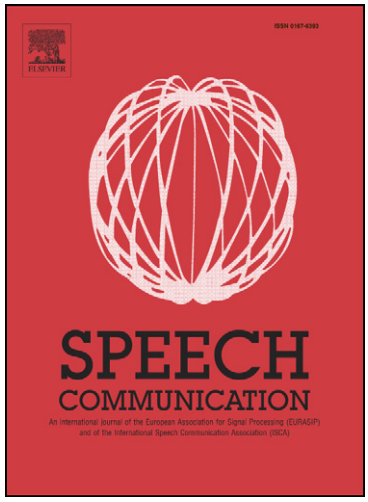
Stabilised Weighted Linear Prediction

Carlo Magi, Jouni Pohjalainen, Tom Bäckström, Paavo Alku



Stabilised Weighted Linear Prediction

Carlo Magi, Jouni Pohjalainen, Tom Bäckström, Paavo Alku

PII: S0167-6393(08)00179-9
DOI: 10.1016/j.specom.2008.12.005
Reference: SPECOM 1765

To appear in: *Speech Communication*

Received Date: 26 September 2007
Accepted Date: 1 December 2008

Please cite this article as: Magi, C., Pohjalainen, J., Bäckström, T., Alku, P., Stabilised Weighted Linear Prediction, *Speech Communication* (2009), doi: 10.1016/j.specom.2008.12.005

Stabilised Weighted Linear Prediction

Carlo Magi *,[1], Jouni Pohjalainen [2], Tom Bäckström, Paavo Alku

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

Abstract

Weighted linear prediction (WLP) is a method to compute all-pole models of speech by applying temporal weighting of the square of the residual signal. By using short-time energy (STE) as a weighting function, this algorithm was originally proposed as an improved linear predictive (LP) method based on emphasising those samples that fit the underlying speech production model well. The original formulation of WLP, however, did not guarantee stability of all-pole models. Therefore, the current work revisits the concept of WLP by introducing a modified short-time energy function leading always to stable all-pole models. This new method, stabilised weighted linear prediction (SWLP), is shown to yield all-pole models whose general performance can be adjusted by properly choosing the length of the STE window, a parameter denoted by M.

The study compares the performances of SWLP, minimum variance distortionless response (MVDR), and conventional LP in spectral modelling of speech corrupted by additive noise. The comparisons were performed by computing, for each method, the logarithmic spectral differences between the all-pole spectra extracted from clean and noisy speech in different segmental signal-to-noise ratio (SNR) categories. The results showed that the proposed SWLP algorithm was the most robust method against zero-mean Gaussian noise and the robustness was largest for SWLP with a small M-value. These findings were corroborated by a small listening test in which the majority of the listeners assessed the quality of impulse-train-excited SWLP filters, extracted from noisy speech, to be perceptually closer to original clean speech than the corresponding all-pole responses computed by MVDR. Finally, SWLP was compared to other short-time spectral estimation methods (FFT, LP, MVDR) in isolated word recognition experiments. Recognition accuracy obtained by SWLP, in comparison to other short-time spectral estimation methods, improved already at moderate segmental SNR values for sounds corrupted by zero-mean Gaussian noise. For realistic factory noise of low pass characteristics, the SWLP method improved the recognition results at segmental SNR levels below 0 dB.

Key words: Linear prediction, All-pole modelling, Spectral estimation.

1 Introduction

Linear prediction (LP) is the most widely used all-pole modelling method of speech (Makhoul, 1975). The prevalence of LP stems from its ability to estimate the spectral envelope of a voice signal and to represent this information by a small number of parameters. By modelling the spectral envelope, LP captures the most essential acoustical cues of speech originating from two major parts of the human voice production mechanism, the glottal flow (which is the physiological source behind the over-all spectral envelope structure) and the vocal tract (which is the cause of the local resonances of the spectral envelope, the formants). In addition to its ability to express the spectral envelope of speech with a compressed set of parameters, LP is known to guarantee the stability of the all-pole models, provided that the autocorrelation criterion is used. Moreover, implementation of the conventional LP can be done with a small computational complexity. LP analysis, however, also suffers from various drawbacks, such as the biasing of the formant estimates by their neighbouring harmonics (El-Jaroudi and Makhoul, 1991). This is caused by aliasing that occurs in the autocorrelation domain and the phenomenon is, in general, most severe for high-pitch voiced speech. Additionally, it is well-known that the performance of LP deteriorates in the presence of noise (Sambur and Jayant, 1976). Therefore, several linear predictive methods with an improved robustness against noise have been developed (Lim and Oppenheim, 1978; Zhao et al., 1997; Shimamura, 2004). However, it is worth noticing that most of these robust modifications of LP are based on the iterative update of the prediction parameters. Weighted linear prediction (WLP) uses time-domain weighting of the square of the prediction error signal (Ma et al., 1993). By emphasising 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 (Magi et al., 2006). In contrast to many other robust methods of LP, the filter parameters of WLP can, importantly, be computed without any iterative update.

When the order of LP increases, the spectral envelopes given by LP might overestimate the underlying speech spectrum (Murthi and Rao, 2000). This occurs especially in the analysis of voiced speech of sparse harmonic structure, in which case LP models not only the spectral envelope but also the multiples of the fundamental. The minimum variance distortionless response (MVDR) method tries to cope with this problem by providing a smooth spectral envelope even when the model order is increased. MVDR is popular in array processing but it has recently also attracted increasing interest in speech processing where it has been used, for example, in the

* Corresponding author. Tel.: +358-9-451-2479; fax: +358-9-460-224.
Email addresses: carlo.magi@acoustics.hut.fi (Carlo Magi), jpohjala@acoustics.hut.fi (Jouni Pohjalainen), tom.backstrom@tkk.fi (Tom Bäckström), paavo.alku@tkk.fi (Paavo Alku).
[1] Supported by Academy of Finland (project number 205962) and TKK.
[2] Supported by Academy of Finland (project number 107494).

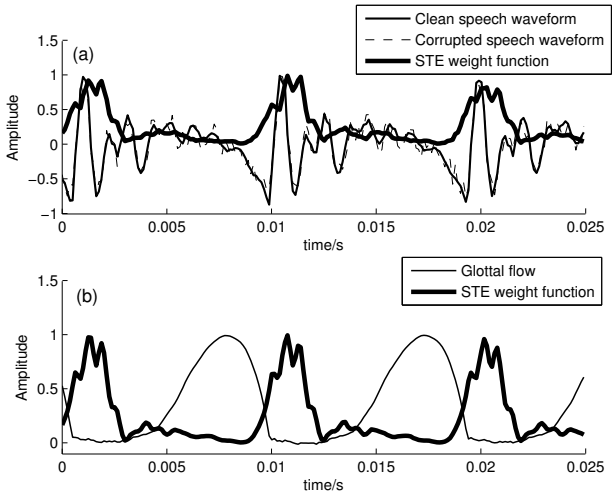


Fig. 1. Upper panel: Time-domain waveforms of clean speech (vowel /a/ produced by a male speaker), additive zero-mean Gaussian white noise corrupted speech (SNR=10dB), and short-time energy (STE) weight function ($M = 8$) computed from noisy speech according to Eq. 7. Lower panel: Glottal flow estimated from the clean vowel /a/ together with STE weight function ($M = 8$) computed also from the clean speech signal.

feature extraction of speech recognition (Wölfel et al., 2003; Dharanipragada et al., 2007; Wölfel and McDonough, 2005; Yapanel and Hansen, 2003).

