EPS MODELS OF AM-FM VOCODER OUTPUT FOR NEW SOUNDS GENERATIONS

J. Timoney and T. Lysaght

Department of Computer Science
NUI Maynooth
Maynooth
Co. Kildare
Ireland
jtimoney@cs.may.ie , tom.lysaght@may.ie

ABSTRACT

The Phase Vocoder [1] was originally introduced as an alternative approach to speech coding but has won much greater acceptance among the music community as a tool both for sound analysis and composition [2]. Although dormant for some time, there has been a resurgence of interest in AM-FM speech signal descriptions in the last ten years [3], [4]. With the intention of building on some of the new ideas proffered, the aim of this work is to first consider their application to musical signals. It then demonstrates how paramaterisation of the extracted AM-FM information using EPS (Exponential Polynomial Signal) models allows modification of the individual large- and small-grain features of the AM and FM components, thus providing a new way for generating audio effects.

1. INTRODUCTION

The basic principle of the Phase Vocoder is to first split the incoming speech or music signal into a number of sub-bands using a fixed filterbank, followed by demodulation of each filter output from which their envelope (AM) and instantaneous frequency (FM) components are computed. Two drawbacks with this approach are: (1) Important signal components may lie outside the passband of the filterbank and be severely attenuated, and (2) if more than one frequency component lies within the passband of a filter, the instantaneous frequency of its output can often appear as an erratic function whose range may extend from negative to positive infinity. This behavior is contrary to what is intuitively understood about frequency and can render modeling or manipulation difficult to implement or understand. In an effort to address these problems, contemporary work in the area of speech vocoding [3] modeled the speech signal as the sum of a number of formants, which were extracted using a tracking scheme. The DESA (Discrete Energy Separation Algorithm) was used to estimate the AM and FM information associated with each one. The results presented appeared to show good performance, however, the tracking scheme depended heavily on decision-based logic, which is computationally intensive and is not always reliable. Additionally, judicious post-estimation smoothing of the AM and FM components using both median and binomial filters was required. More recent work in [4] proposed an alternative data-driven speech formant tracking scheme. A two-stage adaptive filterbank structure was used: a frequency-varying set of Dynamic Tracking Filters was used to estimate and track the dominant spectral components present in the signal, while a set of all-zero filters, operating in a cross-coupled fashion, was used to suppress interference from other strong neighbouring components while the estimation procedure was being carried out. To produce an instantaneous frequency estimate without the large fluctuations or the significant post-filtering requirements, decomposition of each filter output into minimum phase (MinP) and all-pass (AllP) components was performed. The MinP component has the property that all its zeros lie inside the unit-circle and is completely characterised by its envelope alone, while the AllP signal has a positive definite instantaneous frequency. An algorithm for this MinP-AllP decomposition was introduced in [4], named as Linear Prediction in the Spectral Domain (LPSD). The signal model assumed under which this decomposition was to be carried out was given by [4]

$$s(n) = \underbrace{A_c e^{\alpha(n)+\beta(n)+j\left(\hat{\alpha}(n)+\hat{\beta}(n)\right)}}_{\text{MinP}} \underbrace{e^{j\left(\omega_c n - 2\hat{\beta}(n)\right)}}_{\text{AllP}} \qquad (1)$$

where the "hat" stands for the Hilbert transform. A_c is a complex amplitude parameter, of the form $a_0 e^{j\phi}$. $\alpha(n)$ and $\beta(n)$ denote modulating quantities, and ω_c is the carrier frequency of the AllP component.

From (1), an AM-FM signal description in terms of its log-envelope and positive instantaneous frequency (PIF) could then be created. The log-envelope is equivalent to the absolute value of the real part of the MinP component normalised by the complex amplitude, and the positive instantaneous frequency is the differential of the angle of the AllP component. These quantities have properties of boundedness and smoothness that are unattainable using the previous methods. This means then that relatively simpler modeling procedures can be applied to these quantities.

