overall signal power in the middle of long gaps [3]. Non-optimal algorithms have been devised to counter this problem, but a simple post-processing stage is used here instead. Specifically, the reconstructed samples in the output signal $y[n]$ are scaled by a tapered window that maintains continuity at the gap edges while boosting the amplitude in the middle of the gap. A window satisfying these constraints is of the form $h[n] = 1 + \alpha g[n]$, where $g[n]$ is a Tukey window and α is a gain factor.

A meaningful yet quantitative measure of interpolation quality is necessary for parameter selection and objective performance comparisons. We use Bark Spectral Distortion [10], which incorporates a simple auditory model to determine the Euclidean distance between the two signals in a perceptual space. Using this measure, a smaller BSD value corresponds to less perceived distortion, with zero indicating no distortion.

The only parameter needed for FB interpolation is the AR order used to model the segments of data on either side of the gap. Similar to the experiments conducted by Etter [2], we found that one AR order, equivalent to an 80 ms modeling segment, performed the best for most gap lengths in a 16 KHz classical music test signal. For very long gaps (150 to 250 ms), however, a longer modeling interval closer to 500 ms was needed.

Table 1. Interpolation parameters obtained from BSD measures on a classical music test signal.

t_L (ms)	10	20	40	80	150	250
K	64	64	128	128	512	512
t_M (ms)	80	80	80	80	500	500

CM, HM, and CC interpolation require filterbank settings in addition to a preset AR order. A Hamming window was used for the STFT filterbank, with a downsampling factor of $R = K/8$. Confining K to powers of two, the best K was found to scale with the gap size, with low BSD scores tending to cluster around smaller K. Similar results were found between the three decompositions, so a common sequence of K values was chosen for all methods. A summary of the best interpolation parameters for classical music recorded at 16 KHz is given in Table 1, where t_L is the time duration of the gap and $t_M = M/f_s$ is the time duration equivalent to the AR order.

The parameters in Table 1 were used to interpolate gaps in four test signals: classical music at 16 KHz (the same used when choosing the parameters), speech at 16 KHz, a flute and bass duet at 44.1 KHz, and a jazz ensemble at 44.1 KHz. For the 44.1 KHz signals, the number of subbands used was $3K$. In each test signal, 11 gaps of constant duration were distributed evenly throughout a 13-second interval. Then, the average BSD was calculated for each signal after interpolating all eleven gaps. This was repeated for different gap sizes ranging from 10 ms to 500 ms. To repair the 500 ms gaps, the values in Table 1 were extrapolated to K = 1024 and t_M = 500 ms. Quantitative results appear in Figure 1, which displays the improvement in BSD for each decomposition (i.e., the BSD improvement is the log BSD of the corrupted signal minus the log BSD of the repaired signal).

5. DISCUSSION

The most striking observation from Figure 1 is that the FB curves end prematurely at 250 ms for 16 KHz signals and at 80 ms for 44.1 KHz signals. Since the conventional interpolator in [2] requires the inversion of an $L \times L$ matrix, interpolation becomes infeasible, even in non-real-time and on a modern Pentium-class PC, for large gaps using FB interpolation. This is not a problem, however, when interpolating the downsampled modulators in either of the CM, HM, or CC methods.

For the same reason, run-time is greatly shortened using

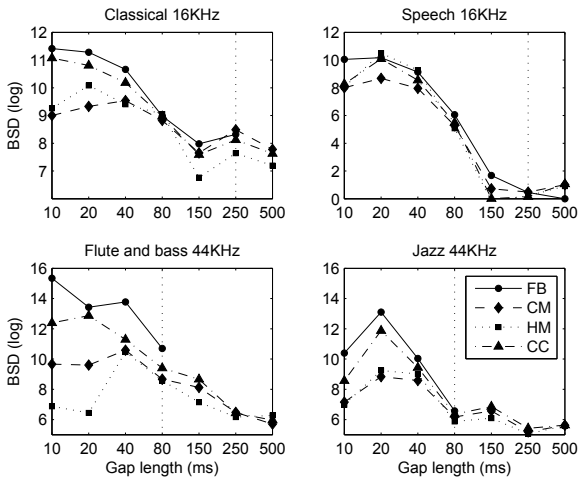

Fig. 1. Improvement in average BSD relative to the corrupted signal, for music and speech in two sampling rates. The vertical dotted lines indicate the maximum gap size for which FB interpolation is still feasible.

