Robust formant detection using group delay function and stabilized weighted linear prediction

Dhananjaya Gowda, Jouni Pohjalainen, Mikko Kurimo, Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Finland
dhananjaya.gowda@aalto.fi, jpohjala@acoustics.hut.fi, mikko.kurimo@aalto.fi, paavo.alku@aalto.fi

Abstract

In this paper, we propose a robust spectral representation for detecting formants in heavily degraded conditions. The method combines the temporal robustness of the stabilized weighted linear prediction (SWLP) with the robustness of group delay (GD) function in the frequency domain. Weighting of the cost function in linear prediction analysis with the short-time energy of the speech signal improves the robustness of the resultant spectrum. It also improves the accuracy of the estimated resonances as the weighting function gives more weightage to the closed phase of the glottal cycle, which is also the high SNR region of the signal. The group delay spectrum computed as the sum of individual resonances denoted by the roots of the SWLP coefficients, improves the robustness of weaker higher order resonances. The proposed SWLP-GD spectrum performs better than the conventional LP spectrum and the STRAIGHT spectrum in terms of spectral distortion measure and formant detection accuracies.

Index Terms: stabilized weighted linear prediction, SWLP, group delay, formant detection, robust spectrum estimation

1. Introduction

Linear prediction (LP) is widely used for the detection and tracking of formants in speech signals [1–3]. One of the main problems with the conventional LP is its robustness, and its performance in providing a smooth spectral representation is known to deteriorate heavily with degradation [4]. Several methods have been proposed to improve the robustness of the LP based spectral representation [5–8]. Unlike other, typically iterative methods [6–8], weighted linear prediction (WLP), proposed by Ma et al. [5], provides a non-iterative way of computing the LP coefficients. It weights the prediction error cost function by the short-time energy (STE). It has been shown that the WLP provides better spectral envelopes than the conventional LP in degraded environments [9].

The WLP method provides robustness against degradation, but it does not guarantee the stability of the all-pole model estimated. Weight modification for WLP-based methods, which ensures the stability of the all-pole model, has been proposed by Magi et al., [10]. It has also been shown that this stabilized weighted linear prediction (SWLP), used in conjunction with STE as the weight function, provides better spectral envelopes than the conventional LP and the minimum variance distortionless response (MVDR) models in the case of additive noise degradations [10, 11]. However, it has also been noted [10] that the use of SWLP broadens the spectral peaks, which may affect its performance in formant detection and tracking in degraded environments.

In this paper, we study the effect of using the STE weight function on the smoothness of the spectral envelope, as well as its ability in providing evidence for detecting formants in the face of degradations. A new spectral representation is proposed by combining the advantages of the SWLP method [10] and the group delay function [12, 13]. The additive property of the group delay function is used to improve the robustness of the SWLP spectra in the detection of formants in degraded environments. While there are several methods proposed for reliable formant tracking, most of these methods do not clearly discriminate between improvements provided by the tracking algorithm and those provided by the underlying representation [14–20]. In this paper, we do not focus on the tracking part, but focus only on the different spectral representations that can provide evidence for formant tracking.

The paper is organized as follows: Section 2 gives an overview of the SWLP method using short-time energy as the weight function. Section 3 describes the computation of the group delay function from the LP coefficients. A brief discussion of the group delay spectrum computed using the SWLP method is given in Section 4. In Section 5, average log-spectral distortion is used to evaluate the robustness of the proposed SWLP-GD spectrum compared to two popular spectral representations. The performance of the SWLP-GD spectrum in formant detection evaluated against the reference methods is given in Section 6. Section 7 gives a summary of the work.

2. Stabilized weighted linear prediction

In conventional linear prediction (LP) the given speech signal is modeled by minimizing the error between the current sample \( x_n \) and its estimate \( \hat{x}_n \) obtained as a linear combination of the past \( p \) samples, given by

\[
e_n(a) = (x_n - \hat{x}_n) = x_n + \sum_{i=1}^{p} a_i x_{n-i},
\]

where \( a = [a_0\ a_1\ \cdots\ a_p]^T \), with \( a_0 = 1 \), denotes the coefficients of an all-pole filter \( H(z) = 1/(\sum_{i=0}^{p} a_i z^{-i}) \) in the source-system approximation of the speech signal given by \( X(z) = H(z)E(z) \). \( X(z) \) and \( E(z) \) are the z-transforms of the speech signal \( x_n \) and the error signal \( e_n \), respectively. \( H(z) \) denotes the all-pole synthesis filter representing the vocal tract system and \( E(z) \) denotes the excitation source signal.

