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

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

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

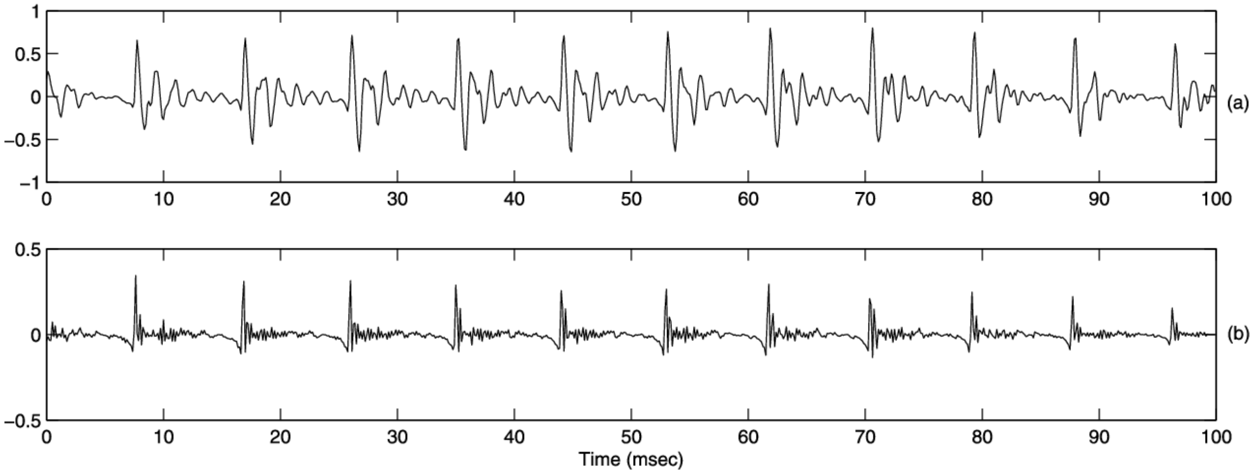


Figure 1: The speech signal (a) and its corresponding LP residual (b)

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.

