Audio Engineering Society

Convention Paper 9668

Presented at the 141st Convention
2016 September 29 – October 2, Los Angeles, CA, USA

Preserving Reverberation in a Sinusoidally Modeled Pitch Shifter

Sarah R. Smith[1] and Mark F. Bocko[1]

[1] *University of Rochester, Department of Electrical and Computer Engineering*

Correspondence should be addressed to Sarah Smith (sarahsmith@rochester.edu)

ABSTRACT

Many pitch shifting algorithms suffer when the signal contains reverberation. In general, it is possible to preserve the spectral envelope of the original sound, however, an appropriate phase response can only be estimated for minimum phase systems such as vocal formants. This paper presents a pitch shifting algorithm that preserves the reverberant qualities of the original signal by modifying the instantaneous amplitude and frequency trajectories of a sinusoidal model. For each overtone, the sinusoidal trajectories are decomposed into correlated and uncorrelated components and a deviation spectrum is calculated. To synthesize the modified sound, the uncorrelated components are adjusted to preserve the deviation spectrum. The resulting trajectories and sounds are then compared with those of a standard pitch shifter.

1 Introduction

There are many applications where it is necessary or desirable to shift the pitch of a recorded sound. Although many methods exist for performing this process, many of them suffer when the original signal contains significant reverberation or a large change in pitch is required. In general, when the recording consists of a source signal that has been filtered by a resonant system, such as a room, the pitch shifter can also shift the filter resonances. When the required shifts are small, correcting only minor pitch errors a musical performance, this effect may be negligible. However, in applications such gaming or synthesizers with limited memory resources, large changes in the pitch of a sample may be required. In such cases, timbral changes often accompany the pitch shift if the signal contains significant reverberant artifacts from resonant room modes. In the case of voice manipulation, a large increase in pitch can give the impression of a smaller vocal tract and is commonly used for intentional effect.

Ideally, the reverberation could be removed from the input signal, and the pitch shift applied to the dry excitation before reapplying the reverberation to the output signal. However, the process of identifying and removing the reverberation from a recorded signal is often exceedingly complex. Although many methods exist to estimate the filter magnitude spectrum from the input signal, estimating an appropriate phase response is far more difficult [1]. The problem is further complicated by the fact that many reverberant filter responses are non minimum phase, meaning that a stable and causal inverse does not exist, even if the filter form is known [2].

While room reverberation filters cannot be easily inverted, the human vocal tract can be reasonably described as a minimum phase filter [cite]. In this case, a suitable phase response can be determined from the amplitude response using Hilbert transform relations, and the original excitation can be separated from the filter response. [3]. As such, methods that decompose the input signal into excitation and filter response have been used in speech processing in order to preserve the vocal tract formants while modifying the glottal excitation [1]. These methods work well for their intended application, but cannot be extended to compensate for reverberation due to the issues previously mentioned.

Without a complete description of the reverberant filter, the best alternative is generally to equalize the output sound to preserve the original amplitude spectrum. Since the human ear is very sensitive to the locations of spectral peaks, This relatively simple adjustment can greatly improve the resulting tone quality of a pitch shifter. Although this compensates for the magnitude response of the reverberant filter, it cannot compensate for the effects associated with the filter's phase response. As such, the proposed system focuses instead on the room's effect on the instantaneous amplitude and frequency trajectories in a sinusoidal model of the reverberant signal. For harmonic sounds, such as musical instruments and voiced speech, the presence of reverberation introduces deviations into the amplitude and frequency time-trajectories [4] [5]. By calculating the deviation of each frequency track from an expected, correlated, frequency trajectory, it is possible to create a frequency deviation spectrum for the tone. This deviation spectrum can then be used to determine the appropriate amount of deviation to introduce into each overtone of the reconstructed, pitch shifted signal. A similar process is used in order to generate the output amplitudes for each overtone while preserving the original spectral envelope.

This paper is organized in five remaining sections. Section 2 defines the signal model used in the paper and introduces the general architecture of a sinusoidally modeled pitch shifter. Section 3 details the effects of reverberation on frequency modulated harmonic tones, with particular attention to the changes in the extracted amplitudes and frequencies in a sinusoidal model. Section 4 describes the proposed algorithm for modifying the sinusoidal parameter tracks to preserve the deviations discussed in section 3. Section 5 presents the results of the proposed system when it is applied to samples of instrumental vibrato in different reverberant conditions. Finally, section 6 summarizes the methods described in this paper and discusses possible extensions of the algorithm.