Weighted linear prediction (WLP) involves estimating the filter coefficients of the all-pole model by minimizing a weighted cost function given by

\[
E(a) = \sum_{n=1}^{N+p} (e_n(a))^2 w_n = a^T Ra,
\]

where \( N \) is the number of speech samples over which the cost function is minimized. \( R \) is the weighted autocorrelation matrix.
given by

\[ R = \sum_{i=1}^{N+p} w_n x_n x_n^T, \]  

where \( x_n = [x_n \ x_n - 1 \ \cdots \ x_{n-M}]^T \). The difference between WLP and conventional LP is the weight function \( w_n \). It should be noted that weighting in WLP is imposed on the square of the residual in the filter optimization and should not be confused for, with, for example, the ordinary Hamming windowing which is used in the autocorrelation computation in conventional LP. One possible option for the weight function is the short-time energy (STE) of the speech signal given by

\[ w_n = \sum_{i=0}^{M-1} x_{n-i-1}^2, \]  

where \( M \) is the length of the STE window, and \( x_n = 0 \) for \( n < 0 \) and \( n \geq N \).

Figure 1 shows a segment of a speech signal and the corresponding electroglottograph (EGG) signal recorded simultaneously. The STE computed using \( M = 10 \) for a sampling rate of 8 kHz and the differentiated EGG (dEGG) signal are also plotted as dotted lines. It can be seen that the choice of STE as the weight function gives more emphasis on the high-SNR closed-phase region of the glottal cycle. While this can provide better weight function gives more emphasis on the high-SNR closed-phase region within one glottal cycle are also marked.

3. Linear prediction group delay function

The group delay function, defined as the negative derivative of the phase spectrum [21] of the all-pole LP filter \( H(z) = 1/A(z) \) is given by

\[ \tau(\omega) = -\frac{\partial}{\partial \omega} \theta(\omega) = -\frac{\partial}{\partial \omega} \tan^{-1} \left\{ -\frac{A_I(\omega)}{A_R(\omega)} \right\}, \]  

\[ \tau(\omega) = \frac{A_R(\omega) A_I'(\omega) - A_I(\omega) A_R'(\omega)}{A_R(\omega)^2 + A_I(\omega)^2}, \]  

where \( \theta(\omega) \) is the phase response of \( H(z) \), and \( A_R(\omega) \) and \( A_I(\omega) \) are the real and imaginary parts, respectively, of the frequency response of the LP inverse filter \( \hat{A}(\omega) = A(\omega) \) \[ \text{e}^{\text{i} \omega} \]. Here \( j = \sqrt{-1} \) denotes the imaginary number.

The group delay function of the LP filter can be expressed as the sum of group delay functions of the individual roots of the polynomial \( A(z) \), and is given by

\[ \tau(\omega) = \sum_{i=1}^{n_r} \tau_i(\omega) + \sum_{j=1}^{n_c} \tau_c(j, \omega), \]  

where \( \tau_i(\omega) \) is the GD function of a real root, \( \tau_c(j, \omega) \) is the GD function of a complex conjugate root pair, and \( n_r \) and \( n_c \) are the number of real roots and complex conjugate root pairs, respectively. Eq. (9) can now be used to compute the individual group delay functions for real and complex conjugate root pairs. The frequency response of the inverse filter for a real root \( z = \beta \) is given by \( \hat{A}(\omega) = A_R(\omega) + j A_I(\omega) = (1 - \beta \cos \omega) + j \beta \sin \omega \). Similarly, the frequency response of the inverse filter for a pair of complex conjugate root pair \( z = (x + j y) \) is given by \( \hat{A}(\omega) = A_R(\omega) + j A_I(\omega) = (1 - 2x \cos \omega + x^2 + y^2) \cos 2\omega + j(2x \sin \omega - (x^2 + y^2) \sin 2\omega) \).

