ECE 417 Multimedia Signal Processing
MP1 - Multiband Excitation Speech Synthesis

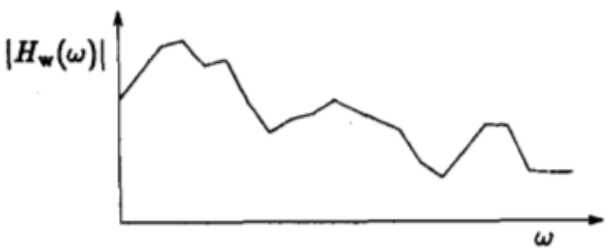
Leda Sari

September 13, 2018

Introduction

- Implement a speech synthesis algorithm based on the multiband excitation synthesis method of Griffin and Lim, 1988
- Speech signal representation: Short time Fourier transform of a windowed signal

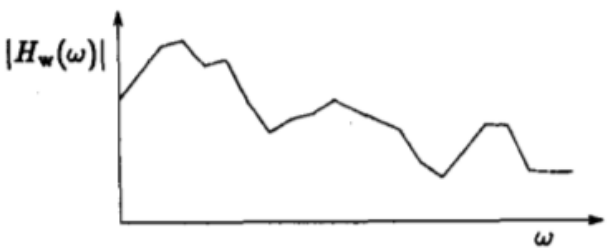
$$|S_w(\omega)| = \underbrace{|H_w(\omega)|}_{\text{Spectral Envelope}} \cdot \underbrace{|E_w(\omega)|}_{\text{Excitation spectrum}}$$



Introduction

- Implement a speech synthesis algorithm based on the multiband excitation synthesis method of Griffin and Lim, 1988
- Speech signal representation: Short time Fourier transform of a windowed signal

$$|S_w(\omega)| = \underbrace{|H_w(\omega)|}_{\text{Spectral Envelope}} \cdot \underbrace{|E_w(\omega)|}_{\text{Excitation spectrum}}$$

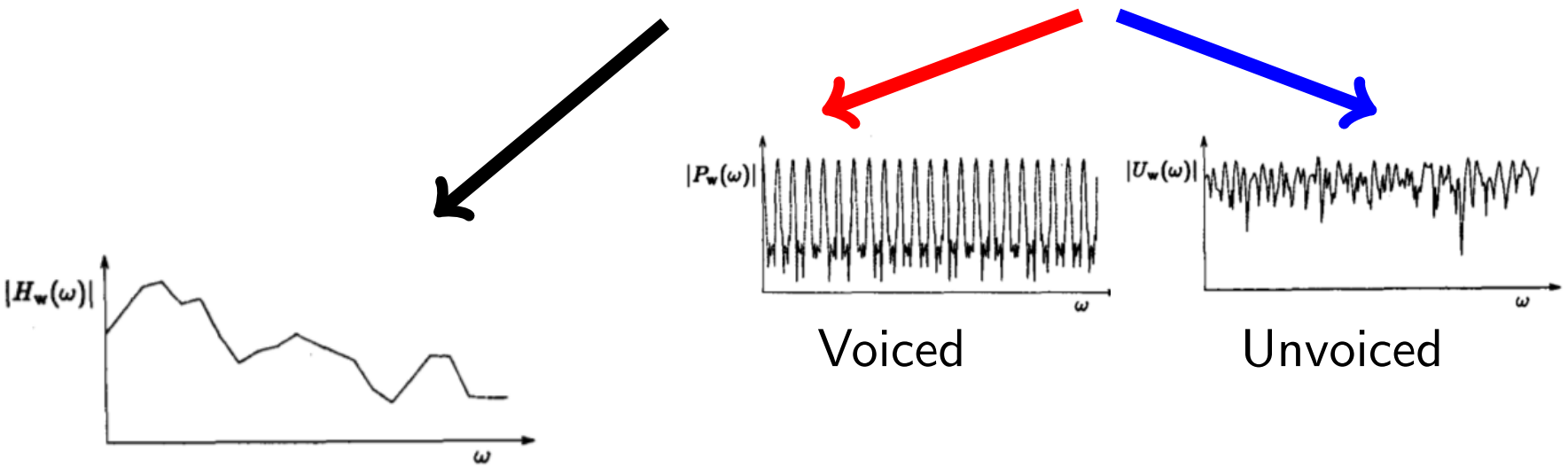


Voiced

Introduction

- Implement a speech synthesis algorithm based on the multiband excitation synthesis method of Griffin and Lim, 1988
- Speech signal representation: Short time Fourier transform of a windowed signal