2 Pitch shifting using a sinusoidal model

Many naturally generated sounds consist of a fundamental frequency component and a series of nearly harmonic overtones. These overtones are created when multiple resonant modes of the system generating the sound are excited simultaneously, as is true of most musical instruments and the human voice [6] [7]. However, these systems are often modulated in semantically important ways. Amplitude and frequency fluctuations in speech convey important information about the context or emotion of the speaker and musicians will often introduce slight modulations into their tone for expressive effect.

Mathematically, this type of signal can be represented in the form of Eq 1. Each overtone in the sound is modeled as a single sinusoid that may be independently modulated in both amplitude and phase. In this form, f_0 is the fundamental frequency of the sound in Hz, and $a_n(t)$ and $\phi_n(t)$ represent the amplitude and phase modulations of the n^{th} overtone respectively.

$$x(t) = \sum_{n=1}^{n=\infty} a_n(t)\cos(2\pi n f_0 t + \phi_n(t)) \tag{1}$$

Although mathematically equivalent, it is often more convenient to represent the phase modulation in the form of a frequency modulation as shown in Eq 2, with the instantaneous frequency, $f_n(t)$, is defined as the time derivative of the instantaneous phase as in Eq 3. Although the first term of the instantaneous frequency expression is governed by harmonic relation to a fundamental, the model itself does not assume harmonicity since any deviation from harmonic spacing can be incorporated into $\phi_n(t)$ as needed.

$$x(t) = \sum_{n=1}^{n=\infty} a_n(t)\cos(2\pi \int_{-\infty}^{t} f_n(\tau)d\tau) \tag{2}$$

$$f_n(t) = nf_0 + \frac{1}{2\pi}\frac{d}{dt}\phi_n(t) \tag{3}$$

2.1 Analysis/synthesis architecture

For quasi harmonic sounds described by the above model, it is possible to make a variety of modifications to the sound by appropriately adjusting the constituent parameter tracks. In this system, an input sound is first analyzed and the instantaneous amplitude and frequency tracks are estimated for each overtone. The number of overtones required to accurately reproduce the sound varies depending on the sound source and the fidelity required. Although the number of overtones required varies depending on the nature of the sound, many tonal musical sounds are well represented with 20-25 overtones. The implementation described in later sections uses a value of N = 25.

Depending on the type of sound and computational resources, a number of different analysis methods can be used to estimate the input trajectories. If the frequency modulation is small compared to the spacing between overtones, the overtones can be isolated using a bandpass filter and the instantaneous amplitudes and frequencies calculated directly from the analytic signal [8]. In order to analyze a broader range of signals with larger modulations, it is common to estimate the parameter trajectories by tracking the peaks in a short time Fourier transform (STFT) [9] [10]. In this case, the amplitude and frequency tracks can be jointly estimated, without needing to isolate the overtones first. Although both methods produce essentially the same result for a clean signal, the figures and examples in this paper were generated using a STFT based analysis.

Regardless of the chosen analysis method, the pitch of a tone can be shifted by scaling the instantaneous frequency trajectories by an appropriate amount and equalizing the amplitude tracks to preserve the original spectral envelope. Depending on the hop size used in the analysis stage, the parameter tracks can then be interpolated to the signal sample rate if needed and used to resynthesize a new tone using the formula of Eq 1.

3 Effect of reverberation on harmonic tones

The signal model and parameter extraction methods discussed in the previous section assume that each channel contains a single modulated sinusoid. Although this model approximates many types of audio sources, the presence of reverberation in a signal can introduce deviations in the extracted parameter tracks. As a simple example, consider the case of a single sinusoid modulated in both amplitude and frequency along with a single echo. If the source signal changes in frequency during the time it takes for the echo to arrive, the resulting sound now contains two sinusoidal components that must be modeled with a single amplitude and frequency trajectory, as illustrated in fig 1. In general, these frequency changes will be small, meaning that the direct and reverberant components cannot be resolved in a short time Fourier transform analysis, or isolated prior to using an analytic signal approach. As such, the calculated instantaneous amplitude and frequency parameters for the input sound depend on both the source signal and the reverberation that is present.

