SINUSOIDAL MODELING PARAMETER ESTIMATION VIA A DYNAMIC CHANNEL VOCODER MODEL

Aaron S. Master[*]

Center for Computer Research in Music and Acoustics,
Stanford University,
Stanford, CA 94305-8180 USA

ABSTRACT

We present a new analysis methodology for extracting accurate sinusoidal parameters from audio signals. The method combines modified vocoder parameter estimation with currently used peak detection algorithms in sinusoidal modeling. The current system processes input frame by frame, searching for peaks like a sinusoidal analysis model, but also dynamically selects vocoder channels, through which smeared peaks in the FFT domain are processed. This way, frequency trajectories of sinusoids of changing frequency within a frame may be accurately parametrized. We note that the computational expense incurred by using the new model is offset by the reduced frame rate allowed, and describe possible applications for the new model. We demonstrate that the current model is able to follow the changing frequencies in a long analysis frame, with accuracy greater than that found in a conventional sinusoidal model.

1. INTRODUCTION

In the field of audio signal processing, the ability to extract accurate sinusoidal parameters is of fundamental importance. Audio coding, speech enhancement, sound source separation, and pitch tracking applications all rely to some extent on the quality of sinusoidal parameters. In many cases, it is also highly desirable to know the general direction and specific path of travel of frequency components within a signal. These components are most correctly called nonstationary quasi-sinusoidal signals, though we will refer to them herein as *nonstationary sinusoids* for convenience.

In the past, two methods of data-reduced signal representations have been commonly used in audio analysis to extract sinusoidal parameters, though neither has completely addressed the issue of the direction and path of travel of sinusoidal components. The two techniques are spectral modeling [1, 2, 3] and vocoder modeling [4]. In spectral modeling, the signal is segmented in time into frames. In vocoder parameter modeling, the signal is assumed to be well-modeled as a set of responses of bandpass filters. In each case, the segmentation of the model limits parameters from being sufficiently dynamic in both time and frequency.

Though these basic models do not completely address the nonstationarity issue, some research has done so within certain limits. In [5], for example, the authors use an entire spectral peak – containing some information about the nonstationarity of the sinusoid – when performing pitch shifting and a variety of other modifications to the input signal. In [6], the problem of concatenating nonstationary sinusoids in synthesis is addressed, though the approach relies on a linear *chirp* approximation to the frequency trajectories. In [7], the possibility of linear-only variation in the amplitude of identified sinusoids is addressed in an efficient synthesis context. This work does not, however, allow nonlinear variation in amplitude or consider instantaneous frequency of nonstationary sinusoidal components. In [8], it is shown that the history of a sinusoid's parameters may be used to optimally inform a peak tracking procedure. While this does bear relevance to accurately tracking the instantaneous parameters of sinusoids, it is limited by the quality of the parameters that are originally detected at the frame level.

Another approach to the nonstationary sinusoid issue has been to set up the system to minimize the effect of nonstationarity. In [1, 2], for example, high frame rates are used to ensure quasistationarity. In [9, 3], any spread out peak portions are modeled, but as bandlimited noise. We hypothesize that doing this does not sufficiently capture the nature of nonstationary sinusoids.

Given the importance in identifying and synthesizing nonstationary sinusoids and a lack of systems that do so, we present a system that fills the void. Our system hybridizes the two basic models mentioned above – spectral and vocoder modeling – but does so in a way that allows us to relax their respective assumptions, namely that sinusoids must remain constant in frequency within a frame and that sinusoids vary only within the limits of a fixed channel bandwidth. Our system demonstrates the ability to identify and follow sinusoidal frequency and amplitude trajectories within a frame, and to match these trajectories between frames.

2. CURRENT MODEL

To obtain the time varying sinusoidal parameters, the current system performs five major operations: time windowing, spectral parsing, peak isolation / Hilbert transform / IFFT, vocoder parameter estimation, and frequency trajectory matching. The overall system acts much as a sinusoidal modeling system, but with a vocoder technique applied to spectral peaks that are too wide to be accurately modeled

[*] Supported by the National Science Foundation.

as quasistationary sinusoids.

2.1. Time Windowing

The system begins by segmenting the signal in time into Hamming-windowed frames, much as in spectral modeling. It uses longer frames than traditionally used with spectral modeling, however, since we will be able to achieve time varying parameters within a frame, something impossible in conventional sinusoidal modeling. The lower frame rate also offsets the computational expense introduced by the vocoder technique described below. To ensure constant overlap and add for the Hamming window, 50% window overlap is used.

2.2. Spectral Parsing

Give the time-windowed and zero-padded signal, an FFT is applied to generate a spectrum for a given frame. We then apply a technique called *spectral parsing* to determine the location of nonstationary sinusoids, and quasistationary sinusoids within the spectrum.

