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).

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

$$e(n) = \sum_{k=1}^{length of the signal} x(n) - \hat{x}(n) \tag{2}$$

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

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

The e(n) is called as LP residual. A sample speech signal and the LP residual obtained form LP analysis with $P = 10$.

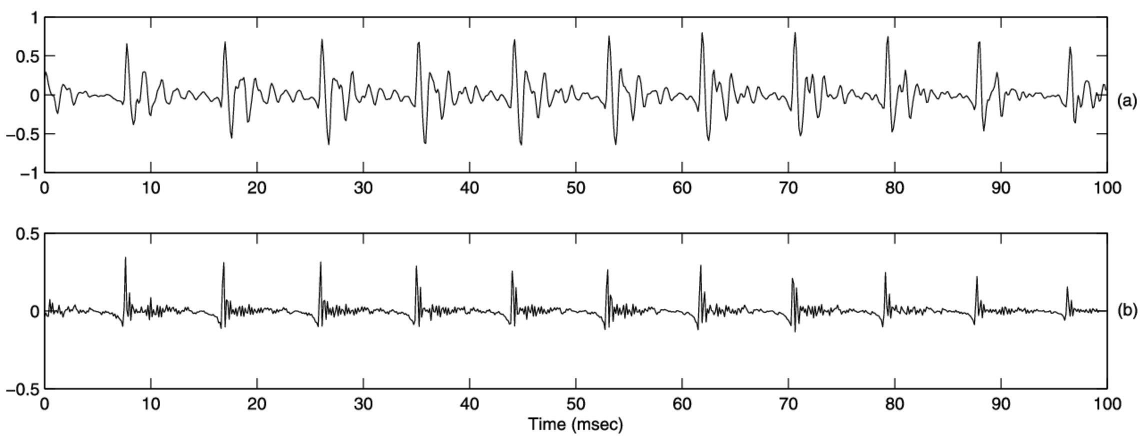

Figure 1: The speech signal (a) and its corresponding LP residual (b)

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.