3.1 Example trajectories

The effects described above can be observed by analyzing and resynthesizing a set of reverberant sounds with no modifications. Fig 2 compares the instantaneous frequencies of a trumpet tone recorded in an anechoic chamber with those drawn from the same sound after it was convolved with a recorded impulse response. Both the initial tone and the impulse response were drawn from available on-line databases [11] [12]. In this example, the reverberation had a T60 decay time of just over 2 seconds. Notice that the overtone trajectories for the anechoic sound closely follow the expected harmonic pattern (indicated with a dotted line in fig 2), while those from the reverberant sound deviate in a significant manner. In particular, the range of estimated frequencies for the 1st overtone is reduced in the reverberant version. More significantly, the reverberation introduces additional large ripples into the frequency trajectory of the 7th overtone.

Furthermore, when both of these sounds are reconstructed from the estimated parameter trajectories, the reverberation in the second tone is clearly audible. However, if the amplitude and frequency deviations are removed and their values are replaced with the average changes in time (denoted a_0 and f_0 in the following sections), the reverberant quality is noticeably absent. This result demonstrates that these deviations encode information about the reverberation and must be considered when making transformations to the original sound.

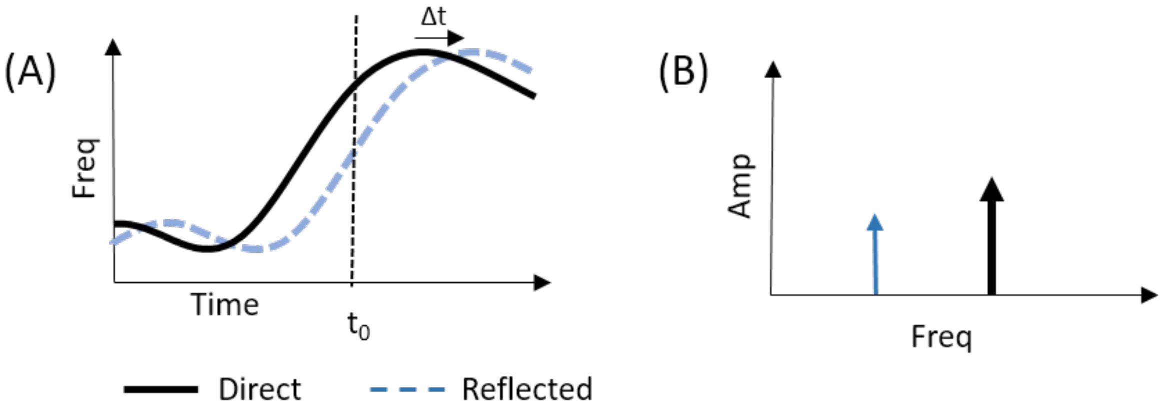

Fig. 1: (A) The instantaneous frequency vs time of a sound with slowly varying frequency is shown for both the direct signal (solid line) and a delayed reflection (dotted line). (B). The spectrum of the signal at time t_0, where the direct and reverberant sounds are at different frequencies. These two components can rarely be resolved during analysis.

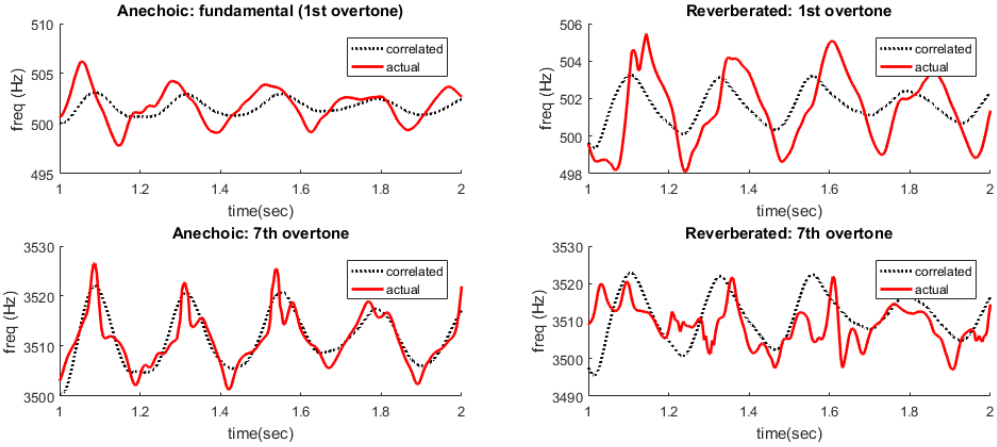

Fig. 2: Comparison of frequency tracks from an anechoic trumpet sound (left), with those from the same sound after reverberation was applied (right). The overtones of the anechoic sound closely follow the correlated trajectory (dotted lines), however significant deviations are introduced after reverberation is added.