To perform this parse, we identify peaks and valleys in the magnitude FFT. This includes a test to ensure that not every spectral ripple will be identified as a smeared peak. Specifically, a valley will not be considered a border between peaks unless its height is less than

$$.45 * (\min(P_1, P_2)),$$

where P_1 and P_2 are the heights of the peaks to the left and right of the valley at hand. Also, a test is applied to ensure that the sidelobes of a sinusoid's peak are not treated as separate peaks. Once the spectrum has been parsed this way, we note the width of each peak i at the higher of its two bordering valleys as D_i. We note the corresponding height in terms of dB down from the peak to the higher valley as H_i.

We now compare D_i to the width of a reference stationary sinusoid's spectrum at H_i. Since all stationary sinusoids will have the same width for a given windowing function and number of points in the FFT, we need only consider the window transform, defined as the FFT of the N-point windowing function $w(n)$:

$$W(\Omega) = \sum_{n=0}^{N-1} w(n) e^{2\pi i n \Omega / N}.$$

We notate the height of $W(\Omega)$ at H_i as D_i^r.

This comparison of D_i^r and D_i will determine the next step taken by the system. If $D_i < .7 * D_i^r$ a spurious peak has been detected [10]; this will later be modeled as noise or excluded from the model. If $.7 * D_i^r < D_i < 1.05 * D_i^r$ then the sinusoid will be considered quasistationary. In this case, conventional sinusoidal modeling as mentioned above may be used. Thus the system will use minimal change criteria for trajectory matching and synthesis. If $D_i > 1.05 * D_i^r$, then the nonstationary model described below is applied.

2.3. Peak Isolation, Hilbert Transform, and IFFT

Assuming the system has decided to model a sinusoid as nonstationary using the criteria above, it is necessary to prepare each corresponding peak separately for the vocoder model. As described in [4], we may use the complex analytic representation of a time domain signal as input to the phase vocoder. To obtain this signal, the peak isolation technique, Hilbert transform, and IFFT must be performed.

To do this, we begin by setting the spectrum to zero outside the peak of interest. (We retain both the positive and negative frequency versions of the peak.) We then take the Hilbert transform of this spectrum, and then IFFT both the original and Hilbert transformed spectra to obtain two time domain signals, which are 90 degrees out of phase with each other. We call these signals $x_k^r(n)$ and $x_k^i(n)$, and we will use them to form the analytic signal used in vocoder analysis.

2.4. Vocoder Model Application

We may now consider a vocoder channel model with a bandwidth corresponding to that of the dynamically selected peak area, and a center frequency corresponding to the peak of the channel. Thus, the traditional vocoder model of a k^{th} channel sinusoid is used:

$$x_k(t) = a_k(t) cos[\omega_k t + \phi_k(t)]$$

where ω_k is the fixed channel center frequency.

To obtain $a_k(t)$ and $\phi_k(t)$, we use the complex analytic representation $x_k^a(t)$ of $x_k(t)$ obtained in the previous step. Thus we have:

$$x_k^a(n) = x_k^r(n) + i x_k^i(n)$$

We estimate the instantaneous amplitude by:

$$a_k(t) = \|x_k^a(n)\|$$

and the instantaneous frequency deviation by:

$$\phi_k(t) = \frac{x_k^r(n)\dot{x}_k^i(n) + \dot{x}_k^r(n) x_k^i(n)}{(x_k^r(n))^2 + (x_k^i(n))^2}$$

where the dot represents a derivative. See [4] for a complete derivation.

These parameters may also be estimated using other time domain techniques, which may be more or less error-prone, depending on the center frequency, the amount of frequency variation and interference from other sinusoids. This may be explored in a future work, and has already been dealt with for select cases [11].

2.5. Matching of Frequency Trajectories

In the current paradigm, the matching of frequency trajectories in consecutive frames may be performed with a modification of a technique used in conventional spectral modeling, frequency proximity matching. This technique uses a simple recursive approach given in [1] whereby a frequency estimate in a given frame is matched to the closest frequency in neighboring frames. In the current system, we must make a slight modification, because we now have an

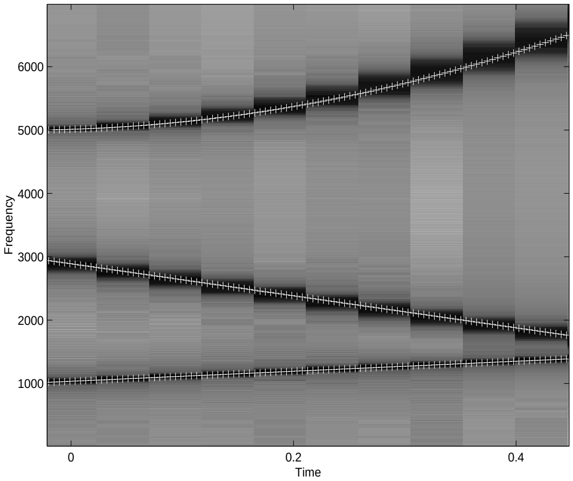

Fig. 1. This spectrogram shows the accuracy of parameter estimation achievable with conventional sinusoidal modeling. The frequency tracks overlaid in white "+" symbols were obtained with the new model, using the same frame size as the spectrogram.