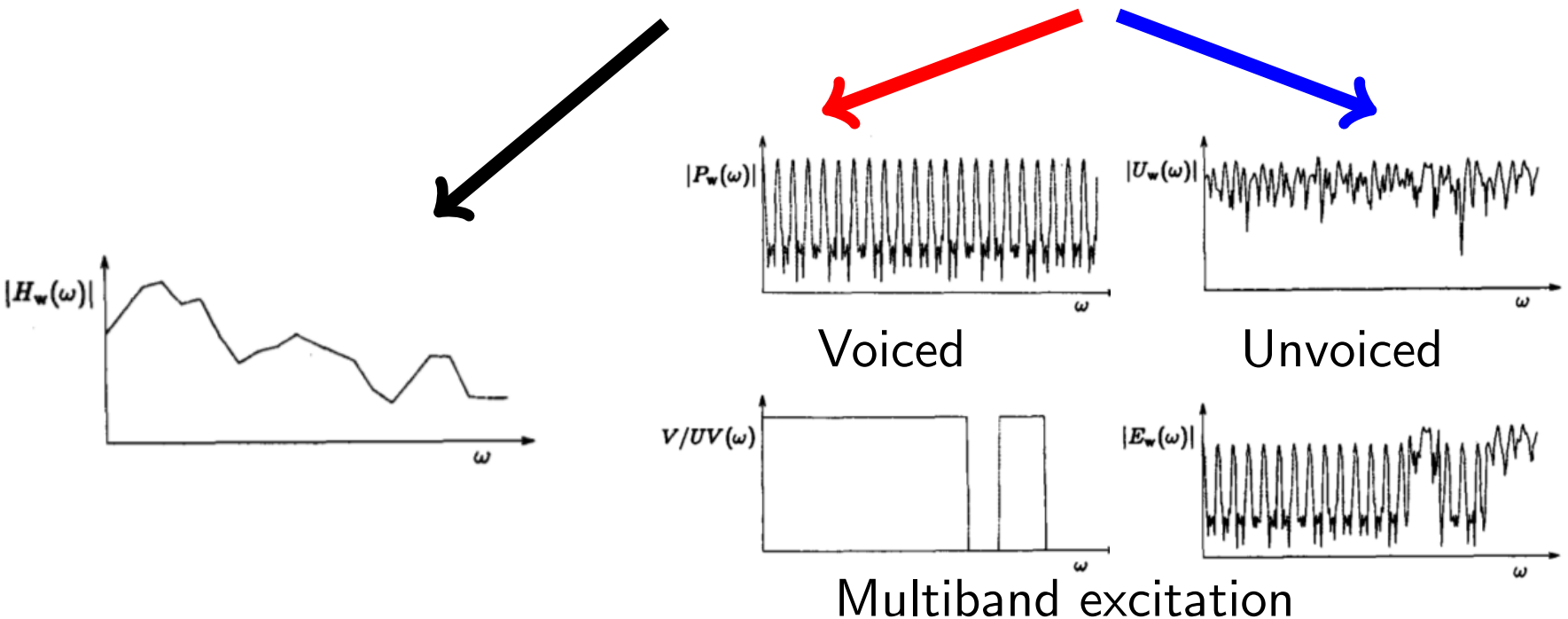
$$|S_w(\omega)| = \underbrace{|H_w(\omega)|}_{\text{Spectral Envelope}} \cdot \underbrace{|E_w(\omega)|}_{\text{Excitation spectrum}}$$



Introduction

- Implement a speech synthesis algorithm based on the multiband excitation synthesis method of Griffin and Lim, 1988
- Speech signal representation: Short time Fourier transform of a windowed signal

$$|S_w(\omega)| = \underbrace{|H_w(\omega)|}_{\text{Spectral Envelope}} \cdot \underbrace{|E_w(\omega)|}_{\text{Excitation spectrum}}$$