2. AM-FM COMPONENT TRACKING IN MUSIC SIGNALS

Some form of tracking procedure is necessary to extract the most significant AM-FM components from the signal if the limitations of the fixed filterbank approach are to be avoided. The tracking method presented in [4] for determining the time-varying spectral locations of the formants is hierarchical in concept and operates on the assumption that the number of formants in the speech are known. The basic principles of the algorithm are as follows: the signal is first divided into short-time frames. The algorithm is initialised by finding a pre-determined number of prominent spectral peaks in the first frame. In the following frame, the all-zero filter bank is used to null the spectral regions corresponding to all of the peaks in the first frame bar one, after which the dynamic-tracking filter measures the frequency of the remaining peak. This process is then carried out for all the other peaks previous from the previous frame. With the new set of spectral peaks for the second frame, the centre frequencies of the all-zero filter bank are updated and the same procedure is repeatedly applied to all successive frames. Using empirical examples, this method was shown to perform well in tracking speech formants, providing a good level of accuracy [ICASSP paper]. However, direct application of this algorithm to musical signals was found not to be as successful. This could be explained by the fact that musical signals are better described by their harmonic content rather than their spectral resonances. A particular obstacle was that the frequency response of the all-zero filter was very poor for musical signals containing four harmonics or more. It was thought that this was because the all-zero filterbank was more suited to nulling formants because they typically encompass a greater bandwidth than the harmonics of musical signals, and are more widely spaced apart in the spectrum, giving rise to less interference between the filters' passbands. This problem was overcome by replacing the all-zero filters with an equivalent cascaded bank of second-order IIR notch filters of system function

$$H(z)=\prod_{\substack{k=1\\ k\neq m}}^{M}\frac{1-2\cos\omega_{Nk}z^{-1}+z^{-2}}{1-2\rho\cos\omega_{Nk}z^{-1}+\rho^{2}z^{-2}} \qquad (2)$$

where M is the total number of harmonics in the signal, ω_{Nk} is the frequency of the k^{th} notch, ρ is the radius of the pole from the unit circle, and *m* is the number of the harmonic being measured. Figure 1 shows a block diagram of the basic filterbank structure,

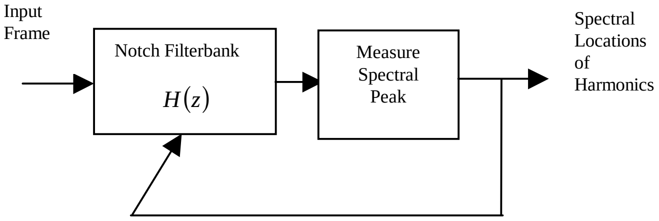

Figure 1. *Block Diagram of Harmonic Tracking Scheme.*

This substitution was found to give far superior performance. Figure 2 shows an example of the tracking scheme applied to a violin note that contains eight harmonics, where the tracks in the figure are marked by black lines with crosses. The number of notch filters in the cascade was set to seven and a pole radius of $\rho = 0.5$ was found to give the best trade-off between filter selectivity and bandwidth.

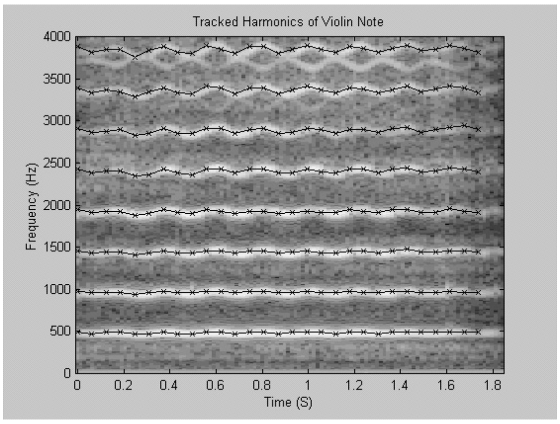

Figure 2 *Harmonic Tracking Scheme applied to a Violin Note*

Following the tacking and extraction of the harmonics from the sound, the log-envelopes' and PIFs' for each harmonic in every frame could then be estimated. This MinP/AllP decomposition of each harmonic segment was achieved using the LPSD algorithm described in [6]. Since in practice successive frames or segments of the signal are processed, the algorithm operates under the assumption that, the carrier frequency ω_c and the amplitude A_c of the signal model, as given in (1), are slowly varying quantities over the total signal length, in comparison to the modulations $\alpha(n)$ and $\beta(n)$. The LPSD algorithm produces a signal $h(n)$ given by

$$h(n)\approx e^{-(\alpha(n)+\beta(n))}e^{-j(\hat{\alpha}(n)+\hat{\beta}(n))} \qquad (3)$$

where $1/h(n)$ is an approximation of the MinP component of the k^{th} harmonic $s_k(n)$.

