Optimized Phase-Preservative Speech Reconstruction and Enhancement

UNIVERSITY OF CENTRAL FLORIDA

Tobiah Bower
Electrical Engineering

College of Engineering and Computer Science

Abstract

The Signal Reconstruction Problem.

While Mel Spectrograms are powerful tools to visualize auditory spectral energy content, there is an unavoidable loss of phase information caused by the shortcomings of the Short-Time Fourier Transform (STFT). Many methods try to insert random stochastic phase information to retain speech quality; however, this research shows a recovery of the original phase information through optimal estimation.

Methodology

For a finite sequence $x[n]$ of digital speech, the STFT for frame t is defined as [5]:

$$X_t[\text{f}] = \sum_{n=0}^{N-1} x[n+tH]\omega[n]e^{-2\pi i f n/N}$$

The Spectrogram utilizes the Mel Scale to create the frequency bins [4]:

$$mel = 2595\log_{10}(1 + \frac{f}{700})$$

A one-shot Griffin-Lim Algorithm (GLA) uses alternating projections to exploit the STFT redundancy [1]:

$$X[k+1] = P_C(P_A(X[k]))$$

Iterative GLA

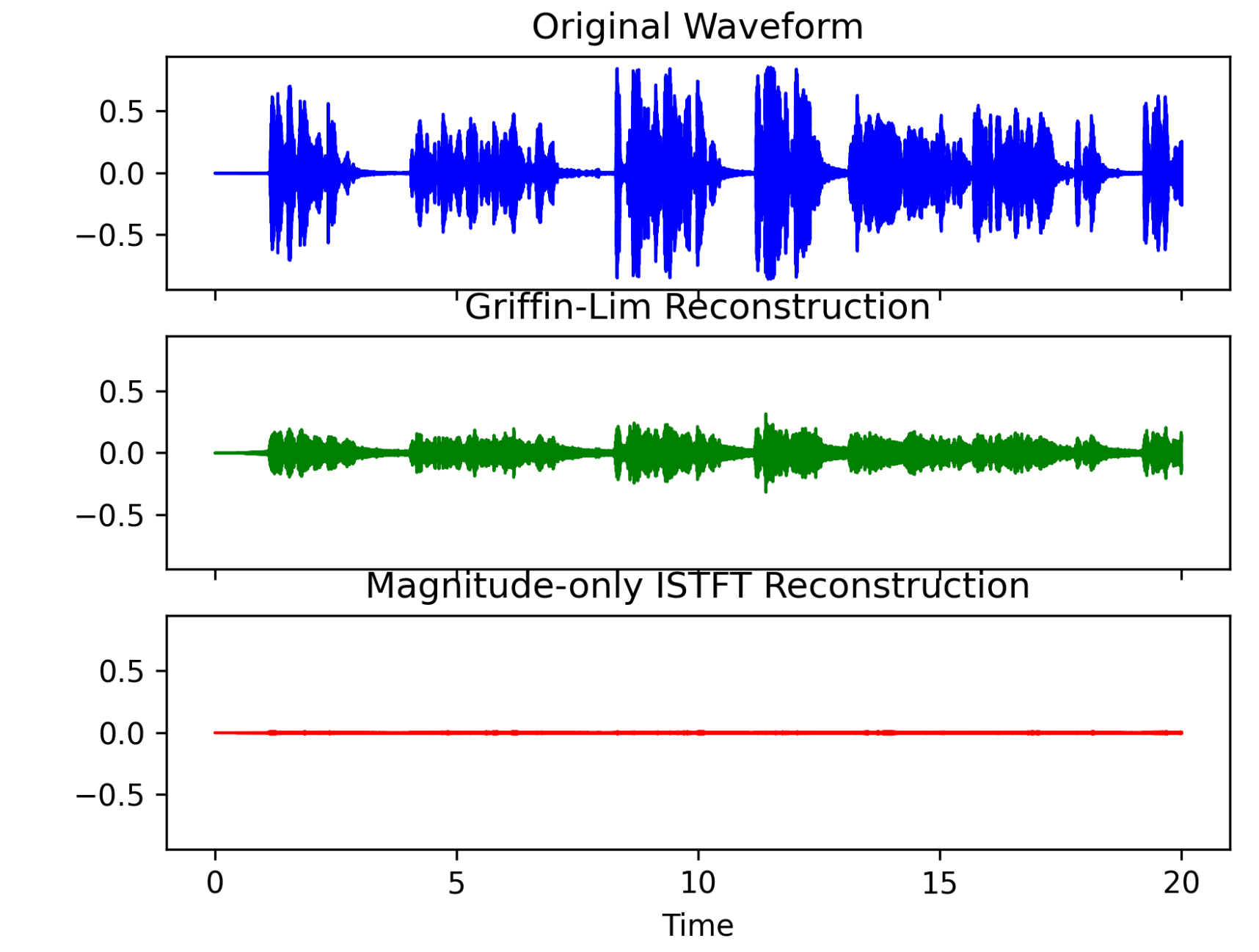
The breakthrough showcased in this presentation was obtained through an iterative implementation [2], which improved the quality of the reconstructed speech.

