Adaptive Estimation of Speech Parameters

J.A.L. Basson and J.A. du Preez
Department of Electrical and Electronic Engineering
University of Stellenbosch

Abstract

Linear predictive coding (LPC), and transformations of it, is currently the most popular way of analysing speech signals. Major limitations of using a frame-based technique are that each frame is analysed in isolation of the rest while assuming the excitation source to be a white noise process. In order to reduce computation time, an all pole model is usually employed.

In this project an adaptive algorithm is proposed for speech signal analysis. The algorithm is based on the recursive least squares method with a variable forgetting factor. A pole-zero model is used to estimate the anti-formants present in certain sounds (i.e. nasals and nasalized vowels). This method offers better detection of poles and zeros in stationary environments and faster tracking of pole and zero frequencies in nonstationary signals than other sequential methods. An effective input estimation algorithm eliminates the influence of pitch on the parameter estimates by assuming the input to be a white noise process or a pulse sequence.

1. Introduction

The accurate estimation and tracking of pole and zero frequencies and their bandwidths has long been recognised as important subjects in both speech recognition and speech synthesis. Most parametric estimation algorithms assume separate models for the excitation and the vocal tract response. The vocal tract is usually modelled by analysing the speech with a linear predictive coding (LPC) technique. At the moment the most popular way to address the problem of extracting the information from a speech signal is to use frame based spectrum analysis (usually Marple's [7] technique) with an all-pole (autoregressive) filter model. Only one set of coefficients is obtained for each data frame. An estimate of the glottal excitation waveform can then be obtained by inverse filtering.

Several factors influence the accuracy of the parameters estimated with an AR-LPC method:

- The placement of the analysis window.
- The length of the analysis window.
- The influence of the fundamental frequency, especially when it lies close to the first formant.
- Spectral valleys due to anti-formants in nasal sounds cause the formant estimates to deviate from their actual values.
- Rapid changes in the formant positions occur at some vowel/consonant transitions, which cannot be followed by LPC methods.

Sequential methods offer many advantages over traditional frame-based methods, since they overcome most of the problems mentioned. The main goal of the project is to eliminate the problems encountered in the above-mentioned block segmentation approach. The basic idea is to obtain a time-varying model that is unaffected by pitch pulse locations and placement or length of the analysis window. The inclusion of zeros in the current all-pole model will also be investigated.

Ting and Childers [4] designed the weighted recursive least squares algorithm with a variable forgetting factor (WRLS-VFF) to estimate the ARMA parameters of the vocal tract. An effective input estimation algorithm uses the variable forgetting factor (VFF) to decide on white noise and pulse excitation.

A summary of the advantages of using a sequential algorithm such as the WRLS-VFF, instead of a block approach, is as follows:

- The WRLS-VFF can accurately estimate and track both formant and anti-formant frequencies and their bandwidths.
- The limitations of using an analysis window of fixed length are removed by employing a variable forgetting factor.
- The influence of the fundamental frequency on the parameter estimates is eliminated with the use of an effective input estimation algorithm.
- Spectral valleys due to anti-formants in nasal and some fricative sounds can be modelled by the zeros in the pole-zero estimation model.

0-7803-1998-2/94/$4.00 © 1994 IEEE

- A slight modification to the WRLS-VFF algorithm allows it to follow rapid changes associated with some vowel/consonant transitions.

Section 2 summarises the weighted recursive least squares algorithm with a variable forgetting factor (WRLS-VFF), developed by Ting and Childers [4]. Section 3 provides the reader with practical procedures for the implementation of the proposed algorithm. A comparison between the proposed sequential algorithm and a popular frame-based method is given in section 4. The conclusion follows in section 5.

2. WRLS-VFF Algorithm

WRLS Algorithm

Suppose the unknown vocal tract system can be modelled as an ARMA process, then the output sequence y_k can be generated according to the following equation:

$$y_k = -\sum_{i=1}^{p} a_k(i) y_{k-i} + \sum_{i=1}^{q} b_k(i) u_{k-i} + u_k$$
$$k = 0 \ldots N-1 \qquad (1)$$

The input to the filter ,u_k , is a zero mean white noise process (k is a time index), and a_k and b_k are the AR and MA parameters, respectively. The orders of the AR and MA processes are p and q respectively. Prewindowing is assumed, because all data before $k=0$ and after $k=N-1$ is assumed to be zero. It is noted that the output of the filter is dependent on the input signal, u_k. Unfortunately this input signal is not available to us in speech processing. In order to provide accurate estimates of the ARMA parameters, the intended algorithm will have to include a reliable input estimation algorithm. Such an algorithm was developed by Ting and Childers [4]. For now, we assume that the estimate of u_k is a known quantity at instant k. The estimated value of u_k will be called $\hat{u}_k$. With the parameter vector (θ_k) and the data vector (ϕ_k) defined as:

$$\theta_k = [a_k(1) \;\; a_k(2) \;\; \cdots \;\; a_k(p) \;\; b_k(1) \;\; b_k(2) \;\; \cdots \;\; b_k(q)] \qquad (2)$$

$$\phi_k^t = [-y_{k-1} \;\; -y_{k-2} \;\; \cdots \;\; -y_{k-p} \;\; u_{k-1} \;\; u_{k-2} \;\; \cdots \;\; u_{k-q}] \qquad (3)$$

