LPC-10 Speech Encoder Implementation

December 16, 2020

Karndeep Singh Rai-Bhatti

Key take-aways are highlighted in blue.

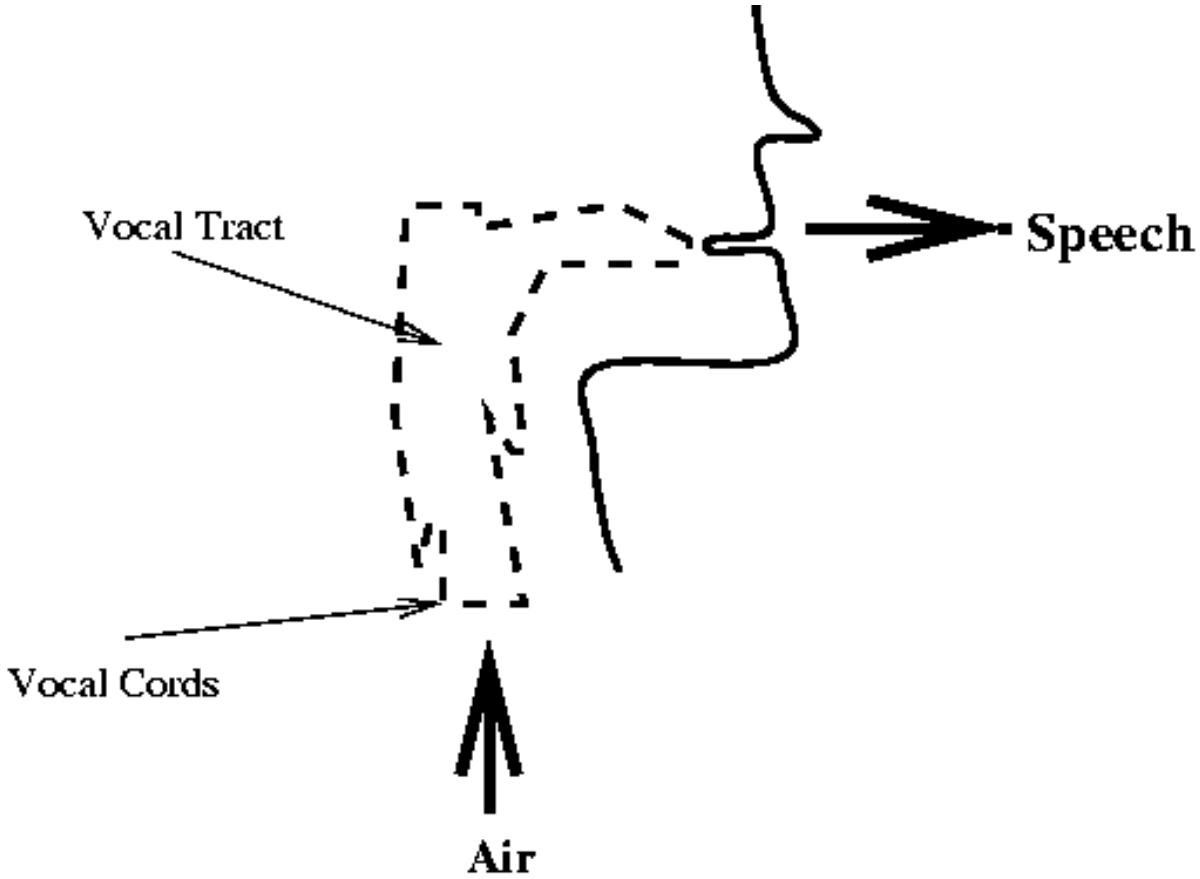
Motivation

- Speech compression for communication systems
  - Telephones
  - Voice over IP
  - Videoconferencing
- Maximize audio data compression
  - Without corrupting the words being spoken
  - Retaining an intelligible voice
  - Additional Security - encryption
- Model that Emulates Human Speech Generation – Physical Model

We want to maximally compress audio while preserving voice and meaning.

Physical Model of Speech and Synthesis

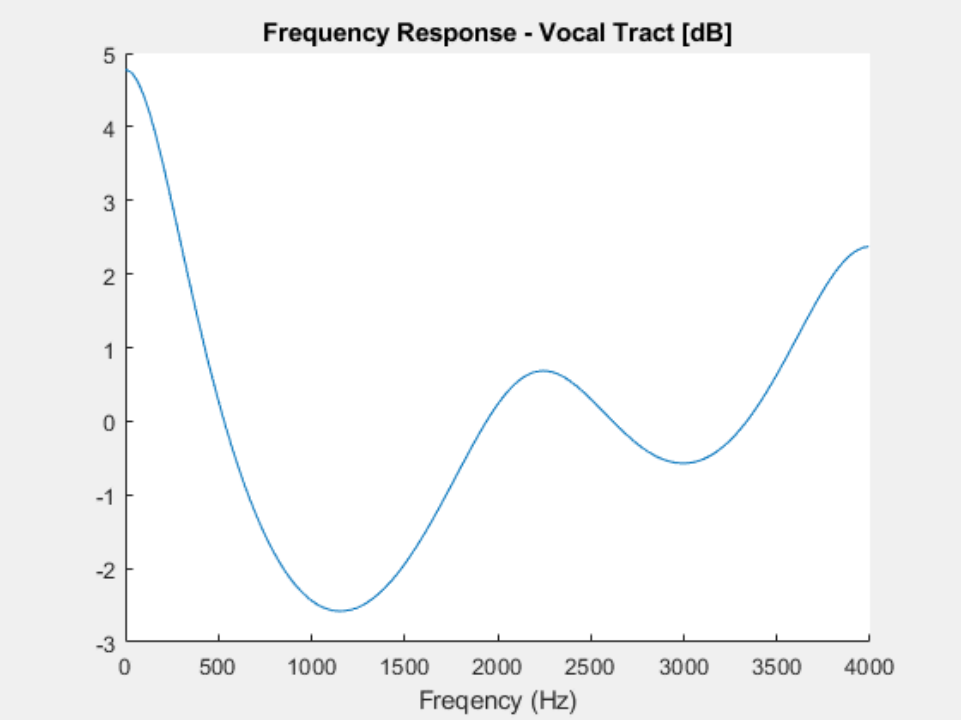
- Air ~ Speech volume
  - Gain
- Vocal Cords ~ Excitation signal
  - Impulse train
  - Period = fundamental Frequency of voice
- Vocal Tract ~ Linear Transformation
  - Modeled by All Pole System
  - Our Linear Predictor
- Speech ~ Output signal
  - Pass excitation through filter with appr