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 (Ma et al., 1993) does not guarantee stability of the all-pole model, the idea of WLP is revisited by developing weight functions which always result in a stable all-pole model. It will be shown that with a proper choice of parameters the proposed stabilised WLP method yields spectral envelopes similar to those given by low order MVDR model but with improved robustness against additive background noise.

2 Weighted Linear Prediction

The discussion is begun by briefly presenting the optimisation of the filter parameters in WLP. Both in conventional LP and in WLP, sample x_n is estimated by a linear combination of the p past samples. This estimate can be formulated as

$$\hat{x}_n = -\sum_{i=1}^{p} a_i x_{n-i}, \tag{1}$$

where coefficients $a_i \in \mathbb{R}$, $\forall i = 1, \ldots, p$. The prediction error $\varepsilon_n(\mathbf{a})$, the residual, is defined as

$$\varepsilon_n(\mathbf{a}) = x_n - \hat{x}_n = x_n + \sum_{i=1}^{p} a_i x_{n-i} = \mathbf{a}^T \mathbf{x}_n, \tag{2}$$

where $\mathbf{a} = [a_0\ a_1\ \cdots\ a_p]^T$ with $a_0 = 1$ and $\mathbf{x}_n = [x_n\ \cdots\ x_{n-p}]^T$. The goal is to find the coefficient vector $\mathbf{a}$, of a p:th order FIR predictor, which minimises the cost function $\mathscr{E}(\mathbf{a})$, also known as the prediction error energy. This problem can be formulated as the constrained minimisation problem:

$$\begin{aligned} &\text{minimise } \mathscr{E}(\mathbf{a}) \\ &\text{subject to } \mathbf{a}^T\mathbf{u} = 1, \end{aligned} \tag{3}$$

where the unit vector $\mathbf{u}$ is defined as $\mathbf{u} = [1\ 0\ \cdots\ 0]^T$. This minimisation depends on the nature of the cost function $\mathscr{E}(\mathbf{a})$. The cost function in the WLP method is defined as

$$\mathscr{E}(\mathbf{a}) = \sum_{n=1}^{N+p} (\varepsilon_n(\mathbf{a}))^2 w_n. \tag{4}$$

In matrix notation, Eq. 4 can be written as

$$\mathscr{E}(\mathbf{a}) = \mathbf{a}^T \left(\sum_{n=1}^{N+p} w_n \mathbf{x}_n \mathbf{x}_n^T \right) \mathbf{a} = \mathbf{a}^T \mathbf{R} \mathbf{a}, \tag{5}$$

where $\mathbf{R} = \sum_{n=1}^{N+p} w_n \mathbf{x}_n \mathbf{x}_n^T$. Here the signal x_n is assumed to be zero outside the interval $[1, N]$, $\mathbf{R}$ corresponds to the autocorrelation matrix if and only if $\forall n = 1, \ldots, N+p,\ w_n = 1$. According to Eq. 4, the formulation allows us to temporally emphasise the square of the residual signal. It should be noticed that in difference to conventional LP the autocorrelation matrix $\mathbf{R}$ is *weighted*.

Matrix $\mathbf{R}$, defined in Eq. 5, is symmetric but does not posess the Toeplitz structure. However, it is positive definite, thus making the minimisation problem in Eq. 3 convex. Using the Lagrange multiplier minimisation method (Bazaraa et al., 1993), it can be shown (Bäckström, 2004) that $\mathbf{a}$, which solves the minimisation problem in Eq. 3, satisfies the linear equation

$$\mathbf{R}\mathbf{a} = \sigma^2 \mathbf{u}, \tag{6}$$

where $\sigma^2 = \mathbf{a}^T \mathbf{R} \mathbf{a}$ is the error energy. The corresponding WLP all-pole filter is obtained as $H(z) = 1/A(z)$, where $A(z)$ is the z-transform of $\mathbf{a}$.

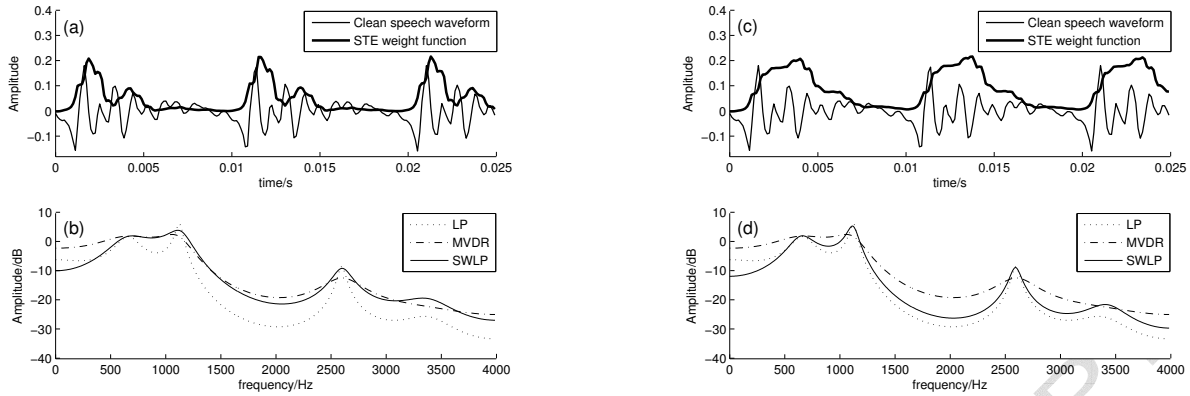


Fig. 2. Time-domain waveforms of clean speech (vowel /a/ produced by a male speaker) and short-time energy (STE) weight function (upper panels) and corresponding all-pole spectra of order $p = 10$ computed by LP, MVDR, and SWLP (lower panels). SWLP analysis was computed by using two different values for the length of the STE window: $M = 8$ (left panels) and $M = 24$ (right panels).

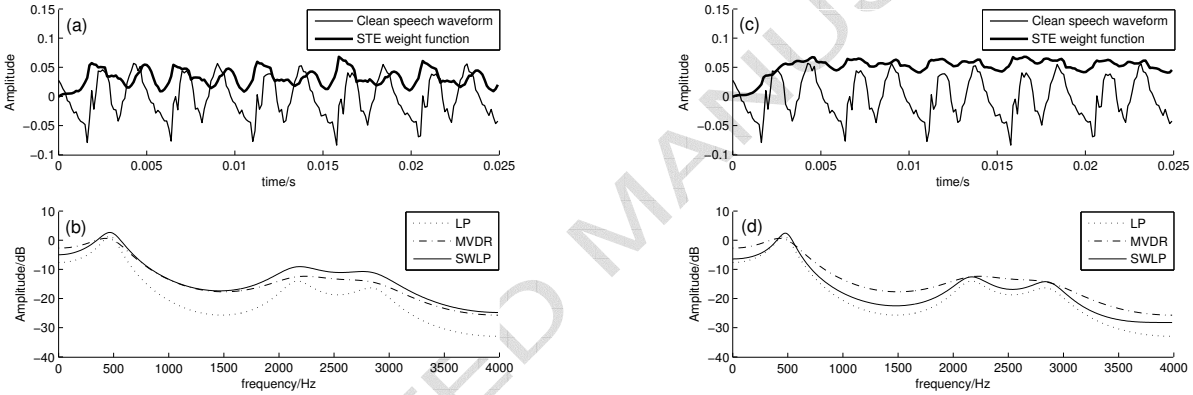


Fig. 3. Time-domain waveforms of clean speech (vowel /e/ produced by a female speaker) and short-time energy (STE) weight function (upper panels) and corresponding all-pole spectra of order $p = 10$ computed by LP, MVDR, and SWLP (lower panels). SWLP analysis was computed by using two different values for the length of the STE window: $M = 8$ (left panels) and $M = 24$ (right panels).