The log envelope is then given by

$$1/|h(n)| = \left|e^{\alpha(n)+\beta(n)}\right| \qquad (4)$$

Also, an error signal can be formed as

$$e(n)=s_k(n)h(n)=A_c e^{j(\omega_c n-2\hat{\beta}(n))} \qquad (5)$$

which provides an approximation to the AllP component. The positive instantaneous frequency of the AllP component can then be computed as

$$d\angle e(n)/dt=\omega_c-2\dot{\hat{\beta}}(n) \qquad (6)$$

For each harmonic segment of approximately 64 msec in duration, the LPSD algorithm was applied to decompose it into its log-envelope and PIF. It was found however that a smoother estimate of the PIF could be obtained by using the estimator given in [7] rather than the using the expression given by (6). The estimator given in [7] can be written as

$$\hat{f}_i(n) = \frac{F_S}{2\pi} \arg\left[\sum_{k=0}^{N-1} B(k) e(n-k+1) e^*(n-k)\right] \qquad (7)$$

where F_s is the sampling frequency and the smoothing function is

$$B(k) = \begin{cases} 1 & k = 0,\ldots,M-1 \\ 0 & k = M,\ldots,N-1 \end{cases}$$

with the value of *M* determining the degree of smoothing. It was found that a value of *M*=5 was sufficient to smooth the PIF without sacrificing too much detail.

Figure 3 shows a plot of the extracted log envelopes of the harmonics in various colours for ease of identification, and Figure 4 shows the PIF traces of these harmonics overlaid on a spectrogram. The boundedness and smoothness properties of these quantities are apparent from the figures.

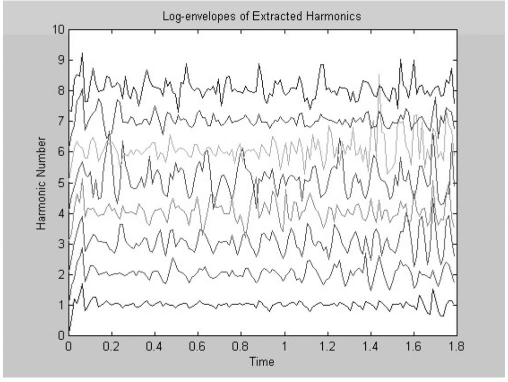

Figure 3. *Log-Envelopes of the Harmonics.*

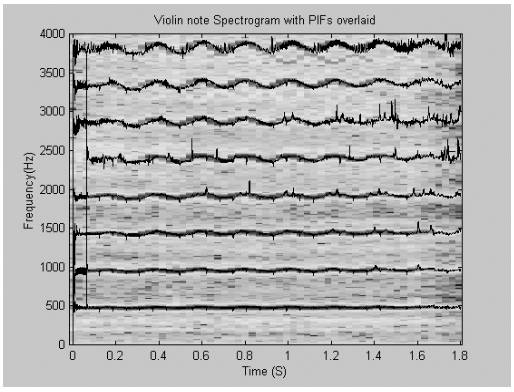

Figure 4. *PIFs of the Harmonics.*

3. EPS MODELLING

Once the log-envelopes and PIFs of each harmonic were calculated, they could then be passed to the modelling stage. The Exponential Signal Model (EPS) [5] appeared to be a natural choice for representing these quantities. Exponential Polynomial Signals are complex waveforms whose amplitude and phase vary over time according to a finite-order polynomial [5].

$$s(n) = \exp\left(a_0 + a_1 n + a_2 \frac{n^2}{2!} + \ldots + a_M \frac{n^M}{M!}\right) \qquad (8)$$

where the coefficients of the polynomial are unknown complex parameters.