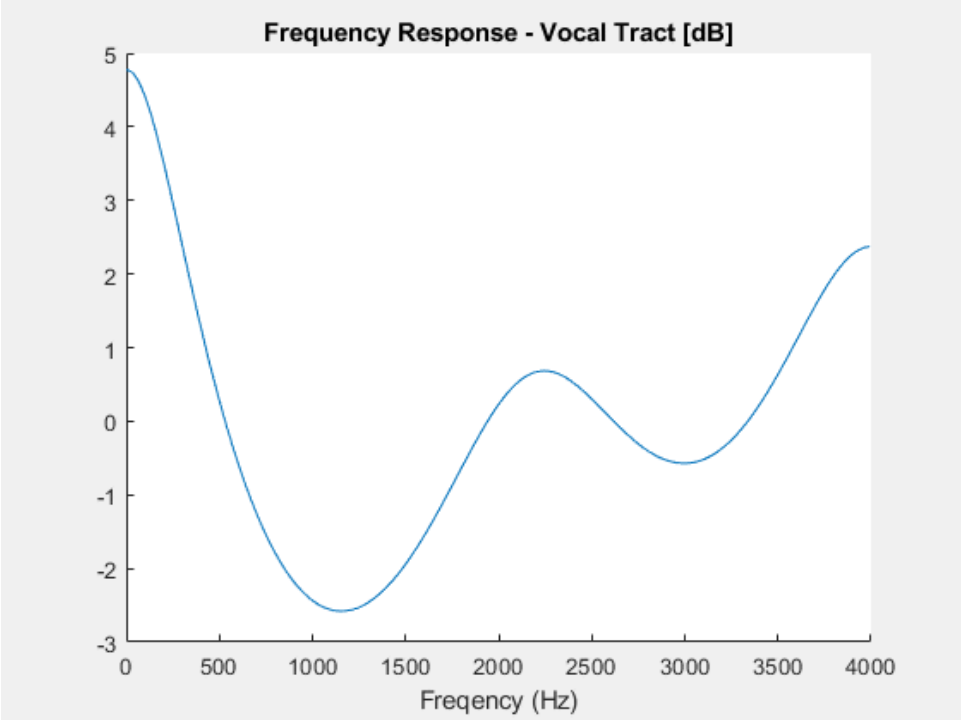


LPC10 models speech synthesis based on the human vocal tract.

Modeling Speech as a LTI system

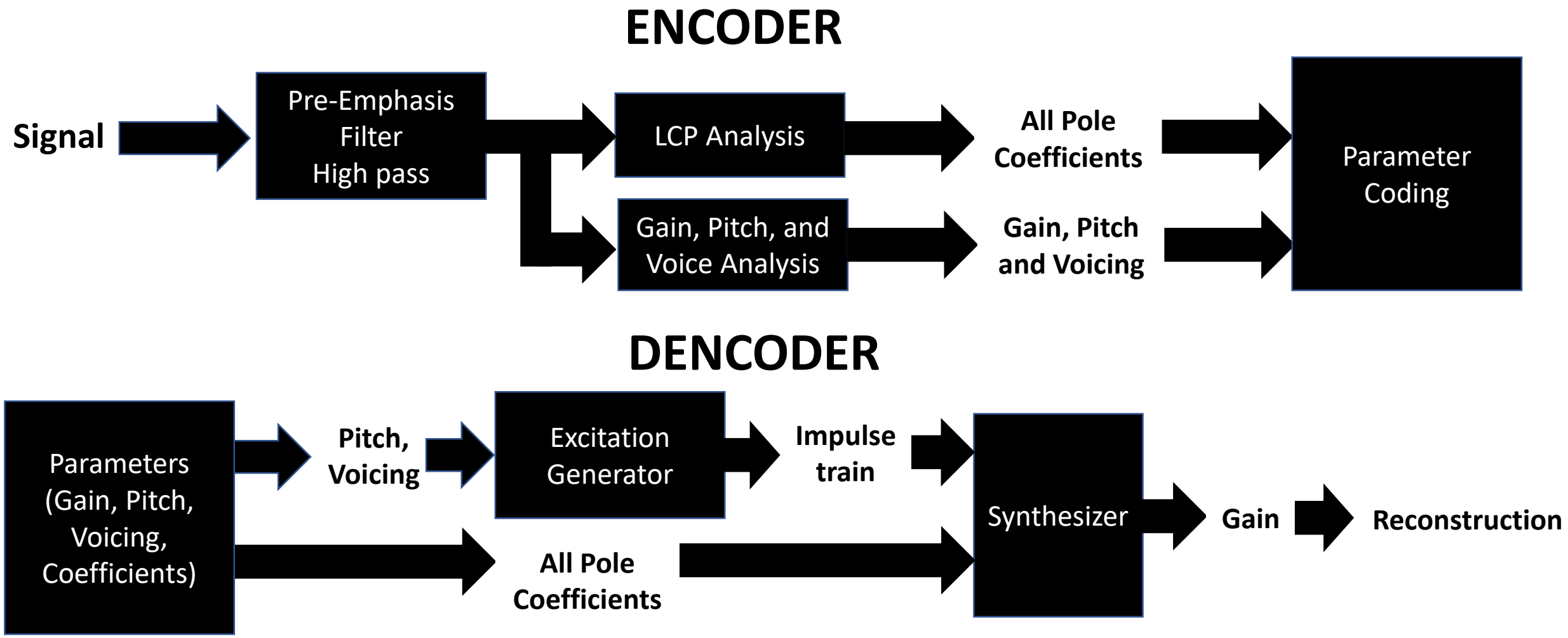
- Model holds true at small intervals (25ms)
- Unvoiced is approximated as 4^{th} order all-pole filter
- Voiced is approximated as 10^{th} order all-pole filter





The Vocal Tract is modeled as an LTI all-pole filter.

LPC-10 workflow



LPC10 is an encoder-decoder network.

What did I do?

- Implemented LPC-10
  - Look at system holds up to different users and emotions
- Attempted to Improve LPC-10
  - Voiced detection
  - All-Pole Filter
  - Pitch Interpolation
  - Gain Interpolation
- Implemented LPC-10 in real-time using MATLAB

I implemented and qualified LPC10, attempted to improve the algorithm and implemented it in real-time.

Data Set: EmoV-DB

- Databased for characterizing how Emotionally Expressive a Voice Generation System is
- 4 Speakers
  - Bea & Jenie (female)
  - Sam & Josh (male)
  - All Bea data recorded at 44k, so it was excluded
- 5 emotions
  - Neutral, Sleepiness, Anger, Disgust, Amusement
  - We will look at Neutral, Anger, Sleepiness
- Use to assess the quality of our LPC-10 Reconstruction

https://github.com/numediart/EmoV-DB

The encoder was tested on both genders and for 3 extremes of emotion.

