Stabilised Weighted Linear Prediction - A Robust All-Pole Method for Speech Processing

Carlo Magi, Tom Bäckström, Paavo Alku

Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology
P.O. Box 3000, FI-02015 TKK, Finland

carlo.magi@acoustics.hut.fi, tom.backstrom@tkk.fi, paavo.alku@tkk.fi

10.21437/Interspeech.2007-240

Abstract

Weighted linear prediction (WLP) is a method to compute all-pole models of speech by applying temporal weighting of the residual energy. By using short-time energy (STE) as a weighting function, the algorithm over-weight those samples that fit the underlying speech production model well. The current work introduces a modified WLP method, stabilised weighted linear prediction (SWLP) leading always to stable all-pole models whose performance can be adjusted by changing the length (denoted by M) of the STE window. With a large M value, the SWLP spectra become similar to conventional LP spectra. A small value of M results in SWLP filters similar to those computed by the minimum variance distortionless response (MVDR) method. The study compares the performances of SWLP, MVDR, and conventional LP in spectral modelling of speech sounds corrupted by Gaussian additive white noise. Results indicate that SWLP is the most robust method against noise especially with a small M value.

Index Terms: linear prediction, all-pole modelling, spectral estimation

1. Introduction

Linear prediction (LP) is the most widely used all-pole modelling method of speech [1]. LP analysis, however, suffers from various drawbacks, such as the biasing of the formant estimates by their neighbouring harmonics [2]. Additionally, it is well-known that the performance of LP deteriorates in the presence of noise [3]. Therefore, several linear predictive methods with an improved robustness against noise have been developed (see for instance [4]). However, it is worth noticing that most of these robust modifications of linear prediction are based on the iterative update of the prediction parameters. The weighted linear prediction (WLP) tries to tackle the problem caused by glottal closure excitation by introducing a time-domain weight of the energy of the prediction error [5]. By emphasizing those data segments that have a high signal-to-noise ratio (SNR), WLP has been recently shown to yield improved spectral envelopes of noisy speech in comparison to the conventional LP analysis [6]. In contrast to many other robust methods of linear prediction, the filter parameters of WLP can, importantly, be computed without any iterative update.

The minimum variance distortionless response (MVDR) method is popular in array processing but it has recently also attracted increasing interests in speech processing where it has been used, for example, in the feature extraction of a speech recognition [7]. In [8], the following three refinements to the original form of MVDR were proposed: frequency warping, scaling of the spectral envelope, and speaker-independent model order selection. It was shown that the scaling procedure suggested in [8] improved the robustness of the MVDR spectral models against additive noise in the frequency domain.

This study addresses the computation of spectral envelopes of speech from noisy signals by comparing three all-pole modelling methods, the conventional LP, MVDR, and WLP. Because the original version of WLP presented in [5] does not guarantee stability of the all-pole model, the idea of WLP is revisited by developing weight functions which always give stable all-pole models. It will be shown that with a proper choice of parameters the proposed stabilised WLP method yields spectral envelopes similar to those given by MVDR but with improved robustness against white zero-mean Gaussian background noise.

2. Stabilised weighted linear prediction

2.1. Model formulation

The discussion is begun by shortly presenting the optimisation of the filter parameters in stabilised weighted linear prediction. The goal is to find the coefficient vector $\mathbf{a} = (a_0\ a_1\ \cdots\ a_p)^T$, of a p:th order FIR predictor, which minimises the cost function $E(\mathbf{a})$ (also known as the prediction error energy) subject to $a_0 = 1$. The corresponding all-pole filter is obtained as $H(z) = 1/A(z)$, where $A(z)$ is the z-transform of $\mathbf{a}$. The cost function in the WLP method is defined as

$$E(\mathbf{a}) = \sum_{n \in I} (\varepsilon_n(\mathbf{a}))^2 w_n, \tag{1}$$

where $\varepsilon_n(\mathbf{a}) = x_n + \sum_{i=1}^{p} a_i x_{n-i} = \mathbf{a}^T \mathbf{x}_n$ is the prediction error. (Note that according to Eq. 1, the formulation allows us to temporally emphasize the residual energy). The constrained minimisation problem defined above leads to the normal equation