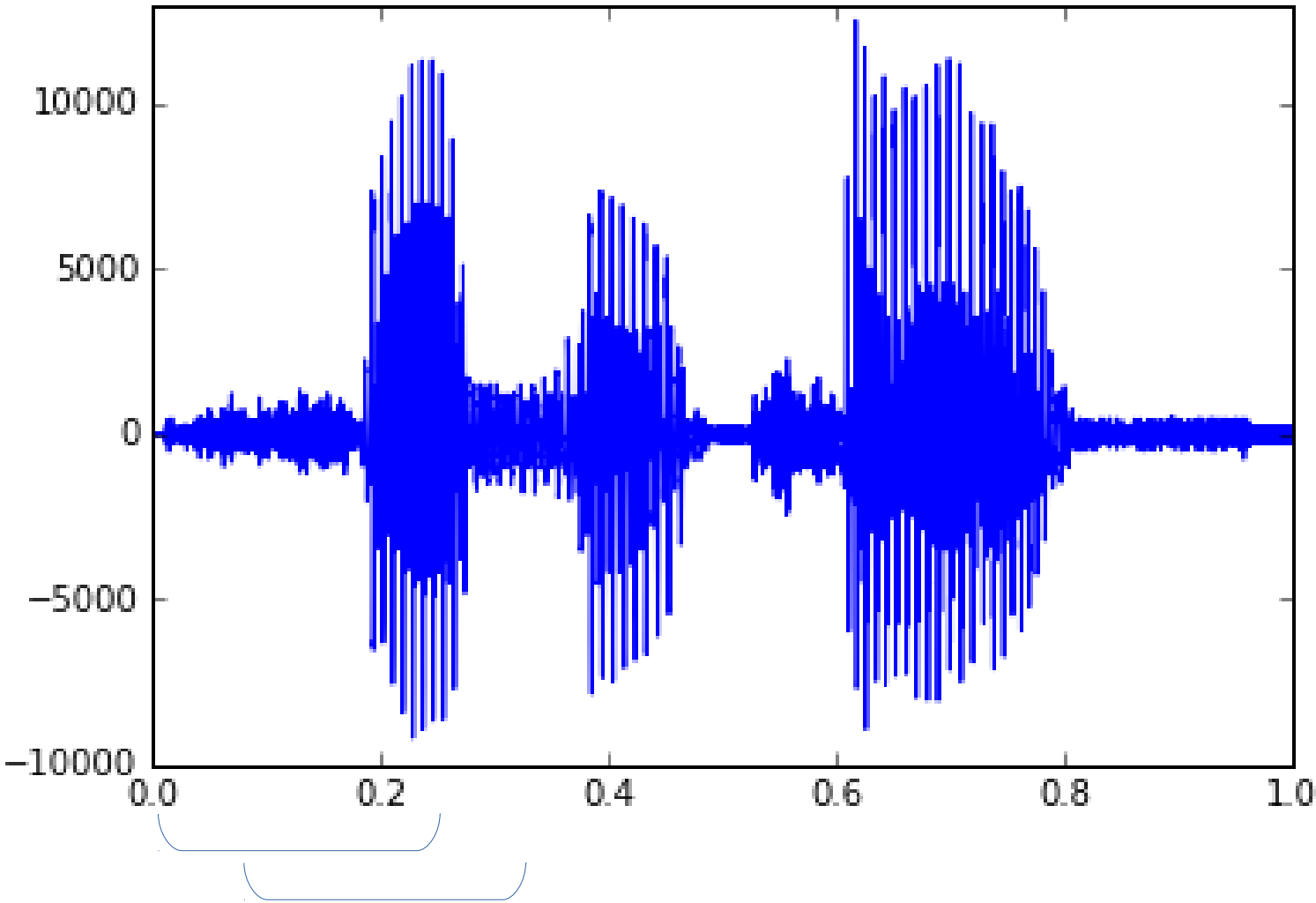


Overview

1. Analysis
   - 1.1 Apply short-time processing of speech signal
   - 1.2 Estimate the parameters of the spectral envelope and the excitation at each frequency band for each speech frame
2. Synthesis
   - 2.1 Voiced speech: time-domain reconstruction
   - 2.2 Unvoiced speech: frequency-domain reconstruction
3. Apply the synthesis algorithm to
   - 3.1 Given clean signal (*s5.wav*)
   - 3.2 Noisy version of *s5.wav* - additive Gaussian noise
   - 3.3 Your own signal