The group delay function of the all-pole filter can also be computed by differentiating the unwrapped phase spectrum of the inverse filter \( A(z) \). Computing the group delay function as a sum of group delays of the individual roots eliminates the need for phase unwrapping which can be problematic at times. It also provides the flexibility to handle unstable roots in the case of conventional LP or WLP methods by converting a mixed-phase system to a minimum-phase system by flipping the poles outside the unit circle radially inward.

4. Stabilized weighted linear prediction

group delay spectrum

SWLP provides a stable and robust representation of the instantaneous shape of the vocal tract system. The choice of STE as the weight function provides robustness by emphasizing the high-SNR regions in the time domain [9]. However, it is also observed that it smoothens the spectrum compared to the conventional LP with an increase in the bandwidth of the individual resonances, as can be seen from Figures 2(b) and 2(c). The magnitude spectra of linear predictive models may not show prominently some weak resonances at higher frequencies or close to a strong resonance [12]. This can be seen in the case of the third formant at around 2.5 kHz in Figures 2(b) and 2(c). Detection of such weak formants becomes difficult especially in degraded environments. These weak formants are better highlighted by computing the group delay function of the SWLP spectrum, as can be seen from Figure 2(d). We refer to this as the stabilized weighted linear prediction group delay (SWLP-GD) spectrum. The additive property of the group delay function provides better resolution in the frequency domain.

Figure 1: (a) Speech signal and STE (dotted line). (b) EGG, dEGG, and STE (for \( M = 10 \)) signals. Approximate closed and open phase regions within one glottal cycle are also marked.
which can improve the robustness of the representation against degradation.

5. Robustness of SWLP-GD spectrum

The robustness of the SWLP-GD spectrum is measured in terms of average log-spectral distortion between spectral estimates of the vocal tract $H[k]$ computed from the clean signal and $\hat{H}[k]$ computed from the same signal corrupted with additive noise degradations. The average log-spectral distortion is defined as

$$d = \frac{1}{L} \sum_{m=1}^{L} \frac{1}{N_m} \sum_{n=1}^{N_m} \left[20 \log_{10}(H[k]) - 20 \log_{10}(\hat{H}[k])\right],$$

(11)

where $L$ is the number of utterances, $N_m$ is the number of frames in the $m^{th}$ utterance, $K$ is the FFT length, and $H[k]$ and $\hat{H}[k]$ denote clean and degraded spectra proportional to the magnitude spectrum.

Since the SWLP-GD spectrum has negative values and is proportional to the square magnitude spectrum, it is processed to avoid logarithm of negative values and to make it proportional to the magnitude spectrum, as given by

$$H_p[k] = (H[k] - \min(H[k]) + 1)^{1/2},$$

(12)

where $H_p[k]$ denotes a positive spectrum proportional to the magnitude spectrum. The positive spectrum is normalized to unit energy before computing the log-spectral distortion, and is given by

$$H_n[k] = \frac{H_p[k]}{\left(\sum_{k=0}^{K/2} H_p[k]^{1/2}\right)^{1/2}}.$$

(13)

where $H_n[k]$ denotes the normalized spectrum.

SWLP-GD method is compared against the SWLP, the conventional LP, and the STRAIGHT method which provides a smoother spectral envelope and is widely used in speech synthesis applications [22]. The magnitude spectra of the SWLP, LP, and STRAIGHT methods are also normalized to unit energy similar to Eq. (13) before computing the log-spectral distortion. The data used for computing the spectral distortion is part of the TIMIT corpus [23] denoted as coretest containing a total of 192 utterances, 8 utterances each from 24 different speakers (8 female and 16 male). Four different types of degradation, namely, babble, factory, volvo (inside car/vehicle) and white noise from the NOISEX database [24] are added to the clean TIMIT utterances to obtain the desired average segmental SNR.

