Enhancing Noise Robustness in Automatic Speech Recognition Using Stabilized Weighted Linear Prediction

*Jouni Pohjalainen, Carlo Magi, Paavo Alku*
Helsinki University of Technology
Dept. of Signal Processing and Acoustics, Espoo, Finland
Contents

1. Background
2. The SWLP method
3. Recognition experiments
4. Conclusions
1. Background

Conventional linear prediction (LP):

- The most widely used auto-regressive (AR) modeling technique of speech.
- Vulnerable to background noise.

=>

Several robust LP methods have been developed during the past decades.
Some examples of previous works on robust LP methods:


Statistical approach based on maximum a posteriori estimation


Effects of uncorrelated white noise are removed by modifying the autocorrelation term $R(0)$

Selection of samples with a two-stage analysis to decrease the effects of glottal excitation


Optimization of the predictor by minimizing the sum of weighted residuals: cost functions are used to give more weight to smaller residuals and down-weight large residuals

Weighting of the square of the residual with the short time energy function


The use of higher order statistics: LP parameters are computed from cumulants
Present study:

The concept of weighted linear prediction (WLP), based on Ma et al. (1993), is revisited by:

- Introducing Stabilized Weighted Linear Prediction (SWLP), a WLP method which guarantees the stability of the all-pole filter

- SWLP is used in recognition of noisy speech as a feature extraction technique
2. The Stabilized Weighted Linear Prediction (SWLP) Method

- Both in the conventional LP and WLP, speech sample $x_n$ is estimated as a linear combination of the $p$ past samples:

$$\hat{x}_n = -\sum_{i=1}^{p} a_i x_{n-i}$$  \hspace{1cm} (1)

- Prediction error, the residual, is defined as:

$$\varepsilon_n(a) = x_n + \sum_{i=1}^{p} a_i x_{n-i} = a^T x_n$$  \hspace{1cm} (2)
- Minimization problem:

\[
\minimize \ E(a) \ \text{subject to} \ \mathbf{a}^T \mathbf{u} = 1
\]

where \( \mathbf{u} \) is the unit vector

- In weighted LP, the prediction error is defined:

\[
E(a) = \sum_{n=1}^{N+p} (\varepsilon_n(a))^2 w_n
\]

(3)

\[
= \mathbf{a}^T \left( \sum_{n=1}^{N+p} w_n \mathbf{x}_n \mathbf{x}_n^T \right) \mathbf{a} = \mathbf{a}^T \mathbf{R} \mathbf{a}
\]

where \( \mathbf{R} \) is a weighted autocorrelation matrix is:

\[
\mathbf{R} = \sum_{n=1}^{N+p} w_n \mathbf{x}_n \mathbf{x}_n^T
\]
- Minimization results in normal equations:

\[ R_a = \sigma^2 u \]  \hspace{1cm} (4)

- Weighting is computed by the short-time energy (STE) function:

\[ w_n = \sum_{i=0}^{M-1} x_{n-i-1}^2 \]  \hspace{1cm} (5)

where \( M \) denotes the length of the STE window.
Example of STE-weighting:

STE emphasizes the closed phase of the glottal cycle, that is, the time span during which formants are prominent.
- Problem with WLP: stability of the all-pole filter not guaranteed!

- Solution:

Weighted autocorrelation matrix can be expressed:

$$ R = Y^T Y $$  \hspace{1cm} (6)

where:

$$ Y = [y_0 \, y_1 \, \ldots \, y_p] \quad y_0 = [\sqrt{w_1} x_1 \ldots \sqrt{w_N} x_N \, 0 \ldots 0]^T $$

Columns $y_k$ of matrix $Y$ can be generated:

$$ y_{k+1} = B y_k \quad , \ k=0, \ 1, \ldots, \ p-1 $$  \hspace{1cm} (7)
where matrix $B$ is defined:

$$B = \begin{bmatrix}
0 & 0 & \ldots & 0 & 0 \\
\sqrt{w_2 / w_1} & 0 & 0 & \ldots & 0 \\
0 & \sqrt{w_3 / w_2} & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \ldots & \sqrt{w_{N+p} / w_{N+p-1}} & 0
\end{bmatrix}$$
It can be shown (see Magi et al., Proc. Interspeech 2007) that the all-pole filter given by the weighted LP is stable if the elements of matrix $B$ are defined:

$$B_{i+1,i} = \begin{cases} \sqrt{w_{i+1} / w_i} & \text{, if } w_i \leq w_{i+1} \\ 1 & \text{, if } w_i > w_{i+1} \end{cases}$$
Examples of SWLP spectra (vowel /a/, p=10):

M=8 => smooth envelope

M=24 => “LP-type of envelope”
3. Recognition experiments: feature extraction

○ Computation of the mel frequency cepstral coefficients (MFCC):
  1. magnitude spectrum estimation
  2. logarithmic energies from triangular bandpass filters
  3. discrete cosine transform

○ Methods for computing the short-time magnitude spectrum:
  ○ FFT
  ○ Linear prediction (LP)
  ○ Minimum variance distortionless response (MVDR)
  ○ Stabilized weighted linear prediction (SWLP)
Speaker-independent classification of isolated words

