# SPECTRAL ENVELOPE ESTIMATION AND REGULARIZATION

*L. Anders Ekman, W. Bastiaan Kleijn* *

Dept. Signals, Sensors and Systems
KTH (Royal Institute of Technology), Sweden

*Manohar N. Murthi* †

Dept. Electrical and Computer Engineering
University of Miami, USA

## ABSTRACT

A well-known problem with linear prediction is that its estimate of the spectral envelope often has sharp peaks for high-pitch speakers. These peaks are anomalies resulting from contamination of the spectral envelope by the spectral fine structure. We investigate the method of regularized linear prediction to find a better estimate of the spectral envelope and compare the method to the commonly used approach of bandwidth expansion. We present simulations over voiced frames of female speakers from the TIMIT database, where the envelope modeling accuracy is measured using a log spectral distortion measure. We also investigate the coding properties of the methods. The results indicate that the new regularized LP method is superior to bandwidth expansion, with an insignificant increase in computational complexity.

## 1. INTRODUCTION

Linear prediction (LP) is commonly used to estimate the parameters of an autoregressive (AR) model describing the spectral envelope. The LP estimation procedure is suited for estimating the spectral envelope since it emphasizes local spectral peaks but is insensitive to global spectral variations [1]. However, it is not clear that the criterion that is minimized in the LP estimation method is optimal for speech processing. In particular, a commonly observed problem is the contamination of the spectral envelope by spectral fine-structure [2].

Over the years, numerous attempts have been made to improve all-pole spectral envelope estimation for speech processing applications. Examples of such attempts include the use of alternate spectral estimation techniques [2], methods to improve the spectral fit for periodic excitation signals, e.g., [3, 4], methods that account for the amplitude of the excitation, e.g., [5], and methods that perform a frequency warp on the AR model, e.g., [6]. However, simple methods to prevent sharp spectral peaks, such as the lag window method [7] and particularly bandwidth expansion [8] have had the most significant impact on speech processing.

In bandwidth expansion, the radii of the poles of the AR model are scaled by a factor $\gamma < 1$, resulting in a bandwidth expansion

$$\Delta B \approx -\frac{ln(\gamma)}{\pi T}, \tag{1}$$

where $T$ is the sampling interval. The method forms an ad-hoc post-processing stage after the LP estimation procedure. While it works well, it cannot be claimed to be optimal in any sense.

Using a more formal approach, Cappe and Moulines [9] used a modified criterion to obtain a broadening of the formant bandwidth for a cepstral representation. In later work [10], it was shown that such a *regularization* approach can also be used to reduce the peaky behavior obtained by the LP estimation approach.

The main disadvantage of the method of [10] was that it was computationally expensive, namely the computation of the penalty term requires an iterative procedure. In [11], Murthi and Kleijn eliminated this problem by modifying the penalty term. Following the work of [11], we minimize a composite cost function of the prediction error variance and a penalty measure, constructed to penalize non-smooth behavior of the spectral envelope. By varying the contribution of the penalty term, different degrees of regularization are attained.

The ability of a particular LP-based method to estimate the spectral envelope is particularly well illustrated for high AR model order. The usage of conventional LP estimation for a high order of the AR model results in a modeling of the fine structure of the speech frame. That is the harmonics of the power spectrum are clearly visible [2, 12]. Conventional LP estimation avoids modeling the spectral fine-structure by having a low prediction order. In this paper, we show that regularized linear prediction solves the problem of spectral-envelope estimation at a fundamental level, independently of prediction order. The regularization technique gives an extra degree of freedom in the spectral modeling, compared to bandwidth expanded LP. The present paper shows that the formal regularization method can provide a better estimate of the spectral envelope than bandwidth expansion at an insignificant additional computational cost.

## 2. REGULARIZED LINEAR PREDICTION

For an AR model $1/A(z)$ of order $M$ with $A(z) = a_0 + a_1 z^{-1} + \ldots + a_M z^{-M}$, the spectral envelope is defined by

$$S(\omega, \mathbf{a}) = \frac{1}{|A(e^{j\omega})|^2}, \tag{2}$$

in which $\mathbf{a} = [a_1, a_2, \ldots, a_M]$. The concept of regularized LP [10, 11] is to minimize the cost function

$$\mathcal{D}(S(\omega, \mathbf{a}), S(\omega)) + \lambda \mathcal{R}(S(\omega, \mathbf{a})), \tag{3}$$

where $\mathcal{D}$ is a cost function between the spectral envelope of the model $S(\omega, \mathbf{a})$ and the original speech power spectrum $S(\omega)$. $\mathcal{R}(S(\omega, \mathbf{a}))$ is the penalty measure, which increases for rapid changes in the contour of the spectral envelope. The factor $\lambda$ controls the trade-off between the fit of the speech spectrum and the smoothness measure.