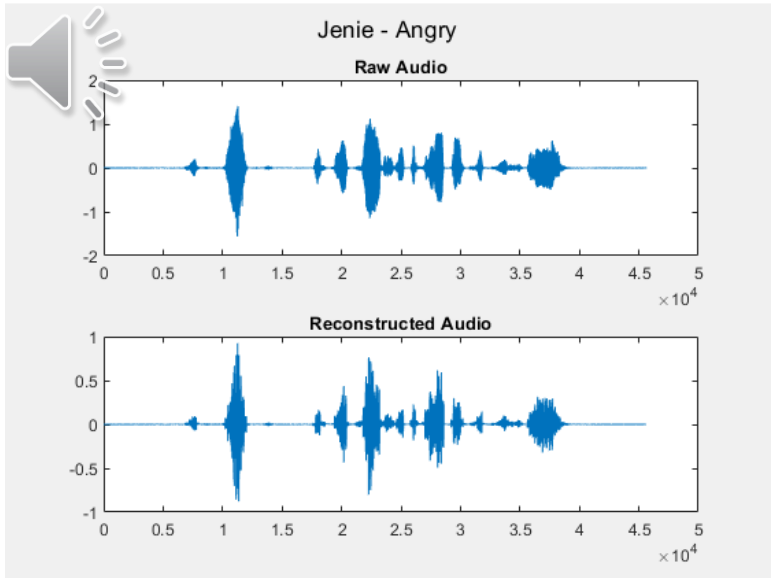
Emotions and Signal Periodicity

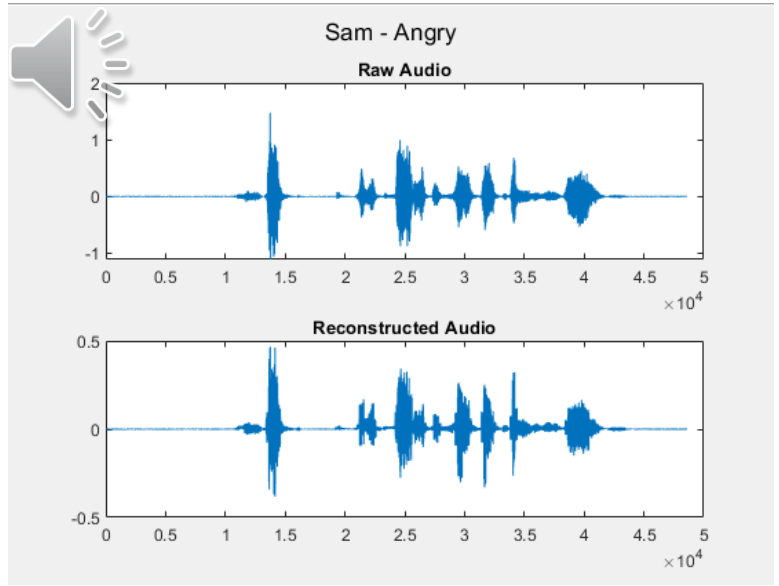
- Emotions tied to periodicity of speech signal
- Anger/Strong Emotions: Expect More Periodicity
  - Expection: Ideal case for model
- Fatique/Weak Emotions: Speech becomes more Aperiodic
  - Expectation: Model will perform more poorly
- LPC-10 relies on the assumption that speech is periodic at small time scales (ms)
- Test the robustness of the algorithm

Emotion is tied to the periodicity of a speech waveform.

Angry Audio – LPC10

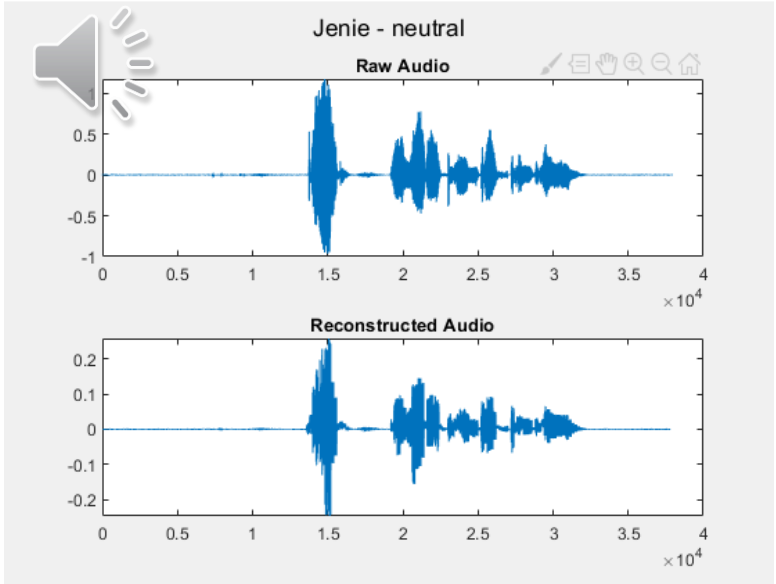
- Periodic Signal
- Waveforms matched closely
- Audio suffers from distortion, primarily when volume peaks

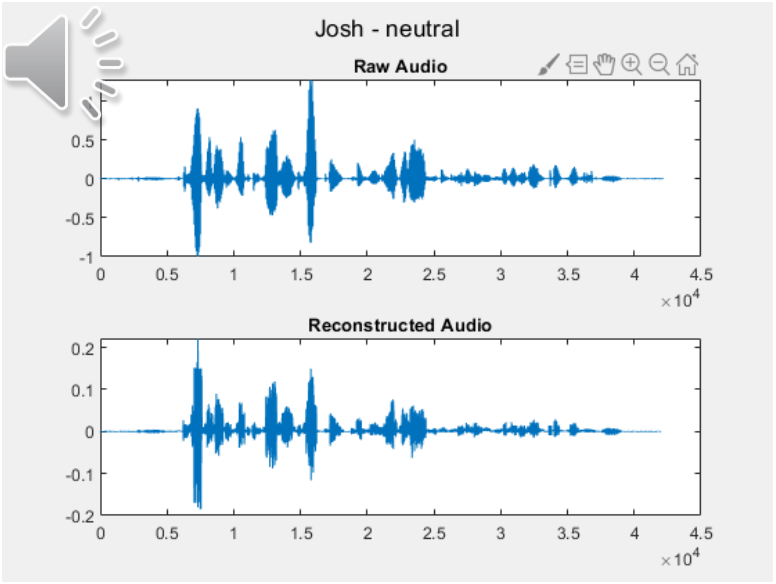


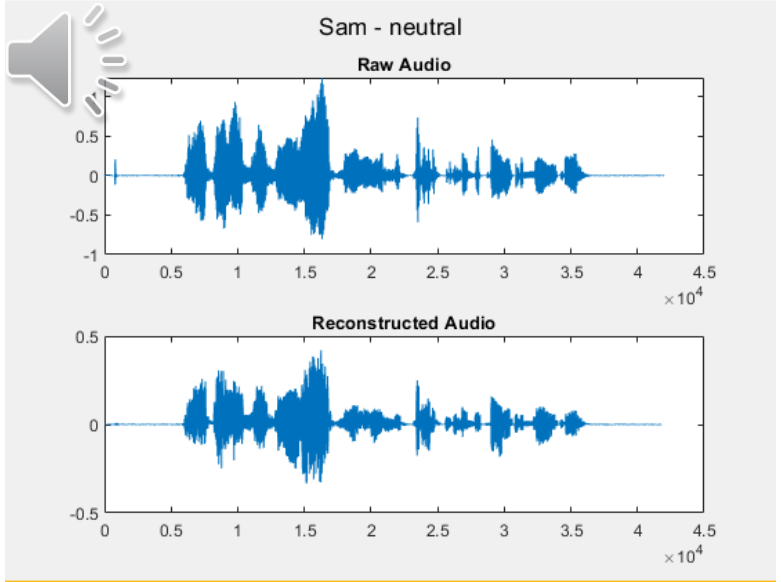


Neutral Audio – LPC10

- Appears to work a lot better on female voices
- Cadence of Voice significant factor in clarity
- Reduced distortion compared to Angry speech

