Estimation of Instantaneous Pitch Frequency in Speech Signals

Team Curiosity

Chepuri Shivani 2018122004

D Sai Siddarth 2018122006

Contents and Workflow

- **Objectives**
- **About Pitch Estimation and its importance**
- **Abstract of the Main paper**
- **Algorithm outline**
- **Methodology**
- **More technical details**
- **Results and Plots**
- **Comparisons**

New (final eval) Video presentation Link:
https://drive.google.com/file/d/172FMeHiIYv3QkISM3NPGyAMbdFpBuUf1/view?usp=sharing

Overall Objectives

1. To Study and Explore the Algorithm proposed in the Main paper and cited papers for instantaneous pitch estimation.
2. To explore and implement VMD and VMD based optimisation algorithm in matlab.
3. To achieve proper input parameters for VMD.
4. To plot and compare results with other methods for pitch estimation
5. Strike some comparisons (VMD vs EMD)
6. Video Presentation

Datasets

- CMU Arctic DB - Scottish Male, US Female - 16KHz
- Cmu_us_awb_arctic-0.90-release, cmu_us_slt_arctic-0.95-release

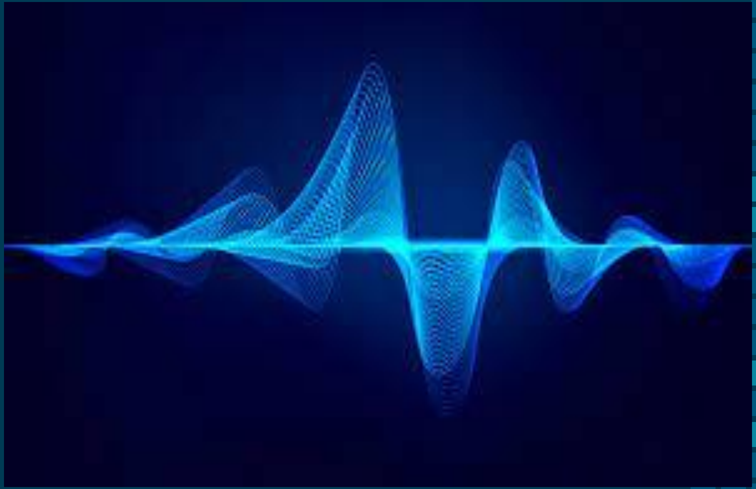
User Inputs and Outputs

- A 1D (mono channel) speech signal
- VMD parameters - K, alpha, tau, K, DC, init, tol

- IMFs or VMFs, residuals (of all iterations)
- Component Centre Frequencies, F_0 component, F_0 envelope
- V/UV separated components
- Instantaneous Pitch Frequency

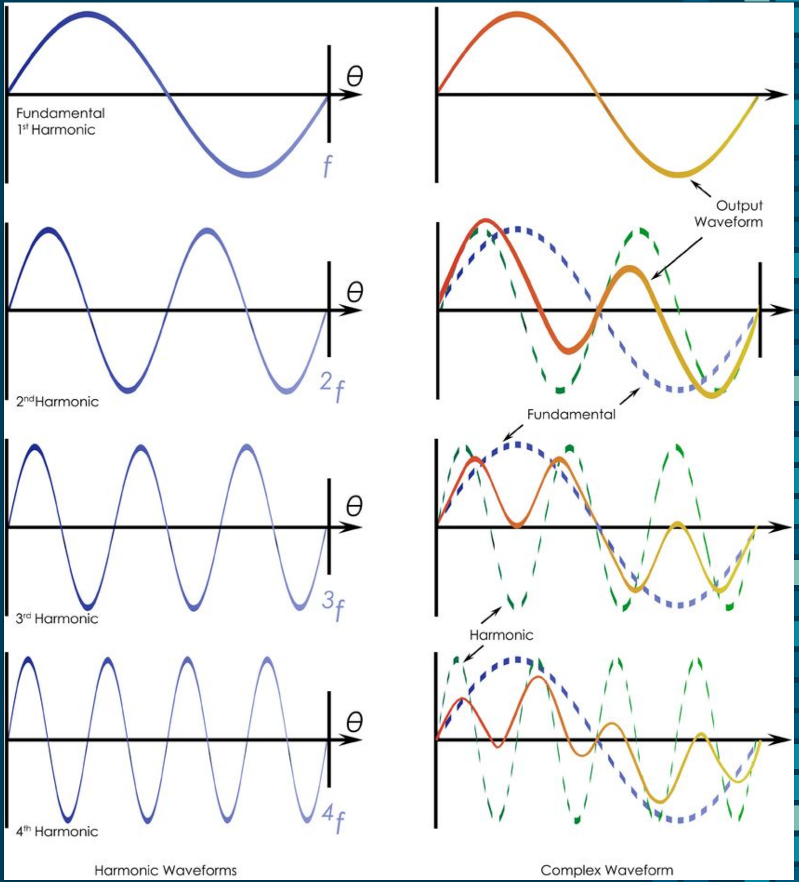
Why is pitch very important?

- Frequency determines it
- Perceptual Property - periodicity
- Important Speaker feature
- Pitch Shifting, Time scaling
- Speech processing
- **Def:** It is the fundamental frequency of the vocal cord vibrations.
- Perceived by Human Brain



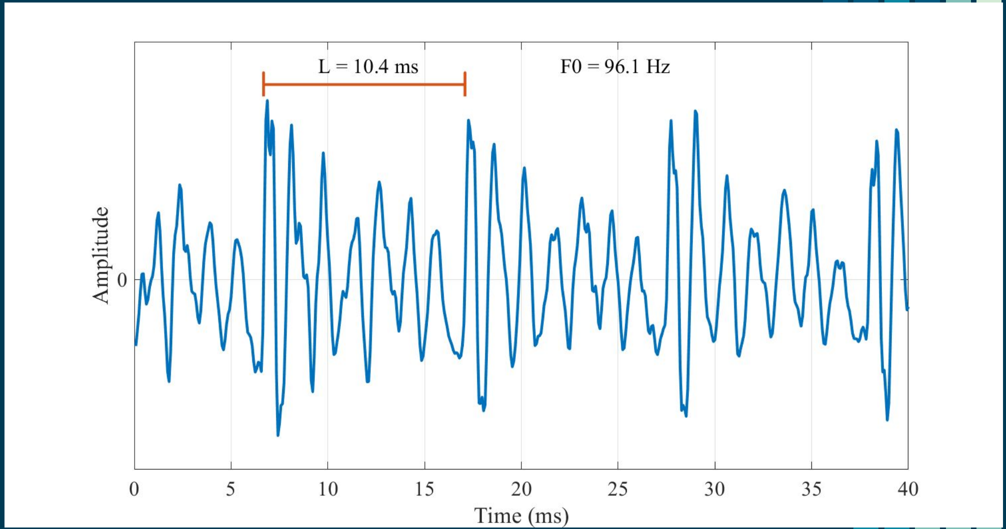
Fundamental Frequency

- The lowest frequency of a periodic waveform.
- The average number of oscillations per second (Hz).
- F_0 of speech can vary from 40 Hz for low-pitched voices to 600 Hz for high-pitched voices.



Fundamental Frequency

- Segment of a speech signal
- Fundamental period length T
- $F_0 = 1/T$.
- Fs samples in one T
- $L = Fs*T = Fs/F_0$.



Instantaneous Pitch Frequency

- The temporal derivative of the oscillation phase φ
- Useful for describing polychromatic (multiple frequencies) signals.
- The instantaneous frequency of a sinusoidal signal is constant and equals the oscillation frequency

$$z(t) = d(t) + jd_H(t) = A(t)e^{j\phi(t)} \tag{4}$$

In (4), $d_H(t)$ is the Hilbert transform of signal $d(t)$. The amplitude envelope and instantaneous phase of analytic signal $z(t)$ denoted as $A(t)$ and $\phi(t)$, respectively and can be computed as follows [31]:

$$A(t) = \sqrt{d^2(t) + d_H^2(t)} \tag{5}$$

$$\phi(t) = \arctan\left[\frac{d_H(t)}{d(t)}\right] \tag{6}$$

The instantaneous pitch frequency of the analytic signal $z(t)$ is determined as follows:

$$\omega(t) = \frac{d\phi(t)}{dt} \tag{7}$$

Pitch Estimation History/Techniques – PDAs

Time domain detection

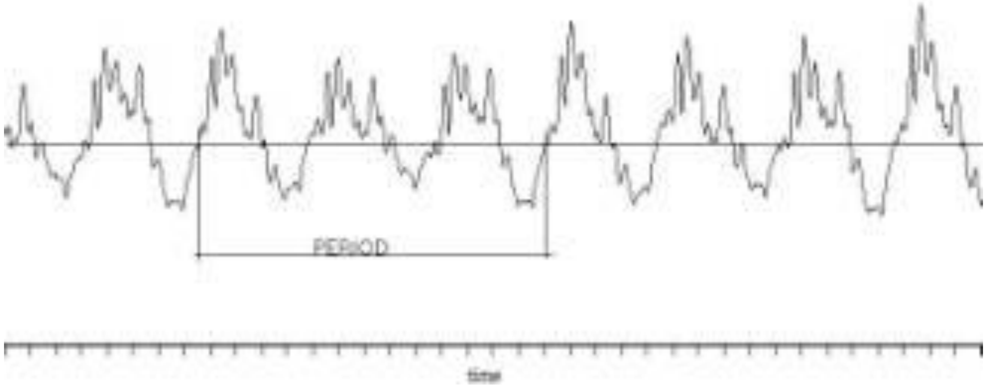
- ZCR
- Autocorrelation
- crosscorrelation
- Adaptive Filters Based
- Super Resolution Pitch Determination
- more

Frequency Domain Detection

- Harmonic Product Spectrum
- Cepstrum
- Maximum Likelihood
- more

Pitch Estimation History - PDAs

Intuitive Approach: Get the *zero crossing rate*. Does not work with overtones or with noise or quasi stationary signals.



Autocorrelation Algorithms: Highly accurate but prone to false detections. Bad with polyphonic and noisy signals

*used this method for comparison with VMD based method

Abstract of the Main Paper

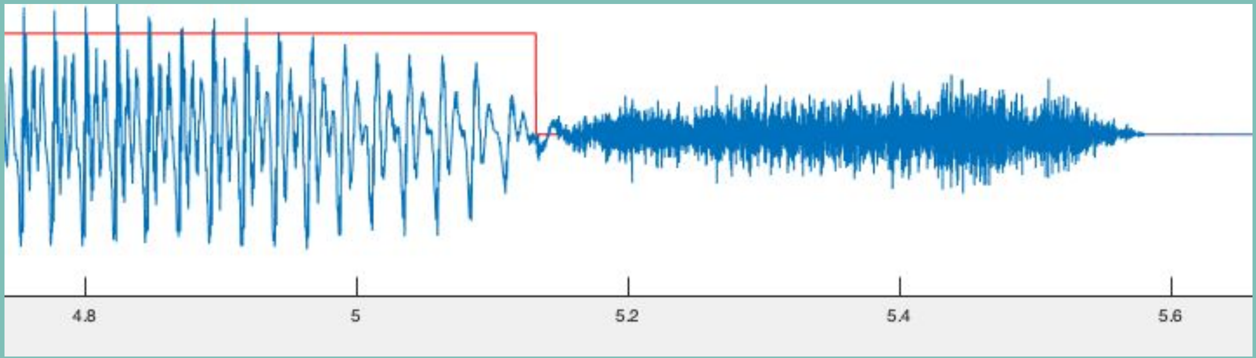
- An algorithm based on VMD and Hilbert Transform.
- VMD is applied iteratively until centre frequencies converge.
- Specific input parameters are needed.
- Fundamental frequency component and its envelope are extracted.
- Voiced and Unvoiced (V/UV) regions from a given speech signal are detected for F_0.
- Instantaneous pitch Freq is obtained by Hilbert Transform of Voiced speech component of F_0.

