REGULARIZED OPTIMIZATION WITH SPECTRAL SMOOTHING FOR SPEECH SPECTRAL ESTIMATION

Karsten Vandborg Sørensen and Søren Vang Andersen

Department of Communication Technology
Aalborg University, DK-9220 Aalborg Ø, Denmark
{kvs,sva}@kom.auc.dk

ABSTRACT

This paper presents a minimization approach based on regularized optimization [1] for use in spectral envelope estimation for speech in the presence of noise. The objective function that is minimized consists of the sum of a data fitting term and a linear combination of the following regularization terms: Minimum perturbation energy, minimum estimate energy, peak preservation, and spectral smoothing. Regularization weights are used to control the trade-off between the regularization terms such that the estimates obtain a lower bias and low variance, while preserving a shape that is natural for speech.

1. INTRODUCTION

When estimating auto-regressive (AR) models of speech signals in additive noise an ordinary linear prediction will estimate coefficients that model the combined signal of speech and noise. The formulation as an unconstrained minimization problem by means of regularized optimization is made because a number of meaningful cost functions exist for which a trade-off should be minimized. This regularization approach was used by Murthi and Kleijn [2] to ensure smoothness of a linear prediction spectral envelope of clean speech. In this paper we combine a number of regularization terms to address the problem of spectral envelope estimation in the presence of noise.

1.1. One-Step Prediction Model with Structured Perturbation

We consider the following one-step prediction model

$$\underbrace{\begin{bmatrix} 0 & 0 & \dots & 0 \\ \vdots & \vdots & \iddots & x_1 \\ \vdots & 0 & \iddots & \vdots \\ 0 & x_1 & & x_N \\ x_1 & \vdots & \iddots & 0 \\ \vdots & x_N & \iddots & \vdots \\ x_N & 0 & \dots & 0 \end{bmatrix}}_{\mathbf{X}\in\mathbb{R}^{(N+p)\times p}} \underbrace{\begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}}_{\mathbf{a}\in\mathbb{R}^{p\times 1}} = \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_N \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{\mathbf{x}\in\mathbb{R}^{(N+p)\times 1}} + \underbrace{\begin{bmatrix} r_1 \\ \vdots \\ r_N \\ r_{N+1} \\ \vdots \\ r_{N+p} \end{bmatrix}}_{\mathbf{r}\in\mathbb{R}^{(N+p)\times 1}},$$

where $\mathbf{X}$ is the noisy speech matrix, $\mathbf{x}$ is the noisy speech vector, $\mathbf{a}$ is the AR coefficient vector, and $\mathbf{r}$ is the residual vector. N is the frame length and p is the order of the AR model. The noisy data matrix and vector has been filled with zeros to ensure stable filter coeffecients in $\mathbf{a}$.

The method of total least squares (TLS) [3] allows a perturbation of the noisy data with a perturbation vector $\mathbf{\Delta x} \in \mathbb{R}^{(N+p)\times 1}$ and matrix $\mathbf{\Delta X} \in \mathbb{R}^{(N+p)\times p}$ of perturbations Δx_i for $i = 1,\dots,$ $(N+p)(p+1)$ without structure, i.e.

$$(\mathbf{X} + \mathbf{\Delta X})\mathbf{a} = \mathbf{x} + \mathbf{\Delta x} + \tilde{\mathbf{r}},$$

where $\tilde{\mathbf{r}} \in \mathbb{R}^{(N+p)\times 1}$ is the residual vector of the perturbed system of equations. In a TLS solution $\mathbf{\Delta X}$ and $\mathbf{a}$ are obtained using the singular value decomposition (SVD) [3][4]. $\mathbf{\Delta x}$ is trivial to find once $\mathbf{\Delta X}$ and $\mathbf{a}$ are available. The TLS solution has the property of minimizing $||[\mathbf{\Delta X}|\mathbf{\Delta x} + \tilde{\mathbf{r}}]||_F$.