$$\mathbf{R}_I \mathbf{a} = \sigma^2 \mathbf{u}, \tag{2}$$

where σ^2 is the error energy, $\mathbf{R}_I = \sum_{n \in I} w_n \mathbf{x}_n \mathbf{x}_n^T$ and $\mathbf{u} = (1\,0\,\cdots\,0)^T$. By defining the index set as $I := \{1, \ldots, N+p\}$ and assuming that the signal x_n is zero outside the interval $[1, N]$, $\mathbf{R}_I$ corresponds to the autocorrelation matrix if and only if $\forall n \in I,\ w_n = 1$. Matrix $\mathbf{R}_I$ can be expressed as $\mathbf{R}_I = \mathbf{Y}^T\mathbf{Y}$, where the matrix $\mathbf{Y} = (\mathbf{y}_0\ \mathbf{y}_1\ \cdots\ \mathbf{y}_p) \in \mathbb{R}^{(N+p)\times(p+1)}$ and $\mathbf{y}_0 = (\sqrt{w_1}x_1\ \cdots\ \sqrt{w_N}x_N\ 0\ \cdots\ 0)^T$. The columns $\mathbf{y}_k$ of the matrix $\mathbf{Y}$ can be generated via the formula: $\mathbf{y}_k = \mathbf{B}\mathbf{y}_{k+1}$, where $\mathbf{B}(i,j) = \delta_{j-i-1}\sqrt{w_i/w_{i+1}}$ (where δ refers to the Dirac delta function).

The stability of the WLP method with the STE weight function as proposed in [5], however, can not be guaranteed. Therefore, a formula for a modified weight function to be used in

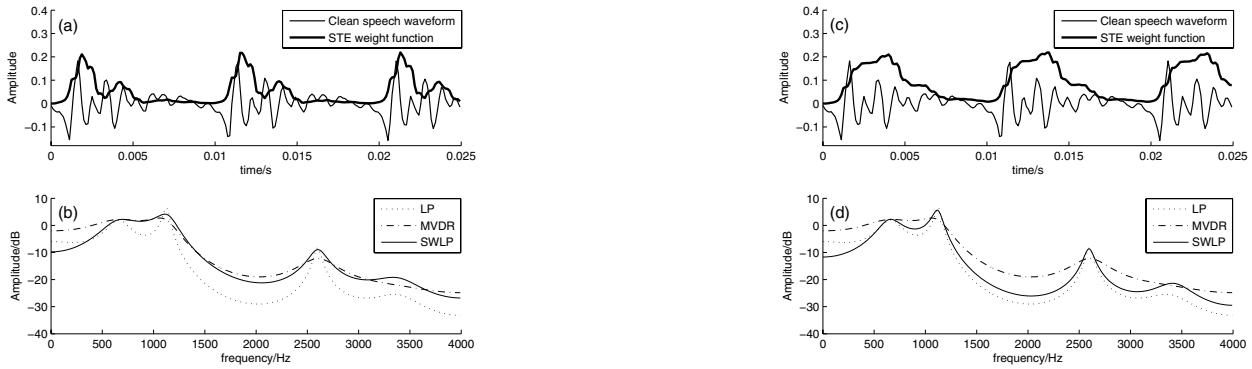

Figure 1: *Time-domain waveforms of clean speech (vowel /a/ produced by a male speaker) and short time energy (STE) weight function (upper panels) and corresponding all-pole spectra of order $p = 10$ computed by LP, MVDR, and SWLP (lower panels). SWLP analysis was computed by using two different values for the length of the STE window: $M = 8$ (left panels) and $M = 24$ (right panels).*