V/UV speech components

Speech can be decomposed into numerous V/UV segments.

Voiced Speech - Lower ZCR, higher energy & amplitude, almost constant frequency tones, periodic (therefore can be identified & extracted)

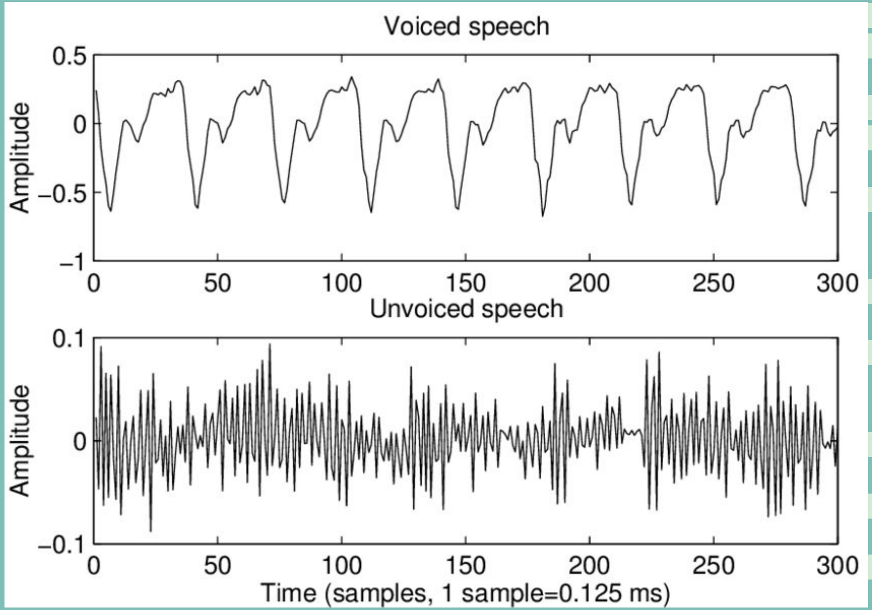
Unvoiced Speech - High ZCR, not periodic, random-like, energy & amplitude much lower



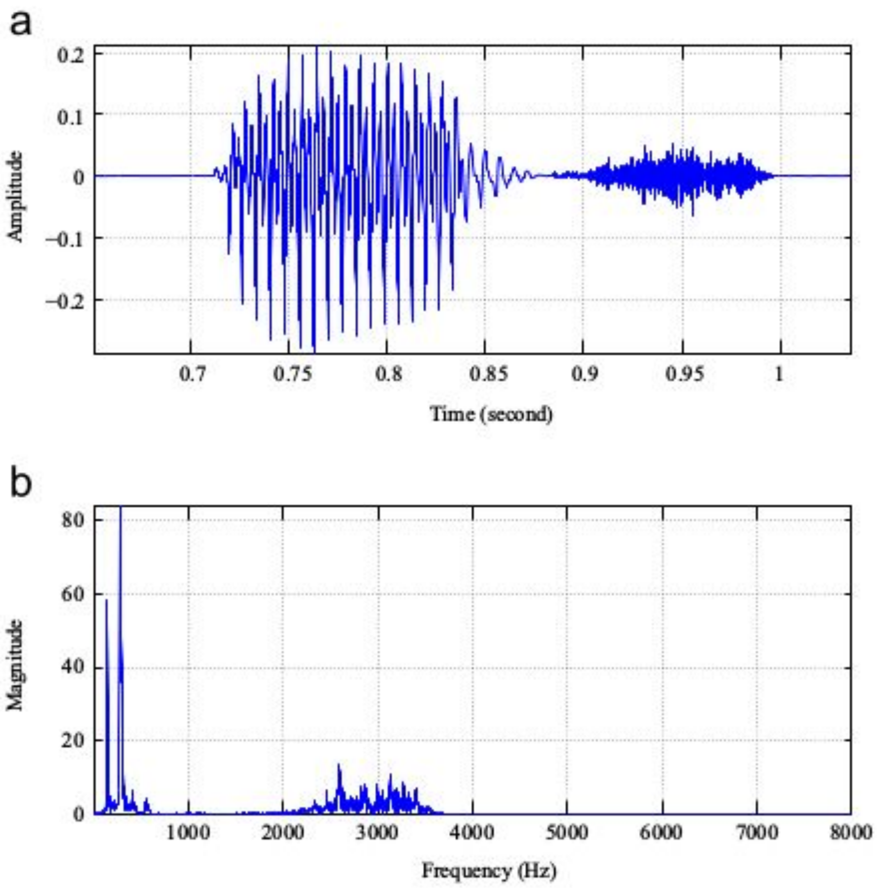
V/UV speech components

Voiced speech – produced when periodic pulses of air generated by the vibrating glottis resonate through the vocal tract, at frequencies dependent on the vocal tract shape.

Unvoiced speech – caused by air passing through a narrow constriction of the vocal tract as when consonants are spoken.



V/UV segments amplitude(speech signal)/magnitude spectrum

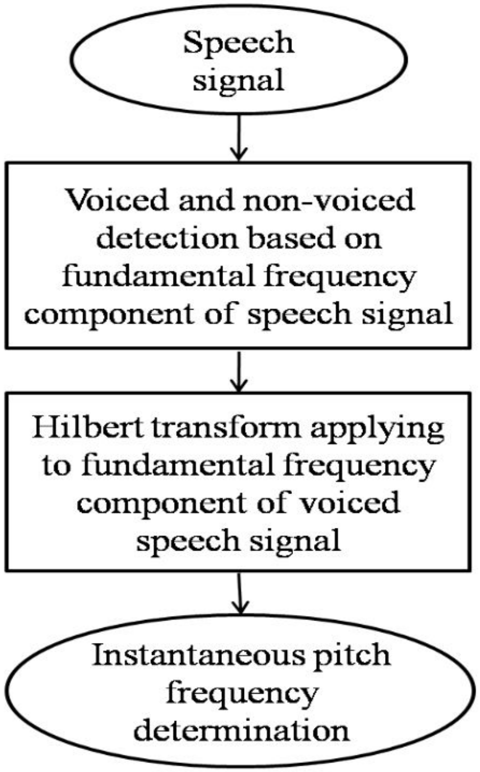


Algorithm Flow chart & Methodology

First Step: Extraction of Voiced Segment of F_0 of the given speech segment (most rigorous, involves more sub-steps) using VMD

Second Step: Hilbert Transform to get Analytical Signal

Third Step: Derivative of phase to get Instantaneous Pitch



Variational Mode Decomposition

- VMD is used to decompose a real valued signal to sub-signals or modes (yk). It replaces traditional EMD.
- Unlike the IMFs which are AM-FM signals with non decreasing phase, these modes are almost considered as compact around the corresponding center frequencies.

- Performance depends on the number of modes given and other input parameters
- It is an adaptive signal processing method (used in time varying systems)

Variational Mode Decomposition

1. The real signal is converted to the analytic signal using the Hilbert transform such that one-sided frequency spectrum of the signal is obtained.
2. The frequency spectrum of the component is shifted to baseband regions, using modulation property to the respective estimated center frequencies.
3. The bandwidth of a component is estimated through the H1 Gaussian smoothness of the demodulated signal, i.e. the squared L2-norm of the gradient.

Variational Mode Decomposition

The resulting constrained variational problem is,

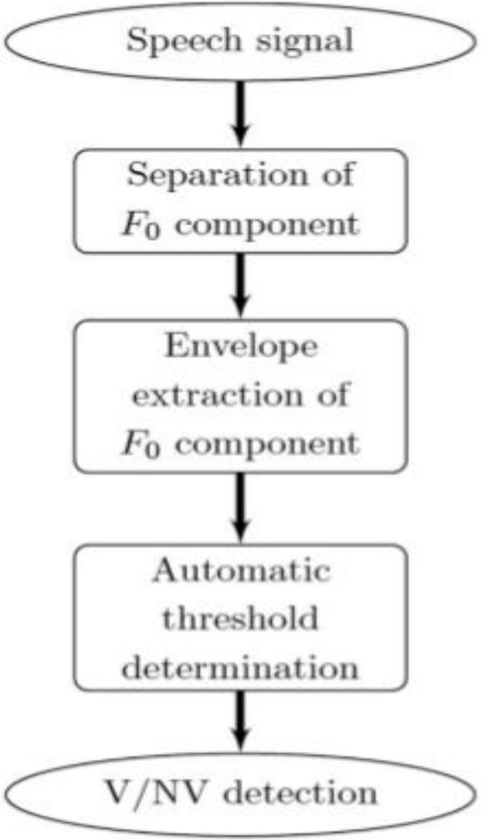
$$\min_{\{y_k\},\{\omega_k\}} \left\{ \sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * y_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}$$

$$\text{such that} \sum_{k=1}^{K} y_k(t) = y(t)$$

Here yk denotes kth component of the set of modes {yk }. Wk represents center frequency of Kth mode of the signal and {Wk} represents the set of center frequencies

Step 1 a- Determination of Fundamental Frequency Component

- Apply VMD on the signal with K = 2 (we will have 2 components as the output) and choose the component with min central frequency for next iteration.
- We go on with the iterations until the center frequencies converge.
- Now, how to achieve the center frequencies is the question. For that we have ADMM optimisation algorithm (alternating direction method of multipliers)



Parameters for the proposed Algorithm

The input parameters (#6):

1. The balancing parameter of the data-fidelity constraint (α)
2. The time step of the dual ascent (tau)
3. The number of components (K) to be extracted
4. The tolerance of convergence criterion (tol)
5. Number of DC components
6. The initialization of center frequencies ω (init)

The values of input parameters namely tol, tau, init, α, and K are fixed to 10^20 , 0, 0, 80 and 2, respectively.

Algorithm

Constrained variational problem with parameters and lagrangians

$$\mathcal{L}(\{y_k\},\{\omega_k\},\lambda) := \alpha \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * y_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \left\| y(t) - \sum_k y_k(t) \right\|_2^2 + \left\langle \lambda(t), y(t) - \sum_k y_k(t) \right\rangle$$

From ADMM, we get the following updates (alpha is low)

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega |\hat{Y}_k(\omega)|^2 \, d\omega}{\int_0^\infty |\hat{Y}_k(\omega)|^2 \, d\omega}$$

$$\hat{Y}_k^{n+1}(\omega) = \frac{\hat{Y}(\omega) - \sum_{i \neq k} \hat{Y}_i(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)^2}$$

Step 1 a- Optimisation Algorithm

Minimisation is done wrt u_k and w_k

Note $u_k \sim Y_k$

Convergence Criteria:

- The final selected component should posses the energy (or center frequency) nearly equal to the component in the previous iteration.
- Parseval relation to convert the signals and computations from time to frequency domains
- In this case, center frequency less than 50 Hz components are discarded

Algorithm 1: ADMM optimization concept for VMD

Initialize $\{u_k^1\}, \{\omega_k^1\}, \lambda^1, n \leftarrow 0$
repeat

$n \leftarrow n + 1$

for $k = 1 : K$ **do**

Update u_k:

$$u_k^{n+1} \leftarrow \arg\min_{u_k} \mathcal{L}\left(\{u_{i<k}^{n+1}\}, \{u_{i \geq k}^{n}\}, \{\omega_i^n\}, \lambda^n\right) \quad (16)$$

end for

for $k = 1 : K$ **do**

Update ω_k:

$$\omega_k^{n+1} \leftarrow \arg\min_{\omega_k} \mathcal{L}\left(\{u_i^{n+1}\}, \{\omega_{i<k}^{n+1}\}, \{\omega_{i \geq k}^{n}\}, \lambda^n\right) \quad (17)$$

