A STOCHASTIC STATE-SPACE PHASE VOCODER FOR SYNTHESIS OF ROUGHNESS

Doug Van Nort, Philippe Depalle

Sound Processing and Control Laboratory
Music Technology Area
Schulich School of Music
McGill University, Montreal, Canada
{doug|depalle}@music.mcgill.ca

ABSTRACT

This paper presents an implementation of the phase vocoder within a Gaussian state-space framework. Rather than formulate the problem as a deterministic evolution of frequencies centered around a given bin, this evolution is treated stochastically by introducing noise into the dynamics matrix of the recursive state equation. This produces effects on the roughness of the input sound, which vary depending on the position within the matrix where the noise is added, how it is propagated throughout the matrix and further by the variance of the noise input.

1. INTRODUCTION

The phase vocoder is a widely used tool for the analysis, transformation and synthesis of audio signals. It began as a way to efficiently code and transmit voice signals using filterbanks [1], was later represented by the Short-Time Fourier Transform (STFT) [2] and then began to find use in musical applications [3],[4]. The most common effects generated by the use of the phase vocoder are pitch shifting and time scaling, which are achieved through altering the time/frequency block increment size between the analysis and synthesis step and then interpolating. If the step increment for both analysis and synthesis is subjected to certain constraints based on the type of windowing function used in the STFT, then the input signal is perfectly reconstructed upon re-synthesis. However the phase vocoder becomes musically interesting when the signal is distorted by transformations such as the aforementioned pitch/time scaling or others in which the amplitude and phase of each frequency bin are modified over time. As the representation itself is purely deterministic and able to capture the signal entirely, these distortions are externally applied to the spectral data in an intermediate (*i.e.* between analysis and synthesis) step. While this technique is very powerful and is the basis for most spectral processing used in computer music compositions, we have found that some interesting effects can be had by embedding a stochastic representation within the phase vocoder itself. This is achieved through the use of a state-space representation.

2. STOCHASTIC STATE-SPACE VOCODER

Rather than model a signal as an Autoregressive-Moving Average (ARMA) process, a stochastic process $x[n]$ that is governed by a linear dynamical system can be expressed by the state-space equations

$$s[n+1] = As[n] + w[n] \tag{1}$$

and

$$x[n] = Bs[n] + v[n] \tag{2}$$

where the sequence $s[n]$ is the state of process x at time n, and Equation (1) represents the internal dynamics of the process as governed by dynamics matrix A. Equation (2) transfers the state vector, which may be hidden, into a vector of observable output variables. Both w and v are Gaussian white-noise processes. The first affects the progression of the state while the second is additive noise present in the output process x. This latter use of noise in the modeling of an audio signal can be found in the Spectral Modeling Synthesis (SMS) approach [5]. Our interest here is in the effect of including noise in the state process equation of a representation based on the the phase vocoder. In particular, we include noise in the state *matrix* rather than simply in the state vector as is done in equation (1). Therefore, we may re-write the state equation as follows

$$s[n+1] = (A + W_n)s[n] \tag{3}$$

where W_n is a time-varying matrix of Gaussian random variables. This matrix will be described in more detail in section 2.2.

2.1. Related work

A state-space approach to analysis/synthesis was presented in [6] in which the real and imaginary components of p sinusoidal partials, tracked over time, were represented in the state vector. The observation matrix summed across the real components of the partials, and the addition of observation noise generated a sinusoid+noise re-synthesis. Thus, this model represents a hybrid state-space / sinusoidal model.

Similarly, a recursive state-space formulation is presented in [7] in which the state is comprised of the real and imaginary components for N evenly spaced frequency bins. Thus, this implementation maintains all of the data from the phase vocoder while the aforementioned work tracks only partials and is closer to a sinusoidal model. The motivation differs in this latter work as well, with the goal being the interpolation of missing audio samples whereas the former research was concerned with building an analysis/synthesis scheme for audio transformations. This current work is situated between these two in the sense that our motivation is towards musical transformations, yet we work on the lower-level representation of the phase vocoder's spectral frames. However our state-space representation differs from [7] in reflection of the differing motivations: the desire to track time-domain signals and interpolate missing values lead to a stochastic representation as in equation (1), in order to build uncertainty into the time-varying signal. We add noise to the dynamics matrix in order to perturb the structure of the *system itself* in order to explore the complex couplings that result.

