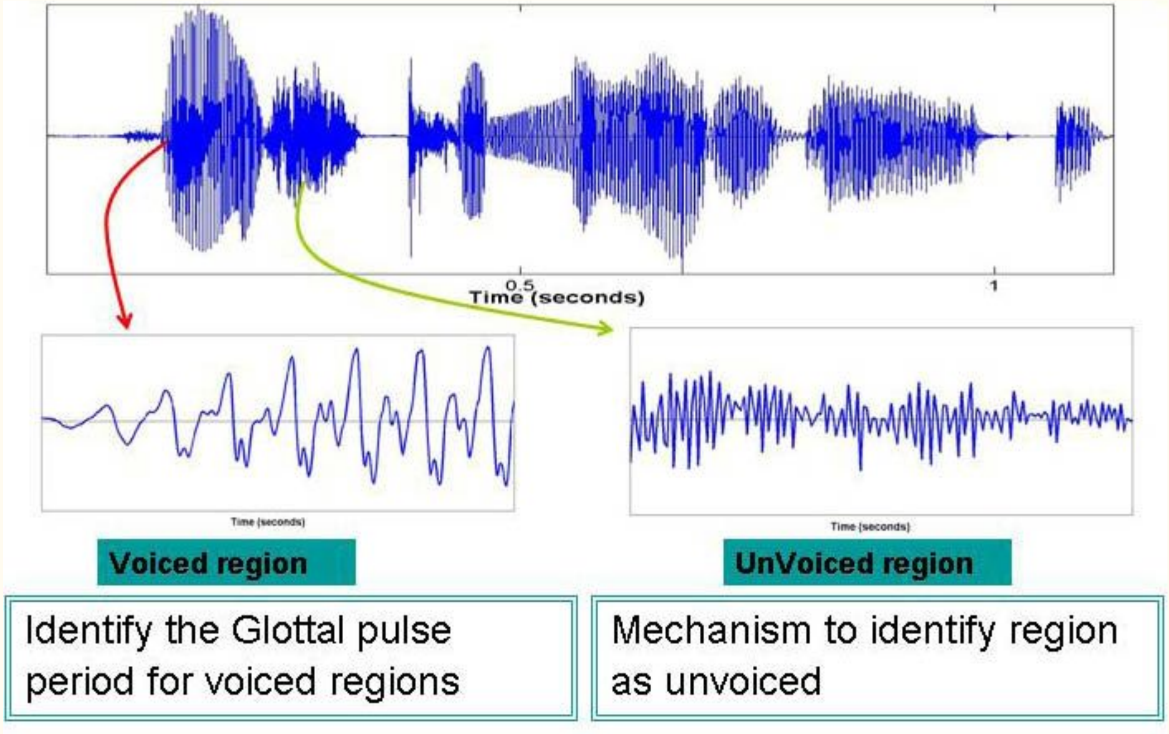
VOICED/UNVOICED DETECTION OF SPEECH SIGNAL USING EMPIRICAL DECOMPOSITION MODEL

Gowri L
Shivani C

What is voiced and unvoiced ?

Voiced signals : produced when the vocal cords vibrate during the pronunciation of phonemes.

Unvoiced signals: Signal which are not voiced, noise, silence etc.



Voiced region

Identify the Glottal pulse period for voiced regions

UnVoiced region

Mechanism to identify region as unvoiced

Need of Separation of Voiced and Unvoiced:

- For pre-processing; Most of the information is in Voiced part.