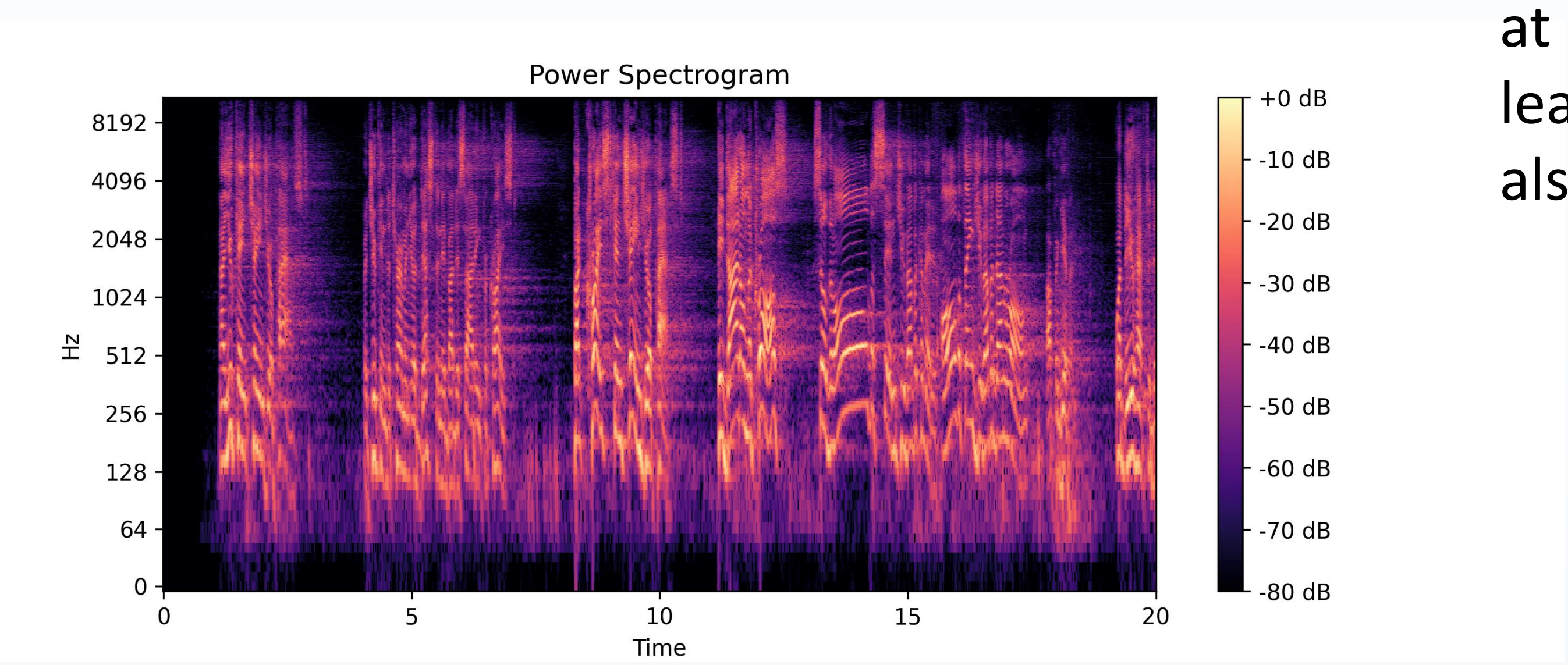
```
Fix initial phase $\angle c_0$
Initialize $c_0 = s \cdot e^{i\angle c_0}$
Iterate for $n = 1, 2, \ldots$
    $c_n = P_{C_1}(P_{C_2}(c_{n-1}))$
Until convergence
    $x^* = G^\dagger c_n$
```