WLP is developed here so that the stability of the resulting all-pole filter is always guaranteed. This can be done by changing the elements of the secondary diagonal of the matrix $\mathbf{B}$ as

$$\mathbf{B}_{i,i+1} = \begin{cases} \sqrt{w_i/w_{i+1}}, & \text{if } w_i \leq w_{i+1} \\ 1, & \text{if } w_i > w_{i+1} \end{cases} \quad (3)$$

Henceforth, the WLP method computed using matrix $\mathbf{B}$, defined above, is called the *stabilised weighted linear prediction* (SWLP) model, where the stability of the corresponding all-pole filter is guaranteed because the zeros of the z-transform of the vector $\mathbf{a}$ solving Eq. 2 belong to the numerical range of the matrix $\mathbf{B}$ [9]. In this case the matrix $\mathbf{B}$, is a nilpotent operator with power of nilpotency $n = N+p$. Moreover, the norm of the matrix $\mathbf{B}$ is clearly equal to $\|\mathbf{B}\| = \max_{i\in\mathbb{I}} \mathbf{B}(i, i+1) \leq 1$. Finally, it has been proved in [10] that the numerical range of the nilpotent operator, with power of nilpotency n, is a circle (open or closed) with centre at the origin and radius ρ not exceeding $\|\mathbf{B}\| \cos(\frac{\pi}{n+1})$.

2.2. Time-domain weight function

The key concept of WLP, introduced in Eq. 1, is the time-domain weight function w_n. By choosing an appropriate waveform for w_n, one can either temporally emphasize or attenuate the weight of the residual energy prior to the optimisation of the filter parameters. In [5] the weight function was chosen based on the short-time energy (STE): $w_n = \sum_{i=0}^{M-1} x_{n-i-1}^2$, where M is the length of the STE window.In difference to [5], the idea of weight, in the current study, is motivated from the point of view of computing linear predictive models of speech that are more robust against noise than the conventional LP. This perspective, illustrated in Figs. 1(a) and 1(c) , is based on the fact that the STE function over-weights those sections of the speech waveform which comprise samples of large amplitude. These segments of speech are less vulnerable to additive uniformly distributed noise in comparison to values of smaller amplitude.

3. Results

3.1. Shape of the all-pole spectrum

The behaviour of SWLP in spectral modelling of speech is demonstrated in Fig. 1. In this figure, the analysed speech sound is shown together with the STE weight functions in the upper panel. The lower panel shows the spectra of parametric all-pole models of order $p = 10$ computed with three techniques: conventional linear prediction with the autocorrelation criterion, minimum variance distortionless response, and the proposed stabilised weighted linear prediction. In order to demonstrate the effect of the weight function length, the SWLP analysis was computed using $M = 8$ (left panel) and $M = 24$ (right panel). The examples depicted demonstrate two characteristic features of SWLP. First, the weight function computed by the STE clearly emphasizes those segments of speech where the data values are of large amplitude while segments of small amplitude values are given lesser weights. Second, the shape of the all-pole spectrum computed by SWLP is, in general, smooth. However, the behaviour of the SWLP spectrum depends on the length of the STE window: with $M = 8$, the SWLP shows a very smooth spectral behaviour reminiscent of MVDR, but for the larger M value the sharpness of the resonances in the SWLP spectrum increases and its general spectral behaviour approaches that of LP. The reason behind this is evident by referring to Eq. 3: the larger the value of M the more elements of matrix $\mathbf{B}$ are equal to unity. In other words, the general spectral shape of the SWLP filter can be made similar to MVDR by selecting a small value of M and it can be adjusted to behave in a manner close to LP by using a larger value of M.

3.2. Spectral behaviour in presence of noise

The main focus in the experiments of this study was to measure how the proposed SWLP method works for speech corrupted by additive noise and, in particular, to compare the performance of SWLP to that of LP and MVDR in spectral modelling of noisy speech. All the experiments reported in this study were conducted using a bandwidth of 4 kHz. The prediction order in all methods tested was set to $p = 10$. The frame length was 25 ms (200 samples) and no pre-emphasis was used. Noise-corrupted signals were generated by adding zero-mean Gaussian white noise to clean speech sounds. Speech data were taken from the TIMIT database [11], consisting of 12 American English sentences from four different dialect regions produced by an equal number of female and male speakers. The total number of speech frames analysed in tests was 1304, comprising both voiced and unvoiced speech sounds.