The output (speech signal) can then be rewritten as:

$$y_k = \phi_k^t \theta_k + u_k \qquad k = 0 \ldots N-1 \qquad (4)$$

The estimated output signal is then:

$$\hat{y}_k = \phi_k^t \hat{\theta}_k \qquad (5)$$

with the estimated parameters:

$$\hat{\theta}_k = [\hat{a}_k(1) \;\; \hat{a}_k(2) \;\; \cdots \;\; \hat{a}_k(p) \;\; \hat{b}_k(1) \;\; \hat{b}_k(2) \;\; \cdots \;\; \hat{b}_k(q)] \qquad (6)$$

The estimation error is defined as:

$$e_k = y_k - \hat{y}_k \qquad (7)$$

Define the cost function (or weighted recursive least squares criterion) for the prewindowed case, and introduce a weighting factor (or forgetting factor):

$$E_{N-1}(\theta) = \sum_{k=0}^{N-1} \lambda^{N-1-k} |e_k|^2 \qquad (8)$$

By working through the normal procedure for deriving the RLS algorithm [6], the estimate for updating the parameters can be obtained as:

$$\hat{\theta}_k = \hat{\theta}_{k-1} + K_k \left(y_k^* - \phi_k^H \hat{\theta}_{k-1} \right) \qquad (9)$$

K_k is the gain vector in the RLS estimation.

Variable Forgetting Factor (VFF)

During the derivation of the weighted RLS with constant exponential weighting factor we assumed that $0 < \lambda \leq 1$ and $\lambda = \lambda_k = \lambda_{k-1}$ for $k = 0 \ldots N-1$.

A VFF will enable us to select a forgetting factor close to unity for stationary signals or a smaller forgetting factor for non-stationary signals. The first step will be to obtain a recursive equation for the cost function or error information based on the estimation error (e_k). Isolate the term corresponding to N-1 in the equation for the cost function.

$$E_{N-1}(\theta) = \lambda E_{N-2}(\theta) + |e_{N-1}|^2 \qquad (10)$$

Write the estimation error in terms of the gain vector (K_k) and the innovation (α_k).

$$e_k = \left(1 - \phi_k^t K_k\right) \alpha_k \qquad (11)$$

The equation for the cost function now becomes:

$$E_k(\theta) = \lambda E_{k-1}(\theta) + \left(1 - \phi_k^t K_k\right)^2 \alpha_k^2 \qquad (12)$$

Ting and Childers defined the variable forgetting factor, λ_k, so that it will compensate for the new error information at each step k. Thus we have $E_k = E_{k-1} = \ldots = E_0$. Set $\lambda = \lambda_k$ and isolate λ_k on the left hand side. Remember that $E_k = E_{k-1} = \ldots = E_0$ so that the VFF (variable forgetting factor) is given by:

$$\lambda_k = 1 - \alpha_k^2 \left(1 - \phi_k^t K_k\right)^2 \Big/ E_0(\theta) \qquad (13)$$

Notice that, when the estimation error ($e_k << E_0$) is small, the value of λ_k will be close to unity. For a large estimation error the value of λ_k will be smaller than unity, implying a shorter memory length. This will allow faster tracking in non-stationary environments.

White noise & pulse input algorithm

When deriving the classical RLS algorithm it was assumed that the input signal (u_k) to the filter is a zero mean, white, gaussian noise process. This is however not true when modelling the speech process. Two input models are commonly used for modelling speech. A pulse input signal is generally assumed for vowel sounds and a white noise input signal for generating fricative sounds. Define the symbol u_k^w to represent a white noise input signal and u_k^p for a pulse input sequence. Thus the total resulting input signal is:

$$u_k = u_k^w + u_k^p \qquad (14)$$

With the estimation error defined as follows:

$$e_k = y_k - \hat{y}_k - \hat{u}_k \qquad (15)$$

The new equation for updating the parameter vector is:

$$\hat{\theta}_{N-1} = \hat{\theta}_{N-2} + K_{N-1}\left(y_{N-1}^* - \phi_{N-1}^H \hat{\theta}_{N-2} - \hat{u}_{N-1}^{*p}\right) \qquad (16)$$

Note that by subtracting the pulse input signal from the new estimation for the parameters, the influence of pitch pulses can be removed. Miyanaga *et al.* [2] showed that the magnitude of the pulse is approximately the same as that of the innovation. By using the VFF, a decision can be made on the choice of the excitation source for the vocal tract.