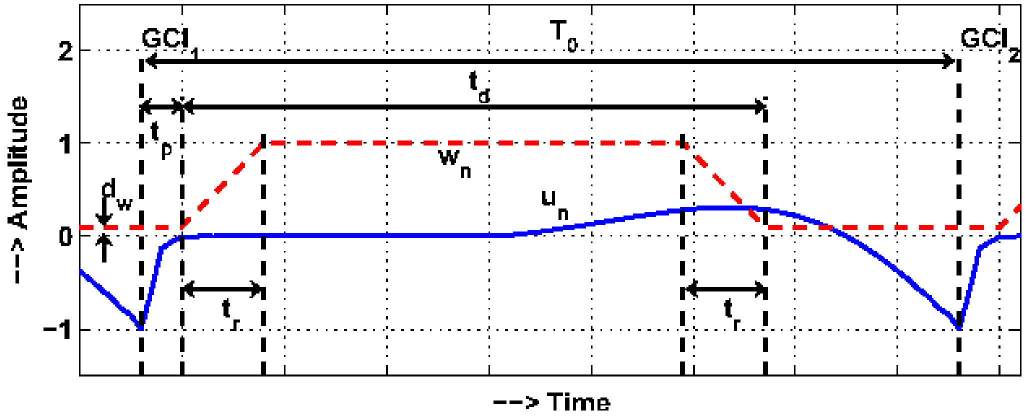


Figure 2: The weight function (dotted line) and glottal flow derivative signal

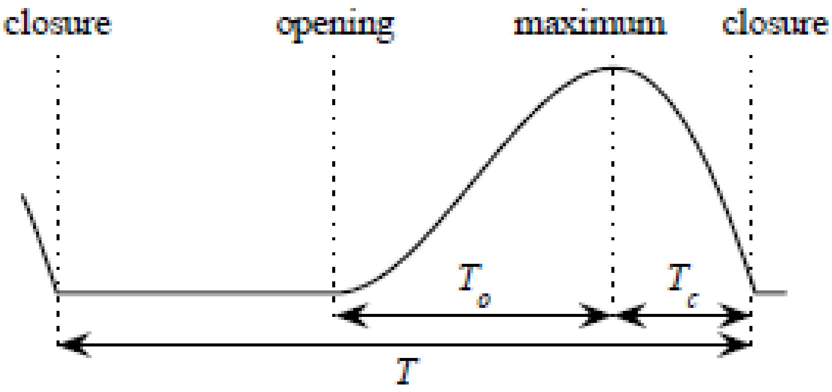


Figure 3: (a) speech signal, The residual obtained from inverse filtering and the corresponding GVV signal (b) and (c) respectively .

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.