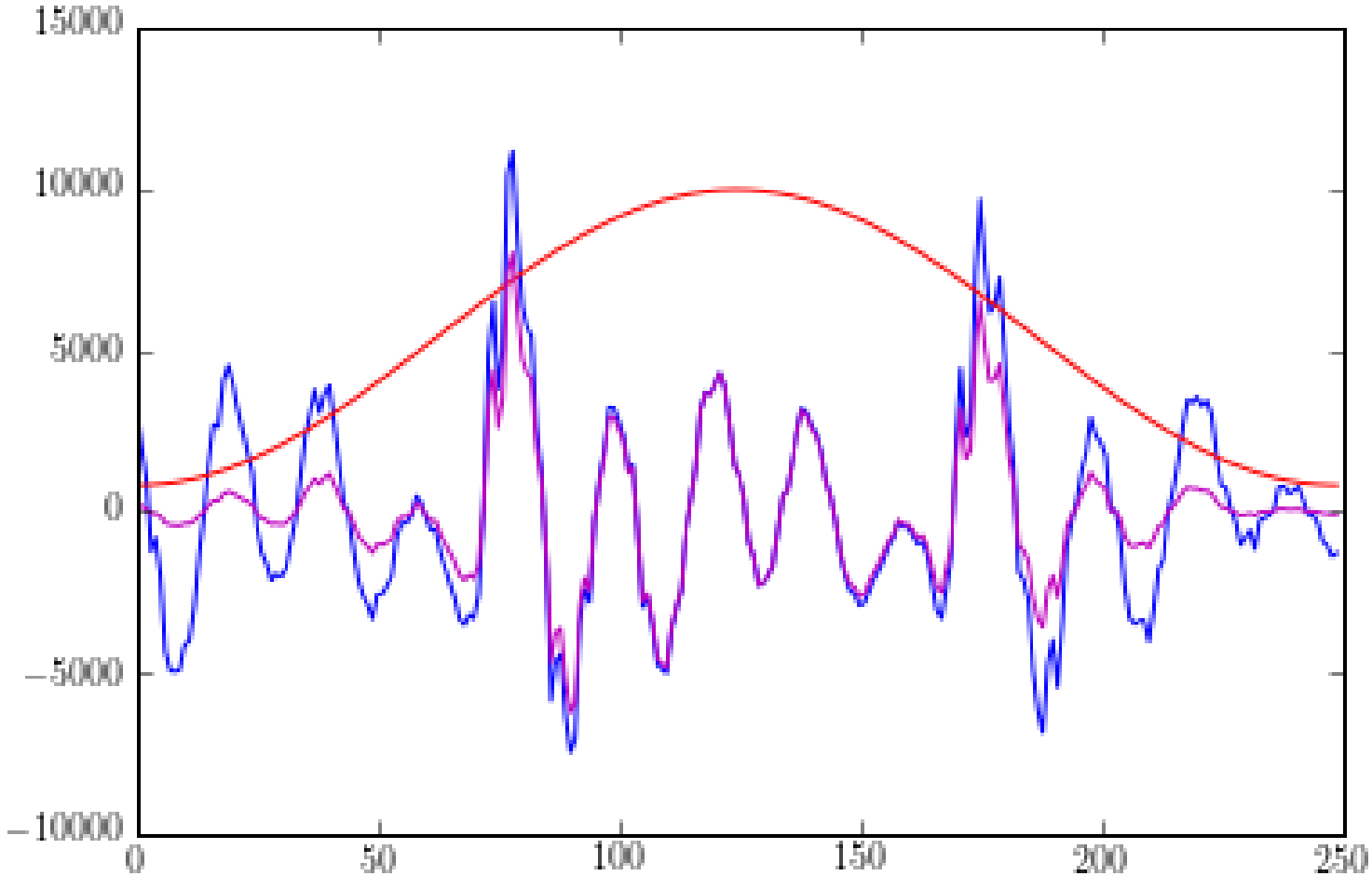
Speech Signal

scipy.io.wavfile.read (Python) or audioread (MATLAB)
25ms frames with 10ms shift



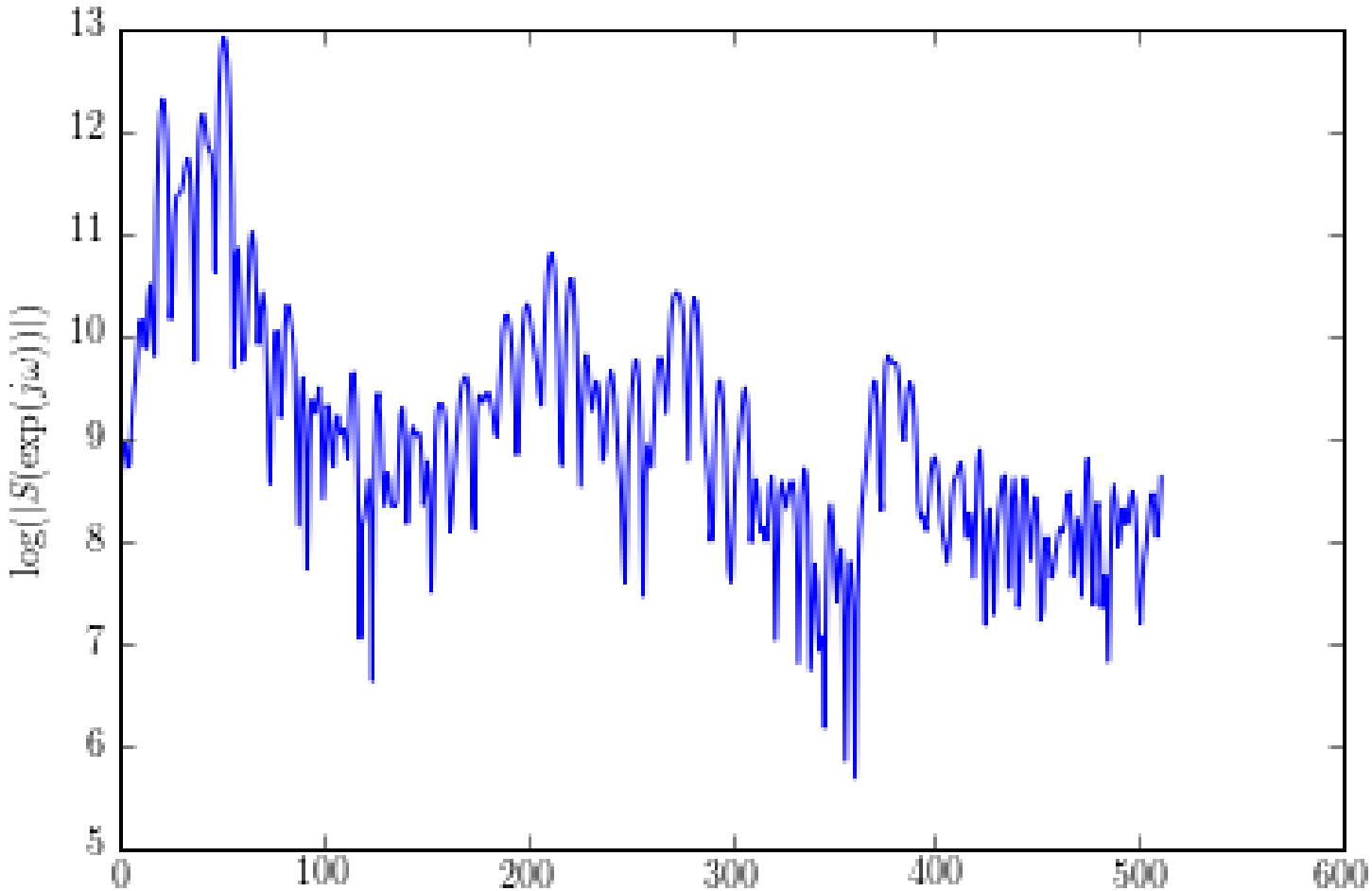
Windowed Frames

Use Hamming window

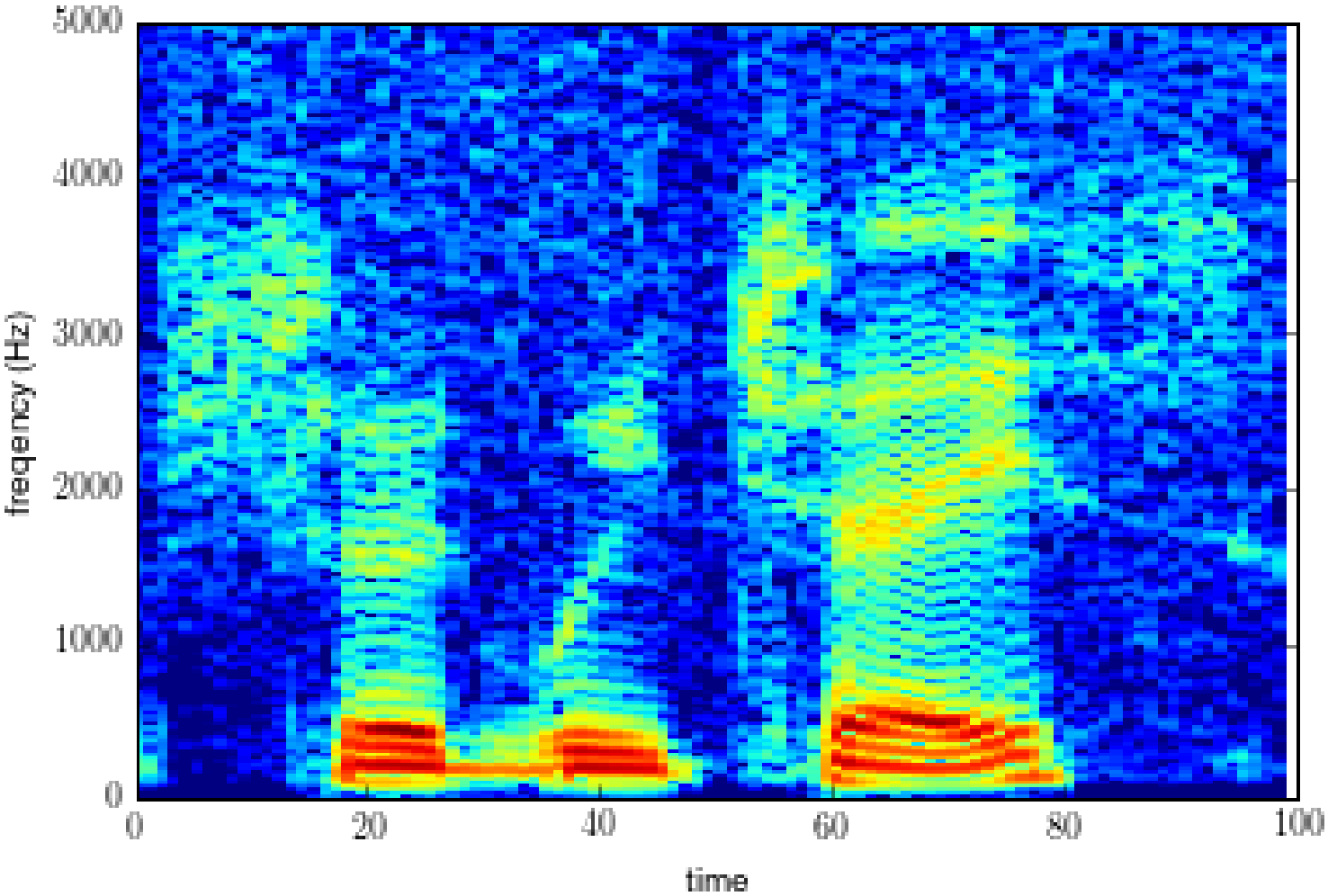


Short-time Spectrum

scipy.fftpack.fftpack.fft or fft



Spectrogram



Analysis - Pitch Estimation

For a given speech frame s and a window w,
compute the autocorrelation of $w^2[n]s[n]$:

$$\phi(m) = \sum_{n=-\infty}^{\infty} w^2[n]s[n]w^2[n-m]s[n-m] \tag{1}$$

Get a rough integer estimate for the pitch:

$$P_0 = \arg\max_P \Phi(P) = \arg\max_P P \sum_{k=-\infty}^{\infty} \phi(kP) \tag{2}$$

