Foreground Speech Segmentation using Zero Frequency Filtered Signal

Deepak K. T., Biswajit Dev Sarma and S. R. Mahadeva Prasanna

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati
Guwahati 781039, India
{deepakkt, s.biswaikut, prasanna}@iitg.ernet.in

Abstract
A method for the robust segmentation of foreground speech in the presence of background degradation using zero frequency filtered signal (ZFFS) is proposed. The speech signal from the desired speaker collected over a mobile phone is termed as foreground speech and the acoustic background picked by the same sensor that includes both speech and non-speech sources is termed as background degradation. The zero frequency filtering (ZFF) of speech allows only information around the zero frequency to pass through. The features from the resulting ZFFS, namely, the normalized first order autocorrelation coefficient and the strength of excitation of ZFFS are observed to be different for foreground speech and background degradation. A method for foreground speech segmentation is developed using these two features. The evaluation using utterances containing isolated words of foreground speech and background degradation collected in a real environment shows a robust foreground speech segmentation.

Index Terms: Foreground speech, background degradation, ZFFS, segmentation.

1. Introduction
In the current era of wireless speech communication, there is no restriction on the environment from which the speakers access the speech processing systems like speech recognition and speaker recognition. Due to this, the received signal contains both the desired speaker’s speech and the acoustic background picked by the sensor. In this work, the desired speaker’s speech is termed as foreground speech and the rest as background degradation. In many cases, the background degradation is of comparable level to that of the foreground speech and includes both speech and non-speech like components. Robust segmentation of foreground speech is crucial for further processing. The end-point or speech / non-speech detection used for selecting speech regions based on energy, zero crossing rate (ZCR) and periodicity of speech will be ineffective due to high similarity between the foreground speech and background. One approach is to preprocess the received signal through the speech enhancement methods to minimize the effects of degradation and then extract the foreground speech [1, 2, 3, 4]. The speech enhancement methods have to deal with the frequency components of speech and degradation present at all the frequencies to minimize the effect of degradation. Alternatively, if the nature of speech and degradation can be observed at a particular frequency, then methods can be developed by analyzing only the characteristics of the particular frequency component. The zero frequency filtering (ZFF) is proposed earlier for the detection of epochs in speech [5]. The ZFF allows only the signal components around the zero frequency and significantly attenuates all other frequency components. The output of ZFF is termed as the zero frequency filtered signal (ZFFS) whose characteristics can therefore be analyzed for foreground speech and degradation cases.

The present work focuses on a practical scenario of isolated word recognizer (IWR) present on a voice server and accessible remotely over mobile phones. For instance, enquiring the prices of agricultural commodities by calling to a voice server over phone and saying the required commodity name in isolated word fashion. In such applications, while the users give speech data from different places, there may be different background degradations like vehicle noise and speakers in the background. Since the nature of degradation is like speech, it becomes very difficult to process such signals for segmentation of foreground speech. However, there will be significant difference in the signal characteristics for the foreground speech and the background degradation [6]. This can be exploited for segmentation, but dealing with all the frequency components still makes the approach difficult. Alternatively, processing by ZFF will give the essence of the characteristics of foreground speech and the background degradation in the ZFFS.

The ZFFS signal shows nearly periodic behavior in the voiced speech regions and absence of such behavior in the unvoiced, silence and degraded regions [7]. Further, it is observed that there will be distortion in the nearly periodic behavior in the voiced speech regions of background speakers. The autocorrelation analysis of ZFFS shows significant change in its characteristics of the voiced speech of foreground speech and the background degradation. The value of the first largest peak (excluding the center) normalized with respect to the center value in the autocorrelation sequence defined as normalized first order autocorrelation coefficient in [8] is found to be large in the voiced speech regions of foreground speech compared to the background degradation. The strength of excitation of ZFFS is also significantly large in the foreground speech relative to other background degradation. Thus the features extracted from the ZFFS show discrimination between the foreground speech and the background degradation regions. A method can therefore be developed using these two features for foreground speech segmentation.

The significance of proposed foreground speech segmentation method is demonstrated in the vowel-like region onset point (VLROP) detection task. The existing VLROP detection method works well, only when speech of the desired speaker is present, even under constant background degradation of any level. However, these methods show significant degradation in performance due to spurious detection when the background degradation is non-stationary and in particular, speech-like. In such cases the proposed foreground speech segmentation can
be used as a preprocessing step to identify only the foreground speech regions and then use for VLROP detection. As a result, significant reduction in the spurious VLROPs detection can be achieved.

The rest of the paper is organized as follows: Section 2 describes the analysis of foreground speech and background degradation regions using ZFFS. The proposed foreground speech segmentation method using ZFFS is described in Section 3. The modified VLROP detection method using proposed foreground speech segmentation is described in Section 4. The experimental results and performance evaluation of both foreground speech segmentation and VLROP detection are given in Section 5. The summary and conclusions of the present work, and the scope for future work are mentioned in Section 6.