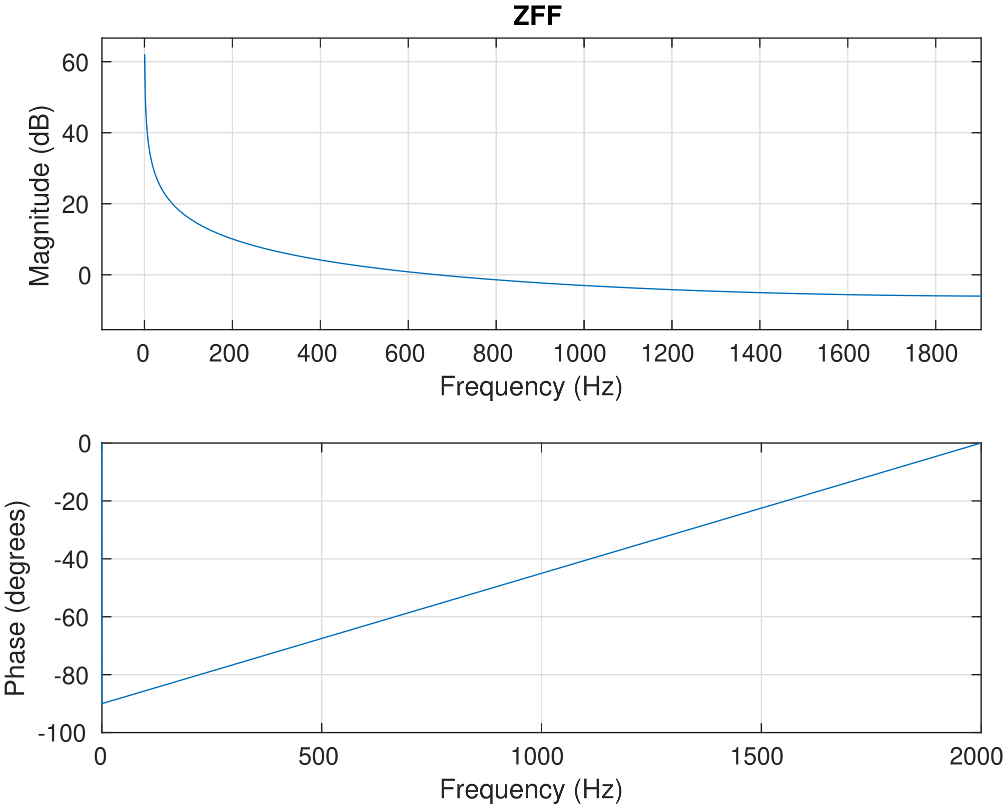


Figure 4: The frequency response of ZFF

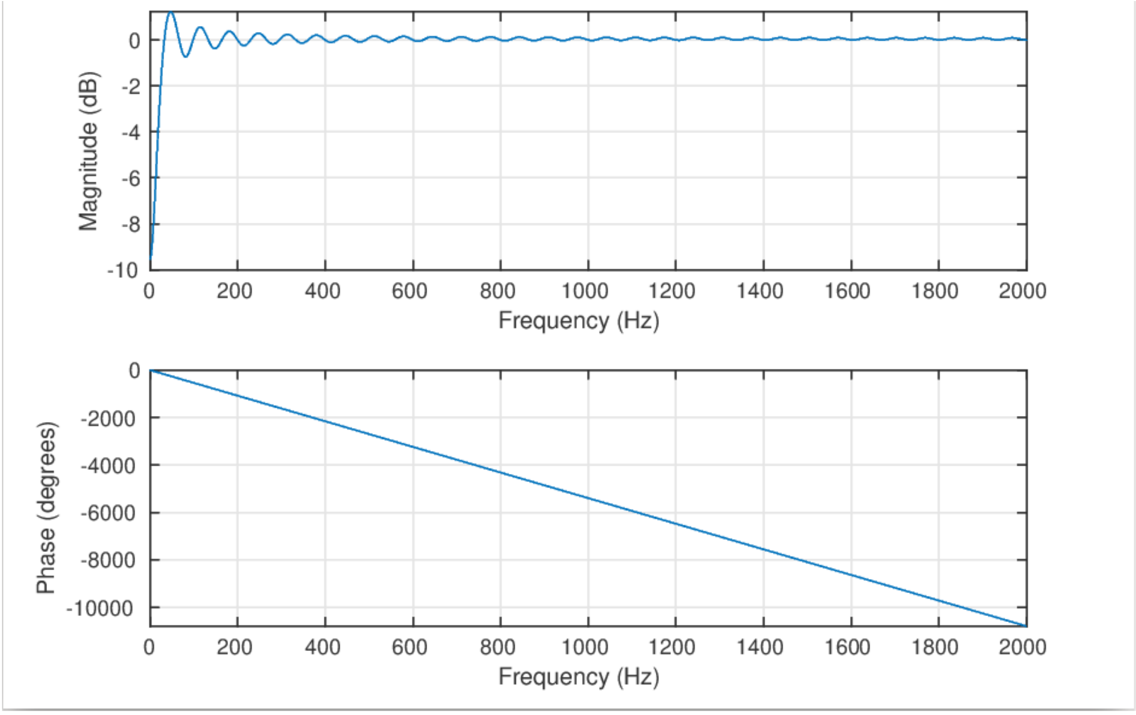


Figure 5: The frequency response of cascaded ZFF and Trend removal filter.

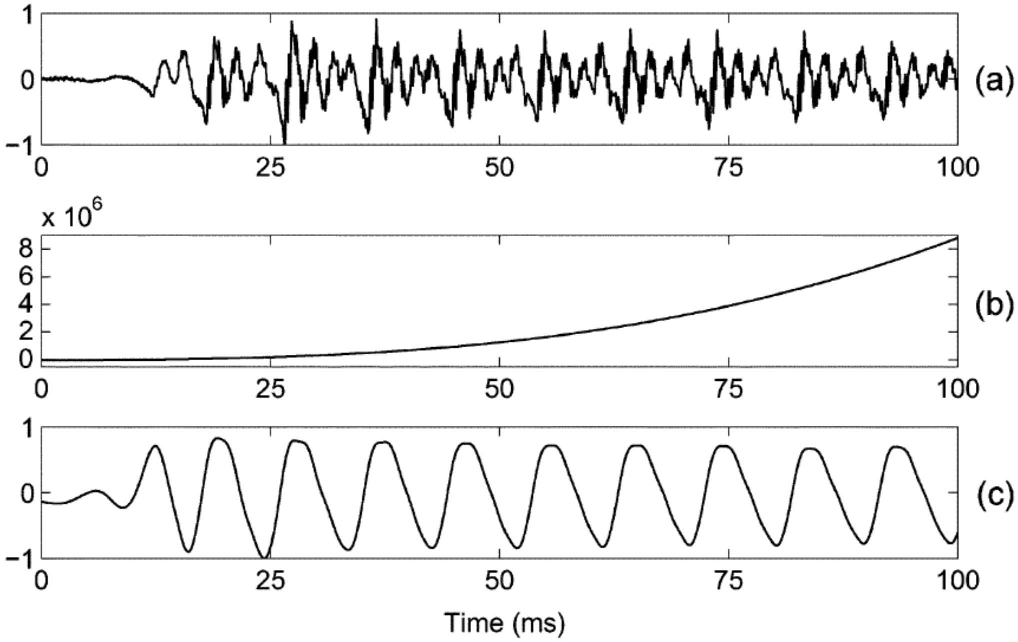


Figure 6: (a) speech signal (b) The ZFF output (c) ZFF output after trend removal referred as ZFF evidence

