INTRODUCTION

Pitch shifting is a common audio processing tool which can be implemented for a variety of audio applications such as transposition, creating harmony tracks and chorusing to be used in either live or studio applications [1]. When a sound is pitch shifted, its inherent characteristics (formants) are also shifted and the character of the sound changes, deviating from the intention and original sound identity. The aim is to create an effect which maintains the formants of the original signal and imposes this spectral envelope on the pitch shifted signal.

Pitch shifting is the inverse process of time stretching, and both are related to sample rate conversion by the imposed scaling factor which determines the amount of shift [2]. These processes can be accomplished in either the frequency domain or the time domain; pitch shifting in the time domain is called time domain harmonic scaling (TDHS) or pitch synchronous overlap and add (PSOLA), where pitch shifting in the frequency domain is referred to as a phase vocoder [3]. Regardless of the domain, the integral component of pitch shifting is interpolation which is the process of reconstructing the original waveform at desired locations; these may be inbetween the previously sampled data points in order to alter the sample rate conversion to either shift the pitch up or down [2].

The basic process of a phase vocoder begins by time scaling the signal by multiplying the signal by a constant giving the degree of pitch shift. To multiply by less than one indicates the signal is pitch shifted down, conversely multiplying by more than one indicates the signal is being pitch shifted up. This signal is then resampled (which is the interpolation process) by the following equation:

$$x_s(n) = \frac{f_s}{m_c} \times x_t(n)$$

Where \(x_s(n)\) is the resampled signal, \(f_s\) is the original sample rate, \(m_c\) is the multiplying constant, and \(x_t(n)\) is the time scaled signal. This returns the signal length to its original duration, however due to being resampled the signal will have more or less samples than it originally did depending on its being shifted up or down [4].

EFFECT OVERVIEW

The method of pitch shifting this paper proposes is a block-by-block frequency domain based approach using the cepstrum to find the spectral envelope (formants) of the signal. An overview of this process is shown in figure 1. First the signal is transformed into the frequency domain and the spectral envelope of the original signal is computed. The signal is then pitch shifted and the spectral envelope of the transformed signal is computed. The formants of the original signal are kept by using a correction factor applied to the pitch shifted signal which subtracts the ‘new formants’ from pitch shifted signal and applies the original spectral envelope to the pitch shifted signal via FFT convolution which is
multiplication in the frequency domain. Next the signal is interpolated (resampled) and the IFFT calculated transforming the signal back into the time domain.

![Diagram](image)

**Figure 1. Overview of the process of pitch shifting in the frequency domain while keeping the formants of the original signal. The spectral envelope is calculated via the cepstrum before being imposed on the pitch shifted signal.**

**PROCESSES OF EFFECT**

**FFT**

The process of transforming the data from the time domain into the frequency domain is done by the fast Fourier transform (FFT) which is a type of discrete Fourier transform (DFT) calculated with length \( N \) being a power of 2, increasing the computational speed [5].

The FFT is calculated by

\[
X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N}
\]

Where \( x(n) \) is the input signal, \( k \) is the frequency bin number from 0, 1, 2...\( N-1 \), and \( N \) is the total number of samples in the signal. Once the signal is in the frequency domain, its spectral envelope is calculated and the pitch shifted spectral envelope is calculated, both via the cepstrum.

**Spectral envelope calculation**

The spectral envelope is calculated from the cepstrum which is defined as the IFFT of the log magnitude of the Fourier transform [6] for which the process is shown in figure 2. As the FFT of the signal has already been computed, the next step is to calculate the logarithm of the magnitude which is done by

\[
\mathcal{X}(k) = \log X(k)
\]

Then the IFFT is calculated by

\[
x_r(n) = \frac{1}{N} \sum_{k=0}^{N-1} \mathcal{X}(k)e^{-j2\pi kn/N}
\]

The real part is taken of the resulting IFFT which is the real cepstrum. This is then multiplied by a lowpass window which is determined by the cut frequency. The cut frequency is considered to be the upper limit of the spectral envelope in samples and should be less than the period of the analysed sound [7]. The lower this value, the smoother the spectral envelope will be showing the overall shape which is desired as this is what the spectral envelope is intended to show. After the cepstrum has been multiplied by the lowpass window, the FFT is computed to give the spectral envelope.
Figure 2. Block diagram of how the spectral envelope is calculated via the cepstrum. The real part of the IFFT of the log magnitude of the FFT of the signal is taken to give the real cepstrum. Next this is windowed by the cut frequency and the real FFT performed to give the spectral envelope of the signal.

This process of calculating the spectral envelope is done for the original signal and for the pitch shifted signal where the spectral envelope of the latter is subtracted from the former by the equation

\[ c(f) = c_1(f) - c_2(f) \]

Where \( c(f) \) is the resulting spectral envelope to be imposed on the signal, \( c_1(f) \) is the spectral envelope of the original signal, and \( c_2(f) \) is the spectral envelope of the pitch shifted signal. The spectral envelope is to be imposed via FFT convolution (which means it is multiplied) with the pitch shifted signal to impose on the signal the original formants without the formants created by the pitch shifted signal.

