VIBRATO AND TREMOLO PRESERVATION DURING PHASE VOCODER TIME-DILATION

Dr. Ted Apel

Boise State University
Department of Music, Boise, Idaho
tedapel@boisestate.edu

ABSTRACT

The phase vocoder technique has proved to be an effective method of "time stretching" musical sounds, however, some musical features of the original sounds are not maintained during the transformation. We consider here maintaining the vibrato and tremolo characteristics within the phase vocoder. The vibrato and tremolo (sub-audio modulation) rates of a sound are maintained by modeling each channel of the phase vocoder in a higher order spectral domain, removing the modulations in this domain, and re-imposing them after time-dilation.

1. SUB-AUDIO MODULATIONS

Phase vocoder time-stretching affects the rate of vibrato and tremolo in musical sounds by slowing them proportionally to the amount of time-stretching. This behavior is for many applications an acceptable or preferred outcome. However, in many musical instrument performances, the particular vibrato and tremolo rates are important acoustic features that are critical to the musical perception of the performances. These sub-audio rate modulations, both in amplitude and frequency, are an important part of musical instrument identification, a performer's style, and musical expression. By preserving the original rates of vibrato and tremolo after a sound has been time-scaled, these important musical perception cues may be retained. Here we will develop a novel method of maintaining the original vibrato and tremolo rates of a recorded sound using second order analysis of the phase vocoder representation of sound.

2. EXISTING METHODS

Carl Seashore performed extensive study of musical vibrato in the 1930's, in which he defined and characterized vibrato and tremolo of western musical instruments and voice [11]. He characterizes vibrato as a 4 to 8 Hz pulsation of pitch, loudness, and timbre, in which pitch change is approximately a semitone, and in which pitch and amplitude modulation rate are approximately constant. Seashore classifies both tremolo and vibrato as aspects of a single phenomenon he calls vibrato. Here we will separate vibrato as only a frequency modulation phenomenon and tremolo as an amplitude modulation phenomenon. When we wish to discuss any periodic or quasi-periodic low frequency modulation, whether amplitude, frequency or a combination or them, we will call them sub-audio modulations. Further discussion of the vibrato and tremolo characteristics of individual musical instruments can be found in Timmers and Dresian [13].

Research into sub-audio modulation of musical sounds typically involves automated detection of the existence of vibrato and or tremolo in a musical signal and estimation of the rate and extent of the modulation. Rossignol et al. suggest five methods of detecting and estimating the vibrato of a musical signal [9]. Two of the methods operate on the DFT spectrum of the signal: the first, comparing its shape to a synthetic spectrum, and the second comparing the shape to the shape at a different time. Three other methods operate on the fundamental frequency track, f_0, extracted by various methods, typically from the analytic signal generated by the Hilbert transform method. This f_0 signal is analyzed for its vibrato by calculating the DFT of the f_0 track, by extracting its fundamental frequency, or by computing the distance between local maxima in the f_0 trajectory. This first of these three methods, was originally developed by Herrera and Bonada [4].

Prior efforts to remove or modify vibrato from a recorded sound are generally based on sinusoidal analysis synthesis systems, that is, methods based on the MQ or SMS analysis synthesis systems [8, 12]. However, one method based on cepstral analysis will be presented here before sinusoidal based methods are reviewed.

2.1. Arfib and Delprat Cepstral Method

Daniel Arfib and Nathalie Delprat suggest a method of vibrato alteration of a sampled sound based on DFT, cepstral, and Hilbert pitch tracking techniques [1, 2]. Their method decomposes the pitch curve calculated in the cepstral domain into a mean frequency and an oscillatory frequency. The oscillatory frequency is modified before recombination with the mean frequency and transformation out of the cepstral domain.

Specifically, their technique is as follows. Starting with a phase vocoder representation of sound, each amplitude spectrum is converted to the cepstral domian. Each cepstral domain spectrum is separated into two parts (cepstral "liftering") consisting of the low coefficients which typically contain the spectral envelope spectrum and the

high coefficients which contain the so called excitation spectrum. The amplitude of the excitation spectrum is converted back to the spectral domain, and the low spectral peaks are used to track the fundamental frequency and the modulation rate or vibrato. This time varying fundamental frequency is subtracted from the modulation rate to produce a vibrato rate signal. The frequency and amplitude of this signal is in turn calculated with a Hilbert transform based analytic signal method. These rates are modified before recombination with the mean frequency and transformation obtained from the cepstral domain. Finally, the inverse cepstrum is combined with the original phase estimates of the phase vocoder to create a new sound. They report successful alteration of vibrato rates with errors produced from the pitch detection method and extraction of the vibrato.