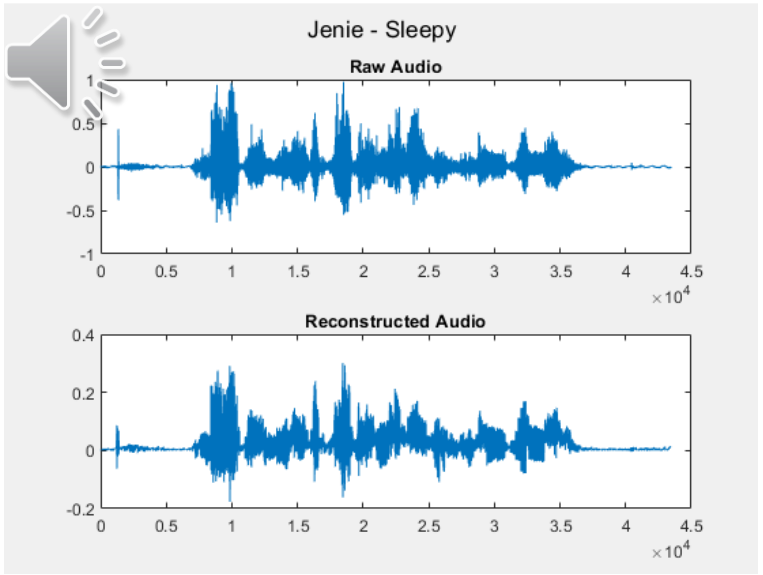


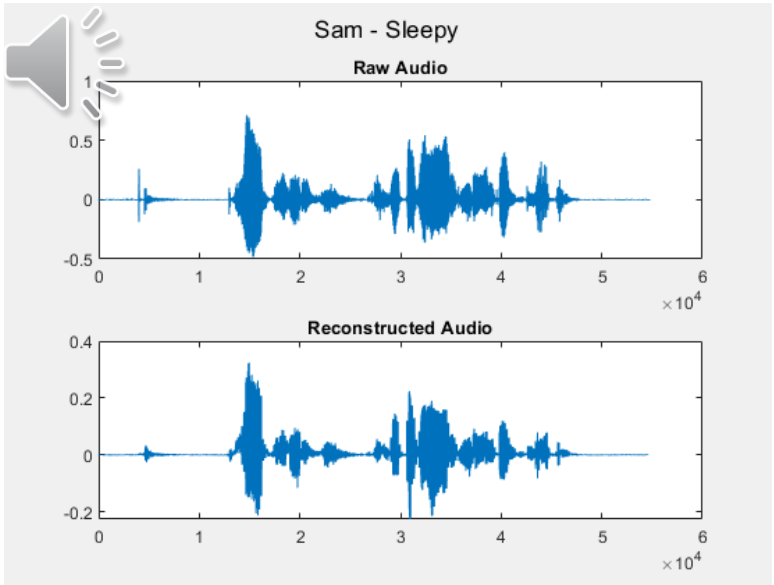


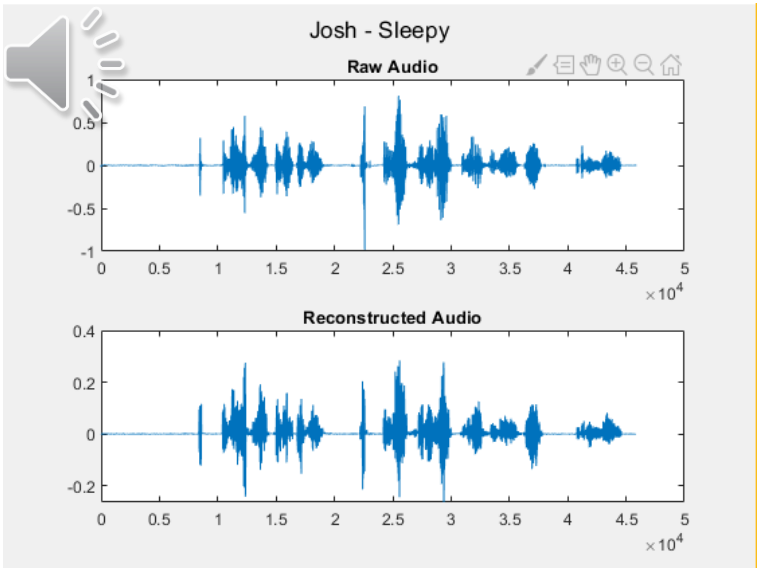


Sleepy Audio – LPC10

- Non-vocal noises handled poorly (high frequency jitter)
  - Reconstruction is very ‘Breathy’
- Slow speech significantly easier to understand

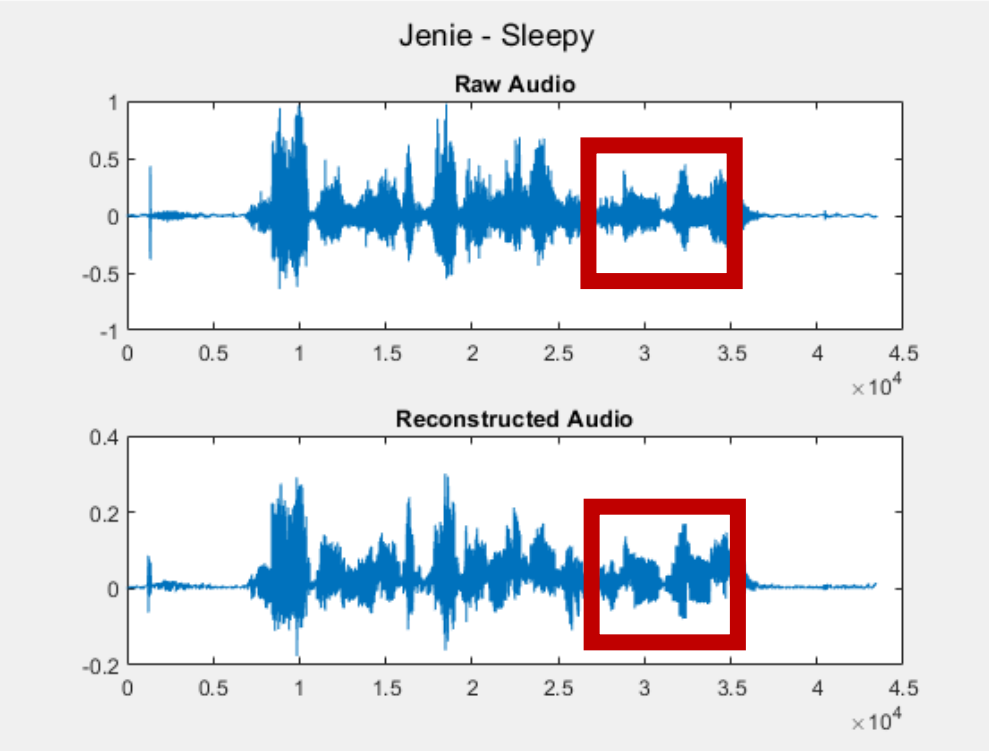


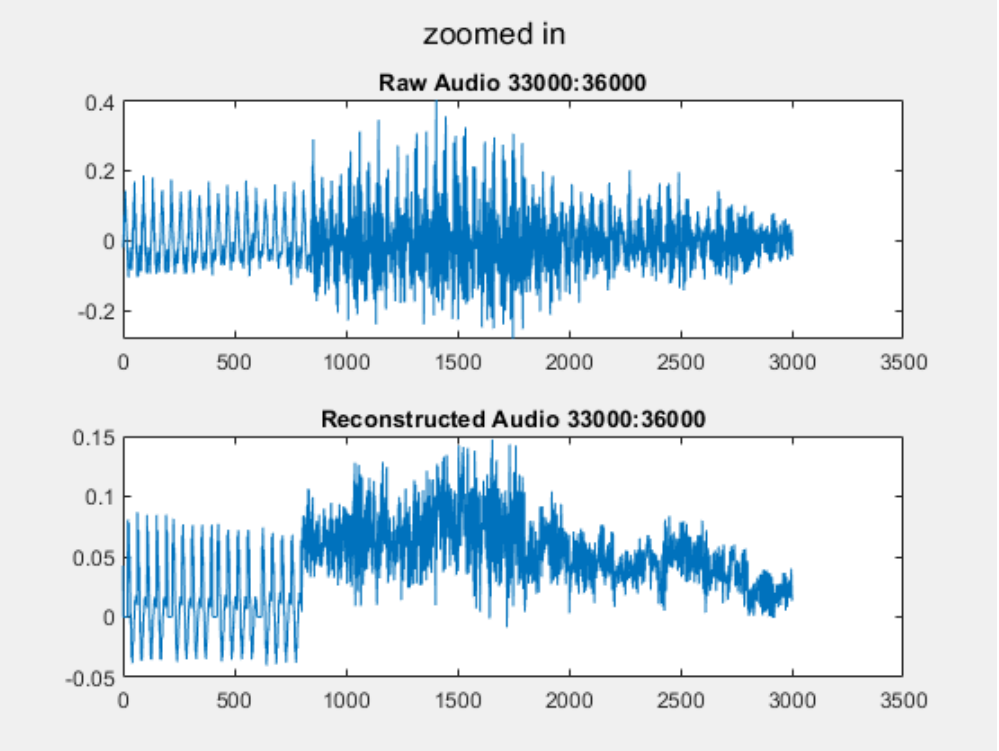




Limitation of LPC?