3 Model Formulation

The key concept of WLP, introduced in Eq. 4, is the time-domain weight function w_n. By choosing an appropriate waveform for w_n, one can either temporally emphasise or attenuate the square of the residual signal prior to the optimisation of the filter parameters. In (Ma et al., 1993) the weight function was chosen based on the short-time energy (STE):

$$w_n = \sum_{i=0}^{M-1} x_{n-i-1}^2, \tag{7}$$

where M is the length of the STE window. The performance of WLP was analysed in the original study by Ma et al. (1993) by using only clean speech represented by a small set of both synthetic and natural vowels. In the current study, however, the idea of weighting is motivated from the point of view of computing linear predictive models of speech that are more robust against noise than the conventional LP. From this perspective, the use of the STE window can be justified by two arguments. Firstly, as illustrated in Fig. 1(a), the STE function over-weights those sections of the speech waveform which consist of samples of large amplitude. It can be argued that these segments of speech are less vulnerable to additive, uniformly distributed noise in comparison to values of smaller amplitude. Hence, by emphasising the contribution of these strong data values in the computation of all-pole models one is expected to get spectral models which show better robustness in noisy conditions. Secondly, there is plenty of evidence in speech science indicating that formants extracted during the closed phase of a glottal cycle are more prominent than those computed during the glottal open phase due to the absence of sub-glottal coupling (Wong et al., 1979; Yegnanarayana and Veldhuis, 1998; Childers and Wong, 1994; Krishnamurthy and Childers, 1986). Hence, emphasis of the contribution of the samples occurring during the glottal closed phase is likely to yield more robust acoustical cues for the formants. Especially in the case of wideband noise, this kind of emphasising should improve modelling of higher formants in comparison to spectral models such as the conventional LP, which treat all data samples equally.

Figure 1(b) illustrates how the STE weight function focuses on the glottal closed phase. In this example, the STE function was computed from the clean /a/ vowel shown in the upper panel of Fig. 1. The glottal flow was estimated from the same clean vowel using the inverse filtering algorithm presented in (Alku, 1992). Even though WLP enables emphasising the contributions of samples occurring during the closed phase, it is worth noticing that the goal of the method *is not* to try to define the vocal tract filter precisely during the closed phase, as is the case in the so-called closed phase covariance method of glottal inverse filtering (Wong et al., 1979; Huiqun et al., 2006).

The stability of the WLP method with the STE weight function, as proposed in (Ma et al., 1993), however, can not be guaranteed. Therefore, a formula for a generalised weight function to be used in WLP is developed here so that the stability of the resulting all-pole filter is always guaranteed. The autocorrelation matrix from Eq. 5 can be expressed as

$$\mathbf{R} = \mathbf{Y}^T\mathbf{Y}, \tag{8}$$

where $\mathbf{Y} = [\mathbf{y}_0\ \mathbf{y}_1\ \cdots\ \mathbf{y}_p] \in \mathbb{R}^{(N+p)\times(p+1)}$ and $\mathbf{y}_0 = \left[\sqrt{w_1}x_1\ \cdots\ \sqrt{w_N}x_N\ 0\ \cdots\ 0\right]^T$. The columns $\mathbf{y}_k$ of the matrix $\mathbf{Y}$ can be generated via the formula

$$\mathbf{y}_{k+1} = \mathbf{B}\mathbf{y}_k \quad k = 0, 1, \ldots, p-1, \tag{9}$$

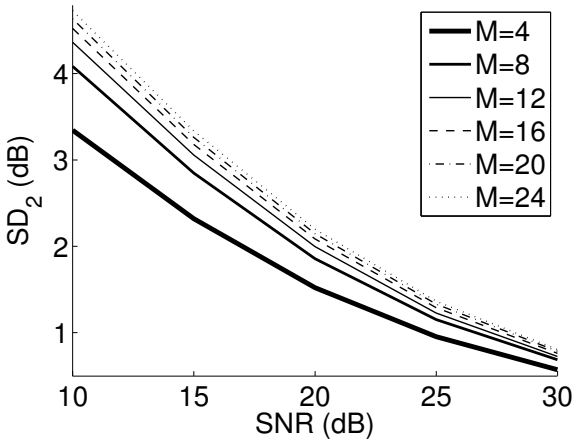


Fig. 4. Spectral distortion values (SD_2) between SWLP envelopes of order $p = 10$ computed from clean and noisy speech. The length of the STE window was varied in six steps from $M = 4$ to $M = 24$. Speech was corrupted by additive zero-mean Gaussian white noise in five segmental SNR categories. SD_2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

where

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \sqrt{w_2/w_1} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{w_3/w_2} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \sqrt{w_{N+p}/w_{N+p-1}} & 0 \end{bmatrix}. \tag{10}$$

The derivation of this stabilised WLP method can be expressed as follows. The weights are first calculated using Eq. 7 such that if $w_i = 0$ a small constant is added to the coefficient ($w_i = 10^{-6}$) and, before forming the matrix $\mathbf{Y}$ from Eq. 8, the elements of the secondary diagonal of the matrix $\mathbf{B}$ are defined (observe this difference in comparison to the original study by Ma et al. (1993)) for all $i = 1, \ldots, N + p - 1$ as

$$\mathbf{B}_{i+1,i} = \begin{cases} \sqrt{w_{i+1}/w_i}, & \text{if } w_i \leq w_{i+1}, \\ 1, & \text{if } w_i > w_{i+1}. \end{cases} \tag{11}$$

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 due to Eq. 11 (see Appendix).

4 Results

The behaviour of SWLP in spectral modelling of speech is demonstrated in the two examples shown in Figs. 2 and 3. In these figures, the analysed speech sounds (vowels /a/ and /e/ in Figs. 2 and 3, respectively) are shown together with the STE weight functions in the upper panels. The lower panels show spectra of parametric all-pole models of order $p = 10$ computed with three techniques: conventional LP with the autocorrelation criterion, minimum variance distortionless response, and the proposed SWLP. In order to demonstrate the effect of the weight function length, the SWLP analysis was computed using M values equal to 8 (left panels) and 24 (right panels). The examples depicted demonstrate two characteristic features of SWLP. First, the weight function computed by the STE clearly emphasises 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 low order ($p = 10$) 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. 10: 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.

The following result section is divided into three major parts. First, objective spectral distortion measurements were computed for LP, MVDR, and SWLP by using the spectral distortion criterion, SD_2. Next, small scale subjective tests were organised in order to obtain subjective evidence for the performance of low order MVDR and SWLP. It is well known that the SD_2 measure favours smooth spectra. Therefore, automatic speech recognition tests were conducted as the third experiment to get evidence on the performance of the different short-time spectral estimation methods in the 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 the sampling frequency of 8 kHz and the bandwidth of 4 kHz. The prediction order in all methods tested was set to $p = 10$, thereby fulfilling the known rule between the bandwidth and the prediction order (Markel and Gray, 1976). In addition, MVDR was also computed using a high model order ($p = 80$) which is a typical choice in studies in which MVDR has been used in automatic speech recognition (Wölfel and McDonough, 2005; Dharanipragada et al., 2007). Corrupted signals with desired segmental signal to noise ratios (SNR) were generated by adding noise to

clean speech sounds. Two types of noise were used: white zero mean Gaussian sequences produced by random number generator and factory noise recorded in realistic circumstances (Varga et al., 1992). Segmental SNR was computed as an average SNR over all 20ms frames in the speech signal (Kleijn and Paliwal, 1995).

