1 LP residual

Linear prediction (LP) analysis uses the past P number of samples to predict the current sample. Minimizing the mean squared error gives LP coefficients (a_k's).

$$\hat{x}(n) = \sum_{k=1}^{P} a_k x(n-k) \tag{1}$$

$$e(n) = \sum_{k=1}^{lengthofthesignal} x(n) - \hat{x}(n) \tag{2}$$

Minimizing the squared error of e(n) would give optimal a_k's

$$argmin(e^2(n))_{a_k} \tag{3}$$

so in frequency domain the output of filtering speech signal with the obtained coefficients can be seen as

$$E(z) = H(z)S(z) \tag{4}$$

where $H(z)$ is the frequency response of filter obtained from LP analysis, and $S(z)$ is the frequency response of speech signal and $E(z)$ is frequency response of $e(n)$. The $e(n)$ is called as LP residual. A sample speech signal and the LP residual obtained form LP analysis with $P = 10$ is shown in Figure 1.

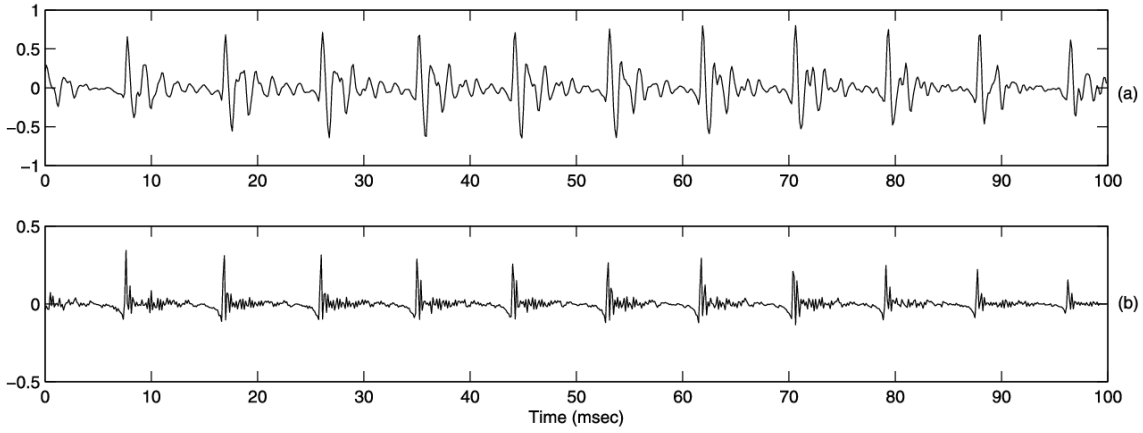

Figure 1: The speech signal (a) and its corresponding LP residual (b)

2 Glottal volume velocity

Glottal volume velocity (GVV) signal was extracted using Quasi Closed Phase (QCP) analysis. The vocal tract transfer function is estimated using a weighted linear prediction analysis of closed phase regions of Glottal cycle. The obtained estimate of vocal tract system is then used for inverse filtering to obtain the GVV signal. Figure 3 shows a sample speech signal and the GVV signal obtained from it.