Different methods for separation:

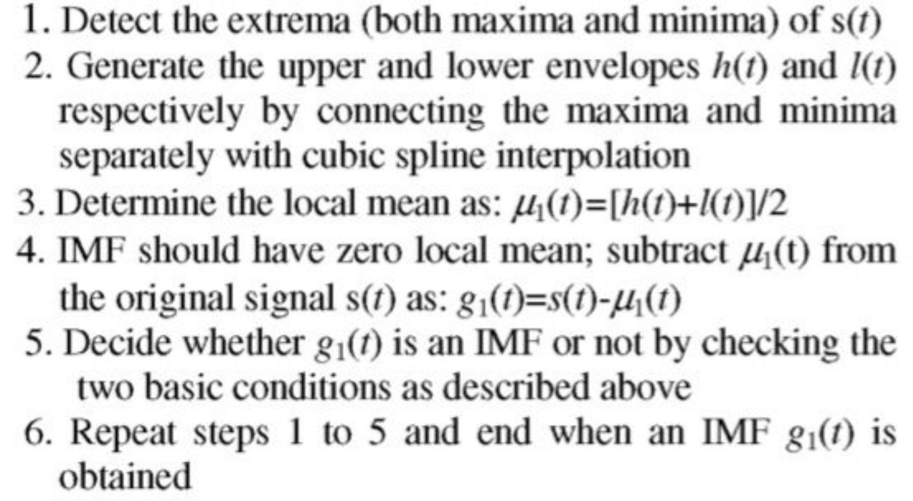
- Energy of the sequence: Computes the energy of each time frame and sets a threshold above which the frames are marked as voiced. Unvoiced part in a speech signal can sometimes have comparable frequency or energy component similar to voiced parts.
- Similarly, Use of autocorrelation/variance contour and a threshold etc. methods. Silenced parts are highly auto correlated like voiced parts(harmonics and periodicity).
- If the threshold(energy) is set very low, both the voiced and unvoiced parts will be marked as voiced.
- Due to changes in pitch/frequency/amplitude the threshold is not robust in the above methods.

Empirical Mode Decomposition

- Signal is broken down into bandlimited components called intrinsic mode functions (IMFs).IMF is any function with the same number of extrema and zero crossings, whose envelopes are symmetric with respect to zero.
- Without leaving the time domain, EMD is adaptive and highly efficient.
- Since the decomposition is based on the local characteristic time scale of the data, it can be applied to nonlinear and nonstationary processes.

- EMD is generally used to decompose a signal into its harmonics

IMF Extraction - Sifting:



1. Detect the extrema (both maxima and minima) of s(t)
2. Generate the upper and lower envelopes $h(t)$ and $l(t)$ respectively by connecting the maxima and minima separately with cubic spline interpolation
3. Determine the local mean as: $\mu_1(t)=[h(t)+l(t)]/2$
4. IMF should have zero local mean; subtract μ_1(t) from the original signal s(t) as: $g_1(t)=s(t)-\mu_1(t)$
5. Decide whether $g_1(t)$ is an IMF or not by checking the two basic conditions as described above
6. Repeat steps 1 to 5 and end when an IMF $g_1(t)$ is obtained

EMD:

$$s(t) = \sum_{m=1}^{M} C_m(t) + r_M(t)$$

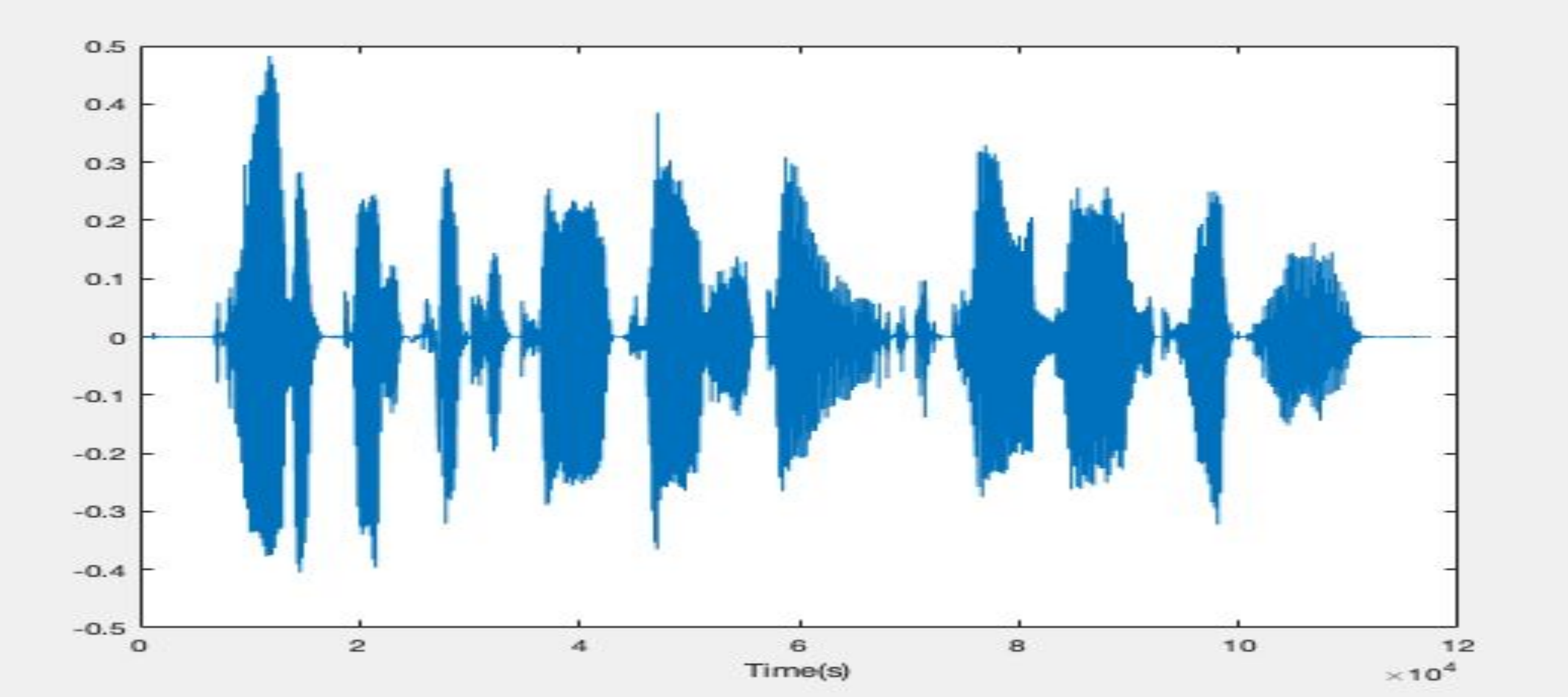
r_1(t)=s(t)-C_1(t);

r_2(t)=r_1(t)-C_2(t) ;...; r_m(t)=r_m-1(t)- C_m(t);

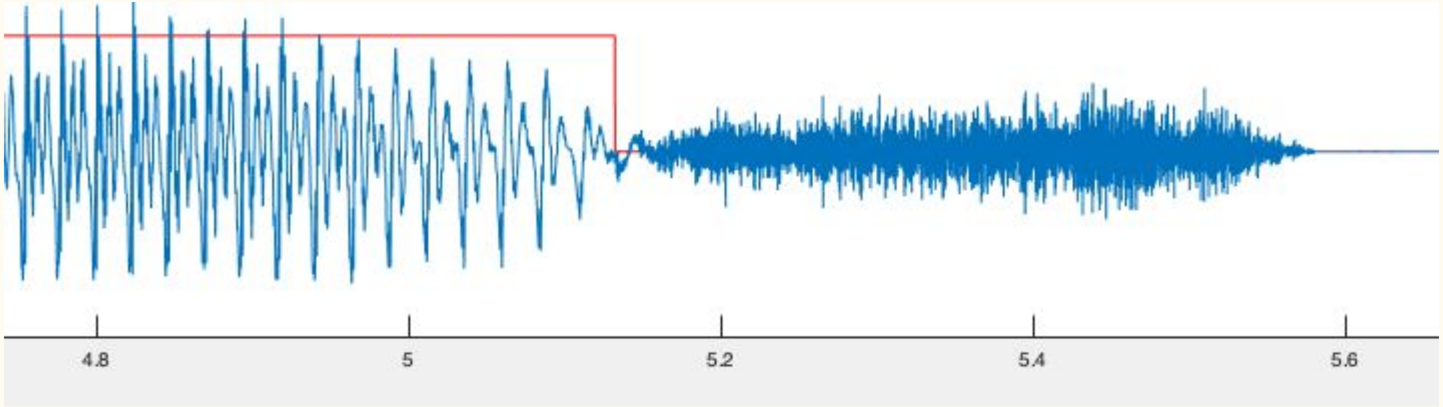
Where s(t) : speech signal, r_m : mth residue function, C_m :mth order IMF