Each of these approaches are based on a recursive description of the Discrete Fourier Transform. That is, the complex exponentials of the DFT can be expressed as

$$e^{jn\theta} = e^{j\theta} e^{j(n-1)\theta} \tag{4}$$

for time n and frequency θ. Thus the DFT matrix and its inverse can be expressed as a first-order recursion, a necessity in order to work within the state-space framework.

2.1.1. A Note on "Roughness"

The term "roughness" has taken on a specific meaning in the psychoacoustics literature [8]. Its introduction in this context can be dated back to Helmholtz in the late 19th century where it was linked to the subjective notion of dissonance. While this latter term has somewhat changed itself with musical periods, psychoacoustical roughness relates to dissonance in the classical sense of beating/modulating sounds such as those that result from certain pitch ratios (*e.g.* a minor second). In this sense it has been shown to change as a function of the depth and frequency of modulation of a sound in both amplitude and frequency. In this work, we do no use the word in this strict sense, and yet our results relate qualitatively to those sounds that would be considered psychoacoustically "rough". Certain results could as easily be labeled as "noisy" or "textural."

2.2. Current Implementation

Again, our state-space phase vocoder is built from a state vector comprised of the real and imaginary components of the audio signal. The nature and size of said state varies depending on whether the analysis or synthesis step is being performed. For the analysis step, given an input block of real signal $x = \{x_1, ..., x_N\}$, the state vector is thus

$$s = [x_1, 0, ..., x_N, 0]^T \tag{5}$$

which represents the initial state vector for the current block of N samples. The state is re-initialized with a new input block at each signal boundary (each N samples), and during the state recursion s is propagated by the dynamics matrix

$$A = \mathbf{DIAG}(R(\theta_0), ..., R(\theta_{N-1})) \tag{6}$$

where $\mathbf{DIAG}$ represents a block diagonal matrix and

$$R(\theta_k) = \begin{pmatrix} \cos(\frac{2\pi k}{N}) & \sin(\frac{2\pi k}{N}) \\ -\sin(\frac{2\pi k}{N}) & \cos(\frac{2\pi k}{N}) \end{pmatrix}. \tag{7}$$

The observation matrix

$$B = \begin{pmatrix} 1 & 0 & ... & 1 & 0 \\ 0 & 1 & ... & 0 & 1 \end{pmatrix} \tag{8}$$

produces an output vector[1]

$$\hat{s} = (s_{0,r}, s_{1,r}, s_{1,i}, ..., s_{\frac{N}{2}-1,r}, s_{\frac{N}{2}-1,i}, s_{\frac{N}{2},r}) \tag{9}$$

which is comprised of the real and imaginary components of the spectrum of input block x. We assume that x is real, and so only the first $\frac{N}{2}$ frequency bins are generated by the analysis state equations.

Now, the observed process $\hat{s}$ becomes the state vector for the synthesis step, where the synthesis dynamics matrix is defined by

$$\hat{A} = \mathbf{DIAG}(1, R^{-1}(\theta_1), ..., R^{-1}(\theta_{\frac{N}{2}-1}), 1). \tag{10}$$

The new observation matrix

$$\hat{B} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & ... & 1 & 0 & 1 \end{pmatrix} \tag{11}$$

produces output signal $\hat{x}$. In the absence of noise added to the state or observation equations for analysis and synthesis, this two-step recursion provides a perfect reconstruction. However, the addition of noise into the state equations at various points in the analysis/synthesis process and in various ways can introduce different roughness qualities into the input sound that can then be controlled.

2.2.1. Introduction of Process Noise

In the standard state-space formulation the noise that is added into the state and/or observation equation is a vector w_n whose dimension is the same as that of the state. In our implementation we introduce an $M \times M$ matrix of Gaussian noise, where M is the size of the state. In particular the matrix is decomposed as follows:

$$W_n = \alpha W_n^d + \beta W_n^r \tag{12}$$

where W_n^d is the block diagonal that corresponds to the non-zero values of the dynamics matrix A and W_n^r provides the Gaussian values for the remaining upper and lower triangular parts of the matrix. α and β are free parameters that allows one to tune the contribution of these two parts of the matrix. We do not include an observation noise vector as this simply adds white noise to the output. Rather, the above matrix is added to the state dynamics as in equation (3), which can produce interesting frequency effects.