Algorithms implemented in Python. There are other methods of de-phasing a waveform, but as with the inverse STFT, they are fallible.

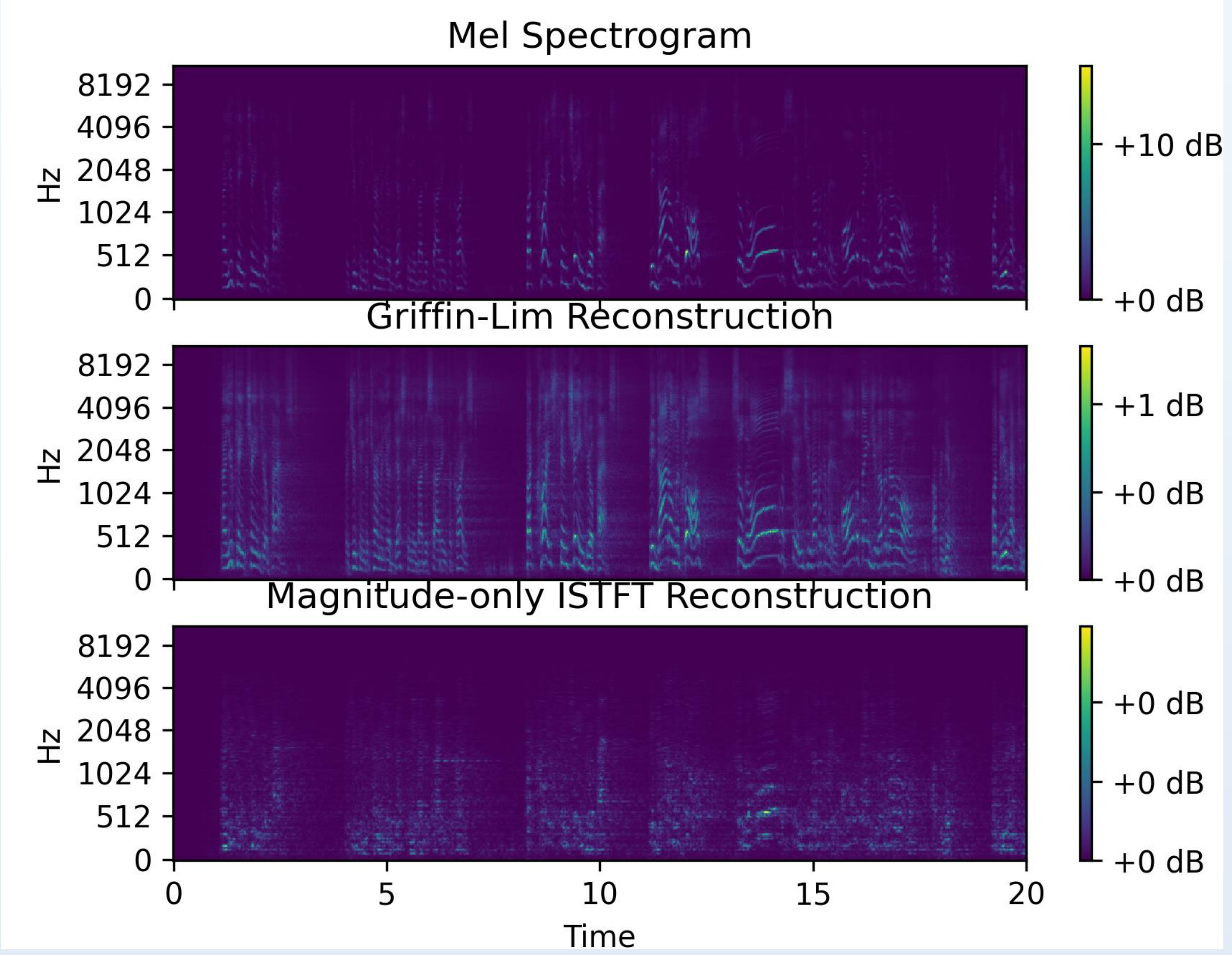
Results

For an input sequence of digital speech sampled at 22kHz, the one-shot GLA is an insufficient reconstruction method. For the iterative GLA, we use the tuning parameters $\alpha = 0.99$ and $\lambda = 0.01$. Additionally, we use 2048 FFT bins and 256 Mel bins.





The difference is audible and apparent through the comparison of spectrograms. The one-shot method results in phasing problems that make it distracting to