2.2. Maher and Beauchamp Vibrato Method

Maher and Beauchamp created a method of removing vibrato from vocal sound recordings based on the MQ sound representation [5]. In their system, the frequency of sinusoidal analysis tracks are smoothed to remove the vibrato. The system uses frequency modulated wavetables to synthesize tones with vibrato, but does not allow for modification of vibrato rates or re-introduction of vibrato to time stretched sounds. However, it does appear to be the first time that vibrato had been modified independently of other musical parameters in a sampled sound.

2.3. Marchand and Raspaud Order-2 Modeling

A recent system proposed by Sylvian Marchand and Martin Raspaud creates a second order sinusoidal system [6] that is conceptually similar to Schumacher's 2DFT analysis introduced below. Their method analyzes the sinusoidal tracks of a sinusoidal modelling system using sinusoidal analysis, i.e. each amplitude and frequency track of a sinusoidal analysis is in turn analyzed as a series of sinusoidal tracks. This method is used in a time-stretching algorithm which preserves the vibrato and tremolo of the recorded sound. As the modulations of each amplitude and frequency spectral track follow the sub-audio variations in each track, a sinusoidal analysis of each of these partials will capture this sub-audio modulation.

Their method uses the standard method of time-stretching with a sinusoidal model, (i.e., interpolating between the amplitude and frequency of spectral track peaks) performed in this new second order domain. As these tracks do not represent the amplitude and frequency of spectral energy, but rather, the amplitude and frequency of individual amplitude and frequency tracks. As will be seen, the idea of a second order spectral representation will be used in our PV based system presented below.

3. RELATED DFT TECHNIQUES

Before our sub-audio retention method is presented, two signal analysis techniques that are incorporated into the method will be reviewed. In our system, the 2DFT second order analysis technique of Schumacher will be expanded to an analysis synthesis system by using the long-term discrete fourier transform (LTDFT) of Hammer and Sundt in the second order domain.

3.1. Schumacher 2DFT Analysis

Robert Schumacher defined the 2DFT method of analyzing low frequency and sub-audio periodicities [10]. His method performs a single DFT along each of the time channels of the spectrogram of a sound. The amplitude spectrum of this second order DFT shows the spectrum of the low frequency and sub-audio periodicity of each channel of the input sound. Averaging the second order DFTs across all of the first order spectral channels produces a spectrum in which low frequency and sub-audio amplitude modulation of the original sound can be identified as spectral peaks. Schumacher used this method to analyze the aperiodicities in periodic waveforms. Schumacher's 2DFT analysis method will serve as part of the analysis, transformation, synthesis system developed here to maintain sub-audio modulation rates. As will be seen, the 2DFT analysis allows us to model sub-audio amplitude modulations as spectral peaks that are maintained during phase vocoder time-stretching operations.

3.2. Long-Term Discrete Fourier Transform

The idea that a single Fourier Transform can be performed over long durations of time-variying sound in order to create novel transformations was formalized by Øyvind Hammer and Henrik Sundt [3]. Their research was inspired by the composition "Z" by Paul Pignon in which a single DFT was taken from a guitar and saxophone improvisation, and manipulations to the DFT coefficients were performed before inverting the DFT. Hammer and Sundt call this transformation the long-term discrete fourier transform or LTDFT and suggest several manipulations that are possible in this domain. Among these, they suggest (i) remapping amplitude spectra values to differing bins in order to create time dispersion of frequency bands or frequency sweeps and (ii) multiplying the phase spectra by a constant in order to move frequency bands in time. Most all of the manipulations in the LTDFT domain create sounds which disrupt the integrity of the original sonic events. That is, sounds become unrecognizable because their frequency components have been moved to radically different locations in the output sound. Unlike the phase vocoder representation of sound as amplitude and frequency spectra which localize sonic events in time, the LTDFT does not encode time in a perceptually relevant manner. This makes it it particularly difficult to predict the outcome of transformations performed in the LTDFT domain.