4.1 Objective spectral distortion measurements

Objective evaluation of the effect of noise on all-pole modelling was computed by adapting the widely used spectral distortion criterion, SD_2 (Rabiner and Juang, 1993; Gray and Markel, 1979). With this measure, the difference between all-pole spectra computed from clean and noisy speech is computed as follows:

$$V(\omega) = \log_{10} P_1(\omega) - \log_{10} P_2(\omega), \tag{12}$$

where P_1 and P_2 denote power spectra of the all-pole filters computed from clean and noisy speech, respectively:

$$P_i(\omega) = \frac{\sigma_i^2}{|A_i(e^{j\omega})|^2} \qquad i = 1, 2. \tag{13}$$

In Eq. 13, the gains σ_i of the all-pole filters are adjusted so that the impulse response energies of the filters become equal. Since power spectra are computed using FFT, the discrete version of SD_2 must be used:

$$\mathrm{SD}_2 = \sqrt{\frac{1}{N_\mathrm{s}} \sum_{i=0}^{N_\mathrm{s}-1} |V(2\pi f_i)|^2}, \tag{14}$$

where N_s is the length of the discrete FFT spectra.

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. Speech data, taken from the TIMIT database (Garofolo, 1993), consisted of 12 American English sentences from four different dialect regions produced by six female and six male speakers. The frame length was 25 ms (200 samples) and no pre-emphasis was used. 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 segmental SNR categories by using SD_2. The experiments were conducted by 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. 4. 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 examples shown in Figs. 2 and

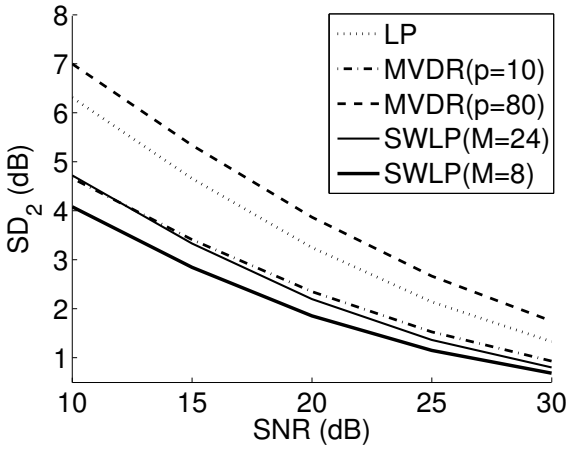


Fig. 5. Spectral distortion values (SD_2) between all-pole envelopes computed from clean and noisy speech with LP ($p = 10$), MVDR ($p = 10$ and $p = 80$) and SWLP ($p = 10$ with $M = 8$ and $M = 24$), where p is the model order and M is the length of the STE weight function. Speech was corrupted by additive zero-mean Gaussian white noise in five segmental SNR categories. SD_2 values were computed as an average over all the analysed segments consisting of 654 frames from the TIMIT database.

3, 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. 2(a), 2(c) and Figs. 3(a), 3(c)). Consequently, the weighting in the case of a small M value emphasises 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 models are computed by emphasising speech samples of larger local SNR over those with small local SNR. Hence, the resulting SWLP model computed with a small M value is less vulnerable to additive Gaussian noise. The results shown in Fig. 4 can also be understood from the point of view of the general shape of the SWLP filter (see Figs. 2(b), 2(d), 3(b), and 3(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. It is understandable that an all-pole filter which has a smooth spectral envelope is less sensitive to noise than a model with sharp resonances, which also explains why Fig. 4 shows the best performance for the lowest value of M.

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 by using two different values for this parameter: a large value of $M = 24$ corresponding to SWLP which behaves similarly to the conventional LP, and a small value $M = 8$, yielding SWLP filters

of smooth spectral shape similar to those computed by low order ($p = 10$) MVDR. The greatest M value used in previous experiments was 24, and hence it was selected to represent the SWLP with a large M value. The selection of the small M value was accomplished by running a special experiment in which the value of M that yielded the largest similarity between the all-pole spectra given by SWLP and MVDR ($p = 10$) was searched for. This was done by running an experiment where SD_2 was computed between the MVDR and SWLP all-pole envelopes by varying the STE window length M from 4 to 24. The SD_2 values were computed as an average over the entire (uncorrupted) training data consisting of 650 frames from TIMIT. The result of the experiment showed that the smallest spectral distortion value between SWLP and MVDR spectra was achieved with $M = 8$. Hence, all the further comparisons between SWLP and low order ($p = 10$) MVDR were computed by using the parameter value $M = 8$.

Performance of LP, MVDR, and SWLP was compared by measuring, for each method, how much the all-pole models computed from clean speech differ from those computed 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 done by adding zero-mean Gaussian noise to the clean utterances with five segmental SNR levels. Data consisted of 12 sentences, produced by 6 females and 6 males, taken from the TIMIT database. (These utterances were different from those used in the search of the M value yielding the largest similarity between SWLP and MVDR spectra). The total number of speech frames was 650. The SD_2 value for each method in each segmental SNR category was computed as an average over the SD_2 values obtained from individual frames.

The results obtained in comparing the robustness of the three all-pole modelling techniques are shown in Fig. 5. As a general trend, all methods show an increase in SD_2 when segmental SNR decreases. This over-all trend implies, naturally, that the spectral difference between the clean all-pole model and the one computed from noisy speech increases for all the methods analysed when the amount of noise is raised. 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 (Magi et al., 2006). The behaviour of SWLP, however, shows the best robustness against additive Gaussian noise. In particular, SWLP with a small M value is able to tackle the effect of additive Gaussian noise more effectively than any of the other methods tested.

4.2 Small scale subjective tests

Next, in order to get tentative subjective evidence for the performance of low order MVDR and SWLP in the modelling of both clean and noisy speech, a small listening test was organised. In this test, subjects ($n = 13$) listened to 200 ms

sounds synthesised 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. The utterances consisted of eight Finnish vowels produced by one male and one female subject. The test involved a perceptual comparison between three sounds (the reference sound, sound A and sound B). The reference was always the original, clean vowel. Sounds A and B were synthesised utterances produced, in random order, by impulse train excited MVDR and SWLP filters. In order to involve no pitch difference between the three sounds, the impulse train was always extracted from the reference signal. In addition, the loudness of the three sounds were normalised by adjusting the intensity levels of the sounds to be equal. The listener was asked to evaluate which one of the two alternatives (A or B) sounded more like the reference. In case the listener found that the quality difference between sound A and the reference was equal to that of sound B and the reference, she or he replied with *No preference*. The listener was allowed to listen to the three sounds as many times as she or he wished. The procedure was then repeated for all the vowels including both clean and noisy speech.

Table 1
Subjective evaluation between impulse train exited SWLP ($M = 8$) and MVDR filters of order $p = 10$. All-pole filters were computed from clean and noisy (SNR $= 10$ dB) male and female vowels.