listen. The iterative method de-phases and results in a more pleasant waveform as verified by the perceptual evaluation of speech quality (PESQ):



Male wideband

Algorithm	PESQ
One-Shot GLA	2.816
Iterative GLA	3.615

Female narrowband

Algorithm	PESQ
One-Shot GLA	1.609
Iterative GLA	1.896

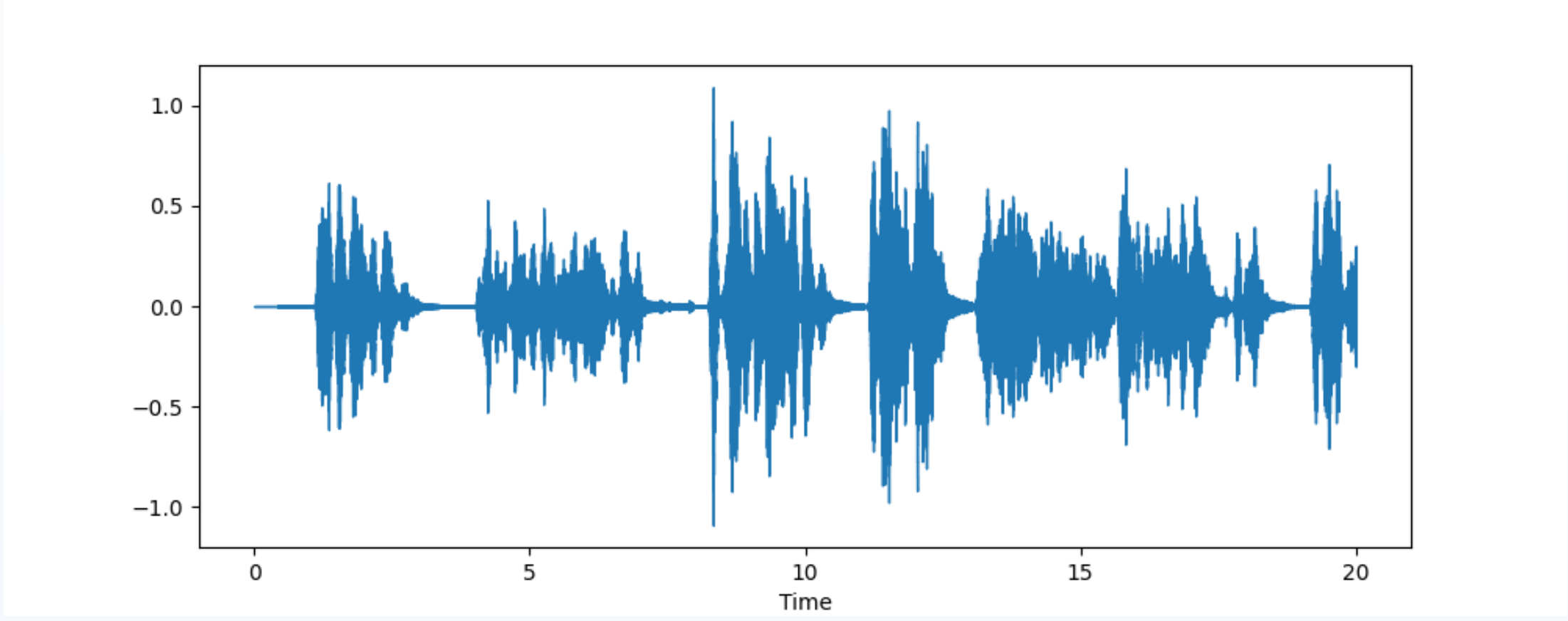
The iterative approach works well at low sampling frequencies, making it advantageous for its low energy. This is verified by poorer performance when using a higher frequency input waveform. There is a need for more powerful algorithms to address this drop off in performance.