We will see in our sub-audio modulation retention method that applying the idea of the LTDFT in a higher order context helps us to model the globally relevant vibrato and tremolo rates. In general, a technique that produces abstract sounds in the first order LTDFT domain produces

a useful model of sub-audio modulation in a higher-order domain.

4. SECOND-ORDER PV BASED SUB-AUDIO MODULATION SYSTEM

We are now in a position to introduce our sub-audio modulation analysis/synthesis method based on the 2DFT analysis and LTDFT analysis/synthesis methods.

The phase vocoder typically employes a frame length of 512 or 1024 samples. Using a sampling rate of 44 100 Hz and a frame length of 1024 samples, the lowest frequency, f, that can successfully be analyzed is found by

$$f = \frac{SR}{N} = \frac{44\,100}{1024} = 43.066\,\text{Hz}. \tag{1}$$

Sub-audio frequencies are not captured explicitly as energy in low-frequency bins but as a pattern of changes across multiple spectral frames. The frame length required to capture modulations as low as 4 Hz, is

$$N = \frac{SR}{f} = \frac{44\,100}{4} = 11\,025\,\text{samples}. \tag{2}$$

A phase vocoder analysis with such a long frame length will smear note onsets and transients under sound transformation such as time-stretching. However, a phase vocoder with this frame length does model sub-audio modulations as spectral energy, and consequently PV time-stretching extends these frequencies at their original rates. If the sound to be analyzed is legato or does not have sharp transients this method may be appropriate.

In order to maintain the frame rate of a traditional phase vocoder and analyze sufficiently long periods of sound to characterize sub-audio modulation, we analyze *each spectral channel* of a PV analysis with a single long DFT (LTDFT). In each of these "2LTDFT" spectra the sub-audio modulation can typically be seen as a prominent peak in the 4 to 8 Hz range. As will be seen, this method allows us to modify or remove the sub-audio modulation by manipulating this 2LTDFT peak. Our method operates by (i) removing the energy in 2LTDFT bins, (ii) PV time-stretching the unmodulated result, (iii) performing another 2LTDFT analysis of the lengthened sound, (iv) and imposing the shape of the original sub-audio energy on the 2LTDFT spectra. The analysis/synthesis system is now presented in detail along with results from differing sound types.

4.1. Removal of Sub-Audio Modulation

This section will present our method of sub-audio modulation removal as the first step in a three-part process of removing the sub-audio modulation, time-stretching the sound with the modulation removed, and adding the modulation back into this time-stretched version of the sound. As noted above, the removal and re-introduction of sub-audio modulation is performed in a second order spectral domain. This domain allows us to model and manipulate the sub-audio modulation of an audio signal. Fig 1 shows all the components of the second order analysis/synthesis system.

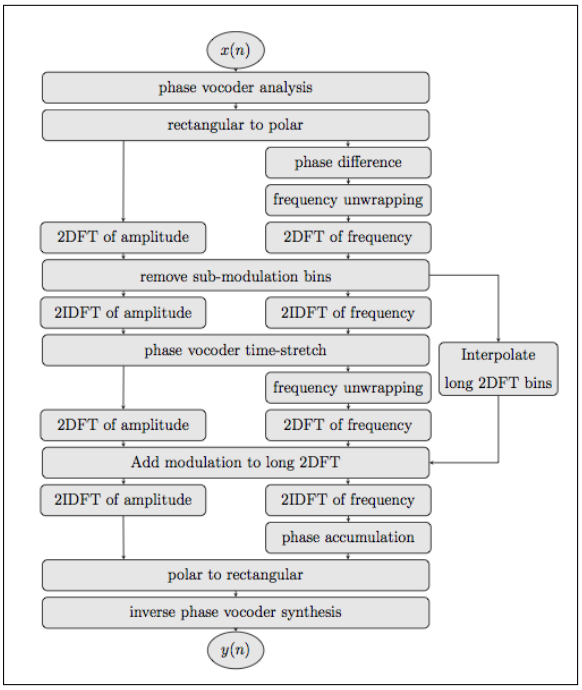

Figure 1. sub-audio analysis/synthesis method.