In the method of structured total least squares (STLS) [4] the same expression is minimized but with a structure constraint (e.g. Hankel structure as in this case) imposed on $\mathbf{\Delta X}$ or on the augmented matrix $[\mathbf{\Delta X}|\mathbf{\Delta x} + \tilde{\mathbf{r}}]$. For feasibility and because $||\tilde{\mathbf{r}}||_2$ will go to zero in the STLS solution we impose the structure on $[\mathbf{\Delta X}|\mathbf{\Delta x}]$, i.e.

$$\mathbf{\Delta x} = [\Delta x_1, \dots, \Delta x_N, 0, \dots, 0]^T,$$

and with a perturbation matrix $\mathbf{\Delta X}$ that is defined using the Hankel operator $\mathcal{H}$ to equate

$$\mathcal{H}([0,\dots,0,\Delta x_1,\dots,\Delta x_N],[\Delta x_N,0,\dots,0]).$$

By exploiting the Hankel structure of $[\mathbf{\Delta X}|\mathbf{\Delta x}]$ the residual $\tilde{\mathbf{r}}$ of the perturbed system of equations can be rearranged as follows, where we use that $\mathbf{\Delta X a} = \mathbf{F \Delta x}$ [4], i.e.

$$\begin{aligned}\tilde{\mathbf{r}} &= (\mathbf{X} + \mathbf{\Delta X})\mathbf{a} - \mathbf{x} - \mathbf{\Delta x} = \mathbf{r} + \mathbf{\Delta X a} - \mathbf{\Delta x} \\ &= \mathbf{r} + \mathbf{F \Delta x} - \mathbf{\Delta x} = \mathbf{r} + \tilde{\mathbf{F}}\mathbf{\Delta x},\end{aligned}$$

with $\tilde{\mathbf{F}} = \mathbf{F} - \mathbf{I} \in \mathbb{R}^{(N+p)\times(N+p)}$ given by

$$\tilde{\mathbf{F}} = \mathcal{T}([-1, a_p, \dots, a_1, 0, \dots, 0], [-1, 0, \dots, 0]),$$

where $\mathcal{T}$ denotes the Toeplitz operator. Due to the structure the signal estimate $\hat{\mathbf{s}}$ is simply the sum of the leading length-N vector in the noisy speech vector and the perturbation vector, i.e.

$$\hat{\mathbf{s}} = \mathbf{x}_N + \mathbf{\Delta x}_N.$$

Our quest is now to find a perturbation vector $\mathbf{\Delta x}_N$ that reduces power when added to $\mathbf{x}_N$. The power should be reduced with an amount equal to what we expect is the power of the noise vector. We base our pursuit to this quest on an assumption of orthogonal signal and noise vectors, i.e. $\mathbf{s}^T\mathbf{n} = 0$.

The authors would like to thank Joachim Dahl, Aalborg University for helpful advice on convex optimization. This work was supported by The Danish National Centre for IT Research, Grant No. 329.

Where the STLS method sets $||\tilde{\mathbf{r}}||_2$ to zero our approach is to identify an AR model with minimum, but non-zero, $||\tilde{\mathbf{r}}||_2$ under the constraint that $||\mathbf{\Delta x}_N||_2$ equate the expectation of $||\mathbf{n}||_2$. We find this approach more natural since our goal is to estimate the AR coefficients of the noise-free speech, and since these coefficients correspond to an AR model with nonzero residual. Furthermore the STLS method calls for joint optimization of the perturbation vector and the coefficient vector. This type of problem has a very high complexity [4] and cannot be solved exact in finite time. It can however be approximated by means of convergence towards a local stationary point, e.g. using alternating minimization where one vector is fixed when the other is estimated and vice versa until the squared sum of changes in the vectors becomes smaller than a small value ϵ. In this paper we will use two steps of this alternating minimization, initially with a fixed prediction coefficient estimate. Due to a very short frame length of 5 ms and a fifty percent overlap between frames we choose to initialize this vector as the estimated AR coefficients from the previous frame and in the first frame we initialize with the coefficient vector of the noisy speech. We then find the perturbation that gives the smallest residual vector, i.e.

