MONSOON 2022

SPEECH SIGNAL PROCESSING

TEMPORAL SPEECH ENHANCEMENT FOR LANGUAGE IDENTIFICATION IN NOISY ENVIRONMENTS

P10

Shruti Kolachana 2020102053
Arshini Govindu 2020102009
Sankeerthana Venugopal 2020102008

02

INTRODUCTION

This report presents a noisy speech enhancement method that does Linear Prediction (LP) Residual Weighting in the time domain. The effectiveness of this method is then tested using a Language Identification model.

The basis for this enhancement technique is that human beings perceive speech by capturing features present from the high signal-to-noise ratio (SNR) regions and then extrapolating the features in the low SNR regions.

Accordingly, a weight function is derived for the residual signal that will reduce the energy in the low SNR regions relative to the high SNR regions of the noisy signal. The residual signal samples are multiplied with the weight function and the weighted LP residual is used to excite the time-varying all-pole filter derived from the noisy speech to generate the enhanced speech.

03

GROSS LEVEL TEMPORAL PROCESSING

The main objective of the gross level processing is to identify and enhance the speech components at the sound units (100–300 ms) level.

The high SNR regions at gross level are identified by using the sum of ten largest peaks in the DFT spectrum and smoothed HE of the LP residual of the noisy speech signal.

The peaks in the DFT spectrum predominantly represent the vocal tract information while the LP residual predominantly contains information about the excitation source.

SUM OF DFT PEAKS

The DFT magnitude spectrum of a voiced frame has the same harmonic structure as the
excitation source spectrum, but the amplitudes of the harmonics have been shaped according to the frequency response of the vocal tract. It has peaks at pitch and harmonic locations and also stronger peaks at formant locations.
Hence, the sum of amplitudes of the major peak locations will be higher in high SNR regions than low SNR regions.

STEPS:

1. **Framing:** Divide the signal into 20 ms frames with 10 ms overlap
2. **Windowing:** Multiply each frame with a Hamming window of the same length
3. **DFT:** Run 512-point DFT on each frame

$$Y(k,l) = \sum_{n=0}^{N-1} y(n)w(n-lR)e^{-\frac{j2\pi nk}{N}}$$

4. **Sum peak amplitudes:** Sum amplitudes of 10 major peak locations frame-wise

$$s_d(l) = \sum_{m=1}^{10} |Y(k_m,l)|$$

5. **Repeat:** Repeat the computed value for each frame 80 times to make the indicator length equal to that of the speech signal

04

SMOOTHED HILBERT ENVELOPE OF LP RESIDUAL

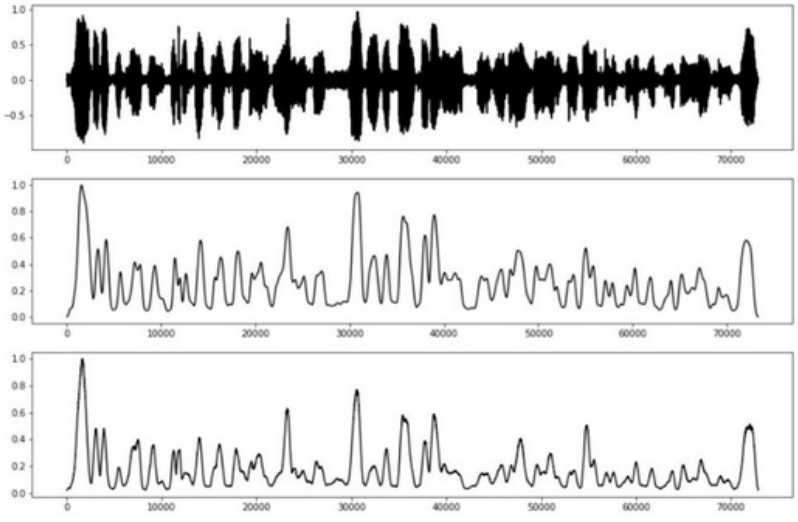
The instants of significant excitation for speech production correspond to instants of glottal closure (GC) or epochs during voiced speech and onset of events like burst and frication during unvoiced speech. The energy associated with these instants will be locally high. These are manifested as large errors in the LP residual of the speech.
However, locating these instants directly is difficult due to the bipolar nature of the LP residual. This limitation is overcome by computing the Hilbert envelope of the LP residual.
The Hilbert envelope is a unipolar function and shows large amplitudes around the instants where the residual error is large, that is, in high SNR regions compared to the low SNR regions.