The first step in our removal algorithm is a phase vocoder analysis. Windowed and overlapped FFT frames are calculated, the complex rectangular spectra are converted to amplitude and phase spectra, and the instantaneous frequency is calculated for each spectral bin with the phase difference method. We call the amplitude spectrum $|X_n(k)|$, and the instantaneous frequency spectrum $\Delta\theta_n(k)$ for time frame n and bin number k.

Typically, higher order spectral analysis considers the spectral frame as the indivisible object of further analysis. For example, cepstral analysis performs a second DFT on the Spectral frame after calculating the log magnitude of the spectrum [7]. Instead we will consider the *time-evolution* of each spectral channel as our indivisable objects for further analysis, not the spectral frame. In order to emphasize that the frame number n is the independent variable, the k amplitude channels are denominated $|X_k(n)|$, and the instantaneous frequency channels are notated $\Delta\theta_k(n)$. We now analyze each of these amplitude and frequency channels using a single FFT for each channel. Each second order spectrum of the amplitude channel is defined as,

$$\hat{X}_a(k,m) = \text{DFT}\big(|X_k(n)|\big), \tag{3}$$

where $\hat{X}_a(k,n)$ is the complex valued spectrum of each amplitude channel k and frequency bins m, and the hat symbol is borrowed from cepstral processing to denote a higher order spectra. Hereafter, we will call these spectra the "2LTDFT amplitude spectra" following Schumacher's nomenclature for defining the second order spectrum of

amplitude channels. It is worth noting here that these 2LTDFT amplitude spectra are complex valued and can be converted to their corresponding amplitude, $|\hat{X}_a(k,m)|$, and phase, $\Delta\hat{\theta}_a(k,m)$, spectra if necessary. It is also worth noting that it is relatively easy to confuse the amplitude spectra of a frequency channel with the phase spectra of an amplitude channel or other combination of first and second order spectra.

As the arctangent function necessary to calculate phase from the complex spectrum produces principal values that are bounded by π and $-\pi$, the instantaneous frequency calculations are also bounded to this range. In order to calculate the corresponding second-order spectra of the frequency channels, the instantaneous frequencies must be transformed into a continuous function by unwrapping since discontinuities in the instantaneous frequency channels would be interpreted as false periodicities or noise in our 2LTDFT frequency analysis. After frequency unwrapping, each second order spectrum of the frequency channel is defined as,

$$\hat{X}_f(k,m) = \mathrm{DFT}\big(\Delta\theta_f(k,n)\big), \tag{4}$$

where $\hat{X}_f(k,m)$ is similarly the complex valued spectrum of each frequency channel, defined here as the 2LTDFT frequency spectrum.

The center frequency of each of the 2LTDFT bins can be calculated as

$$f_k = \frac{SR}{(H)(N)} \tag{5}$$

where SR is the sampling rate of the sampled sound, H is the hop overlap of the first-order window and N is the number of of frames in the first-order spectra. As an example, Figure 2 shows a time-domain signal of a saxophone note with prominent tremolo and vibrato. Visible in the 3 second sample is the slow amplitude variation at approximately 4 Hz.

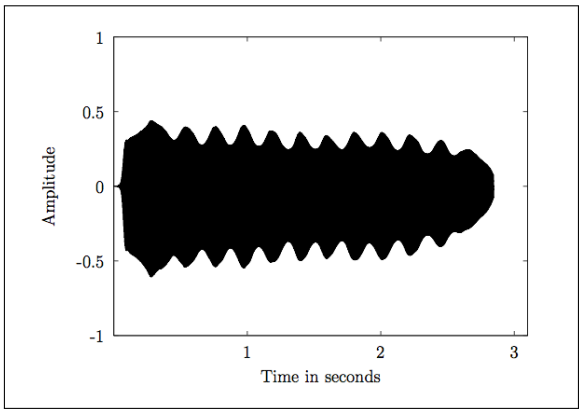

Figure 2. Saxophone sound showing amplitude modulation.