- Autocorrelation functions of Higher frequency components(Lower order IMFs) are neglected and the last IMF’s Autocorrelation function is considered as the oscillation with fundamental frequency which is of a damping cosine form.
- This function’s mean fractional energy is taken as the threshold above which the frames are labeled “voiced”.

RESULTS

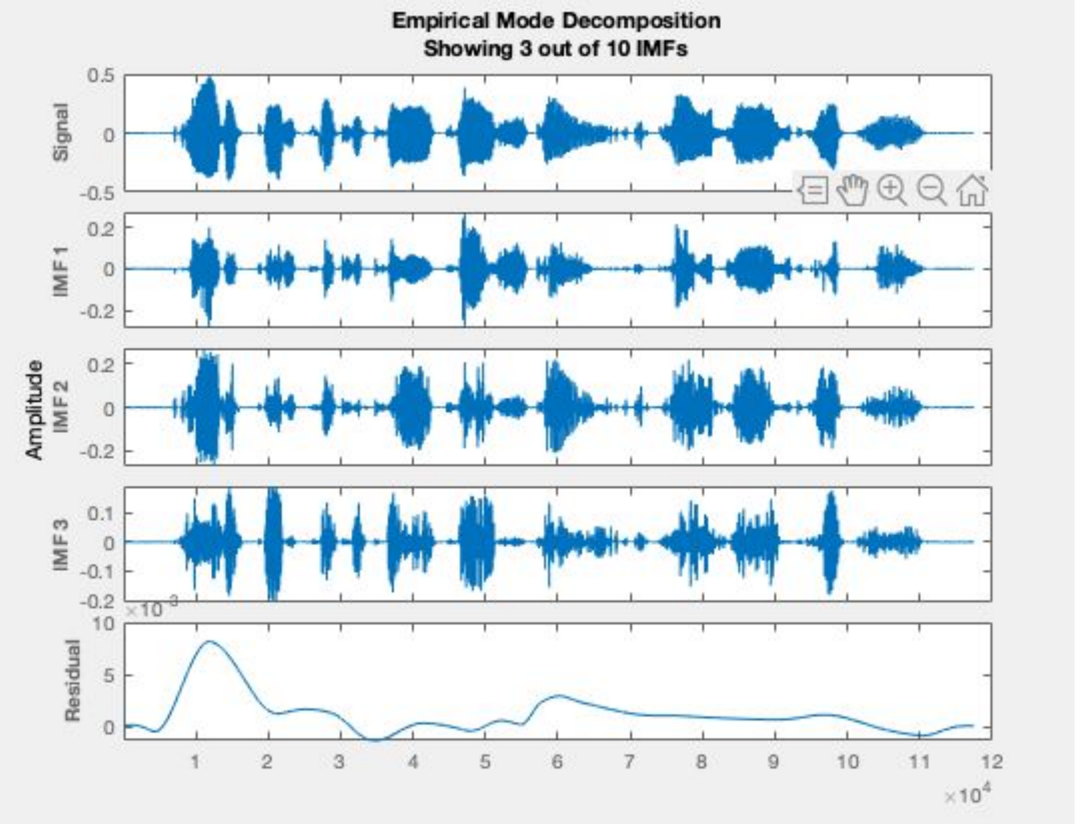


Speech signal

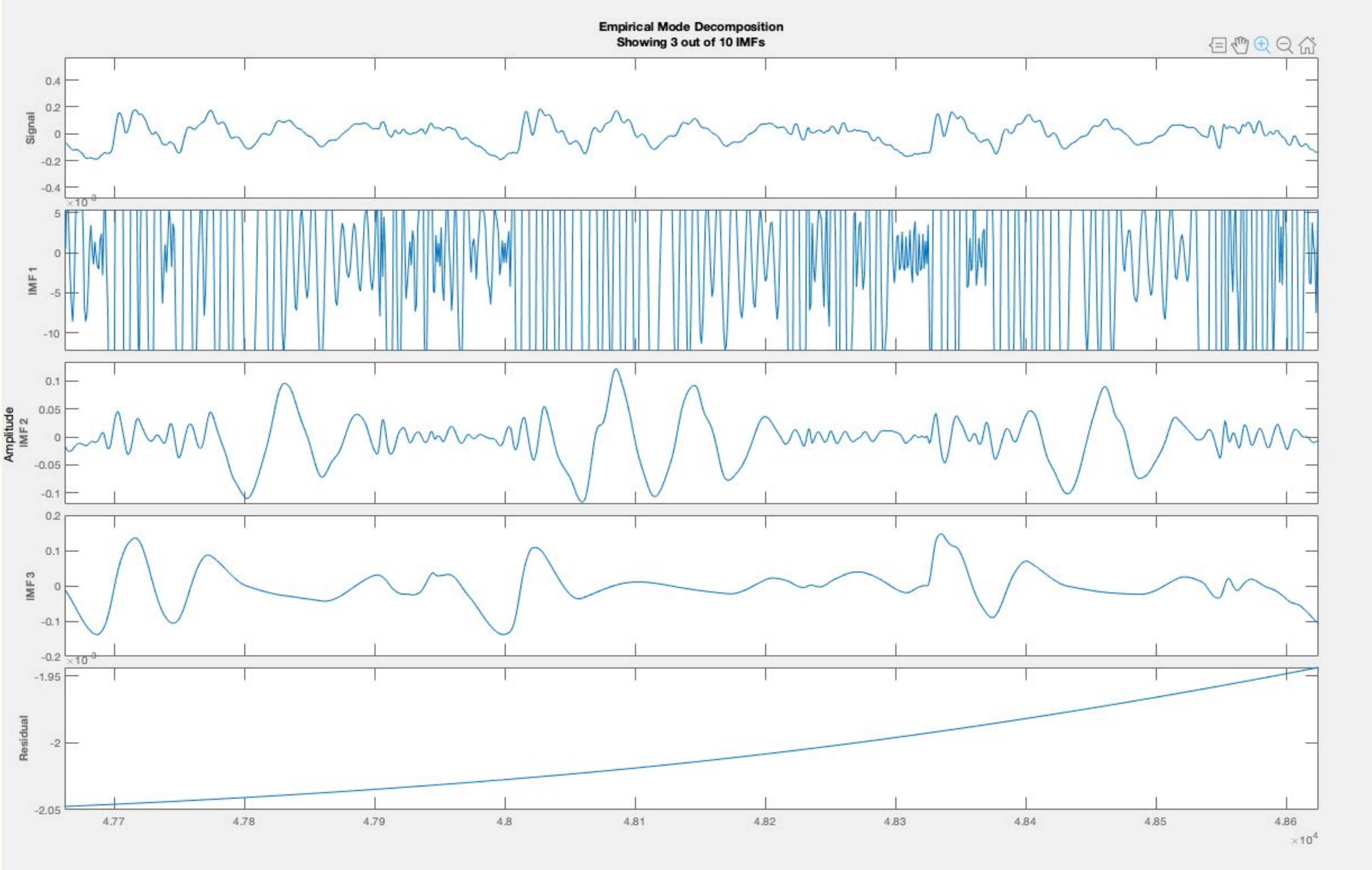


