The Short-Time Periodicty Transform and its Applications to Audio Synthesis*

JACOB SUNDSTROM
(jlsundst@ucsd.edu)

University of California, San Diego, Department of Music
La Jolla, California, United States

This paper discusses the application of the so-called periodicity transform as described by William Sethares and Thomas Staley to analyze and resynthesize a time-domain audio signal. In so doing, considerations unique to its application to time-domain audio are examined and a short-time periodicity transform is developed. By analyzing a signal using a window of a carefully selected length to produce a set of nonorthogonal periodic basis vectors—which are then "naively inverted" by summation—a highly accurate reconstruction can be created from purely periodic, non-sinusoidal signals with overlapping spectra. A modified version of the M-best γ algorithm is used to analyze the signal in the short-time implementation to account for fractional periods.

0 INTRODUCTION

In [1], W. A. Sethares and T. W. Staley introduced the concept of the "periodicity transform" (PT) whereby a signal can be decomposed into periodic basis vectors by projection onto periodic subspaces. The PT excels in situations where the time series is described best in terms of period rather than frequency. This technique and its variations have been applied to astromonical data ([2]), machine vibration ([3]), gene sequencing ([4]), and musical rhythms ([5]); however, its applcation to time-domain audio signals directly has been lacking. Since the basis vectors calculated with the PT often have overlapping spectra, it seems a natural extention to apply the PT to both analysis and synthesis of audio, and especially to musical audio. In doing so, a short-time periodicity transform was developed in order to capture the changing periodic nature of musical signals and is applied to analyze and resynthesize an excerpt of a well-known electronic work.

Section 1 gives a brief overview of the periodicity transform and the M-best γ algorithm. Section 2 describes the particular variations used for audio analysis, including a description of the short-time transform. Section 3 details an analysis and resynthesis of a section from Charles Dodge's *Speech Songs* compared with an FFT analysis and resynthesis.

1 THE PERIODICITY TRANSFORM[1]

The periodicty transform (PT) has been described multiple times in various flavors, notably in [1], [6], and [3]; this paper works with the PT as implemented in [1]. In a nutshell, the PT decomposes a signal into a set of periodic basis vectors by projecting the signal onto a set of "periodic subspaces", P_p. As such, the set of basis vectors which best describe the signal are not required a priori but are instead derived numerically. In contrast to the FFT, the periodic subspaces are nonorthogonal and the transform in general is based in large part on the projection theorem as defined by Luenberger [7]. As noted by Sethares and Staley: "It is not a transform in a strict sense, rather it is a transform by analogy with wavelet or Fourier transforms." [1]

A sequence of real numbers $x(k)$ is called p-periodic if there is an integer p with $x(k+p) = x(k)$ for all integers k. In practice, we will consider signals of finite length N, which are always periodic with period P_N. Smaller periodicities are found by projecting x_N onto the subspaces P_p for $p < N$. When x_N is "close to" a periodic subspace P_p then there is a p-periodic element x_p that is contained within x_N. The fundamental formula for projection is:

$$\alpha_s = \frac{1}{\lfloor N/p \rfloor} \sum_{n=0}^{\lfloor N/p \rfloor - 1} x_{N_i}(s+np) \tag{1}$$

*Jacob Sundstrom; e-mail: jlsundst@ucsd.edu

[1] Following the conventions in [1], let P denote the set of all periodic sequences, P_p denote the p-periodic subspace where p is an integer, and N denote the length of the input signal x. Note that N is an integer and P is a set.