STEPS:

1. **LP Residual:** Subtract initial signal from predicted signal
2. **Hilbert Envelope:** Apply Hilbert Transform on LP residual and take the Hilbert envelope

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)}$$

3. **Mean Smoothing:** Smoothen the envelope by using a mean smoothing filter of 50ms duration



SIGNAL

SUM OF DFT PEAKS

HILBERT ENVELOPE OF LP RESIDUAL

05

GROSS LEVEL WEIGHT FUNCTION

The two parameters described above are due to different aspects of speech production. They exploit different information to provide evidence for the presence of high SNR regions. Therefore these may be effectively combined in order to obtain gross weight function, which is robust and also identifies the high SNR regions better compared to the individual parameters.

For both the extracted features:

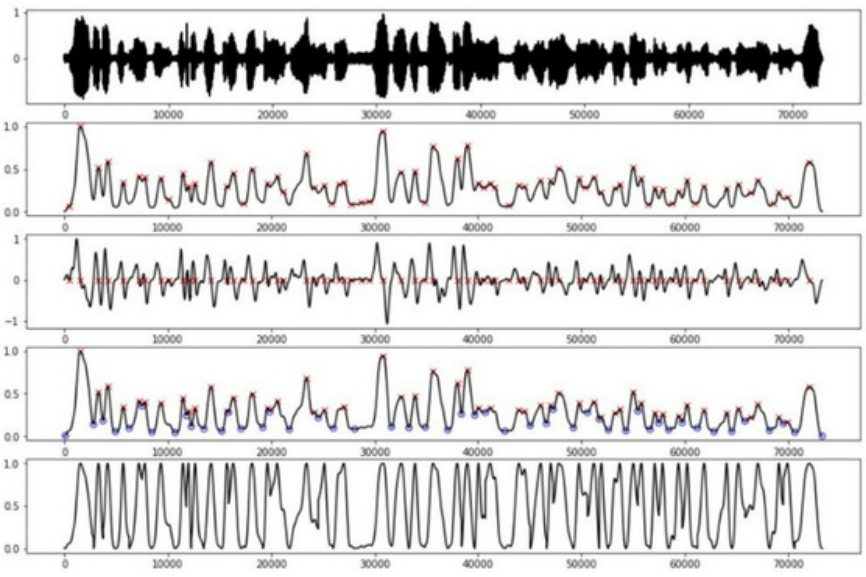
1. **Find Peaks:** Find locations of peaks
2. **FOD:** Calculate First Order Difference values

$$f[n] = y[n] - y[n-1]$$

3. **Filter Peaks:**
   - Calculate absolute FOD values for a duration of 5ms on either side with reference to each peak location. If less than the threshold, filter peak.

$$Threshold = 0.5 * mean(f)$$

   - If two successive peaks occur within 50 ms, the peak with a lower FOD value is eliminated.
4. **Find nearest Zero transitions:** Identify the nearest negative to positive going zero transition points on either side for each peak
5. **Normalize:** Normalize the regions between zero transitions