The real parts of the polynomial coefficients specify the envelope of the signal, while the imaginary parts specify its phase, i.e.

$$|s(n)| = e^{\mathrm{Re}\left\{a_0 + a_1 n + \ldots + a_M \frac{n^M}{M!}\right\}} \qquad (9)$$

and

$$\angle s(n) = \mathrm{Im}\left\{a_0 + a_1 n + \ldots + a_M \frac{n^M}{M!}\right\} \qquad (10)$$

A number of approaches were tried to find the simplest and most robust method for determining the coefficients of the EPS polynomial description of the log-envelope and PIF components of each harmonic. For the method described in [5] it was found that modelling errors occurring when estimating the higher-order coefficients would propagate, because of the reliance on demodulation, to the lower-order coefficients and cause gross errors in their estimates. Similar difficulties were encountered in applying the DPT algorithm [8] to estimate only the signal phase coefficients as it again employed a demodulation procedure. Following this, the RLS-based method of [9] was applied. However, while it provided good tracking of the PIF as it varied within the segment, the final values of the coefficients would not always provide a reasonable reflection of the overall variation. Lastly, it was found that the most straightforward and robust approach was to use multivariate least square regression to obtain an initial estimate for the coefficients, where the coefficients describing a segment's log-envelope and the PIF were estimated separately, i.e.

$$\begin{bmatrix} \mathrm{Re}(\hat{a}_0) \\ \mathrm{Re}(\hat{a}_1) \\ \vdots \\ \mathrm{Re}(\hat{a}_M) \end{bmatrix} = \left(\sum_{m=0}^{N-1} \boldsymbol{d}_m \boldsymbol{d}_m^T\right)^{-1} \sum_{m=0}^{N-1} \boldsymbol{d}_m \ln \frac{1}{|h(n)|} \qquad (11)$$

where

$$\boldsymbol{d}_n^T = \left[1 \;\; n \ldots \; \frac{n^M}{M!}\right]$$

with *N* being the segment length, and

$$\begin{bmatrix} \mathrm{Im}(\hat{a}_1) \\ \mathrm{Im}(\hat{a}_2) \\ \vdots \\ \mathrm{Im}(\hat{a}_M) \end{bmatrix} = \left(\sum_{m=0}^{N-1} \boldsymbol{d}_m \boldsymbol{d}_m^T \right)^{-1} \sum_{m=0}^{N-1} \boldsymbol{d}_m \frac{F_s\left(\omega_c - 2\dot{\hat{\beta}}(n)\right)}{2\pi} \qquad (12)$$

where F_s is the sampling frequency.

The initial estimates could then be passed to an iterative Gauss-Newton procedure for refinement. This combination of methods appeared to provide a set of coefficients that gave a reasonable and reliable approximation to variations in the log-envelope and the PIF. Finally, an estimate of the gain $|A_c|$ was found as the mean of the absolute value of the LPSD error signal. For modelling purposes, the phase shift ϕ was ignored as accurate estimation was not possible, and further work determined that it did not have a significant perceptual impact.

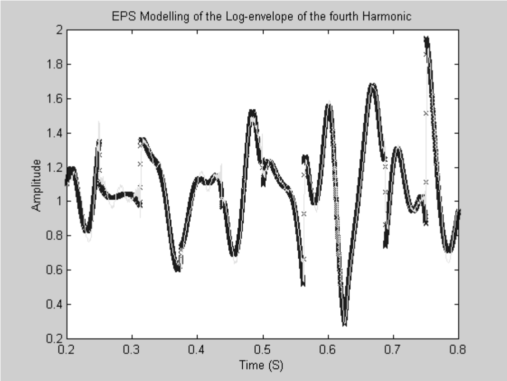

Figure 5. *Original and Modeled Log-envelopes.*

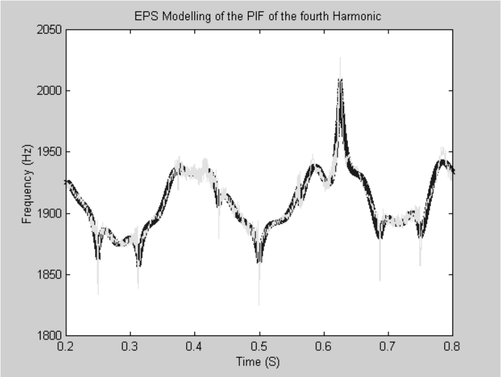

Figure 6. *Original and Modeled PIFs.*

Figures 5 and 6 respectively show the original log-envelope and positive instantaneous frequency signals for the fourth harmonic of a violin sound, along with the reconstructed versions using EPS modelling. The original signals are traced by the yellow line, while the reconstructed ones are shown by the thicker blue line. The order of the EPS model was chosen to be 6. A close match can be seen in both cases.