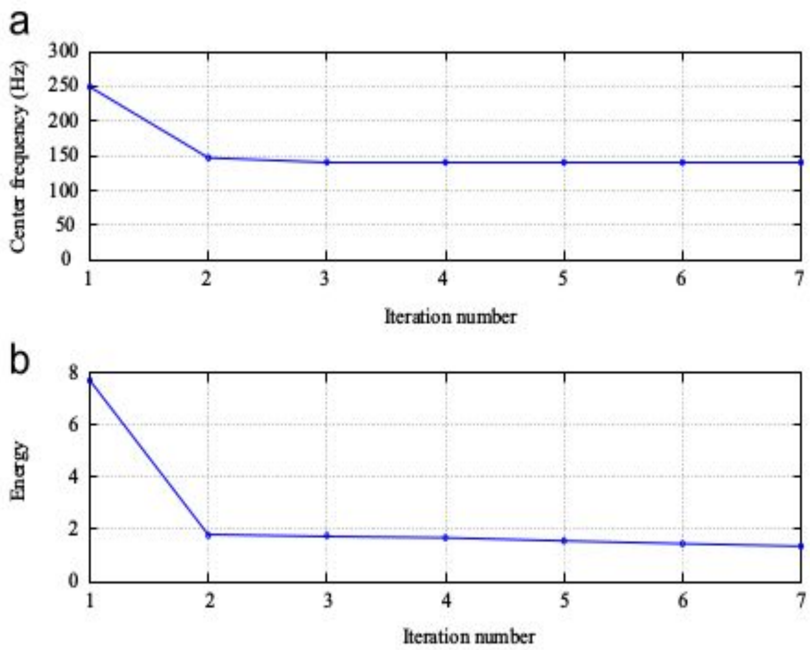
end for

Dual ascent:

$$\lambda^{n+1} \leftarrow \lambda^n + \tau\left(f - \sum_k u_k^{n+1}\right) \quad (18)$$

until convergence: $\sum_k \|u_k^{n+1} - u_k^n\|_2^2 / \|u_k^n\|_2^2 < \epsilon$.

- Center frequency and energy of selected components of iter3 and iter4 are nearly equal.
- So, iter4 selected component is chosen as F_0



Methodology

Step 1b - Envelope detection

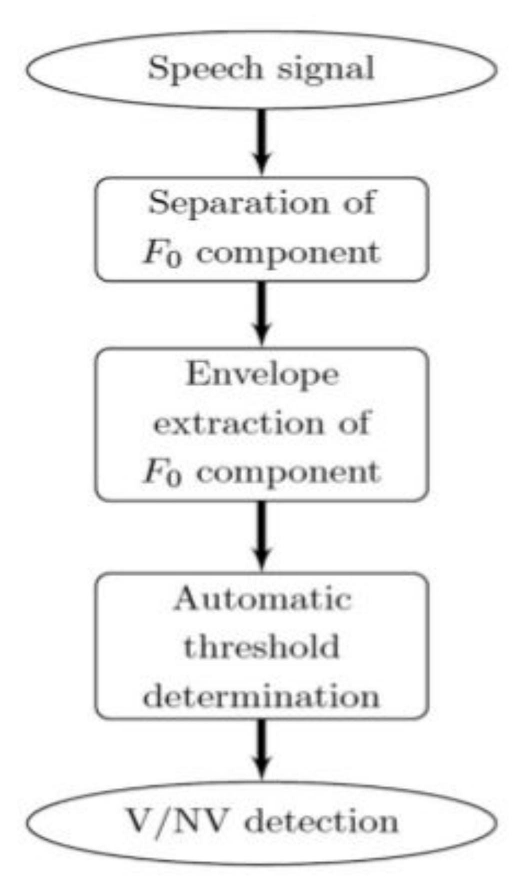
Step 1c - The threshold for V/NV detection. d(t)

Step 2 - Apply Hilbert transform on the voiced part of the speech ($d_h(t)$)

Step 2a - Compute the analytical signal of the d(t)

Step 3 - Compute the phase of the analytical signal.

Step 4 - Compute the derivative of the phase. This will be the instantaneous pitch.



Results – F_0 and center freq convergence is achieved

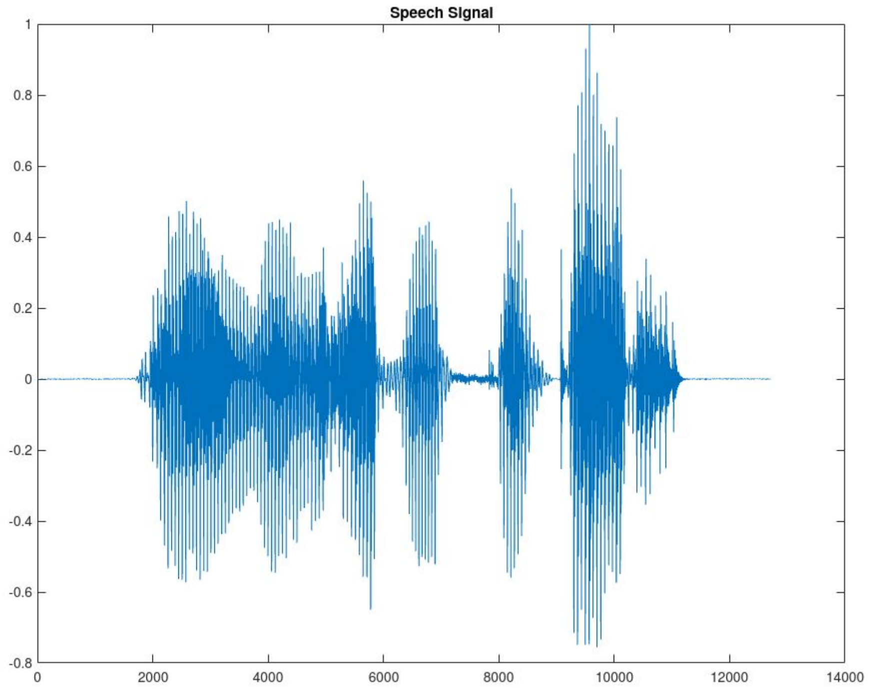
We got the F0 component from VMD. We had written an elaborate code for estimation of F0 from VMD.

We also computed the envelope of F0

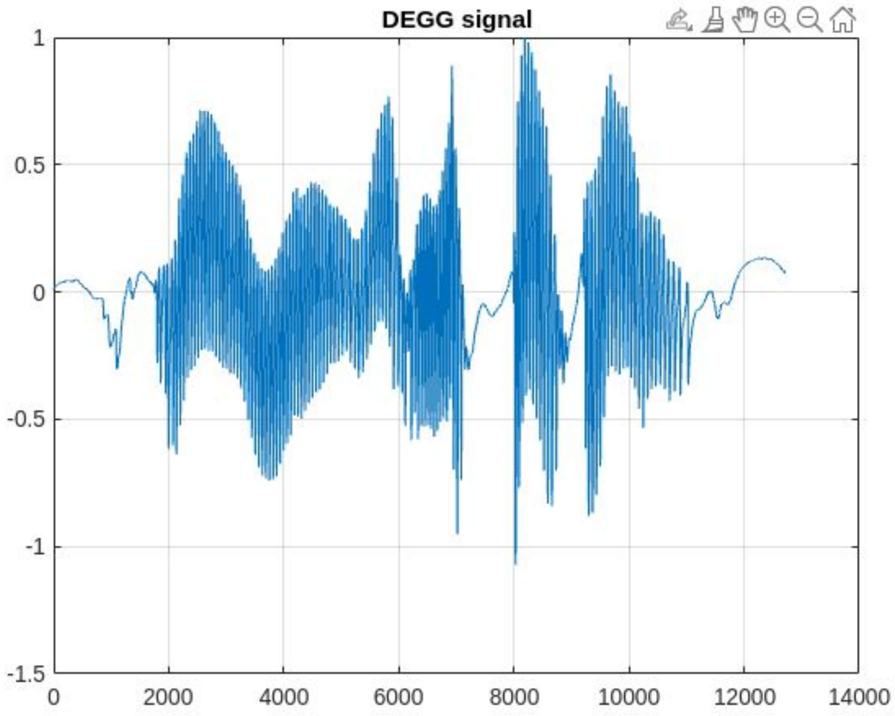
We need to find automatic threshold next for V/UV and main paper scope will be done.

We need to verify energy convergence (it will as center freq are converging already)

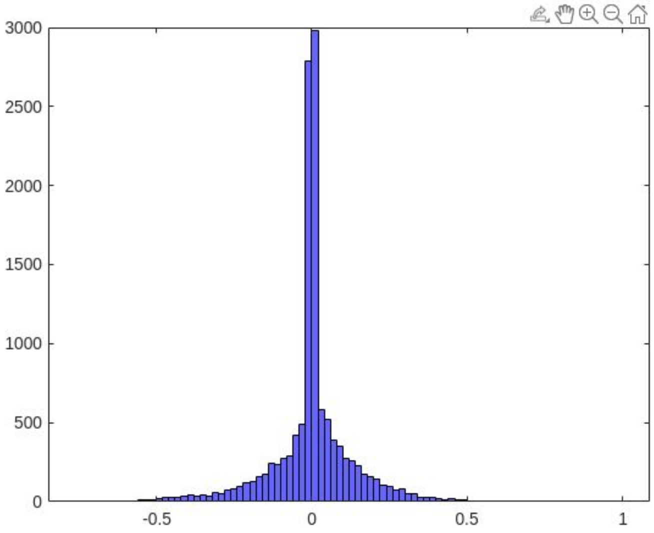
Need to plot Hilbert Transform, Analytical signal etc

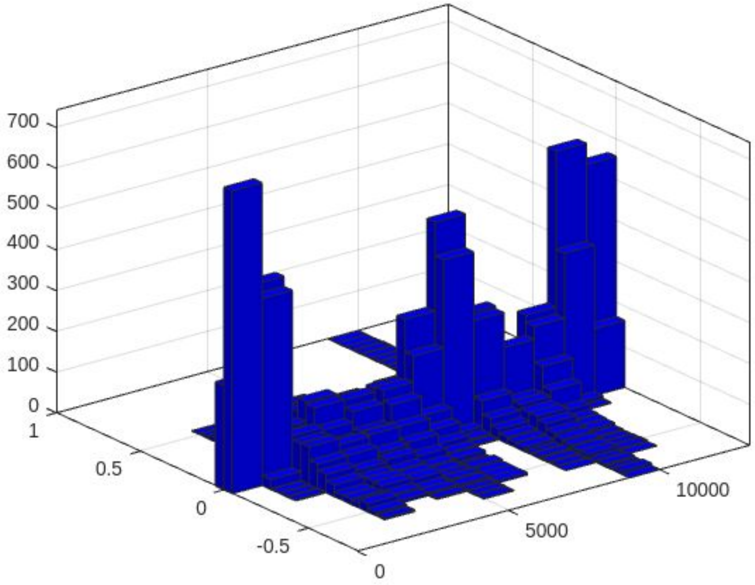


- DEGG Signal and the Histogram of Speech Signal



Histograms of the speech signal





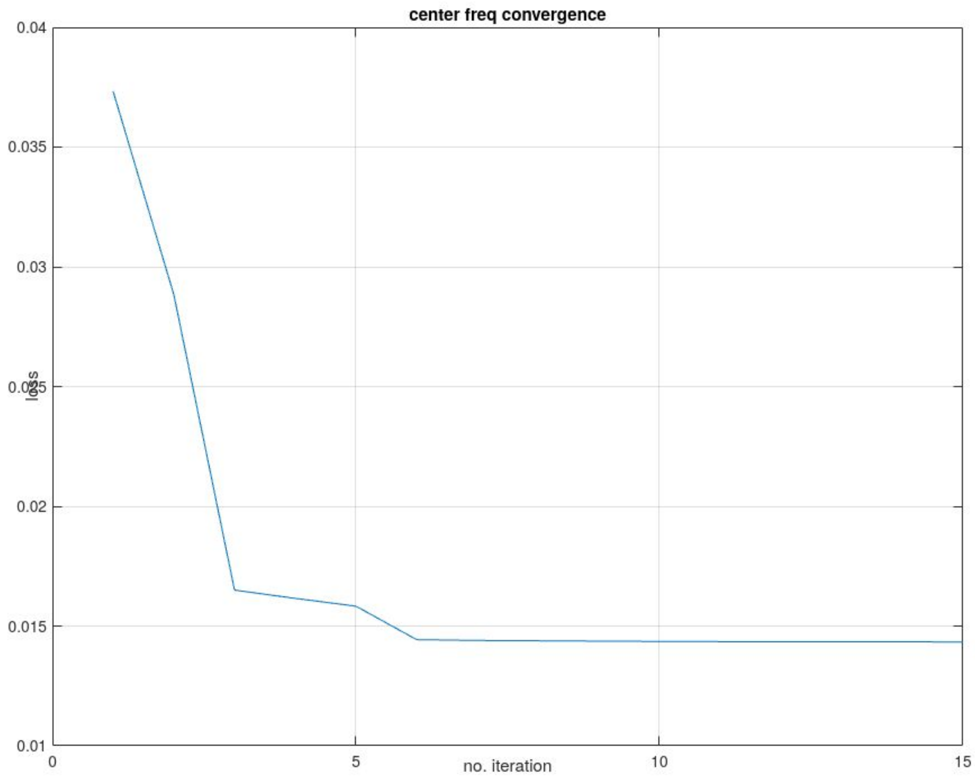
Plots - Center Frequency Convergence

The convergence parameter difference = 1e-6 for these plots.