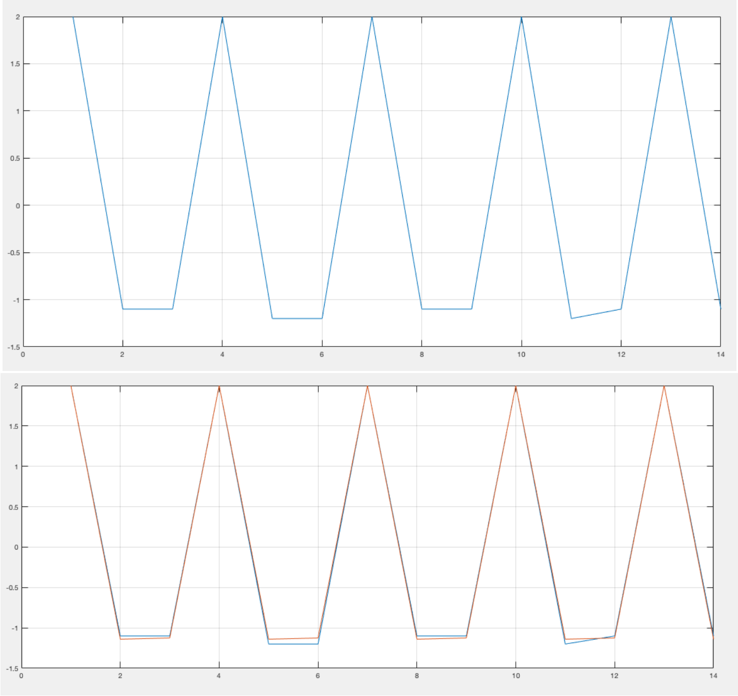

Fig. 1. *Top*: The signal x. *Bottom*: $\pi(x,P_3)$ overlaid onto x.

where $N_i = p\lfloor N/p \rfloor$ to deal with partial sequences.[2] Most notably, this transform works only with *integer* periods.

This processes is most easily understood in the context of a short example. Let $x = \{...,2,-1.1,-1.1,2,-1.2,-1.2,2,-1.1,-1.1,2,-1.2,-1.1,2,-1.1,...\}$ where $N = 14$. The signal is plotted in Fig.1, top. Let $p = 3$ which means $s = \{0,1,2\}$. For $s = 0$:

$$(s+np) = [1,4,7,10,13]$$

$$x(s+np) = [2,2,2,2,2]$$

Therefore:

$$\alpha_0 = \frac{1}{\lfloor 14/3 \rfloor} \sum_{n=0}^{\lfloor 14/3 \rfloor - 1} x(s+np) = 2$$

Then replace all values at indicies $(s+np)$ with 2, giving $[2,2,2,2,2]$. Similarly, $\alpha_1 = [-1.14,-1.14,-1.14,-1.14,-1.14]$ and $\alpha_2 = [-1.125,-1.125,-1.125,-1.125]$. Interleaving the results, we find $x_3 = \{...,2,-1.14,-1.125,...\}$. This projection is plotted over the original signal in the bottom of Fig.1.

More generally,

$$x_p = \pi(x,P_p) = \sum_{s=0}^{p-1} \alpha_s \delta_p^s \tag{2}$$

where, $\pi(x,P_p)$ represents the projection of x onto P_p, δ_p^s are the p-periodic basis elements of P_p, and s is as defined above. (2) describes the method of projection used in this paper and the M-best γ algorithm. Of note is the fact that these periodic basis vectors as acquired by (2) are *not* orthogonal. Muresan and Parks in [3] presented a method of finding orthogonal periodic basis vectors but their method was not implemented for this paper (and may in fact be unncessary when resynthsizing audio).

1.1 The M-Best γ Algorithm

Since the operation described in (2) merely *projects* a signal onto a periodic subspace, the process by which decomposition is performed must be described seperately. Of the four algorithms presented in [1], the M-best γ algorithm most consistently produces the best results in terms of decomposition and resynthesis.

M-best γ is a greedy algorithm which works in two stages: initially by finding the M-best periodic basis functions (where M is the number of desired basis elements), then by examining each of the factors of the M-best periods in order to find the best representation. Both stages are detailed further in subsequent paragraphs, although it is important to discuss exactly what measure of "best" is used.