Figure 3 shows one amplitude channel (channel 10) after the phase vocoder analysis of this same saxophone sound, and Figure 4 shows the corresponding unwrapped instantaneous frequency channel. These two figures clearly show the sub-audio modulation that will be modeled by the 2LTDFT analysis.

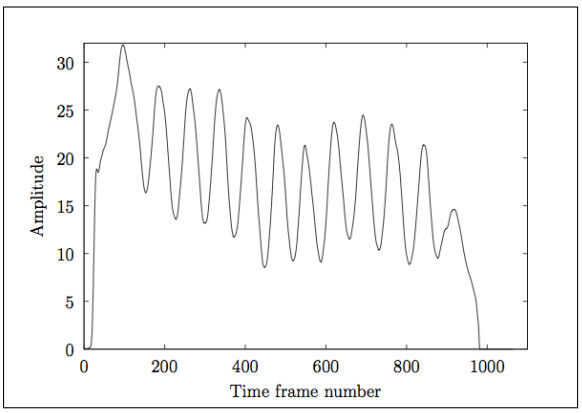

Figure 3. A single amplitude channel (channel 10) of the phase vocoder analysis of saxophone sound showing amplitude modulation.

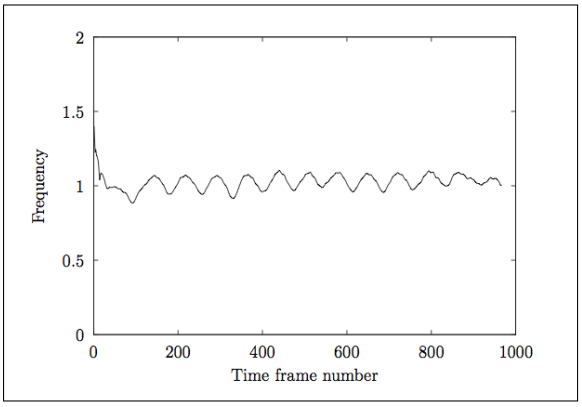

Figure 4. A single frequency channel (channel 10) of the phase vocoder analysis of saxophone sound showing frequency modulation.

Figure 5 shows the 2LTDFT amplitude spectrum of the amplitude channel. As can be seen in these figures a prominent spectral peak at approximately bin number 16 is visible. The center frequency for this bin is approximately 4 Hz. This spectral shape consisting of a large DC component and a prominent spectral peak around the vibrato and tremolo rate is characteristic of many musical instruments which feature this low frequency modulation.

We can now remove this low frequency modulation by "spectrally filtering" each of the 2LTDFT frequency and amplitude spectra. Spectrally filtering simply means changing the amplitude of the complex spectral bins. Our method consists of multiplying each 2LTDFT spectrum by an inverted Hamming window centered at the 2LTDFT bin with the maximum sub-audio frequency component, i.e. the peak of the sub-audio modulation in the 2DFT spectra. Figure 6 shows the same amplitude of an 2LTDFT amplitude spectra after an inverted Hamming window centered on the modulation peak is multiplied by the spectrum. As can be seen, the energy in the peak bins is re-

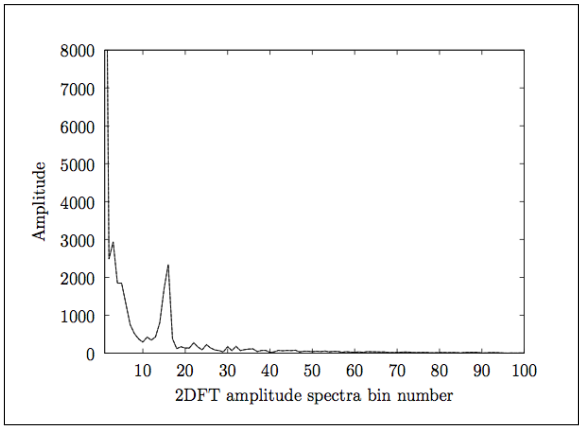

Figure 5. Amplitude of 2DFT amplitude spectrum of a single amplitude channel.

duced to zero.

