Author(s): Jouni Pohjalainen, Heikki Kallasjoki, Paavo Alku, Kalle J. Palomaki and Mikko Kurimo

Title: Weighted linear prediction for speech analysis in noisy conditions

Year: 2009

Version: Final published version

Please cite the original version:

Jouni Pohjalainen, Heikki Kallasjoki, Paavo Alku, Kalle J. Palomaki and Mikko Kurimo. Weighted linear prediction for speech analysis in noisy conditions. In Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009), pages 1315-1318, Brighton, UK, September 2009.

Weighted Linear Prediction for Speech Analysis in Noisy Conditions

Jouni Pohjalainen[1]*, Heikki Kallasjoki*[2]*, Kalle J. Palomäki*[2]*, Mikko Kurimo*[2]*, Paavo Alku*[1]

[1]Dept. Signal Processing and Acoustics, Helsinki University of Technology, FI-02015 TKK, Finland
[2]Adaptive Informatics Research Centre, Helsinki University of Technology, FI-02015 TKK, Finland
jpohjala@acoustics.hut.fi

Abstract

Following earlier work, we modify linear predictive (LP) speech analysis by including temporal weighting of the squared prediction error in the model optimization. In order to focus this so called weighted LP model on the least noisy signal regions in the presence of stationary additive noise, we use short-time signal energy as the weighting function. We compare the noisy spectrum analysis performance of weighted LP and its recently proposed variant, the latter guaranteed to produce stable synthesis models. As a practical test case, we use automatic speech recognition to verify that the weighted LP methods improve upon the conventional FFT and LP methods by making spectrum estimates less prone to corruption by additive noise.

Index Terms: speech analysis, linear prediction, noise

1. Introduction

Most applications of speech and audio technology - whether coding, synthesis, recognition, enhancement or analysis - require some kind of model for the short-time magnitude or power spectrum of the signal. The importance of short-time spectrum models is reflected, for example, by the structure of the auditory system: a certain kind of spectrum analysis takes place on an early stage of auditory processing, the basilar membrane of the cochlea in the inner ear [11].

The standard spectrum analysis method for many applications is the discrete Fourier transform, implemented by FFT. Linear prediction (LP), commonly also referred to as linear predictive coding (LPC), is another prevalent method for modeling the short-time magnitude spectrum of the speech signal [7]. LP and its variants produce all-pole models which represent the magnitude spectrum envelope using only a few parameters.

Neither FFT nor LP have been designed to handle conditions of additive noise, in which another sound source is present in the recording environment or transmission channel. In this paper, we discuss two modifications of LP that have been found to show, in certain ways, more robust behavior in the presence of additive noise: weighted linear prediction (WLP) [5] and stabilized weighted linear prediction (SWLP) [6]. The former is a relatively straightforward generalization of LP that uses a temporal weighting function in order to focus on the less noisy regions of the signal. The latter is a modification of the former that guarantees the stability of the resulting all-pole model and is thus suitable for coding and synthesis applications. After introducing the methods and illustrating some of their properties, we apply them in feature extraction of large vocabulary continuous speech recognition. This provides both a realistic, challenging test problem for robustness, as well as a continuation of the recent applications of SWLP in automatic speech recognition [6] [4].

2. Spectrum Estimation Methods

2.1. Weighted Linear Prediction (WLP)

In linear predictive modeling, it is assumed that each speech sample can be predicted as a linear combination of p previous samples, i.e.,

$$\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k}, \tag{1}$$

where s_n is the digital speech signal, the a_k are the prediction coefficients and p is the prediction order. The difference between the actual speech sample s_n and its predicted value $\hat{s}_n$ is the residual $e_n = s_n - \sum_{k=1}^{p} a_k s_{n-k}$.