Objective evaluation of the effect of noise on all-pole modelling was computed by adapting the widely used spectral dis-

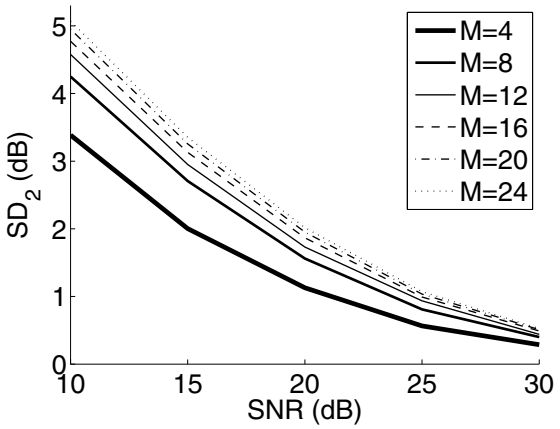

Figure 2: *Spectral distortion values (SD_2) between SWLP envelopes of order $p = 10$ computed from clean and noisy speech. The length of the STE window was varied in six steps from $M = 4$ to $M = 24$. Speech was corrupted by additive zero-mean Gaussian white noise in five SNR categories. SD_2 values were computed as an average over all the analysed segments consisting of 654 frames from the TIMIT database.*

tortion criterion, SD_2 [12]. Before SD_2 calculation the gains σ^2 of the all-pole filters, calculated from clean and noisy speech samples, were adjusted so that the impulse response energies of the filters became equal. The experiments here were begun by running a test to analyse how much the performance of SWLP is affected by additive Gaussian noise for different values of M. The total number of speech frames analysed in this test was 654, comprising both voiced and unvoiced speech sounds. The difference in the SWLP spectral models computed from clean and noisy samples was quantified in five different SNR categories by using SD_2. The experiments were conducted using six different values (4, 8, 12, 16, 20, 24) of the STE window length M. The results obtained from the first experiment are shown in Fig. 2. The data depicted show that the effect of noise on SWLP modelling depends greatly on the choice of the STE window length M: the smaller the value of M the larger the robustness of SWLP against noise. By referring to the example shown in Figure 1, this behaviour can be explained by the effect the value of M has on the shape of the STE function and, consequently, on the general shapes of the SWLP spectral models. In the case of a small M value, temporal fluctuations in the weighting function are greater than those computed with a larger value of M (see Figs. 1(a) and 1(c)). Consequently, the weighting in the case of a small M value over-emphasizes samples of large amplitude more than the weight function defined with a larger M value. In the case of zero-mean Gaussian additive noise this implies that the all-pole modelling is computed by emphasising speech samples of larger SNR over those with small SNR. Hence, the resulting SWLP model computed with a small M value is less vulnerable to additive Gaussian noise. The results shown in Fig. 2 can also be understood from the point of view of the general shape of the SWLP filter (see Figs. 1(b), 1(d)). In the case of a small M value the all-pole model indicates, also in the case of clean speech, a smoother spectral behaviour than the model computed with a larger M value. In other words, the poles of the SWLP filter computed from speech with large SNR tend to be closer to the origin of the z-plane when the STE function is computed with a small M value.