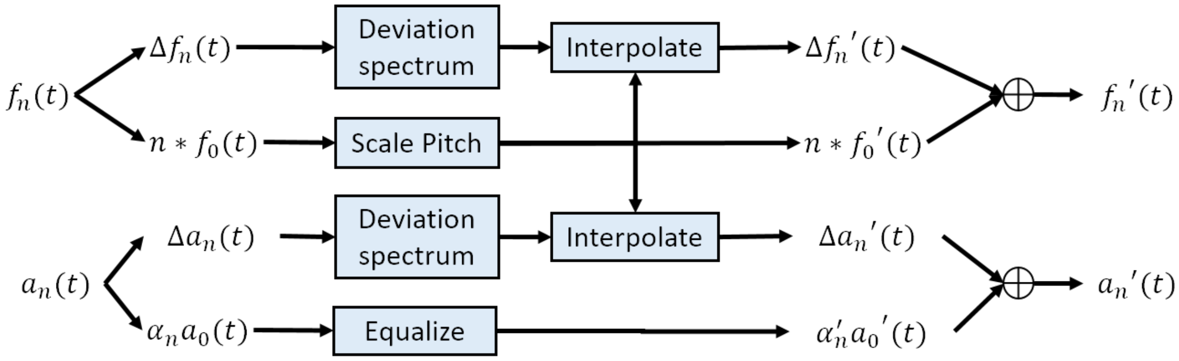

Fig. 3: An overview of the proposed pitch shifting algorithm. The input frequency and amplitude trajectories for each overtone are decomposed into correlated and uncorrelated (or uncorrelated) components. The deviation of the uncorrelated components is then used to calculate deviation spectra for both the amplitudes and frequencies. This deviation spectra is then interpolated and the appropriate amount of decorrelation is added to the output parameter trajectories.

4 System design and overview

A diagram of the proposed pitch shifting algorithm is shown in Fig 3. In a sinusoidal modelling system, any desired modification is performed by modifying the model parameters, the instantaneous amplitudes and frequencies of each overtone. Here, the original amplitudes and frequencies obtained from the sinusoidal analysis are denoted as $a_n(t)$ and $f_n(t)$ respectively. The goal of a pitch shifter is then to generate a set of output parameter tracks $a_n'(t)$ and $f_n'(t)$, such that the pitch is transposed without affecting the timbral characteristics of the original sound. However, when the input sound is reverberant, the initial parameter tracks contain information about both the original source, which we want to shift in frequency, and the reverberant system, which should remain preserved.

As described above and discussed in [4], the presence of reverberation can introduce deviations into the overtone tracks of a harmonic signal. These deviations are generally frequency dependent and therefore affect each overtone differently. The proposed pitch shifting algorithm takes advantage of this property, and the fact that many natural sound sources consist of correlated overtones, by decomposing the amplitude and frequency tracks for each overtone into correlated and uncorrelated components. In this model, the correlated components are taken to represent the source signal while the deviations are considered salient artifacts of the reverberation. In this way, the correlated components can be scaled in frequency and combined with an appropriately generated deviation tracks, as described below.

4.1 Modification of the instantaneous frequencies

Once the frequency tracks $f_n(t)$ have been calculated for the input signal, the algorithm then decomposes these tracks into two parts. First, an estimate of the time varying fundamental pitch $f_0(t)$ is calculated by averaging the normalized frequency tracks of all the overtones, as shown in eq 4. Alternatively, the $f_0(t)$ trajectory could be estimated using any number of available pitch trackers [13]. For each overtone, a relative deviation track $\Delta f_n(t)$ can then be defined as the difference between the actual trajectory and a scaled version of the $f_0(t)$, normalized to the average frequency of the overtone $(\overline{f_n})$ as in eq 5.

$$f_0(t) = \frac{1}{N} \sum_{n=1}^{N} \frac{f_n(t)}{n} \tag{4}$$

$$\Delta f_n(t) = \frac{f_n(t) - n * f_0(t)}{\overline{f_n}} \tag{5}$$

In order to shift the pitch of the sound, the fundamental frequency track is scaled by the desired amount to generate an output fundamental track, $f_0'(t)$. However, generating the output deviation tracks $\Delta f_n'(t)$ is more complex. Unlike the fundamental frequency track, the deviation characteristics should remain in their original frequency locations. For example, when the input signal is shifted up by an octave ($f_0'(t) = 2f_0(t)$), the first overtone of the output signal is at the same frequency as the second overtone of the input signal, and the deviation $\Delta f_1'(t)$ should be equal $\Delta f_2(t)$, the deviation from the second overtone of the input.

