Significance of variable height-bandwidth group delay filters in the spectral reconstruction of speech

Devanshu Arya, Anant Raj, and Rajesh M Hegde

Indian Institute of Technology Kanpur
devanshu,anantraj,rhegde@iitk.ac.in

Abstract

The significance of varying the height and bandwidth of the group delay spectrum is hitherto unexplored in the spectral reconstruction of speech signals. In this paper, a family of variable height-bandwidth filters are designed to evaluate their performance in the reconstruction of speech. The design procedure for higher order group delay filters as a cascade of second order filters is first described. These higher order filters enable the synthesis of speech sounds by simultaneously varying the height and bandwidth of the group delay spectrum. The group delay filter response is corrected by removing zeros in close proximity to the unit circle which give rise to abrupt phase transitions at points of significant excitation. Experiments on spectral reconstruction and perception of speech using variable height bandwidth group delay filters are conducted to compute the optimal height and bandwidth of the group delay filter. The experimental results indicate that the optimal height bandwidth obtained from a family of variable height bandwidth group delay filters does indeed improve the spectral reconstruction and perception of speech sounds when compared to fixed height bandwidth group delay filters.

1. Introduction

Group delay filters are generally used to realize a given phase response in linear systems. They are also used in phase equalization for non linear phase systems. The Group delay function [1], is defined as the negative derivative of the Fourier transform phase as

$$\tau(\omega) = -\frac{d(\theta(\omega))}{d\omega} \quad (1)$$

where the phase spectrum $\theta(\omega)$ of a signal is a continuous function of $\omega$. However, spectral phase is discontinuous and needs to be unwrapped to obtain a continuous function. But the group delay function can be computed directly from the speech signal [1] without unwrapping the short-time phase spectrum as

$$\tau_x(\omega) = \frac{X_R(\omega)Y_I(\omega) + X_I(\omega)Y_R(\omega)}{|X(\omega)|^2} \quad (2)$$

where the subscripts $R$ and $I$, denote the real and imaginary parts of the Fourier transform. $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x(n)$ and $nx(n)$, respectively. The group delay spectrum obtained herein is considered complementary to the magnitude spectrum due to the difference in both the height and bandwidth of the formants in speech sounds. But the significance of varying the height and bandwidth of the group delay spectrum is hitherto unexplored in the processing of speech and audio signals. In this paper, significance of variable height-bandwidth group delay filters in the spectral reconstruction of speech is explored. The rest of the paper is organized as follows. A brief discussion on recursive method of higher order group delay filter design in the frequency domain is described in Section 2. In Section 3 and Section 4, a method to reconstruct vowel and speech signals is described. The performance of the optimal height bandwidth group delay filter is then evaluated by conducting experiments on perceptual evaluation of reconstructed vowels from Hillenbrand vowel database [2] in Section 5 to illustrate the significance of the method. Section 6 presents brief conclusion.

2. Design of Variable Height-Bandwidth Group delay filter

In general group delay filters are used to realize a given phase response in a linear phase system and in the phase equalization of a non linear phase system. The recursive design procedure followed in this work uses an equal ripple approximation of the group delay filter [3]. The design of higher order variable height bandwidth group delay filters is described in the subsequent section.

2.1. Design of a basic variable height-bandwidth group delay filter [4]

Group delay filters are generally designed as all pass filters to preserve the magnitude response of the original signal while providing linear phase in the processing of speech sounds. A typical representation of a group delay filter [5] of order $2N$ is

$$H(z) = \sum_{i=1}^{2N} A_i z^{-i} + \sum_{i=1}^{2N} A_i z^{i-1} \quad (3)$$

where, $z = \exp(-j\omega)$, $\omega$ is the normalized frequency in radians per second, and $A_i$ are the filter co-efficients. The basic building block of a single resonance (peak) group delay filter is given in [4]. For an $n^{th}$ order group delay filter (all pass) design, the group delay function must satisfy the following equation [6]

$$\int_0^\pi \tau_n(\omega) d\omega = n\pi \quad (4)$$

As the group delay is the negative derivative of the phase response with respect to frequency, its integral around the unit circle is simply the negative of the phase accumulated during one traversal of the unit circle which is equal to $2\pi$ per pole, regardless of the position of pole inside the unit circle. Note that a pole corresponds to the resonance (peak) in the group delay spectrum. A single resonance (peak) group delay filter can be designed for a fixed height and bandwidth only by using a basic second order filter as can be seen from Equation 4. The
height and bandwidth of this filter cannot be controlled simultaneously. This is because the integral of the group delay of the second order all pass filter around the unit circle is always $4\pi$. But the height-bandwidth product can be controlled by varying the height only in the case of a second order group delay filter which necessitates the design of higher order group delay filters. Higher order group delay filters are realized as a cascade of second order group delay filters to utilize the additive property of the group delay function [7], which is described in the context of this work in the next Section.