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 . The instant of closure is followed by a relatively short closed phase with nearly constant flow. Then, the flow starts to increase gently. This phase ends abruptly at a knee that begins a segment with more rapid flow increase. Thus, there are two instants that could be considered instants of glottal opening. From now on, these instants are referred to as the primary opening, which is the end of the horizontal phase, and the secondary opening, which is the instant of abrupt increase of flow derivative.

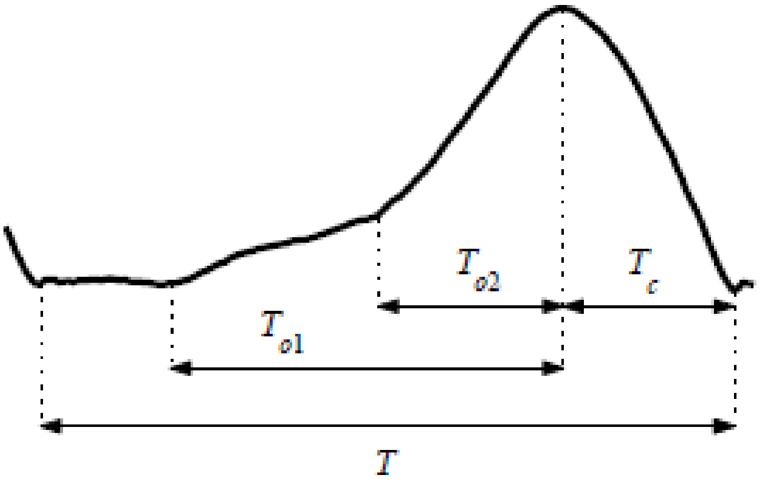


Figure 7: Glottal flow pulse with primary and secondary opening. The length of the glottal cycle is denoted by T. The interval from the primary opening to the instant of maximum flow is indicated by To1 and the interval from the secondary opening to the instant of maximum flow by To2. The closing phase length is denoted by Tc.

Open quotient (OQ1): It is defined as the ratio of the length of primary opening phase to the total length of the glottal cycle.

$$OQ = \frac{T01 + Tc}{T} \tag{7}$$

Open quotient (OQ2): It is defined as the ratio of the length of secondary opening phase to the total length of the glottal cycle.

$$OQ = \frac{T02 + Tc}{T} \tag{8}$$

Closing quotient (CIQ): It is the ratio of the closing phase duration to the glottal cycle length.

$$CIQ = \frac{Tc}{T} \tag{9}$$

Speed quotient (SQ1): It is defined as the ratio of the length of the primary opening phase to the closing phase length.

$$SQ1 = \frac{T01}{Tc} \tag{10}$$

Speed quotient (SQ2): It is defined as the ratio of the length of the secondary opening phase to the closing phase length.

$$SQ2 = \frac{T02}{Tc} \tag{11}$$

Amplitude quotient (AQ): It is defined as ratio between the AC-amplitude of the flow signal and the amplitude of the minimum of the differentiated flow.