In [11], the authors show that the penalty measure

$$\mathcal{R}(S(\omega, \mathbf{a})) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ \frac{d}{d\omega} \log |S(\omega, \mathbf{a})| \right]^2 d\omega \tag{4}$$

*This work was supported by Ericsson Research & Development.

†The work of M.N. Murthi was supported in part by the National Science Foundation via CAREER Award CCF-0347229

can be approximated by

$$\hat{\mathcal{R}}(S(\omega, \mathbf{a})) = \mathbf{a}^T \mathbf{DFDa}, \tag{5}$$

where $\mathbf{D}$ is a diagonal matrix in which each diagonal element consists of the row number, and $\mathbf{F}$ is a Toeplitz autocovariance matrix for a suitably windowed autocorrelation sequence of a speech frame. The similarity between (5) and the cost function of conventional LP

$$\mathcal{D}(S(\omega, \mathbf{a}), S(\omega)) = \mathbf{a}^T \mathbf{Ra} + 2\mathbf{a}^T \mathbf{r} + \mathbf{r}^T \mathbf{r}, \tag{6}$$

leads to a composite cost function that has a simple form and yields the compact solution

$$\mathbf{a}_{opt} = -(\mathbf{R} + \lambda \mathbf{DFD})^{-1} \mathbf{r}, \tag{7}$$

where $\mathbf{R}$ is the autocovariance matrix of the speech frame and $\mathbf{r}$ is the vector of speech autocovariance values $[r(1), r(2), \ldots, r(M)]^T$. The matrix $(\mathbf{R} + \lambda \mathbf{DFD})$ is Hermitian and positive definite, so efficient algorithms for solving (7) exist, e.g., Cholesky decomposition. The optimal AR filter coefficients can thus be calculated with an insignificant increase in computational complexity over conventional LP.

## 3. DISTORTION MEASURE

An often used distortion criterion is the log spectral distortion at the harmonics between the power spectrum of the speech frame and the modeled spectral envelope (e.g., [3], [13]). In our case, this measure is less suitable because it does not indicate whether the AR model of the envelope is contaminated by the spectral fine structure (the harmonics). Instead, we evaluate the envelope modeling performance as the log spectral distortion between the envelope obtained by linear interpolation between the harmonic peaks of the logarithmic periodogram, and the spectral envelopes of the AR models:

$$LSD = \sqrt{\frac{1}{\pi} \int_0^\pi \left[10 \log_{10} S_{lin}(\omega) - 10 \log_{10} S(\omega, \mathbf{a})\right]^2 d\omega}. \tag{8}$$

$S_{lin}(\omega)$ is the interpolated periodogram, which can be seen as a coarse estimate of the spectral envelope. If we ignore the spectral tilt of the excitation signal, this envelope can be seen as a reasonable vocal tract transfer function.

Figure 1 illustrates the approach. The periodogram of a voiced frame is shown, together with the harmonics and the linear interpolation between them. Also shown are the spectral envelopes of bandwidth expanded LP with $\gamma = 0.985$, and regularized LP with $\lambda = 0.0050$. The values of the constants are chosen to minimize the criterion in (8). The bandwidth expanded LP envelope follows several of the harmonic peaks, whereas the regularized LP gives a smoother, more plausible, envelope. The poles of the bandwidth expansion approach are restricted to having the same angles as those of conventional LP, and therefore the bandwidth expanded LP envelope has a shape similar to that of the conventional LP envelope (not shown in the figure).

## 4. FINDING THE OPTIMAL $\lambda$

Methods for regularized linear prediction can be divided into two classes: those using a constant $\lambda$ and those using an adaptive $\lambda$. In the first class, $\lambda$ is kept constant through all frames. In the second class, $\lambda$ is allowed to vary in each frame. In the simulations we have considered candidates from both classes, and they are listed in table 1, along with the abbreviations used in the upcoming simulation plots.

