
National Institute of Technology
Karnataka, Surathkal

20 Oct 2017

# Digital Signal Processing Laboratory

TALKING TOM

*Submitted To:*
Krishnan CMC
Asst. Professor
EE Department

*Submitted By :*
Rupashi Sangal (15EE239)
Vilohit Kaza (15EE151)
Raveesh Sinha (15EE142)
Shreyansh (15EE155)

# 1 Introduction

Talking Tom is a virtual pet app in which a virtual animal squeakily repeats anything that the user says into the device's microphone, while interacting with the animal by tapping and swiping on the screen. The app records the voice of the user through microphone, in digital format, which is then modulated to a certain pitch/tempo and played through the microphone, creating a **Chipmunk-like effect**.

In this project, various possible techniques for voice conversion have been explored. These techniques can be categorized as : **Time Domain and Frequency Domain based Pitch Shifting**.

# 2 Implemented Methods

In the standard time-domain pitch-shifting technique, for pitch-shifting the signal by a factor of $\alpha$, the input signal is first resampled by a resampling factor equal to $1/\alpha$. Since resampling changes the length of the signal, a timescale modification method is used to preserve the time duration of the original signal. The time duration of the resampled signal should be scaled by a factor equal to $\alpha$.

In the **Phase Vocoder** technique, first the signal is converted to its frequency-domain representation using a short-time Fourier transform (STFT). After modification of the frequency-domain parameters according to the pitch-shifting factor, the signal is converted back to its time-domain waveform.

**Pitch Synchronous Overlap and Add Method (PSOLA)** uses the overlap and add method to stretch the sample wherein, the input signal is divided into overlapping segments which are shifted with respect to each other according to the time-scaling factor. After the convolution of signals, we get the final output signal.

The **Delay based method** works in the time domain and crossfades between two channels with different varying delays and gains to produce a smoothly transitioned pitch shifted signal.

Another method for time stretching, **Sinusoidal Spectral Modeling**, relies on a spectral model of the signal. In this method, peaks are identified in frames using the STFT of the signal, and sinusoidal "tracks" are created by connecting peaks in adjacent frames. The tracks are then re-synthesized at a new time scale. This method can yield good results on both polyphonic and percussive material, especially when the signal is separated into sub-bands.

In **Frame Based Approach**, the first step is to split the signal into short analysis frames of fixed length. The analysis frames are spaced by a fixed number of samples, called the analysis hopsize. To achieve the actual time-scale modification, the analysis frames are then temporally relocated to have a synthesis hopsize. This frame relocation results in a modification of the signal's duration by a stretching factor of $\alpha$.

# 3 Our Project

We implemented the following 4 algorithms:

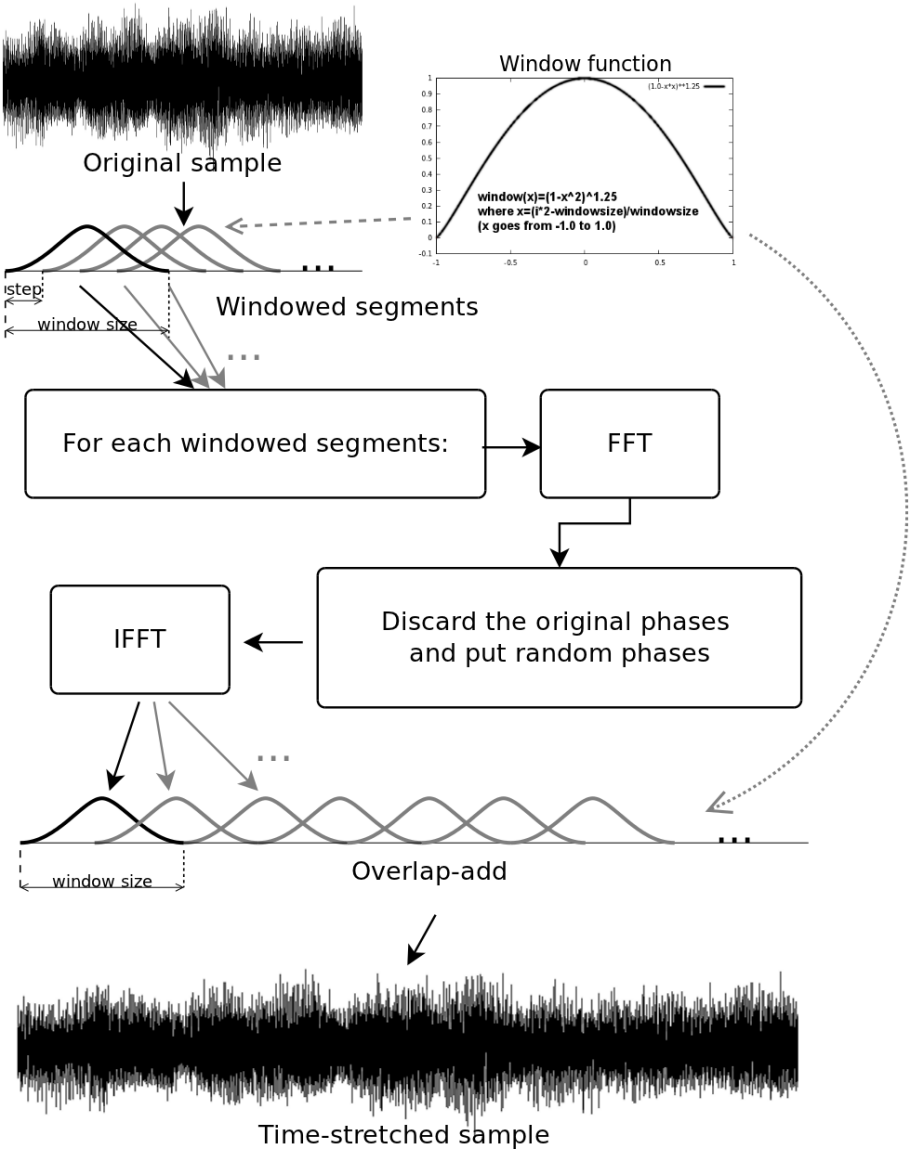
## 3.1 Resampling

The simplest way to change the duration or pitch of a digital audio clip is through sample rate conversion. This is a mathematical operation that effectively rebuilds a continuous waveform from its samples and then samples that waveform again at a different rate. When the new samples are played at the original sampling frequency, the audio clip sounds faster or slower. The frequencies in the sample are always scaled at the same rate as the speed. Hence, slowing down the recording lowers the pitch, speeding it up raises the pitch.

*Drawback*: Using this method, the two effects cannot be separated.

## 3.2 Phase Vocoder

A phase vocoder is a type of vocoder which can scale both the frequency and time domains of audio signals by using phase information. This is one way of stretching the length of a signal without affecting the pitch.




Basic steps:

- The phase vocoder has an analysis section that performs an overlapped short-time FFT (ST-FFT), which is the discrete Fourier transform of a short, overlapping and smoothly windowed block of samples;
- Magnitudes and phases of Fourier Transform are randomized;
- The synthesis section performs an inverse STFT by taking the inverse Fourier transform on each block and adding the resulting waveform segments.

To time stretch a signal, the phase vocoder uses a larger *hop size* for the overlap-add operation in the synthesis section than the analysis section. Here, the hop size is the number of samples processed at one time. As a result, there are more samples at the output than at the input although the frequency content remains the same.

Now, we pitch scale this signal by playing it at a higher sample rate, which produces a signal with the original duration but a higher pitch.

