1 Voice disorder

Voice disorders may occur due to poor respiratory system, or incomplete glottal closure or extra lession on vocal fold or irregularity in the vibration of vocal fold or abnormal vocal fold closure or weakness in muscle which are responsible for voicing or it may be due to psychological reason. The voice disorder can be organic or functional.

1.1 Organic Voice Disorder

Organic voice disorders (OVD) are physiological voice disorder,due to anatomic abnormality in the larynx or muscles stain, which results in incomplete glottal closure. OVD in broad sense can be categorized into two sub-types: Structural and Neurogenic.

1.1.1 Structural voice disorder

Structural disorders are due abnormal or extra growth on vocal folds, which cause irregular glottal phase. Vocal cord polyp, nodules, leukoplakia,laryngitis, are some of the structural voice disorder.

1.1.2 Neurogenic voice disorder

Spasmodic dysphonia and recurrent laryngeal nerve palsy are the main common disorder which falls into category of neurogenic voice disorder. Spasmodic dysphonia, is also known as laryngeal dystonia. Dystonia is neurological disorder in which sudden, involuntary movements (spasm) muscles control occur in body parts. Recurrent laryngeal nerve palsy (RLNP) is a paralysis of RLN either on one side or both sides of vocal folds. If there is no movement at all it is known as paralysis and if movement slows down it is called paresis. The resulting effect of RLNP is that vocal folds do not move close to each other. Voice may sound like breathy and rough.

1.2 Non-organic Voice Disorder

Non-organic also most common known as functional disorder which may be due to increased tension of muscle which cause abnormal laryngeal movement. The phonation in this case is characterize by excessive laryngeal activity, tension, reduced vocal capacity, impaired voice all without any organic abnormality.

1.2.1 Functional voice disorder

Muscle tension voice disorders (MTVD) are also another name of functional voice disorder. MTVD is due to improper coordination of the laryngeal muscle and breathing pattern responsible for producing normal speech.

2 ZFF evidence

The zero frequency filter or ZFF is a low pass filter [1]. 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

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

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

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

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.

