ROBUST LINEAR PREDICTION ANALYSIS OF VOWELS

David J.M. Weenink

1. INTRODUCTION

Many a time we want a description of vowels in terms of formants and bandwidths. Determining these parameters is a tedious matter, especially when these vowels are uttered by women and children. Linear prediction analysis seems to be the best way to estimate these parameters. However, formant and bandwidth estimations often fail on these voices, either formants are completely missed or not positioned well. A new algorithm which should obtain better estimates was suggested by Lee (1987). In order to get more insight in the way it will behave on natural speech, we tested the algorithm on some artificially generated signals.

2. AUTOREGRESSIVE MODELLING OF SPEECH

Linear prediction has been a widely employed method in the past decades in the modeling of the speech signal. At the core of linear prediction lies the assumption that a speech sample s_n can be considered as a linear combination of past samples and a certain input u_n.

$$s_n=-\sum_{k=1}^{p} a_k s_{n-k} + u_n \qquad (1)$$

Given a particular signal $\{s_1,...,s_N\}$ the problem is to determine the predictor coefficients a_k in some manner.
In speech the source signal u_n is unknown and some model for it has to be assumed. This means that the signal s_n can only be estimated from a linear summation of past samples. Let this approximation of s_n be s_n', where

$$s_n'=-\sum_{k=1}^{p} a_k s_{n-k}$$

Then the error between s_n and its predicted value s_n' is given by

$$e_n=s_n-s_n'=s_n+\sum_{k=1}^{p} a_k s_{n-k}$$

This equation can be rewritten as

$$s_n=-\sum_{k=1}^{p} a_k s_{n-k} + e_n \qquad (2)$$

and we see that the only input signal u_n that will result in the signal s_n as output, is that where $u_n=e_n$. That is, the error signal is an estimation of the source signal.

In standard linear prediction a solution for the a_k is determined by minimizing a function $\rho(e_n)$ of the error.

$$\text{minimize } \Sigma \rho(e_n), \text{ where } \rho(e_n)=e_n^2 \qquad (3)$$

When this function is used we call this minimalization Least Squares (LS) minimalization. This minimalization is optimal when the errors e_n are normally distributed, i.e. when the error signal is gaussian white noise, and it is equivalent to minimizing the variance of the error signal (source signal). We can solve this LS problem with either the autocorrelation or the covariance method (Makhoul, 1975).
This method of least squares has been used for many years. Nevertheless, it is well known that outliers have an uncomfortably large influence on the resulting LS estimators. For voiced speech the source is of a quasi-periodic nature with spiky excitations at glottal opening and closing which interact with the filter A(z). This type of interaction results in systematic errors in LPC-derived formants and bandwidths estimates and becomes more severe as the fundamental frequency is raised (Atal, 1975). One solution to this problem that has been suggested is to restrict the analysis interval to the region of glottal closure. During the closed glottis interval, the speech wave consists of free decaying oscillations whose frequencies and decay rates are not being influenced by the glottal pulse. Yet, when we only have the speech signal itself at our disposition, such an interval of appropriate length is difficult to locate in natural speech, especially in speech uttered by females and children where the pitch is high. And even if such an interval can be defined, stability problems can arise because the number of samples available for closed glottis analysis is limited for signals with a short pitch period.
A great deal of these problems can be overcome with a better model for the source function, or, equivalently, with a better error minimalization criterium, which reduces the influence of outliers. Accordingly, robust procedures have been created to modify the LS scheme so to deweight the influence of outliers on the final estimation (Lee, 1987; Miyoshi et al., 1987).