	Male vowels		Female vowels	
Preferred method	clean	noisy	clean	noisy
MVDR	17%	1%	19%	5%
SWLP	71%	73%	46%	45%
No preference	12%	26%	35%	50%

The results, shown in Table 1, indicated that for clean male vowels, the listeners preferred the quality of the all-pole filters computed by SWLP over that given by MVDR: in 71% of all comparisons, they rated the vowels synthesised by SWLP to be closer in quality to the original speech, while only in 17% of the cases the listeners were in favour of MVDR. However, there were differences between the vowels: for /a/, /e/, /o/, /ä/, and /ö/, all the listeners preferred the sound synthesised by SWLP while for /i/ and /u/ SWLP was preferred only in approximately 10% of the cases. For these two vowels, both SWLP and MVDR failed to model the second formant properly. MVDR, however, modelled the over-all spectral envelope of the original vowel sound slightly better which might have explained the higher preference of MVDR. When listening to the sounds synthesised from noisy speech, the responses were even more in favour of SWLP: in 73% of all the cases, the vowels produced by SWLP filters were preferred, while those synthesised by MVDR filters were preferred in only 1% of the responses. For clean female vowels, the listeners preferred SWLP in 46% of the cases while MVDR was assessed better

in 19% of the comparisons. Again, when listening to the sounds synthesised from noisy speech, the listeners favoured the sounds synthesised by SWLP: in 45% of the cases, it was considered to yield quality closer to the original speech, while MVDR was preferred only in 5% of the cases.

4.3 Automatic speech recognition tests

As the third main part of the experiments, the performance of the proposed SWLP method was tested in feature extraction of a speech recogniser. In the field of automatic speech recognition (ASR), the mel-frequency cepstral coefficient (MFCC) representation is, by far, the most popular method of feature extraction. The stages of the MFCC computation for one speech frame can be outlined as follows (O'Shaughnessy, 2000): 1) estimation of the short-time magnitude spectrum; 2) computation of logarithmic mel-filterbank energies using triangular bandpass filters in the frequency domain; 3) discrete cosine transformation of the logarithmic filtered energies. In the first stage, simple FFT (periodogram) spectrum estimation is typically used; however, it is not the best spectrum estimation method in terms of robustness when the signal is corrupted by noise. Indeed, it has been argued that both LP and MVDR spectrum estimation, when substituted as the first stage of the MFCC computation, improve noise robustness of the features in certain cases (de Wet et al., 2001; Dharanipragada et al., 2007). This raises the question of whether SWLP could also offer improvement to the robustness of ASR systems.

The performance of six different spectrum estimation methods was evaluated in ASR: FFT, LP ($p = 10$), MVDR with $p = 10$ and $p = 80$, and SWLP ($p = 10$) with $M = 8$ and $M = 24$. This resulted in six slightly different 12-dimensional MFCC feature vectors, which were tested in isolated word recognition (IWR). The goal was to focus on the effect that the short-time spectrum in itself has on robustness. This means that the information given to the recogniser only involved the *shape* of the short-time spectrum. For this reason, neither the zeroth MFCC coefficient, which reflects frame energy, nor the inter-frame $\Delta/\Delta\Delta$-coefficients were included in the feature vector. It is well known that the inclusion of $\Delta/\Delta\Delta$-coefficients, which characterise the temporal changes of the spectrum, in the feature vector generally improves the performance of an ASR system (O'Shaughnessy, 2000). The $\Delta/\Delta\Delta$-coefficients are, however, based on short-time spectral estimation methods. Hence, it is reasonable to assume that whenever the spectrum estimation is distracted by noise, this will also have a negative effect on the obtained $\Delta/\Delta\Delta$-coefficients, resulting in lower recognition performance.

The use of IWR as the test problem can be justified by two reasons. First, state of the art continuous speech recognisers rely heavily on language models to improve their performance. Because language modelling compensates for shortcomings in the acoustic modelling, it may in an unpredictable fashion mask or distort the rel-

ative performance differences between the different features. Second, the acoustic modelling in both continuous and connected speech recognition benefits from long-time temporal structure. Instead, by focusing on IWR with vocabularies consisting of fairly short and common words, which might differ by just one phoneme, it can be argued that the feature evaluation concentrates more effectively on the importance of the correct identification of phonetic units based on the short-time spectrum.

The IWR system used in the present study is based on dynamic time warping (DTW) (O'Shaughnessy, 2000). DTW has been widely replaced by HMM methods in continuous speech recognition, the main focus of current ASR research. However, DTW is still well suited for IWR tasks and provides a good test bench for the present purpose of feature evaluation.

The idea of DTW is to compute a meaningful time-aligned distance between two templates, a test template $T(n)$ consisting of N_{T} feature vectors and a reference template $R(n)$ consisting of N_{R} feature vectors, by warping their time axes in order to synchronise acoustically similar segments in the templates. The time alignment is accomplished by finding the minimum-cost path through a grid of $N_{\mathrm{T}} \times N_{\mathrm{R}}$ nodes, where each node (i, j) corresponds to a pair of feature vectors $(T(i), R(j))$ and has an associated cost $d(T(i), R(j))$. In the current implementation, $d(T(i), R(j))$ was chosen to be the squared Euclidean distance between the two MFCC feature vectors $T(i)$ and $R(j)$. The optimised DTW distance was given by the sum of the node costs along the best path. The current system uses the so-called constrained endpoints version of DTW, where the path is required to start from grid node $(1, 1)$ and end at node $(N_{\mathrm{T}}, N_{\mathrm{R}})$ (O'Shaughnessy, 2000). The local continuity constraints of the present implementation dictate that along any permitted path any grid node (i, j) can be reached by one move only from one of the nodes $(i-1, j)$, $(i, j-1)$, or $(i-1, j-1)$. Exceptions naturally occur at grid boundaries where $i = 1$ or $j = 1$. In addition to these constraints, at most two consecutive moves from $(i, j-1)$ to (i, j) are permitted, except at the grid boundary where $i = N_{\mathrm{T}}$.

The training templates were clustered (Rabiner and Juang, 1993) using complete link agglomerative clustering (Theodoridis and Koutroumbas, 2003). This involves computing pairwise DTW distances between all training templates corresponding to the same vocabulary word. For each word in the vocabulary, ten clusters were generated and one reference template was chosen from each cluster. The representative template for each cluster was chosen as the one with the minimum average distance between it and every other template in the same cluster. During the recognition phase, each test template (test word) was recognised as follows. DTW distances were computed between the test template and each reference template (of which there are ten for each word in the vocabulary). For each vocabulary word, the average of the three smallest DTW distances was computed. The recognition decision was then made based on the smallest such averaged distance. Similar averaging is suggested in (Rabiner and Juang, 1993).

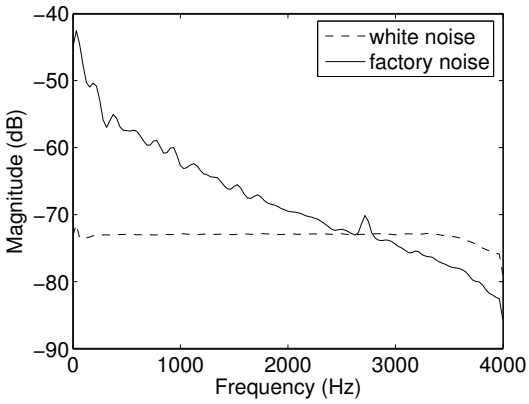


Fig. 6. The averaged power spectra of white noise and car factory noise from the Noisex-92 database (Varga et al., 1992), after re-sampling the signal to 8 kHz. The spectra were estimated using Welch's method with a 20 ms window.