We experimented with 0.1, 0.01 also but it was not quite converging.

Value of Convergence is around 0.11 rad/s

We are plotting only time domain plots. Magnitude Spectrum, Threshold function, etc will plot in later ppt



Code Snippets

```
if v3.CentralFrequencies(1) <= v3.CentralFrequencies(2)
    comp = 1;
else
    comp = 2;
end

if comp == 1
    y_iter = v1(:,1);
else
    y_iter = v1(:,2);
end

if i == 1
    CF_arr = v3.CentralFrequencies(comp);
else
    CF_arr = [CF_arr; v3.CentralFrequencies(comp)];
end

if (i >= 2)
    if abs(CF_arr(i) - CF_arr(i-1)) <= diff
        y_F0 = y_iter;
        y_CF = CF_arr(i);
        break;
    end

    if(i>100)
        y_F0 = y_iter;
        y_CF = CF_arr(i);
        break;
    end

end
```

```
yh = hilbert(voiced);
z = voiced + 1j*yh;

phi=zeros(1,length(z));
for j=1:length(z)
    phi(j)= atan(imag(z(j))/real(z(j)));
end

% omega=diff(phi)/(1/fs);

% differentiation of phase to get IF
omega = zeros(1,length(phi));
for i=1:length(phi)-1
    omega(i) = (phi(i+1) - phi(i))/(t(i+1) - t(i));
end
omega(length(phi)) = omega(length(phi)-1);

%% IF extraction using EMD (HHT)
```

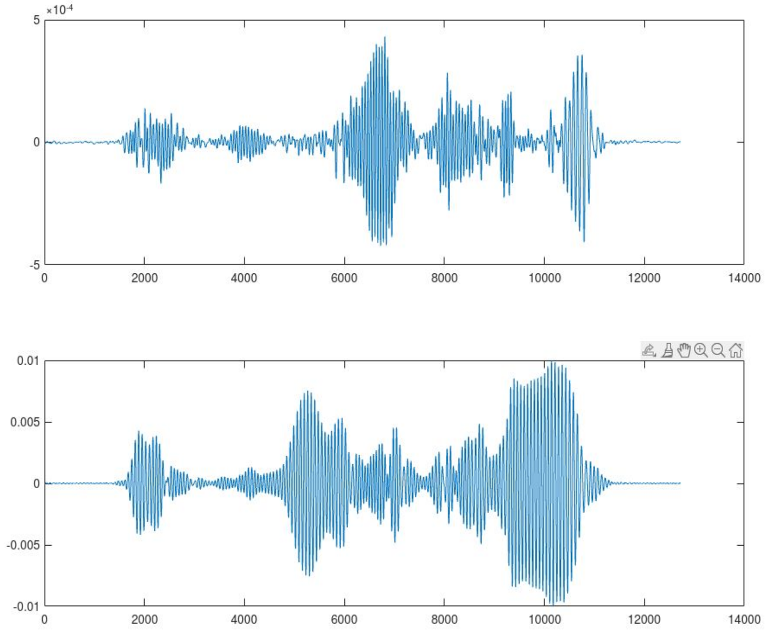
VMFS of all Iterations – Plot of F0

Number of IMF or modes achieved for each iteration of called vmd is 2 (k=2)

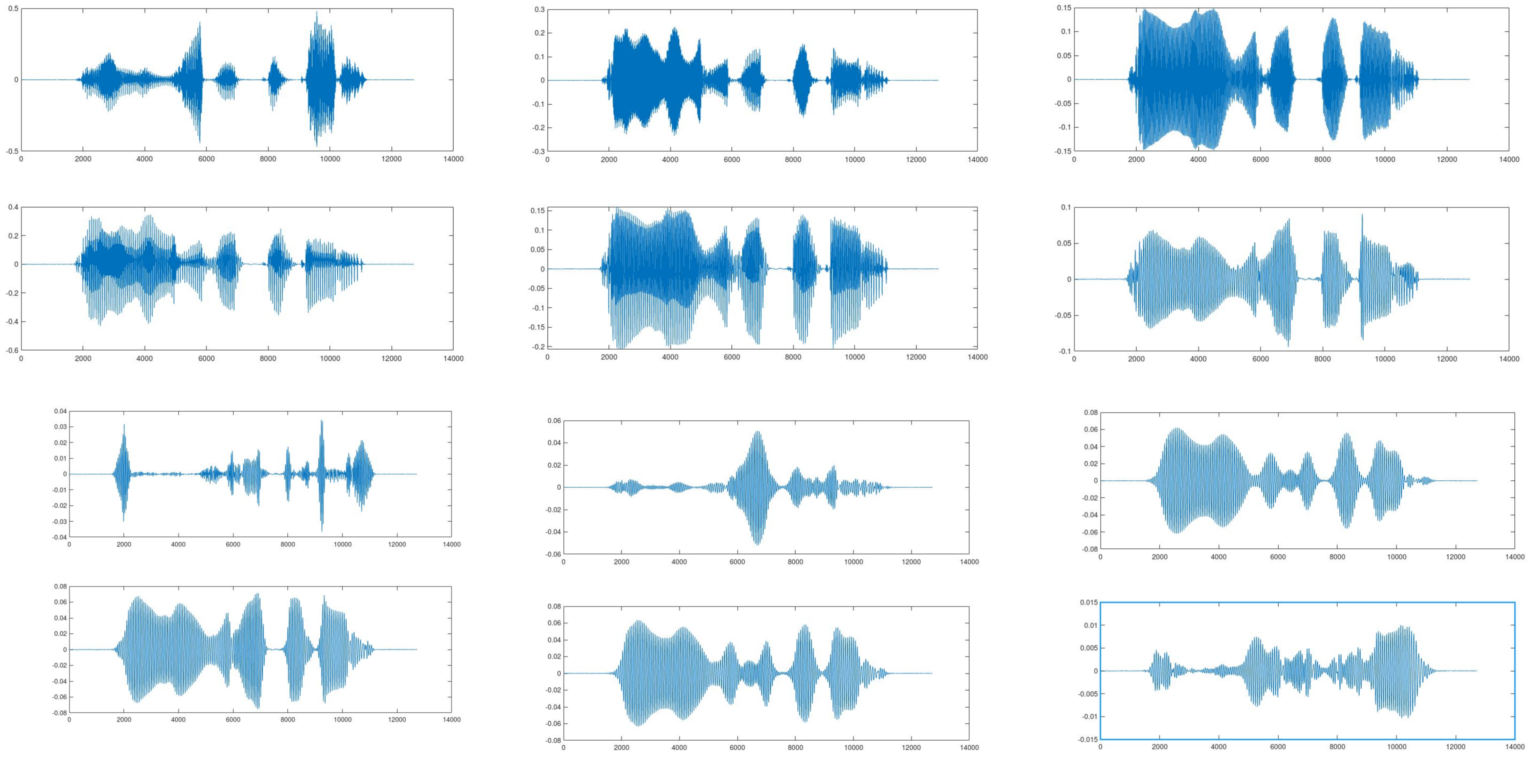
Number of iterations depends upon the convergence criteria.

The component with least center frequency of all these achieved components is the F_0.

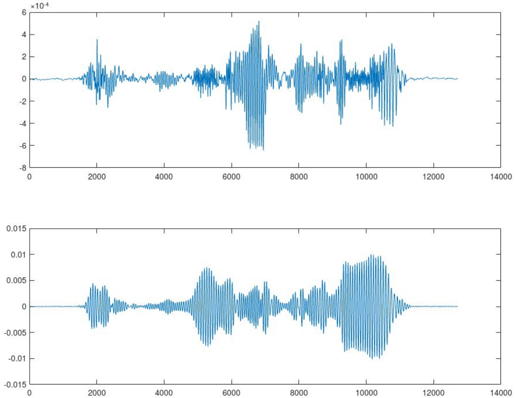
Note: This F_0 may not be fully accurate and it depends on 'diff' parameter. In this case, it is pretty good

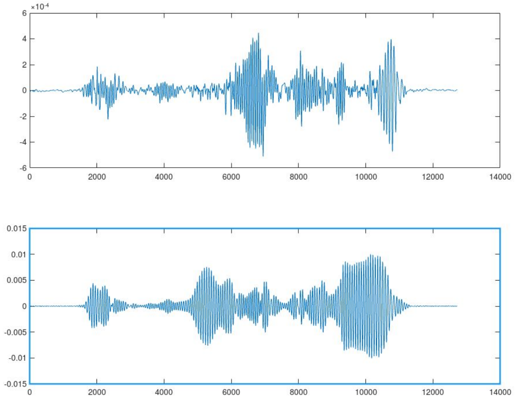


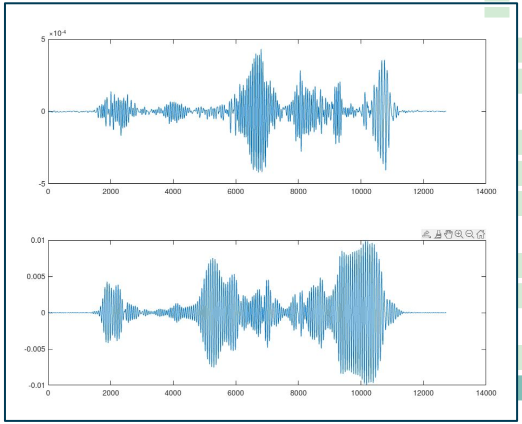
Intermediate VMF plots (iteration 1-6)

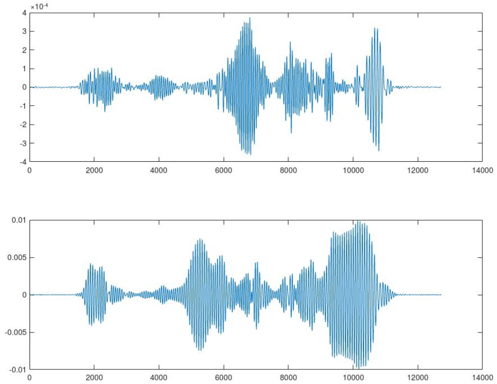


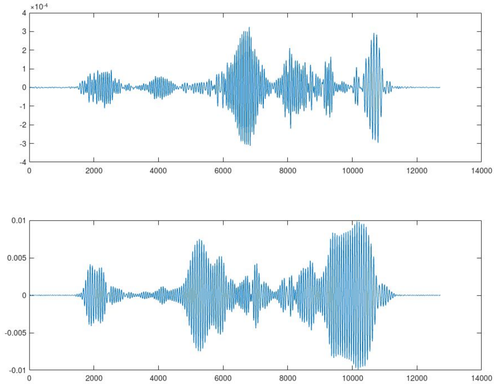
Intermediate VMF plots (iteration 6-12)