3. ROBUST ESTIMATION

We are faced with the following problem: We have a sequence of independent and identically distributed errors $\{e_1,...,e_N\}$ with distribution function G and we are looking for a function $\rho(e_n)$ which minimizes $\Sigma \rho(e_n)$. The function $\rho(e_n)$ which has been widely used is $\rho(e_n)=e_n^2$ and this function is optimal when G is a normal distribution, which means that the minimalization we perform is a least squares minimalization. What happens to the estimators when the distribution of $\{e_1,...,e_N\}$ is not strictly normal but is contaminated by some amount ε with an unknown distribution H with a much bigger variance? We can imagine that in this case our parameters are poorly estimated. Let us therefore firstly restrict our attention to a distribution
$F= (1-\varepsilon)\Phi + \varepsilon H$, where Φ is the standard normal distribution, H is an unknown contaminating distribution and ε is a known number $0 \leq \varepsilon \leq 1$. Under the assumption that H puts most of its weigth in the tails of Φ , Huber (1964) showed that the most robust function $\rho(e_n)$, i.e. least sensitive to outliers, which minimizes $\Sigma \rho(e_n)$ is:

$$\begin{aligned} &\rho(e_n)=1/2\ e_n^2 && |e_n| < k \\ &\rho(e_n)=k|e_n| - 1/2\ k^2 && |e_n| \geq k \end{aligned} \qquad (4)$$

The k, which depends on ε, is the number of standard deviations of the standard normal distribution from whereon the influence of the contaminating distribution is

assumed.
When we take the derivative of this $\rho(e_n)$ and denote it by $\psi(e_n)$, then

$$\psi(e_n)=\min[\, k, \max(\, e_n, -k\,)\,] \tag{5}$$

Then our minimalization problem can be stated by putting the derivative of ρ equal to zero:

$$\Sigma\, \psi(e_n)=0 \tag{6}$$

In the above formulation scale S (standard deviation), location T (mean) and the amount of contamination ε are known beforehand (S=1, T=0). Huber also showed how to proceed in those cases where we do not know the amount of contamination nor the scale of the normal distribution. First of all, equation (6) has to be modified to

$$\Sigma\, \psi(e_n/S)=0 \tag{7}$$

to make it scale invariant. S is the most robust estimate for the scale and can be calculated in the following way:
Choose beforehand a number k and determine S such that

$$\Sigma\, \psi^2(e_n/S) = N\beta, \text{ with} \tag{8}$$
$$\beta = E_\Phi\, \psi^2(x),$$

the expected value of ψ^2 over the distribution Φ. The value of k hidden in equation (8) is not too critical, any value between 1 and 2 will do, so we choose k=1.5 since this value performs well in practice. The above equation can be solved in the following way: Let m_1, m_2 and m_3 be the numbers of observations satisfying $e_n \le -kS$, $|e_n| < kS$ and $e_n \ge kS$ respectively. We can then write equation (8) as

$$m_1k^2+ \Sigma'(e_n/S)^2 + m_3k^2 = N\beta, \text{ which gives}$$

$$S^2 = \Sigma' e_n^2/(N\beta-(m_1+m_3)k^2), \tag{9}$$

where the summation Σ' is only over the observations m_2. This can be used to compute S by iterations: Start with some initial value S_0 and determine the numbers m_1, m_2 and m_3. Compute a new S (S_1) with formula (9). Determine new m_1, m_2 and m_3 and compute S_2. Repeat this scheme untill the numbers m_1, $m_2$and m_3 do not change any more from one iteration to the following.
The obtained value of S is the most robust estimate for the variance of a normal distribution with mean (T=0) and an unknown amount of contamination ε. Equation (7), now with known S, can be used for obtaining a solution for the problem at hand, in our case auto regressive modeling.

4. GENERALIZATION TO AUTOREGRESSIVE MODELING

For the generalization of robust estimation to auto regressive modeling of the speech signal we will follow the scheme developed by Lee (1987).
We assume that a time series $\{ s_1,\ldots,s_N \}$ is generated by an auto regressive model of order p. We can express the predictor residuals for any LP coefficient vector $\underline{a}$ as

$$e_n(\underline{a})=s_n + \sum_{k=1}^{p} a_k s_{n-k}, \qquad n=p+1,\ldots,N \tag{10}$$

The above formulation for minimalization can be applied to solve for an estimate of the LP coefficients. An estimate for the LP coefficients is obtained by solving the following optimalization problem

$$\text{minimize} \sum_{n=p+1}^{N} \rho(e_n(\underline{a})) \tag{11}$$

For $\rho(t)$ we take we take the above defined function with derivative $\psi(t)$. The corresponding system of estimating equations then becomes

$$\sum_{n=p+1}^{N} s_{n-j}\psi(e_n(\underline{a})/S) =0 \qquad j=1,\ldots,p \tag{12}$$