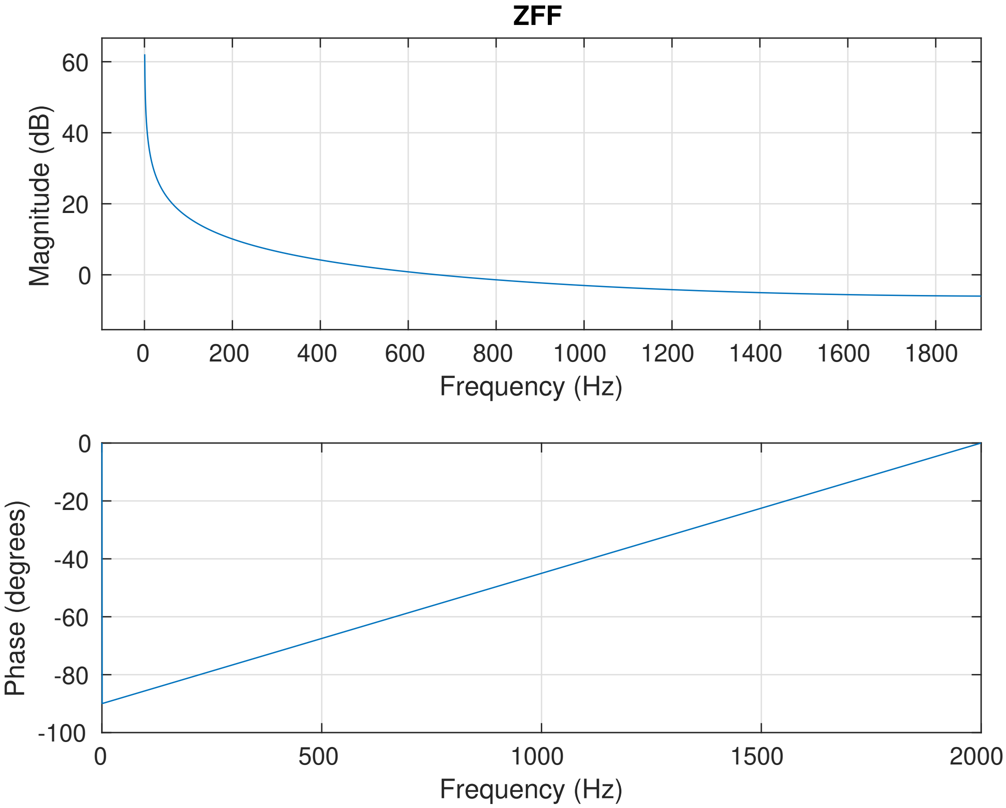


Figure 1: The frequency response of ZFF

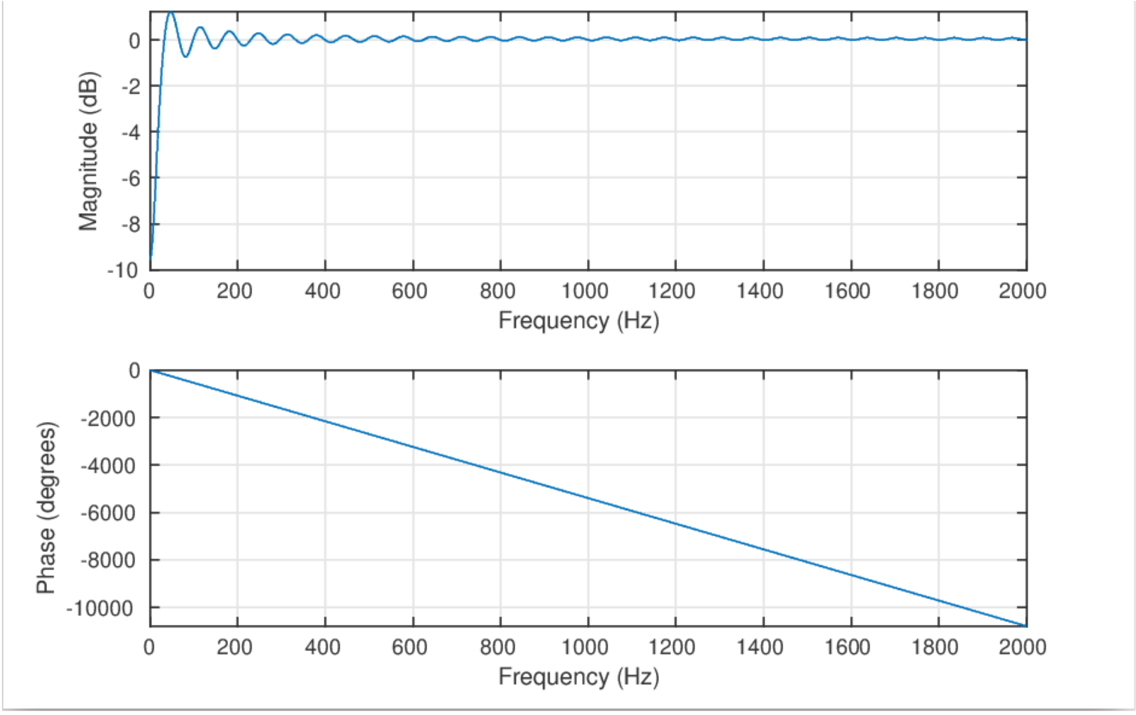


Figure 2: The frequency response of cascaded ZFF and Trend removal filter.