Speech Enhancement (Future Work)

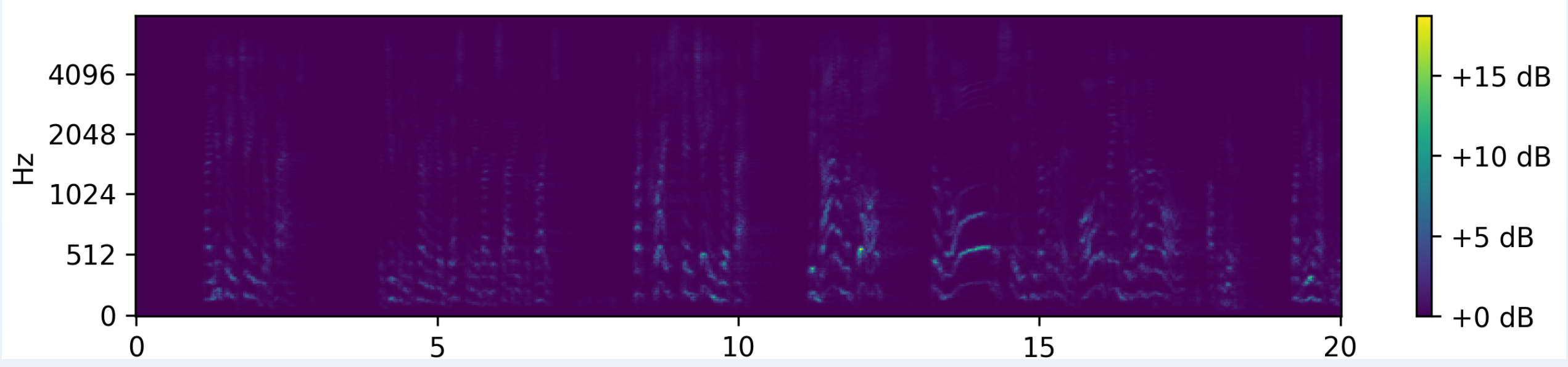
The Kalman filter can be used to remove artifacts and noise from the output waveform, pictured in the top right. It can also intelligently fuse speech waveforms, which may be different versions of the same source. This is a possible remedy to the performance gaps of GLA.

Discussion

As we see, the iterative GLA results in not only a higher fidelity waveform, but the phase is considerably more accurate to the original speech. It is important to note that at this complexity, higher sampling frequency may actually lead to poorer algorithm performance. More iterations may also reduce the quality of the PESQ scores



Additionally, we see the similarity in the Mel Spectrogram, as artifacts and noise have been removed. Further work will be done to refine the wideband performance for higher frequency vocals, and to allow for CNN learning to produce an output without access to the original speech.



In future studies, we will examine computational benefits of local GLA runs on multi-recordings of same source fused with Kalman speech enhancement, with end-to-end refined iterative GLA. This methodology is relevant to the speech to text effort, used by translation and home assistants.

References

[1] Masuyama et al. "Signal reconstruction from Mel-spectrogram based on bi-level consistency of full-band magnitude and phase“

[2] Perraudin et al. “A Fast Griffin-lim Algorithm,” 2013.

[3] Orchisama Das, “Kalman Filter in Speech Enhancement,” 2016.

[4] Natsiou, O’Leary, ‘A Sinusoidal signal reconstruction method for the inversion of the mel-spectrogram’

[5] Bruce Sharpe, “Invertibility of overlap-add processing”

Acknowledgements

Dr. Wasfy B. Mikhael inspired me with this project idea, and I have learned so much under his teaching and mentorship.

Dr. Chinwendu Enyioha has helped me make more sense of all the concepts in optimization and estimation.