2. Analysis of Foreground Speech and Degradation Regions using ZFFS

Let \( x(n) \) be the given signal that contains both foreground speech and background degradation. The first order difference of this signal is given by

\[
d(n) = x(n) - x(n - 1)
\]  
(1)

The difference signal is passed twice through an ideal resonator at zero frequency given by [5].

\[
z_1(n) = -\sum_{k=1}^{\alpha_1} a_k z_1(n - k) + d(n)
\]  
(2)

and

\[
z_2(n) = -\sum_{k=1}^{\alpha_2} a_k z_2(n - k) + z_1(n)
\]  
(3)

where \( \alpha_1 = -2 \) and \( \alpha_2 = 1 \). This is equivalent to successive integration four times, termed more commonly as filtering at zero frequency. The trend in \( z_2(n) \) is removed by subtracting the average over a pitch period at each sample. The resulting signal

\[
z(n) = z_2(n) - \frac{1}{2N + 1} \sum_{m=-N}^{N} z_2(n + m)
\]  
(4)

is called the zero frequency filtered signal (ZFFS) \( z(n) \). Here \( 2N + 1 \) corresponds to the number of samples in one pitch period.

Figure 1(a) shows a signal that contains foreground speech (Region A) and the background degradations, like background speaker’s speech (Region B) and vehicle noise (Region C) collected over a mobile phone. The nature of degradation components match closely with the foreground speech. Figure 1(b) shows the ZFFS of the signal. The nature of ZFSS is different in the foreground speech and background degradation regions in terms of significant change in amplitude. Alternatively, the energy and ZCR computed on the signal shown in Figures 1(c) and 1(d), respectively show relatively less discrimination between the foreground and background regions.

Figures 2(a), 2(b) and 2(c) plots 50 ms segments taken from the foreground speech, background speech and vehicle noise, respectively taken from Figure 1. The nature of signal characteristics are different for the three cases. Even though 2(b) is also speech segment, it is more dispersed due to the background effect, that is, distance from the sensor. Figures 2(d), 2(e) and 2(f) plot corresponding ZFFS. The ZFFS of foreground speech

![Figure 1](image1.png)

![Figure 2](image2.png)
malized first order correlation coefficient can be used as a parameter for discrimination between foreground speech and the background degradation. The observation of 50 ms segments of ZFFS indicate that the strength of excitation defined as the slope of the ZFFS at the positive zero crossings [7] as given in Figures 2(j), (k) and (l) are also significantly different among the three cases. The strength of excitation is significantly larger in the foreground speech regions relative to that of the other background cases. Thus these two characteristics of the ZFFS can be combined for discriminating the foreground speech and background degradation regions.

3. Foreground Speech Segmentation using ZFFS

The given signal containing both foreground speech and the background degradation is processed by the ZFF method to extract the ZFFS as discussed in the previous section. The ZFFS is processed in blocks of 50 msec with the shift of one sample to extract the normalized first order correlation coefficient and strength of excitation features. The features show high values in the foreground speech regions compared to the rest as plotted in Figure 3. A linear combination of these two features results in the combined evidence that shows significant discrimination between the foreground speech and the other regions as in Figure 3(d). The discriminative evidence is further enhanced by taking the exponential of the combined evidence and is shown in Figure 3(e).

The maximum evidence region present in the combined evidence ($E(n)$) is further subjected to the endpoint detection algorithm to perform segmentation of foreground speech [9]. The endpoint detection algorithm works on the basis of lower ($T_l$) and upper threshold ($T_u$) values of the evidence. The $T_l$ and $T_u$ are computed using the maximum value of the evidence ($E_{\text{max}}$), and the minimum value of the evidence ($E_{\text{min}}$). The following relations are used to obtain the $T_l$ and $T_u$:

$$I_1 = 0.03 \times (E_{\text{max}} - E_{\text{min}}) + E_{\text{min}} \tag{5}$$

$$I_2 = 4 \times E_{\text{min}} \tag{6}$$

$$T_l = \text{min}(I_1, I_2) \tag{7}$$

$$T_u = 2 \times T_l \tag{8}$$

Let $L_{\text{max}}$ (in ms) is the location of $E_{\text{max}}$ in $E(n)$. The starting points for the search on either side from $L_{\text{max}}$ is given by $L_{\text{max}} \pm 1000$. These points are initially labeled as the begin or end points of the foreground speech, unless the energy falls below $T_l$ before it rises above $T_u$. If this happens a new begin point is obtained and the process is repeated. The endpoints obtained are used to segment the foreground speech from the rest of the background as shown in Figure 3(a).

4. VLROP Detection using Foreground Speech Segmentation

The VLROP is defined as the instant at which the onset of vowel-like region takes place [10]. The vowel-like regions include vowels, diphthongs, semivowels and nasals. The existing method for VLROP detection mainly uses the excitation source information for detection [10]. This method works fine as long as there is no speech-like background degradation. Alternatively, if the speech is affected by speech-like background degradation, then it produces spurious VLROPs. The proposed foreground speech segmentation can therefore be used as a pre-processing step to the existing VLROP detection method. The incoming speech signal is first passed through the proposed foreground speech segmentation method to mark the boundaries of foreground speech. The detected foreground speech regions are only passed through the VLROP detection method. As a result the spurious VLROPs detected earlier in the background regions are eliminated. Figures 4(a) and (b) show the VLROPs detected prior and after applying the proposed foreground speech segmentation, respectively, where in the latter case, the spurious VLROPs are eliminated.