Weighted linear prediction (WLP) can be viewed as a generalization of LP. In contrast to conventional LP, WLP allows non-uniform weighting of the squared residual to emphasize some temporal regions and de-emphasize others in terms of modeling error energy. WLP minimizes the energy of the weighted squared residual [5]

$$E = \sum_n e_n^2 W_n = \sum_n (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2 W_n, \tag{2}$$

where W_n is the temporal weighting function. The range of summation of n, although not explicitly written in the formulas, is chosen in this work to correspond to the autocorrelation method of linear prediction [7]. According to the autocorrelation criterion, the signal s_n is considered to be zero outside the analysis interval. By setting the partial derivatives of E with respect to each a_k to zero, we arrive at the WLP normal equations

$$\sum_{k=1}^{p} a_k \sum_n W_n s_{n-k} s_{n-i} = \sum_n W_n s_n s_{n-i}, \qquad 1 \le i \le p, \tag{3}$$

which can be solved for the coefficients a_k to obtain the WLP all-pole model

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{4}$$

If W_n is a finite nonzero constant for all n, it becomes a multiplier of both sides of Eq. 3 and cancels out, leaving the LP normal equations [7]. The conventional autocorrelation LP method is guaranteed to produce a stable all-pole model. However, we know of no such guarantees to exist for autocorrelation WLP when the weighting function W_n is arbitrary [5] [6].

2.2. Stabilized Weighted Linear Prediction (SWLP)

Because of the importance of the model stability in coding and synthesis applications, a *stabilized* version of WLP, named stabilized weighted linear prediction (SWLP), was developed [6]. The WLP normal equations (Eq. 3) can alternatively be written in terms of partial weights $Z_{n,j}$ as

$$\sum_{k=1}^{p} a_k \sum_n Z_{n,k} s_{n-k} Z_{n,i} s_{n-i} = \sum_n Z_{n,0} s_n Z_{n,i} s_{n-i}, \quad (5)$$
$$1 \le i \le p,$$

where $Z_{n,j} = \sqrt{W_n}$ for $0 \le j \le p$.

As shown in [6], the stability of SWLP is guaranteed if the partial weights $Z_{n,j}$ are, instead, defined recursively as

$$Z_{n,0} = \sqrt{W_n} \quad (6)$$

and

$$Z_{n,j} = \max(1, \frac{\sqrt{W_n}}{\sqrt{W_{n-1}}}) Z_{n-1,j-1}, 1 \le j \le p. \quad (7)$$

Substitution of these values in Eq. 5 gives the SWLP normal equations. For a proof of stability of the SWLP all-pole model, or an alternative matrix-based representation of the method, the interested reader is referred to [6].

2.3. Choosing the Weight Function

The idea of temporal weighting in linear predictive analysis is to emphasize the contribution of less noisy speech samples in optimizing the filter coefficients. We follow Ma et al [5] in choosing short-time energy (STE) as the weight function. STE is computed using a sliding window of length M samples, as

$$W_n = \sum_{i=n-M+1-k}^{n-k} s_i^2, \quad (8)$$

where k is the offset with respect to the weighted sample index. In the current paper, we always choose $k = 1$.

STE weighting emphasizes 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 stationary additive noise than segments consisting of samples of smaller amplitude.

3. Examples

Figure 1 shows, for the vowel /u/ (spoken by a male speaker and sampled at 16 kHz), magnitude spectra as modeled by four different spectral estimation methods (FFT, LP, WLP and SWLP). The left panel shows the spectra for a clean sample and the right panel shows the spectra for the same sample artificially corrupted by noise recorded in a real factory (from NOISEX-92 database). The spectra have been computed from 20 ms Hamming-windowed frames using parameters $p = 20$ for the prediction order and $M = 16$ (corresponding to one millisecond) for the STE window length. The figure demonstrates a case in which WLP is able to separate the important two lowest formants F1 and F2 while conventional LP fails to do so, even on clean speech. WLP is able to separate the formants also in the noise-corrupted case. Regarding model robustness, it is noteworthy that the spectrum of SWLP is perhaps the one least affected by noise corruption, although SWLP does not separate F1 and F2.

Figure 2 shows a similar analysis for the vowel /a/, spoken by another male speaker. In this case, LP, WLP and SWLP all perform well in modeling the clean vowel. In the right panel, the same vowel is severely corrupted by noise. In the corrupted case, only WLP is able to clearly find F1 and F2. Again, although SWLP does not model formants very sharply, the spectrum of SWLP is arguably less affected by noise corruption than that of conventional LP.