Re-synthesis of the sound is achieved by generating the appropriate polynomials for each segment, applying a Binomial filter to smooth minor discontinuities between segments and then summing across the harmonics. Note that in the case of the modeled PIF, an integration procedure must be carried out to get the appropriate phase angle. Also, the angle of the MinP component can be computed before reconstruction by taking the Hilbert transform of the log-envelope and then forming the complex exponential.

4. AM-FM MODIFICATION VIA EPS PARAMETERISATION

Given the smoothness of the AM-FM components produced, the EPS model was an ideal parameterisation that would allows new types of signal modification at both a coarse and fine granularity to be carried out. For example, vibrato can be introduced or removed by modulating or smoothing the first term of the EPS description of the PIF. More fine grain effects can be performed by affecting changes to the higher orders of the imaginary EPS coefficients that will manifest themselves in the frequency domain as variations in the structure of the signal bandwidth. Another possibility is non-linear time-stretching and compression effects that can be carried out by specifying a non-linear time vector in the resynthesis stage. Figure 7 shows the result of modulating the first coefficients of a 6th order EPS model for the first seven harmonics with a cosinusoidal function of the form $0.095\cos(\omega' n') + 1$, where n' corresponds to the frame index and the normalised frequency $\omega' = 1.1184 rad/s$. This had the perceptual effect of transforming one note into many.

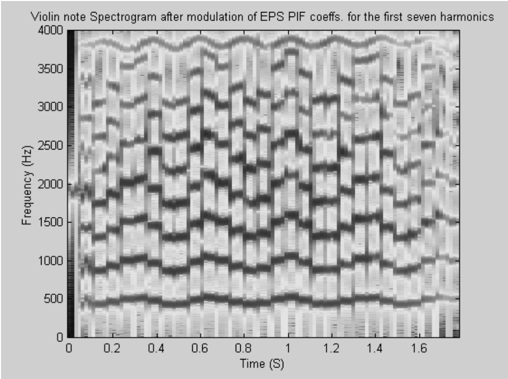

Figure 7. *Spectrogram of Violin note whose EPS coefficients are modified by a cosine function.*

The next example, in Figure 8, shows the spectrogram of the same note, with the first half of the note being stretched in time by a factor of two, with the remaining half left untouched. Perceptually this lengthens the attack of the note giving it a more gradual rise in intensity and produces a strong sense of shimmer within the vibrato of the note.

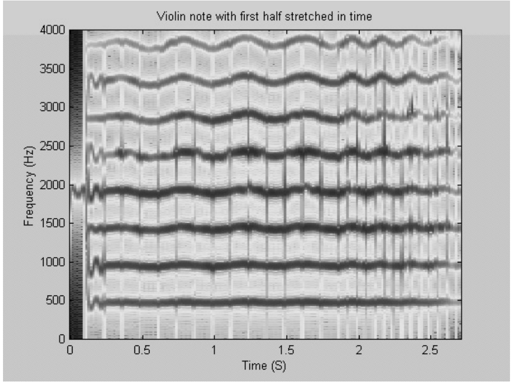

Figure 8. *Spectrogram of Violin note with the first half stretched in time.*

Figure 9 shows the spectrogram of the note when a time-varying non-linear time stretching function of sinusoidal shape, shown in Figure 10, is used.