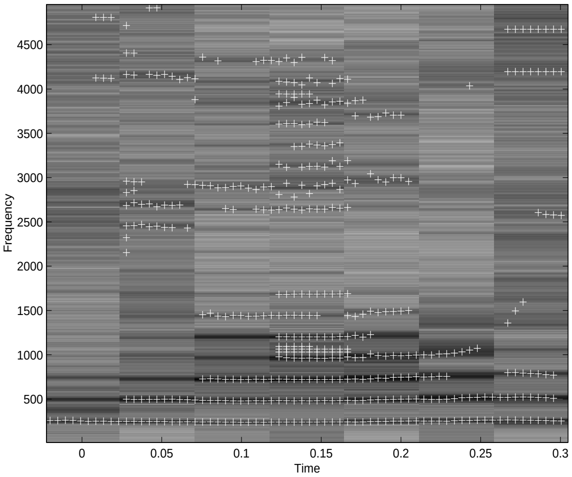

Fig. 2. This spectrogram of the female speech "you always" shows the accuracy of parameter estimation achievable with conventional sinusoidal modeling. The frequency tracks overlaid in white "+" symbols were obtained with the new model, using the same frame size as the spectrogram.

entire frequency trajectory in each frame rather than a single estimate. We thus attempt to match only the frequency estimates at the beginning and end of the frame. Because we use 50% overlapping frames in the current system, we use the frequency estimates taken at 25% and 75% of the way into each frame as the "beginning" and "end" estimates. Because the frequency trajectory introduces more accurate information about instantaneous frequency than the single frequency estimate used in spectral modeling, we must also demand that matching of frequency trajectories occur only within a very small epsilon. In the current system, we choose to allow 5% variation, a liberal amount by this standard.

3. RESULTS

To compare the algorithm with conventional spectral modeling and show its unique strengths, we present two examples below. The first contains three nonstationary sinusoids and the second consists of speech. In the first example, we show one spectrogram, and in the second, two. The first spectrogram for each example is constructed using the same (longer than typical) frame size, frame rate, and windowing function as used in the current system. The current system's frequency trajectories are overlaid as a series of white "+" symbols. For the second example, the second spectrogram uses frame sizes that are one-sixth the size of that used in the current system, and shows the identical overlaid frequency trajectory in white "+" symbols.

A look at the spectrogram (figure 1) for the first example gives us an idea of the accuracy of parameters that may be obtained by conventional sinusoidal modeling at the current system's frame rate (47 ms) and frame size (94 ms). In this example, we have three sinusoids: two which vary linearly in frequency, from 3000 to 1700 Hz and 1000 to 1400 Hz, and one which varies quadratically, as $5000 + (40n/N)^2$ Hz (where n are the sample indices and N is the number of samples in the signal). Only ten frequency estimates for a given sinusoid can be made using conventional sinusoidal modeling, because there are only ten frames. We could estimate these frequency parameters visually from the spectrogram: frequency peaks occur at dark points.

Interestingly, we see that the current system was able to track the exact frequency curves of the input sinusoids, while using the same number of frames. Especially important is the correctly followed quadratic curve, something impossible with conventional sinusoidal modeling at this frame rate. (Sinusoidal modeling uses linear frequency interpolation or a linearly-biased cubic fit of frequency and phase [1, 2].) On a more basic level, we note that the current system has also correctly identified the direction of travel of each sinusoid in each frame, something that may be useful in applications where the acoustic origin of each peak is sought.

The second example contains the speech utterance "you always" from a female speaker [12]. Here, we present two spectrograms, the first of which shows us how little data conventional sinusoidal modeling has to offer at this frame rate. Nonetheless, the system is able to track several sinusoids within and between frames, again shown as a series of "+" symbols. The second spectrogram, using a more conventional frame rate (8ms) and frame size (16 ms), verifies the frequency trajectory estimates of the current system, again overlaid.