**Pros**

- Best Quality
- Customizable
- Ideal for speech

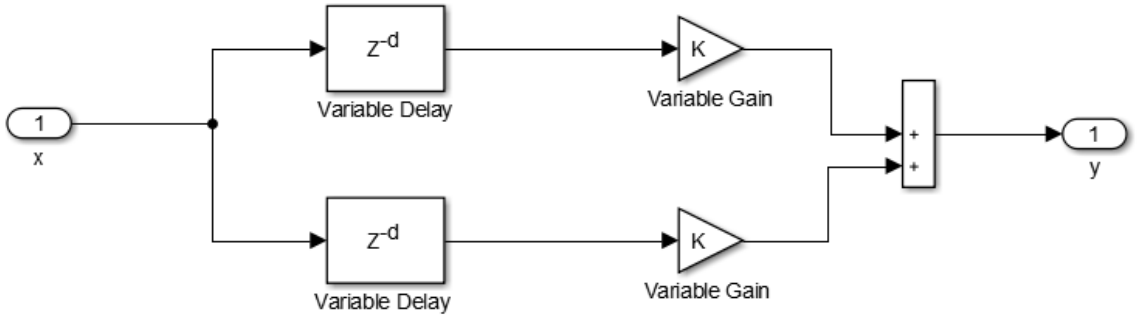
**Cons**

- Lots of Computations
- Slow
- Complex Math

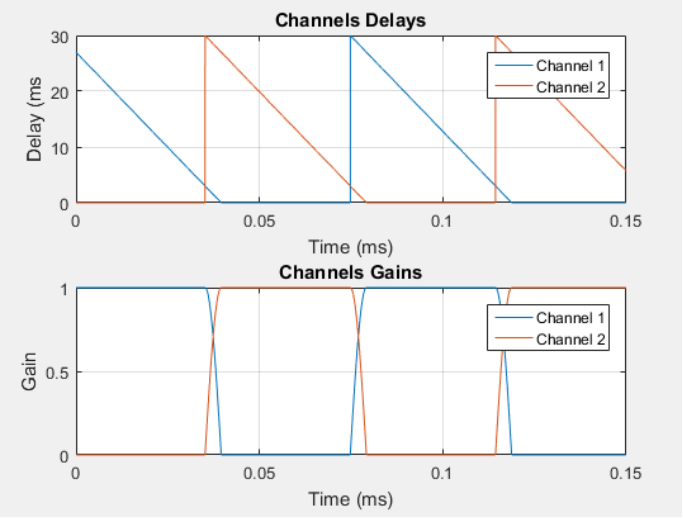
## 3.3 Delay Based Pitch Shifting

The delay based method works in the time domain and crossfades between two channels with different varying delays and gains to produce a smoothly transitioned pitch shifted signal.

To create an upward pitch change, we used a 30-millisecond delay and steadily decrease it at a rate that yields the desired pitch change. As the delay approaches 0, a second delay channel is started at 30 milliseconds and sweeps in a similar manner. A quick crossfade from the first to the second channel is applied, making sure the first channel is completely faded out before its delay reaches 0. This process is repeated, going back and forth between the delay channels.




A downward pitch change is achieved in a similar manner, only the delay channels are started with a near zero initial delay, and the delay is increased out to around 30 milliseconds, at which time the alternate channel is started and the cross-fade performed. The changing delays and cross fading pattern is shown below:




The desired output pitch may be controlled by varying the rate of change of the channel delays. Cross-fading reduces the audible glitches that occur during the transition between channels.

**Pros**

- Fast to process
- No filtering
- Better quality than SOLA

**Cons**

- Unwanted smeared frequencies
- Unintuitive

## 3.4 Changing Pitch with PSOLA for Voice Conversion

*PSOLA* (Pitch-Synchronous Overlap and Add) is a method used to manipulate the pitch of a speech signal to match it to that of the target speaker.

The basic algorithm for the PSOLA technique consists of three steps.

- First, the speech signal is divided into separate but overlapping smaller signals. This is accomplished by windowing segments around each pitch mark or peak amplitude in the original signal. The windowed segments usually contain two to four pitch periods.
- Secondly, the smaller signals are modified by either repeating or leaving out speech segments, depending on whether the pitch of the target speaker is higher or lower than the pitch of the source speaker. This modifies the duration of the signal, therefore changing the fundamental frequency.
- Lastly, the remaining smaller segments are recombined through overlapping and adding. The result is a signal with the same spectrum as the original but with a different fundamental frequency. Thus, the pitch changes, but the other vocal qualities remain the same.