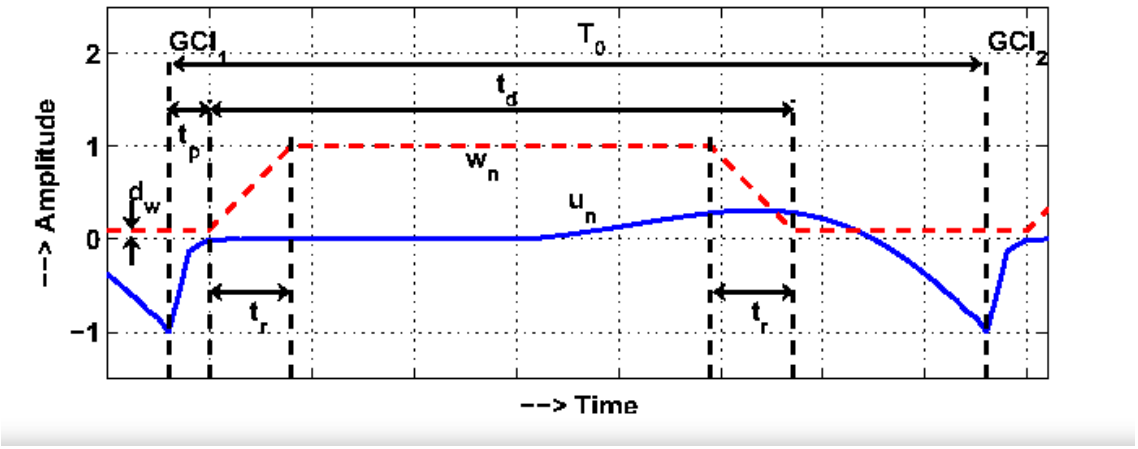

Figure 2: The weight function (dotted line) and glottal flow derivative signal

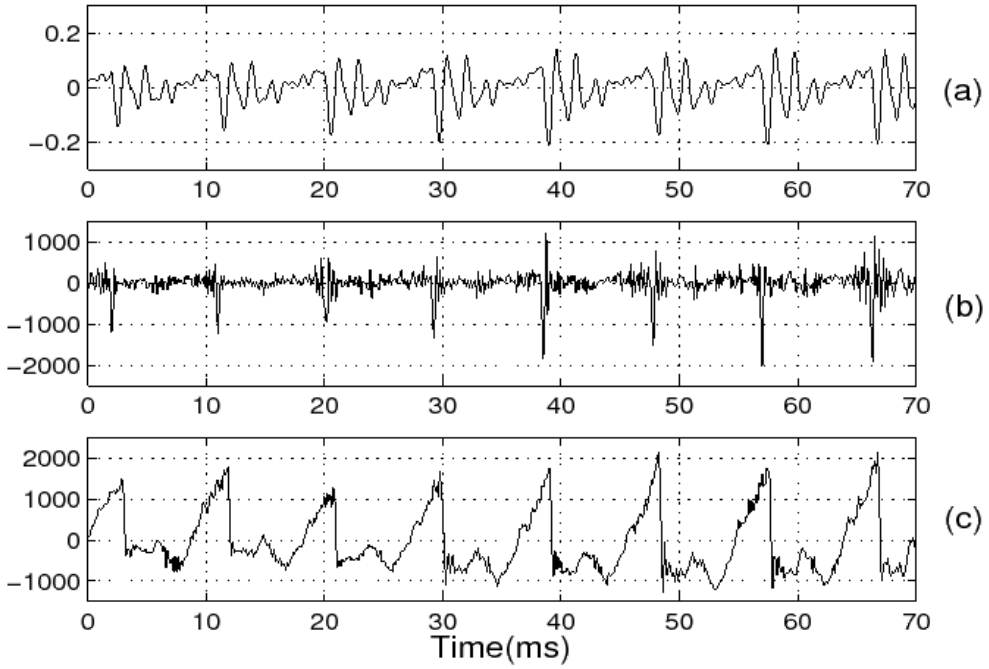

Figure 3: (a) speech signal, The residual obtained from inverse filtering and the corresponding GVV signal (b) and (c) respectively .

3 ZFF evidence

The zero frequency filter or ZFF is a low pass filter. A careful insight will reveal that it is just an cascaded integrator, which is defined by the impulse response of a ramp function. it frequency response is given by the equation

$$H(z) = \frac{1}{(1 - z^{-1})^2} \tag{5}$$

The frequency response of ZFF is shown in Figure 4. The output of ZFF filter is passed through a trend removal filter. A trend removal filter calculates the average across the window, symmetric about one sample, and subtracts it from the every sample. Its transfer is shown in equation

$$h(n) = \delta(n) - \frac{1}{N} \sum_{i=n-\frac{N-1}{2}}^{n+\frac{N}{2}} x(i) \tag{6}$$

Where the 'N' is the average periodicity of signal calculated from the auto-correlation function. Its frequency response is shown in Figure 5. The combination of the ZFF and trend removal filter is an ban pass filter. which has peak at frequency obtained form the calculated average periodicity across the signal.