$$\mathbf{\Delta x} = \arg\min_{\mathbf{\Delta x}} ||\tilde{\mathbf{r}}||_2^2 = \arg\min_{\mathbf{\Delta x}} \left|\left|\tilde{\mathbf{F}}\mathbf{\Delta x} + \mathbf{r}\right|\right|_2^2.$$

When the perturbation vector is found we discard the last p elements and use it to create the perturbation matrix and insert both in the model and solve in a least squares sence for the coefficient vector. This will cause the perturbed data to be described closely by the AR model parameters $\mathbf{a}$ (in $\tilde{\mathbf{F}}$) when $||\tilde{\mathbf{r}}||_2^2$ is small and in our case thereby using information from the previous frame to obtain the current estimate of the speech spectral envelope. The principle of allowing a small change using this definition of cost has previously been used by Jensen, Jensen, and Hansen [5]. If the signal is not quasi-stationary over at least two frames we recommend that the coefficient vector of the noisy speech is used instead of the estimated coefficient vector from the previous frame. We denote this term the data fitting term, and it will be the first term in the objective function of the minimization.

2. OBJECTIVE FUNCTION

The perturbation vector we want to find is the argument that minimizes the objective function, i.e.

$$\mathbf{\Delta x} = \arg\min_{\mathbf{\Delta x}} f(\mathbf{\Delta x}),$$

where the quadratic objective function $f(\mathbf{\Delta x})$ that consist of several cost-defining regularization terms is given by

$$\begin{aligned} f(\mathbf{\Delta x}) = &\left|\left|\tilde{\mathbf{F}}\mathbf{\Delta x} + \mathbf{r}\right|\right|_2^2 + \nu_1 ||\mathbf{\Delta x}||_2^2 + \nu_2 ||\mathbf{x} + \mathbf{\Delta x}||_2^2 \\ &+ \nu_3 ||\mathbf{W}_1 \boldsymbol{\mathcal{F}} \mathbf{\Delta x}||_2^2 + \nu_4 ||\mathbf{W}_2 \mathbf{D} \boldsymbol{\mathcal{F}}(\mathbf{x} + \mathbf{\Delta x})||_2^2, \quad (1) \end{aligned}$$

with the regularization weights ν_1, ν_2, ν_3, and $\nu_4 \in \mathbb{R}_+$, i.e. the non-negative real numbers.

2.1. Power Reduction

The second term $||\mathbf{\Delta x}||_2^2$ and the third term $||\mathbf{x} + \mathbf{\Delta x}||_2^2$ drag the solution in opposite directions so the relation between their weights will determine how much of the power in $\mathbf{x}$ that will remain after adding $\mathbf{\Delta x}$. If the estimated relation between noise and noisy speech power is given by $||\mathbf{n}||_2^2 = k\,||\mathbf{x}||_2^2$, then we use the following relationship between the weights

$$\nu_1 = \nu_2 \left(\frac{1 - \sqrt{k}}{\sqrt{k}} \right), \quad (2)$$

which for high weights, and $\mathbf{s}^T\mathbf{n} = 0$, will ensure that

$$||\mathbf{\Delta x}||_2^2 = k\,||\mathbf{x}||_2^2 = ||\mathbf{n}||_2^2.$$

For feasibility we will neglect the amount of power that will be distributed among the last p elements in $\mathbf{\Delta x}$. If necessary a regularization term could be added to the objective function that would provide zero perturbation of the last p elements.