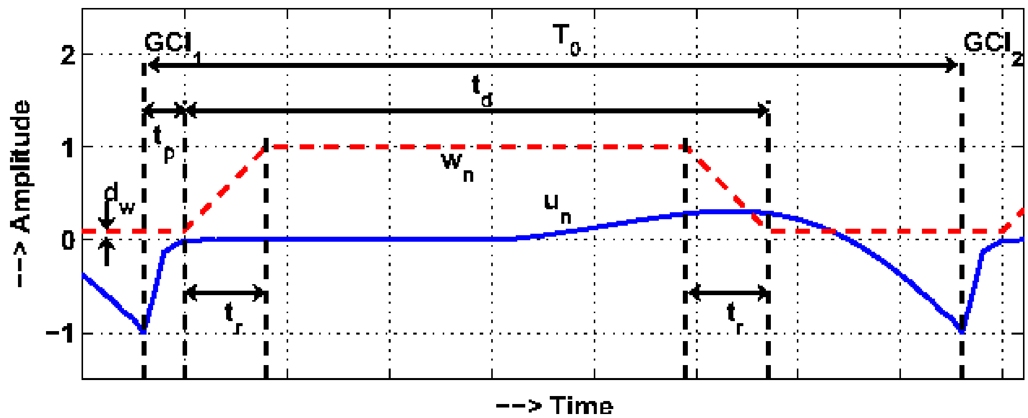

Figure 2: The weight function (dotted line) and glottal flow derivative signal

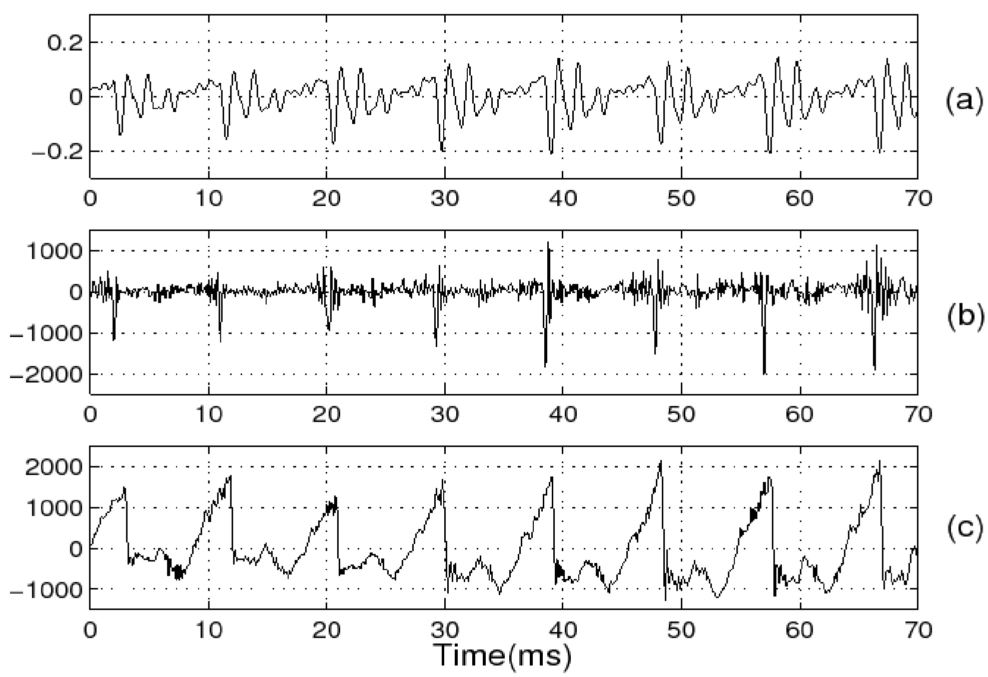

Figure 3: (a) speech signal, The residual obtained from inverse filtering and the corresponding GVV signal (b) and (c) respectively .

3 ZFF evidence

The zero frequency filter or ZFF is a band 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. Its 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.