CM, HM, or CC interpolation. Although the filterbank creates a multitude of K interpolations, each interpolation deals with a gap that is reduced in length by a factor of $K/8$. The complexity of repairing an L-length gap is $O(L^3)$, whereas repairing K modulators requires on the order of $8^3L^3/K^2$ operations. Given the values of K in Table 1, the latter represents a significant decrease in arithmetic operations compared to FB interpolation.

According to the BSD measure, FB interpolation consistently performs the best for small gaps between 10 and 40 ms. This is partly explained by the disproportionate increase in gap size due to the filterbank stage, which extends the gap length to $L + K - 1$ samples. For larger gaps, however, CM and CC interpolation yield almost equivalent signal quality compared to FB interpolation. HM interpolation suffers noticeably in the classical and flute/bass test signals. A possible cause for this disparity is the fact that the Hilbert modulators are magnitude signals with sharp inflections at the zero-crossings, and are therefore poorly modeled as stationary AR processes.

It is interesting to note that the speech test signal exhibits far less signal improvement for large gaps when compared to the music signals. This suggests two possible explanations: that the chosen interpolation parameters are incompatible with speech signals, or that speech is inherently less stationary and thus more difficult to interpolate over long gaps.

6. CONCLUSION

Using a least-squares AR-based interpolator, a modulation decomposition is clearly beneficial in terms of computational complexity, especially for long gaps. In terms of quality, a perceptually-motivated distortion measure shows that standard full-band interpolation performs the best for small gaps. However, interpolations of the full-band signal, constant-carrier modulators, and coherent modulators give similar results for gaps longer than about 80 ms. A remaining question, to be investigated via listener testing, is whether BSD accurately reflects changes in modulation as perceived by human listeners. Also, we have reason to suspect that the filterbank structure described in Section 2 imposes a possibly substantial bias on carrier detection, which could result in errors in demodulation and in subsequent gap interpolation. This matter is currently the topic of ongoing research.

7. ACKNOWLEDGEMENTS

The authors would like to acknowledge helpful discussions with Prof. Bishnu Atal, Kaibao Nie and the members of ISDL of the University of Washington.

8. REFERENCES

[1] A.J.E.M. Janssen, N.J. Veldhuis, and L.B. Vries, "Adaptive interpolation for discrete time signals that can be modeled as autoregressive processes," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 2, pp. 317–330, April 1990.

[2] W. Etter, "Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters," *IEEE Trans. Signal Processing*, vol. 44, no. 5, pp. 1124–1135, May 1996.

[3] P.A.A. Esquef and L.W.P. Biscainho, "An efficient model-based multirate method for reconstruction of audio signals across long gaps," *IEEE Trans. Speech and Audio Processing*, vol. 14, no. 4, July 2006.

[4] G. Cocchi and A. Uncini, "Subband neural networks prediction for on-line audio signal recovery," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 867–876, July 2002.

[5] L. Atlas and C. Janssen, "Coherent modulation spectral filtering for single-channel music source separation," *Proc. IEEE ICASSP*, vol. 4, pp. 461–464, 2005.

[6] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, February 1994.

[7] Z.M. Smith, B. Delgutte, and A.J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Letters to Nature*, , no. 416, pp. 87–90, March 2002.

[8] P.J. Loughlin and B. Tacer, "On the amplitude- and frequency-modulation decomposition of signals," *Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1594–1601, September 1996.

[9] A. Papoulis, "Random modulation: A review," *IEEE Trans Acoustics, Speech, and Signal Processing*, vol. ASSP-31, no. 1, February 1983.

[10] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal Selected Areas Communications*, vol. 10, no. 5, pp. 819–829, June 1992.