2.2. Peak Preservation

Because voiced speech signals contain a high amount of power at formant frequencies we want to pick out the highest peaks in the noisy speech spectrum and preserve the power at these frequencies, i.e. we neglect the noise at the peaks due to high signal-to-noise ratio at these frequencies. The fourth term is given by

$$||\mathbf{W}_1 \boldsymbol{\mathcal{F}} \mathbf{\Delta x}||_2^2,$$

where $\boldsymbol{\mathcal{F}} \in \mathbb{R}^{(N+p)\times(N+p)}$ is a matrix with Fourier basis vectors in its rows and $\mathbf{W}_1 \in \mathbb{R}^{(N+p)\times(N+p)}$ is a diagonal weight matrix. $\mathbf{W}_1$ will have the property of assigning a high cost to perturbations at peak frequencies. How it is obtained is described by the pseudo-code in Algorithm 1.

Algorithm 1 Calculation of weight matrix $\mathbf{W}_1$

```
for each frame do
  calculate the magnitude spectrum
  find index with peaks
  peakcount ← 0
  for every peak (in decreasing magnitude order) do
    if peak magnitude > 0.1 times the largest peak then
      if peakcount < 3 then
        save index of peak
        increase peakcount by 1
      end if
    end if
  end for
  assign weight 1 at five neighboring frequencies centered at the peak
end for
```

2.3. Smoothness

The last regularization term is used to ensure a smooth magnitude spectrum of the estimated speech signal. The preferred term would be similar to the one formulated by Murthi and Kleijn [2], that is

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{d}{d\omega} \log |S(\omega, \mathbf{x} + \mathbf{\Delta x})| \right|_2^2 d\omega, \quad (3)$$

but since we are looking for an objective function that can be optimized analytically (in closed form) and is computationally efficient we propose a more tractable approximation of (3) given by

$$||\mathbf{W}_2 \mathbf{D} \boldsymbol{\mathcal{F}}(\mathbf{x} + \mathbf{\Delta x})||_2^2,$$

where $\mathbf{D} \in \mathbb{R}^{(N+p-1)\times(N+p)}$ is a discrete first order differentiator, i.e.

$$\mathbf{D} = \mathcal{T}([-1, 0, \ldots, 0], [-1, 1, 0, \ldots, 0]).$$

$\mathbf{W}_2 \in \mathbb{R}^{(N+p-1)\times(N+p-1)}$ is a diagonal matrix that enables us to use e.g. frequency dependent smoothness. Since we have left out the absolute value operator we depend on the neighbouring Fourier coefficients to have approximately the same angle in order for the approximate differentiation (and squaring from the squared norm) to be correct. In practice this has turned out to be an acceptable assumption. The weight matrix $\mathbf{W}_2$ contains nothing but unit weight between peaks.

2.4. Summary

When we combine all the previously described terms to a regularized objective function to be minimized they act together to create meaningful estimates; We reduce an amount of (noise) energy in a way that preserves high power spectral regions (that we assume to be formants) while ensuring a smooth spectrum of the estimate with a shape that is slightly fitted to the estimated spectrum from the previous frame. The main principle is illustrated in Figure 1.

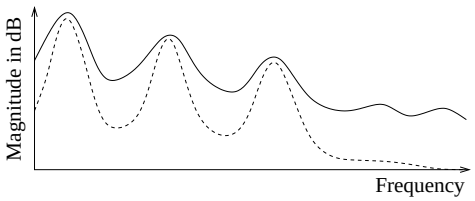

Fig. 1. The noisy speech spectral envelope (solid line) and what we want to achieve (dashed line) by the optimization procedure.