S is the robust estimator of the scale parameter (the standard deviation) of the residuals. In general this system of equations is nonlinear and iterative methods are required to solve for the coefficients a_k's.
For solving the above equation we substitute

$$\psi(e_n/S)=e_n w_n \tag{13}$$

From $\psi(x)= \min[\, k, \max(\, x, -k\,)\,]$ it follows that the weigths w_n are:

$$\begin{aligned} w_n &= 1 && \text{for } |e_n| < kS \\ w_n &= kS/|e_n| && \text{for } |e_n| \geq kS \end{aligned} \tag{14}$$

Equation (11) can now be written in matrix form as

$$R\underline{a} =-\underline{c} \qquad \text{with solution } \underline{a} = -R^{-1}\underline{c} \tag{15}$$

where R is the weighted covariance matrix with elements

$$R_{ij} = \sum_{n=p+1}^{N} s_{n-i}s_{n-j}w'_n \qquad 1 \leq i,j \leq p, \tag{16}$$

and R^{-1} is its inverse. The weights w'_n are computed based on the residuals obtained from a preliminary estimate $\underline{a}'$ for the prediction coefficients.

5. THE ALGORITHM

The structural steps in the algorithm are displayed in figure 1. The iteration loop is entered with all weigths set equal to unity and the covariance matrix R is calculated (16). Next the coefficients a_k are determined and used for the calculation of the error signal (10). From this error signal the scale S is determined via an iterative process and the weigths are calculated (14). These new weights w'_n enter in the next iteration step for calculating the new covariance matrix. Convergence is reached when the following condition is fulfilled: $\rho_{i-1}-\rho_i < \alpha\rho_i$. The value for α at this moment is 0.001 but can be made smaller.
At the implementation stadium reached now, the algorithm is rather time consuming,

despite the fact that already some optimalizations have taken place: the algorithm for the calculation of the covariance matrix, responsible for 80% of the time consumption, was optimized by calculating the complete covariance matrix only once for the first iteration. In the next iteration steps only action is undertaken when a weight has changed. The amount of time saved depends of course on the number of iterations, the more iterations the more effective this method becomes (e.g. a saving factor 5 in computing time when the mean number of iterations is 7).

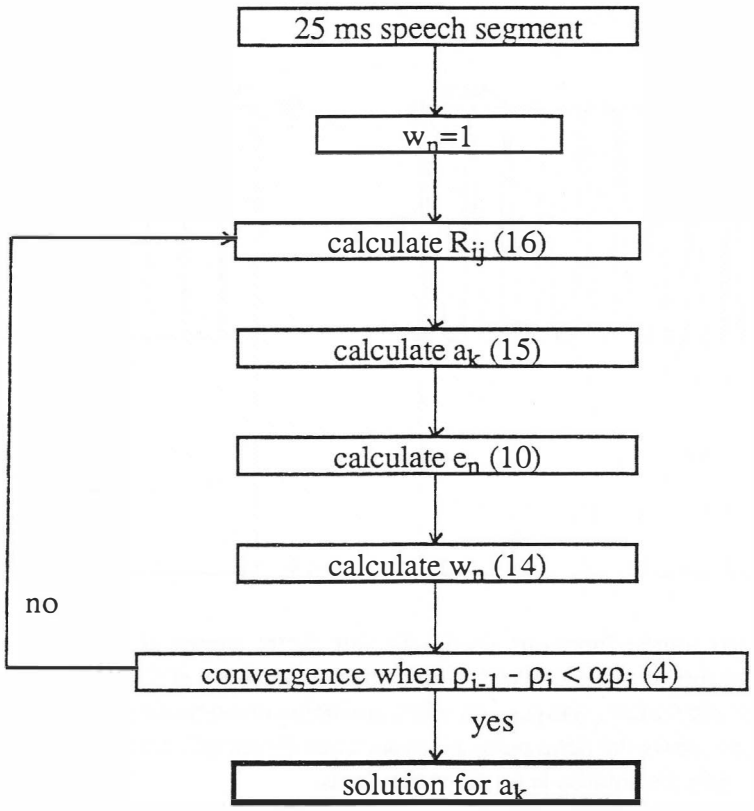

Fig.1. Flow chart of the robust linear prediction algorithm.