Figure 3. Spectral envelope obtained from calculating the cepstrum. The over shape is seen for a xylophone sound with the fundamental formant frequency the first major peak in the spectrum.

Pitch shifting

The basic method for pitch shifting via the frequency domain is shown in the block diagram figure 3. The main processes causing the change in pitch is the time stretching ratio and the resampling ratio.
The process of pitch shifting by a phase vocoder begins with transforming the signal into the frequency domain. As previously mentioned, this is done via the FFT. After this, the phase is extracted to be unwrapped to obtain the instantaneous frequency for every frequency bin at a given time instant [8]. This is calculated from the difference between two successive frames by

\[ f_i = \frac{f_s (\varphi_u - \varphi_p)}{2\pi R_a} \]

Where \( f_i \) is the instantaneous frequency, \( f_s \) is the sampling frequency, \( R_a \) is the analysis hop size, \( \varphi_u \) is the unwrapped phase and \( \varphi_p \) is the previous frame’s phase value. The unwrapped phase is calculated by

\[ \varphi_u = \varphi_t + (\varphi_m - \varphi_t) \]

Where \( \varphi_t \) is the target phase and \( \varphi_m \) is the measured phase. The measured phase is the existing phase value and the target phase is the expected unwrapped phase value [8]. The unwrapped phase gives a continuous phase without discontinuities. Usually, the phase value is between 0 and \( 2\pi \) and when it exceeds this value there are discontinuities in the response causing inaccurate calculations. By unwrapping the phase, all values exceeding \( 2\pi \) are included appropriately [9]. An example of a signal with unwrapped phase can be seen in figure 4 below.

Figure 4. Block diagram of pitch shifting algorithm by time stretching to increase or decrease pitch then resampling to return signal to its original duration.

Figure 5. An example of a signal with unwrapped phase. The values lie between \(-\pi\) and \(\pi\) where multiples of \(2\pi\) have been adjusted for. Figure from [9].
After the phase is unwrapped, it is multiplied by the frequency shift ratio which is calculated by

\[ f_{\text{shift}} = \frac{R_s}{R_a} \]

Where \( f_{\text{shift}} \) is the frequency shift ratio and \( R_s \) is the synthesis hop size. If the ratio were to equal two, this would indicate a pitch shift upwards of one octave, and vice versa if it were equal to 0.5 this would indicate a pitch shift downwards of one octave; a ratio equal to one indicates no transformation of the signal took place. If \( R_a < R_s \) the signal is pitch shifted upwards, inversely if \( R_a > R_s \) the signal is pitch shifted downwards. \( R_s \) is to be a divisor of the window length of the FFT to ensure resynthesis of the signal is properly calculated.

The next process is interpolation, which is the process of reconstructing the signal at desired locations to obtain new sample points in between previously sampled data. This process makes the signal its original duration, but sampled at different intervals. This is principally calculated by

\[ f_{\text{resample}} = \frac{R_a}{R_s} \]

The interpolation process involves creating a sinc filter function to identify the new sample points which is defined by

\[ \frac{\sin(x)}{x} \]

This filter function is designed to have one peak to be centred on the sample point to be obtained and zero-crossings at the sampling rate so that only the sample point at the centre of the sinc filter function is calculated and the surrounding sample points are at the zero-crossings, therefore they do not factor in to the equation. This filter is convolved with the signal to obtain the sample points at the necessary intervals. The new sample points are then output at the same sample rate, creating pitch shifting [2].

**Signal reconstruction**

To reconstruct the signal, the magnitude and phase responses are calculated by

\[ \text{synthesized} = m_g \times e^{i\theta} \]

Where \( m_g \) is the magnitude of the signal and \( \theta \) is the phase of the signal. Then the inverse fast Fourier transform (IFFT) is calculated transforming the resulting signal into the time domain.
Figure 6. Magnitude spectrum of xylophone sound, which is the original signal. The fundamental frequency and harmonics are shown in the graph.

Figure 7. Magnitude spectrum of pitch shifted xylophone sound. This graph shows that the fundamental and harmonics have been pitch shifted upwards. This is a pitch shift upwards of an octave (frequency shift ratio = 2), where the formants of the original have not been imposed on this signal.
CONCLUSION

In conclusion, the aim to pitch shift a signal while maintaining the original signal’s formants can be accomplished by pitch shifting in the frequency domain and calculating the spectral envelope via the cepstrum. The cepstrum is found by the IFFT of the log magnitude of the FFT of the signal. From this, the spectral envelope is found by lowpass weighting the cepstrum to obtain the smoothed response revealing the formants. The process of pitch shifting begins with unwrapping the phase of the signal and multiplying by the time stretch ratio which determines the amount of pitch shift. The signal is then interpolated by the resample ratio. The formants of the original signal is then multiplied with the pitch signal, after the latter’s spectral envelope has been removed. Following this, the signal is reconstructed before the IFFT is performed returning the signal to the time domain.

REFERENCES