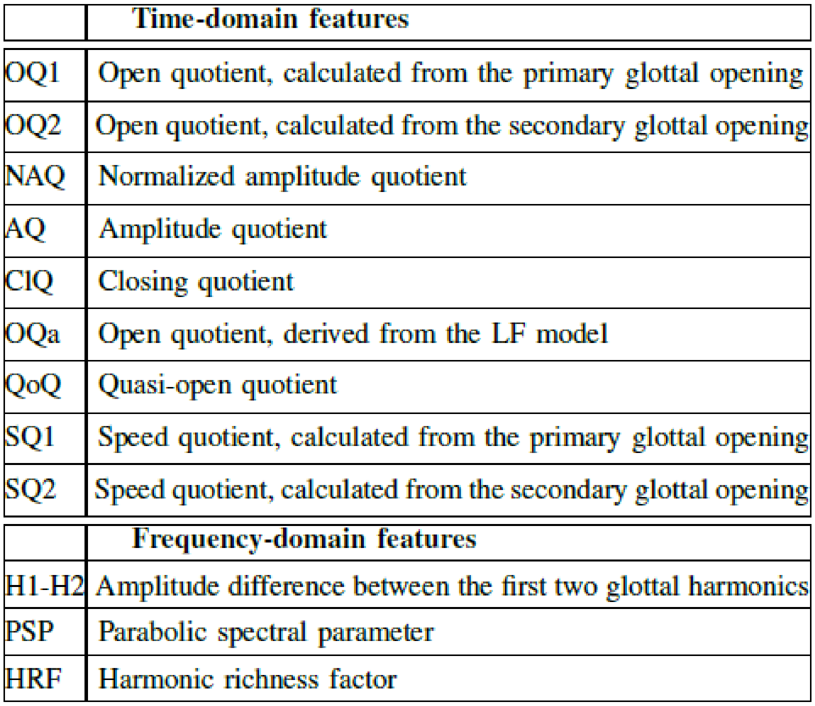
$$AQ = \frac{Aac}{dmin} \tag{12}$$

H1-H2: It is the difference between the values of the first and second harmonic of glottal signal.

Parabolic spectral parameter: PSP is based on fitting a parabolic function to the low-frequency part of a pitch-synchronously computed spectrum of the estimated glottal flow. PSP gives a single numerical value that describes how the spectral decay of an obtained glottal flow behaves with respect to a theoretical limit corresponding to maximal spectral decay.

Harmonic richness factor: It relates the first harmonic (H1) with the sum of the energy of the other harmonics (Hk).

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

5 Intonation Features

Intonation features captures the characteristic of phonation. The knowledge of epoch locations obtained from the speech signal has been used to obtain the mentioned intonation features. In this work, zero-phase zero frequency filtering has been used to extract the epoch locations from speech signal. A 76-dimensional intonation feature vector has been derived using fundamental frequency, harmonic to noise ratio, jitter and shimmer. Strength of excitation, and Energy of excitation are measured around the epoch locations. Statistical measures : 5 features

- Mean
- Median
- Standard deviation
- Minimum
- Maximum

Jitter-shimmer quotients : 22 features

- Mean-absolute-Difference (M-abs-D) of successive cycles : 1
- Ratio of (M-abs-D) and mean of successive cycles : 1

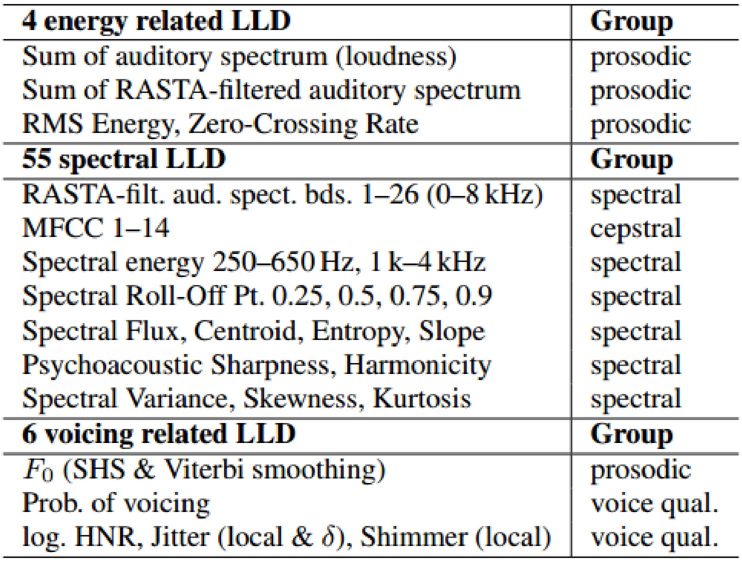
- Perturbation coefficients (Classical Schoentgen, Classical Baken, generalized Schoentgen) for 3 cycle samples :3
- Perturbation coefficients (Classical Schoentgen, Classical Baken, generalized Schoentgen) for 5 cycle samples : 3
- Perturbation coefficients (Classical Schoentgen, Classical Baken, generalized Schoentgen) for 11 cycle samples : 3
- Zeroth order perturbation :1
- Shimmer : 1
- Ratio of mean of absolute of first order difference and mean :1
- Mean, Standard deviation, prictle (5,25,50,75) of Teager-Kaiser energy operator features(TKEO) : 6
- max-min/(max+min) : 1
- TKEO prictle(75)- TKEO prictle(5) : 1