PEAKS

FIRST ORDER DIFFERENCE

FILTERED PEAKS

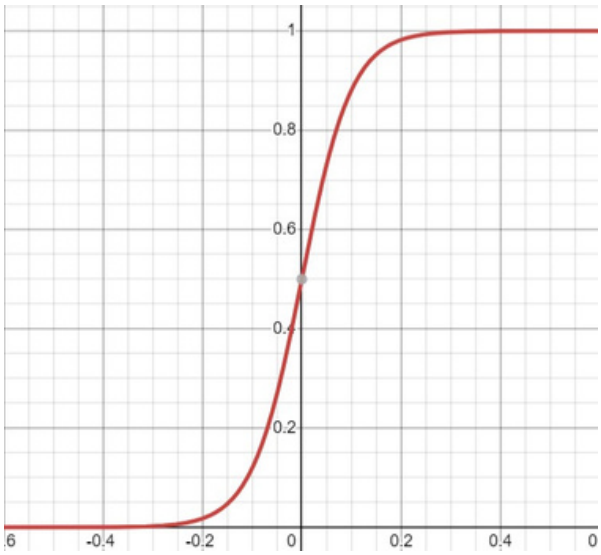
NORMALIZED SIGNAL

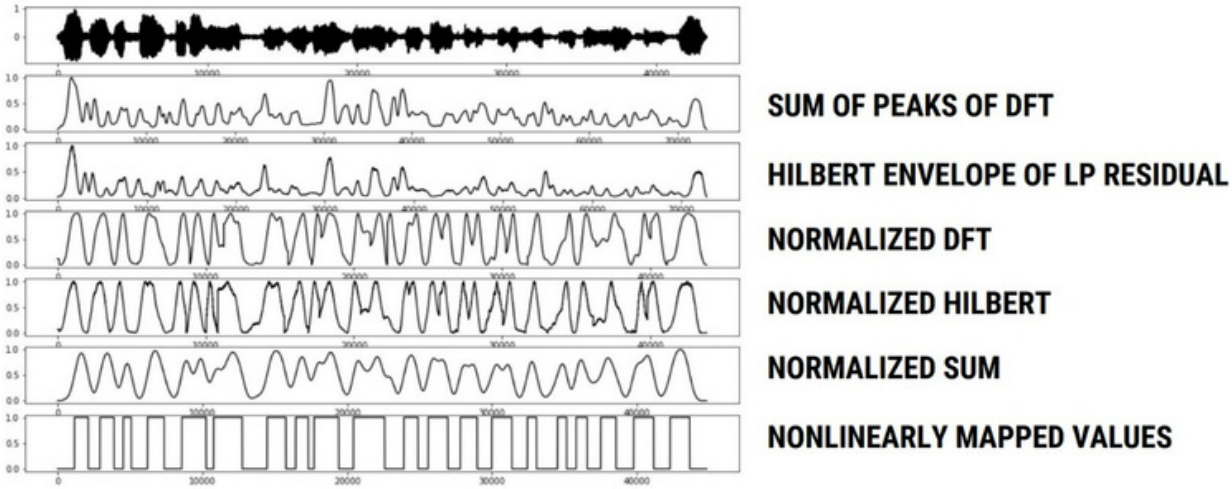
06

STEPS:

1. **Sum:** Add both the signals
2. **Normalize:** Normalize the sum
3. **Non-Linear Mapping:** Pass the normalized sum through a sigmoid function

$$w_g(n) = \frac{1}{1 + e^{-\lambda(S_i(n)) - T)}}$$





07

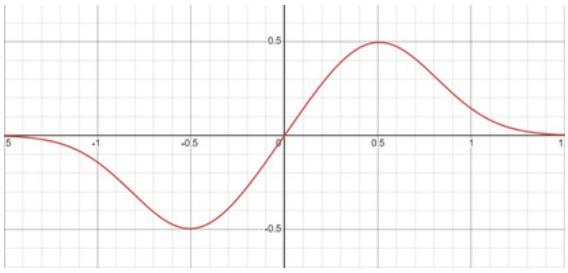
FINE LEVEL TEMPORAL PROCESSING

The objective of the fine level processing is to identify and enhance the speech-specific features at the subsegmental (2–3 ms) level.
The basis for the fine level temporal enhancement is that voiced speech is produced as a result of excitation of quasi-periodic glottal pulses and unvoiced speech is produced as a result of excitation of onset of events like burst, frication and aspiration. The significant excitation in each glottal cycle takes place at the instant of glottal closure.
The relative spacing between the GC events is not affected by degradations. Therefore by locating the instants of significant excitation, it is possible to enhance speech around the instants relative to other regions.

