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

1.1 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,

$$H_{SFFB}(z) = \{H_1(z),\ H_2(z),\ ....\ H_k(z)\ ....\ H_M(z)\} \tag{1}$$

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 1, 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.

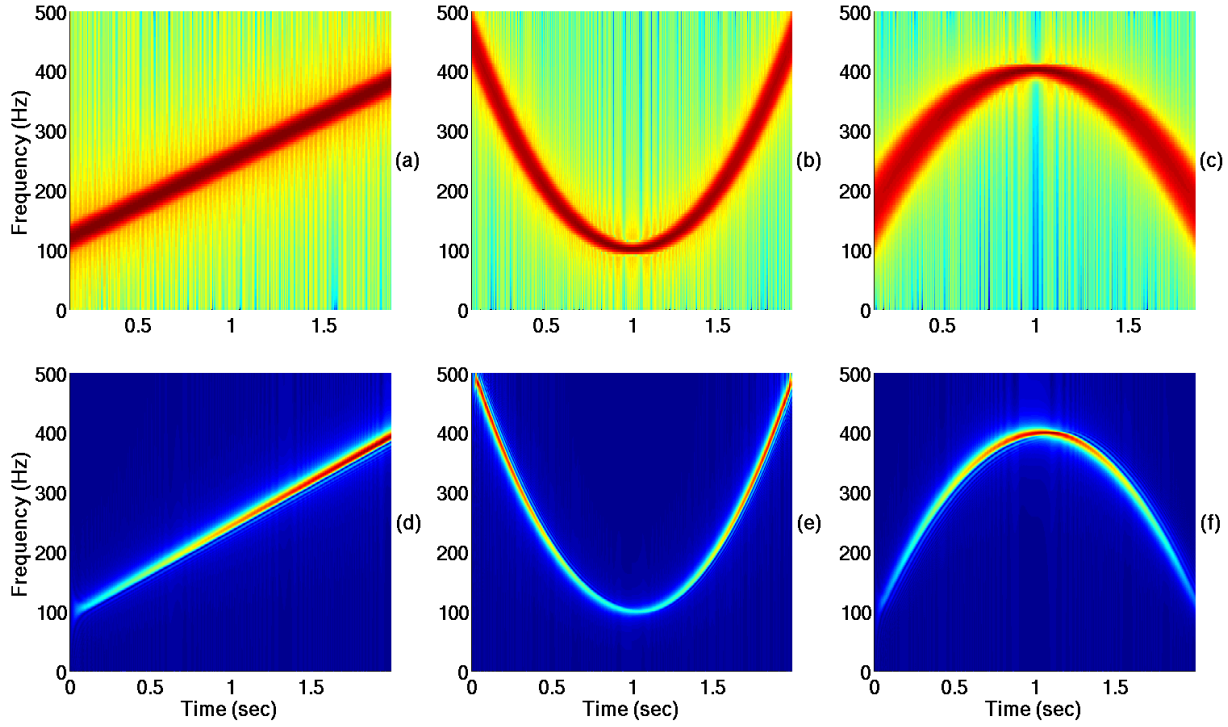

Figure 1: Time-frequency representation of synthesized linear, quadratic, and convex chirp signals. Short time Fourier transform (Top row: (a)-(c)). Single frequency filtering (Bottom row: (d)-(f)).

1.2 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.

2 Auditory filter bank based long-term average spectral features

The long-term average spectral features captures atypical average spectral information in the signal. First, an auditory filter bank has been used to decompose the speech signal into 9 octave bands with center frequencies of approximately 30, 60, 120, 240, 480, 960, 1920, 3840, and 7680 Hz. Further, 10 components (9 octave band and full-band speech signals) together used to extract the LTAS features. Finally, for each of the 10 components following spectral features are computed and these are referred to as auditory 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, $\frac{mean(\{S_{RMSi}[k]\}-\{S_{RMSi}[k-1]\})}{rms\{s_0[n]\}}$

By including full-band speech signal a total 10 components will be present. This results in $99-$dimensional AFB-LTAS feature vector.