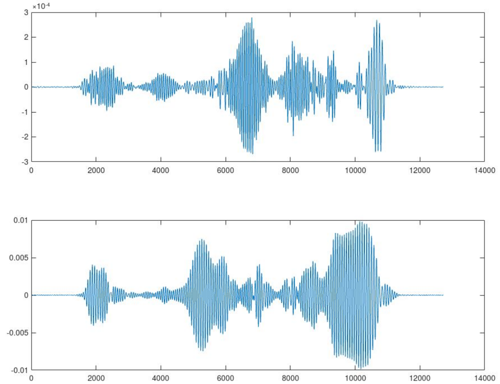




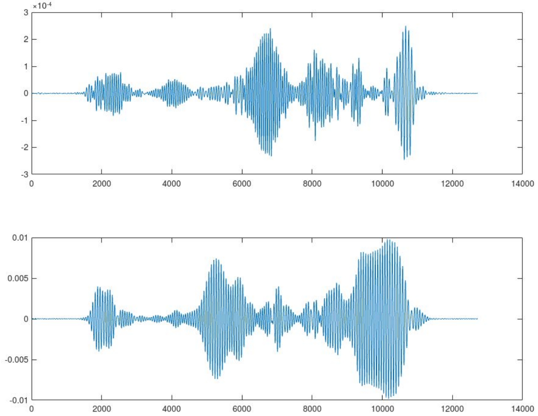


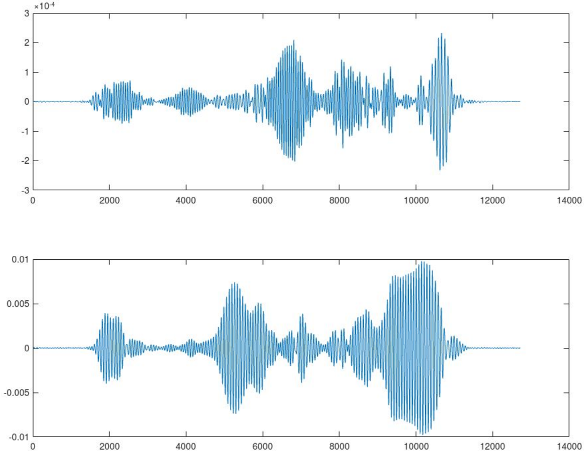


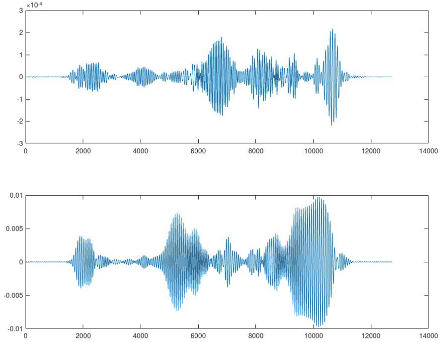




Intermediate VMF plots (iteration 12-15)



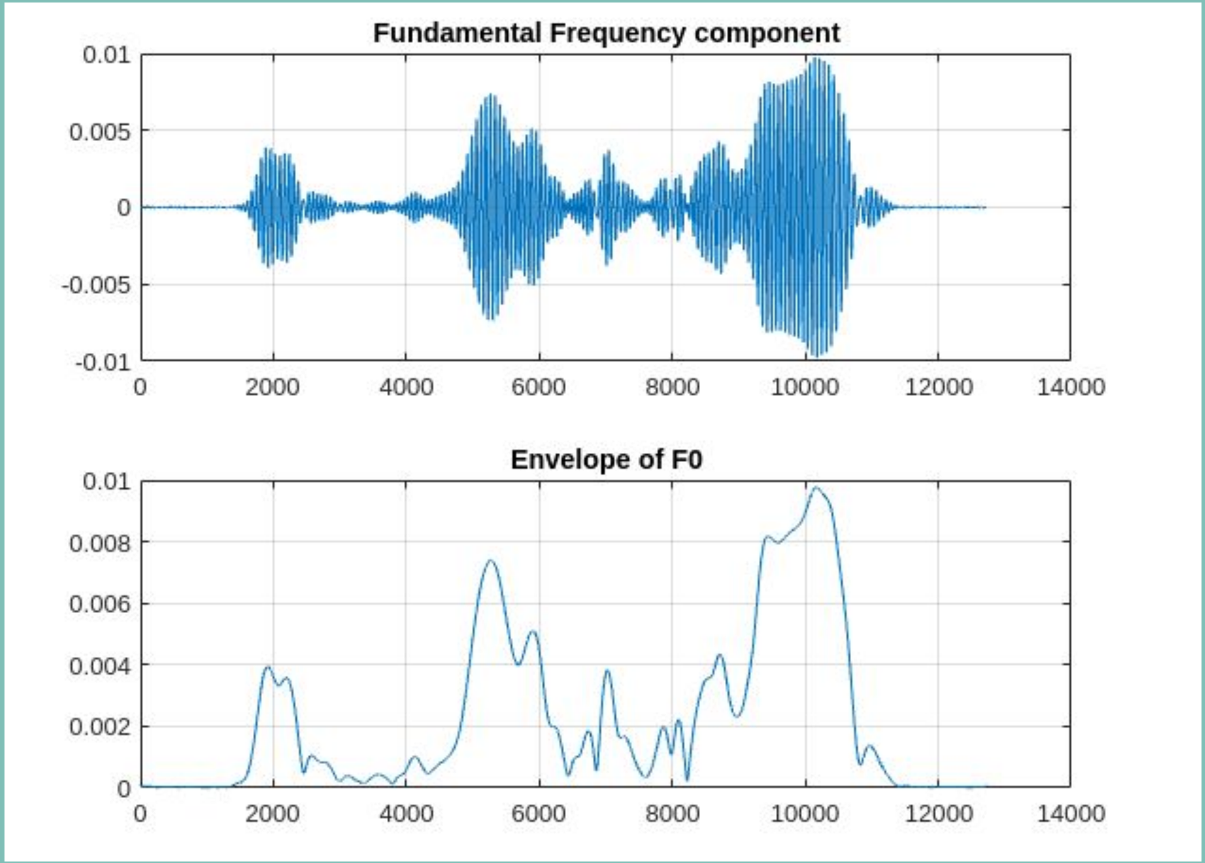




Observations

We can cleary notice that with increased number of iterations, the frequency of the components is reducing (F0 is basically the smallest frequency in the signal)

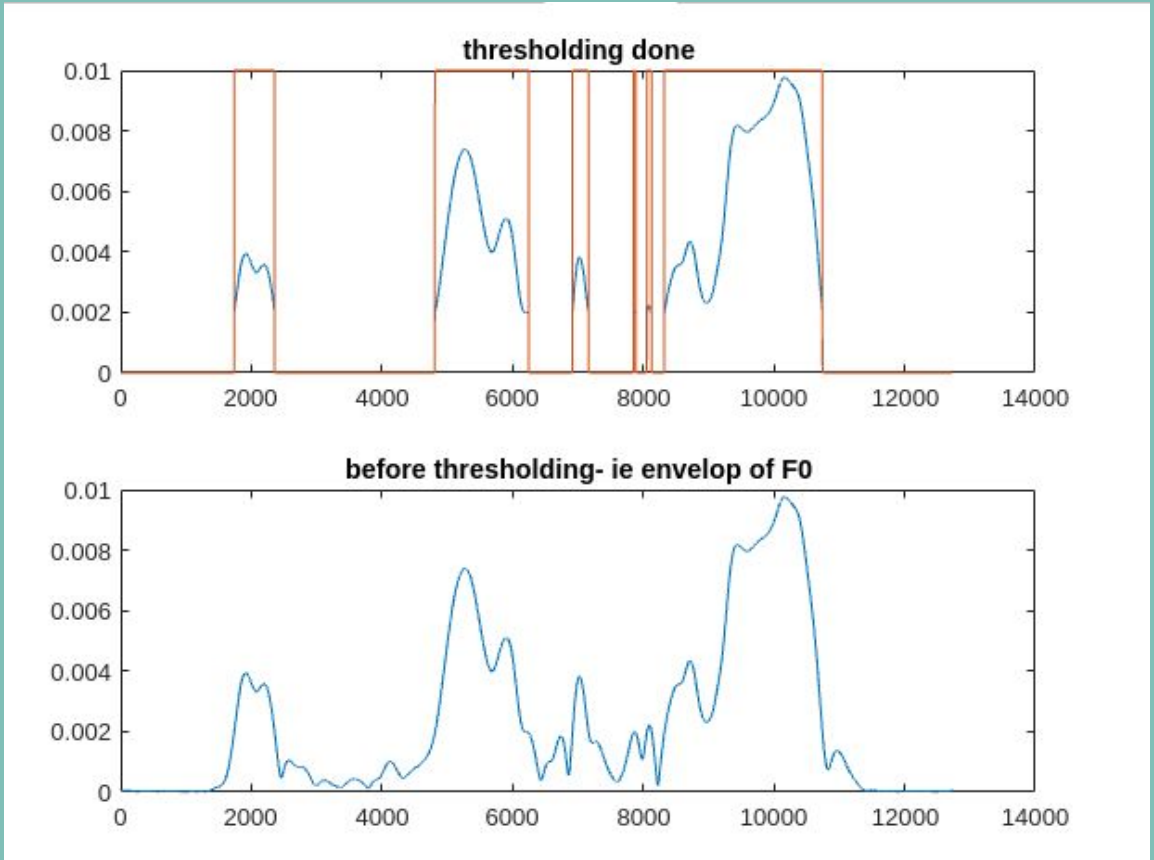
F0 envelope


Fundamental Frequency component
0.01
0.005
0
-0.005
-0.01
0
2000
4000
6000
8000
10000
12000
14000
Envelope of F0
0.01
0.008
0.006
0.004
0.002
0
0
2000
4000
6000
8000
10000
12000
14000

Extraction of voiced and unvoiced components

The first threshold can be set as 0.5% of the maximum value of the envelope of the F0 component of speech signal.
Difference between the probability density function (PDF) of the appended silence region and that of the noise in entire speech can be computed by Kullback–Leibler divergence (KLD) , the average value of which comes out to be less or approximately equal to 1%. These threshold are given in the main paper.

