Voice Pitch Correction using a Simplified Phase Vocoder

BE 3010 Final Project
Pedram Bayat
December 16, 2025

1 Introduction

Pitch correction systems (e.g. Antares Auto-Tune [2]) are widely utilized in modern audio processing to refine intonation in vocal recordings. At their core, pitch correction systems estimate the fundamental frequency of the input vocals and map them to the nearest note in a target musical scale, stabilizing a singer's pitch for greater accuracy and stylistic effect.

From a signals and systems perspective, vocal recordings are non-stationary signals, meaning their frequency evolves continuously over time. Consequently, a traditional Fourier Transform (FT), which provides a global frequency representation, is insufficient for analysis. Each time-localized segment of the recording must be analyzed individually, which can be accomplished by the Short-Time Fourier Transform (STFT). [1] The STFT segments a time-domain signal (e.g. a vocal recording) into short, overlapping windows, allowing for the local analysis of frequency content at specific time intervals.

In practice, the STFT is most commonly applied through a Phase Vocoder approach, which can modify pitch while maintaining temporal coherence. [3] Specifically, a standard Phase Vocoder preserves smooth frequency evolution over time by observing how phase changes from frame to frame of an STFT and estimating the true instantaneous frequency of each spectral component. Each modified frame is finally transformed back into the time domain and overlap-added to produce the output signal. Correctly managing phase differences preserves temporal coherence and timbral structure of a vocal signal over time, producing an accurate pitch correction. Simplified implementations, such as the one explored in this project, modify the magnitude spectrum while retaining the original phase of the signal, which allows for pitch shifting without complex phase unwrapping but may introduce phase discontinuity artifacts.

The aim of this project is to implement a simplified phase vocoder system in MATLAB, designed to tune a vocal input to the C Major scale.

2 Methods

The pitch correction system was designed to correct the pitches of a C Major scale sung a cappella with intentional errors, such as singing some notes flat (slightly lower frequency) and some notes sharp (slightly higher frequency). This was accomplished with a STFT framework and spectral analysis akin to the analysis done by a Phase Vocoder.

2.1 Signal Analysis

The input audio signal $x[n]$ was first converted to mono audio and segmented into overlapping frames. An STFT window size of $N = 2048$ was selected as a practical trade-off between frequency resolution and time resolution. Digitally recorded audio has a sampling rate of $f_s = 44.1$ kHz, meaning a 2048-sample window spans approximately 46 ms, long enough to assume the signal is locally stationary. The corresponding frequency resolution is 21.5 Hz, which is sufficient to resolve harmonic structure in vocal signals. To reduce spectral leakage, a Hanning window $w[n]$ was also applied to each frame before applying a Fast Fourier Transform (FFT), defined as

$$w[n] = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right), \quad 0 \le n \le N-1$$

To ensure smooth reconstruction of the audio signal, a hop size of $N/4$ (512 samples) resulting in 75% overlap was chosen. At each time frame, the frequency $X[k]$ of each frame was computed using the Fast

Fourier Transform (FFT):

$$X[k] = \sum_{n=0}^{N-1} (x[n] \cdot w[n]) e^{-j\frac{2\pi kn}{N}}$$

where k represents the frequency bin index.

2.2 Pitch Detection & Modification Pipeline

The pitch correction logic is carried out through three steps, primarily relying on identifying the fundamental frequency (f_0) of the input signal of a specific frame and shifting it to a target grid.

1. **Peak Detection**: The algorithm identifies the index k_{max} corresponding to the maximum magnitude in the spectrum. The current pitch $f_{current}$ is calculated as:

$$f_{current} = \frac{k_{max} \cdot f_s}{N}$$

2. **Quantization**: The system compares $f_{current}$ against a predefined array of target frequencies corresponding to the C Major scale starting at C4 (261.63 Hz, 293.66 Hz, ...). The nearest target frequency f_{target} is selected, and a shift factor α is derived:

$$\alpha = \frac{f_{target}}{f_{current}}$$

3. **Spectral Remapping**: The magnitude spectrum is frequency-warped by mapping the energy from original bin k to a new bin k':

$$k' = \text{round}(k \cdot \alpha)$$

Crucially, the original phase information $\phi[k]$ was preserved and recombined with the shifted magnitude $|X_{new}|$. This simplified phase-vocoder approach avoids complex phase unwrapping but introduces phase incoherence artifacts.

2.3 Noise Reduction Strategy

The correction pipeline aggressively manipulates the input signal, causing unnatural, "robotic" artifacts in the pitch-corrected audio. A two-stage filtering strategy was implemented to reduce these artifacts and additional background noise:

1. **Harmonic Masking**: a "comb filter" mask was generated in the spectral domain. Based on the target frequency f_{target}, the mask preserves frequencies at harmonics of the fundamental frequency and attenuates other frequencies.

$$f_{harmonic} = n \cdot f_{target} \pm \Delta f$$

where $n = \{1, 2, \ldots, 8\}$ and bandwidth $\Delta f = 60$ Hz. This effectively zeros out noise between the harmonic partials. The range of integers represents the fundamental frequency and seven additional harmonics (overtones) of the voice.

2. **Global Band-Pass Filter**: a 4th-order Butterworth band-pass filter with cutoffs at 100 Hz and 3500 Hz was applied to preserve the fundamental frequency and its harmonics while attenuating background noise.