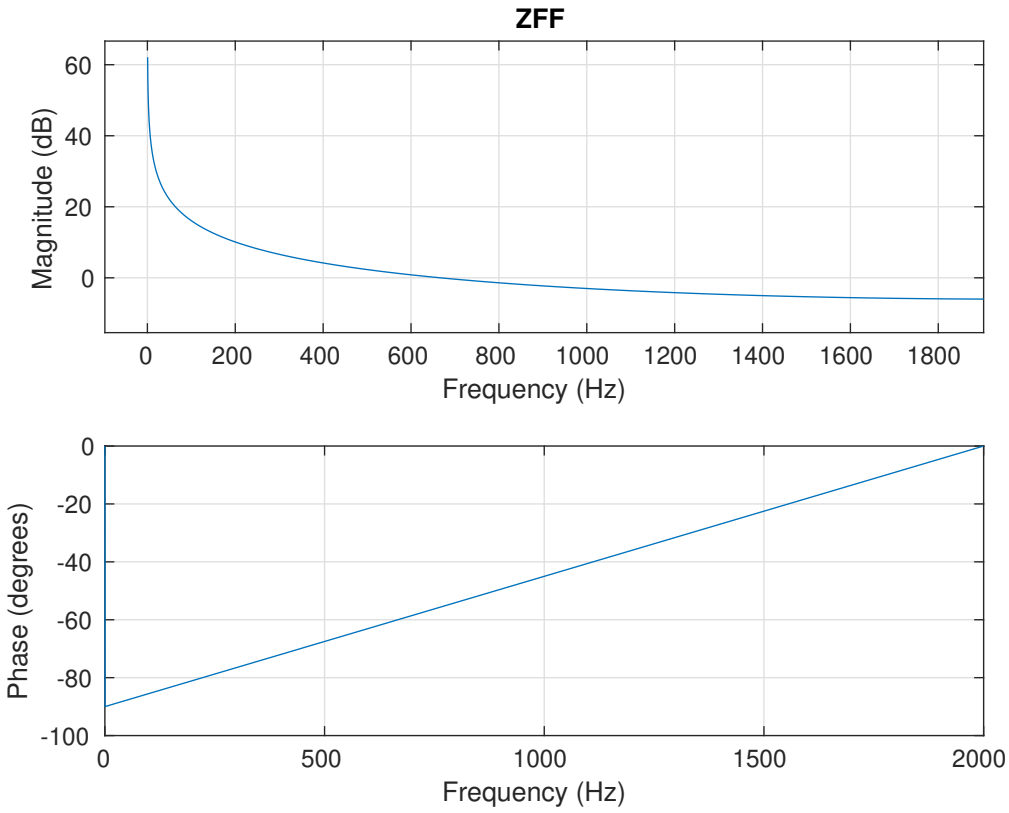

Figure 4: The frequency response of ZFF

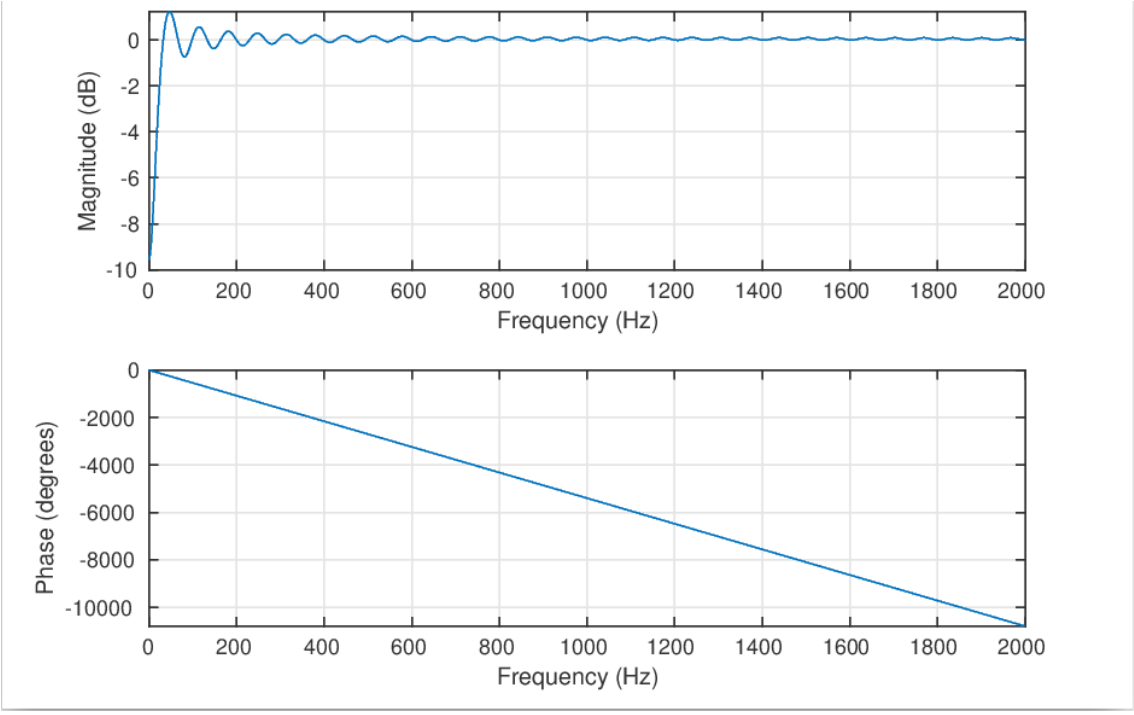

Figure 5: The frequency response of cascaded ZFF and Trend removal filter.