6. TEST OF THE LINEAR PREDICTION ALGORITHM

In order to get some insight in the meritus of the algorithm it was tested on some artificially generated signals, since this is the only way to know the exact values of all formants and bandwidths. Three different 4-formant vowels of 60 ms duration and 10 kHz sample frequency were synthesized. Vowels /u/, /a/ and /i/ were taken,being at the corners of the vowel triangle. Bandwidths were choosen 10% of the formant frequency values. Table I shows the formant frequencies and bandwidths of these vowels.

Table I. Formant frequencies and bandwidths of the generated vowels, all values in Hz.

vowel	F_1	B_1	F_2	B_2	F_3	B_3	F_4	B_4
/u/	320	32	670	67	2330	233	3500	350
/a/	800	80	1340	134	2650	265	3500	350
/i/	340	34	2200	220	2970	297	3500	350

For each vowel besides the variation in formant frequencies and bandwidths two other variations were introduced, namely, variation in fundamental frequency and variation in source function. Each vowel was generated with 3 different fundamental frequencies, 125, 225 and 325 Hz being approximately the mean fundamental frequencies for respectively male, female and children voices.
Two different source functions were used, a spike function (delta pulse) and a polynomial source function according to Rosenberg (1971). In the Rosenberg source function the durations of glottal opening and closing times were chosen as 0.4 and 0.16 of the duration of a pitch period.

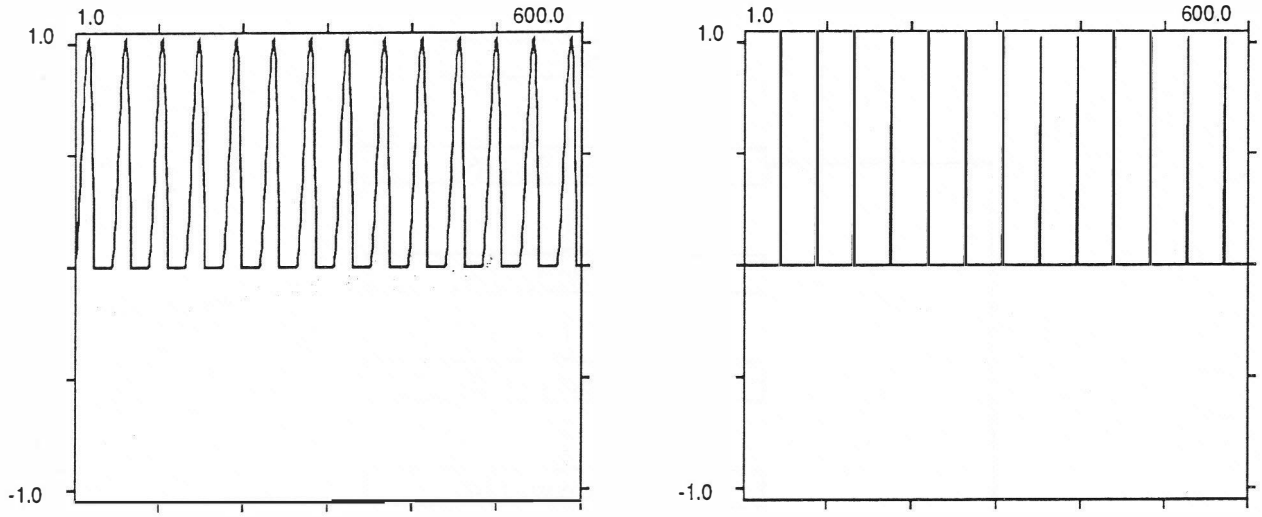

Fig. 2 The two source functions used. The left figure shows the polynomial Rosenberg function for the signals generated with a fundamental frequency of 225 Hz. Durations of glottal opening and closing times relative to the pitch period are respectively 0.40 and 0.16.
The right figure shows the delta pulse function. Time (in sample numbers) is displayed on the horizontal axis and the amplitude on the vertical axis.