5. Experimental Results and Discussion

The speech data for this work is collected for an IWR task running on a voice server. The speakers called the voice server using their mobile phone from a remote place having varied acoustic background and tested the IWR system. The error analysis of the IWR system showed that around 60 files had high background consisting of either speech or non-speech like degradation like bird chirping, car reverse horn, car horn, coughing, breathy noise etc. The 60 files were manually labeled for both foreground speech segmentation and VLROPs.
using the wavemover. The listening and visual modes for time
domain waveform and spectrogram were employed to manually
mark the segments. These labels are used as the reference for
evaluation. The average length of files considered for evalua-
tion is 16.35 seconds. Broadly the files are categorized into
two sets based on speech or non speech like background noise
present. There are 19 speech files having speech like and 41
files having non speech like background noise present predo-
nomarily in the test set. The Correct Detection Rate (CDR), Miss
Detection Rate (MDR) and Spurious Detection Rate (SDR) are
used as parameters to evaluate the performance of the proposed
method. The CDR, MDR and SDR are defined as follows:

$$\text{CDR} = \frac{N_{CF}}{N_{FG}}, \quad \text{MDR} = \frac{N_{MF}}{N_{FG}}, \quad \text{and} \quad \text{SDR} = \frac{N_{SF}}{N_{FG}} \quad (9)$$

where $N_{CF}$ is number of correctly detected, $N_{MF}$ is number
of missed, $N_{SF}$ is number of spurious and $N_{FG}$ is total
number of foreground regions. The $N_{CF}$ has the average
deviation of 15.8 ms and 20.64 ms for the begin and end sample,
respectively, with reference to manually marked foreground re-

Table 1 represents the performance of the foreground speech
segmentation as compared to the manually marked seg-
mentation. The failure cases were mainly in speech-like noise
cases. The analysis of these cases revealed that the correspond-
ing acoustic sources were very close to the sensor. The VL-
ROPs are spuriously detected at different regions of the speech
signal due to the presence of background degradation. The spu-
rious VLROPs that are detected outside the foreground speech
segment are eliminated using proposed segmentation method.
Since the objective is to demonstrate the significance of pro-
posed foreground speech segmentation method, the following
evaluation is done. The VLROP method is evaluated using three
cases, namely, missing VLROPs in the foreground speech re-
gions, spurious VLROPs in the foreground speech regions and
spurious VLROPs in the background degradation regions. Table
2 shows the performance evaluation of the proposed algorithm
evaluated on 60 files mentioned above. The table represents
the actual VLROPs missed and the spurious VLROPs detected
due to wrong segmentation of the foreground speech. Back-
ground speech and other non-speech like degradation present
in the speech signal contribute to the detection of large num-
ber of spurious VLROPs which are not part of the foreground
speech. The proposed foreground segmentation eliminates all
these spurious ones.

6. Summary and Conclusions

This work proposed a foreground speech segmentation method
using ZFSS. The ZFSS features, namely, the normalized first
order autocorrelation coefficient and strength of excitation
showed significant discrimination among foreground speech
and background degradation regions. A method is developed
using two features for foreground speech segmentation. The
significance of proposed foreground speech segmentation is
demonstrated in VLROP detection task. The large number of
spurious VLROPs in the background degradation regions are
significantly reduced due to foreground speech segmentation.

The proposed method performance degrades, if the back-
ground speaker is close to the sensor. Future work should look
along this direction. Some more features from the ZFSS can be
explored for foreground speech segmentation task.

7. Acknowledgements

The present work is part of the ongoing DIT project on "Devel-
opment of prosodically guided phonetic engine in Indian lan-
guages" (2011-2012).

8. References

enhancement by temporal and spectral processing,” IEEE Trans.
Audio, Speech, and Language Processing, vol. 17, no. 2, pp. 253–
266, February 2009.

vol. 27, no. 2, pp. 113–120, April 1979.

separation by lp residual weighting and harmonics enhancement,”
International Journal of Speech Technology (Springer), vol. 13,

[4] Y. Ming and O’Shaughnessy, “Speech enhancement based con-
ceptually on auditory evidence,” IEEE Trans. Signal Processing,

speech signals,” IEEE Trans. on Audio,Speech, And Language

IEEE Trans. on speech and audio processing, vol. 2, no. 4, pp.

terization of glottal activity from speech signals,” IEEE Signal


the endpoints of isolated utterances,” Bell system Tech. Journal,

regions for speaker verification under degraded condition,” IEEE
Trans. Audio, Speech, and Language Processing, vol. 19, no. 8,

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Without foreground segmentation</td>
<td>147</td>
<td>171</td>
<td>29</td>
<td>742</td>
<td>527%</td>
</tr>
<tr>
<td>With foreground segmentation</td>
<td>147</td>
<td>171</td>
<td>29</td>
<td>9</td>
<td>29%</td>
</tr>
</tbody>
</table>