2.5. Minimization of the Objective Function

The objective function (1) can be expanded and rewritten as

$$f(\mathbf{\Delta x}) = \mathbf{\Delta x}^T\mathbf{A}\mathbf{\Delta x} + 2\mathbf{b}^T\mathbf{\Delta x} + \mathbf{c},$$

where $\mathbf{A} \in \mathbb{R}^{(N+p)\times(N+p)}$, $\mathbf{b} \in \mathbb{R}^{N+p}$, and $\mathbf{c} \in \mathbb{R}$ consist of coefficients given by

$$\begin{aligned}\mathbf{A} &= \tilde{\mathbf{F}}^T\tilde{\mathbf{F}} + (\nu_1+\nu_2)\mathbf{I} + \nu_3\boldsymbol{\mathcal{F}}^H\mathbf{W_1}^T\mathbf{W_1}\boldsymbol{\mathcal{F}} \\ &\quad + \nu_4\boldsymbol{\mathcal{F}}^H\mathbf{D}^T\mathbf{W_2}^T\mathbf{W_2}\mathbf{D}\boldsymbol{\mathcal{F}} \qquad (4)\\ \mathbf{b} &= \tilde{\mathbf{F}}^H\mathbf{r} + \nu_2\mathbf{x} + \nu_4\boldsymbol{\mathcal{F}}^H\mathbf{D}^T\mathbf{W_2}^T\mathbf{W_2}\mathbf{D}\boldsymbol{\mathcal{F}}\mathbf{x} \\ \mathbf{c} &= \mathbf{r}^T\mathbf{r} + \nu_2\mathbf{x}^T\mathbf{x} + \nu_4\mathbf{x}^T\boldsymbol{\mathcal{F}}^H\mathbf{D}^T\mathbf{W_2}^T\mathbf{W_2}\mathbf{D}\boldsymbol{\mathcal{F}}\mathbf{x}.\end{aligned}$$

Note that $\mathbf{W}_1, \mathbf{W}_2$, and $\mathbf{D}$ are highly sparse matrices which enables us to calculate the matrix products very efficiently. The gradient is given by

$$\nabla f(\mathbf{\Delta x}) = 2\mathbf{A}\mathbf{\Delta x} + 2\mathbf{b},$$

thus the argument that minimizes $f(\mathbf{\Delta x})$ can be found analytically if we set this gradient vector equal to the zero vector and solve for $\mathbf{\Delta x}$. The solution to this problem is the minimum norm least-squares solution, i.e.

$$\mathbf{\Delta x} = -\mathbf{A}^\dagger\mathbf{b},$$

where † denotes Moore-Penrose pseudo-inverse. Note that we do not need to calculate $\mathbf{c}$ as it only adds a constant offset to the function that is minimized and therefore not will be a part of the solution. When ν_1 and ν_2 are chosen to have relative large values they will cause a high valued constant diagonal matrix (second term in Equation 4) to be a part of the square matrix $\mathbf{A}$ and it will in practice always have full rank which enables us to solve much more efficiently for $\mathbf{\Delta x}$, i.e.

$$\mathbf{\Delta x} = -\mathbf{A}^{-1}\mathbf{b}.$$

3. MONTE CARLO SIMULATION

A Monte Carlo simulation with 1000 runs per frame has been used to evaluate the developed method. An estimate of the noise-free AR coefficient vector is obtained in each run, thus we obtain a *set* of estimates for each frame. The average of the estimates is denoted the *centroid* and the distance between the noise-free coefficient vector and this centroid is used as a measure of estimator bias and the average distance of the estimates in a set to the centroid is used as a measure of estimator variance.

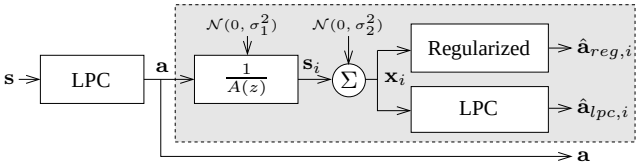

Fig. 2. Monte Carlo scenario as it is used for evaluating the proposed regularized method. The shaded region within the dashed square uses the same coefficients in the synthesis filter in each frame but replaces the noise realizations used as filter excitation signal and additive noise in each run.