The 76 dimentaional Intonation feature vector includes are:

1. Statistical measures of F0 – (5)
2. Jitter quotients of F0 – (22)
3. Shimmer quotients of strength of excitation (SOE) – (22)
4. Shimmer quotients of Energy of excitation (EOE) – (22)
5. Harmonic to noise ratio and noise to harmonic ratio – (4)
6. F0 dispersion – (1)

6 openSMILE- ComParE Feature set

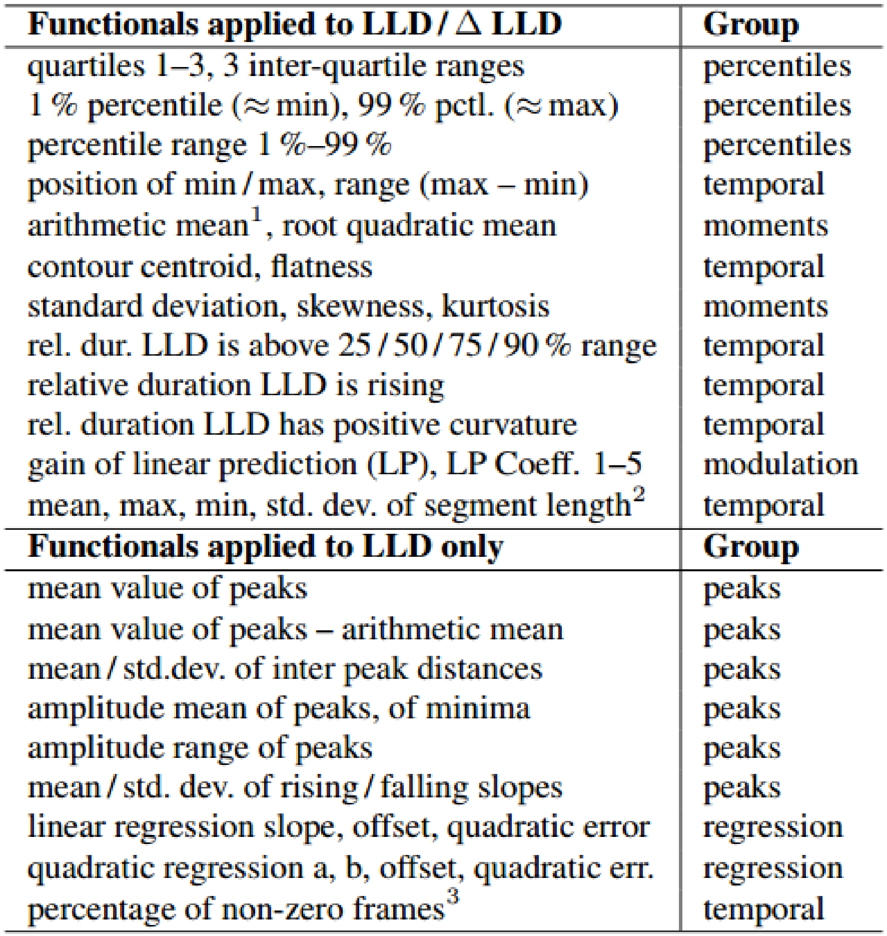
Speech contains both linguistic and non-linguistic or paralinguistic information. Paralinguistic information includes accent, pitch, loudness, speech rate, modulation, intonation, fluency etc. in speech. For voice disorder identification paralinguistic information plays very important role. The 2013 Interspeech Computational Paralinguistics Challenge features set (ComParE) is large-scale (high dimension) brute-forced acoustic feature set contains 6373 static features resulting from the computation of various functional over low-level descriptor (LLD) contours. The low-level descriptors cover a broad set of descriptors (features) from the fields of speech processing, Music Information Retrieval,and general sound analysis. LLDs are feature which are related to low level description of audio information like temporal, spectrum related, voice quality related features. In this set, supra segmental features are obtained by applying a large set of statistical functional to acoustic low-level descriptors. There are 4 energy related parameter(like zero crossing rate,RMS energy,loudness),55 spectral features(MfCC,spectral energy,Spectral variance, skewness, kurtosis) and 6 voicing related features(Jitter,Shimmer,HNR), The statistical functionals applied to the LLD include the mean, standard deviation, percentiles and quartiles, linear regression functionals, quadratic regression and minima/maxima related functionals.