The test material consisted of words extracted from continuous speech in the TIMIT database. The vocabulary of the recognition task was the 21 words in the two "SA" sentences spoken by every speaker in TIMIT. These sentences were "She had your dark suit in greasy wash water all year" and "Don't ask me to carry an oily rag like that". The training set consists of these words spoken by 136 randomly chosen male and 136 female speakers in the "train" subset of TIMIT (this number was chosen because it is the number of female speakers in the TIMIT "train" subset). The testing set has the words spoken by 50 randomly chosen male and 50 randomly chosen female speakers in the TIMIT "test" subset (which has completely different speakers from the "train" subset). Thus, the training and testing sets contained totals of 5712 and 2100 word tokens, respectively. A similar TIMIT-based corpus (albeit with slightly different sizes of the training and testing sets and non-balanced male-female speaker populations) was used in (Wu and Chan, 1993), where the best evaluated HMM-based recognisers using single-frame acoustic features achieved a word recognition performance of 91.0 %.

The speech material was down-sampled to 8 kHz for the evaluation. All features were computed using a frame length of 20 ms and a frame shift of 10 ms. No preemphasis was used. Noise corruption was done by adding prerecorded, down-sampled noise from the Noisex-92 database (Varga et al., 1992) to the test data with seven different segmental SNR levels. Two types of noise were used: white noise and factory noise recorded in a carproduction hall. The averaged power spectra of these two noise signals are shown in Fig. 6. It can be seen that the two noise types have very different characteristics, as the spectrum of the factory noise has a steep downward slope.

The correct recognition rates for the two noise types are shown in Tables. 2-3. For each noise type and segmental SNR level, the two best scores are shown in boldface. With clean speech, the two most conventional methods, FFT-MFCC and LP-MFCC, showed the best performance. The results for FFT-MFCC and LP-MFCC

are in agreement with a previous study, which found LP-based MFCC features to be more robust than their FFT-based counterparts in moderate noise conditions (de Wet et al., 2001). MVDR-MFCC with $p = 10$ slightly outperformed FFT-MFCC in white noise with some segmental SNR levels, while MVDR-MFCC with $p = 80$ showed, in general, modest improvement over FFT-MFCC in factory noise. Considering that the factory noise used here is of a low-pass type, like most other real-world noises used in other studies, the latter observation appears to be well in line with the findings reported in the literature, e.g. (Dharanipragada et al., 2007). The SWLP-MFCC features were superior to the other methods in white noise conditions, in particular when used with the parameter value $M = 8$. With factory noise, SWLP-MFCC became the best method when speech was severely corrupted by noise (that is SNR < 0 dB), and in these cases SWLP-MFCC was on an average 10 percentage units better than the baseline FFT-MFCC.

Table 2
Correct recognition rates (%) with white noise. Two best scores are shown in boldface

	Signal to noise ratio (dB)							
Feature vector	CLEAN	20	15	10	5	0	-5	-10
FFT-MFCC	**90.9**	86.5	78.3	61.7	42.1	24.6	13.3	**8.7**
LP-MFCC	**91.6**	**87.8**	80.0	65.9	49.9	**32.7**	**15.7**	7.2
MVDR-MFCC, p=10	89.5	84.8	75.8	60.3	44.2	28.0	13.1	6.9
MVDR-MFCC, p=80	89.7	85.2	76.6	61.7	45.0	25.8	12.3	6.2
SWLP-MFCC, M=8	88.7	86.7	**82.5**	**73.7**	**58.0**	**38.5**	**19.2**	**9.5**
SWLP-MFCC, M=24	90.3	**87.8**	**84.3**	**73.6**	**54.0**	32.0	15.4	7.2

Table 3
Correct recognition rates (%) with car factory noise. Two best scores are shown in boldface

	Signal to noise ratio (dB)							
Feature vector	CLEAN	20	15	10	5	0	-5	-10
FFT-MFCC	**90.9**	89.3	88.0	86.0	78.8	65.5	43.2	22.9
LP-MFCC	**91.6**	**91.2**	**90.5**	**88.4**	**83.3**	**69.7**	49.3	26.8
MVDR-MFCC, p=10	89.5	87.9	85.6	82.0	73.3	57.2	38.4	21.5
MVDR-MFCC, p=80	89.7	**89.8**	**88.1**	**86.4**	**81.1**	**68.1**	45.9	24.4
SWLP-MFCC, M=8	88.7	88.4	87.1	83.5	78.6	67.1	**50.9**	**30.4**
SWLP-MFCC, M=24	90.3	89.2	87.3	85.3	79.4	67.9	**51.9**	**34.8**

The results indicate that SWLP-based feature extraction outperformed the other techniques in recognition of speech corrupted by white noise already at segmental SNR value of 20 dB. In the case of factory noise, the major improvements achieved by SWLP occurred at clearly smaller segmental SNR values of -5 dB

and -10 dB. The difference in the performance of SWLP between the two noise types can be explained by the fact that in the case of white noise, upper frequencies of voiced speech are masked by noise already at reasonably high segmental SNR levels. This, in turn, implies that traditional spectral modelling techniques, such as LP, cannot model upper formants properly from speech corrupted by white noise. The proposed SWLP, however, emphasises the contribution of speech samples during the closed phase of the glottal cycle and thereby models formants during the time span inside the fundamental period when the resonances are more prominent (see Sec. 3). This implies that higher formants modelled by SWLP are less likely to be masked by additive noise as severely as those modelled by LP and, consequently, the acoustical cues embedded in them will be more effectively used in the feature extraction. The spectral envelope of factory noise, however, is of a low-pass nature and reminds that of voiced speech. Therefore, higher formants of speech corrupted by factory noise are not distorted severely until at the lowest segmental SNR categories below 0 dB. Hence, the improved recognition accuracy achieved by the proposed SWLP method takes place at the lowest values of the segmental SNR range in the case the additive noise is of low-pass nature.

5 Summary

LP was analysed in this study by using temporal weighting of the residual energy. The work is based on the previous study by Ma et al. (1993) where the concept of WLP 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. This new method, named stabilised weighed linear prediction, was then compared to two known all-pole modelling methods, conventional LP and minimum variance distortionless response, 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 SD_2. 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, SWLP was compared to other short-time spectral estimation methods in isolated word recognition experiments. It was shown to improve recognition accuracy already at moderate segmental SNR values for sounds corrupted by white noise. For realistic factory noise of low pass characteristics, the proposed method improved the recognition results at segmental SNR levels below 0 dB.

In difference to the original work by Ma et al. (1993), 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 WLP. 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 makes SWLP an attractive method for speech processing because it enables, with the same method, the computation of stable all-pole filters that yield spectral envelopes which are either smooth or of large dynamics. In particular, we believe that the proposed SWLP method when combined with a properly chosen value of M might become a potential technique in the development of new feature detection methods for recognition of noisy speech. This argument is justified by the increasing interest shown recently in the speech recognition community towards the MVDR technique, due to its promising performance in producing cepstral features for the recognition of noisy speech (Wölfel et al., 2003; Dharanipragada et al., 2007). The current study, however, shows evidence that MVDR is outperformed in robustness by the proposed SWLP in cases when the level of noise corruption is moderate to high. Hence, there are promising areas of future study in examining how the concept of WLP affects the recognition of noisy speech, when used in a state-of-the-art HMM-based continuous speech recognition framework.

A Stability of SWLP All-Pole Filter

In this section, a proof is presented for the minimum phase property of the SWLP inverse filter $A(z) = 1 + a_1 z^{-1} + \ldots + a_p z^{-p}$, where the coefficients a_i are solved from Eq. 6. The structure of the proof is similar to that given in (Delsarte et al., 1982), but for the sake of completeness a more detailed treatment is given in the following.