- $P(E_n) <= Th2 = 0.99$
- $Th1 = 0.005 \times \max\{E_n\}$
- $X = \max\{Th1, Th2\}$



thresholding done
0.01
0.008
0.006
0.004
0.002
0
0
2000
4000
6000
8000
10000
12000
14000
before thresholding- ie envelop of F0
0.01
0.008
0.006
0.004
0.002
0
0
2000
4000
6000
8000
10000
12000
14000

Hilbert transform followed by analytical signal

- Instantaneous pitch Freq is obtained by Hilbert Transform of Voiced speech component of F_0 say $d_h(t)$
- Analytical signal of voiced speech component of is F_0 created as follows

$$z(t) = d(t) + j d_H(t) = A(t) e^{j\phi(t)} \tag{4}$$

In (4), $d_H(t)$ is the Hilbert transform of signal $d(t)$. The amplitude envelope and instantaneous phase of analytic signal $z(t)$ denoted as $A(t)$ and $\phi(t)$, respectively and can be computed as follows [31]:

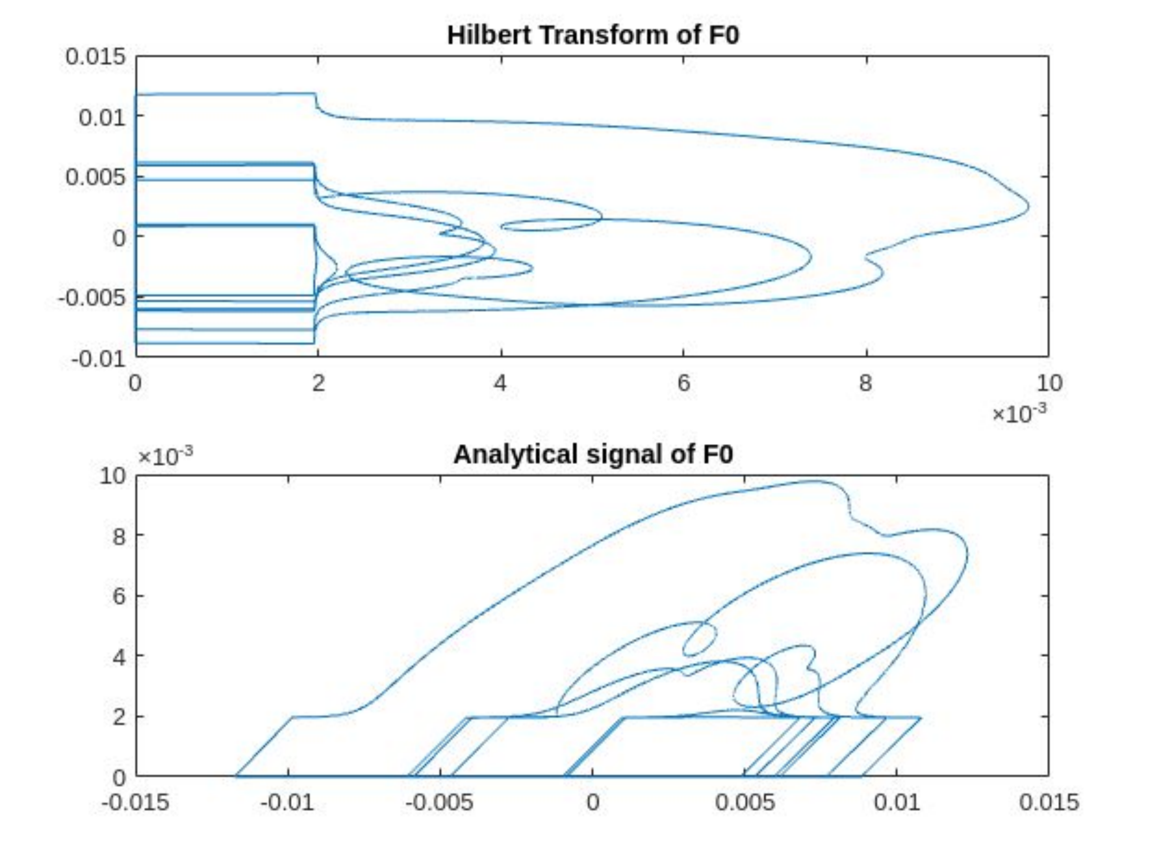
$$A(t) = \sqrt{d^2(t) + d_H^2(t)} \tag{5}$$

$$\phi(t) = \arctan\left[\frac{d_H(t)}{d(t)}\right] \tag{6}$$

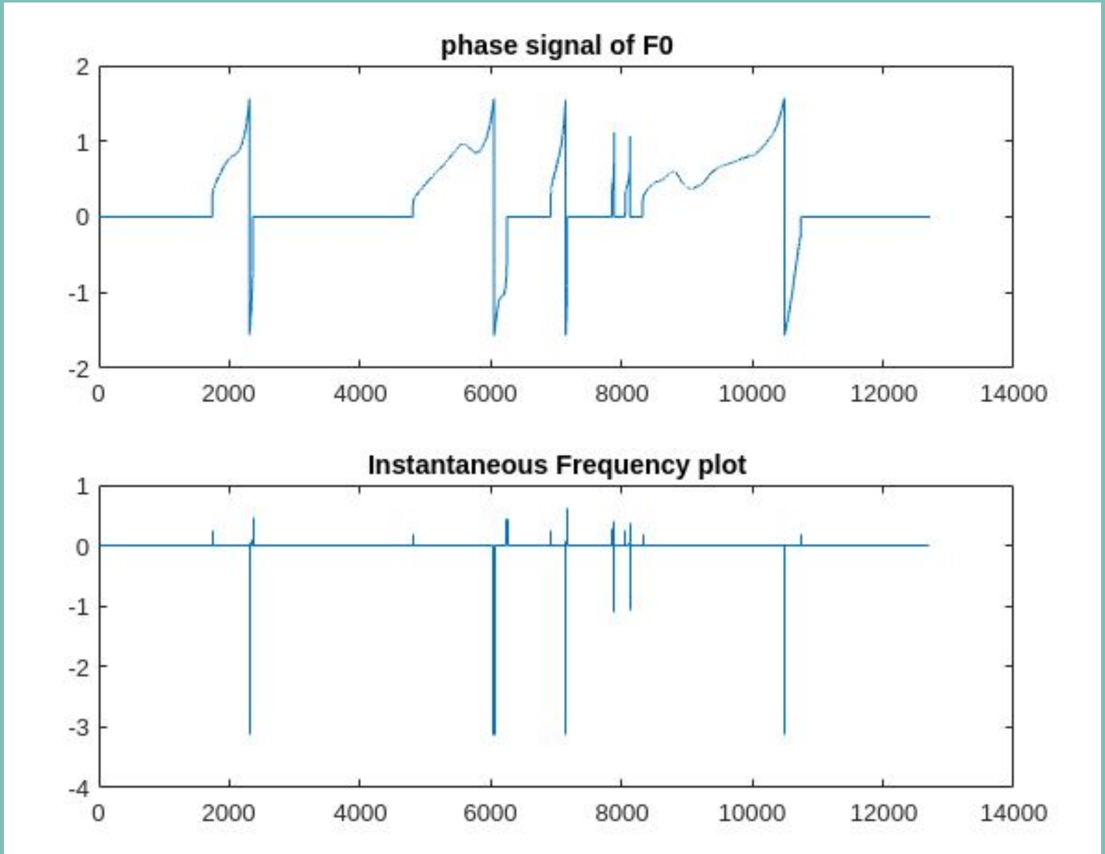
The instantaneous pitch frequency of the analytic signal $z(t)$ is determined as follows:

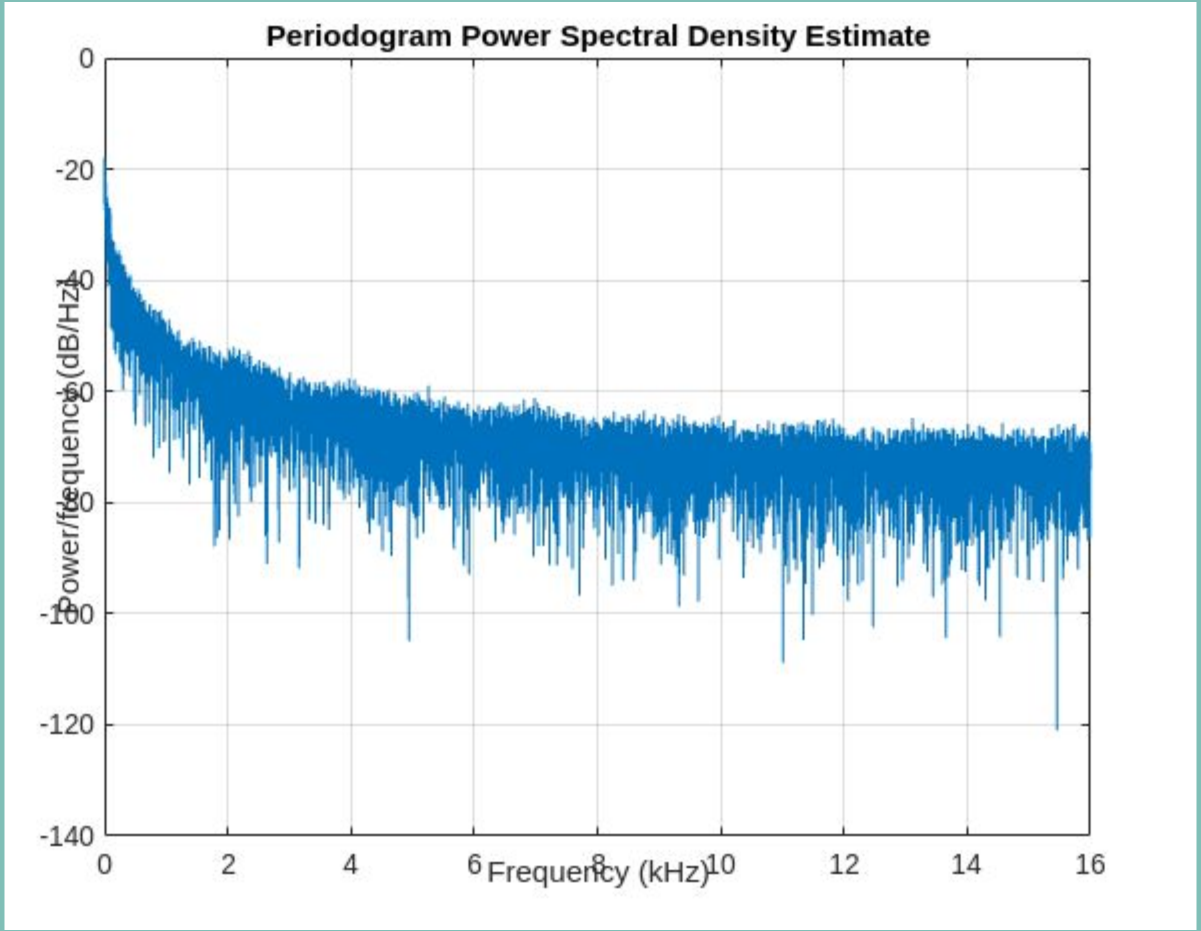
$$\omega(t) = \frac{d\phi(t)}{dt} \tag{7}$$