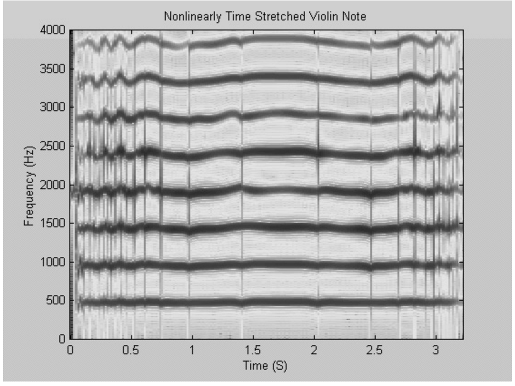

Figure 9. *Spectrogram of Violin note with non-linear time stretching applied.*

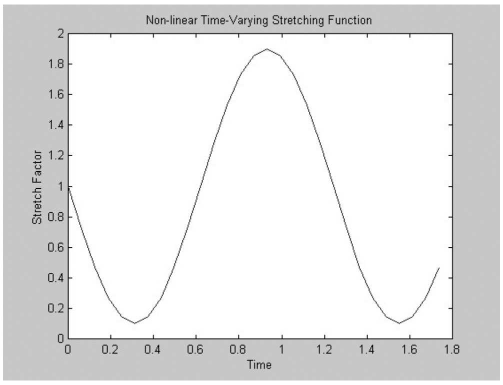

Figure 10.*Non-linear time stretching function applied to the note.*

The perceptual impressions of the attack and decay portions of the sound are of a more rapidly executed note with the time shortening resulting in the appearance of a more intense vibrato. In the middle portion of the sound the vibrato is drawn out, to a degree beyond what is commonly expected are vibrato, creating a sound that is unusual in time but not in timbre.

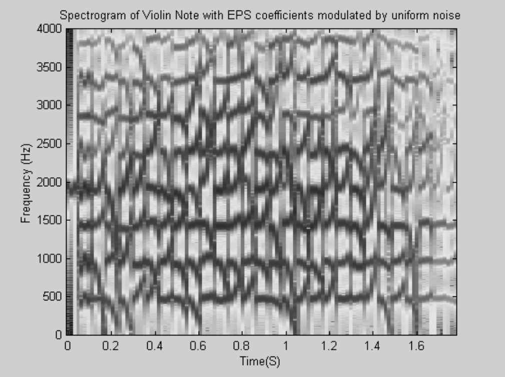

Figure 11. *Spectrogram of Violin note with the EPS coefficients modulated by uniform noise.*

The last example in Figure 11 shows the spectrogram of the resulting sound where all the EPS coefficients of a 5th order EPS model for the harmonics are modulated using a uniformly-distributed random number sequence. In this example, only the first EPS coefficient in the set representing the PIF centre frequency of each harmonic in each segment left untouched. The sound appears the contain the original violin note with an interesting set of extraneous frequencies that twist through it, giving the impression of two sounds imposed on one another.

5. CONCLUSIONS AND FUTURE WORK

This paper has outlined methods for the extraction and subsequent modeling of the log-envelope/PIF quantities of the harmonics of music signals. Exponential Polynomial Signal modeling was shown to be a reasonable approach for representation of these quantities and some of the various methods for determining the coefficients of this model were discussed. Examples were then given of how the EPS parameterisation allowed some new audio effects to be created. Future work intends to investigate more of the myriad of effects possible and to assess which ones are most interesting for composers of new electro-acoustic music.

6. REFERENCES

[1] J.L. Flanagan, *Speech analysis, synthesis and perception*, Berlin: Springer-Verlag, 2nd ed., 1972.

[2] T. Wishart, *Computer sound transformation*, http://www.trevorwishart.co.uk/transformation.html.

[3] A. Potamianos, '*Speech processing applications using an AM-FM modulation model*,' PhD. Thesis, Harvard University, Massachusettes, MA, Aug.1995.

[4] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech and Audio Proc.*, Vol.8, No. 3, May 2000, pp. 240-254.

[5] S. Golden and B. Friedlander, "Estimation and analysis of Exponential Polynomial Signals," *IEEE Trans. Sig. Proc.*, Vol.46, No.11, Nov. 1998, pp. 3127-3130.

[6] A. Rao and R. Kumaresan, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *Journl. Acoust. Soc. Amer.*, vol. 105, no. 3, March1999, pp. 1912-1924.

[7] G.W. Lank, I.S. Reed and G.E. Pollon, " A semicoherent detection statistic and doppler estimation statistic," *IEEE Trans. Aerospace and Electr. Syst.*, vol. AES-9, no. 2, 1993, pp. 151-165.

[8] S. Peleg et al.," The Discrete Polynomial Transform: its properties and applications", *Proc. 25th Asilomar Conf,*. 1991, pp. 763-767.

[9] B. Slocumb et al, "A polynomial phase parameter estimation-phase unwrapping algorithm", *ICASSP 1994*, IV, pp. 129-132.