**Pros**

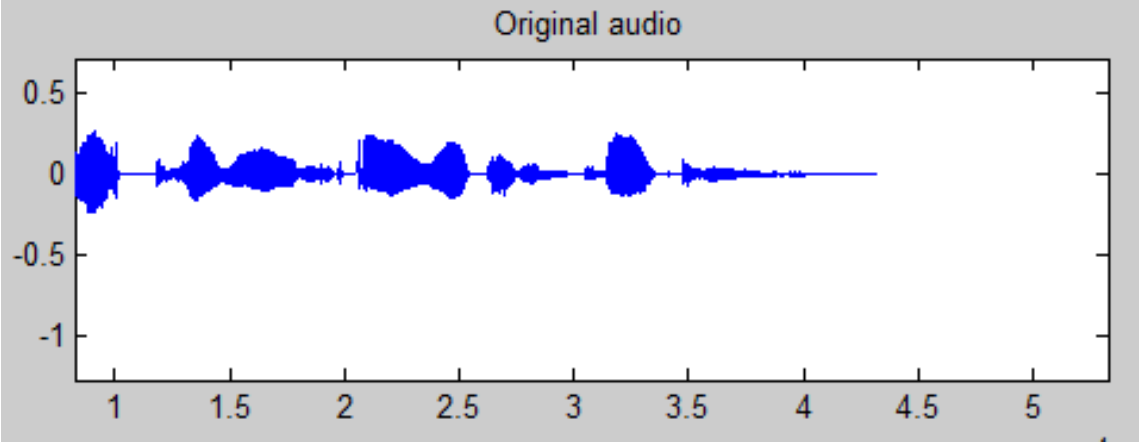
- Fast to Process
- Simple
- Ideal for tones

**Cons**

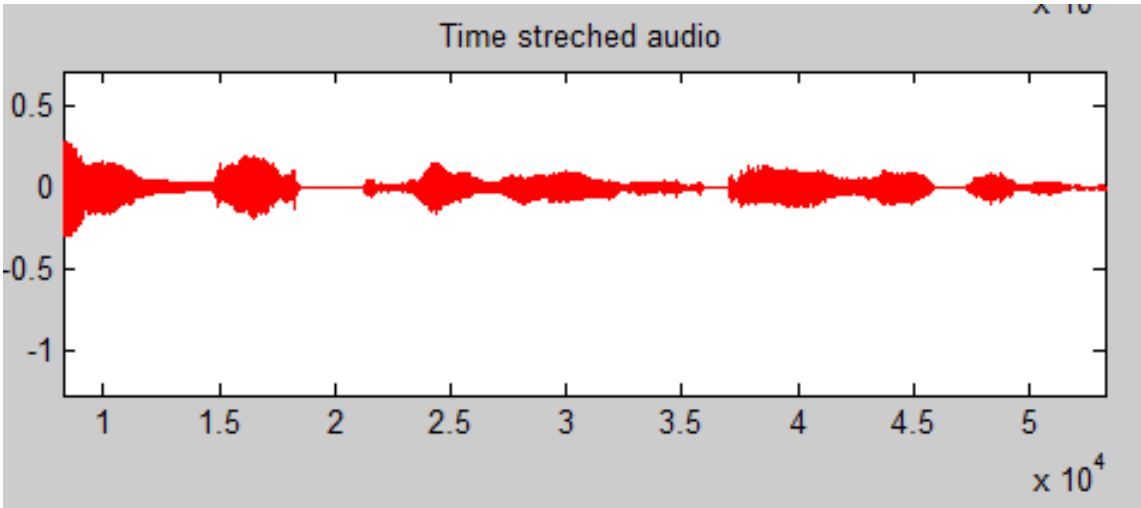
- Reverberation
- Lower Quality
- Resampling Algorithm

# 4 Plots Obtained

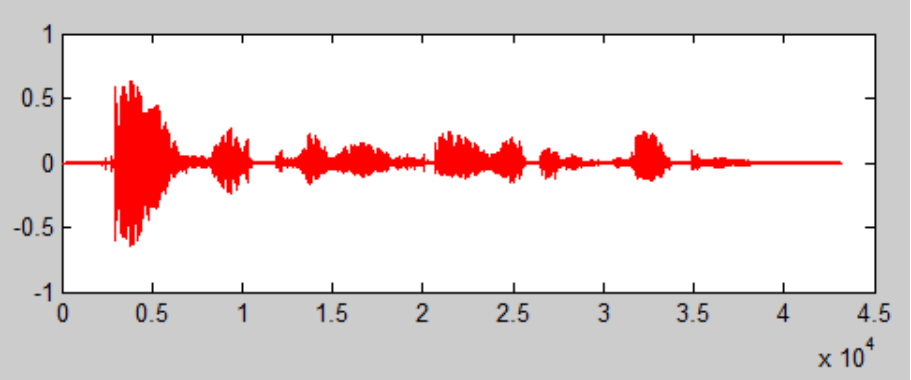
Input audio :



