Department of Electronics and Electrical Communication Engineering, IIT Kharagpur

Experiment 4: Speech Recognition with Primarily Temporal Cues

Samyak Sheersh, Anubhav Mitra

Roll Numbers: 22EC30045, 22EC30007
Group Number: 24

2 October 2024

1 Objectives

1. Generation of multiple frequency bands from the primary signal

2. Processing the bands via Hilbert Transform for envelope detection, passing the FFT through an appropriate low pass filter and finally adding noise of the corresponding frequency band to the signal

3. Qualitative comparison of audio files before and after processing to gauge intelligibility.

2 Procedure

2.1 Generation of bands

We generate bands on a logarithmic scale. The number of bands is predicted to have a correlation with the intelligibility of the output, we experiment by varying this figure from 1 to 16, in a range from 90Hz to 5.76kHz. Notably, a single band will have one filter of range 90Hz to 5.76Hz, two bands will have ranges of 90Hz to 720Hz and 720Hz to 5.76kHz.

2.2 Processing

We apply a Hilbert Transform to the outputs in their individual bands and take its magnitude to obtain the envelope of the resultant waveforms in their respective bands, since the Hilbert transform itself in the time domain can be conceptualized as the sum of a sine and a cosine component with identical arguments resulting in a rotating phasor. We then apply a further low-pass filter of 240Hz, this time on the frequency domain itself to smoothen out any irregularities present. Finally before recombination, we add white noise whose band corresponds to the band we are working with.

2.3 Qualitative Analysis

The various resultant outputs, sorted by the number of their bands was then checked for clarity and compared.