The method is evaluated on a 500 ms segment of voiced speech sampled at 16 kHz. The frame length is N=80 (5 ms) and the frames are half overlapping. The speech segment is shown in Figure 3.

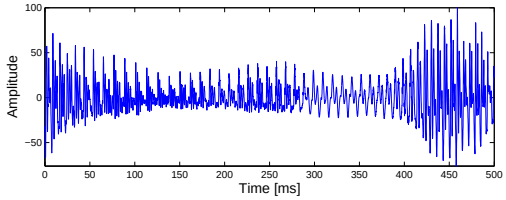

Fig. 3. The 500 ms noise-free speech segment used in the experiments.

A signal-to-noise ratio of 10 dB and a model order of p=16 have been used in the experiments. The regularization weights are given by ν_2=10, ν_3=1, and ν_4=0.1 (ν_1 is calculated in run-time using Equation 2 and an estimate of k).

3.1. Distance Measures

We use the sum-squared error (SSE) and the Itakura distance measure [6, p328] as measures of distance between the estimated AR coefficient vector $\hat{\mathbf{a}}$ and the noise-free coefficient vector $\mathbf{a}$, i.e.

$$d_{SSE}(\mathbf{a},\hat{\mathbf{a}}) = (\mathbf{a}-\hat{\mathbf{a}})^T(\mathbf{a}-\hat{\mathbf{a}}),$$

$$d_{Itakura}(\mathbf{a},\hat{\mathbf{a}},\mathbf{s}) = \frac{\hat{\mathbf{a}}^T\mathbf{R}_{ss}\hat{\mathbf{a}}}{\mathbf{a}^T\mathbf{R}_{ss}\mathbf{a}},$$

where $\mathbf{R}_{ss}$ is the autocorrelation matrix of the signal $\mathbf{s}$ and $\mathbf{a}$ is the coefficients from a linear prediction coding (LPC) of $\mathbf{s}$ using the same window as in the regularized method. Note that SSE equals squared Euclidean distance.

4. EXPERIMENTAL RESULTS

Table 1 and Figure 4 contain objective measurements of the performance in the form of estimator bias and variance averaged across frames. Note that estimator bias in general is much smaller for the developed method than the LPC and that estimator variance is larger (but still small in the Itakura distance measure which is the most refined measure of the two). Two spectral estimates are shown in Figure 5 to illustrate the importance of the weight matrices and the performance once the number of peaks used in the peak preservation weight match the actual number.

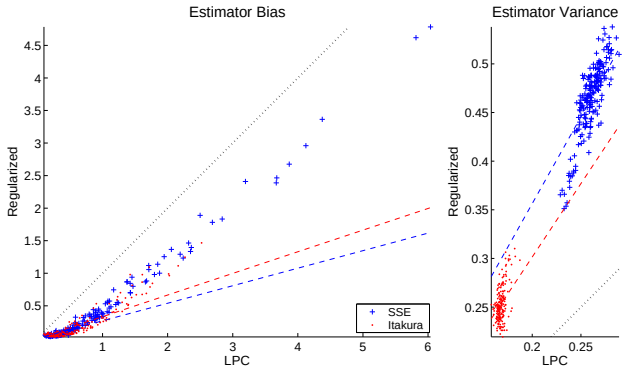

Fig. 4. Estimator bias (left) and estimator variance (right) in each frame. The dashed lines have slopes equal to the division of the bias (averaged across frames) from this regularized method and the LPC listed in Table 1. The estimator bias for the developed method is much lower than for the LPC and the estimator variance is slightly higher than for the LPC.

Preliminary informal listening tests on a signal constructed from $\mathbf{x}_N + \mathbf{\Delta x}_N$ indicate good noise reduction but with some remaining waterfall noise.

	Estimator Bias		Estimator Variance	
Distance Measure	SSE	$Itakura$	SSE	$Itakura$
LPC	0.8371	0.6188	0.2607	0.1680
Regularized	0.4069	0.2031	0.4651	0.2534

Table 1. Estimator bias and variance averaged across frames for the speech segment in Figure 3.