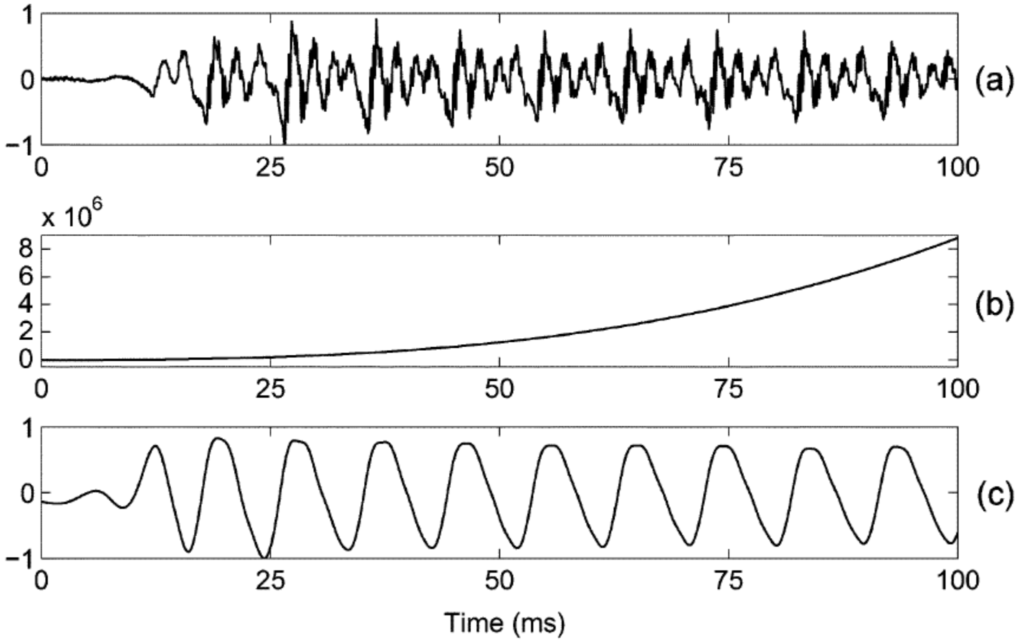


Figure 3: (a) speech signal (b) The ZFF output (c) ZFF output after trend removal referred as ZFF evidence

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.

3 LP residual

Linear prediction (LP) analysis uses the past P number of samples to predict the current sample [2]. Minimizing the mean squared error gives LP coefficients (a_k's).

$$\hat{x}(n) = \sum_{k=1}^{P} a_k x(n-k) \tag{3}$$

$$e(n) = \sum_{k=1}^{length of the signal} x(n) - \hat{x}(n) \tag{4}$$

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

$$argmin(e^2(n))_{a_k} \tag{5}$$

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

$$E(z) = H(z)S(z) \tag{6}$$

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.

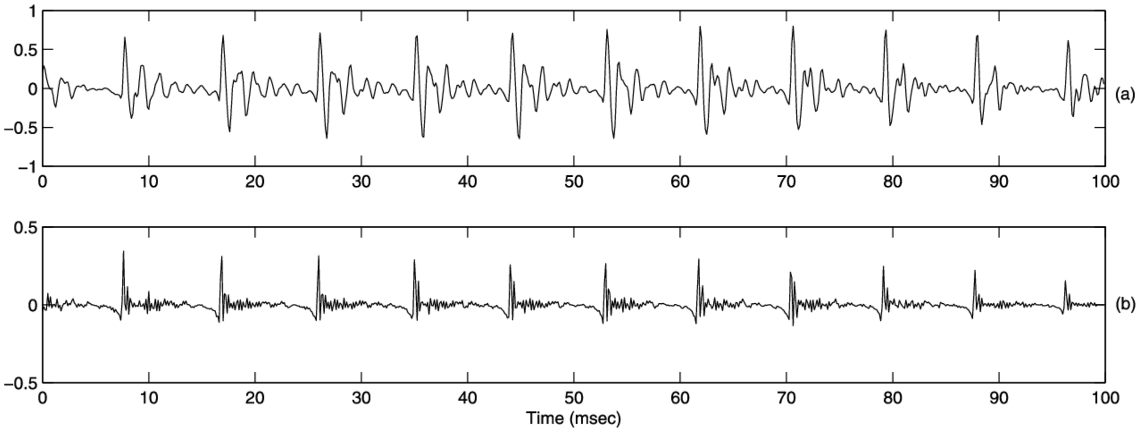


Figure 4: The speech signal (a) and its corresponding LP residual (b)