2.2. Design of higher order variable height-bandwidth group delay filter

The desired group delay filter with simultaneous control over height and bandwidth can be realized by cascading $N$ number of second order filters [4] with different pole location radii ($r_k$) and phase angle ($\theta_k$) as illustrated in Figure 1 (top). Figure 1 (bottom) illustrate three second order group delay filter responses with three different radii ($r_k$) and phase angle ($\theta_k$). Figure 1 (top) shows the response of thirty second order filters cascaded together. As order of the filter $N$ is increased, the area under the curve of the group delay function increases. Hence the height and bandwidth can be simultaneously changed albeit with a constant linear phase delay as can be seen from Equation 4.

3. Reconstruction of Vowel Sounds using Variable Height Bandwidth Group Delay Filters

In the reconstruction of vowel sounds using variable height bandwidth group delay filters, it is always feasible to design a robust single peak variable height bandwidth group delay filter using different center frequencies. The center frequencies need to be aligned with the formants of the vowel that needs to be reconstructed. The group delay spectrum corresponding to a single formant (peak) function is given by [8]

$$D(f) = \tau_0 \exp\left(-\pi \frac{(f-f_c)^2}{f_b}\right)$$

(5)

where $f_c$ is the center frequency, $f_b$ is the parameter that defines bandwidth, $\tau_0$ is the group delay value at a particular center frequency. The bandwidth parameter $f_b$ is expressed as a function of equivalent rectangular bandwidth (ERB) which indicates the bandwidth of the auditory filter as

$$f_b = c_e \ast ERB$$

(6)

where $c_e$ is a fixed parameter named as Bandwidth Coefficient. The ERB is computed as a function of the center frequency as

$$ERB = 24.7(4.37f_c/1,000 + 1)$$

(7)

The reconstruction of vowels with variable height-bandwidth group delay is performed using formant synthesis. The vowels are reconstructed in this work by passing an impulse train of certain pitch period through this single peak group delay filter centered at the formants of the vowel.

3.1. Appearance of spurious peaks at instance of significant excitation

The output signal reconstructed is found to replicate a synthetic vowel but with spurious peaks of high magnitude which appear at the instants of significant excitation (pitch). These spurious peaks cause considerable distortion in the reconstructed vowel and are primarily due to the zeros of the group delay filter that are in close proximity to the unit circle. Figure 2a illustrates the pole zero plot of the desired filter. It can be observed that some of the poles and zeros are shifted more towards the unit circle. This implies that the second order filter corresponding to these poles and zeros have a higher peak value. Hence these contribute largely in the determination of peak of the designed filter and while the other poles and zeros contribute in the determination of bandwidth. Figure 2b, illustrates the impulse response of the designed group delay filter for a fixed pitch period. As can be seen from this Figure, there are spurious peaks at the instants of significant excitation.

3.2. Removal of spurious peaks in the reconstructed vowel

When the zeros are completely removed as can be seen in Figure 3a, the vowel quality improves. This can be observed from Figure 3b, where the spurious peaks are completely removed when the same impulse response with the same pitch period is passed through the group delay filter with the zeros. In order to
analyze this mathematically we consider a group delay filter (all pass) with the transfer function

$$H(z) = \frac{\sum_{k=0}^{N} A(N-k) z^{-k}}{\sum_{k=0}^{N} A_k z^{-k}}$$

The phase response of the filter is given by

$$\alpha = N\omega + 2\theta(\omega)$$

where the contribution of \( \theta \) is due to the zeros and \(-\theta\) is due to the poles which adds up 2\( \theta \) in terms of spectral phase. The group delay function can therefore be written as

$$\tau(\omega) = N + 2\theta'(\omega)$$

When zeros are removed from the group delay filter the group delay function can be written as

$$\tau(\omega) = \theta'(\omega)$$

It can be observed herein that by removing the zeros from the group delay filter, its all pass property is lost but the parameter (formant) specification for the desired group delay filter can still be realized.