However, if the frequency is not being shifted by an even multiple such as an octave, and appropriate estimate of the deviation at the new set of overtone frequencies must be found. In order to generate the output deviation track, two pieces of information are necessary: the extent of deviation expected at the new overtone frequency and the evolution of these deviations in time. Both of these quantities can be found based on a interpolation from the nearby frequency regions of the original tone. Estimating the extent of deviation is accomplished by generating a deviation spectrum based on the original tone. The deviation spectrum is calculated as the standard deviation of $\Delta f_n(t)$ as a function of the overtone frequency $\sigma(\Delta f_n(t))$. This collection of points $(\overline{f_n}, \sigma(\Delta f_n(t)))$ can then be interpolated to estimate the extent of deviation expected at the new set of overtone frequencies. When an overtone in the new sound has a frequency outside of the originally analyzed spectrum, the deviation spectrum is assumed to go to zero at DC and half the sampling rate. While this assumption may not be strictly accurate, the generated overtone trajectories will be more correlated than with other assumptions and are less likely to introduce unpleasant artifacts into the sound.

Although the deviation spectrum specifies the amount of deviation that should be added to the output trajectories, it does not describe the time evolution of the deviation process. In the current implementation, the shape of the deviation track is found by linearly interpolating the normalized deviations of the adjacent overtones of the input sound. This tine varying process is then scaled to generate and output deviation track $\Delta f_n'(t)$ of the appropriate width. The final output frequencies $f_n'(t)$ are then calculated as the sum $f_n'(t) = n * f_0'(t) + \Delta f_n'(t)$.

4.2 Modifying the instantaneous amplitudes

The process for determining the instantaneous amplitude trajectories of the pitch shifted sound is very similar to that used for the instantaneous frequencies with a few noticeable differences. Whereas the correlated frequency tracks are shifted in pitch, the amplitude tracks are equalized based on an estimated amplitude spectrum. As before, an estimate of the correlated amplitude evolution is generated and the amplitude deviations are calculated as the difference between the estimated trajectory and a scaled version of the correlated estimate. The standard deviation of the uncorrelated components are computed to form an amplitude deviation spectrum. Similarly, the average amplitude for each overtone is used to determine the overall spectral envelope. In order to generate the output amplitude trajectories, both the spectral envelope and amplitude deviation spectrum are linearly interpolated to the new frequency positions. The output amplitude tracks are then generated as before, by adding the appropriate amount of deviation to the scaled correlated tracks.

5 Results and discussion

To evaluate the proposed method, the results were compared to those of a traditional implementation. Although both methods preserve the spectral envelope of the original sound, only the new method adjusts the parameter deviations. In listening to the resulting samples, there is an audible difference between the two sounds. The proposed method seems to produce a fuller and more natural sounding resonance when compared with the alternate implementation.

In addition to the subjective listening criteria, it is also possible to compare the instantaneous frequency tracks that result from the two algorithms. Fig 4 shows the instantaneous frequency trajectories of the first three overtones of an example sound. The analyzed frequency tracks of the original sound are shown in column A. Notice that the third overtone (around 1500 Hz) exhibits a different and asymmetric modulation pattern compared to the first two overtones. Using the traditional method (column B), this pattern remains in the third overtone after the pitch has been shifted, moving it to a higher frequency band, closer to 2250 Hz. However, in the proposed method (column C), this deviation remains in the original frequency band of 1500 Hz, which now corresponds to the second overtone. Finally, the nearly constant frequency profile of the third overtone of the new sound (at 2250 Hz) has been generated based on the deviations contained in the fourth and fifth overtones of the original sound (located at 2000 and 25000 Hz).