In deciding which periodic basis vectors to choose, the "induced norm" of x is used:

$$||x|| = \sqrt{\langle x,x \rangle} \tag{3}$$

where $\langle x,x \rangle$ is an inner product so that

$$\langle x,x \rangle = \lim_{k\to\infty} \frac{1}{2k+1} \sum_{i=-k}^{k} x^2(i) \geq \frac{\varepsilon}{p} > 0$$

Of interest is the fact that (3) gives the same value whether x is considered to be an element of P_p, P_kp (for all positive integers k), or P. In the M-best γ algorithm, the value used to evaluate the quality of the vector it is taken

[2] This is slightly modified in this paper, described in Section 2.2.1.

one step further: $\frac{||x||}{\sqrt{p}}$. This is the measure of "energy" used in all analyses presented in this paper.

The first stage, as noted above, merely compiles a list, x_{q_i}, of the M-best periodicies by searching through all possible (allowed) periods in x and keeping the basis vector with the largest induced norm. For the sake of clarity, let x_{q_1} be the strongest basis vector in x with a period of p_1. Then, the residual is taken so that $r_p = x - x_{q_1}$. This process is repeated using r_p in place of x until the initial list of M-best periodicies (along with their representative basis functions) are found.

In the second stage, for each periodicity p_i in x_{q_i}, the factors are computed with the exception of 1 and itself (since, in this case, all basis vectors are periodic with both 1 and itself) so that $Q \in p_i$. The reason for this is that by Theorem 3.2 in [1], the projections onto subspace np necessarily contains the projections onto subspace p. The sequence x_{q_i} is then projected onto each factor creating the basis vector x_Q and the induced norm is computed. Then, if $||x_Q|| > \min(||x_{q_i}||)$, and the sum of $||x_{q_i}||$ would increase as a result of replacing x_{q_i}, then x_{q_M} is removed and x_Q is added to the end of the list. This process is repeated for all members of x_{q_i}, including newly added members, until no further changes take place. Also of note is the fact that the M-best γ algorithm will *always* return M periodicities and thus sometimes returns "false positives". A modification to avoid this and improve the computed basis vectors has been implemented but was not included in this paper.

In this paper, a slight modification is made to the projection method within the algorithm. In the case where p is not an factor of N, the signal is extended by wrapping and linearly crossfading so that the "missing" portion of the period is concatenated to the end of x. Doing so gives a sightly better representation of the period p in x. This effect is exaggarated as $N/p \to 0^+$ given a steady period. This is detailed in Section 2.2.1.

2 APPLICATION OF THE PERIODICITY TRANSFORM TO AUDIO

As seen in [1], the PT as described has been applied to musical signals. However, in that case it was applied to musical *rhythms* as binary sequences in order to extract meter and tempo information as opposed to time-domain audio to extract other time-domain signals. It is a natural extension, then, to apply this technique digital audio and natural again to attempt to resynthesize the original signal. This section describes the various extentions and details of how the PT can be applied to applied to musical audio.

2.1 Considerations Unique to the Audio Domain

The largest disadvantage with the PT as described thus far is its inability to find fractional periods; the reliability and efficiency of finding fractional periods is a source of ongoing research. However, since the PT as described in Section 1 does not produce orthogonal basis vectors, it is possible to recover fractional periods in the reconstruction by summing enough basis vectors from an analysis so that the weaker periods "correct" the basis vector of those periods which contribute much more energy to the original signal. This phenomenon is the basis of the method of analysis used in this paper and will be described analytically in future work.

In utilizing the M-best γ algorithm, one can ask for the return of an arbitrary number of basis vectors. In practice, it was found that using "nested" algorithms produced better results in terms of the (lack of) relative power of the residual signal. That is, instead of asking the M-best γ algorithm to return 15 periodic basis vectors, it is better to run 3 serialized M-best γ calls for 5 periods each (a 5x3 structure), passing the residual signal after decomposition to each subsequent iteration of the algorothm, retaining all basis vectors created throughout the process. There is, of course, a trade-off between speed and accuracy so choosing a good nested structure is important.