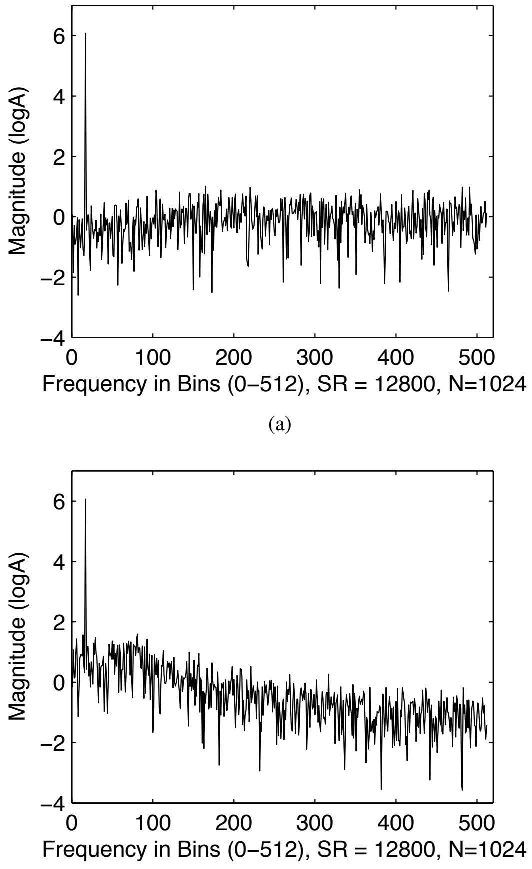
In particular, the matrix W_n^d is added to the sinusoidal components of the dynamic rotation matrix, causing an uncorrelated and random fluctuation of amplitude and/or phase in each element of the state. This effect can be very precise and localized in certain cases. For example, if the same noise values $\hat{N}$ are added to a sub-block along the diagonal of the synthesis-step state matrix $\hat{A}$ as follows

$$R^{-1}(\theta_k) + \hat{N} = \begin{pmatrix} \cos(\frac{2\pi k}{N}) + n_1 & -\sin(\frac{2\pi k}{N}) + n_2 \\ \sin(\frac{2\pi k}{N}) + n_1 & \cos(\frac{2\pi k}{N}) + n_2 \end{pmatrix} \tag{13}$$

then there will be a random modulation in the amplitude of partial k while the phase remains unchanged. The amplitude or phase of a given frequency can also be modified by converting the corresponding members of the state vector into polar form, acting on the appropriate values and then converting back to rectangular form before re-inserting them into the state equation. In this way one can *e.g.* introduce concurrent random modulations between partials that can affects a given sound's "texture" [9].

Now, the matrix W_n^r causes a random fluctuation which behaves quite differently. Random values added in this part of the matrix introduce a non-linear distortion, and noise can be added at specific matrix locations in order to introduce a coupling between

[1] Strictly speaking, the analysis step produces a 2 × N matrix. These values are then rearranged and the trivial imaginary values at $\theta_0, \theta_{\frac{N}{2}}$ discarded in order to form $\hat{s}$.



(a)

(b)

Figure 1: *Log FFT plot for stochastic state-space processed sinusoid. Gaussian noise added to analysis dynamics matrix in outer triangular portion (a) and to diagonal (b).*