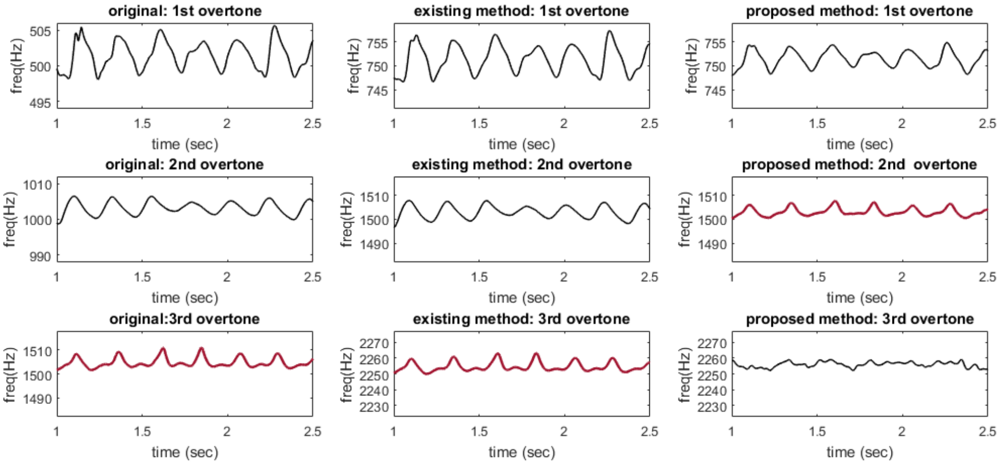

Fig. 4: A comparison of frequency trajectories for the first three overtones of a reverberated sound before and after a change in pitch. Column A is the original sound. Column B was shifted up a fifth using traditional methods. Column C was shifted using the proposed method. The original tone contains atypical frequency deviations in its third overtone. Using traditional methods, this pattern of deviation remains in the third overtone, shifted in pitch. In the proposed method this characteristic deviation now occurs at the original frequency (now the 2nd overtone)

6 Conclusions and future work

This paper describes a method for preserving reverberation when sounds are shifted in pitch. When a harmonic signal is affected by reverberation and analyzed using a sinusoidal model, noticeable deviations appear in the instantaneous amplitude and frequency tracks. By focusing explicitly on these deviations, it is possible to generate an appropriate amount of frequency deviation at the shifted overtone locations. As opposed to estimating the filter response directly from the recorded signal, the current implementation modifies the overtone tracks directly based on the assumption that the underlying modulation is correlated across overtones. Although the present work has focused on the problem of pitch shifting, a similar process could be used to modify the reverberation in these types of signals without changing the pitch of the note itself. Furthermore, there are many extensions and improvements that can be made to the current implementation. Specifically, when data is available from multiple notes, it should be possible to combine information from multiple deviation spectra in order to better determine the appropriate amount of decorrelation in the output tracks.

References

[1] Quatieri, T. and McAulay, R., "Speech Transformations Based on a Sinusoidal Representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(6), pp. 1449–1464, 1986.

[2] Neely, S. and Allen, J., "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, 66(1), pp. 165–169, 1979.

[3] Oppenheim, A. V. and Schafer, R. W., "Discrete-time signal processing," 2010.

[4] Smith, S. R. and Bocko, M. F., "Effect of Reverberation on Overtone Correlations in Speech and Music," in *Audio Engineering Society Convention 139*, Audio Engineering Society, 2015.

[5] Arroabarren, I., Rodet, X., and Carlosena, A., "On the Measurement of the Instantaneous Frequency and Amplitude of Partials in Vocal Vibrato," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 14(4), pp. 1413–1421, 2006.

[6] Sundberg, J., *The science of the singing voice*, Northern Illinois University Press, 1987.

[7] Fletcher, N. H. and Rossing, T., *The physics of musical instruments*, Springer Science & Business Media, 2012.

[8] Cohen, L., *Time-frequency analysis*, volume 299, Prentice hall, 1995.

[9] Serra, X. and Smith, J., “Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition,” *Computer Music Journal*, 14(4), pp. 12–24, 1990.

[10] Quatieri, T. F. and MacAulay, R., “Audio Signal Processing Based on Sinusoidal Analysis/Synthesis,” in M. Kahrs and K. Brandenburg, editors, *Applications of Digital Signal Processing to Audio and Acoustics*, chapter 9, pp. 343–416, Springer US, 2002.

[11] Fritts, L., “University of Iowa musical instrument samples,” *on-line at http://theremin. music. uiowa. edu/MIS. html*, 1997.

[12] Murphy, D. T. and Shelley, S., “Openair: An interactive auralization web resource and database,” in *Audio Engineering Society Convention 129*, Audio Engineering Society, 2010.

[13] De Cheveigné, A. and Kawahara, H., “YIN, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, 111(4), pp. 1917–1930, 2002.