A filter of the form $1-z^{-1}$ is employed on the generated vowel to get a better spectral roll-off. Figure 2 shows both source functions for a vowel with a pitch frequency of 225 Hz.
All the synthesized vowels were analyzed with the robust linear prediction algorithm. The results for the vowels with a puls-type excitation are summarized in table II.
As we can see from this table, the estimations of the formant/bandwidths are excellent. The formant frequency estimations are generally within 0.5% of the true values. Bandwidths of the formants are somewhat less exact but deviate generally no more than 4% (see also Golstein Brouwers, 1987). However, when the first formant lies approximately halfway between the first and the second harmonic of the fundamental frequency, the bandwidth estimation of this formant is not so well and estimated too wide: 17% for the /i/ and 50% for /u/. The algorithm iterates much slower to the best estimation, than when the formant lies closer to one of the harmonics.
We must note that the values obtained with this algoritm are significantly better than those obtained with the conventional linear prediction algorithm (table III).

Table II. Deviations of formant frequencies and bandwidths from the true values of the vowels generated with a puls source function. Robust analysis used. All values are in Hz.

vowel	F_0	ΔF_1	ΔB_1	ΔF_2	ΔB_2	ΔF_3	ΔB_3	ΔF_4	ΔB_4
	125	1	5	1	1	0	0	-1	2
/u/	225	4	16	0	1	0	2	-2	2
	325	4	4	3	-1	-1	-4	-3	14
	125	0	0	0	0	0	-1	0	0
/a/	225	1	0	0	0	0	-1	0	2
	325	3	2	1	-3	-2	-2	-2	9
	125	0	2	0	0	0	0	0	2
/i/	225	1	6	0	0	0	0	0	2
	325	1	0	0	-1	-1	0	-1	5

Table III. Deviations of formant frequencies and bandwidths from the true values of the vowels generated with a puls source function. Conventional covariance analysis with order 10 used. All values are in Hz.

vowel	F_0	ΔF_1	ΔB_1	ΔF_2	ΔB_2	ΔF_3	ΔB_3	ΔF_4	ΔB_4
	125	137	243	7	180	-14	-44	-16	99
/u/	225	79	682	4	-26	-31	-41	-26	207
	325	17	1	10	-33	20	-117	-185	343
	125	47	30	36	-30	-45	-36	-97	269
/a/	225	88	77	40	52	-36	-27	-69	186
	325	65	190	3	-78	-23	-79	-49	231
	125	78	195	23	-57	-70	41	-90	100
/i/	225	160	503	0	-32	-35	19	-36	67
	325	2	6	-10	-7	-35	113	-59	284

More serious problems arise when we look at table IV and V, where the analysis results of the signals generated with the Rosenberg puls are displayed. We see that the exact correspondence between estimated parameters and system parameters has disappeared; the estimate has become worser. The effects noted above for the signals with a deltapuls as excitation are magnified. First of all, the formant frequency estimation errors have increased. The greatest relative error in frequency is 8% and it occurs for the first formant of the vowel /i/ with fundamental frequency of 325 Hz. In general, the sign of the deviations in formant frequency is such, that the estimated frequency is close to the nearest harmonic of the fundamental frequency. Despite the worse estimations of formant frequencies, we can say, that the relative errors never exceed 8% and, that this case will only happen when the first formant and the fundamental frequency have approximately the same values.

We now come to the critical point of the robust analysis: bandwidth estimation from signals generated with a speech like source function. Especially the bandwidth estimations are very inexact, whenever there is a low frequency first formant. Two cases occur: bandwidths are estimated too small whenever the first formant is close to the fundamental frequency, and bandwidths are estimated too big when the first formant frequency is halfway between the first and second harmonic of the fundamental.