Likewise, it is possible to limit the range of periods to those that span the audible range and, in a musical context, to those frequencies that might be considered "musically useful". There is no sense, for instance, in computing for $p = 2$ or $p = 3$, or even $p = 100$ at a sampling rate of 44.1 kHz, as those periods represent extremely high frequencies which, if present, are likley a harmonic of some longer period and can thus be captured therein by exploiting the nonorthogonality of the periodic basis vectors. Future versions that use orthogonal basis vectors will need new techniques to overcome this.

2.2 The Short-Time Periodicty Transform

Analagous to the short-time Fourier transform (STFT), the short-time periodicity transform (STPT) is a way to measure changes in the periodic components of a signal through time by means of windowing. In order to more accurately capture the changes in time, the STPT also uses overlapping windows. Unlike the STFT however, taking anything but a rectangular window will distort the periods the PT is able to find and thus compromise a resynthesis.

The size of the window has no practical limit but was found to best be a function of the largest-sought period (lowest pitch) and expected rate of change of periods. In an auditory context, the lower limit of human hearing is approximately 20 Hz which gives a period of 2205 samples at a samplerate of 44.1 kHz. Therefore, if we want to limit the largest period to 1/3 of the window size as in the case of the default in the M-best γ algorithm, we find the smallest window necessary to capture the fundamental of the lower limit of human hearing to be $N = 6615$ at 44.1 kHz, or 0.15 seconds. In practice, however, and especially with regard to a musical context, a window between 0.02 and 0.05 seconds in the M-best γ algorithm has proven to be sufficient except in the most extreme cases.

As a demonstration of the STPT's ability to capture changing periods, Fig.2 shows the results from a sine sweep from 100 Hz ($p = 441$ @ 44.1 kHz) to 1000 Hz ($p = 44.1$ @ 44.1 kHz) as a heatmap of the powers of all the periodic components found in each window. The change in the sine wave is linear in period, decreasing from

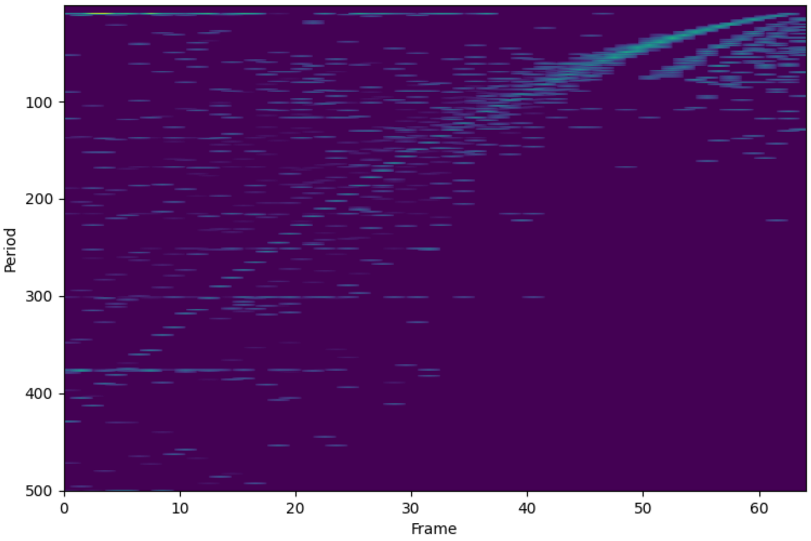

Fig. 2. Plot showing the change in periodicity captured by the STPT of a sweeping sine tone. Note the smattering of “false” periods when the actual period captured in the window is not an integer (and in fact shifts over the course of the window).

441 to 44.1 over 2 seconds ($N = 88200$). The analysis consists of a window size of 1500 samples (approximately 0.034 seconds), overlapping each neighboring window by 150 samples in a 2x10 strucutre (see Section 3 for a description of the nested strucure used on audio). Note the smattering of “false” periods when the actual period of the window is not an integer. Additionally, the current method of plotting the results of a PT do not necessarily represent the accuracy of a resynthesis, since as noted in the first part of this section, some of the weaker basis vectors help to correct the errors of others and M-best γ reutnrs M basis vectors regardless.