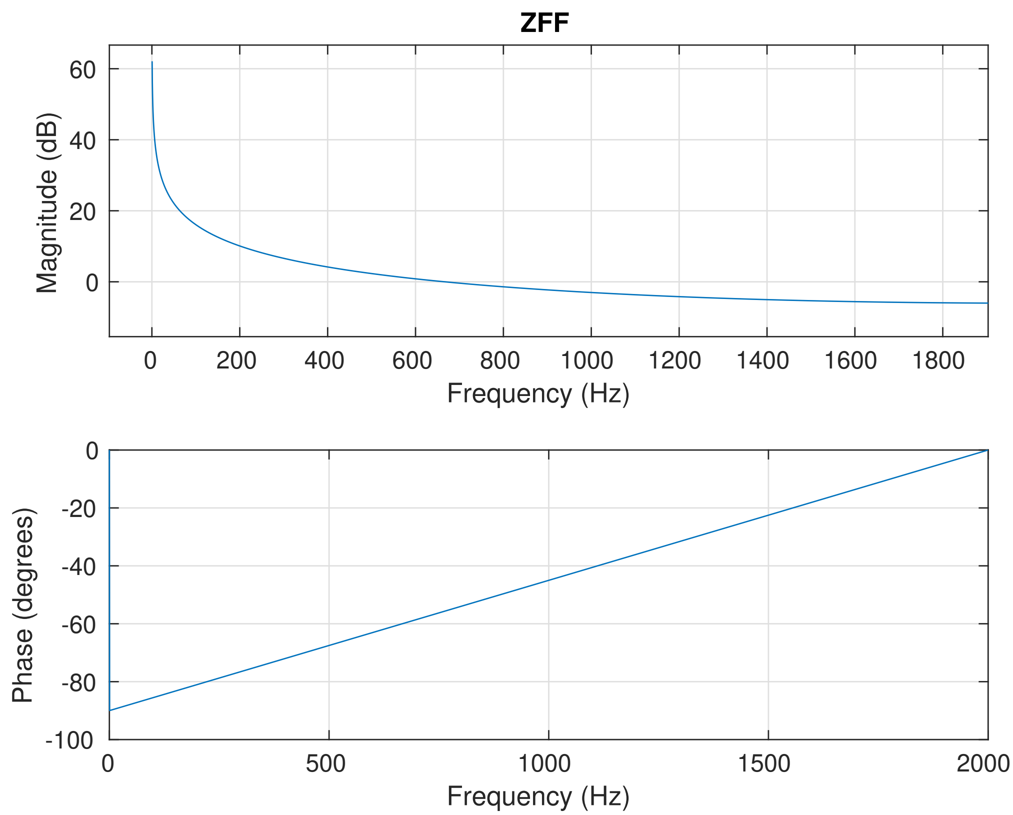

Figure 4: The frequency response of ZFF

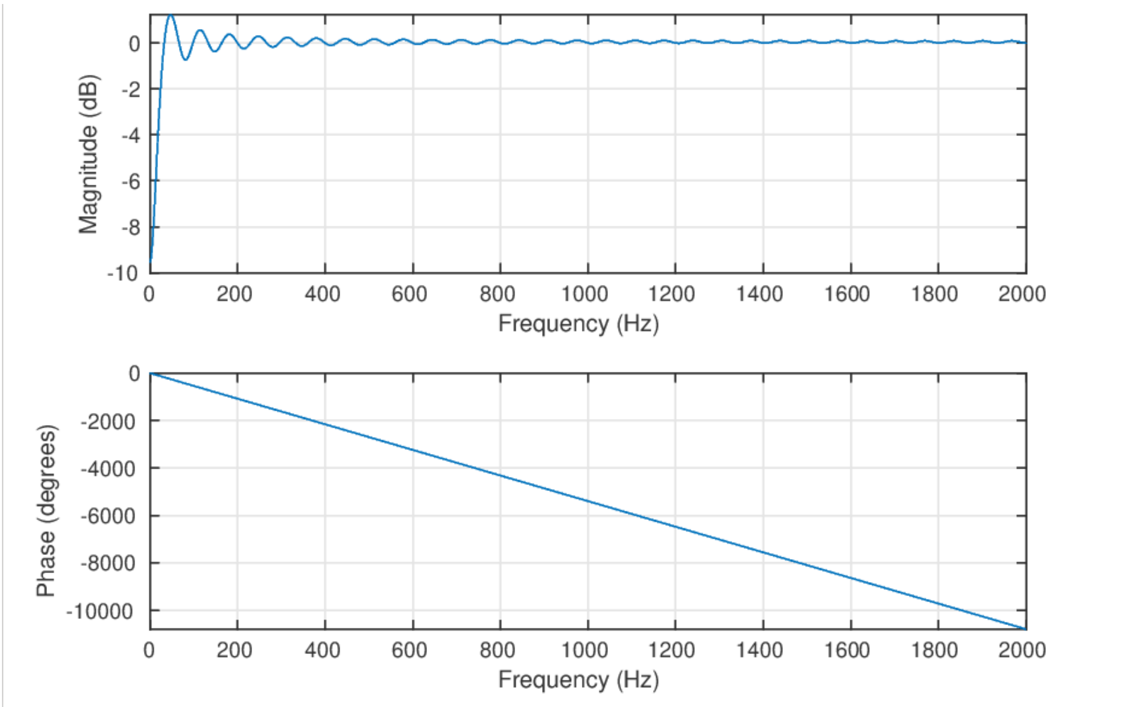

Figure 5: The frequency response of cascaded ZFF and Trend removal filter.