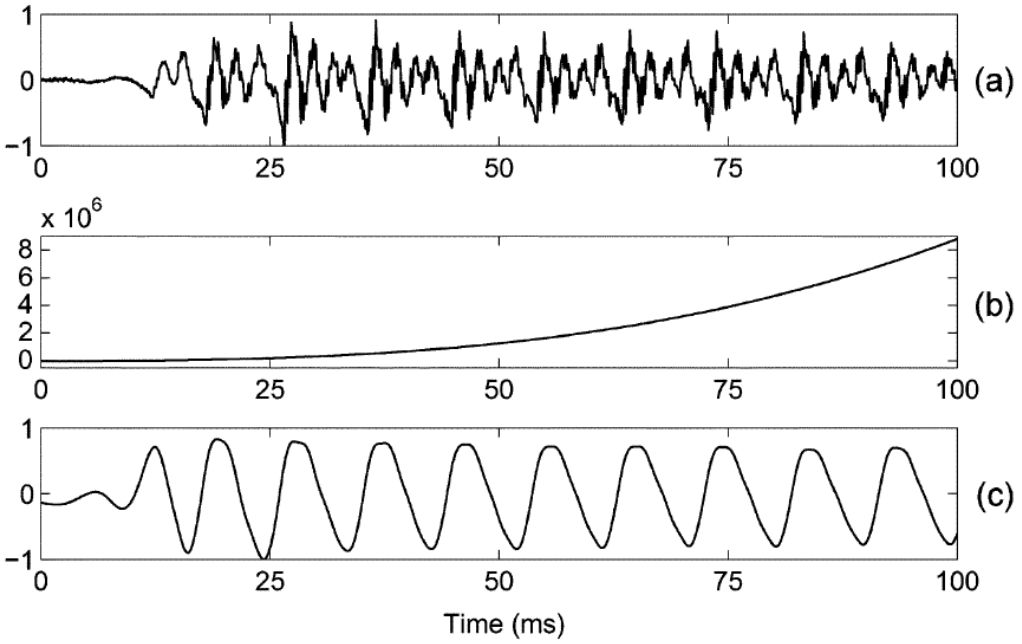

Figure 6: (a) speech signal (b) The ZFF output (c) ZFF output after trend removal referred as ZFF evidence