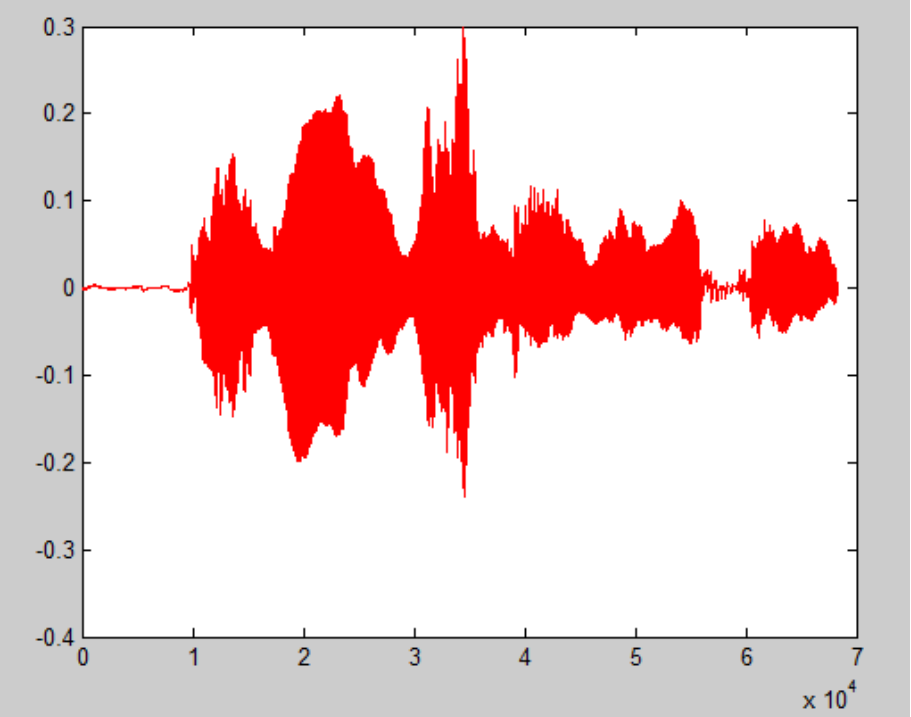

Phase Vocoder :



Delay Based :




TD PSOLA :



# 5 Conclusion

We started off with simply playing the input audio file at different sampling rates and observed that as the pitch increases, the duration of the output audio file was significantly reduced. In order to overcome this problem, we looked into time stretching and pitch shifting algorithms. This led to the implementation of TD-PSOLA technique wherein the output was observed to be a little phasey and diffused but the length of audio was maintained. Also, it seemed to work only for single audio channeled inputs.

On further exploration, delay based pitch shifting algorithm was found to give clearer results than other time domain techniques.

In the frequency domain, the Phase Vocoder technique was executed which yielded better quality results at all hop ratios but a residual smearing effect still remained.

Of all the above discussed algorithms, the **Phase Vocoder** approach was found to be of the *best quality* and ideal for speech.

Hence, in general, time-domain techniques are simple and fast, and work fine for periodic and quasi-periodic signals. However, their quality is not good for signals which contain a lot of non-harmonic components. On the other hand, frequency-domain algorithms are more suitable for complex signals, but the price of the high-quality is the computational complexity.

# 6 References


- Delay-Based Pitch Shifter - MATLAB and Simulink - MathWorks India, in.mathworks.com/help based-pitch-shifter.html.
- Jarvis, M. (2017). PaulStretch: An Interview with Paul Nasca — Mat Jarvis. [online] Microscopics.co.uk. Available at: http://www.microscopics.co.uk/blog/2010/paulstretc an-interview-with-paul-nasca/
- Songar, Ashwini, and Mrs B. Harita. "Matlab based Voice Conversion Model using Psola Algorithm." International Journal of Digital Application and Contemporary research 1.8 (2013).
- Allam. "Voice conversion using pitch shifting algorithm by time stretching with PSOLA and re-sampling." Journal of electrical engineering 61.1 (2010): 57-61.
- Garrison, Jake, and Jisoo Jung. "Pitch Shifter."