Hilbert Transform and Analytical Signal of Voiced F0



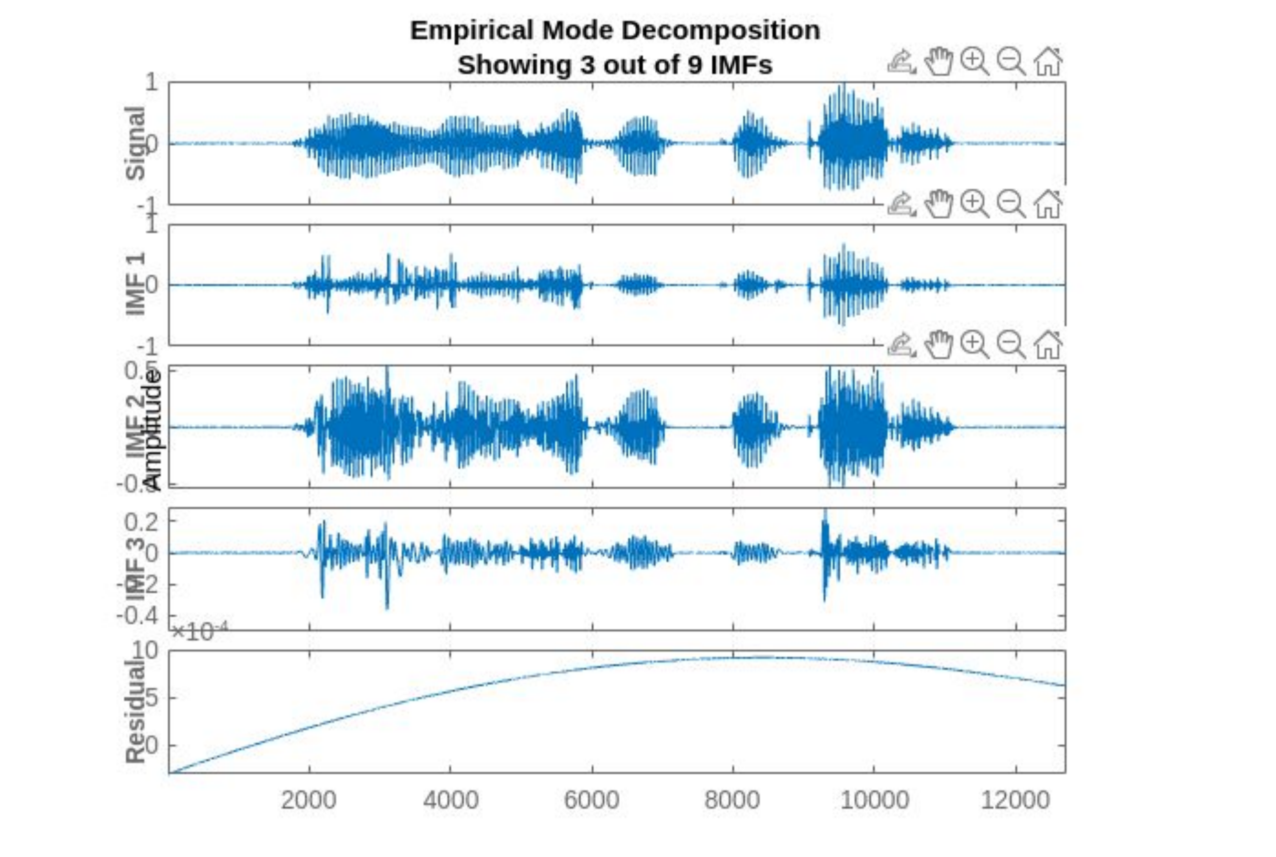
Phase of Analytical Signal and IF

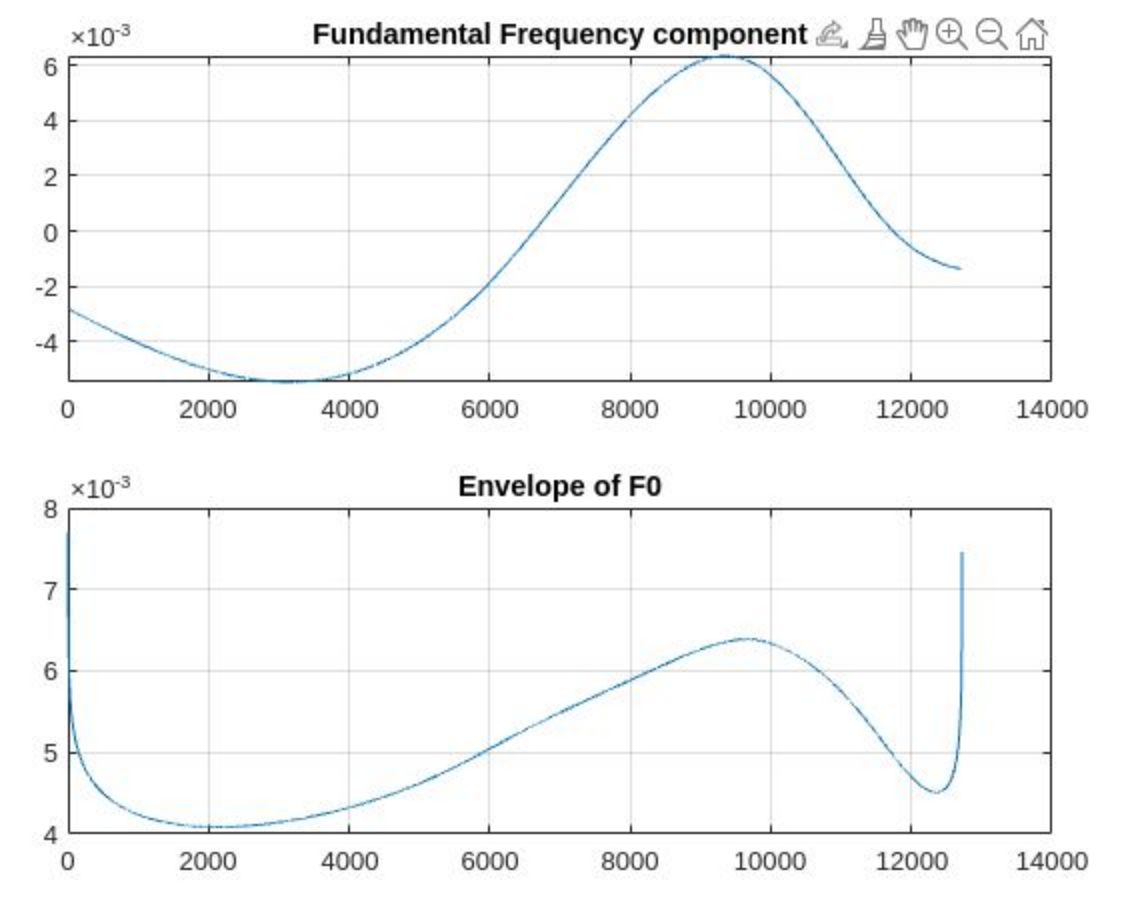


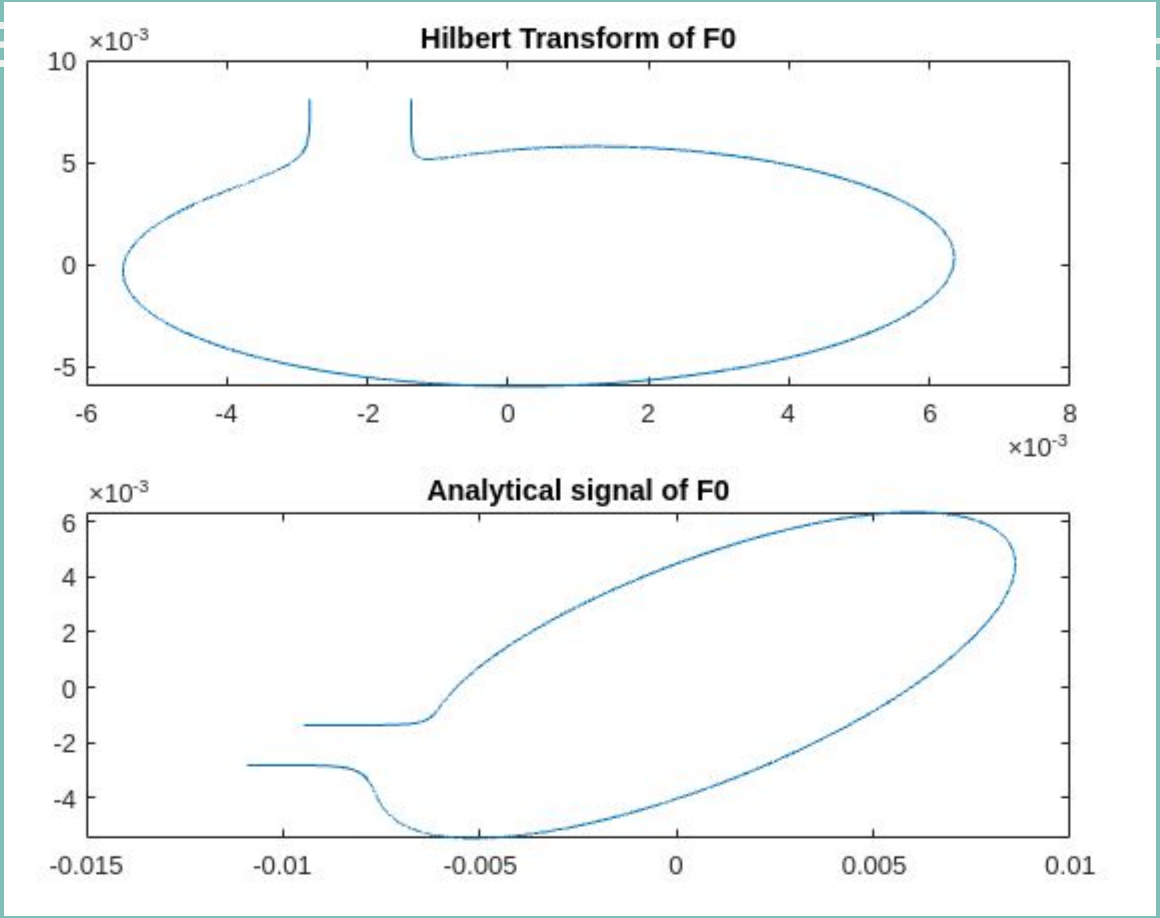


Periodogram Power Spectral Density Estimate
Power/frequency (dB/Hz)
Frequency (kHz)
0
-20
-40
-60
-80
-100
-120
-140
0
2
4
6
8
10
12
14
16