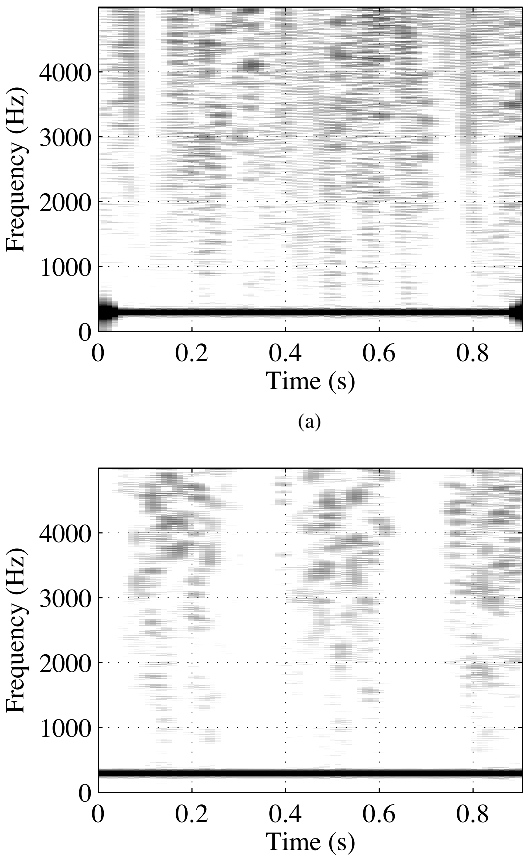
two frequencies. This can be non-physical — such as if the frequency at bin i is coupled to bin j but j is not coupled to i — or it can maintain a physical meaning if frequencies remain coupled in a bi-directional manner. We experimented with different process noise behaviors towards the end of creating musically interesting roughness effects.

3. RESULTS

Different roughness effects are observed depending on several factors: where in the dynamics matrix noise was introduced, if and how it was propagated in time through the matrix, and whether it was added during the analysis or synthesis step. Some results are summarized below.

3.1. Noise in Analysis Step

Given an input sinusoid at frequency f, and when noise was added to the outer triangular regions of the matrix only — that is, when[2] $W_k = \beta W_k^r$ — a nearly white noise component was added to the



(a)

(b)

Figure 2: *Spectrogram of sinusoid affected by noise propagated through analysis state matrix at one column alone (a) and coupled between a column/row pair (b). Sampling frequency = 11.025 kHz.*

entire signal, with a slight increase in energy at higher frequencies. In contrast, when noise was added to the matrix diagonal and $W_k = \alpha W_k^d$, a band of noise was introduced whose energy was concentrated around frequency f and fell off at higher frequencies. This difference is illustrated in Figure 1.

In order to create more interesting roughness qualities, we propagated noise through various parts of the matrix to create a time-varying effect. In particular, noise was propagated through a given column or row of the matrix. When noise was propagated down a single row or column, a beating noise with several small peaks was introduced. The rate of the beating can be controlled by the speed at which the Gaussian scalar noise value is sent through the given row/column. This modulating behavior can be seen in Figure 2b.

While this time-varying single perturbation produced a more musical result, it was not physically accurate: the noise value caused an interaction between the frequency located at the given column where the propagation occurred and each other frequency bin at the instant that the noise was swept past it in the matrix. However, this was not truly a coupling between frequencies as it did not occur in both directions. Thus, to make the effect more physical, we sent the same noise value down both column and row. Therefore, at time t if the input noise value was added to matrix

[2] We use index k here to underscore the fact that the state recursion in the analysis step is a function of frequency rather than time.

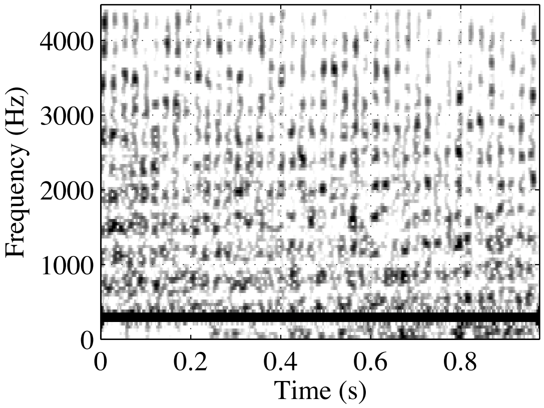


Figure 3: *Spectrogram of sinusoid at 300 Hz affected by noise propagated through synthesis state matrix. Sampling frequency = 8.5 kHz.*

value $A(i, j)$, it was further added to the value at $A(j, i)$. This coupled time-varying effect proved to create a much more sophisticated stochastic component to the sound — one that possessed more “texture" and resembled the sound of fire. For input sinusoid with frequency f, this effect was most prominent when it occurred at the column/row associated with the highest-energy frequency bin, namely $k = \frac{N*f}{F_s}$ where F_s is sampling frequency and N is the size of the input signal block. The difference between the “abstract" and physical roughness effects can be seen in Figure 2. Beyond having more high frequency content, the coupled example of Figure 2a possesses a spectral fine-structure that is present throughout the spectrum and which likely contributes to the overall textural quality.