3 Results

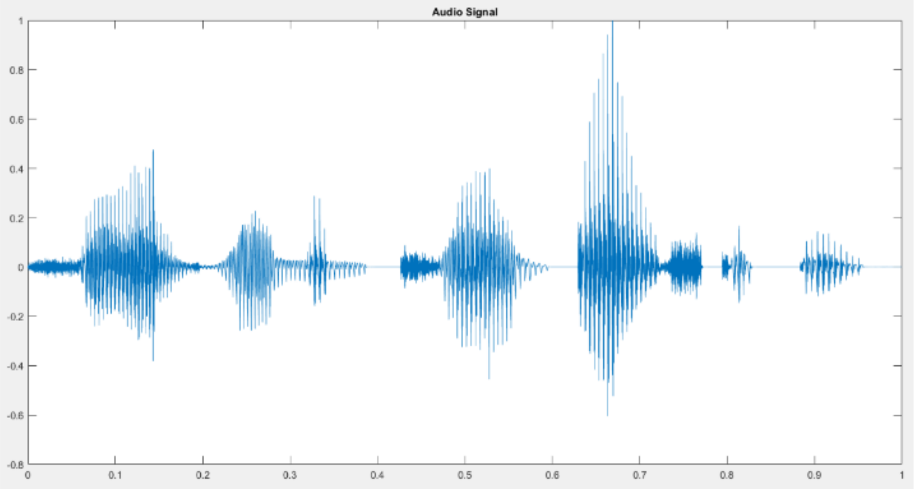

Figure 1: Time plot of the original signal

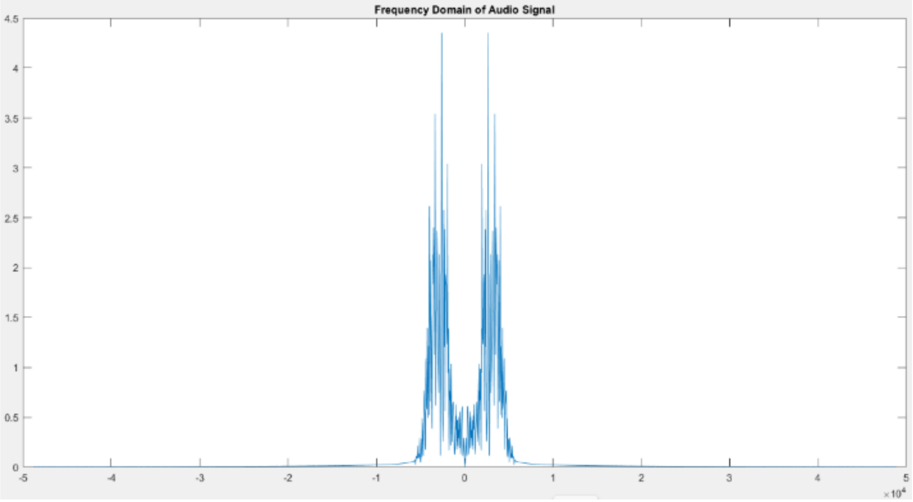

Figure 2: Fourier Transform of the original signal

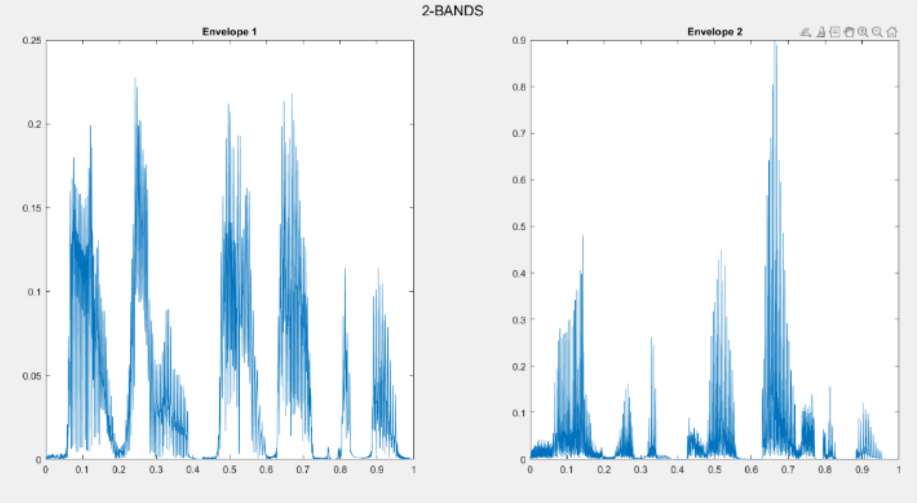

Figure 3: Time representation of the separated two bands after processing

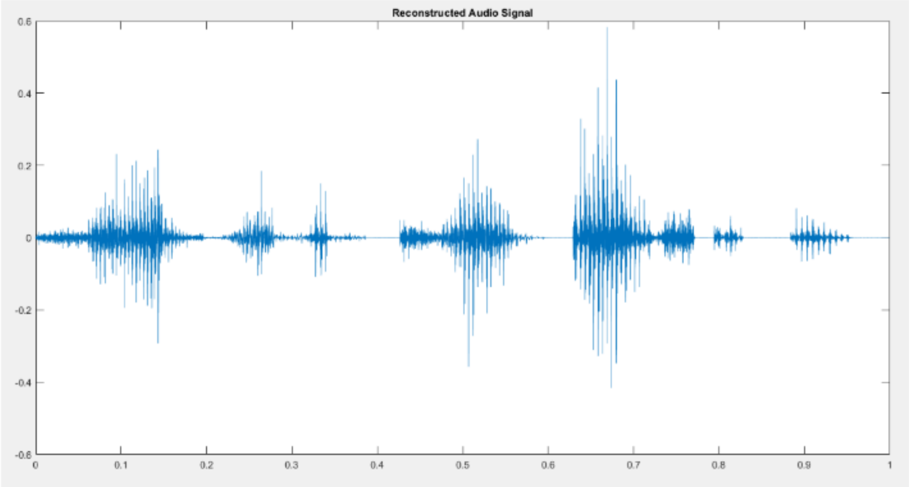

Figure 4: Time representation of the reconstructed signal from those two bands

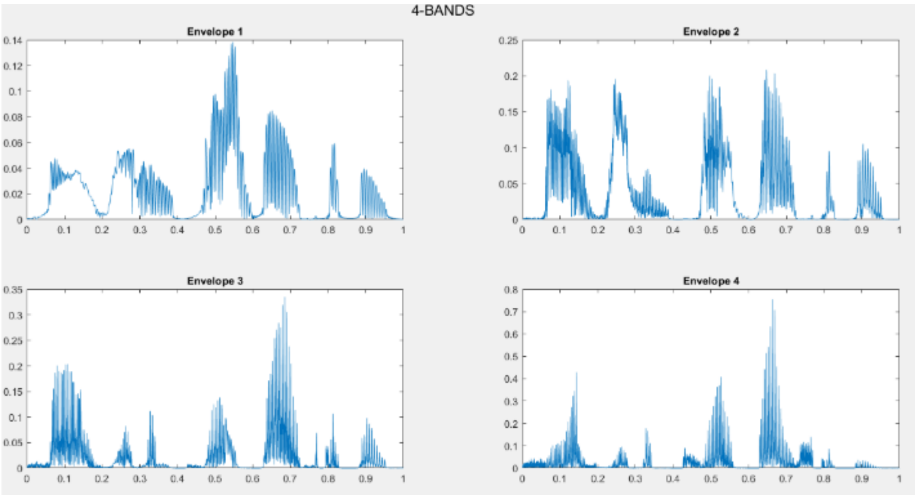

Figure 5: Time representation of the separated 4 bands post processing

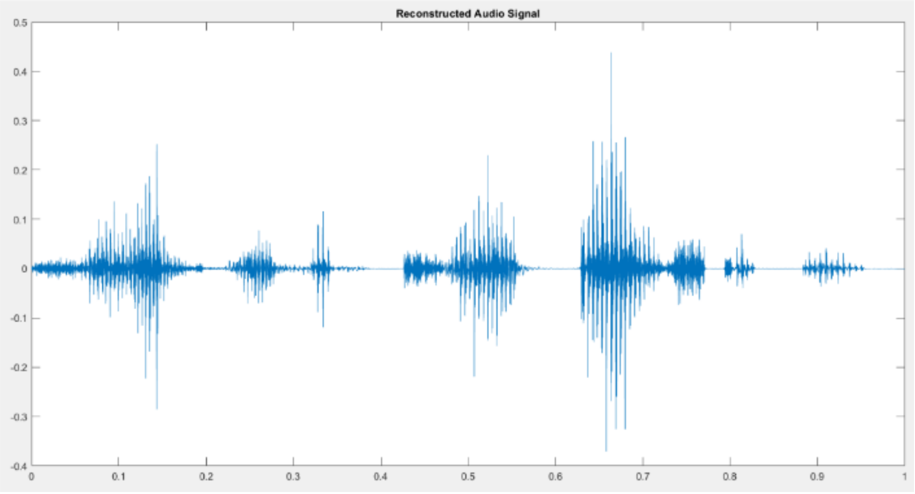

Figure 6: Time representation of the reconstructed signal from those 4 bands

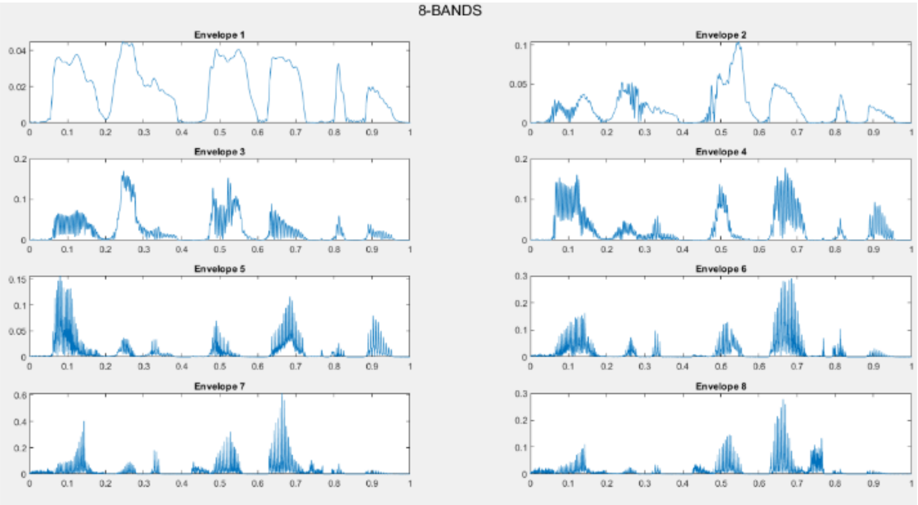

Figure 7: Time representation of the separated 8 bands post processing

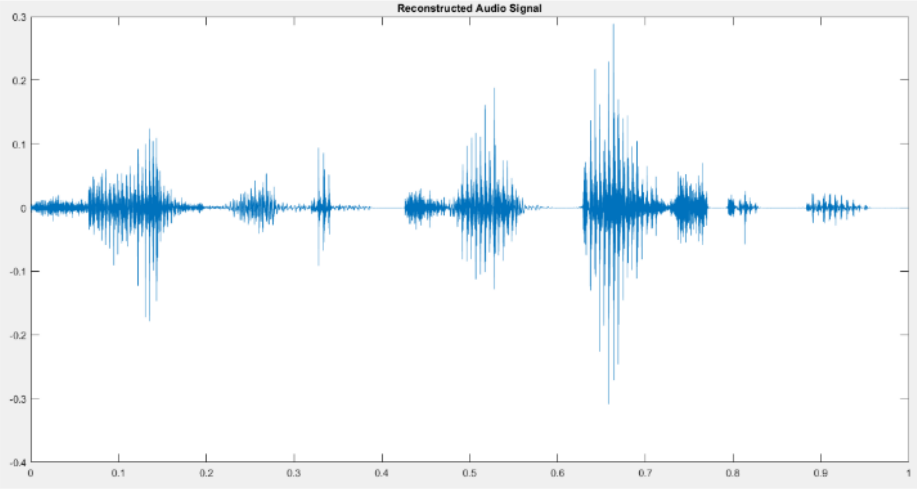