- Distortion – failed prediction
- Voiced Determination? (all prior used zero crossing)
- Increase number of coefficients?





Linear Predictor performed poorly at times, investigated a solution in the next slides.

Voiced Detection – Energy vs Zero-Crossing

- Energy Based Approach
  - For a speech clip:
    - Absolute value ()
    - Summation ()
    - Normalize to clip length
  - If energy > energy Threshold
    - Speech
  - Else
    - Noise

Tests Energy or Volume

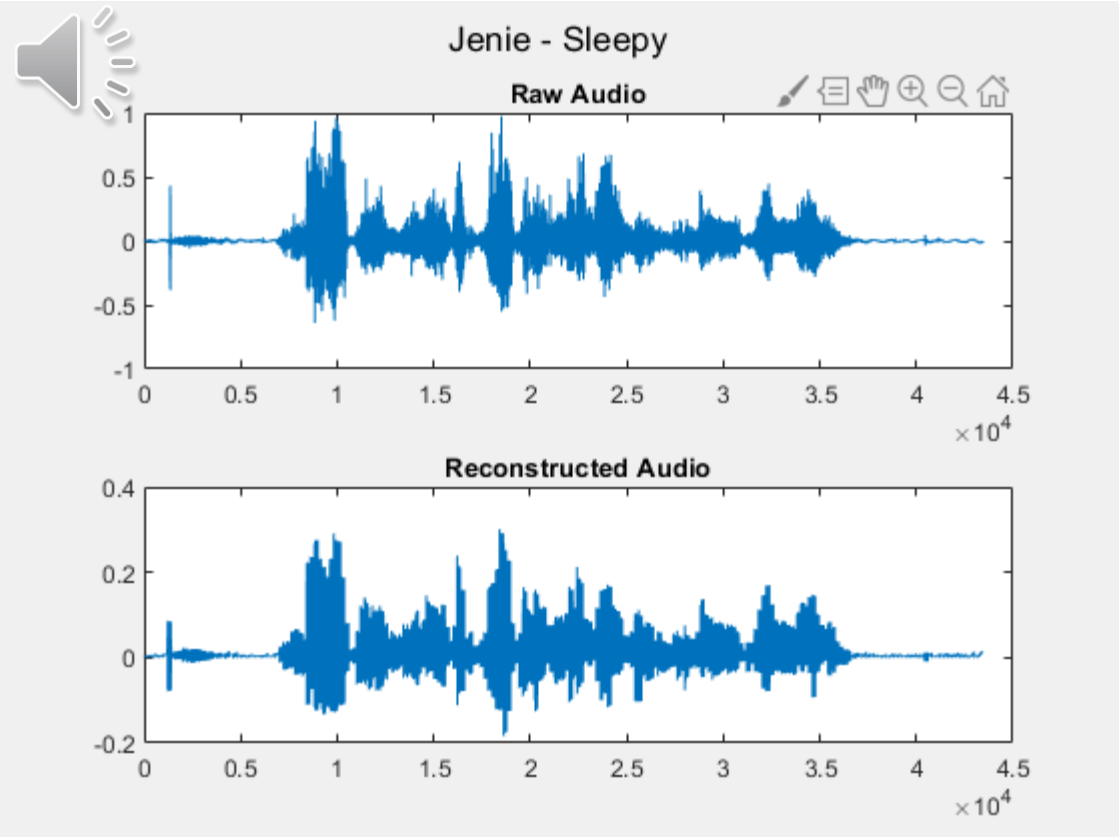
- Zero-Crossing
  - For a speech clip:
    - Turn in 1s -1s depending on sign
    - Take derivative
    - Count number of flips (zero-crossings)
    - Normalize to length of clip
  - If # zero crossings > crossing Threshold
    - Noise
  - Else
    - Speech

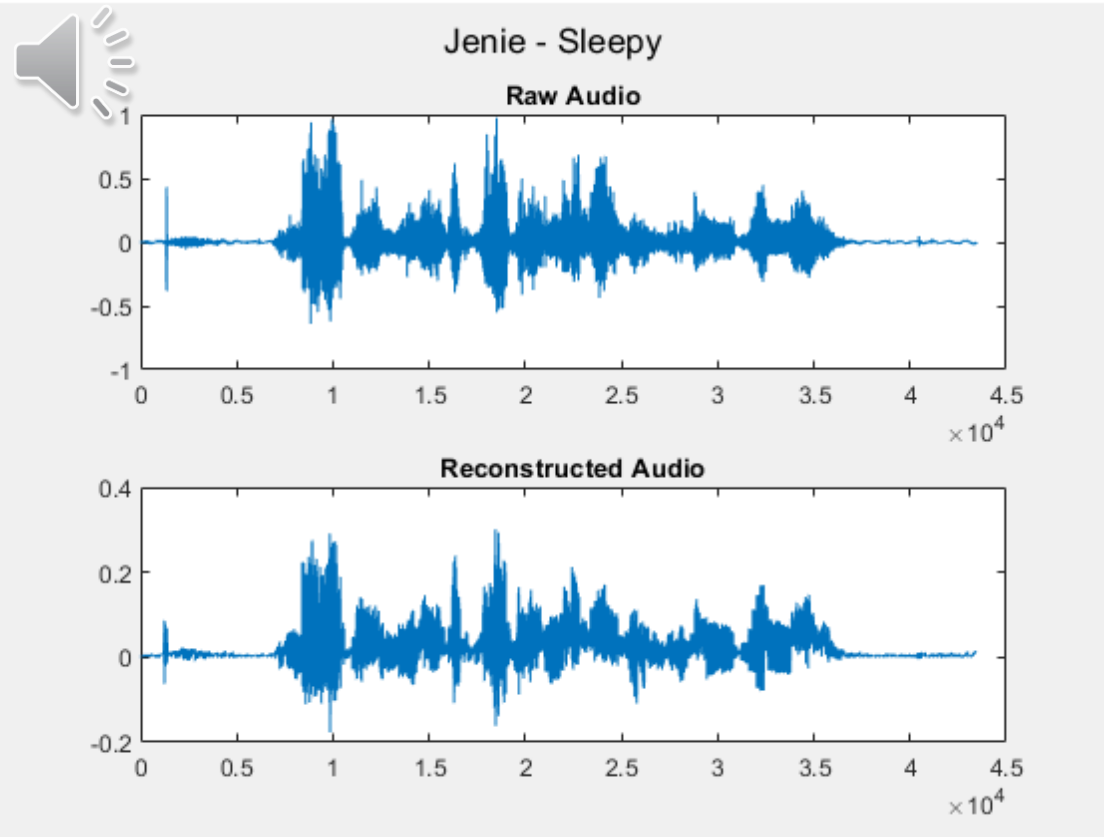
Test Periodicity

Voicing detection was done using two different metrics.
Each resulted in a different audio signature.

Voiced Detection – Jenie-Sleepy

- Energy Based Approach
  - Better preserve of Pitch
- Zero-Crossing
  - Less noise/artifacts





Coefficient Estimation – Number Coefficients

- N = 10
  - Used in official Standard
- Tested on very clear sample
  - N = 5
  - N = 20
  - N = 40
- Test using both Voicing methods

Explored how the filter order impacted synthesis quality.