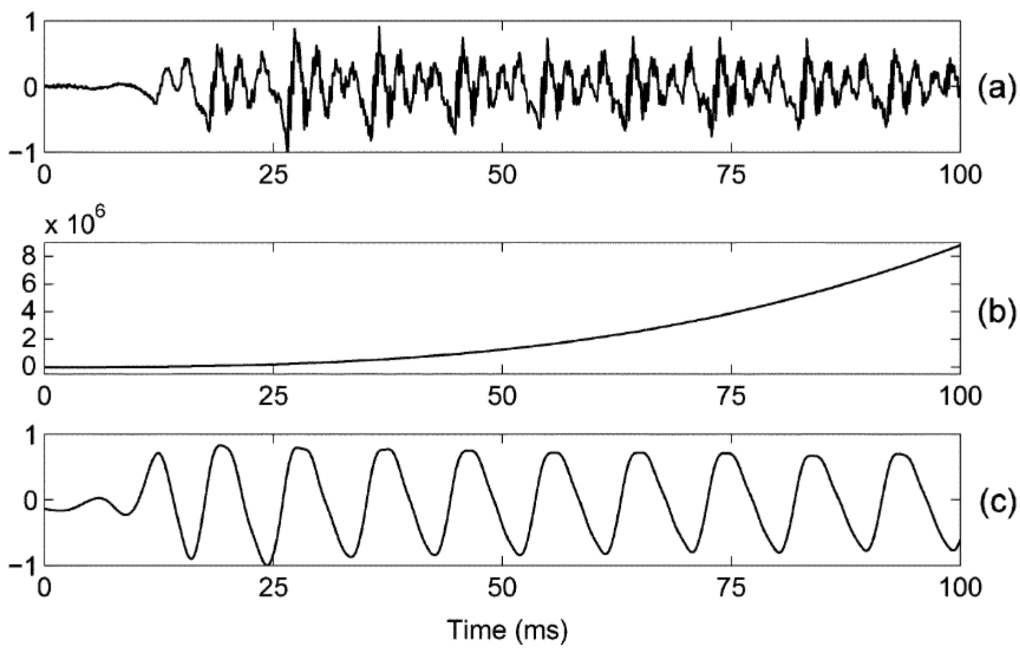

Figure 6: (a) speech signal (b) The ZFF output (c) ZFF output after trend removal referred as ZFF evidence

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 [1].

4 Glottal feature

Glottal volume velocity waveform (GVV) is calculated from QCP method as discussed in section . From GVV Glottal features are set of 12 features (9 time-domain and 3 frequency-domain features) are used in this study to characterize the glottal flow waveforms estimated by glottal inverse filtering methods. Statistics are applied to time domain, frequency domain features and on its difference, they are: Mean, standard deviation, median, minima, maxima, range, skewness and kurtosis. 16 statistics are applied to time 12 dimensional features makes total of 192 features. The list of glottal features are shown in figure below.

	Time-domain features
OQ1	Open quotient, calculated from the primary glottal opening
OQ2	Open quotient, calculated from the secondary glottal opening
NAQ	Normalized amplitude quotient
AQ	Amplitude quotient
ClQ	Closing quotient
OQa	Open quotient, derived from the LF model
QoQ	Quasi-open quotient
SQ1	Speed quotient, calculated from the primary glottal opening
SQ2	Speed quotient, calculated from the secondary glottal opening
	Frequency-domain features
H1-H2	Amplitude difference between the first two glottal harmonics
PSP	Parabolic spectral parameter
HRF	Harmonic richness factor

Figure 7: Time-domain and Frequency-domain glottal features derived from glottal flows estimated by QCP analysis

5 openSMILE- ComParE Feature set

COMPARE is a large brute-forced acoustic feature set containing 6373 static features (i. e. functionals) of low-level descriptor (LLD) contours. 65 LLD contours used in this set are shown in figure. The functionals applied to the LLD contours include the mean, standard

4 energy related LLD	**Group**
Sum of auditory spectrum (loudness)	prosodic
Sum of RASTA-filtered auditory spectrum	prosodic
RMS Energy, Zero-Crossing Rate	prosodic
55 spectral LLD	**Group**
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	spectral
MFCC 1–14	cepstral
Spectral energy 250–650 Hz, 1 k–4 kHz	spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	spectral
Spectral Flux, Centroid, Entropy, Slope	spectral
Psychoacoustic Sharpness, Harmonicity	spectral
Spectral Variance, Skewness, Kurtosis	spectral
6 voicing related LLD	**Group**
F_0 (SHS & Viterbi smoothing)	prosodic
Prob. of voicing	voice qual.
log. HNR, Jitter (local & δ), Shimmer (local)	voice qual.