Since the resulting reconstruction of a PT is in the time domain directly, overlapping windows must have a linear crossfade in order to preserve the waveforms at the beginning and end of the windows. This crossfade and overlap does not cause phase interference between subsequent windows and thus provides a means for the smoothest transition between changes in periodicity.

2.2.1 *Whole Period Projection*

The only change made to the M-best γ algorithm is slight but critical. In the case where p is not an integer factor of N, the signal is *extended* so that the end of the signal aligns with $\phi_p = 0$. That is, the signal is extended to a multiple of p.

This is accomplished by wrapping and linearly crossfading so that the ”missing” portion of the period is concatenated to the end of x by the number of samples necessary to let $N = \left\lceil \frac{N}{p} \right\rceil p = N_{p^+}$ where N_{p^+} is the new number of samples in the window. The sample on which to begin the crossfade is then $N_{p^-} = \left(\left\lfloor \frac{N}{p} \right\rfloor \times p \right) - 1$ (assuming zero indexing). The total number of samples to crossfade is simply, $N_{p^+} - N_{p^-} + 1$. In doing so, a more accurate representation of the signal as projected onto periodic subspace p is obtained, assuming the period extends towards infinity in both directions.

3 ANALYZING AND RESYNTHESIZING DODGE’S *Speech Songs IV: The Days are Ahead*

The power of the STPT for analyzing and synthesizing audio can be aptly demonstrated using an excerpt from Charles Dodge’s *The Days are Ahead* from his 1974 work, *Speech Songs*. The excerpt used in this paper is from 0:41 to 0:51 (“Nine-hundred twenty-six thousand...”) ([8]). At this point in the piece, the synthesized voices begin to overlap in a series of descending glissandos. The signal is sampled at 44.1 kHz and was converted from 16-bit PCM to 32-bit float for higher accuracy during the analysis and ease of processing in Python.

It was found that using a 5x10 structure (five nested M-best γ iterations where $M = 10$) marked a good compromise between compute time and a satisfactory result. The window size was set according to the period of the lowest fundamental in the excerpt, which is about 84 Hz at approximately 8.5 seconds. Since the maximum longest period by default in the M-best γ algorithm is $N/3$ where N is the length of the input signal, the window length for the STPT was found to be $N = \left\lceil \frac{44100}{84} \right\rceil \times 3 = 525 \times 3 = 1575$.[3] This amounts to approximately 0.036 seconds at 44100 kHz. Note that the ideal input parameters to the

[3] Note that it is possible to adjust the largest possible period in the M-best γ algorithm to be less than or equal to N. However, in the interest presenting a novel technique as simply as possible, the default parameters given by [1] were kept.