4 Features

4.1 STAT features

For capturing of the vocal tract features conventional Mel Frequency Cepstral Coefficients (MFCC), Constant Q Cepstral Coefficient (CQCC), Perceptual Linear Prediction (PLP) [6] and Linear prediction Cepstral Coefficients (LPCC) were obtaiined. The source evidences were obtained from LP-residual and ZFF methods.From these cepstral coefficients were obtained named MFCC-WR and MFCC-ZFF respectively. For all the above features we have calculated statistical averages to extract utterance level information from the speech signal. The statistical parameters are extracted from these 39 coefficients. They are

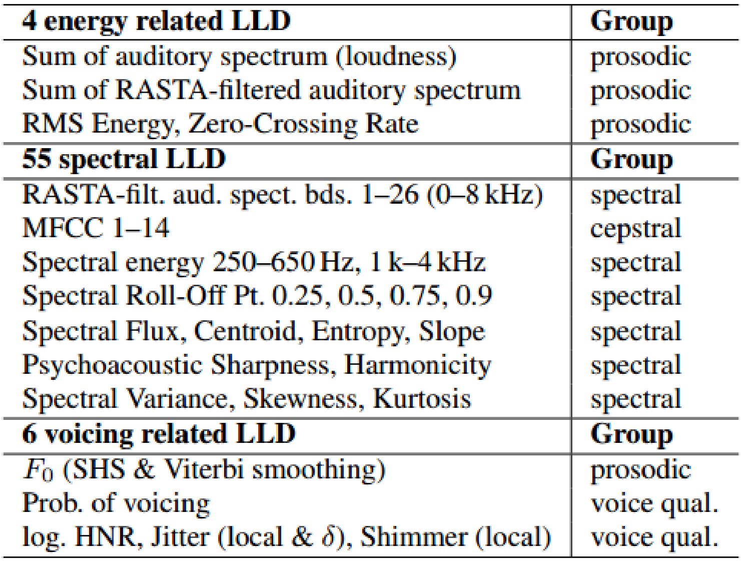
- Mean across the frames
- Standard deviation across all the frames
- Skewness
- Kurtosis

This results in a 156 dimensional feature vector and this is referred as Statistical Averaging Across Time (STAT) in the work. The averages obtained from the system features referred as MFCC-STAT, LPCC-STAT, PLP-STAT. The statistical averages obtained from ZFF method and LP analysis are referred as MFCC-ZFF-STAT and MFCC-Residual-STAT respectively.

4.2 openSMILE- ComParE Feature set

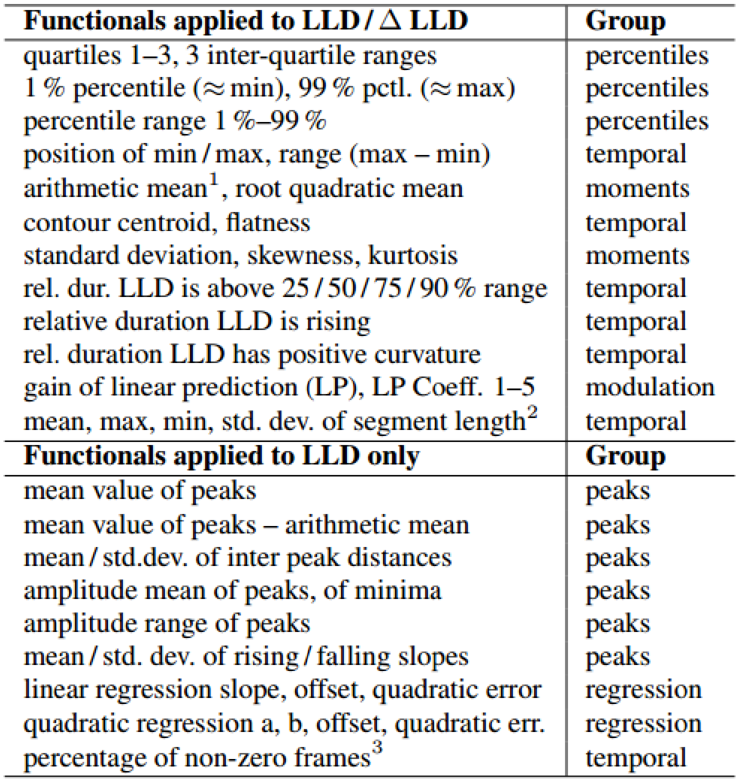
Speech contains both linguistic and non-linguistic or paralinguistic information. Paralinguistic information includes accent, pitch, loudness, speech rate, modulation, intonation, fluency etc. in speech. For voice disorder identification paralinguistic information plays very important role. The 2013 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 [3][4].