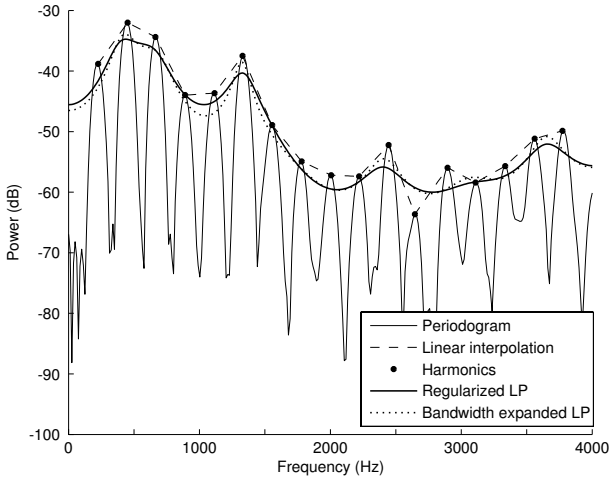


**Fig. 1**. Periodogram of a frame of voiced female speech (solid), using a 20 ms Hamming window. The detected harmonics are marked $(\cdot)$, and the linear interpolation $S_{lin}(\omega)$ is dashed. The bandwidth expanded LP envelope for $\gamma = 0.985$ is shown dotted, and the regularized LP envelope for $\lambda = 0.0050$ is shown as solid. The AR model order is 16.

### 4.1. Experimental setup

Voiced frames were extracted from female speakers of the TIMIT database. We used 9,483 frames as training data, and another 4,001 frames as validation data. The test and validation data sets were from different speakers. Only female speech was considered due to the fact that the pitch frequency generally is higher for female speakers, and it is at high pitch frequencies that the problem of narrow bandwidth envelopes occurs in its most significance. The speech was sampled at 8 kHz. Each frame consisted of 20 ms of speech samples, windowed with a Hamming window of the same length before processing with the different linear prediction methods. The AR model order was chosen to be 16, to clearly capture the harmonic fit problem that LP suffers from. For a 10'th order AR model the results have the same trend, although less pronounced.

### 4.2. Training

For each frame of the training database, we calculated the AR model coefficients using conventional LP, along with regularized LP for a range of different $\lambda$ values, as well as the coefficients of the bandwidth expansion approach for a range of different $\gamma$ values. We estimated the pitch of the current frame and found the pitch harmonics in the power spectrum of the speech samples, and calculated the linear interpolation of the power spectrum, $S_{lin}(\omega)$. The log spectral distortions were calculated between the interpolated power spectrum $S_{lin}(\omega)$ and the conventional LP, regularized LP and bandwidth expanded LP, respectively, using (8). In the training procedure, we could then choose the $\lambda$ and $\gamma$ according to the criteria in table 1. $\mathbf{R_{opt}}$ and $\mathbf{BE_{opt}}$ are included to determine the optimal performance of the methods of regularized LP and bandwidth expansion, respectively, and cannot be used in practical situations.

| | | |
|---|---|---|
| Constant $\lambda, \gamma$ | **LP** | Conventional Linear Prediction is used as a reference method. |
| | **RC** | Regularized LP with the constant $\lambda$ that yields the lowest distortion. |
| | **BEC** | Bandwidth expansion with the constant $\gamma$ that yields the lowest distortion. |
| Adaptive $\lambda, \gamma$ | **MOD** | Regularized LP with $\lambda$ chosen as a function of the pitch frequency. |
| | $\mathbf{R_{opt}}$ | Regularized LP with $\lambda$ changing on a frame-by-frame basis, picked as the $\lambda$ that gives the smallest distortion possible for each particular frame. |
| | $\mathbf{BE_{opt}}$ | Bandwidth expansion with $\gamma$ changing on a frame-by-frame basis, picked as the $\gamma$ that gives the smallest distortion possible for each particular frame. |

**Table 1**. The different approaches of selecting $\lambda$ and $\gamma$, and their abbreviations.

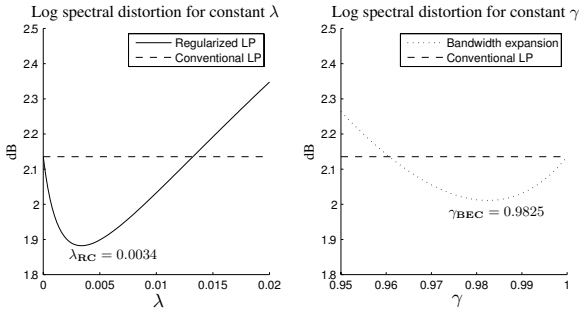


**Fig. 2**. The mean log spectral distortion over the training database of 9,483 frames for conventional LP (dashed), bandwidth expansion (dotted) and regularized LP (solid). The best constant values of $\lambda$ and $\gamma$ are depicted in the figure.