Pitch Estimation

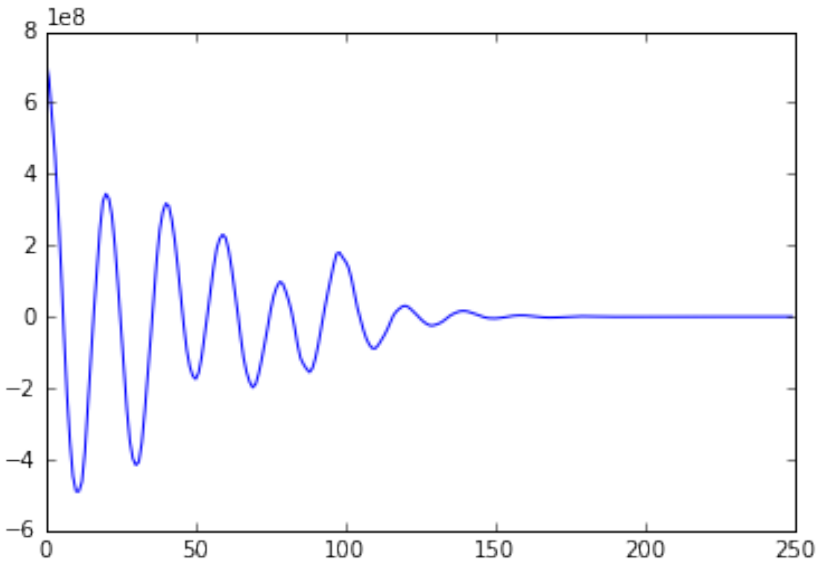


Figure : Autocorrelation

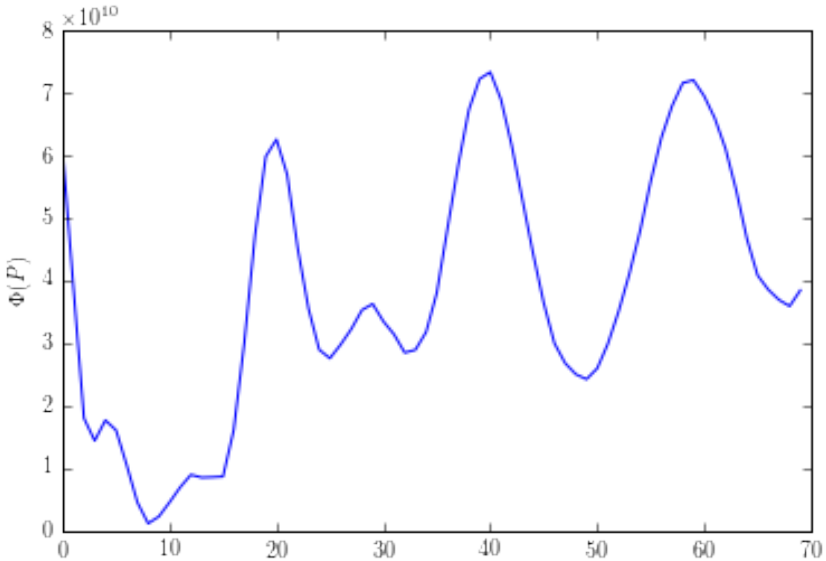


Figure : $\Phi(P)$ versus P

Analysis - Pitch Refinement and Spectral Envelope

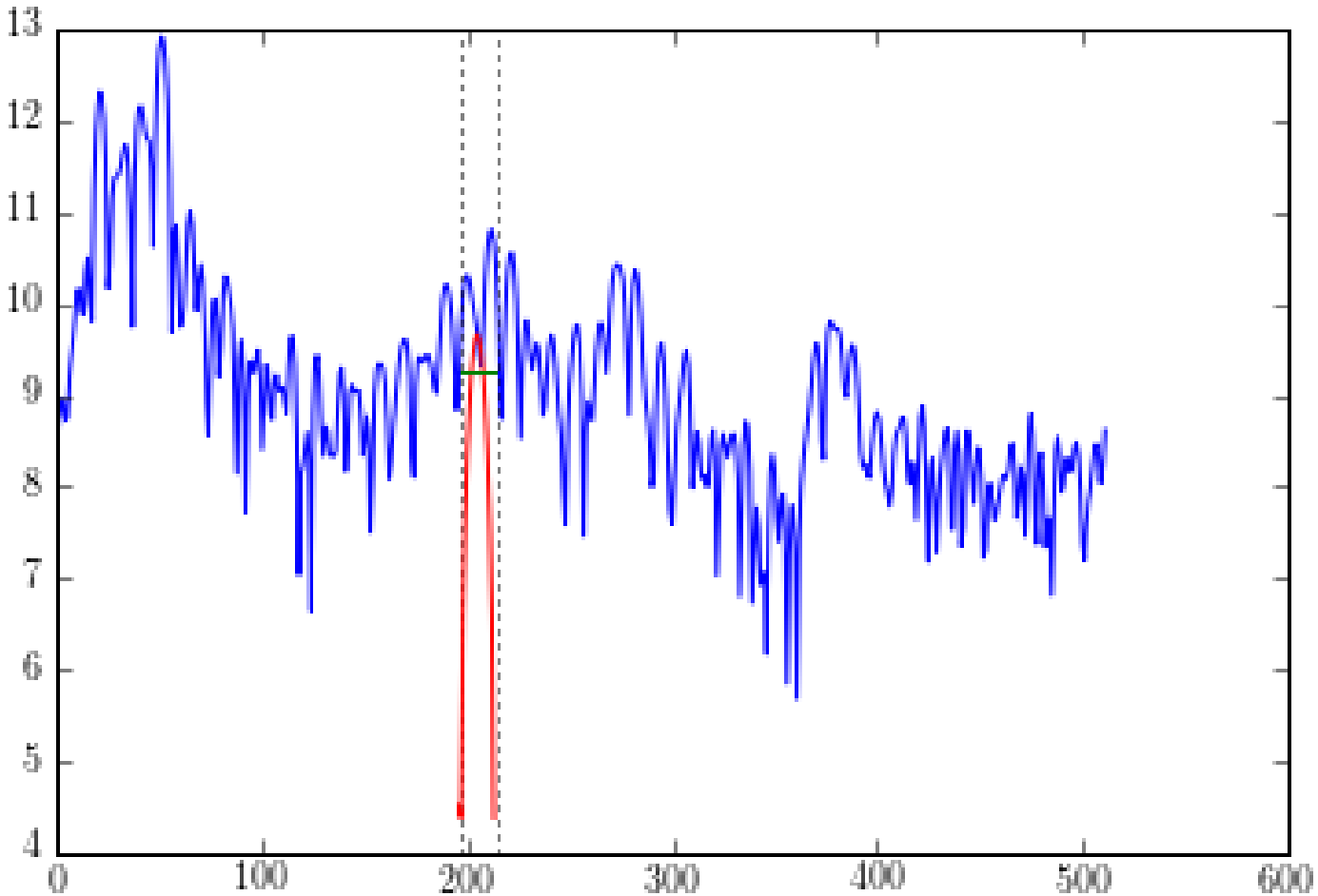
For a given rough pitch estimate P_0,

1. Compute $\omega_0 = 2\pi/P_0$
2. Determine the frequency bands:
   $[a_m, b_m] = \left[\left(m - \frac{1}{2}\right)\omega_0, \left(m + \frac{1}{2}\right)\omega_0\right], m = 1, 2, \ldots, P_0 - 1$
3. Compute $A_m = \frac{\int_{a_m}^{b_m} S_w(\omega)E^*(\omega)d\omega}{\int_{a_m}^{b_m} |E(\omega)|^2 d\omega}$
   - 3.1 Assume that the m-th band is voiced: take $E_w(\omega)$ as the Fourier transform of the window centered around $m\omega_0$
   - 3.2 Assume that the m-th band is unvoiced: take $E_w(\omega) = 1,\ \omega \in [a_m, b_m]$
   - 3.3 For each case, compute the error $\varepsilon_m(P_0) = \frac{1}{2\pi}\int_{a_m}^{b_m} |S_w(\omega) - A_m E_w(\omega)|^2 d\omega$
   - 3.4 If $\varepsilon_{m,\text{voiced}} < \varepsilon_{m,\text{unvoiced}}$, band m is voiced, otherwise it is unvoiced
   - 3.5 Pick the corresponding A_m parameter
4. Compute the total error: $\varepsilon(P_0) = \sum_{m=1}^{P_0-1} \varepsilon_m(P_0)$

Parameter Estimation

Voiced/unvoiced?

A_m?