5. CONCLUSION

We have developed a highly flexible and stable method with a closed-form solution and with a reasonably low computational cost. The developed method has proven to reduce estimator bias in both the SSE and Itakura distance measure. It has a larger estimator variance in the SSE distance measure and only slightly larger estimator variance in the Itakura distance measure. The basic principle where peaks are preserved while noise power is reduced from the remaining smoothened spectrum works very well in voiced regions, but it is not well suited for unvoiced speech. The high amount of fine tuning that are needed to make the method work well is a drawback of the method and a new method with a very limited number of weights is currently being developed. Elements from this method could very well be used as regularization terms for already existing methods to ensure e.g. a smooth or a peak preserved spectrum. Perceptual weights could be incorporated in the weight matrices to improve the perceived quality.

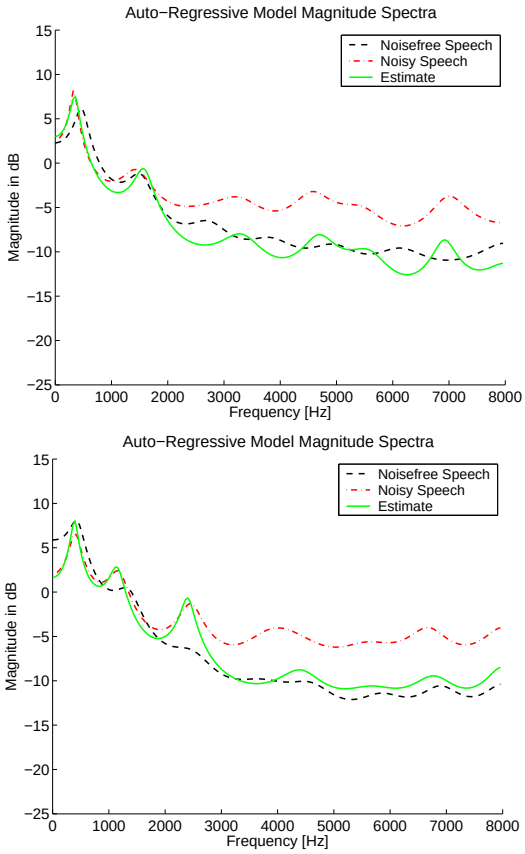

Fig. 5. Effect of two (upper) and three (lower) allowed peaks. Two peaks gives a good spectral envelope estimate and three peaks gives an additional peak in the estimated spectral envelope in this example.

6. REFERENCES

[1] Stephen Boyd and Lieven Vandenberghe, "Convex Optimization," `http://www.ee.ucla.edu/~vandenbe/cvxbook.html`, Dec. 2002.

[2] Manohar N. Murthi and W. Bastiaan Kleijn, "Regularized Linear Prediction All-Pole Models," in *IEEE Workshop on Speech Coding Proceesings*, Lake Lawn Resort, Delavan, Wisconsin, USA, Sept. 2000, pp. 96–98.

[3] Sabine Van Huffel and Joos Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, Frontiers in Applied Mathematics, Vol. 9, Society for Industrial and Applied Mathematics, 1991.

[4] Sabine Van Huffel, Haesun Park, and J. Ben Rosen, "Formulation and Solution of Structured Total Least Norm Problems for Parameter Estimation," *IEEE Transactions on Signal Processing*, vol. 44(10), pp. 2464–2474, Oct. 1996.

[5] Jesper Jensen, Søren Holdt Jensen, and Egon Hansen, "A Perturbation-Based Pre-Processing Algorithm for CELP-Coders," in *IEEE Speech Coding Workshop Proceedings*, Haikko Manor, Porvoo, Finland, June 1999, pp. 153–155.

[6] John R. Deller, Jr., John G. Proakis, and John H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Wiley-Interscience, 2000.