Table 1: Optimal height and bandwidth co-efficient (BWCF) data for different vowels after subjective evaluation where F1, F2, F3 are the 1st, 2nd and 3rd formant of the vowel respectively

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Height</th>
<th>BWCF</th>
<th>Height</th>
<th>BWCF</th>
<th>Height</th>
<th>BWCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>'ah'</td>
<td>42</td>
<td>0.8</td>
<td>50</td>
<td>0.7</td>
<td>32</td>
<td>0.45</td>
</tr>
<tr>
<td>'ih'</td>
<td>35</td>
<td>0.8</td>
<td>50</td>
<td>0.6</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>'ih'</td>
<td>30</td>
<td>0.8</td>
<td>50</td>
<td>0.8</td>
<td>32</td>
<td>0.8</td>
</tr>
<tr>
<td>'ae'</td>
<td>45</td>
<td>0.8</td>
<td>55</td>
<td>0.7</td>
<td>34</td>
<td>0.6</td>
</tr>
<tr>
<td>'oo'</td>
<td>53</td>
<td>0.7</td>
<td>46</td>
<td>0.8</td>
<td>40</td>
<td>0.3</td>
</tr>
<tr>
<td>'iy'</td>
<td>40</td>
<td>0.9</td>
<td>53</td>
<td>0.7</td>
<td>38</td>
<td>1</td>
</tr>
<tr>
<td>'eh'</td>
<td>39</td>
<td>0.8</td>
<td>56</td>
<td>0.7</td>
<td>45</td>
<td>0.6</td>
</tr>
</tbody>
</table>

4. Spectral Reconstruction of speech using optimal height-bandwidth group delay filters

In this Section, we illustrate the significance of this technique on the reconstruction of vowel sounds and on actual speech signals. In Figure 4a, LPC spectra of reconstructed vowel is plotted along with the LPC spectra of original vowel. It can be seen from this figure that both the spectra are closely related with the formants of both the vowels coinciding. Similar results can be seen for the FFT spectra in Figure 4b. Prior work on speech reconstruction using fixed height-bandwidth group delay filter (function) and spectral magnitude can be seen in [8], [9] and [10]. In this work formants are extracted from the speech signal using various methods as described in [11], and a complete speech signal is reconstructed by designing variable height-bandwidth group delay filters around the formants of the signal. Group delay filters at the given formant frequencies of this signal are designed by carefully varying its height and bandwidth simultaneously. The speech signal is then reconstructed using the optimal height bandwidth for each frame. The results of using such an approach on the utterance "where were you while we were away" are shown in Figures 5a, 5b, 5c and 5d, where it can be noted that using an optimal height-bandwidth group delay filter yields better reconstruction as compared to other methods as can be seen from Figures 5b and 5c.

5. Performance Evaluation

In this Section, experiments on perceptual evaluation are performed on the reconstructed signals to investigate the effectiveness of the optimal height-bandwidth group delay filters.

5.1. Experiments on subjective evaluation of the reconstructed vowel sounds

In order to evaluate the quality of reconstructed vowel sound, an established subjective test protocol [12] is followed. Formant data of vowels are taken from Hillenbrand [2] database. The same procedure for vowel reconstruction is followed as in
Section 3.3. The optimal height and bandwidth for each vowel was found by varying the height and bandwidth. Experiments were also conducted by keeping the height and bandwidth of group delay filters at any random value other than the optimal height and bandwidth. The height of each filter is varied in the range of 20 – 60 with the constraint that the bandwidth coefficient and height is kept same for all the formants of corresponding vowel. The reconstructed vowel sounds were compared with corresponding reference signal taken from the respective database. Twenty listeners were involved in this experiment. The parameters used for the experiments were global quality of reconstructed sound, the effective similarity between the reference and reconstructed vowel and percentage noise present in the signal. A listener was first familiarised with different vowel sounds of synthetic vowels, aimed to train the subject for the required test. The subject then rated the overall quality of reconstructed vowel on a range from 0-100. The mean opinion score and standard deviation for the task is calculated from all the candidates. The values are then tabulated as in Table 2.