Analysis - Pitch Refinement

- For each pitch estimate in $[P_0 - 2, P_0 - 1.8, P_0 - 1.6, \ldots, P_0 + 1.8, P_0 + 2]$
- Repeat the previous procedure and compare the errors $\varepsilon(P)$
- Final pitch estimate is $P = \arg\min_P \varepsilon(P)$
- Compute values of A_m and voiced/unvoiced decisions for the refined P

Synthesis - Voiced Part

Let the speech segments be taken at every K samples, to reconstruct the samples of the voiced signal between $[fK, (f+1)K)$

$$s_v[n] = \sum_m A_m[n] \cos(\theta_m[n]) \tag{3}$$

$$\theta_m[n] = \theta_m[n-1] + m\omega_0[n] \tag{4}$$

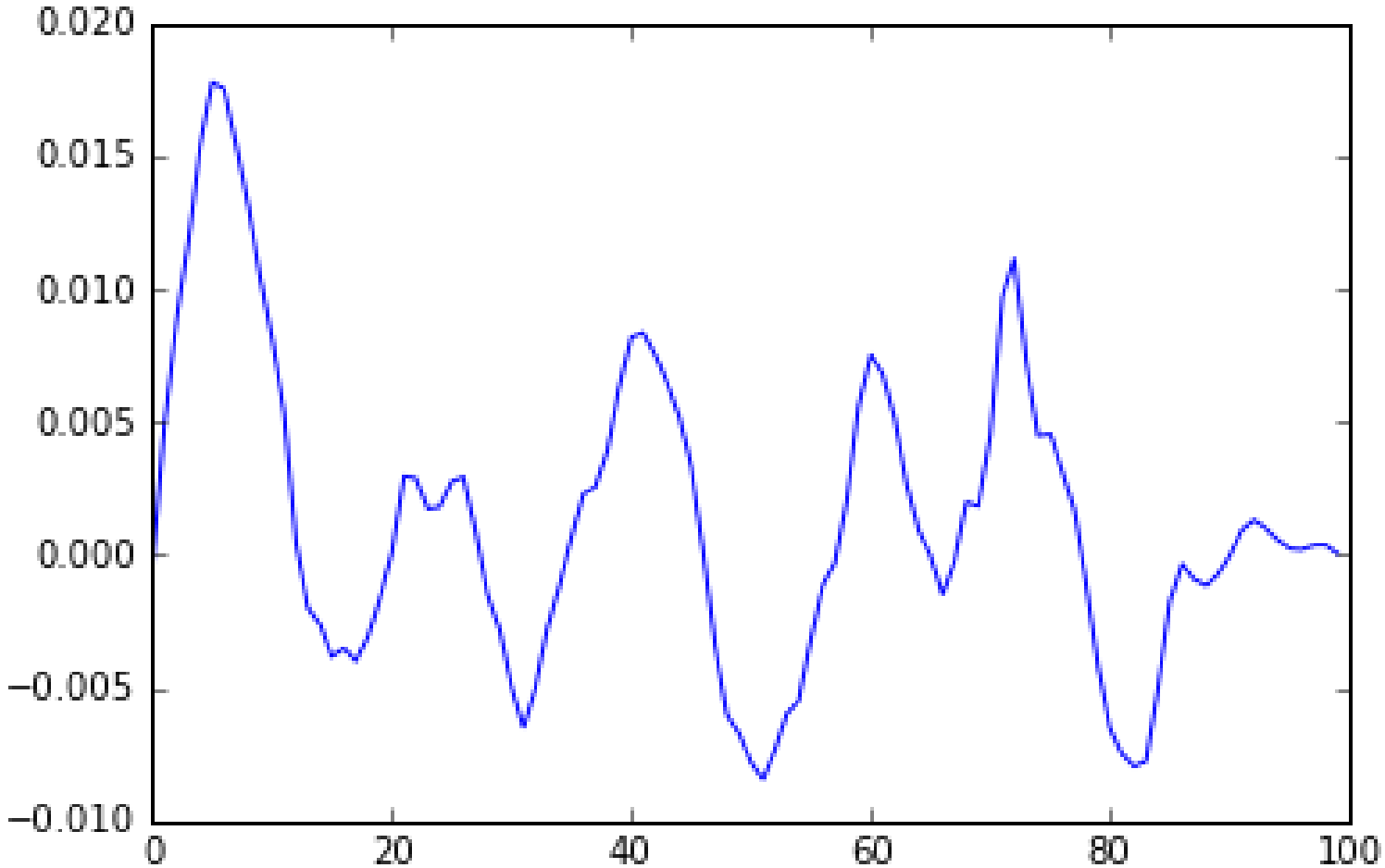
$$\omega_0[n] = \left(f + 1 - \frac{n}{K}\right) \frac{2\pi}{P_f} + \left(\frac{n}{K} - f\right) \frac{2\pi}{P_{f+1}} \tag{5}$$

$$A_m[n] = \left(f + 1 - \frac{n}{K}\right) A_{m,f} + \left(\frac{n}{K} - f\right) A_{m,f+1} \tag{6}$$

Notes :

- Take $A_{m,f} = 0$ for unvoiced bands
- Due to differences between pitch estimates between consecutive frames, the number of bands can change between frames. For the nonexistent bands, assume that $A_m = 0$.
- For the first frame ($f = 0$), assume that $\theta_m[-1] = 0$
- For the last frame, take $\omega_0 = \frac{2\pi}{P_f}, A_m[n] = A_{m,f}$