2.4 Signal Reconstruction

The modified frequency spectrum $Y[k]$ is converted back to the time domain using the Inverse FFT:

$$y_{frame}[n] = \text{Re}\left(\frac{1}{N} \sum_{k=0}^{N-1} Y[k] e^{j\frac{2\pi kn}{N}} \right)$$

To smooth discontinuities caused by spectral modifications, a second Hanning window is applied to the output frame before reconstructing the continuous signal via the overlap-add method.

3 Results

The pitch correction system was evaluated with spectral analysis and by inspection of the processed vocal recording. The spectrograms of the input signal and the pitch-corrected output were compared (Fig. 1).

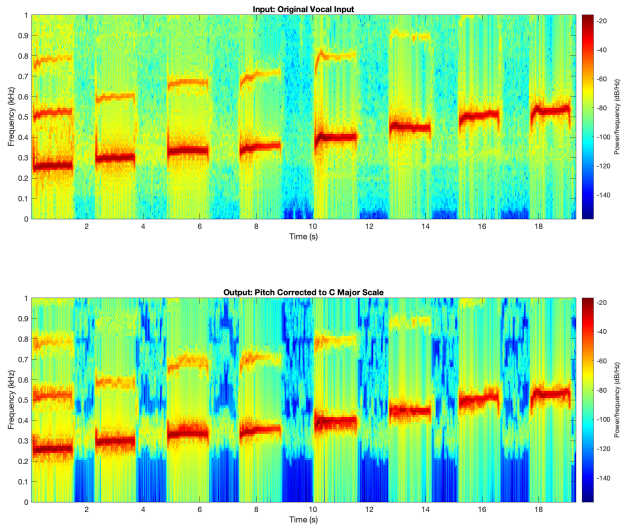

Figure 1: Comparison of Input and Pitch-Corrected Spectrograms. **(Top)** The original vocal input displays the pitch trajectory corresponding to a C Major scale with multiple notes out of tune. **(Bottom)** The pitch-corrected output demonstrates tuning the input to the target C Major scale and signal attenuation between harmonic partials.

The hybrid noise reduction process effectively reduced background noise but left vertical striations across the spectrum, likely corresponding to phase discontinuities at STFT frame boundaries. (Fig. 2)

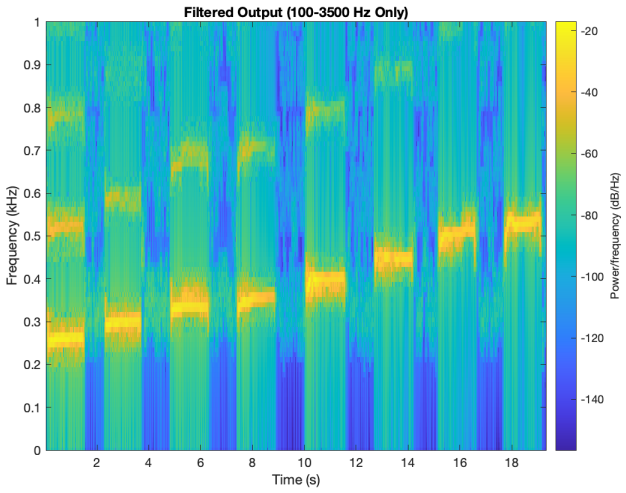

Figure 2: The hybrid noise reduction system indicates the effective suppression of background noise (deep blue regions between harmonic partials) and a global band-pass filter. However, vertical striations (broadband impulses) remain visible across the spectrum, representing the phase incoherence artifacts introduced at the STFT frame boundaries during the pitch-shifting process.

The processed audio signal had noticeably lower quality and clarity than the original signal, with "robotic" artifacts corresponding to phase discontinuities.

4 Conclusion

This project successfully designed and implemented a pitch correction system using a simplified Short-Time Fourier Transform (STFT) with a phase vocoder in MATLAB. The system analyzed a C Major scale sung a cappella with intentional errors by identifying the fundamental frequency of each note and quantizing the pitch to the corresponding target pitch, independent of time stretching.

The system also introduced distortions to the input signal, likely due to misaligned phase propagation between overlapping frames. This was accounted for through harmonic masking and global band-pass filtering. While this hybrid filtering strategy significantly improved the signal-to-noise ratio by removing broadband spectral noise, it could not fully eliminate the phase-induced degradation of the input signal.

To enhance the naturalness of the output in future iterations of a pitch correction system, a standard phase vocoder algorithm can be implemented to explicitly calculate and propagate phase through each frequency frame. This would enforce a continuous phase through frame boundaries and likely reduce the "robotic" striations visible in the output spectrogram. Additionally, a more precise adjustment can be made on the input signal by estimating the entire spectral envelope (e.g. Linear Predictive Coding) and only adjusting the original frequencies by shifting the fine harmonic structure. Finally, data-driven machine learning models could be incorporated to learn pitch trajectories and correction strengths from real vocal performance, which could enable a more adaptive and natural pitch correction.

Overall, this work demonstrates that frequency-domain pitch correction using an STFT-based framework is effective for refining discrete pitch errors and indicates the need for phase-aware or data-driven approaches to achieve more natural vocal processing.

Acknowledgements

I would like to express my sincere gratitude to Allen Yan for his assistance in producing the input data. I am also deeply thankful to Professor Siedlik for his guidance, insight, and continued support over the course of this project and the entire semester.

References

[1] J. Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):235–238, 1977.

[2] Antares Audio Technologies, Santa Cruz, CA. *AutoTune 2026 Product Manual*, 2025.

[3] J. L. Flanagan and R. M. Golden. Phase vocoder. *The Bell System Technical Journal*, 45(9):1493–1509, 1966.