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 .



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

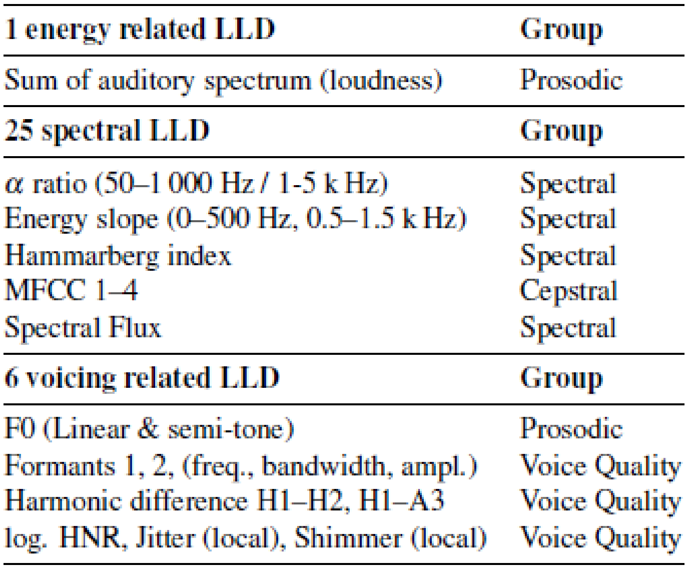


Functionals applied to LLD / Δ LLD	**Group**
quartiles 1–3, 3 inter-quartile ranges	percentiles
1 % percentile ($\approx$ min), 99 % pctl. ($\approx$ max)	percentiles
percentile range 1 %–99 %	percentiles
position of min / max, range (max – min)	temporal
arithmetic mean[1], root quadratic mean	moments
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 6: 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

4.3 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 [5]. 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 7: eGeMAPS acoustic feature set: 42 provided low-level descriptors(LLD)

4.4 Long-term Average Spectral features (LTAS) features

The long term average spectrum features capture the static information like voice quality, gender information and age-related features from the speech signal. To extract these features, first, the speech signal $s[n]$ is passed through the bank of filters (design of various filter banks will be discussed in Section 3) to decompose it into multiple time-frequency components. If $h_i[n]$ is filter's impulse response then the output of the filter is given by

$$s_i[n] = h_i[n] * s[n] \qquad i = 1, 2.....N \tag{7}$$

where N is the number of filters. All the N band signals along with original full-band signal in total $N+1$ components are framed using a non-overlapping rectangular window of 20 ms. Then root mean square energy is calculated for each frame denoted by $s_{RMSi[k]}$ correspond to the k^{th} frame of i^{th} band. Finally, 10 statistical averages like normalized mean, standard deviation, range, skewness and kurtosis are calculated, the resulting $((N+1)*10-1)$ dimension feature vector is denoted as LTAS feature.