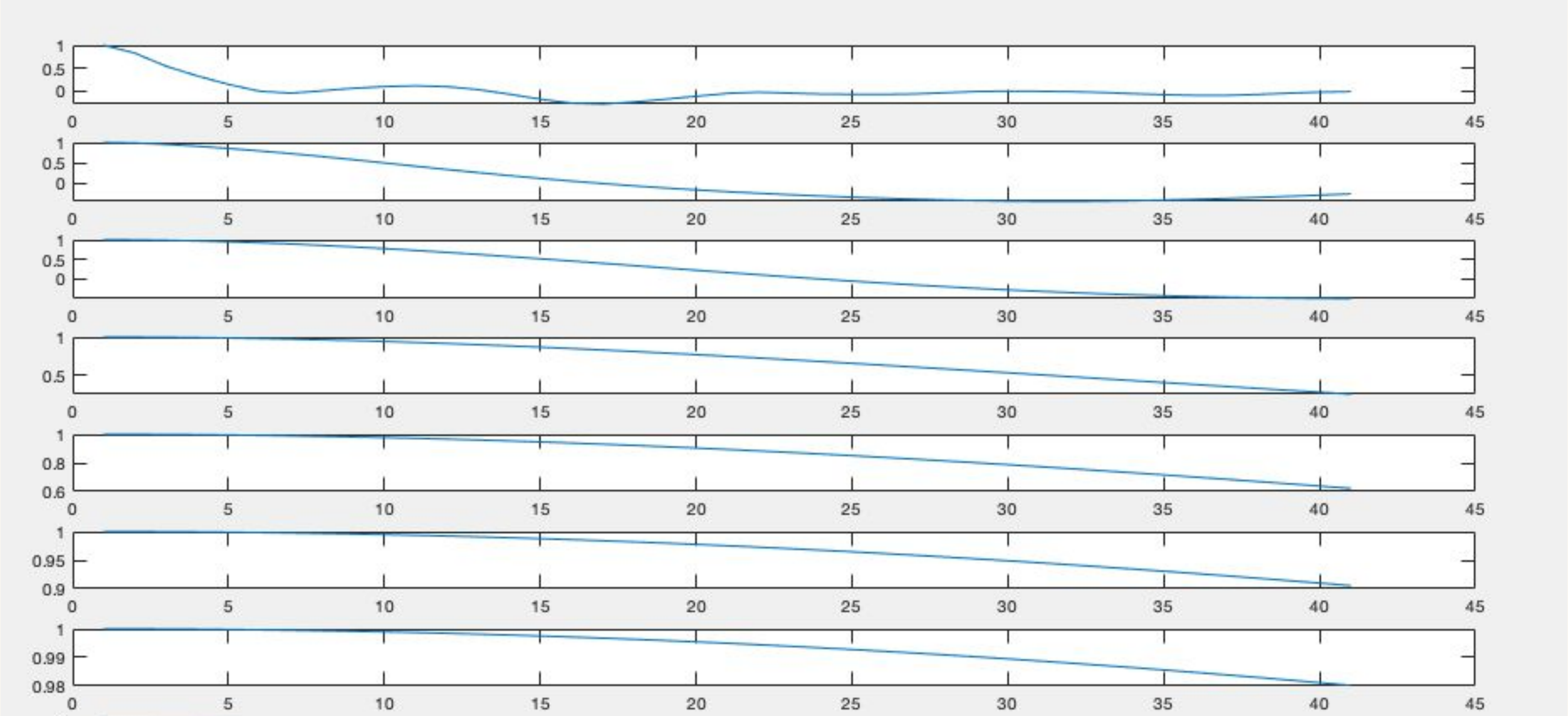
Separation of Voiced and Unvoiced using a threshold

- By using EMD, we find the IMFs
- After applying autocorrelation on IMFs, we find the fractional energy
- Find the decision factor/threshold