Preliminary testing of the model using polyphonic audio input reveals similar results to those obtained using sinusoidal modeling. Thus, estimation of parameters every 8 ms, whether done in a single frame of more than 80 ms or 10 frames of 8 ms frame rate, appear to be of similar quality for a small set of test cases. We note that both models

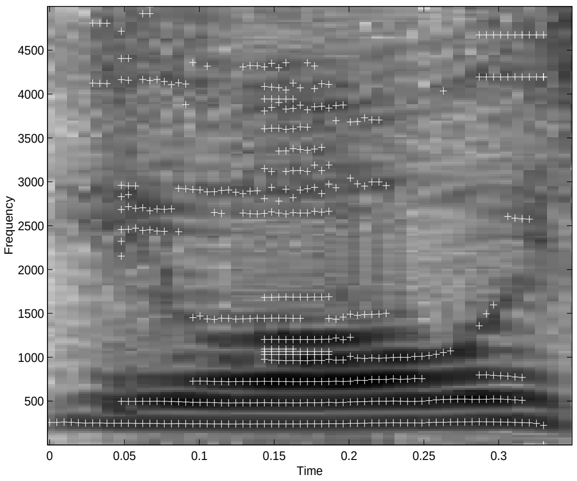

Fig. 3. This spectrogram of the female speech "you always" uses a frame size one sixth the size of that used in the current system. The current system's frequency tracks are overlaid in white "+" symbols.

seem to require that the multiresolution technique of [3] be used to guarantee audio of high perceptual quality over all frequencies.

4. FUTURE DIRECTIONS

The current system presents only one possible implementation of the concept of nonstationary sinusoidal modeling. Future implementations might include multiresolution analysis frames to allow differing degrees of nonstationarity versus frequency. Alternately, the time-frequency resolution tradeoff might be explored in more detail, with various frame lengths used for a variety of audio input. By doing this we could determine the ideal frame length for the nonstationary sinusoid model, and those cases in which a quasistationary spectral model was advantageous.

The frequency trajectory matching portion of the system will also be the subject of future investigation. Specifically, we will apply derivative-based trajectory matching, so that we may incorporate the direction of travel of the frequency trajectory when matching trajectories between frames.

The dynamic vocoder channel selection procedure too will receive future attention. We aim to select channels that yield the most efficient and accurate sinusoidal representation. We believe that using an iterative approach or information from neighboring frames could further our objective.

5. SUMMARY

An audio signal model has been presented that hybridizes spectral modeling and phase vocoder modeling. We noted that by using this model, we are able to increase the accuracy of frequency trajectories and reduce the frame rate in return for increasing computational expense at the frame level. We showed that the current model allows for more accurate representation of speech and nonstationary-sinusoid-based signals.

6. REFERENCES

[1] Julius O. Smith and Xavier Serra, "PARSHL: A program for the analysis/synthesis of inharmonic sounds based on a sinusoidal representation," in *Int Computer Music Conf.* 1987, Computer Music Association, Also available as Stanford Music Department Technical Report STAN–M–43.

[2] R. J. McAulay and T. F. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," Tech. Rep. 693, Lincoln Laboratory, MIT, 1985.

[3] Scott Nathan Levine, *Audio Representations for Data Compression and Compressed Domain Processing*, Ph.D. thesis, Electrical Engineering Dept., Stanford University (CCRMA), December 1998.

[4] Mark Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, Winter 1986.

[5] Jean Laroche and Mark Dolson, "New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing, and other exotic audio modifications," *Journal of the Audio Engineering Society*, vol. 47, no. 11, pp. 928–936, November 1999.

[6] Michael Goodwin and Alex Kogon, "Overlap-add synthesis of nonstationary sinusoids," in *International Computer Music Conference*, 1995.

[7] Jean Laroche, "Synthesis sinusoids via non-overlapping inverse fourier transform," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 471–477, July 2000.

[8] J. J. Wolcin, "Maximum a posteriori estimation of narrowband signal parameters," *Journal of the Acoustical Society of America*, vol. 68, no. 1, pp. 174–178, July 1980, Nusc TM no. 791115, June 21, 1979.

[9] X. Serra, *A System For Sound Analysis Transformation Synthesis Based on a Deterministic Plus Stochastic Decomposition*, Ph.D. thesis, Stanford University, 1989.

[10] Xavier Serra and Julius O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, Winter 1990.

[11] James A. Moorer, "The use of the phase vocoder in computer music applications," *Journal of the Audio Engineering Society*, vol. 26, no. 1/2, pp. 42–45, Jan./Feb. 1978.

[12] E. Wan, A. Nelson, and Rick Peterson, *Speech Enhancement Assessment Resource (SpEAR) Database*, Oregon Graduate Institute of Science and Technology CSLU, Beta Release v1.0, avaliable online at http://ee.ogi.edu/NSEL/.