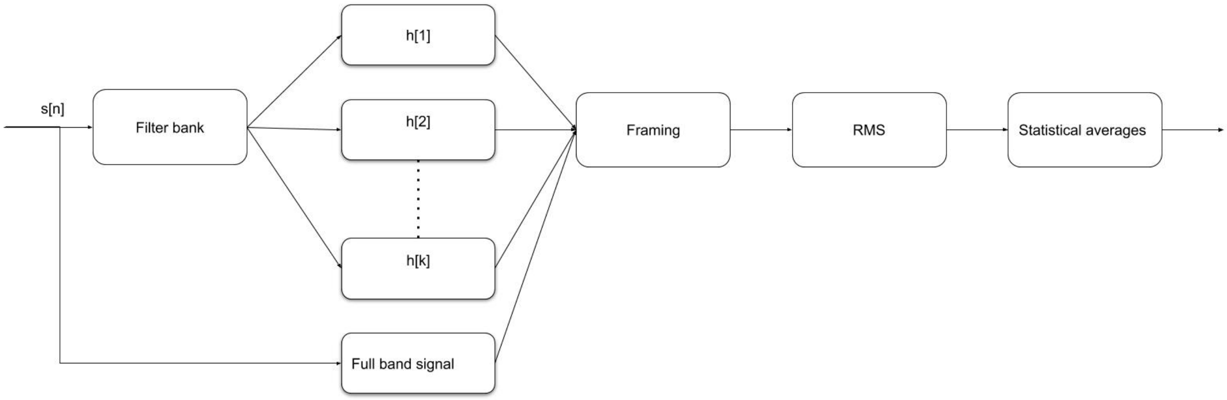


Figure 8: LTAS feature extraction

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, $\frac{mean(\{S_{RMSi}[k]\}-\{S_{RMSi}[k-1]\})}{rms\{s_0[n]\}}$

LTAS features are obtained from critical band, constant Q, gammatone and single frequency filter bank named CBFB-LTAS, CAFB-LTAS, GFB-LTAS and SFFB-LTAS respectively.

Critical band filter bank: Critical band filter bank (CBFB), also referred to as octave band filter bank, is used to mimic human perception. Octave band filters are set of band pass filters in which highest frequency is twice of the lowest frequency. Octave band is mainly used in music, in which one octave is difference between same notes with double it's frequency. The original speech segment is filtered into 9 octave bands with center frequencies of approximately 30, 60, 120, 240, 480, 960, 1920, 3840, and 7680Hz, using eight-order Butterworth filters for the calculation of LTAS features.

Constant Q filter bank: Like the Fourier transform a constant Q transform is a bank of filters, but in contrast to the former it has geometrically spaced center frequencies. Constant Q filter bank (CQFB) is geometrically spaced filter bank with constant Q factor (i.e. ratio of center frequency to the resolution is constant), such that resolution of the filters can be approximated to musical notes. The k^{th} center frequency of constant Q transform is given by

$$f_k = f_0 \; 2^{k/B} \tag{8}$$

where, f_0 is minimum frequency, and B is number of bins per octave. The bandwidth of the filter b is given by

$$b = f_k \; (2^{1/B} - 1). \tag{9}$$

Constant Q filters has high temporal resolution at high frequency and high frequency resolution at low frequency which also mimic the human auditory system [8].

Gammatone Filter bank: Gammatone filters are linear approximation of physiologically motivated processing performed by the cochlea. It is commonly used in modeling the human auditory system and consists of a series of band pass filters. In the time domain, the filter is defined by the following impulse response [10][11]:

$$g(t) = at^{(N-1)}e^{-2\pi bt}cos(2\pi f_c t + \phi) \qquad for \;\; t \geq 0. \tag{10}$$

Here, N is the order of the filter which determines the slope of the filter's skirts, b is the bandwidth of the filter, f_c is center frequency, a and ϕ are the scaling factor and phase of the cosine wave, respectively. In [12] Glasberg and Moore relate center frequency and the ERB of an auditory filter as

$$b = ERB(f_c) = 24.7(4.37f_c + 1) \tag{11}$$

where, b is in Hz and f_c is in kHz.

Single frequency filter bank : The single frequency filter bank (SFFB) [11], is based on single frequency filtering which provides good time-frequency resolution. In single frequency filter bank approach speech signal is passed through a set of complex band pass filters to decompose signal into different frequency bands. The transfer function of the k^{th} filter is given by,

$$H_k(z) = \frac{1}{1 - a_k z^{-1}} \qquad k = 0, 1, 2, ...M \tag{12}$$

where, $a_k = ae^{-jw_k}$, a represents pole location, w_k is k^{th} frequency component, f_s corresponds to sampling frequency and M is total number of filters. The value of a which can be selected in between 0 to 1, determines the bandwidth of the filter. The narrow filters are designed to provide high spectral resolution by choosing the value of 'a' between 0.95 to 0.995.