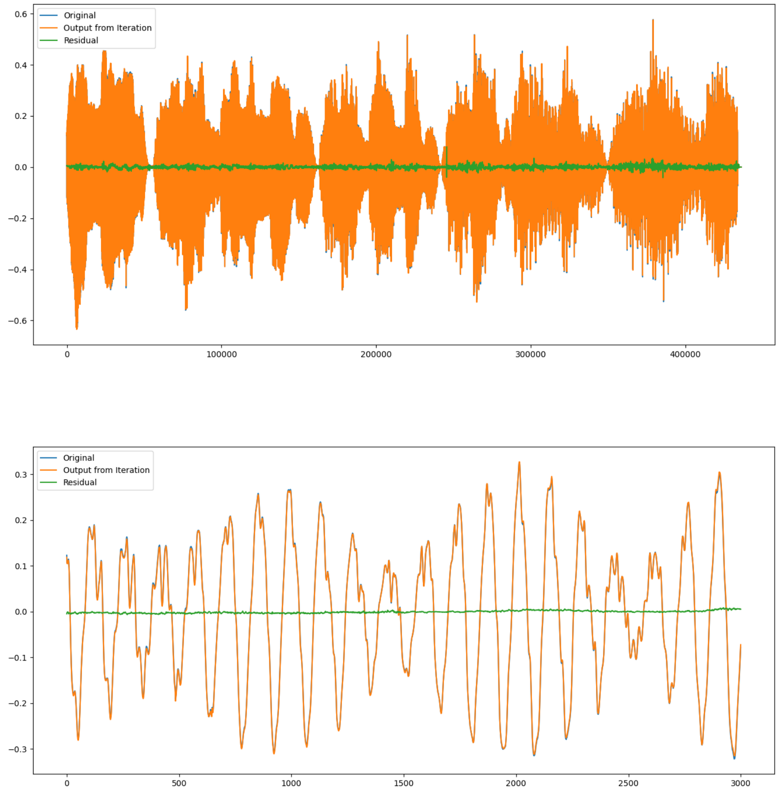

Fig. 3. *Top*: The original signal (blue, mostly hidden), the resynthesis (orange), and the residual (green). The residual signal is the difference between the original and resynthesis. *Bottom*: Detail of samples 65,500 to 68,500.

STPT as a function of the properties of the input audio signal have not yet been determined.

The window was then resynthesized using only periodic sequences by "inverting" the analysis (i.e. summing the basis vectors computed during analysis). Subsequent windows were concatenated using a linear overlap equal to the analysis overlap, set to 0.1 so that 10% of the signal on either side of the window overlaps with adjacent windows. This produced a seamless transition from one resynthesized window to the next. The resulting resynthesis is shown in Fig.3, with audio examples of the original signal, the resynthesized signal, and the residual available at http://notthatintomusic.com/papers/stpt_resynth/. Note that in the bottom of Fig.3, it is impossible to know (without knowing so already) where the windows of the analysis are.

Upon listening to the resynthesis and comparing to the original signal, it is clear that the STPT indeed does a very good job of resynthesizing the orignal signal from only periodic basis vectors.

3.1 Quantifying the Resynthesis

The question, though, is how to quantify the quality of resynthesis. Many methods have been proposed, including the Perceptual Evaluation of Audio Quality (PEAQ) standard ([9]). However, in in interest of clarity for the reader, this paper limits the quantifiable measures to the mean squared error, the correlation coefficient, and the signal-to-noise ratio, calculated in two ways. The formula for these calculations are given in the Appendix.

The mean squared error (MSE) between the resynthesis and the original signal is 3.7972×10^{-11} (this would be 0 if the two signals were identical). The correlation coefficient is perhaps one of the better measures used in this paper to examine the differences between the two signals. It is denoted by $\rho_{x,y}$ where x is the original signal, and y is the signal plus noise. In this case, we find $\rho_{x,y} = 0.99958$ (note that $\rho = 1$ if x and y are identical). SNR_ρ is found to be 30.77. If we take the ratio of the average power of the resynthesis to the average power of the noise, defined here

Table 1. Comparison between the STPT and the STFT

Value	STPT	STFT
ρ	0.99958	0.99981
SNR_ρ	30.776	34.371
SNR_{dB}	30.766 dB	28.499 dB
MSE	3.7972×10^{-11}	6.8563×10^{-11}
MSE / σ_x^2	1.9214×10^{-9}	3.4694×10^{-9}

Note: The STFT analysis was conducted with a Hamming window, 25% overlap, and a window size of 1024 samples.