#### *4.2.1. Constant $\lambda$*

**RC** and **BEC** were compared to conventional LP in figure 2, for a set of different values of $\lambda \in [0, 0.02]$ and $\gamma \in [0.95, 1]$. For the best choice in $\lambda$ ($\lambda_{\mathbf{RC}} = 0.0034$) the regularized LP performed better, with an average improvement (over all the frames) of 0.25 dB over conventional LP, whereas bandwidth expansion performed 0.12 dB better than conventional LP. The best constant $\gamma$ was in this case 0.9825, which corresponds to 45 Hz of bandwidth expansion, as given by (1).

#### *4.2.2. Adaptive $\lambda$*

The problems with linear prediction envelopes exhibiting peaks that are too sharp is mainly noticeable for female speakers, due to the high pitch frequency. Our simulations confirm that at a low pitch frequency, the optimal $\lambda$ is often zero, reducing the method to conventional LP, whereas a larger $\lambda$ is optimal when the pitch frequency is higher. We therefore propose a simple model (**MOD**), $\lambda = \max(0, a \cdot pitch + b)$, where the constants $a$ and $b$ are determined from the training data set with least squares fitting. This method could be integrated into a system, if a pitch estimator is present. As can be seen in figure 3, the model is slightly better than

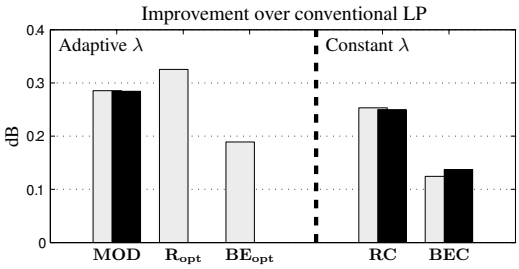


**Fig. 3**. Improvement of log spectral distortion over conventional LP for the investigated methods, over a database of 9,483 female voiced frames of speech. The results of the gray bars are evaluated for the same data as was used in training. The black bars show the result from the validation data set, with 4,001 frames.

using a constant $\lambda$ (**RC**). It is 0.29 dB better than conventional LP.

$\mathbf{R_{opt}}$ is the best performance we can get for the regularized LP with this distortion measure, since the optimal $\lambda$ is chosen for each frame. It corresponds to a log spectral distortion that is 0.33 dB lower than conventional LP.

$\mathbf{BE_{opt}}$ is the corresponding method for bandwidth expansion. The best $\gamma$ (in the log spectral distortion sense) is chosen for each frame. The result is 0.19 dB better than conventional LP. The results from training and validation are summarized in figure 3.

### 4.3. Validation

In the validation session, the first four entries of table 1 are investigated. The best choice of constant $\lambda$, constant $\gamma$ and the model parameters $[a, b]$ are carried over from the training database:

$$\lambda_{\mathbf{RC}} = 0.0034, \quad \gamma_{\mathbf{BEC}} = 0.9825,$$
$$a = 8.9 \times 10^{-5}, \quad b = -1.4 \times 10^{-2}. \tag{9}$$

The results for constant $\lambda$ (**RC**) and $\gamma$ (**BEC**) are similar to those of previous section, and can be found in figure 3. The model of $\lambda$ (**MOD**) receives a slightly higher score than the regularization with constant $\lambda$ (**RC**), which is consistent with the previous result. Figure 4 confirms that the regularization and bandwidth expansion methods are important only at high pitch frequencies (at around 175Hz and above).

### 4.4. Other approaches of selecting $\lambda$

The regularization scheme leads to an increase in the prediction error variance. In regularization techniques, a way to display information about the regularization solution is to plot the solution (or in our case, the penalty measure) versus the norm of the residual vector. Methods for this kind of analysis, e.g., by means of the L-curve [14] to find a "sweet spot" between the increase in prediction error variance and the decrease of the penalty measure in (5) were investigated. We have also investigated if there is an optimal relative increase of the prediction error variance that leads to the best choice in $\lambda$. The results of these tests show that the methods are not satisfactory, which is why they are omitted from the presented results.

## 5. CODING COMPLEXITY

Let us consider the case where regularized linear prediction is to be used in a speech coder. We study the case of high rate entropy coding of the line spectral frequencies (LSFs). We use a 10'th order

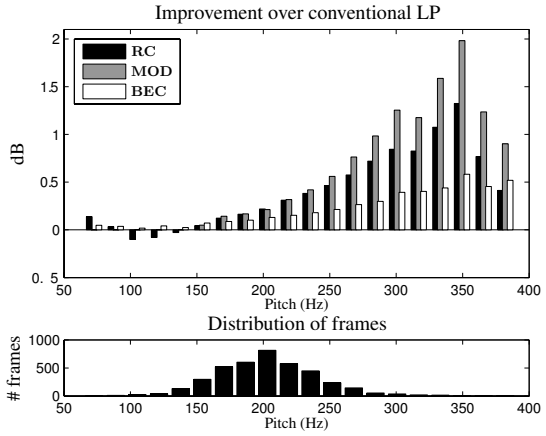


**Fig. 4**. The improvement in log spectral distortion over conventional LP at different pitch frequencies, for the investigated methods. The lower plot shows the distribution of the frames over pitch frequency. The results are from the validation data base.