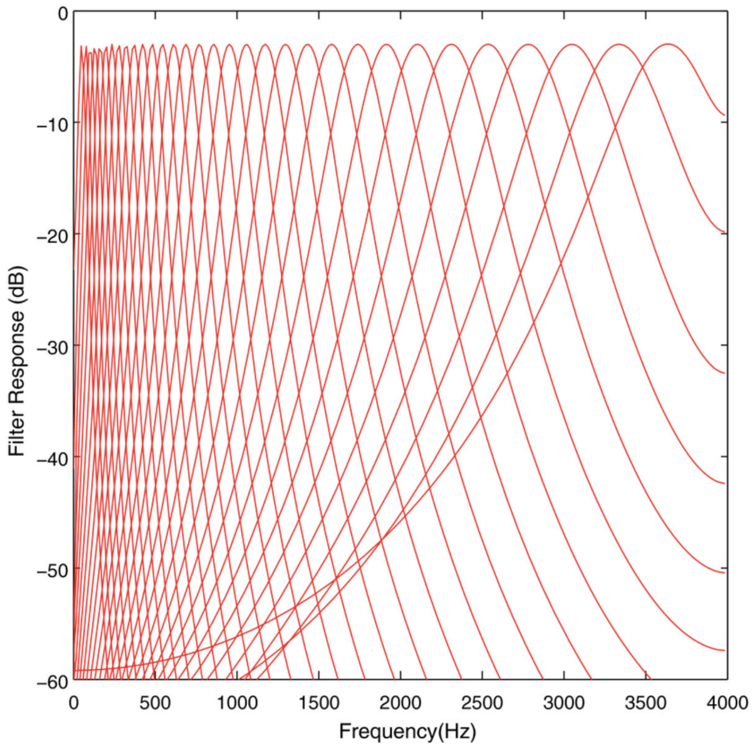


Figure 9: Frequency response of 32 gammatone filter

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. Makhoul, John. ”Spectral linear prediction: Properties and applications.” IEEE Transactions on Acoustics, Speech, and Signal Processing 23.3 (1975): 283-296.

3. Eyben, Florian, Martin Wöllmer, and Björn Schuller. ”Opensmile: the munich versatile and fast open-source audio feature extractor.” Proceedings of the 18th ACM international conference on Multimedia. 2010.

4. Eyben, Florian, et al. ”The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing.” IEEE transactions on affective computing 7.2 (2015): 190-202.

5. Schuller, Björn, et al. ”The INTERSPEECH 2015 computational paralinguistics challenge: nativeness, parkinson’s eating condition.” Sixteenth annual conference of the international speech communication association. 2015.

6. Hermansky, Hynek. ”Perceptual linear predictive (PLP) analysis of speech.” the Journal of the Acoustical Society of America 87.4 (1990): 1738-1752.

7. Mendoza, Elvira, et al. ”Differences in voice quality between men and women: use of the long-term average spectrum (LTAS).” Journal of voice 10.1 (1996): 59-66.

8. Brown, Judith C. ”Calculation of a constant Q spectral transform.” The Journal of the Acoustical Society of America 89.1 (1991): 425-434.

9. Auditory processing-based features for improving speech recognition in adverse acoustic conditions,”EURASIP J. on Audio, Speech, and Music Process.,vol. 2014, no. 1, pp. 21, 201

10. Glasberg, Brian R., and Brian CJ Moore. ”Derivation of auditory filter shapes from notched-noise data.” Hearing research 47.1-2 (1990): 103-138.

11. Gurugubelli, Krishna, and Anil Kumar Vuppala. ”Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment.” ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

12. Glasberg, Brian R., and Brian CJ Moore. ”Derivation of auditory filter shapes from notched-noise data.” Hearing research 47.1-2 (1990): 103-138.