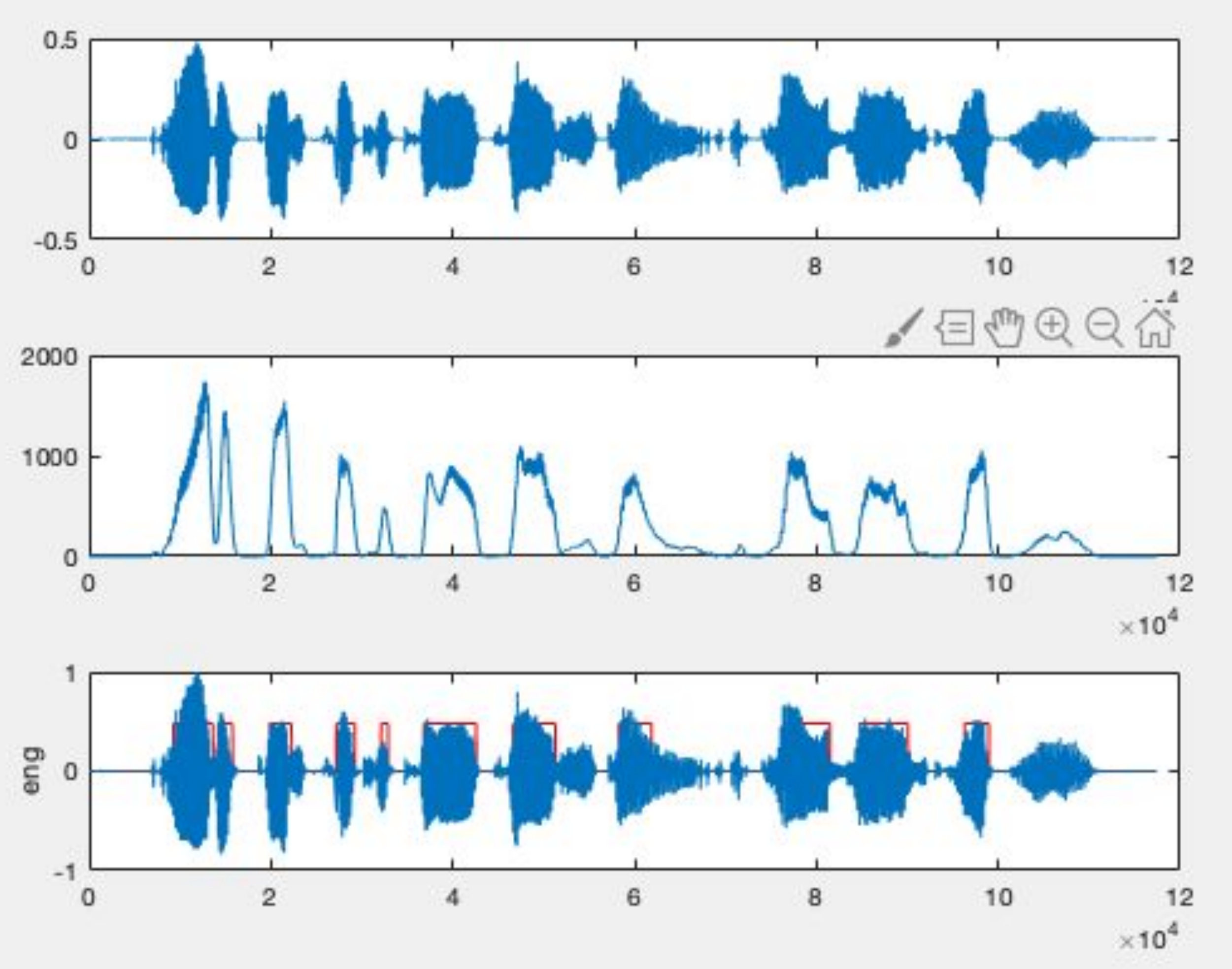


We only use IMFs in the fundamental frequency (discarding the lower order IMFs) to decide the threshold


Empirical Mode Decomposition
Showing 3 out of 10 IMFs
Signal
0.4
0.2
0
-0.2
-0.4
×10⁻³
IMF 1
5
0
-5
-10
Amplitude
IMF 2
0.1
0.05
0
-0.05
-0.1
IMF 3
0.2
0.1
0
-0.1
-0.2
×10⁻³
Residual
-1.95
-2
-2.05
4.77
4.78
4.79
4.8
4.81
4.82
4.83
4.84
4.85
4.86
×10⁴