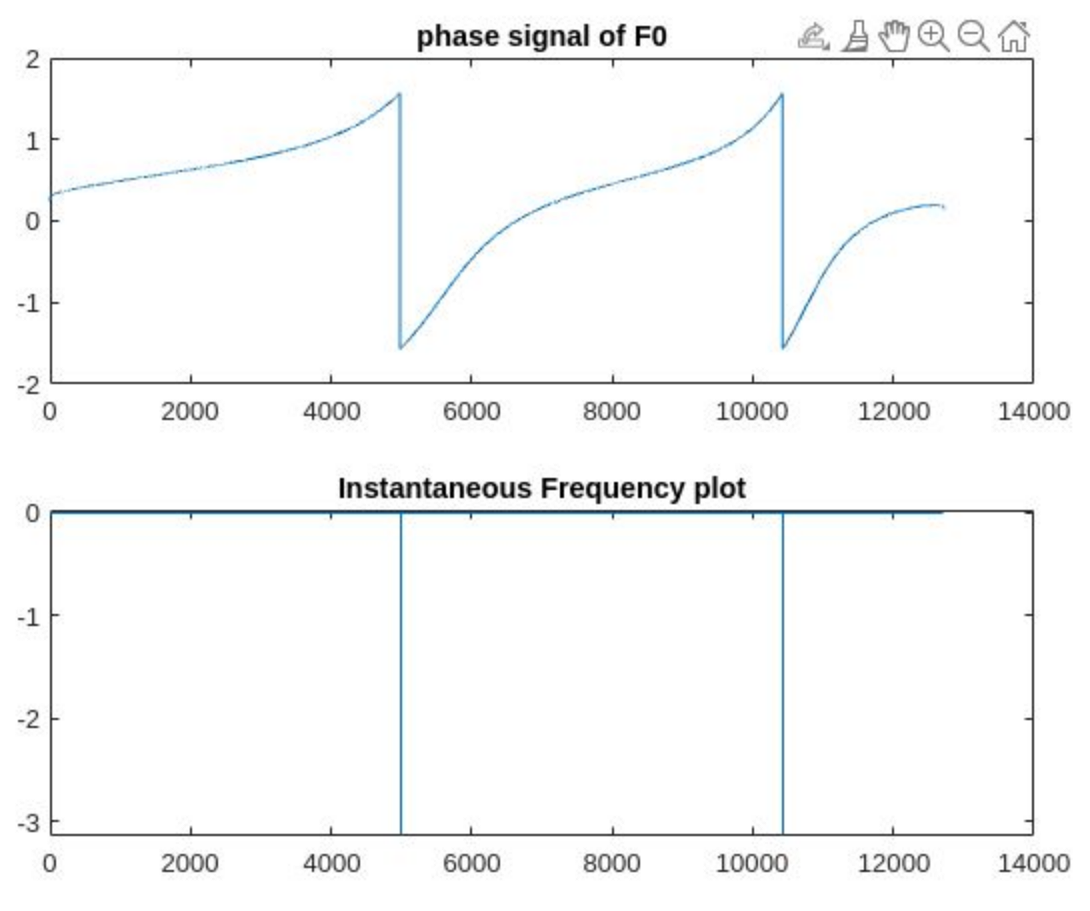
IF Detection using EMD and HHt

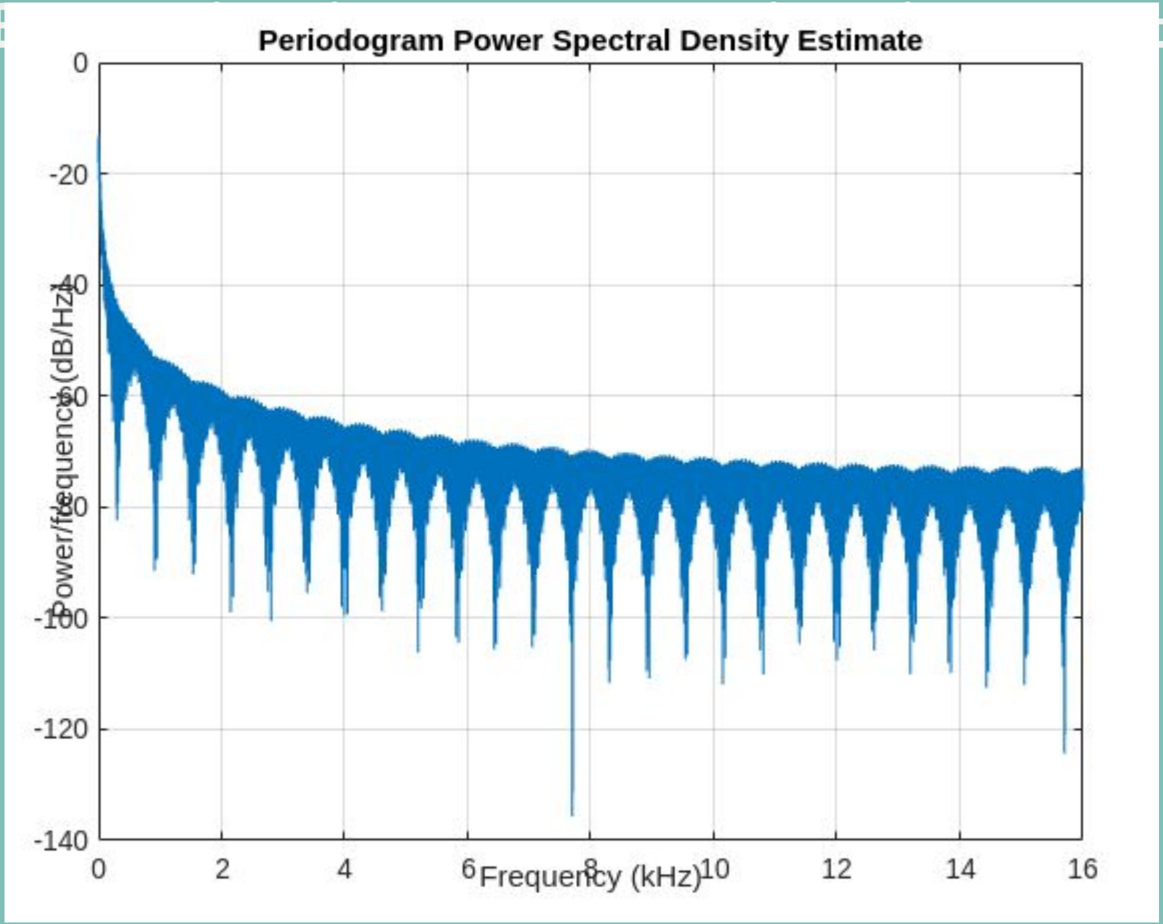



Fundamental Frequency component
×10-3
6
4
2
0
-2
-4
0
2000
4000
6000
8000
10000
12000
14000
Envelope of F0
×10-3
8
7
6
5
4
0
2000
4000
6000
8000
10000
12000
14000


Hilbert Transform of F0
×10-3
10
5
0
-5
-6
-4
-2
0
2
4
6
8
×10-3
Analytical signal of F0
×10-3
6
4
2
0
-2
-4
-0.015
-0.01
-0.005
0
0.005
0.01

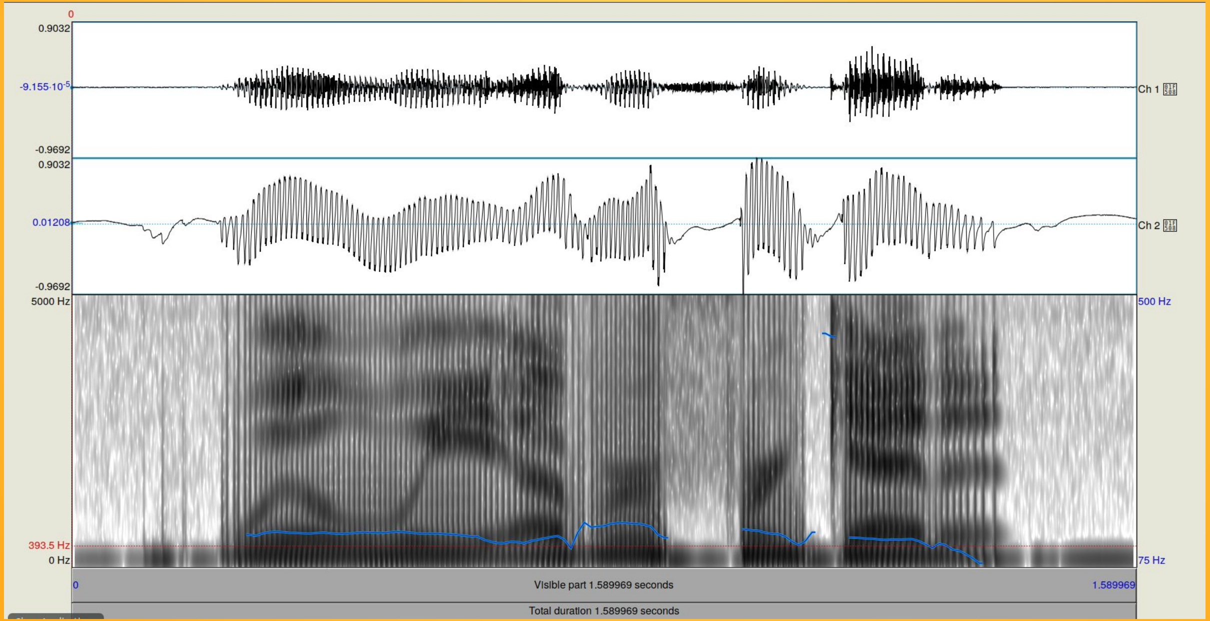
Er … ction


phase signal of F0
2
1
0
-1
-2
0
2000
4000
6000
8000
10000
12000
14000
Instantaneous Frequency plot
0
-1
-2
-3
0
2000
4000
6000
8000
10000
12000
14000


Periodogram Power Spectral Density Estimate
Power/frequency (dB/Hz)
0
-20
-40
-60
-80
-100
-120
-140
Frequency (kHz)
0
2
4
6
8
10
12
14
16

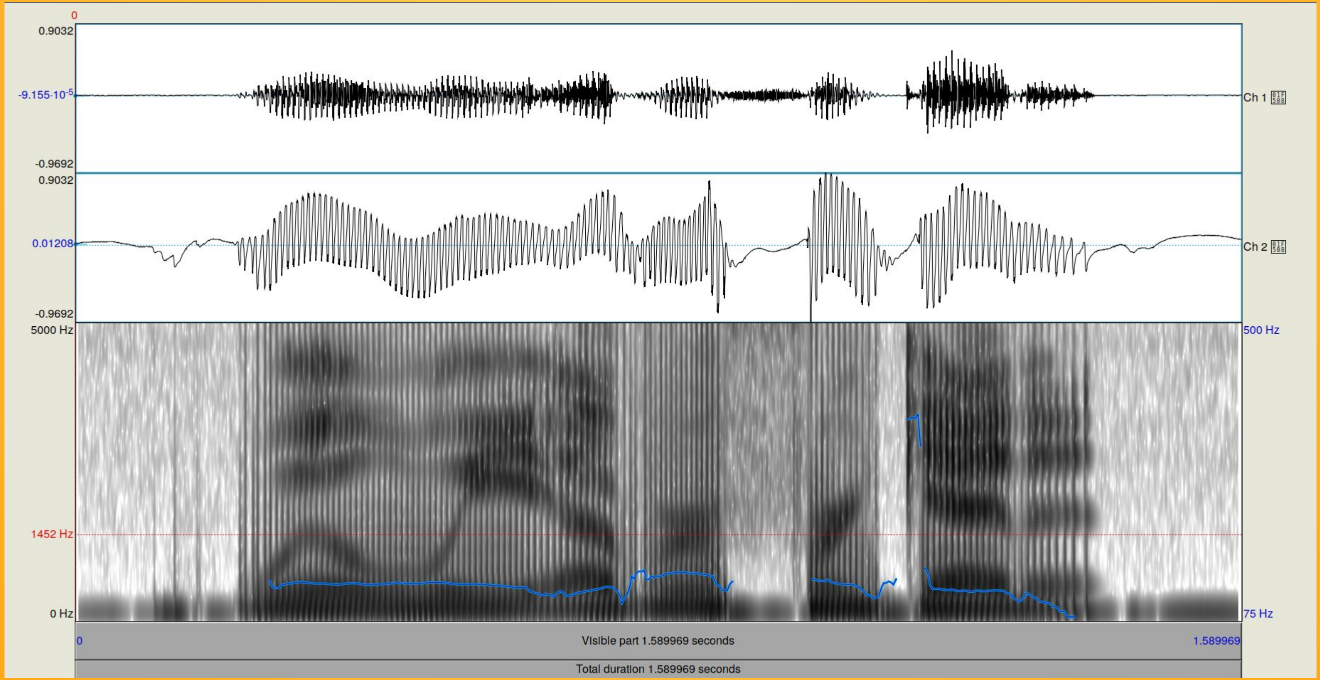
As given in the paper we computed pitch frequency contour using Autocorrelation and cross correlation pitch detection algorithms using **Praat software.**

Autocorrelation



As given in the paper we computed pitch frequency contour using Autocorrelation and cross correlation pitch detection algorithms using **Praat software.**

Crosscorrelation



New (final eval) Video presentation Link:
https://drive.google.com/file/d/172FMeHiIYv3QkISM3NPGyAMbdFpBuUf1/view?usp=sharing

Contributions:

- Matlab Code for VMD part and Optimisation part have been done by both the members equally.
- Both of us experimented with input parameters, results and debugging.
- Both of us helped each other in understanding various concepts involved in the algorithm
- Presentation slides mostly done by Shivani, some are done by Siddharth.

  We both studied all the concepts and steps but for the sake of presenting in the video, we have shared the topics randomly.

References

1. http://www.festvox.org/cmu_arctic/
2. https://ieeexplore.ieee.org/document/7369574 (main)
3. Y. Li, B. Xue, H. Hong, and X. Zhu, “Instantaneous pitch estimation based on empirical wavelet transform,” in 19th International Conference on Digital Signal Processing. IEEE, 2014, pp. 250–253.
4. A. Upadhyay and R.B. Pachori, “Instantaneous voiced/non-voiced detection in speech signals based on variational mode decomposition,” Journal of the Franklin Institute, vol. 352, no. 7,pp.2679–2707, 2015.
5. K. Dragomiretskiy and D. Zosso, “Variational mode decomposition,” IEEE Transactions on Signal Processing, vol. 62, no. 3, pp. 531–544, Feb. 2014.
6. https://en.wikipedia.org/wiki/Pitch_detection_algorithm
7. https://ccrma.stanford.edu/~pdelac/154/m154paper.htm
8. https://sci-hub.do/10.1016/j.jfranklin.2013.01.002

THANK YOU!

Please find 1 matlab code file, named 'curiosity_new.m' (zipped while submission)

Old (proj 1 eval) Video presentation Link
https://drive.google.com/file/d/10FXc8ZIRr_qIFjlleSVzH3vUzpqI7oYO/view

New (final proj eval) Video presentation Link
https://drive.google.com/file/d/172FMeHiIYv3QkISM3NPGyAMbdFpBuUf1/view