The above observations motivate our experimental setup. WLP and SWLP appear to exhibit somewhat different aspects of robustness in speech spectrum modeling: WLP seems to indicate formants most prominently, while SWLP seems to be the one least affected by noise. FFT and LP are the conventional methods, widely used but vulnerable to additive noise. In the present study, we chose to investigate the performance of FFT, LP, WLP and SWLP in the feature extraction stage of automatic speech recognition (ASR), although there are potential applications for the temporally weighted linear predictive methods also in speech analysis (e.g., formant extraction). The speech recognizer was trained using clean speech and the recognition performance was tested in noisy conditions. The purpose was to gain evidence on the robustness of these spectrum modeling methods – especially from the recognition perspective.

4. Automatic Speech Recognition Tests

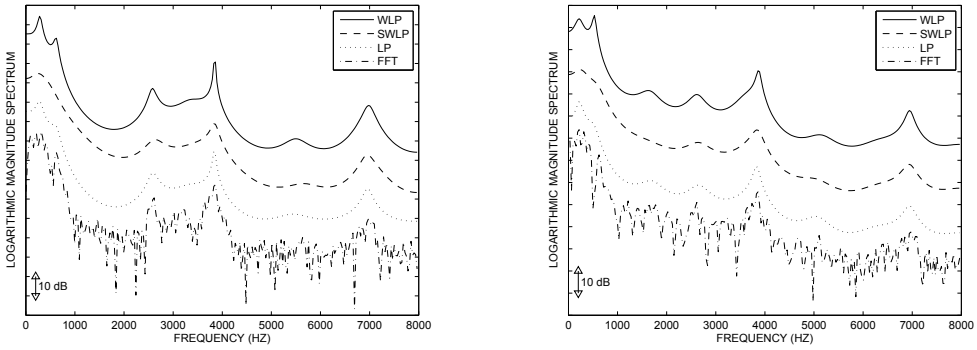
4.1. Feature Extraction

Sampling rate of all audio data was 16 kHz. After a pre-emphasis filter $H_{\rm pre}(z) = 1 - 0.97z^{-1}$, Hamming-windowed frames of 16 ms (256 samples) were generated, using a frame shift interval of 8 ms (125 frames per second).

The baseline method was a straightforward computation of the mel-frequency cepstral coefficients (MFCCs), which are widely used as the feature set for speech recognition [2]. A perceptually smoothed spectrum was computed from a FFT-based short-term magnitude spectrum estimate, using a filterbank of 23 logarithmically spaced triangular filters. Discrete cosine transformation was applied to the logarithm of the perceptual spectrum to get cepstral coefficients.

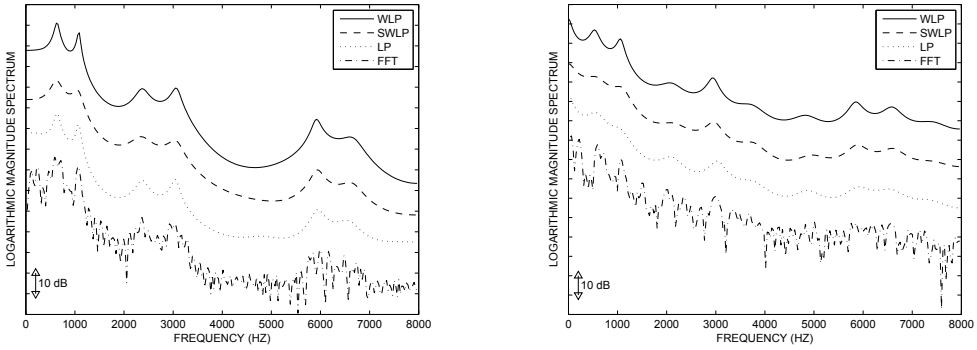
To evaluate the all-pole methods LP, WLP and SWLP, they were substituted (in place of FFT) as the spectrum estimation method in the MFCC computation explained above. To implement this for LP and SWLP, which are both guaranteed to be stable, the impulse response of the all-pole model was used as input to the MFCC computation. In the case of WLP, which is not guaranteed to yield a stable all-pole model, the magnitude spectrum of the WLP inverse filter ($1/H(z)$ from Eq. 4) was first computed via FFT. Next, any spectral sample that had a value smaller than 80 dB below the maximum of the spectrum was clipped to that limit. Next, the WLP synthesis filter spectrum was produced by inverting the modified spectrum. Finally, MFCC computation proceeded directly using this spectrum as input to the filterbank analysis.