5.2. Experiments on objective evaluation of the reconstructed vowel sounds

To determine the quality of the reconstructed vowel sounds on an objective test scale, experiments on perceptual evaluation [13] and [14], are conducted. Formant data from Hillenbrand database was taken for vowels ‘ah’, ‘ih’ and ‘uh’. Experiments are conducted on the reconstruction of vowels through optimal height and bandwidth (OHB) and by varying height of filter at a formant between the range 20 – 60, for different bandwidth coefficients. Meanwhile the height and bandwidth parameters for other formants of the vowel were kept constant at their optimal values. This experiment was conducted to show the variation in quality of vowel sound with the variation of height and bandwidth of group delay filters and its comparison with vowel reconstructed from optimal height bandwidth filters. The overall quality of reconstructed vowels were evaluated. The objective measures used were PESQ (Perceptual Evaluation of Speech Quality) and PSMt (Perceptual Similarity Measure). The mean PESQ and PSMt scores are tabulated in Table 3 and indicate that the OHB group delay filter performs reasonably better than other methods.

![Spectrogram of the original TIMIT utterance](image1)

![Spectrogram of the utterance reconstructed from spectral magnitude](image2)

![Spectrogram of the utterance reconstructed using fixed height-bandwidth group delay filter for different frames](image3)

![Spectrogram of the utterance reconstructed using optimal height-bandwidth group delay filter for different frames](image4)

Figure 5: Spectrograms of the utterance “where were you while we were away” from the TIMIT database using different methods

<table>
<thead>
<tr>
<th>Vowel</th>
<th>BWCF</th>
<th>GQ</th>
<th>ES</th>
<th>PN</th>
</tr>
</thead>
<tbody>
<tr>
<td>'ah'</td>
<td>OHB</td>
<td>97.21</td>
<td>1.34</td>
<td>99.45</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>76.24</td>
<td>2.93</td>
<td>54.34</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>69.31</td>
<td>2.58</td>
<td>53.94</td>
</tr>
<tr>
<td>'ih'</td>
<td>OHB</td>
<td>95.31</td>
<td>1.37</td>
<td>73.91</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>73.22</td>
<td>2.80</td>
<td>55.14</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>71.91</td>
<td>1.83</td>
<td>50.75</td>
</tr>
<tr>
<td>'uh'</td>
<td>OHB</td>
<td>95.33</td>
<td>1.74</td>
<td>88.49</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>74.73</td>
<td>1.77</td>
<td>53.78</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>69.34</td>
<td>1.91</td>
<td>55.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vowel</th>
<th>BWCF</th>
<th>F1 PESQ</th>
<th>F1 PSMt</th>
<th>F2 PESQ</th>
<th>F2 PSMt</th>
<th>F3 PESQ</th>
<th>F3 PSMt</th>
</tr>
</thead>
<tbody>
<tr>
<td>'ah'</td>
<td>OHB</td>
<td>4.40</td>
<td>0.98</td>
<td>4.40</td>
<td>0.94</td>
<td>4.40</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>3.8</td>
<td>0.87</td>
<td>3.9</td>
<td>0.89</td>
<td>3.8</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3.0</td>
<td>0.78</td>
<td>3.3</td>
<td>0.84</td>
<td>3.1</td>
<td>0.80</td>
</tr>
<tr>
<td>'ih'</td>
<td>OHB</td>
<td>4.3</td>
<td>0.94</td>
<td>4.4</td>
<td>0.98</td>
<td>4.3</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>3.2</td>
<td>0.80</td>
<td>3.0</td>
<td>0.79</td>
<td>3.2</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3.5</td>
<td>0.82</td>
<td>3.7</td>
<td>0.81</td>
<td>3.4</td>
<td>0.80</td>
</tr>
<tr>
<td>'uh'</td>
<td>OHB</td>
<td>4.0</td>
<td>0.89</td>
<td>4.2</td>
<td>0.92</td>
<td>4.4</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>3.7</td>
<td>0.87</td>
<td>3.1</td>
<td>0.73</td>
<td>4.4</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3.8</td>
<td>0.88</td>
<td>3.2</td>
<td>0.89</td>
<td>3.6</td>
<td>0.81</td>
</tr>
</tbody>
</table>

6. Conclusion

The significance of varying the height and bandwidth of the group delay spectrum by designing appropriate group delay filters in the frequency domain has been investigated in this paper. Experiments on speech reconstruction and vowel perception indicate that use of an optimal height-bandwidth group delay filter improves the perceptual quality of vowel sounds when compared to fixed height-bandwidth group delay filters. Design of optimal height bandwidth filters adaptively on a frame wise basis for continuous speech is challenging requires further investigation. The use of variable height-bandwidth group delay filters in feature extraction for speech recognition also needs to be explored.
7. References