When a speech signal is passed through ZFF it gives exponentially growing or decaying signal, as zff is just an cascaded integrator. After the trend removal operation it looks like an sinusoidal signal,this will be referred to as ZFF evidence. It is demonstrated using a sample speech signal and shown in the Figure 6. The zero crossings of this ZFF evidence correspond to the epoch locations.

4 Cepstral coefficients and STAT

Cepstrum of the obtained excitation source evidence, LP residual, GVV and ZFF evidence, was derived using a 512 point DFT with 30 ms frame size and 15 ms frame shift. Then a triangular filter bank with 26 filters from 0 to 4000 Hz with Mel-frequency spacing was used to obtain first 12 cpestral coefficients. They are referred as LPRCC, GVVCC, and ZFFCC in our work.

The statistical parameters are extracted from these 26 coefficients. They are

- Mean across the frames -12
- Standard deviation across all the frames -12
- Skewness -12
- Kurtosis -12
- range -12
- mean absolute value of difference 12

This results in a 72 dimensional feature vector and this is referred as Statistical Averaging Across Time (STAT) in the work. The features are obtained are referred as LPCC_stat, ZFCC_stat and GVVCC_stat.

5 Single frequency filter bank based long-term average spectral features

5.1 Single frequency filter bank

The speech signal $s[n]$ is pre-emphasised and passed through the SFFB to decompose it into multiple frequency components. The SFFB is given by,

$$H_{SFFB}(z) = \{H_1(z),\ H_2(z),\ ....\ H_k(z)\ ....\ H_M(z)\} \tag{7}$$

Here M represents the number of frequency components to be decomposed, and $H_k(z) = \frac{1}{1-a_k z^{-1}}$ represent the filter transfer function of $k_t h$ frequency band, where $a_k = ae^{-jw_k}$, represents pole location $z-$plane and w_k denotes the k^{th} frequency component. For the decomposition of speech signal into multiple frequency components, value of $a = 0.98$ and step size of 20 Hz were used. From Figure 7, it can be noticed that the frequency spread is less in case of SFFB based time spectral representation as compared to short-time Fourier representation. It is also observed that better time-frequency localization can be observed in SFFB.