Figure 8: Time representation of the reconstructed signal from those 8 bands

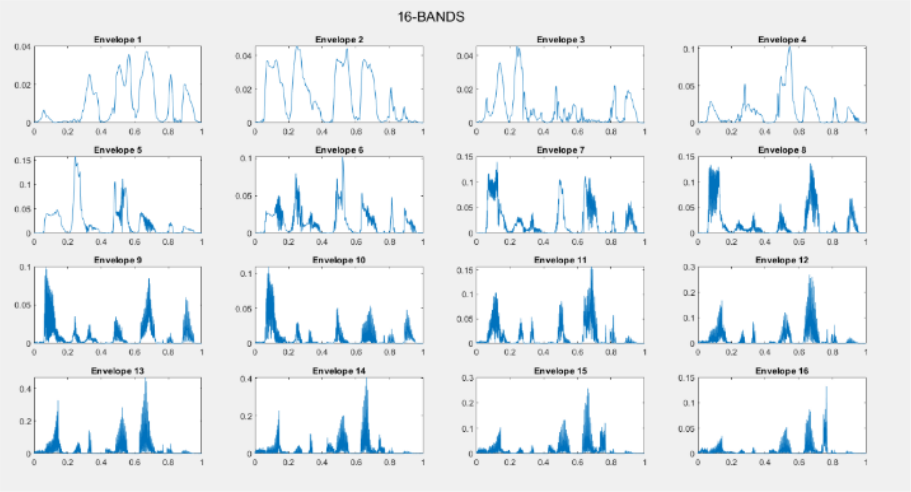

Figure 9: Time representation of the separated 16 bands post processing

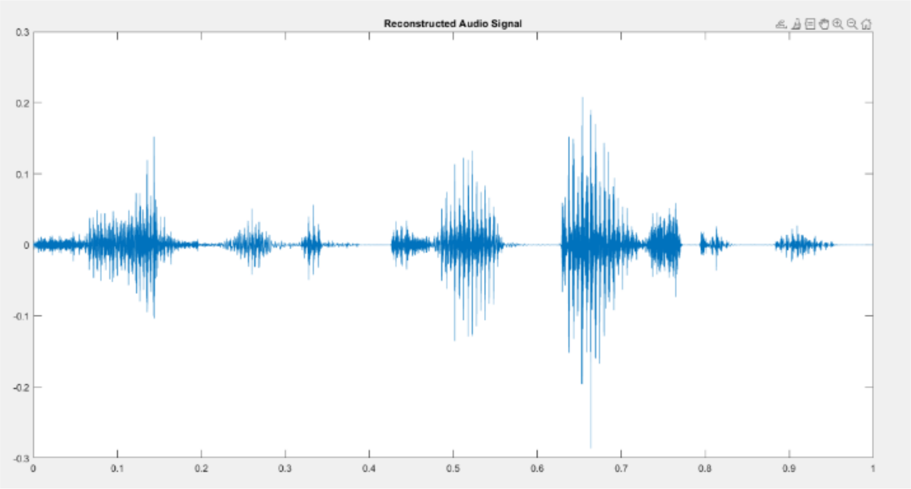

Figure 10: Time representation of the reconstructed signal from those 16 bands

4 Discussion: Samyak Sheersh

1. From the qualitative response (i.e. what we actually heard), we see that as we increase the number of bands, the signal is easier to make sense of. This is because taking a larger number of bands allows us to capture the microstructure of the higher frequencies, and thus we need to reject less actual information and thus the signal is clearer.

2. However, the fact that we can still make out what is being said points to the fact that the envelope of these sounds is what matters to be intelligible rather than the actual soundwave itself

3. We used the Hilbert transform to get the envelope for our signal. This is because Hilbert transform (multiplication with $-j.sgn(f)$ in the frequency domain) allows us to obtain what is called an *analytic signal*:

   $$m_a(t) = m(t) + jm_h(t) \tag{1}$$

   which when we take the modulus, comes out be reasonably accurate since many of the higher order terms become miniscule.

4. We subsequently multiply this envelope by additive White Gaussian Noise to mimic the natural signal representation of human voice, providing the necessary texture without needing to store the microstructure of the original signal.

5. We finally conclude that parallel processing through multiple bands allows us to reproduce human speech in a much more memory efficient manner.