Table IV. Deviations of formant frequencies and bandwidths from the true values of the vowels generated with a Rosenberg source function. All values are in Hz.

vowel	F_0	ΔF_1	ΔB_1	ΔF_2	ΔB_2	ΔF_3	ΔB_3	ΔF_4	ΔB_4
	125	5	8	-10	7	-10	30	-8	193
/u/	225	-3	40	-11	-6	-12	-11	35	325
	325	8	-31	-1	-55	-39	90	14	625
	125	-6	8	-1	7	6	6	7	7
/a/	225	0	41	2	22	21	17	20	7
	325	-48	24	-4	-24	3	-7	6	7
	125	-13	5	6	5	9	-1	8	7
/i/	225	-6	42	21	21	18	-7	22	12
	325	-26	-24	6	28	12	8	14	15

Table V. Deviations of formant frequencies and bandwidths from the true values of the vowels generated with a Rosenberg source function. Analysis with conventional covariance method of order 10. All values are in Hz. A * signals that the formant was not found.

vowel	F_0	ΔF_1	ΔB_1	ΔF_2	ΔB_2	ΔF_3	ΔB_3	ΔF_4	ΔB_4
	125	-15	31	-30	2	-32	72	-44	449
/u/	225	21	85	4	-39	-25	64	22	359
	325	*	*	-2	-66	-85	96	168	900
	125	-29	30	3	4	8	18	10	13
/a/	225	1	96	3	-2	36	14	36	14
	325	-116	24	-13	-91	17	-29	50	57
	125	6	20	15	3	22	9	18	18
/i/	225	43	92	39	18	18	-9	29	21
	325	*	*	56	110	64	-27	79	109

We rather expected the fit not to be optimal since our model did not correspond exactly to the situation: the assumption is that the source function has a big variance and this is true only for a part of the source function. This means that we cannot separate source and filter in the calculations, which results in poorer estimates. Looking at fig 3. where the error signals of the vowel /u/ are displayed we see that we do not get the correct (differentiated) Rosenberg pulse as the error signal.
Notwithstanding these bad bandwidth estimations with a Rosenberg pulse, we believe that this new robust analysis will be of use in speech analysis and can be improved in some ways. First of all, for voiced speech we can make use of the fact that the outliers in the source function are localized and not randomly divided, and use this to deweigth samples just before and after these outliers. Another improvement could be, to use the coefficients as determined by the algorithm for formant frequency calculations, and to afterwards fit the bandwidths (Willems, 1986).