The log-spectral distortion between the clean and degraded spectra computed using SWLP-GD, SWLP, LP, and STRAIGHT methods for different types of degradation and for different SNRs are shown in Table 1. A value of $M = p = 13$ is used for the STE computation in SWLP and SWLP-GD methods. An order of $p = 13$ is used so as to be consistent with the formant detection experiment discussed in the next section, although $p = 10$ gives slightly lower spectral distortion scores for all LP based methods. It can be seen that the SWLP-GD spectrum has the least log-spectral distortion for all degradations and for all SNRs. It can be seen that the SWLP method performs better than the conventional LP, while the SWLP-GD improves upon the SWLP method. One possible reason for the poor performance of STRAIGHT method is the presence of residual harmonic ripples of the fundamental frequency, and requires a reliable pitch estimation algorithm to compensate for the ripples.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Method</th>
<th>SNR (in dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10 dB</td>
</tr>
<tr>
<td>White</td>
<td>SWLP-GD</td>
<td>1.68</td>
</tr>
<tr>
<td></td>
<td>SWLP</td>
<td>3.67</td>
</tr>
<tr>
<td></td>
<td>LP</td>
<td>5.56</td>
</tr>
<tr>
<td></td>
<td>STRAIGHT</td>
<td>5.04</td>
</tr>
<tr>
<td>Babble</td>
<td>SWLP-GD</td>
<td>1.14</td>
</tr>
<tr>
<td></td>
<td>SWLP</td>
<td>1.72</td>
</tr>
<tr>
<td></td>
<td>LP</td>
<td>2.82</td>
</tr>
<tr>
<td></td>
<td>STRAIGHT</td>
<td>2.90</td>
</tr>
<tr>
<td>Factory</td>
<td>SWLP-GD</td>
<td>1.35</td>
</tr>
<tr>
<td></td>
<td>SWLP</td>
<td>2.31</td>
</tr>
<tr>
<td></td>
<td>LP</td>
<td>3.82</td>
</tr>
<tr>
<td></td>
<td>STRAIGHT</td>
<td>3.64</td>
</tr>
<tr>
<td>Volvo</td>
<td>SWLP-GD</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>SWLP</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>LP</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>STRAIGHT</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 1: Log spectral distortion between clean and degraded spectra for different spectrum estimation methods.

6. Formant detection

The performance of the proposed SWLP-GD spectrum in formant detection and its robustness against degradations is studied. At this point we would like to discriminate between the formant detection and formant tracking tasks. While several tracking algorithms have been proposed for tracking formant contours, they can in principle be applied on evidence available from any underlying spectral representation. Therefore, in this paper we evaluate the ability of different spectral representations in providing evidence for formant detection, without the use of any tracking algorithm. The different spectral representations considered are based on the SWLP-GD, SWLP, LP, and STRAIGHT methods.

The performance of formant detection is evaluated on the test utterances of the natural vocal tract resonance (VTR) database [25]. The test data of the VTR database has 192 utterances, 8 utterances each from 24 different speakers (8 female and 16 male), for which the reference formant frequen-
cies are available for the first three formants. The reference formant frequencies have been obtained in a semi-supervised manner, where the formant tracks derived using an LP based algorithm ([26]) is verified and corrected manually based on spectrographic evidence. While there may be a concern on the accuracy of the reference formants, nevertheless they provide a decent reference on natural continuous speech for evaluating formant detection algorithms. The data originally recorded at 16 kHz sampling rate is downsampled to 8 kHz before processing.