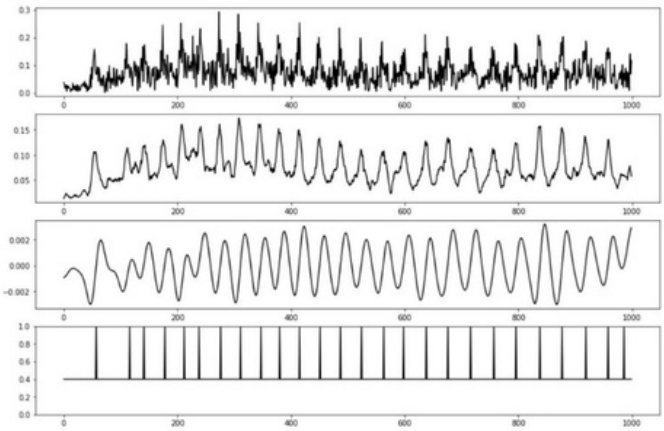
STEPS:

1. **LP Residual:** Subtract initial signal from predicted signal
2. **Hilbert Envelope:** Apply Hilbert Transform on LP residual and take the Hilbert envelope
3. **Mean Smoothing:** Smoothen the envelope by using a mean smoothing filter of 1 ms duration
4. **Negative FOGD:** Convolve the envelope with a negative First Order Gaussian DIfferentiator

$$g_d(n) = \frac{1}{\sigma\sqrt{2\pi}}\left[e^{-\frac{(n+1)^2}{2\sigma^2}} - e^{-\frac{(n)^2}{2\sigma^2}}\right] \qquad 1 \le n \le L_g$$



5. **Zero-Crossings:** Identify ZCs of the convolved signal
6. **Scale and Shifting:** Scale the values to 0.6 and shift by 0.4 to reduce perceptual distortion