Figure 8: ComParE acoustic feature set: 65 provided low-level descriptors(LLD)

deviation, percentiles and quartiles, linear regression functionals, and local minima/maxima related functionals are shown in figure .

6 openSMILE- eGeMAPS Feature set

extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) is a small (low dimensional) knowledge-based acoustic feature sets containing 88 features. Functionals are applied to 45 LLD.

Functionals applied to LLD / Δ LLD	Group
quartiles 1–3, 3 inter-quartile ranges	percentiles
1 % percentile ($\approx$ min), 99 % pctl. ($\approx$ max)	percentiles
percentile range 1 %–99 %	percentiles
position of min / max, range (max – min)	temporal
arithmetic mean[1], root quadratic mean	moments
contour centroid, flatness	temporal
standard deviation, skewness, kurtosis	moments
rel. dur. LLD is above 25 / 50 / 75 / 90 % range	temporal
relative duration LLD is rising	temporal
rel. duration LLD has positive curvature	temporal
gain of linear prediction (LP), LP Coeff. 1–5	modulation
mean, max, min, std. dev. of segment length[2]	temporal
Functionals applied to LLD only	**Group**
mean value of peaks	peaks
mean value of peaks – arithmetic mean	peaks
mean / std.dev. of inter peak distances	peaks
amplitude mean of peaks, of minima	peaks
amplitude range of peaks	peaks
mean / std. dev. of rising / falling slopes	peaks
linear regression slope, offset, quadratic error	regression
quadratic regression a, b, offset, quadratic err.	regression
percentage of non-zero frames[3]	temporal

Figure 9: Functionals applied to ComParE Feature set [1]: arithmatic mean of LLD [2]: not applied to voicing related LLD except F0 [3]: only applied to F0

1 energy related LLD	Group
Sum of auditory spectrum (loudness)	Prosodic
25 spectral LLD	**Group**
α ratio (50–1 000 Hz / 1-5 k Hz)	Spectral
Energy slope (0–500 Hz, 0.5–1.5 k Hz)	Spectral
Hammarberg index	Spectral
MFCC 1–4	Cepstral
Spectral Flux	Spectral
6 voicing related LLD	**Group**
F0 (Linear & semi-tone)	Prosodic
Formants 1, 2, (freq., bandwidth, ampl.)	Voice Quality
Harmonic difference H1–H2, H1–A3	Voice Quality
log. HNR, Jitter (local), Shimmer (local)	Voice Quality

Figure 10: eGeMAPS acoustic feature set: 45 provided low-level descriptors(LLD)

References

1. Murty, K. Sri Rama, and Bayya Yegnanarayana. "Epoch extraction from speech signals." IEEE Transactions on Audio, Speech, and Language Processing 16.8 (2008): 1602-1613.

2. Airaksinen, Manu, et al. "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction." IEEE/ACM Transactions on Audio, Speech, and Language Processing 22.3 (2013): 596-607

3. Kadiri, Sudarsana Reddy, and Paavo Alku. "Analysis and Detection of Pathological Voice using Glottal Source Features." IEEE Journal of Selected Topics in Signal Processing (2019).

4. Alku, Paavo, and Erkki Vilkman. "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering." Speech communication 18.2 (1996): 131-138.

5. Holmberg, Eva B., Robert E. Hillman, and Joseph S. Perkell. "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice." The Journal of the Acoustical Society of America 84.2 (1988): 511-529.

6. Alku, Paavo, Helmer Strik, and Erkki Vilkman. "Parabolic spectral parameter—a new method for quantification of the glottal flow." Speech Communication 22.1 (1997): 67-79.

7. Titze, Ingo R., and Johan Sundberg. "Vocal intensity in speakers and singers." the Journal of the Acoustical Society of America 91.5 (1992): 2936-2946.