Coefficient Estimation – Number Coefficients

Energy Based Voice

- N = 5
- N = 10
- N = 20
- N = 40

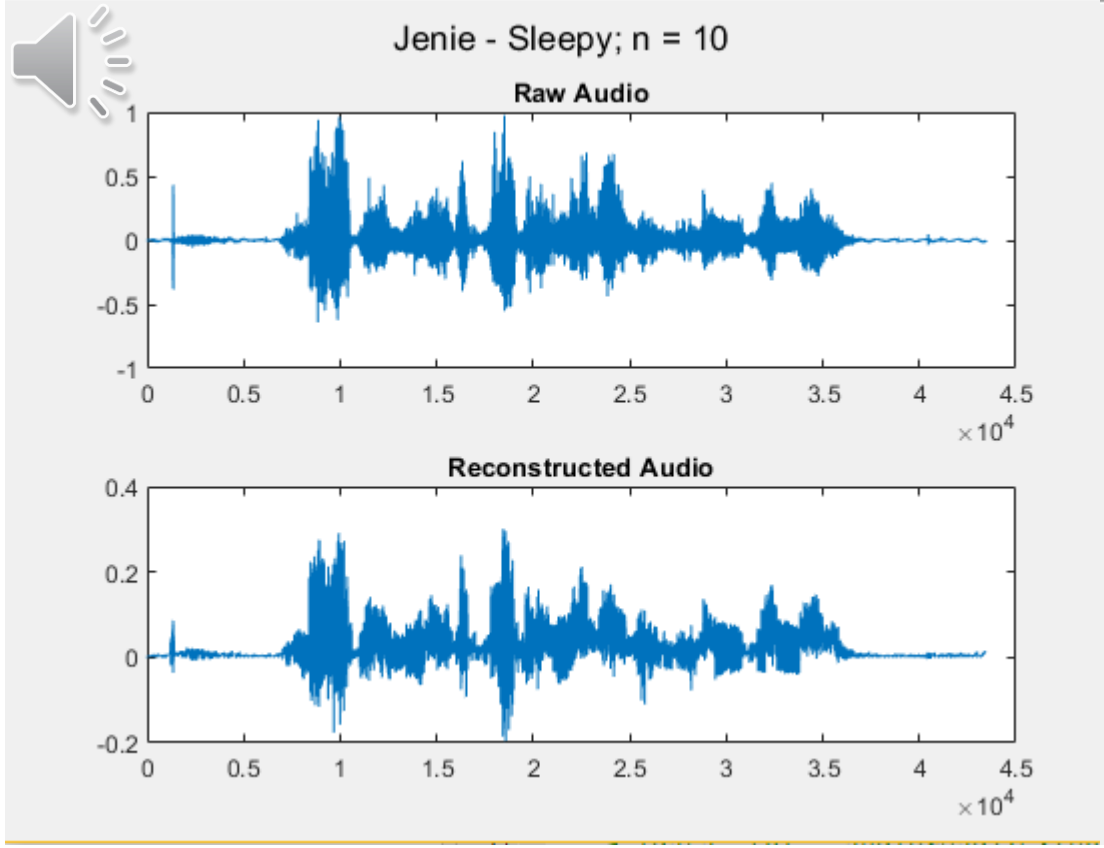
Zero Crossing Based Voice

- N = 5
- N = 10
- N = 20
- N = 40

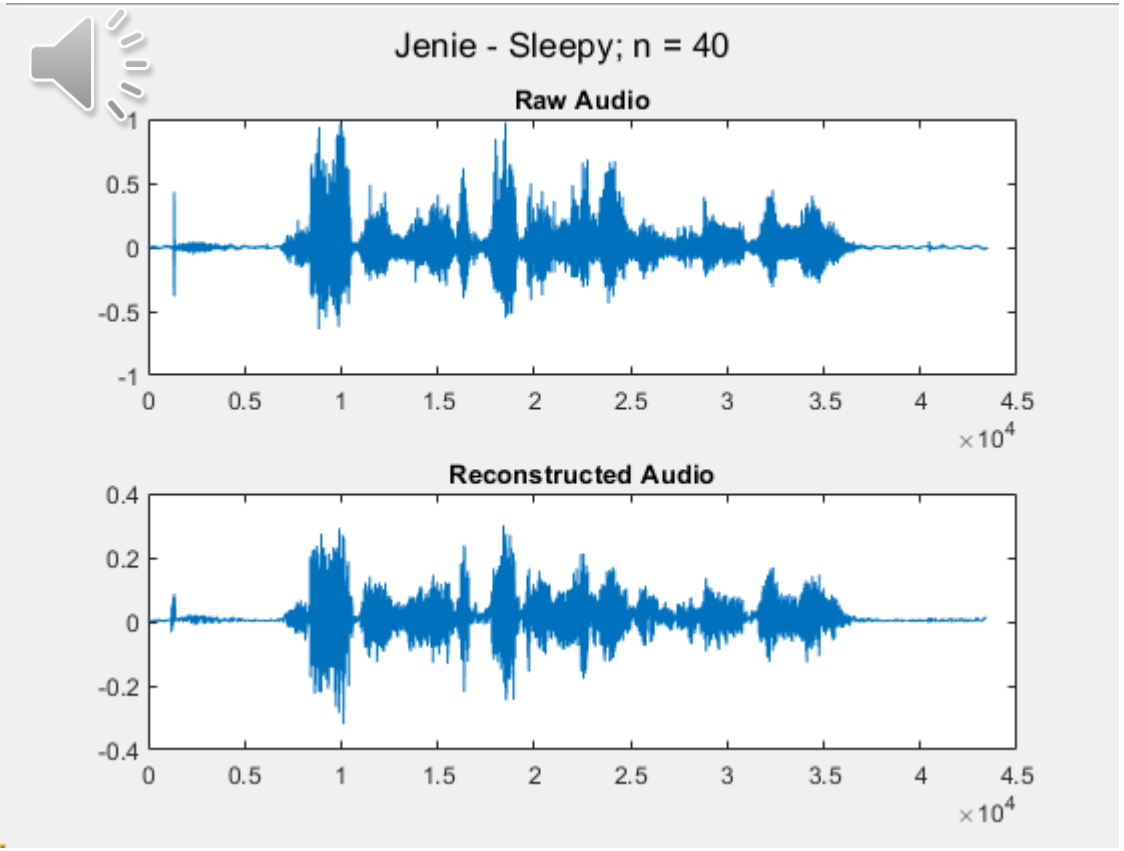
Significant increase in quality for both!