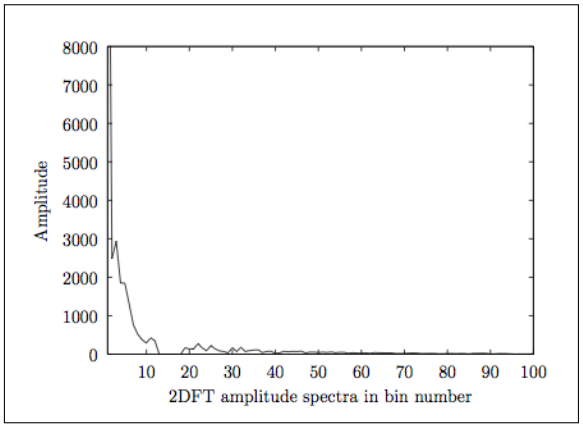

Figure 6. Amplitude of 2LTDFT amplitude spectrum of a single amplitude channel after removal of modulation.

This method smoothes the transitions to the adjacent bins of the 2LTDFT spectra. The original unmodified 2LTDFT bins are stored separately for use in re-introducing modulation after time-stretching.

The set of 2LTDFT amplitude spectra and 2LTDFT frequency spectra are then each converted back to the first-order spectral representation by the inverse FFT. Again, we note that this is a single global operation over the entire length of the spectral channels. No windowing or overlapping is performed to transform to or from the 2LTDFT domain. These new amplitude and frequency channels are now in the standard phase vocoder representation of amplitude and frequency channels which can be transformed back to a time-domain sound by means of phase accumulation, inverse FFT, windowing, and overlapping. When these steps are taken, the resultant sound has the sub-audio tremolo and vibrato removed. Here, we will not convert back to a time-domain sound, but rather continue working with the new sound in PV form.

4.2. De-modulated Lengthening

Next, a traditional phase vocoder time-stretch of this now "de-modulated" sound is performed. As the sound is already in a first-order phase vocoder representation, the initial steps of spectral analysis and instantaneous frequency calculation are already preformed. All that is necessary is to calculate the new phase frames by phase advancement and calculate the amplitude frames by interpolation. Stretch factors or 2, 4, and 8 are used for computational efficiency in our examples.

4.3. Re-introducing Modulation

The final step of our method is to re-introduce sub-audio modulation at the original rate to our de-modulated and lengthened phase vocoder channels $Y_k(n)$. As above, each amplitude and frequency channel is transformed into the second-order 2LTDFT format by taking the FFT of each channel across time. Again, the instantaneous frequency values are unwrapped in time to make a continuous function before the FFT is performed. These new 2LTDFT spectra will be longer in proportion to the amount of time-stretching. Consequently, the original 2LTDFT sub-modulation bins cannot be directly combined into our new 2LTDFT spectra.

In order to add the original modulation rate to the 2LTDFT spectra, the original 2LTDFT bins representing the sub-audio modulation are interpolated to fit the corresponding spectral length in the new longer 2LTDFT spectra. This interpolated peak is substituted for the existing data in the new range of bins. Figure 7 shows these bins interpolated over the new longer 2LTDFT length and substituted for the original 2LTDFT bins in the modulation range.

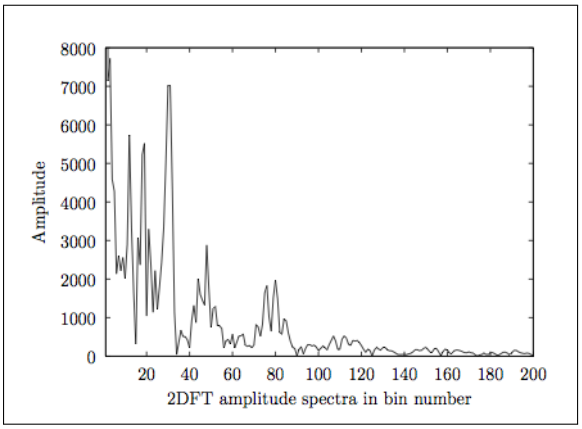

Figure 7. Amplitude of 2DFT amplitude spectrum of a single amplitude channel with addition of interpolated modulation bins.