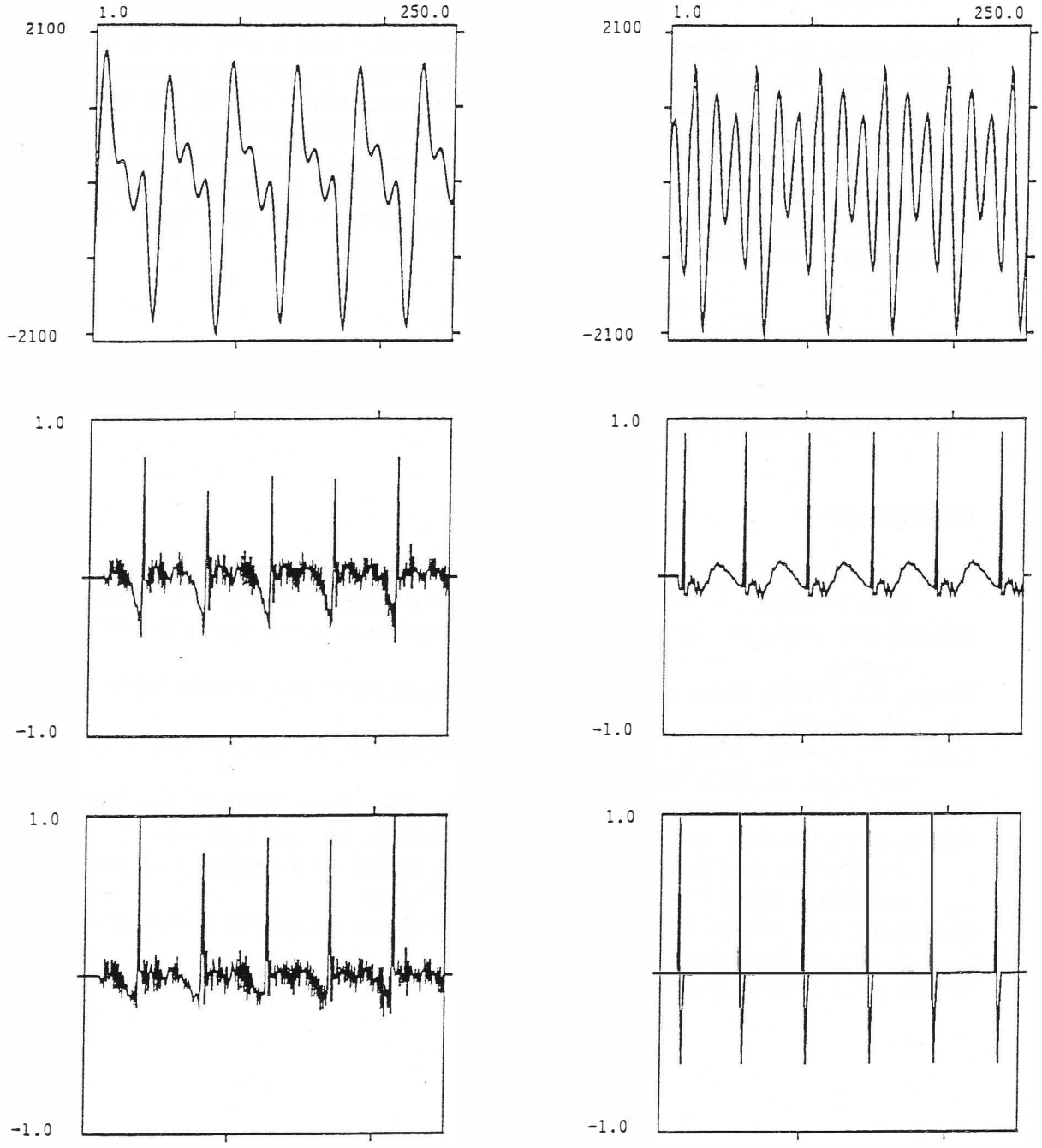

Fig. 3 In both columns from top to bottom respectively 25 ms segment form vowel /u/ with F_0=125 Hz, the error signal from the conventional linear prediction and the error signal from the robust algorithm. In the left column the /u/ has a Rosenberg excitation, in the right column a delta pulse excitation.

7. CONCLUSIONS

The robust algorithm serves well for artificially generated signals when the excitation is a delta pulse. Excellent agreement between estimated and system parameters are obtained. When the excitation is choosen to be a smoother function, i.e. less delta pulselike, the estimations are not so well, especially bandwidths are often seriously underestimated when the first formant is close to the fundamental frequency. It is our belief that the algorithm can be upgraded to perform better in a real speech environment because the basic assumptions underlying the algorithm are more valid than the assumptions for conventional linear prediction.

ACKNOWLEDGEMENTS

Thanks to Wim van Golstein Brouwers for an essential part of the software for the robust analysis and for valuable suggestions.

REFERENCES

Atal, B.S. (1975), Linear prediction of speech - Recent advances with applications to speech analysis, in: D.R. Reddy ed., *Speech Recognition,* Academic Press.

Golstein Brouwers, W. van (1987), Robust LPC-analysis, Iternal report 87 DNL/28 (in Dutch).

Huber, P.J. (1964), Robust estimation of a location parameter, Annals Math. Statist. 35, 73-101.

Lee, C.-H. (1987), Robust linear prediction for speech analysis, Proceedings IEEE Int. Conf. on ASSP 1987, 289-292.

Makhoul, J. (1975), Linear prediction: a tutorial review, Proc. IEEE 63, 561-580.

Miyoshi, Y., Yamato, K., Mizoguchi, R., Yanagida, M. and Kakusho, O. (1987), Analysis of speech signals of short pitch period by a sample-selective linear prediction, IEEE Trans. on ASSP, 35, 1233-1239.

Rosenberg, A.E. (1971), Effect of glottal pulse shape on the quality of natural vowels, J. Acoust. Soc. Am., 49, 583-590.

Willems, L.F. (1986), Robust formant analysis, IPO report 529 (in Dutch).