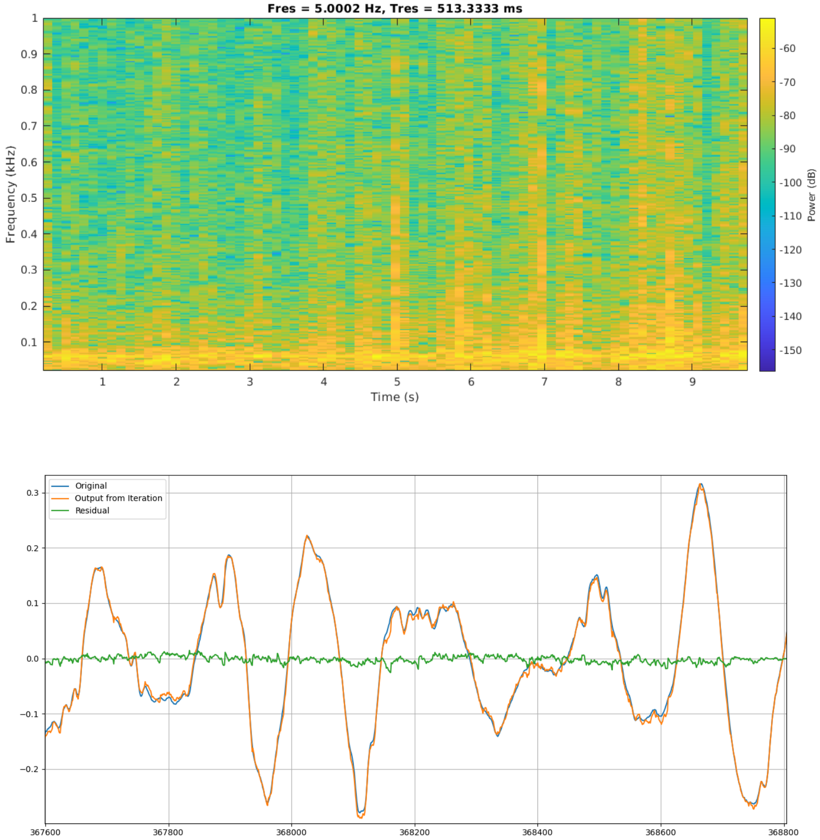

Fig. 4. *Top*: Spectrogram of the residual signal between 0 and 1000 Hz. *Bottom*: Detail between sample 367600 (8.33 seconds) and sample 368800 (8.36 seconds) showing extraneous noise.

as SNR_{dB}, we find that to be 30.77 dB for the STPT resynthesis. These values and those of an STFT resynthesis are printed in Table 1.

An analysis of the residual signal itself is perhaps more revealing. A spectrogram (Fig.4) reveals low frequency energy close to the end of the signal at about 8.5 seconds where the periods of all the voices become long relative to the window size and hence, the maximum allowed period. This shows that, as suspected, most of the energy in the high partials is removed via the nonorthogonality of the periodic subspaces and what is left is primarily the non-periodic portions of the signal *lower* (in pitch) than the largest allowable period. Inspecting closer, there is a clear band of noise around 65Hz or 678 samples throughout which is lower than the largest allowable period in the analysis at 84Hz or about 525 samples. This is expected, since the algorithm was not ”allowed” to reconstruct any periods larger than this. However, this situation could possibly be mitigated in the future by using a dynamic window size at the expense of temporal accurancy. (In the interest of simplicity, the window size remained fixed throughout the analysis.) It is notable, however, that the length of the window as it was set did not cause the STPT to fail to catch the sometimes rapid changes in pitch in the excerpt.

Also in the same area around 8.5 seconds, we find energy in frequencies above the lower limit set in the algorithm. Upon closer examination as shown in Fig. 4, bottom, it is clear that the resynthesis contains significantly more noise that the original signal. This is perhaps a good demonstration of the weakness of using the M-best γ ”as-is”: when tasked with returning M periodicities, it will return M periodicities regardless, even those which are not actually present, but only a function of “false” positives in the noise. This suggests that a dynamic structure could mitigate this, whereby as the reconstruction of a window comes within a power tolerance, the algorithm is ceased. Likewise, if the reconstruction has not yet breached the tolerance, the algorithm continues until another threshold, perhaps maximum iterations, is reached.

Additionally, the analysis/resynthesis was also computed using the same 5x10 structure but with a window size of 1000 samples. While the resulting resynthesis was still rather convicing, the error towards the end of the excerpt was greatly exaggarated when compared with the window of 1575. This is unsuprising but suggests that longer windows in the STPT do not smear information in the same way a STFT does. There will be, of course, an upper limit to the practical window size as a function of the periodic content of the window but this has yet to be derived.