3 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 aforementioned 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)

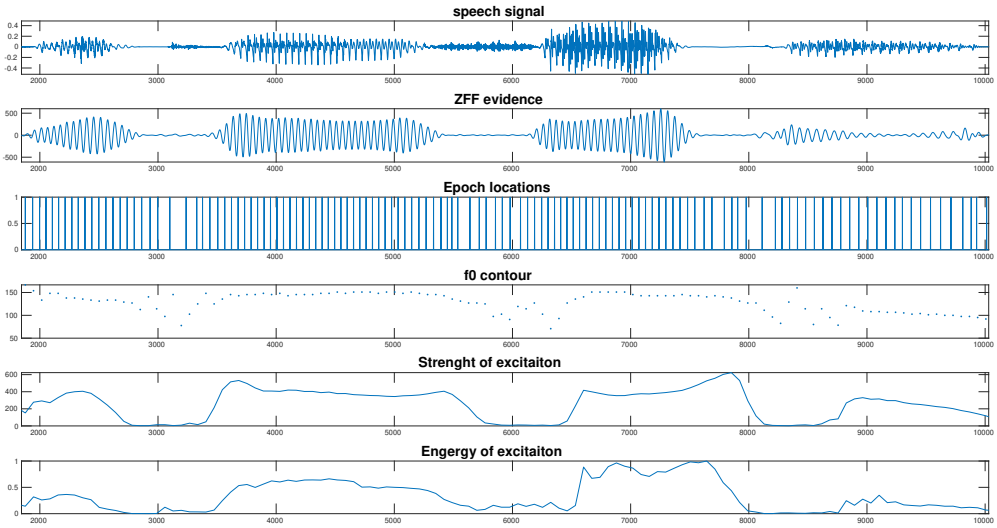

Figure 2: Evidences for extracting Intonation features. (from top to bottom) Speech signal, ZFF evidence, Epoch locations, f0 contour, Strength of excitation, Energy of excitation.

4 Durational features

The duration of the vocalic and intervocalic segments are computed and a series of features are extracted. To extract voiced segments Zero frequency filter(ZFF) is used. The evidence from ZFF is smoothed and a threshold is applied. As the energy of ZFF evidence is relatively very high in voiced regions, the voiced regions can be extracted by a simple threshold. Then the duration of vocalic and non-vocalic intervals is obtained. A 12-dimension durational feature vector has been extracted by using the vocalic and inter-vocalic segment regions.

1. ΔV Standard deviation of vocalic intervals.
2. ΔIV Standard deviation of inter vocalic-intervals.
3. ΔV-IV Standard deviation of vocalic and inter-vocalic intervals.
4. ΔV Standard deviation of vocalic intervals.
5. %V percent of utterance duration composed of vocalic intervals.
6. Varco-V standard deviation of vocalic intervals divided by mean vocalic duration ($\times$100).
7. Varco-IV standard deviation of intervocalic intervals divided by mean intervocalic duration ($\times$100).
8. Mean of the difference between successive vocalic intervals divided by their sum ($\times$100).
9. Mean of the difference between successive intervocalic intervals.

10. Mean of the difference between successive vocalic + intervocalic intervals divided by their sum ($\times 100$).

11. Mean of the difference between successive vocalic + intervocalic intervals.

12. Articulation rate Number of vocalic + intervocalic intervals produced per second excluding pauses.

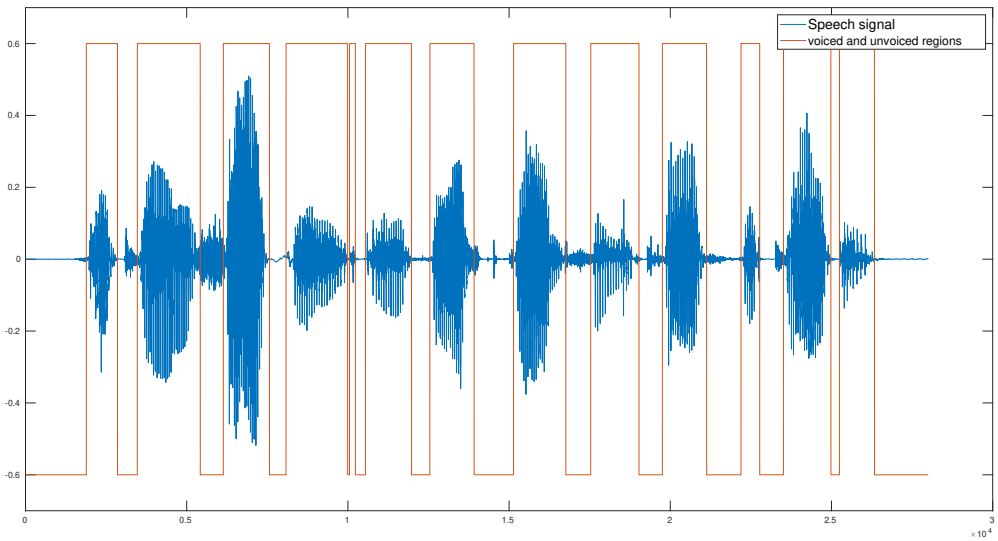

Figure 3: The voiced and unvoiced regions extracted from speech signal using ZFF .

.