Coefficient Estimation – Re-visit Sleepy Jenie

- LPC n = 10
  - From Before



- LPC n = 40
  - Subjectively better, later portion



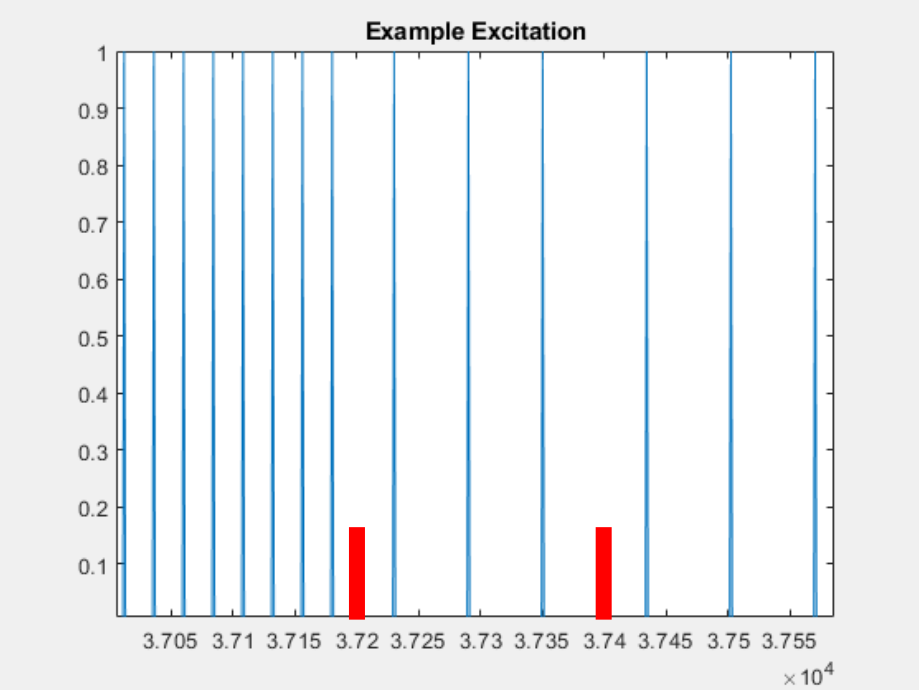
Improvements to LPC-10: Interpolation

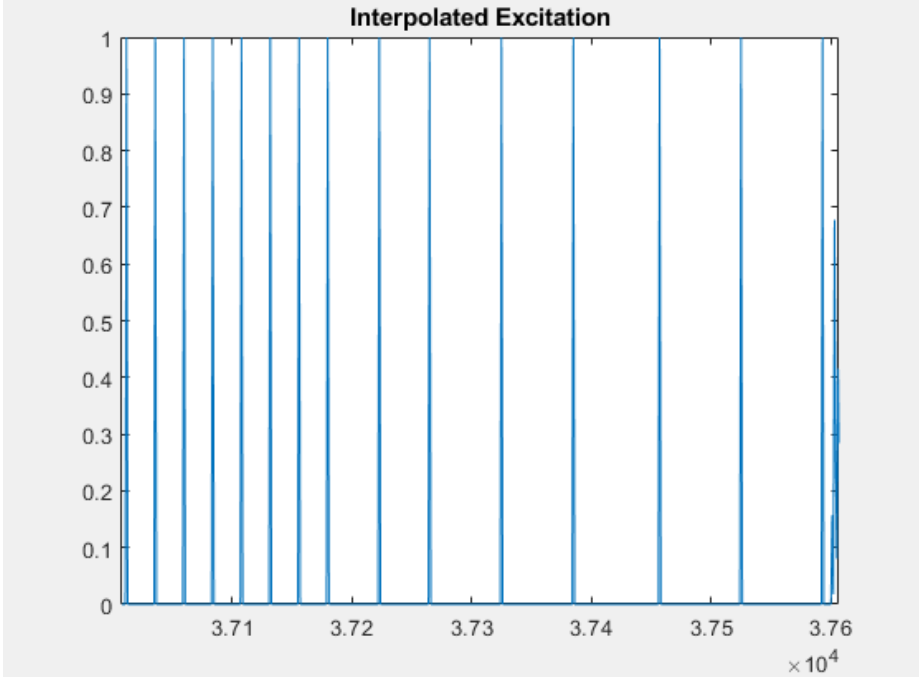
Attempted to improve audio quality by interpolating pitch and gain between windows

Pitch linear interpolation

- Pitch calculated on a window to window (200 samples)
- Idea: Linearly interpolate period from prior window
- Problem: Signal Period is Large Relative to windows
- Shift correction

Phase correction works great!
Pitch interpolation has negligible effect.