The second experiment was conducted to compare the performance of the proposed SWLP method to that of conventional LP and MVDR in spectral modelling of noisy speech. Since the behaviour of SWLP depends greatly on the value of the STE window length M, it was decided to compute the SWLP using two different values for this parameter: a large value of M corresponding to the SWLP which behaves similarly to the conventional LP and a small value of M yielding SWLP filters of smooth spectral shape similar to those computed by MVDR. The selection of the small M value was accomplished by running a special experiment in which the value of $M \in \{4, 8, 12, 16, 20, 24\}$ yielding the largest similarity between the all-pole spectra given by SWLP and MVDR was searched for. The result of the experiment showed that the smallest spectral distortion value between SWLP and MVDR spectra was achieved with $M = 8$. Because $M = 24$ was the greatest value used in previous experiments it was selected to represent the SWLP with a large M value.

Performance of LP, MVDR, and SWLP (with $M = 8$ and $M = 24$) was compared by measuring for each method how much the all-pole models computed from clean speech are different from those extracted from noisy speech. SD_2 was used as an objective distance measure between the all-pole spectra extracted from clean and noisy signals. Again, noise-corruption was measured in five SNR categories. The total number of speech frames in this test was 650. (These utterances were different from those used in the search of the M value yielding the largest similarity between SWLP and MVDR spectra). The SD_2 value for each method in each SNR category was computed as a mean over values obtained from individual frames.

The results obtained in comparing the robustness of the five all-pole modelling techniques are shown in Fig. 3. As a general trend, all methods show an increase in SD_2 when SNR decreases. In comparing conventional LP and MVDR, the results here are in line with previous findings indicating that LP is sensitive to noise while MVDR shows a clearly better performance [6]. The behaviour of SWLP, however, shows the best robustness against noise. In particular, SWLP with a small M value is able to tackle the effect of additive noise more effectively than any of the other methods tested.

Finally, in order to get tentative subjective evidence for the performance of MVDR and SWLP in the modelling of both clean and noisy speech, a small listening test was organized. In this test, subjects ($n = 13$) listened to 200 ms sounds synthesized by exciting MVDR and SWLP filters of order $p = 10$ by impulse trains. The all-pole filters were computed with MVDR and SWLP both from clean and noisy utterances corrupted with additive zero-mean Gaussian noise with SNR $= 10$ dB. Utterances consisted of eight Finnish vowels produced by one male and one female subject. The test involved a perceptual comparison between three sounds (reference and sounds A and B). The reference was always the original, clean vowel. Sounds A and B were synthesized utterances produced, in random order, by impulse train excited MVDR and SWLP filters. The listener was asked to evaluate which one of the two alternatives (A or B) sounded more like the reference. The options were *A*, *B*, or *No preference*.

The results showed that for male vowels listeners clearly preferred the SWLP quality. SWLP was preferred in 71% and 73% of evaluations of clean and corrupted sounds, respectively. MVDR, however, was preferred only in 17% and 1% of evaluations of clean and corrupted sounds, respectively. For female vowels, the preference of SWLP were somewhat smaller; it was preferred in 46% and 45% of evaluations of clean and corrupted

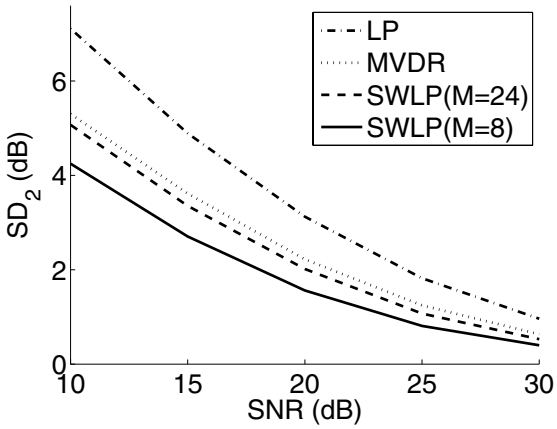

Figure 3: *Spectral distortion values (SD_2) between all-pole envelopes of order $p = 10$ computed from clean and noisy speech with LP, MVDR and SWLP (with $M = 8$ and $M = 24$). Speech was corrupted by additive zero-mean Gaussian white noise in five SNR categories. SD_2 values were computed as an average over all the analysed segments consisting of 654 frames from the TIMIT database.*

sounds, respectively. MVDR of female sounds was assessed better in 19% and 5% of evaluations of clean and corrupted sounds, respectively.