For all three all-pole methods, prediction order $p = 20$ was chosen, following conventional guidelines for LP order selection [2]. STE window length $M = 16$ (1 ms) was used for WLP and SWLP, because this length of STE window has been found to give good results in earlier experiments [6]. 39 features, consisting of the 12 first cepstral coefficients, the logarithmic frame energy, and their first and second derivatives were used in the recognition experiments. Cepstral mean subtraction



(a) (b)

Figure 1: *Spectra of the vowel /u/. a) Clean sample. b) The same sample corrupted by factory noise.*



(a) (b)

Figure 2: *Spectra of the vowel /a/. a) Clean sample. b) The same sample corrupted by factory noise.*

(CMS) was done with a 150-frame window. Using the training data, the features were also normalized to have zero mean and unit variance, and a final maximum likelihood linear transformation (MLLT) step was applied.

4.2. Speech Material

The SPEECON [3] Finnish language corpus was used for the experiments. The training set for the recognizer contained approximately 21 hours of clean speech from 293 speakers. The evaluation of the feature extraction methods was done using speech from two different realistic noisy environments. Three audio channels, corresponding to three different microphones situated at various distances, were used to create tests sets with distinct noise levels.

The "Car" evaluation set consisted of 30 read phonetically rich sentences from 20 speakers, recorded in a moving car. Total length of the evaluation set was 57 minutes, including the leading and trailing silences. Audio channel 0 was recorded with a headset microphone, channel 1 with a lavalier microphone and channel 2 with a medium-distance microphone mounted at the car ceiling. Averages of the SNR estimates provided by the recording system for channels 0, 1 and 2 were 14 dB, 5 dB and 8 dB, respectively.

Utterances of the second evaluation set, labeled "Public places", were recorded both indoors and outdoors, and had background noise of various kinds, such as other speech, footsteps, etc. The set contained 30 read sentences from 30 speakers, with a total length of 94 minutes. Channel 0 and 1 microphones were identical to the "Car" set, while channel 2 was recorded with a different medium-distance microphone placed 0.5-1 meter away from the speaker. In this set, channels 0, 1 and 2 had average SNR estimates of 24 dB, 14 dB and 9 dB, respectively.

4.3. Experiment Setup

Our large vocabulary continuous speech recognizer was used for the experiments. The recognizer uses an n-gram language model trained on a Finnish language data set of approximately

145 million words of book and newspaper data, using a method for growing an n-gram model [10]. The language model uses statistical morphs, learned from the text data with an unsupervised method [1], as language modeling units. The decoder of the recognizer is based on a one-pass time-synchronous Viterbi beam search algorithm [9]. The acoustic modeling employs cross-word triphones modeled with state-clustered hidden Markov models. The model states use a mixture of (on average) 16 Gaussians to model the speech features, and an additional Gamma probability distribution for state duration modeling [8].

The primary performance measure used was the letter error rate. The more common word error rate is shown in the result tables for completeness, but it is not as well suited to Finnish, because single Finnish words are often concatenations of several morphemes, and therefore correspond to more than one word in English. As an example, a word like "kahvin+juoja+lle+kin" translates to "also for a coffee drinker."

4.4. Recognition Results

The recognition results for the "Car" and "Public places" test sets are shown in Tables 1 and 2, respectively. While WLP was slightly worse than the other methods with clean speech, it clearly outperformed them on all noisy channels, with both types of noise. SWLP improved the recognition in select noisy scenarios, as in [6]. However, slightly better SWLP results – yet not outperforming our WLP results – have been achieved by careful tuning of the M parameter for each different training and recognition scenario [4].

Table 3 shows relative improvements of the linear predictive techniques over the baseline FFT-based MFCC in the letter error rates for the different recording channels of a combined "Car" and "Public places" data set. The relative improvements can be seen to become more marked as the analyzed speech becomes more affected by noise. With combined "Car" and "Public places" test set, WLP yielded statistically significant improvement over all the other methods in channels 1 and 2.