Voiced Part



Synthesis - Unvoiced Part

For each unvoiced band in each frame

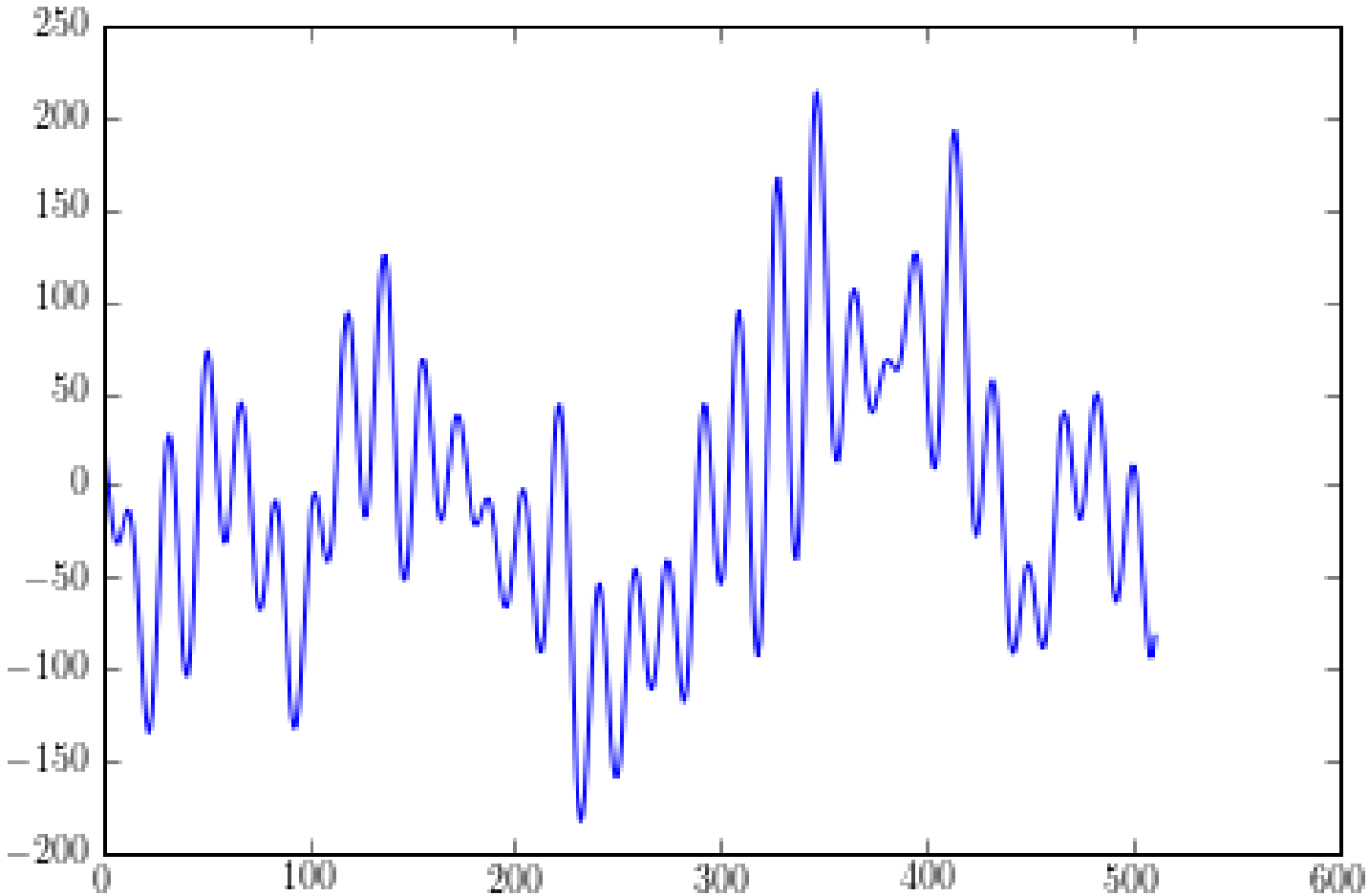
1. Estimate the noise variance $\sigma_m^2 = \frac{1}{b_m - a_m} \int_{a_m}^{b_m} |S_w(\omega)|^2 d\omega$
2. Sample the noise transform:

$$U_f[k] = \begin{cases} 0, & 2\pi k/N_{fft} \text{ is voiced} \\ \mathcal{N}(0, 0.5\sigma_m^2) + j\mathcal{N}(0, 0.5\sigma_m^2), & \text{else} \end{cases} \tag{7}$$

3. Compute $u_f[n]$ by inverse FFT
4. Use linear interpolation to construct $s_u[n]$: for $fK \leq n < (f+1)K$

$$s_u[n] = \left(f + 1 - \frac{n}{K}\right) u_f[n - fK] + \left(\frac{n}{K} - f\right) u_{f+1}[n - (f+1)K] \tag{8}$$

Unvoiced Part



Steps

1. After synthesis, save the output (*scipy.io.wavfile.write* or *audiowrite*) and listen. Can you understand what is being spoken?
2. Compare the spectrograms of the original and the synthetic speech
3. Change your multiband voiced/unvoiced decisions to a single decision and synthesize the signal again
4. Artificially add Gaussian noise to the original speech signal with different variances, repeat previous steps
5. Record your own voice and repeat steps 1-3

Remarks

- For numerical values of the parameters, refer to the MP description
  - Female voice has lower pitch period, so when processing your own speech, you may need to adjust the range of the pitch appropriately
- Please check the Notes section of the MP description
- Reports
  - Include your plots (pitch vs frame, error vs frame, spectrograms) and label the axes properly
  - Also include qualitative comparison of the outputs
- Submission
  - Submit your report (PDF) and codes (zip) to Compass
  - File names must be <Lastname>_<Firstname>_report.pdf and <Lastname>_<Firstname>_code.zip
  - Teams will submit a single report but make sure that all names are included in the report
- Questions
  - Post questions on Piazza
  - Come to the office hours