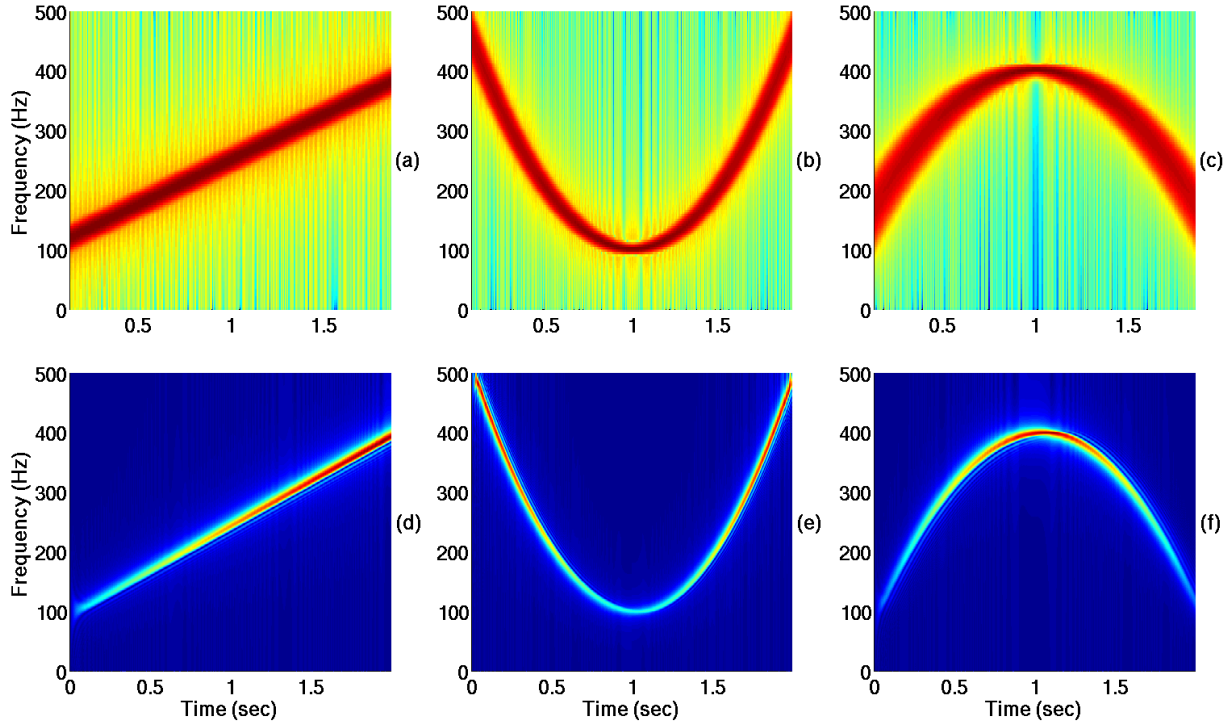

Figure 7: Time-frequency representation of synthesized linear, quadratic, and convex chirp signals. Short time Fourier transform (Top row: (a)-(c)). Single frequency filtering (Bottom row: (d)-(f)).

5.2 Single frequency filter bank based long-term average spectral features (SFFB-LTAS)

The speech signal $s[n]$ is filtered into M components with uniformly spaced center frequencies of 20, 40, 60, ... $fs/2$ Hz using SFFB (with a frequency step of $20Hz$ and $fs = 8000Hz$). Each of the $M + 1$ components (including the raw speech signal $s[n]$), $s_i[n]$, $i = 0, 1, 2, ... M$, is segmented with an non-overlapping window of 20 ms. For each frame, the root mean square (RMS) is calculated and denoted as $S_{RMS_i}[k]$ for k^{th} frame of i^{th} band. Finally, for each of the $M+1$ components following spectral features are computed and these are referred to as single frequency filter bank based long-term average spectral features.

1. The RMS value normalized by the full-band RMS value, $\frac{rms\{s_i[n]\}}{rms\{s_0[n]\}}$
2. The normalized mean frame RMS, $\frac{mean\{S_{RMSi}[k]\}}{rms\{s_0[n]\}}$
3. The standard deviation of frame RMS, $std\{S_{RMSi}[k]\}$
4. The frame standard deviation normalized by full-band RMS, $\frac{std\{S_{RMSi}[k]\}}{rms\{s_0[n]\}}$
5. The frame standard deviation normalized by band RMS, $\frac{std\{S_{RMSi}[k]\}}{rms\{s_i[n]\}}$
6. The skewness of frame RMS, $skew\{S_{RMSi}[k]\}$
7. The kurtosis of frame RMS, $kurt\{S_{RMSi}[k]\}$
8. The range of frame RMS, $range\{S_{RMSi}[k]\}$
9. The normalized range of frame RMS, $\frac{range\{S_{RMSi}[k]\}}{rms\{s_0[n]\}}$
10. Pairwise variability of RMS energy between ensuing frames, $\frac{mean(\{S_{RMSi}[k]\}-\{S_{RMSi}[k-1]\})}{rms\{s_0[n]\}}$

The SFFB decomposes speech into multiple frequency $M = 200$ bands in a frequency range of 0 to $4000Hz$ with a frequency step of $20Hz$. For M components, $(M)*10 - 1$ spectral features will be extracted. By including full-band speech signal a total 201 components will be present. This results in $2019-$dimensional SFFB-LTAS feature vector.

.