Table 1: *Results for the "Car" test set in terms of letter error rate (word error rate in parentheses).*

Ch.	FFT	LP	WLP	SWLP
0	4.0 (14.2)	3.9 (14.4)	4.7 (16.3)	4.2 (15.2)
1	29.6 (51.9)	27.2 (49.7)	23.1 (45.5)	32.3 (54.4)
2	68.6 (84.7)	55.0 (78.9)	51.2 (77.7)	55.9 (77.3)

Table 2: *Results for the "Public places" test set in terms of letter error rate (word error rate in parentheses).*

Ch.	FFT	LP	WLP	SWLP
0	3.3 (13.6)	3.4 (14.2)	4.7 (18.3)	3.6 (14.4)
1	23.4 (41.8)	20.8 (40.4)	20.0 (43.3)	20.4 (41.2)
2	40.8 (56.5)	34.9 (53.2)	31.9 (56.0)	34.3 (53.2)

5. Conclusions

We discussed LP-based all-pole methods with main focus on the temporal weighting of the squared residual. The results on ASR tests support the conclusion that temporal weighting in WLP, implemented with the STE function, leads to spectrum models that are less prone to noise corruption than the spectra given by the conventional methods FFT and LP.

Table 3: *Relative letter error rate improvement (%) of the linear predictive models with respect to the baseline MFCC system using combined "Car"+"Public places" test set.*

Ch.	LP	WLP	SWLP
0	-1.7	-32.0	-7.3
1	9.8	18.1	3.2
2	17.3	23.7	17.4

Applications that require stable all-pole models should use either autocorrelation LP or SWLP; our ASR evaluation showed no clear preference for either method. The WLP method, which was observed to model formants sharply in both clean and noisy conditions, but is not guaranteed to give stable synthesis models, turned out to provide the best overall performance in terms of ASR noise robustness. One interesting direction for future work is the search of new weighting functions, other than the simple STE weighting, that would lead to further performance improvements in ASR and other applications.

6. Acknowledgements

This work was supported by the Academy of Finland within projects 127345 (J. Pohjalainen) and 114369 (H. Kallasjoki and K. Palomäki).

7. References

[1] Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S. and Pylkkönen, J., "Unlimited vocabulary speech recognition with morph language models applied to Finnish", Computer Speech & Language, 20(4):515–541, 2006.

[2] Huang, X., Acero, A. and Hon, H.-W., "Spoken Language Processing", Prentice Hall PTR, 2001.

[3] Iskra, D., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F. and Kiessling, A., "SPEECON - speech databases for consumer devices: Database specification and validation", Proc. LREC, pp. 329–333, 2002.

[4] Kallasjoki, H., Palomäki, K., Magi, C., Alku, P. and Kurimo, M., "Noise robust LVCSR feature extraction based on stabilized weighted linear prediction", Proc. SPECOM, 2009.

[5] Ma, C., Kamp, Y. and Willems, L. F., "Robust signal selection for linear prediction analysis of voiced speech", Speech Communication, 12(2):69–81, 1993.

[6] Magi, C., Pohjalainen, J., Bäckström, T. and Alku, P.,"Stabilised weighted linear prediction", Speech Communication, 51(5):401–411, 2009.

[7] Makhoul, J., "Linear prediction: a tutorial review", Proceedings of the IEEE, 63(4):561–580, 1975.

[8] Pylkkönen, J. and Kurimo, M., "Duration Modeling Techniques for Continuous Speech Recognition", Proc. INTERSPEECH, pp.385–388, 2004.

[9] Pylkkönen, J., "An efficient one-pass decoder for Finnish large vocabulary continuous speech recognition", Proc. 2nd Baltic Conference on Human Language Technologies (HLT'2005), pp. 167–172, 2005.

[10] Siivola, V. and Pellom, B., "Growing an n-gram language model", Proc. INTERSPEECH, pp. 1309–1312, 2005.

[11] Zwicker, E. and Fastl, H., "Psychoacoustics, Facts and Models", Springer-Verlag, 1990.