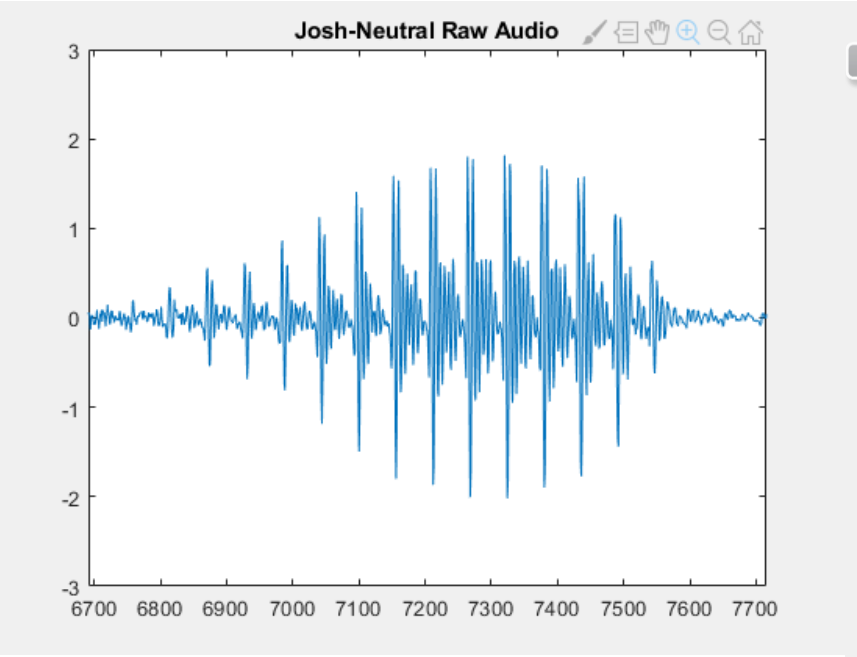


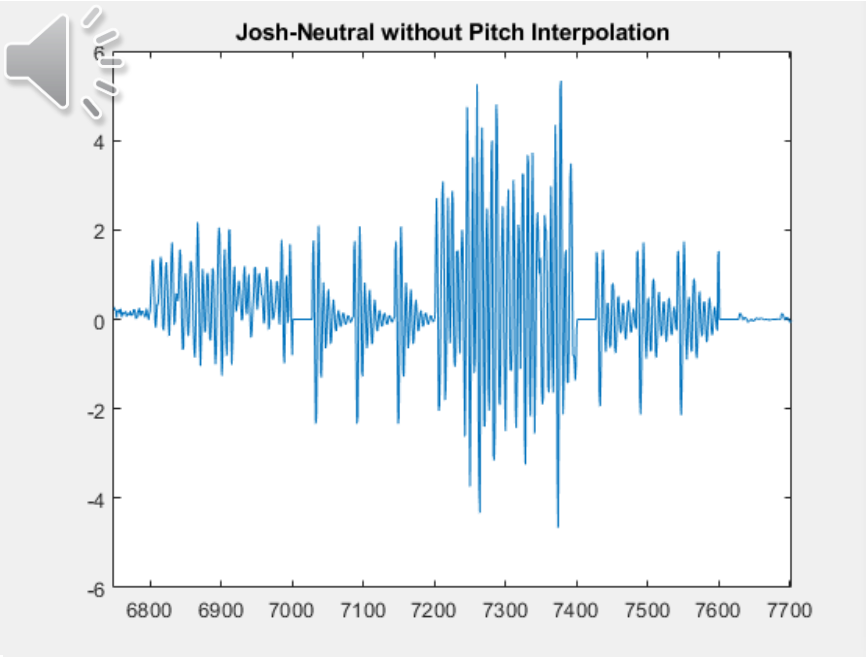


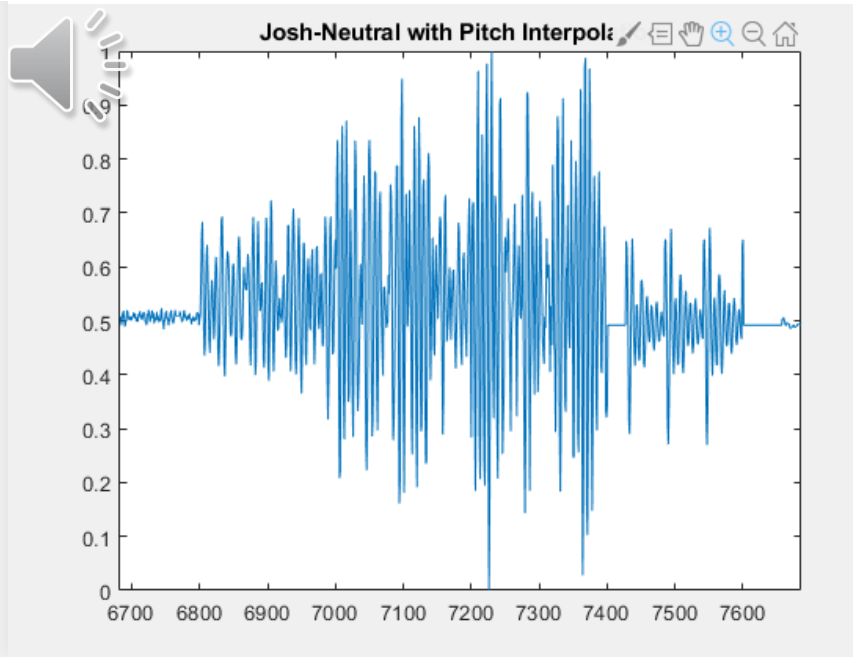
Pitch linear interpolation - Results

Also Tried on Sleepy Sam:
Without With

- Pitch interpolation – Clearer
- Shift correction reduced some artifacts
- Note: Past data has the shift correction.



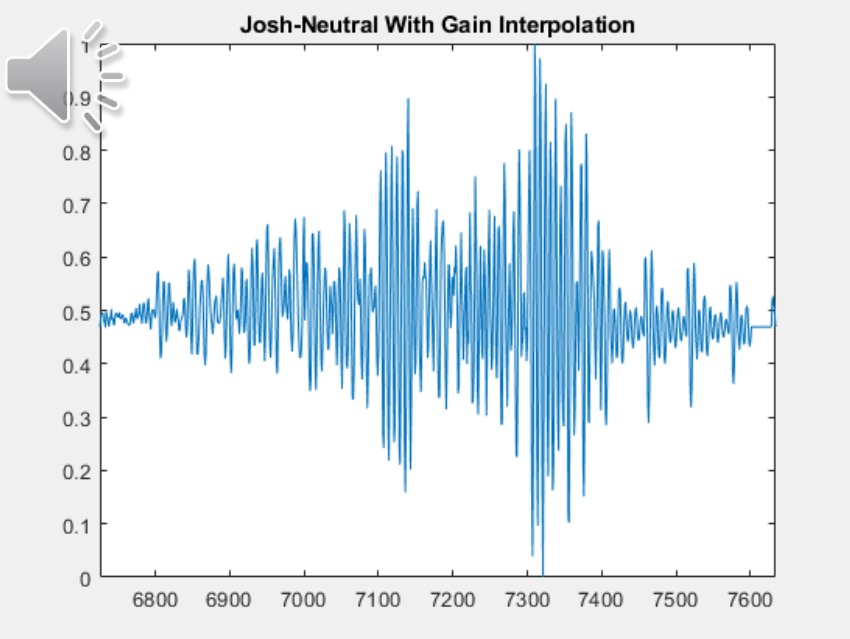


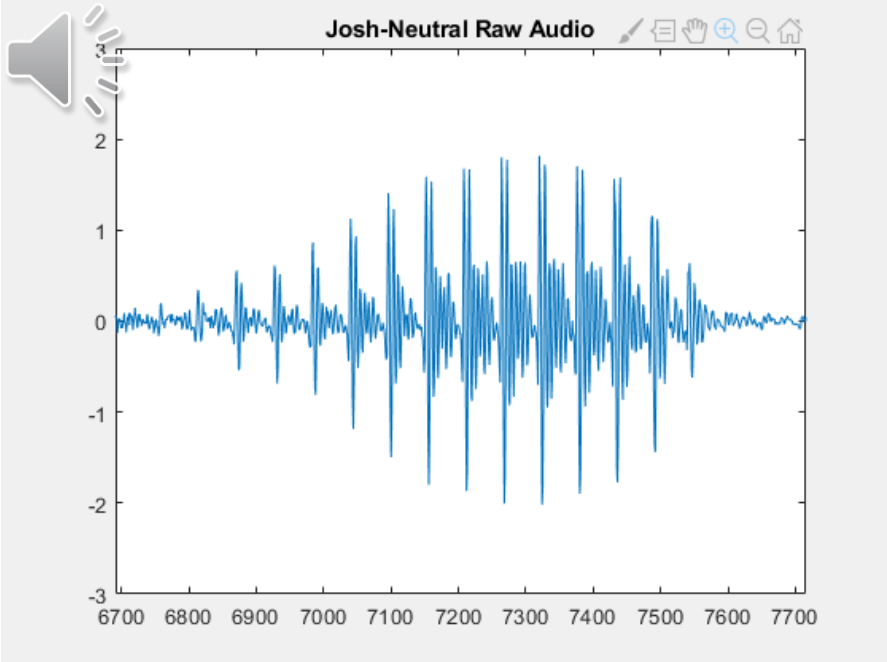


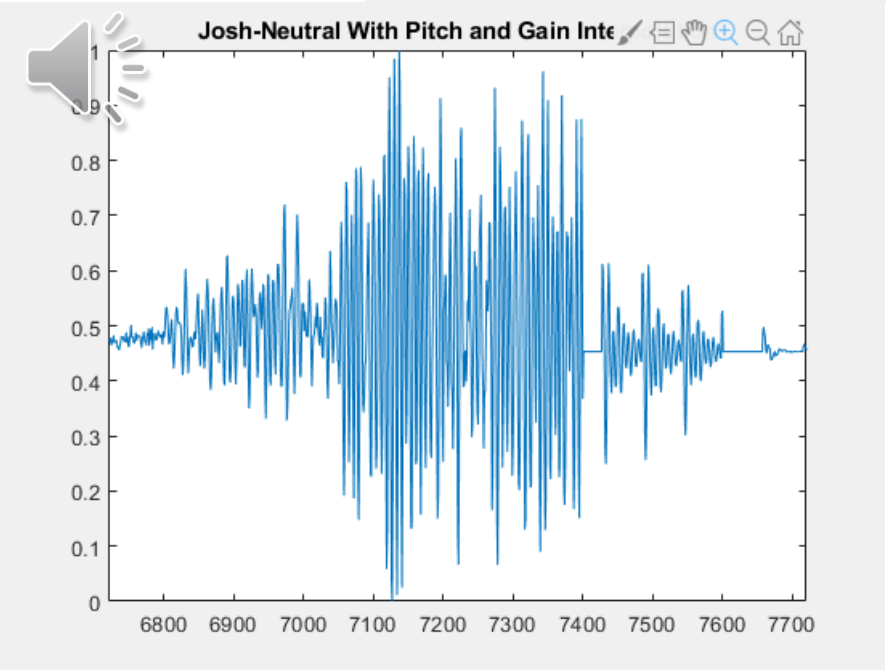
Pitch interpolation only:

Gain Interpolation

- Apply an envelope to signal
- Linearly Interpreting gains of each bin
- Reduces Artifacts







Gain interpolation slightly reduces high frequency artifacts.

Real Time LPC demo

- Compare different filter orders
- Has shift correction implemented

Work Cited

- Spanias, A. (1994). Speech Coding: A Tutorial Review. *Proceedings of the IEEE*, *82*(10), 1541-1582. https://doi.org/10.1109/5.326413
- *Kang, G., & Everett, S. (1985). Improvement of the excitation source in the narrow-band linear prediction vocoder. IEEE Trans. Acoust. Speech Signal Process., 33, 377-386.*