After substituting the new interpolated 2LTDFT values, the inverse FFT can be performed on each spectral channel in order to return to the first-order phase vocoder representation of sound. Phase accumulation followed by the standard phase vocoder inversion is employed to generate the new sound. This sound exhibits the long term development, pitch and timbre of the orignal sound slowed

by the new factor but with the original tremolo and vibrato rates maintained. Figure 8 a shows the final 3 second saxophone sound stretched to 6 seconds, where the original 4 Hz amplitude modulation rate is evident from a visual inspection.

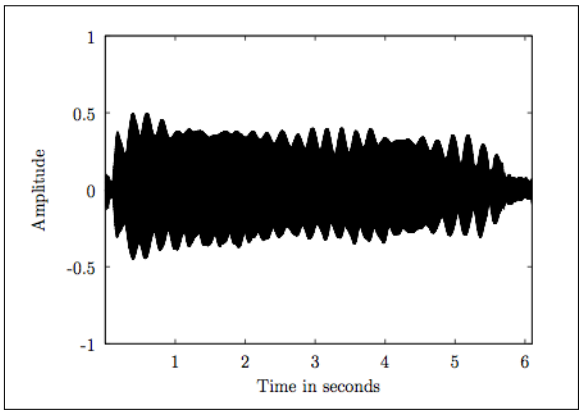

Figure 8. Time-stretched Saxophone sound showing original amplitude and frequency modulation.

5. CONCLUSIONS

We have seen that a method based on a higher order analysis of phase vocoder channel modulations can be used to alter the vibrato and tremolo characteristics of a sampled monophonic musical instrument sound. Our sample sounds show that the method is able to retain the tremolo and vibrato rates of musical sounds with fairly stable vibrato and tremolo rates. The method achieves better results on sounds which have prominent tremolo and less well with sounds with prominent vibrato. This may be attributable to our ability to more easily perceive inappropriate pitch changes than amplitude changes as discussed in the examples section above. The method was unsuccessful at manipulating vibrato and tremolo on human singing tones due to the added complication of formant frequencies present in the human voice.

The current version of the software precludes testing the algorithm on significantly longer sounds than the examples presented due to the computational complexity. Future versions of the software could be developed for more efficient computer languages and machines which would allow testing of the techniques efficacy on longer musical phrases.

A related line of research might be in devising other sonic transformations that are effectively carried out in the second order sound representation presented here.

6. REFERENCES

[1] D. Arfib and N. Delpart, "Selectrive transformations of sounds using time-frequency representations: An application to the vibrato modification," in *104th Convention of the Audio Engineering Society*, 1998, pp. 1–7.

[2] ——, "Alteration of the vibrato of a recorded voice," in *International Computer Music Conference*, 1999, pp. 186–189.

[3] Ø. Hammer and H. Sundt, "Musical applications of decomposition with global support," in *Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99)*, 1999.

[4] P. Herrera and J. Bonada, "Vibrato extraction and parameterization in the spectral modeling synthesis framework," *Proceedings of the Digital Audio Effects Workshop*, Dec 1998. [Online]. Available: http://www.iua.upf.es/mtg/publications/dafx98-perfe.pdf

[5] R. Maher and J. Beauchamp, "An investigation of vocal vibrato for synthesis," *Applied Acoustics*, vol. 30, pp. 219–245, 1990.

[6] S. Marchand and M. Raspaud, "Enhanced time-stretching using order-2 sinusoidal modeling," in *Proceedings of the International Conference on Digital Audio Effects*, 2004.

[7] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Prentice-Hall, Inc., 1998.

[8] T. F. Quatieri and R. J. McAulay, "Speech transformations based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 1449–1464, 1986.

[9] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J. L. Collette, "Vibrato: Detection, estimation, extraction, modification," in *Proceedings of the Workshop on Digital Audio Effects (DAFx-99)*, 1999.

[10] R. T. Schumacher, "Analysis of aperiodicities in nearly periodic waveforms," *The Journal of the Acoustical Society of America*, vol. 91, no. 1, pp. 438–451, 1992.

[11] C. E. Seashore, *Psychology of Music*. Dover Publications, Inc., 1967.

[12] X. Serra, "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Ph.D. dissertation, Stanford University, 1989.

[13] R. Timmers and P. Desain, "Vibrato: Questions and answers from musicians and science." in *Procdeedings International Conference on Music Perception and Cognition*, 2000.