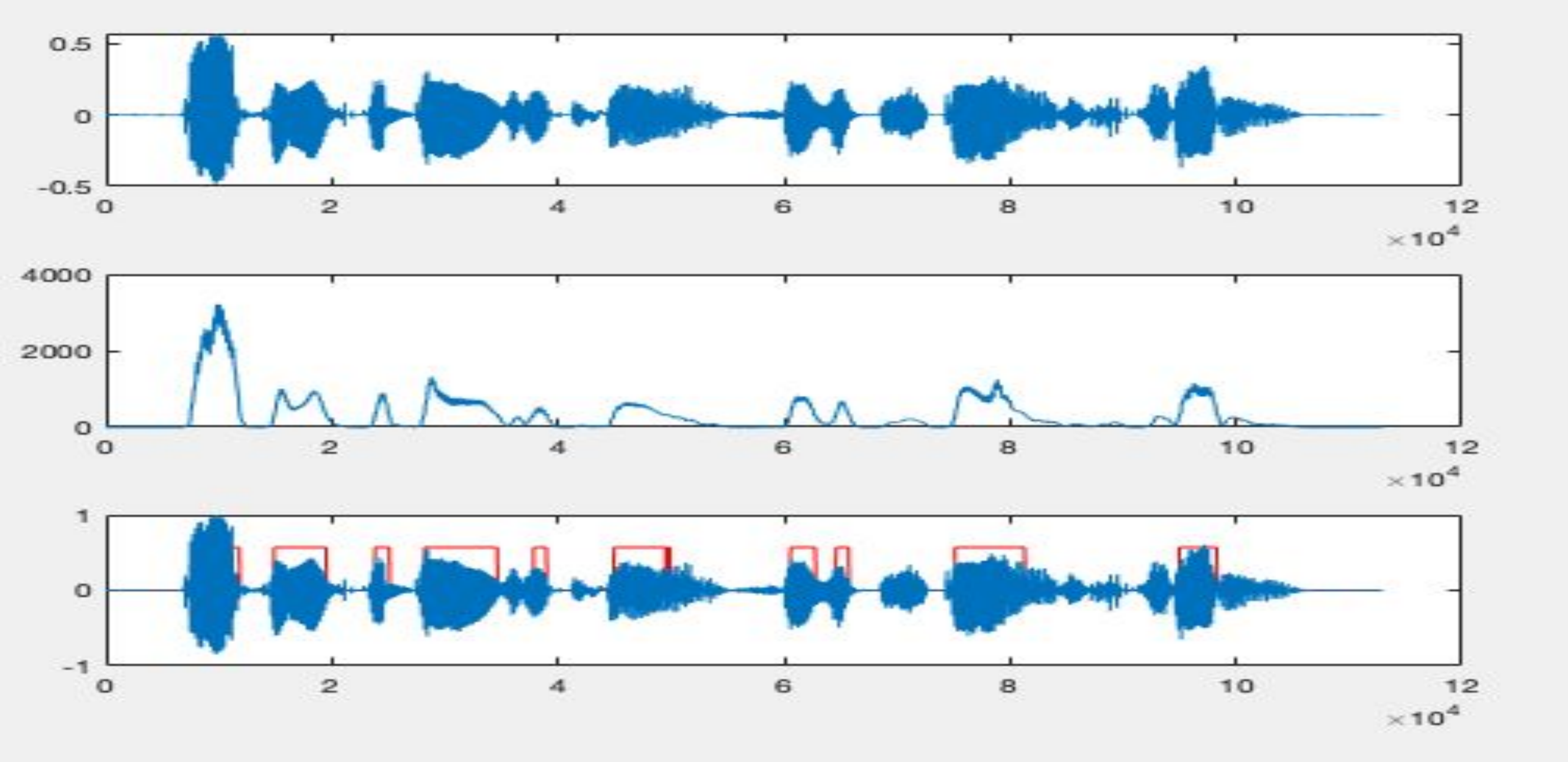


Autocorrelation Functions



From the IMF of the signal, we plot the energy contours and then using frames decide that:

- those above the calculated threshold are “voiced”
- below the threshold are “unvoiced”


0.5
0
-0.5
0
2
4
6
8
10
12
×10^4
4000
2000
0
0
2
4
6
8
10
12
×10^4
1
0
-1
0
2
4
6
8
10
12
×10^4

Thank You