Rewrite Eq. 6 in the case when the autocorrelation matrix $\mathbf{R}$ is factorised as $\mathbf{R} = \mathbf{Y}^T\mathbf{Y}$ and $\mathbf{Y} = [\mathbf{y}_0\ \mathbf{y}_1\ \cdots\ \mathbf{y}_p] \in \mathbb{R}^{(N+p)\times(p+1)}$:

$$\begin{bmatrix} \mathbf{y}_0^T\mathbf{y}_0 & \mathbf{y}_0^T\mathbf{y}_1 & \cdots & \mathbf{y}_0^T\mathbf{y}_p \\ \mathbf{y}_1^T\mathbf{y}_0 & \mathbf{y}_1^T\mathbf{y}_1 & \cdots & \mathbf{y}_1^T\mathbf{y}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_p^T\mathbf{y}_0 & \mathbf{y}_p^T\mathbf{y}_1 & \cdots & \mathbf{y}_p^T\mathbf{y}_p \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{A.1}$$

In the SWLP model formulation, the columns $\mathbf{y}_i$ of matrix $\mathbf{Y}$ were generated via Eq. 9, using matrix $\mathbf{B}$ from Eq. 11. However, the column vectors $\mathbf{y}_i$ of matrix $\mathbf{Y}$ can be expressed by the following reverse equation

$$\mathbf{y}_k = \mathbf{M}\mathbf{y}_{k+1} \quad k = 0, 1, \ldots, p-1, \tag{A.2}$$

where

$$\mathbf{M} := \begin{bmatrix} 0 & 1/\mathbf{B}_{2,1} & 0 & \cdots & 0 \\ 0 & 0 & 1/\mathbf{B}_{3,2} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 1/\mathbf{B}_{N+p,N+p-1} \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \tag{A.3}$$

and $\mathbf{B}_{i+1,i}$ are the elements of matrix $\mathbf{B}$ from Eq. 11. Matrix $\mathbf{M}$, defined in Eq. A.3, is a nilpotent [3] operator with power of nilpotency $n = N + p$. Moreover, the norm of the Hilbert space for the matrix $\mathbf{M}$ is clearly equal to

$$\|\mathbf{M}\|_2 = \max_n\{1/\mathbf{B}_{n+1,n}\} = \max_n\{\sqrt{w_n/w_{n+1}}\}. \tag{A.4}$$

Note that, according to Eq. 11, $1 \leq \mathbf{B}_{n+1,n} < \infty,\ \forall n$ which implies that $\|\mathbf{M}\|_2 \leq 1$.

Defining the matrices $\mathbf{Y}_0 := [\mathbf{y}_0\ \mathbf{y}_1\ \cdots\ \mathbf{y}_{p-1}] \in \mathbb{R}^{(N+p)\times p}$ and $\mathbf{Y}_1 := [\mathbf{y}_1\ \mathbf{y}_2\ \cdots\ \mathbf{y}_p] \in \mathbb{R}^{(N+p)\times p}$ and corresponding subspaces $\mathscr{Y}_0 := \text{span}\{\mathbf{y}_0, \ldots, \mathbf{y}_{p-1}\} \subset \mathbb{C}^{N+p}$ and $\mathscr{Y}_1 := \text{span}\{\mathbf{y}_1, \ldots, \mathbf{y}_p\} \subset \mathbb{C}^{N+p}$ (where the base field is $\mathbb{C}$), respectively. Note that the reverse equation A.2 can be written in a more compact form

$$\mathbf{Y}_0 = \mathbf{M}\mathbf{Y}_1. \tag{A.5}$$

Next, define the symmetric linear projection operator $\mathbf{P} :\ \mathbb{C}^{N+p} \to \mathscr{Y}_1$ as

$$\mathbf{P} := \mathbf{Y}_1(\mathbf{Y}_1^T\mathbf{Y}_1)^{-1}\mathbf{Y}_1^T. \tag{A.6}$$

Thus, for all $\mathbf{v} \in \mathscr{Y}_1$ the projection operator has the property

$$\mathbf{P}\mathbf{v} = \mathbf{v}. \tag{A.7}$$

By rearranging Eq. A.1, the coefficients $\mathbf{a} = [a_1\ \cdots\ a_p]^T$ can be solved from the equation

$$\mathbf{a} = -(\mathbf{Y}_1^T\mathbf{Y}_1)^{-1}\mathbf{Y}_1^T\mathbf{y}_0. \tag{A.8}$$

From this equation yet another important property for the projection operator $\mathbf{P}$ is obtained:

$$\mathbf{P}\mathbf{y}_0 = \mathbf{Y}_1(\mathbf{Y}_1^T\mathbf{Y}_1)^{-1}\mathbf{Y}_1^T\mathbf{y}_0 = -\mathbf{Y}_1\mathbf{a}. \tag{A.9}$$

Lemma 1. *The zeros of the inverse filter $A(z)$ of the SWLP model are the nonzero eigenvalues of linear operator $\mathbf{PM} :\ \mathbb{C}^{N+p} \to \mathscr{Y}_1$.*

[3] Matrix $\mathbf{A}$ is nilpotent with power of nilpotency n if n is the smallest integer such that $\mathbf{A}^n = 0$.

Proof. Take the eigenpair $(\mathbf{v}, \lambda)$ of the linear operator $\mathbf{PM}$, where the eigenvector $\mathbf{v} \in \mathscr{Y}_1$ can be expressed as $\mathbf{v} = \mathbf{Y}_1\boldsymbol{\xi}$, where $\boldsymbol{\xi} = [\xi_1 \cdots \xi_p]^T \in \mathbb{C}^p$ is the coordinate vector with respect to the basis of space $\mathscr{Y}_1$. Using Eqs. A.5, A.7, A.9 gives

$$\begin{aligned}\lambda\mathbf{Y}_1\boldsymbol{\xi} &= \mathbf{PMY}_1\boldsymbol{\xi} = \mathbf{PY}_0\boldsymbol{\xi} \\ &= \begin{bmatrix}\mathbf{Py}_0 & \mathbf{Py}_1 & \cdots & \mathbf{Py}_{p-1}\end{bmatrix}\boldsymbol{\xi} \\ &= \begin{bmatrix}-\mathbf{Y}_1\mathbf{a} & \mathbf{y}_1 & \cdots & \mathbf{y}_{p-1}\end{bmatrix}\boldsymbol{\xi} \\ &= \mathbf{Y}_1\mathbf{C}\boldsymbol{\xi},\end{aligned} \tag{A.10}$$

where

$$\mathbf{C} = \begin{bmatrix}-a_1 & & & \\ -a_2 & & \mathbf{I}_{(p-1)\times(p-1)} & \\ \vdots & & & \\ -a_p & 0 & \cdots & 0\end{bmatrix} \tag{A.11}$$

is the companion matrix of the inverse filter $A(z)$, that is the zeros of $A(z)$ are the eigenvalues of $\mathbf{C}$. According to Eq. A.10

$$\begin{aligned}&\mathbf{Y}_1\mathbf{C}\boldsymbol{\xi} = \lambda\mathbf{Y}_1\boldsymbol{\xi} \\ &\mathbf{Y}_1(\mathbf{C}\boldsymbol{\xi} - \lambda\boldsymbol{\xi}) = 0 \\ &\mathbf{C}\boldsymbol{\xi} = \lambda\boldsymbol{\xi},\end{aligned} \tag{A.12}$$

where the last implication is due to the fact that
$\{\mathbf{x} \in \mathbb{C}^p \mid \mathbf{Y}_1\mathbf{x} = \mathbf{0}\} = \emptyset$. □