- Two word sets extracted from TIMIT:
  - T21: a vocabulary of 21 short words, plenty of training data
  - T22: a vocabulary of 22 slightly longer words, smaller training set

- Word recognition using Dynamic Time Warping (DTW)
  - Time-align sequences of 12-dimensional MFCC feature vectors
    - $\Delta$’s and $\Delta \Delta$’s are not included in the feature vector - concentrate on the aspect of short-time magnitude spectrum estimation
  - Each vocabulary word is represented by ten MFCC sequences (selected using cluster analysis)
Correct recognition rates with white noise corruption

<table>
<thead>
<tr>
<th>T21</th>
<th>MFCC VARIANT</th>
<th>SNR 10</th>
<th>SNR 5</th>
<th>SNR 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT</td>
<td>63.8</td>
<td>45.2</td>
<td>23.5</td>
<td>13.2</td>
</tr>
<tr>
<td>LP</td>
<td>67.9</td>
<td>52.1</td>
<td>33.0</td>
<td>15.4</td>
</tr>
<tr>
<td>MVDR-10</td>
<td>63.2</td>
<td>46.7</td>
<td>27.3</td>
<td>12.2</td>
</tr>
<tr>
<td>MVDR-80</td>
<td>62.5</td>
<td>44.5</td>
<td>27.0</td>
<td>11.2</td>
</tr>
<tr>
<td>SWLP-8</td>
<td>76.3</td>
<td>61.4</td>
<td>39.4</td>
<td>18.4</td>
</tr>
<tr>
<td>SWLP-24</td>
<td>75.5</td>
<td>56.0</td>
<td>33.5</td>
<td>13.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T22</th>
<th>MFCC VARIANT</th>
<th>SNR 10</th>
<th>SNR 5</th>
<th>SNR 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT</td>
<td>68</td>
<td>48</td>
<td>37</td>
<td>29</td>
</tr>
<tr>
<td>LP</td>
<td>66</td>
<td>53</td>
<td>40</td>
<td>31</td>
</tr>
<tr>
<td>MVDR-10</td>
<td>68</td>
<td>57</td>
<td>38</td>
<td>24</td>
</tr>
<tr>
<td>MVDR-80</td>
<td>68</td>
<td>48</td>
<td>35</td>
<td>25</td>
</tr>
<tr>
<td>SWLP-8</td>
<td>75</td>
<td>67</td>
<td>54</td>
<td>35</td>
</tr>
<tr>
<td>SWLP-24</td>
<td>77</td>
<td>67</td>
<td>48</td>
<td>31</td>
</tr>
</tbody>
</table>
Correct recognition rates with pink noise corruption

<table>
<thead>
<tr>
<th>MFCC VARIANT</th>
<th>SNR 10</th>
<th>SNR 5</th>
<th>SNR 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT</td>
<td>80.6</td>
<td>65.0</td>
<td>43.8</td>
</tr>
<tr>
<td>LP</td>
<td>83.5</td>
<td>69.4</td>
<td>48.5</td>
</tr>
<tr>
<td>MVDR-10</td>
<td>79.0</td>
<td>63.0</td>
<td>44.8</td>
</tr>
<tr>
<td>MVDR-80</td>
<td>82.0</td>
<td>65.0</td>
<td>43.7</td>
</tr>
<tr>
<td>SWLP-8</td>
<td>84.7</td>
<td>77.4</td>
<td>60.7</td>
</tr>
<tr>
<td>SWLP-24</td>
<td>84.3</td>
<td>74.9</td>
<td>54.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MFCC VARIANT</th>
<th>SNR 10</th>
<th>SNR 5</th>
<th>SNR 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT</td>
<td>75</td>
<td>66</td>
<td>47</td>
</tr>
<tr>
<td>LP</td>
<td>71</td>
<td>63</td>
<td>50</td>
</tr>
<tr>
<td>MVDR-10</td>
<td>73</td>
<td>63</td>
<td>50</td>
</tr>
<tr>
<td>MVDR-80</td>
<td>74</td>
<td>65</td>
<td>51</td>
</tr>
<tr>
<td>SWLP-8</td>
<td>78</td>
<td>70</td>
<td>57</td>
</tr>
<tr>
<td>SWLP-24</td>
<td>84</td>
<td>76</td>
<td>61</td>
</tr>
</tbody>
</table>

Helsinki University of Technology
4. Conclusions

Stabilized Weighted Linear Prediction (SWLP):
- An all-pole modeling method based on temporal weighting (with STE function) of the square of the residual
- Guarantees stability
- Performance (smoothness of the envelope) can be adjusted by length of the STE window (parameter M)
- Recognition setup:
  - Speaker-independent isolated word recognition using DTW
  - 12-dimensional MFCC vectors (no c(0), no deltas)
  - No feature postprocessing (such as CMS)

- With these conditions, SWLP as part of the MFCC computation outperforms other spectral estimation methods in terms of robustness against white and pink noise