| | LP | BEC | RC | MOD |
|---|---|---|---|---|
| Entropy | 39.3 | 37.5 | 36.5 | 35.9 |
| Improvement | | 1.8 | 2.8 | 3.4 |

**Table 2**. The entropy for the quantization indices for the different LP methods. All numbers are in bits. The LSD criterion is that of equation (8).

AR model, and the AR coefficients are converted to LSFs to obtain a realistic coding scheme. The first LSF coefficient is coded as is, whereas the remaining LSFs are coded as differences between two adjacent LSFs, $L = [\text{lsf}_1, \text{lsf}_2 - \text{lsf}_1, \text{lsf}_3 - \text{lsf}_2, \ldots, \text{lsf}_M - \text{lsf}_{M-1}]^T$.

In high-rate entropy constrained quantization, the optimal quantizer is the uniform quantizer. The distortion is known, and the same, for all the different approaches, here chosen to be 1.00 dB for transparency. Therefore, the components of $L$ are quantized using uniform quantizers. The entropy of the quantizer indices is calculated for each component of $L$ over all the frames of the validation database. Conventional LP, bandwidth expansion, and regularized LP are all tested using this approach, and the result is shown in table 2.

The conventional LP (**LP**) performs worst, followed by bandwidth expansion (**BEC**). Regularized LP with constant $\lambda$ (**RC**) performs better and the adaptive $\lambda$ (**MOD**) scheme provides the best results. The improvement is approximately 2 bits for the regularized LP methods, as can be seen from the middle line of table 2. Since regularized LP increases the prediction error, it remains to be shown that the composite modeling of coefficients and residual also improves compared to conventional linear prediction.

## 6. CONCLUSIONS

Regularized linear prediction yields a smoother and physically more plausible spectral envelope than linear prediction with bandwidth expansion. Whereas bandwidth expansion simply modifies an existing LP estimate of the spectral envelope that may be contaminated by spectral fine structure, regularized LP optimizes the spectral fit with a penalty on nonsmooth behavior. This provides the regularization procedure with an advantage over bandwidth expanded LP at a fundamental level. As a result, regularized LP can provide an accurate estimate of the spectral with a high model order.

Our simulations show that the regularized linear prediction models an approximation to the vocal tract transfer function more accurately than bandwidth expanded linear prediction, even for a constant penalty factor $\lambda$. Regularized prediction has a negligible additional computational cost, and results in a lower coding rate for the spectral description. Thus, regularized linear prediction forms an attractive method for estimation of the AR parameters in speech processing systems.

## 7. REFERENCES


[1] J. Makhoul, "Linear prediction: A tutorial review," *Proc. of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[2] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 221–239, May 2000.

[3] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Processing*, vol. 39, pp. 411–423, February 1991.

[4] P. Kabal and W. B. Kleijn, "All-pole modelling of mixed excitation signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 97–100, 2001.

[5] C. Lee, "On robust linear prediction of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 642–650, 1988.

[6] H. W. Strube, "Linear prediction on a warped frequency scale," *J. Acoust. Soc. Am.*, vol. 68, no. 4, pp. 1071–1076, 1980.

[7] Y. Tohkura, F. Itakura, and S. Hashimoto, "Spectral smoothing techniques in PARCOR speech analysis-synthesis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, pp. 587–596, 1978.

[8] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, pp. 309–321, 1975.

[9] O. Cappé and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing Letters*, vol. 3, pp. 100–102, April 1996.

[10] M. Oudot, O. Cappe, and E. Moulines, "Robust estimation of the spectral envelope for "harmonics+noise" models," in *Proc. IEEE Workshop on Speech Coding*, pp. 11–12, 1997.

[11] M. N. Murthi and W. B. Kleijn, "Regularized linear prediction all-pole models," in *Proc. IEEE Workshop on Speech Coding*, pp. 96–98, September 2000.

[12] J.-H. Chen, "Low-delay coding of speech," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), pp. 209–256, Elsevier, Amsterdam, 1995.

[13] B. Wei and J. D. Gibson, "A new discrete spectral modeling method and an application to CELP coding," *IEEE Signal Processing Letters*, vol. 10, pp. 101–103, April 2003.

[14] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM Review*, vol. 34, pp. 561–580, December 1992.