3.2. Noise in Synthesis Step

It is important to remember that the state vector is not the same between analysis and synthesis steps. During analysis, the state is initialized with real and imaginary components of an input signal block of size N. At the synthesis stage, the state vector is initialized with the real and imaginary spectral values that are generated by the first $N/2$ iterations (assuming a real-valued input) of the recursive analysis process. Therefore, the addition of noise to the state matrix affects the dynamics of either the complex modulation or demodulation process associated with the DFT/iDFT and there is no reason to assume that the addition of the Gaussian noise vector would produce the same sonic result at each stage. Indeed, the addition of noise values in the synthesis state matrix — of coupled noise propagated down a column/row pair — produce a strong modulation effect not present in the previous examples. While the others produced fluctuations and a beating effect, this synthesis-step noise introduces a spectrotemporal modulation illustrated by the spectrogram of figure 3. In this representation one can see a quasi-periodic emergence of strong spectral peaks which modulate throughout the spectrum.

4. CONCLUSION AND FUTURE WORK

We have introduced a phase vocoder in which a stochastic element has been built into the representation via a state-space framework. Through the embedding of noise within the system representation itself (rather than as input to the state or observation equations) a sound can be re-synthesized with an added “roughness" quality. Preliminary results illustrate the potential of using this approach to generate interesting effects. With this simple linear systems framework, musically useful nonlinear distortions can be introduced and different effects can be created and controlled by altering the Gaussian noise matrix over time. One simple physically-inspired effect was suggested. We intend to explore more complex processes by exploiting the interaction between the analysis and synthesis state vectors, by introducing harmonic distortion via couplings between harmonically related frequency components and by consrtucting more elaborate time-based control curves for the noise parameters.

5. REFERENCES

[1] J. L. Flanagan and R. M. Golden, “Phase vocoder,” *Bell System Tech. J.*, vol. 45, pp. 1493–1509, Nov. 1966.

[2] J. B. Allen and L. R. Rabiner, “A unified approach to short-time Fourier analysis and synthesis,” *Proc. IEEE*, vol. 65, pp. 1558–1564, 1977.

[3] J. A. Moorer, “The use of the phase vocoder in computer music applications,” *J. Audio Eng. Soc.*, vol. 26, no. 1, pp. 42–45, 1978.

[4] M. Dolson, “The phase vocoder: A tutorial,” *Computer Music J.*, vol. 10, no. 4, pp. 14–27, 1986.

[5] X. Serra and J. O. Smith, “Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition,” *Computer Music J.*, vol. 14, no. 4, pp. 14–24, 1990.

[6] H. D. Thornburg and R. J. Leistikow, “Analysis and resynthesis of quasi-harmonic sounds: An iterative filterbank approach,” in *Proc. Int. Conf. on Digital Audio Effects (DAFx-03),* London, UK, 2003, pp. 129–134.

[7] A. T. Cemgil and S. J. Godsill, “Probabilistic phase vocoder and its application to interpolation of missing values in audio signals,” in *13th European Sig. Proc. Conf.*, Antalya, Turkey, 2005, [Online] http://www-sigproc.eng.cam.ac.uk/~atc27/papers/cemgil-godsill-em-restore-eusipco.pdf.

[8] P. Daniel and R. Weber, “Psychoacoustical roughness: Implementation of an optimized model,” *Acustica*, vol. 83, pp. 113–123, 1997.

[9] S. Dubnov, N. Tishby, and D. Cohen, “Influence of frequency modulating jitter on higher order moments of sound residual with applications to synthesis and classification,” in *Proc. Int. Comp. Music Conf. (ICMC’96),* Hong Kong, 1996, pp. 378–385.