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

The functionals applied to the LLD contours include the mean, standard deviation, percentiles and quartiles, linear regression functionals, and local minima/maxima related functionals are shown in figure .

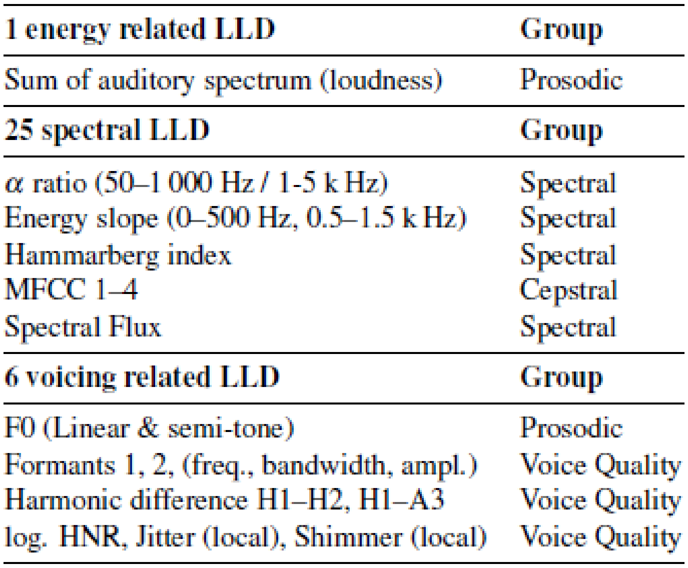


Functionals applied to LLD / Δ LLD	**Group**
quartiles 1–3, 3 inter-quartile ranges	percentiles
1 % percentile (≈ min), 99 % pctl. (≈ max)	percentiles
percentile range 1 %–99 %	percentiles
position of min / max, range (max – min)	temporal
arithmetic mean[1], root quadratic mean	moments
contour centroid, 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 10: 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

7 openSMILE- eGeMAPS Feature set

extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) are small-scale (low-dimension) knowledge-based acoustic feature set contains 88 parameters,these feature set is also designed to extract paralinguistic information from speech with small feature set ComParE to ComParE feature set(6373 features). Functionals are applied to 45 LLD. Frequency related parameter are total of (12) Pitch, Jitter, first three formant frequency and bandwidth of first formant their mean and standard deviations. Energy related parameters are 6 which includes Loudness,Shimmer and Harmonic to noise ratios (HNR) mean and standard deviation. In total it consist of 42 LLD on which two statistical functionals (arithmetic mean and coefficient of variations) is applied makes total of 88 parameters. More details of the feature set is given in



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 11: eGeMAPS acoustic feature set: 42 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.