HILBERT ENVELOPE OF LP RESIDUAL

SMOOTHED HILBERT

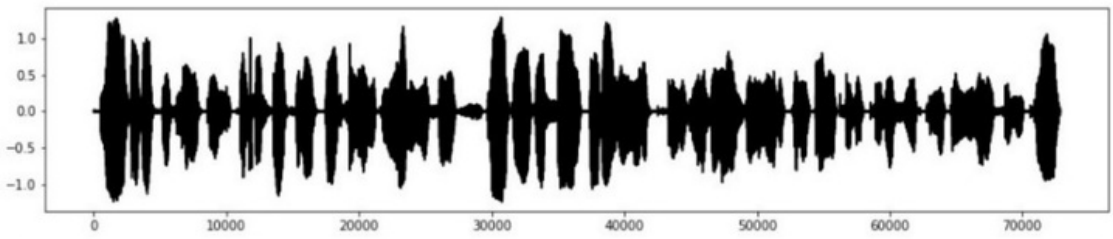
CONVOLVED WITH NEG FOGD

ZCS WITH NEG-POS TRANSITION

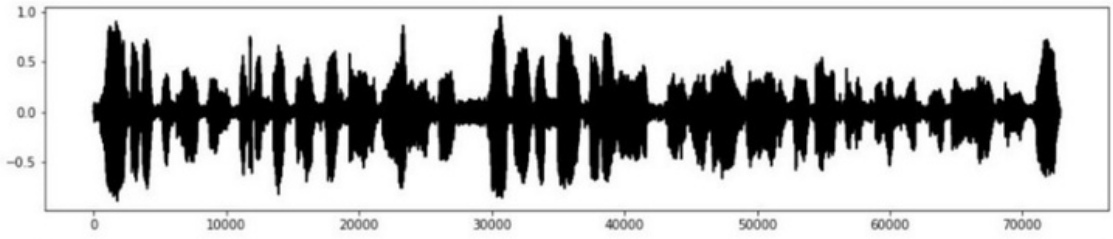
08

TEMPORALLY PROCESSED SPEECH

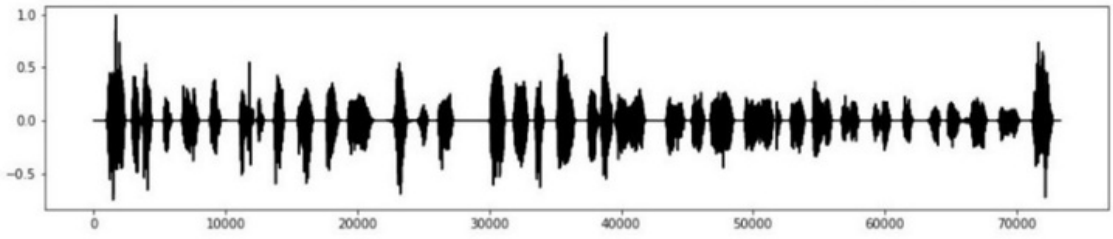
The Final weight function is obtained by multiplying the Gross weight function and the Fine weight function. The noisy speech residual signal samples are then multiplied with the final weight function.
The residual samples are weighted rather than the speech samples mainly because the residual samples are relatively less correlated and hence weighting may lead to less perceptual distortion. LP synthesis is performed using the LP coefficients of the noisy signal on the weighted residual to generate the enhanced speech which is termed as temporally processed speech.



CLEAN SIGNAL



NOISY SIGNAL



OUTPUT SIGNAL

11

LANGUAGE IDENTIFICATION MODEL

FEATURES:

1. **MFCC:** 13 features
2. **GMM:** 32
3. **Training:** 6 languages × 80 files
4. **Testing (Clean Speech):** 6 languages × 20 files
5. **Testing (Noisy Speech):** 6 languages × 4 files
6. **Testing (Temporally Processed Speech):** 6 languages × 4 files

RESULTS:

1. **Accuracy:** 85%
2. **Accuracy for Noisy Speech:** 41.67%
3. **Accuracy for Temporally Processed Speech:** 29.16%
4. **Increase in Input Language Score:**
   - **Babble noise:** in 66.7% of cases
   - **Factory noise:** in 66.7% of cases
   - **Pink noise:** in 33.3% of cases
   - **White noise:** in 50% of cases

10

CONFUSION MATRIX FOR TEMPORAL PROCESSED SPEECH

	Assamese	Bengali	Gujarathi	Manipuri	Marathi	Odia
Assamese	0	3	0	0	0	1
Bengali	0	4	0	0	0	0
Gujarathi	1	3	0	0	0	0
Manipuri	0	4	0	0	0	0
Marathi	0	0	2	0	0	2
Odia	0	2	0	0	0	2

09

CONFUSION MATRIX FOR NOISY SPEECH

	Assamese	Bengali	Gujarathi	Manipuri	Marathi	Odia
Assamese	0	0	1	0	0	3
Bengali	0	1	3	0	0	0
Gujarathi	0	0	3	0	0	1
Manipuri	0	0	2	2	0	0
Marathi	0	0	2	0	0	2
Odia	0	0	0	0	0	4

REFERENCES

- Ananthapadmanabha, T., Yegnanarayana, B., 1979. Epoch extraction from linear prediction residual for identification of closed glottis interval
- Krishnamoorthy, P., Prasanna, S.R.M., 2008. Temporal and spectral processing of degraded speech
- P. Krishnamoorthy, S.R.M. Prasanna, 2011. Enhancement of noisy speech by temporal and spectral processing
- Stack Overflow - Where Developers Learn, Share, & Build Careers
- https://scikit-learn.org/
- https://librosa.org/
- Wikipedia