4. Conclusions

Linear prediction was analysed in this study using temporal weighting of the residual energy. The work is based on the previous study by Ma et al. [5] where the concept of weighted linear prediction was introduced by applying short time energy waveform as the weighting function. In contrast to the original work by Ma et al., the present study established a modified STE weighting which guarantees the stability of the resulting all-pole filter. Moreover the present study also focused on how the length of the STE window, the parameter M, affects the general shapes of the all-pole envelopes given by SWLP. It was shown, importantly, that by choosing the value of M properly, the behaviour of SWLP can be adjusted to be similar to either LP (corresponding to large M values) or to MVDR (corresponding to small M values).

This new method, SWLP, was then compared to two known all-pole modelling methods, LP and MVDR, by analysing speech corrupted by additive noise. It was shown that the proposed SWLP method gave the best performance in robustness against noise when quantifying the difference between the clean and noisy spectral envelopes using the objective spectral distortion measure. This finding was also corroborated by a small subjective test in which the majority of the listeners assessed quality of impulse train excited SWLP all-pole filters extracted from noisy speech to be perceptually closer to original clean speech than the corresponding all-pole responses computed by MVDR.

Finally, in our most resents experiments the proposed SWLP method was applied in speech recognition [13]. The experiments conducted show that the method gives superior performance in comparison to various other feature extraction techniques especially with speech of low SNR. Therefore, it can be argued that SWLP is a potential low order all-pole model for feature extraction.

5. Acknowledgements

This work was supported by the Academy of Finland (project number 205962).

6. References

[1] Makhoul, J., "Linear Prediction: A Tutorial Review", Proceedings of the IEEE, 63(4):561–580, 1975.

[2] El-Jaroudi, A. and Makhoul, J., "Discrete All-Pole Modeling", IEEE Transactions on Signal Processing, 39(2):411–423, 1991.

[3] Sambur, M. and Jayant, N., "LPC Analysis/Synthesis From Speech Inputs Containing Quantizing Noise or Additive White Noise", IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-24(6):488–494, 1976.

[4] Lim, J. and Oppenheim, A., "All-Pole Modelling of Degraded Speech", IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-26(3):197–210, 1978.

[5] Ma, C., Kamp, Y. and Willems, L., "Robust Signal Selection for Linear Prediction Analysis of Voiced Speech", Speech Communication, 12(1):69–81, 1982.

[6] Magi, C., Bäckström, T. and Alku, P., "Objective and Subjective Evaluation of Seven Selected All-Pole Modelling Methods in Processing of Noise Corrupted Speech", CD Proc. NORSIG 2006, Reykjavik, Iceland, 2006.

[7] Dharanipragada, S., Yapanel, U. and Rao, B., "Robust Feature Extraction for Continuous Speech Recognition Using the MVDR Spectrum Estimation Method", IEEE Transactions on Audio, Speech and Language Processing, 15(1):224–234, 2007.

[8] Wölfel, M. and McDonough, J., "Minimum Variance Distortionless Response Spectral Estimation", IEEE Signal Processing Magazine, 22(5):117–126, 2005.

[9] Delsarte, P., Genin, Y. and Kamp, Y., "Stability of Linear Predictors and Numerical Range of a Linear Operator", IEEE Transactions on Information Theory, IT-33(3):412–415, 1982.

[10] Karaev, M., "The Numerical Range of a Nilpotent Operator on a Hilbert Space", Proceedings of the American Mathematical Society, 132(8):2321–2326, 2004.

[11] Garofolo, J. et al, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (CD-ROM)", NIST, 1993.

[12] Rabiner, L. and Juang, B., "Fundamentals of Speech Recognition", Prentice Hall PTR, 1993.

[13] Magi, C., Bäckström, T. and Alku, P., "Stabilised Weighted Linear Prediction", submitted to IEEE Transactions on Speech and Audio Processing.