A common framework is employed for detecting formants using each of the representations, and for evaluating the performance of formant detection. All methods process the preemphasized (using a factor of 0.97) speech signal at a frame rate of 100 frames per second. Hamming windowed segments of size 30 ms are used for SWLP-GD, SWLP, and LP methods, whereas the STRAIGHT uses the default 40 ms frame size recommended by its authors. A prediction order of $p = 13$ is used for all the three LP based methods, and $M = p = 13$ is used for STE computation in the SWLP-GD method. A slightly higher order of prediction is used for the formant detection task to allow for any spurious peaks that may be introduced due to degradations. Nevertheless, each method is allowed to hypothesize only the strongest four peaks as the formants. The peaks in the spectrum are detected by convolving the spectrum with a difference Gaussian window of width 100 Hz and picking the negative zero-crossings. Each of the three reference formants is associated with the nearest hypothesized formant location which is within 300 Hz of deviation and less than 25% of the reference formant frequency. If more than one reference formant is mapped to a single hypothesized formant, the reference formant closest to the hypothesized formant is picked. The other reference formant is allowed to pick the next best match satisfying the condition stated above. The performance of the different spectral representations in providing evidence for formant detection is shown in Table 2 in terms of percentage detection. The performance is measured only over the vowel, diphthong and semivowel regions. It can be seen that the SWLP-GD spectrum provides the best evidence for formant detection compared to SWLP, LP and STRAIGHT spectra. It can be seen that the STRAIGHT spectrum provides better evidence than the SWLP spectrum for all degradations except for white noise degradation. However it should be noted that the SWLP provides better spectral distortion score compared to STRAIGHT as can be seen from Table 1. This may be due to the smoothing of the spectrum caused by the weighting function. The combination of the group delay function and the SWLP improves upon the spectral evidence for formant detection, while at the same time keeping the spectral distortion scores the least.

### Table 2: Performance of formant detection in terms of percentage detection.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>SWLP-GD</th>
<th>SWLP</th>
<th>LP</th>
<th>STRAIGHT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>F2</td>
<td>F3</td>
<td>F1</td>
</tr>
<tr>
<td>Clean</td>
<td>89.16</td>
<td>90.56</td>
<td>85.72</td>
<td>77.95</td>
</tr>
<tr>
<td>White</td>
<td>10</td>
<td>86.23</td>
<td>84.44</td>
<td>71.32</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>82.69</td>
<td>79.61</td>
<td>64.08</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>76.37</td>
<td>72.33</td>
<td>56.88</td>
</tr>
<tr>
<td></td>
<td>-5</td>
<td>67.84</td>
<td>64.74</td>
<td>51.03</td>
</tr>
<tr>
<td>Babble</td>
<td>10</td>
<td>86.40</td>
<td>87.97</td>
<td>80.62</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>83.19</td>
<td>85.09</td>
<td>76.88</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>79.37</td>
<td>81.27</td>
<td>71.11</td>
</tr>
<tr>
<td></td>
<td>-5</td>
<td>74.72</td>
<td>74.73</td>
<td>64.59</td>
</tr>
<tr>
<td>Factory</td>
<td>10</td>
<td>85.95</td>
<td>87.24</td>
<td>77.93</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>82.05</td>
<td>83.44</td>
<td>72.25</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>76.40</td>
<td>78.56</td>
<td>65.20</td>
</tr>
<tr>
<td></td>
<td>-5</td>
<td>67.78</td>
<td>71.60</td>
<td>58.70</td>
</tr>
<tr>
<td>Volvo</td>
<td>10</td>
<td>88.54</td>
<td>90.30</td>
<td>85.77</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>87.62</td>
<td>90.05</td>
<td>85.57</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>85.91</td>
<td>89.40</td>
<td>85.08</td>
</tr>
<tr>
<td></td>
<td>-5</td>
<td>82.46</td>
<td>88.20</td>
<td>84.45</td>
</tr>
</tbody>
</table>

### 7. Conclusions

In this paper, we proposed a robust spectral representation for formant detection by combining the advantages of the SWLP method and group delay function. The SWLP spectrum provides robustness in the temporal domain by giving more weight to the high-SNR glottal closure regions. The resulting broadening of the spectral peaks was countered by computing the group delay function of the SWLP coefficients. The additive property of the group delay function provides better resolution of the spectral peaks and hence provides robustness in the frequency domain. The robustness achieved by combining the advantages of the SWLP method and the GD function was studied by computing the log-spectral distortion between clean spectrum and the spectrum obtained by degrading the speech signal with additive noise. The usefulness of the proposed SWLP-GD spectral representation in formant detection was also demonstrated. The proposed spectral representation shows promise to be used in other speech analysis problems, such as deriving robust features for different classification and recognition systems, especially in adverse environments.

### 8. Acknowledgement

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7) under grant agreement no. 287678 (Simple4All), and The Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170).
9. References