Theorem 1. *The zeros of the inverse filter $A(z)$ of the SWLP model are located inside a circle with centre at the origin and radius*

$$\rho = \max_n\{\sqrt{w_n/w_{n+1}}\}\cos\left(\frac{\pi}{N+p+1}\right).$$

Proof. Take a normalised eigenvector $\mathbf{v} \in \mathscr{Y}_1$ and the corresponding eigenvalue $\lambda \in \mathbb{C}$ of the linear operator $\mathbf{PM}$. Straightforward calculation gives

$$\begin{aligned}\lambda &= \lambda\|\mathbf{v}\|^2 = \mathbf{v}^T\lambda\mathbf{v} = \mathbf{v}^T\mathbf{PMv} \\ &= (\mathbf{Pv})^T\mathbf{Mv} = \mathbf{v}^T\mathbf{Mv} \in \mathcal{F}(\mathbf{M}).\end{aligned} \tag{A.13}$$

Hence, the zeros of the inverse filter $A(z)$ belong to the numerical range $\mathcal{F}(\mathbf{M})$ of nilpotent linear operator $\mathbf{M}$. It has been proved in (Karaev, 2004) that the numerical range of the nilpotent operator $\mathbf{M}$ with power of nilpotency $N + p$ is

a circle (open or closed) with centre at the origin and radius ρ not exceeding $\|\mathbf{M}\|_2 \cos(\frac{\pi}{N+p+1})$. Hence, according to Eq. A.4, the zeros of the inverse filter $A(z)$ of the SWLP model are located inside a circle with centre at the origin and with radius $\rho = \max_n\{\sqrt{w_n/w_{n+1}}\} \cos\left(\frac{\pi}{N+p+1}\right)$. Note that, in the SWLP method, $\max_n\{\sqrt{w_n/w_{n+1}}\} \leq 1$ according to Eq. 11, which guarantees the stability of the corresponding all-pole filter $1/A(z)$. □

References

Alku, P., June 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Communication 11 (2), 109–118.

Bazaraa, M., Sherali, H., Shetty, C., 1993. Nonlinear programming: Theory and algorithms, 2nd Edition. J. Wiley & Sons, Inc., New York.

Bäckström, T., 2004. Linear predictive modelling of speech – constraints and line spectrum pair decomposition. Ph.D. thesis, Helsinki University of Technology (TKK), Espoo, Finland, http://lib.tkk.fi/Diss/2004/isbn9512269473/.

Childers, D., Wong, C.-F., July 1994. Measuring and modeling vocal source-tract interaction. IEEE Transactions on Biomedical Engineering 41 (7), 663–671.

de Wet, F., Cranen, B., de Veth, J., Boves, L., 2001. A comparison of LPC and FFT-based acoustic features for noise robust ASR. In: Proceedings of Eurospeech 2001. Aalborg, Denmark.

Delsarte, P., Genin, Y., Kamp, Y., May 1982. Stability of linear predictors and numerical range of a linear operator. IEEE Transactions on Information Theory IT-33 (3), 412–415.

Dharanipragada, S., Yapanel, U., Rao, B., Jan 2007. 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.

El-Jaroudi, A., Makhoul, J., February 1991. Discrete all-pole modeling. IEEE Transactions on Signal Processing 39 (2), 411–423.

Garofolo, J., 1993. U.S. department of commerce. DARPA TIMIT Acoustic-phonetic Continuous Speech Corpus. Http://www.mpi.nl/world/tg/corpora/timit/timit.html.

Gray, A., Markel, J., October 1979. Distance measures for speech processing. IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-24 (5), 380–391.

Huiqun, D., Ward, R., Beddoes, M., Hodgson, M., March 2006. A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds. IEEE Transactions on Audio, Speech, and Language Processing 14 (2), 445–455.

Karaev, M., February 2004. The numerical range of a nilpotent operator on a Hilbert space. Proceedings of the American Mathematical Society 132 (8), 2321–2326.

Kleijn, W., Paliwal, K., 1995. Speech Coding and Synthesis. Elseviever Science B.V.

Krishnamurthy, A., Childers, D., August 1986. Two-channel speech analysis. IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-34 (4), 730–743.

Lim, J., Oppenheim, A., June 1978. All-pole modelling of degraded speech. IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-26 (3), 197–210.

Ma, C., Kamp, Y., Willems, L., March 1993. Robust signal selection for linear prediction analysis of voiced speech. Speech Communication 12 (1), 69–81.

Magi, C., Bäckström, T., Alku, P., 2006. Objective and subjective evaluation of seven selected all-pole modelling methods in processing of noise corrupted speech. In: CD Proc. of 7th Nordic Signal Processing Symposium NORSIG 2006, Reykjavik, Iceland, June 7-9.

Makhoul, J., April 1975. Linear prediction: A tutorial review. Proceedings of the IEEE 63 (4), 561–580.

Markel, J., Gray, A. H., 1976. Linear prediction of speech. Berlin : Springer.

Murthi, M., Rao, B., May 2000. All-pole modeling of speech based on the minimum variance distortionless response spectrum. IEEE Transactions on Speech and Audio Processing 8 (3), 221–239.

O'Shaughnessy, D., 2000. Speech Communications: Human and Machine, 2nd Edition. IEEE Press.

Rabiner, L., Juang, B.-H., 1993. Fundamentals of Speech Recognition. Prentice Hall PTR.

Sambur, M., Jayant, N., December 1976. 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.

Shimamura, T., 2004. Pitch synchronous addition and extension for linear predictive analysis of noisy speech. In: CD Proc. of 6th Nordic Signal Processing Symposium NORSIG 2004, Espoo, Finland, June 9-11.

Theodoridis, S., Koutroumbas, K., 2003. Pattern Recognition, 2nd Edition. Academic Press.

Varga, A., Steenneken, H., Tomlinson, M., Jones, D., 1992. Noisex-92 database. Http://spib.rice.edu/spib/select_noise.html.

Wong, D., Markel, J., Gray, A., August 1979. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-27 (4), 350–355.

Wu, J., Chan, C., November 1993. Isolated word recognition by neural network models with cross-correlation coefficients for speech dynamics. IEEE Trans. Pattern Analysis and Machine Intelligence 15, 1174–1185.

Wölfel, M., McDonough, J., September 2005. Minimum variance distortionless response spectral estimation. IEEE Signal Processing Magazine 22 (5), 117–126.

Wölfel, M., McDonough, J., Waibel, A., 2003. Warping and scaling of the minimum variance distortionless response. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop, 387-392. pp. 387–392.

Yapanel, U., Hansen, J., 2003. A new perspective on feature extraction for robust in-vehicle speech recognition. In: EUROSPEECH 2003, Geneva, Switzerland,

September 1-4.

Yegnanarayana, B., Veldhuis, R., July 1998. Extraction of vocal-tract system characteristics from speech signals. IEEE Transactions on Speech and Audio Processing 6 (4), 313–327.

Zhao, Q., Shimamura, T., Suzuk, J., 1997. Linear predictive analysis of noisy speech. In: Communications, Computers and Signal Processing. PACRIM'97, Victoria, Canada, August 20-22. Vol. 2. pp. 585–588.