4 CONCLUSION

This paper investigated the application of the periodicity transform as described by Setheres and Staley to time-domain audio using the M-best γ algorithm presented in their paper. In doing so, a short-time periodicity transform was developed which can capture changes of periodic sequences over time. The orignal signal was recoverable by “naively inverting”; that is, by summing each of the basis vectors computed during decomposition.

While it is clear that STFT remains slightly superior in some respects — most notably computation time — the yet-refined technique of STPT analysis/resynthesis of time-domain audio signals is plausible. In particular, the various basis vectors possess overlapping spectra and may ease some of the issues of phase coherence when returning to the time domain after manipulating audio in the frequency domain by operating on each basis vector separately. Additionally, the PT is able to untangle nearly-coincident harmonics in a signal that the Fourier transform is not. This was demonstrated via and analysis and resynthesis of Charles Dodge’s *The Days are Ahead* using only a maximum of 50 basis vectors. In doing so, the nonorthogonality of the basis vectors in the transform described in [1] was exploited to capture harmonics of a given fundamental without having to derive these independently.

5 REFERENCES

[1] W. A. Sethares, T. W. Staley, “Periodicity transforms,” *IEEE transactions on Signal Processing*, vol. 47, no. 11, pp. 2953–2964 (1999).

[2] R. Buccheri, B. Sacco, “Time Analysis in Astronomy: Tools for Periodicity Searches,” in *Data Analysis in Astronomy*, pp. 15–27 (Springer) (1985).

[3] D. D. Muresan, T. W. Parks, “Orthogonal, exactly periodic subspace decomposition,” *IEEE Transactions on Signal Processing*, vol. 51, no. 9, pp. 2270–2279 (2003).

[4] R. Arora, W. A. Sethares, “Detection of periodicities in gene sequences: a maximum likelihood approach,” presented at the *2007 IEEE International Workshop on Genomic Signal Processing and Statistics*, pp. 1–4 (2007).

[5] W. A. Sethares, T. W. Staley, “Meter and periodicity in musical performance,” *Journal of New Music Research*, vol. 30, no. 2, pp. 149–158 (2001).

[6] P. P. Vaidyanathan, S. Tenneti, “Srinivasa Ramanujan and signal-processing problems,” *Philosophical Transactions of the Royal Society A*, vol. 378, no. 2163, p. 20180446 (2020).

[7] D. G. Luenberger, *Optimization by vector space methods* (John Wiley & Sons) (1997).

[8] C. Dodge, “Speech Song IV: The Days are Ahead,” (2010), released on CD with New World Records. Catalog Number NWCRL348. Originally recorded 1976.

[9] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, “PEAQ-The ITU standard for objective measurement of perceived audio quality,” *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29 (2000).

APPENDIX: ERROR MEASURES

In the following formula, x is the orignal signal while y is the resynthesis except in SNR_{dB} where x is the resynthesis and y is the residual.

The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{N}\sum_{i=0}^{N-1}(x[i]-y[i])^2$$

The correlation coefficient ρ as used in this paper is defined as:

$$\rho_{x,y} = \frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y}$$

SNR_ρ is defined as:

$$SNR_\rho = \frac{\rho^2}{1-\rho^2}$$

with ρ and $\rho_{x,y}$ being equivalent.

SNR_{dB} is defined as ratio of the average power of the signal x to the average power of the noise y (in this case taken as the residual signal):

$$SNR_{dB} = 20\log_{10}\left(\frac{x_{rms}}{y_{rms}}\right)$$

where x_{rms} and y_{rms} are the average powers of signals x and y.

THE AUTHOR

Jacob Sundstrom

Jacob Sundstrom is a PhD Candidate at the University of California, San Diego in the Department of Music where he studies under Miller Puckette. As an artist, he is interested in exploiting idiosyncratic properties of mediums as a way to expose hidden and latent layers of meaning. His creative projects have been exhibitioned worldwide and his work on brain-computer-music-interfaces has been published in *Frontiers in Human Neuroscience*. Sundstrom's research interests include audio analysis, signal processing, sound spatialization, and some other nonsense that may or may not be related to music.