If $\lambda_k < \lambda_0$, a pulse input is assumed so that $\hat{u}_k^p = y_k - \phi_k^t \hat{\theta}_{k-1}^*$ and $\hat{u}_k^w = 0$. The white noise input is selected when $\lambda_k \geq \lambda_0$ by using the method of Morikawa and Fujisaki [3] so that $\hat{u}_k^p = 0$ and $\hat{u}_k^w = y_k - \phi_k^t \hat{\theta}_k^* = \alpha_k\left(1 - \phi_k^t K_k\right)$.

A fractal algorithm

In applying the WRLS-VFF to continuous speech it became necessary to develop a way to detect voiced/unvoiced jumps in the speech signal. A discontinuity is defined as a place in the speech signal where the WRLS-VFF will loose track of the signal.

One way to detect these instances is to count the zero crossings in the original speech signal. The start of a region where the count is high can then be defined as a voiced/unvoiced boundary.

The proposed method is based on work done by Boshoff [1]. The idea is to determine the local fractal dimension of the sampled speech signal by using a fast box count algorithm. A value for the fractal dimension greater than a predefined threshold indicates a discontinuity. The complexity of the box count algorithm compares favourably with that of zero crossing rate [1]. A further advantage is that the mean of the signal is not needed in the computation as in the case of zero crossing rate.

Figure 1 shows part of a speech sentence ("Even my sense of humour..."). The corresponding fractal dimension, as estimated by using the above fast box count technique, is shown in figure 2.

From the results it is clear that:

- The fractal dimension becomes a value close to two during silent parts in the speech signal. See for instance the beginning of the sentence.

- Unvoiced sounds like the /s/ in "sense", the /f/ in "of" force the fractal dimension up.

By defining the threshold value as 1.6, it is seen that all the abovementioned unvoiced sounds and silences will be discovered.

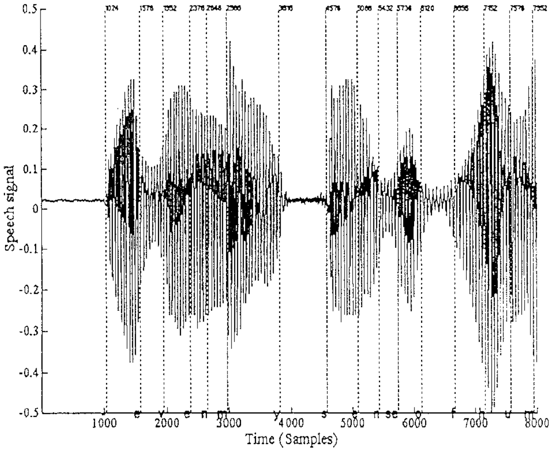

Figure 1 Speech signal of "even my sense of hum-".

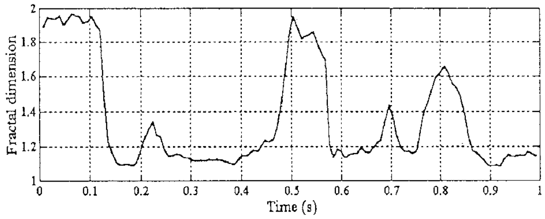

Figure 2 Fractal dimension of "even my sense of hum".

3. Implementation

- Various order determination techniques have been proposed over the years. The interested reader is pointed to work done by Akaike (FPE, AIC), Rissanen (MDL) and Parzen (CAT).
- According to Ting and Childers [4] and Haykin [6] the initial value of the covariance matrix may be set to $P_0 = \sigma I$, with σ a large positive number. A too small value for σ will slow down the rate of convergence. Morikawa and Fujisaki [3] showed that the convergence properties are not significantly affected, as long as P_0 is large compared to the variance of the source signal, ϕ_k.
- The initial estimation of the parameter vector may be set to zero ($\hat{\theta}_0 = 0$).
- The error information, E_0 (sum of the estimation errors), can be calculated before the algorithm is started. Ting and Childers [4] suggested the use of a LPC method on a frame or two to determine a suitable value. When extracting parameters of a large speech database, this is not a practical option. After testing the WRLS-VFF on speech by using a constant value for the error information, we noticed that tracking deteriorated in the unvoiced regions. The problem was that the specific constant value of E_0 was too high for tracking the low energy signals. Although the choice of E_0 could be perfect for the voiced (and thus higher energy) speech, it caused λ_k to be very close to unity when the estimated error (e_k) becomes small. The reader might argue that this is exactly what is desired - a longer memory during times where the estimation error is small. The magnitude of the estimation error is however related to that of the speech signal being followed. Thus, for a softly pronounced part of the speech the estimation error would be less than would be the case if the same segment is spoken in a louder voice. The above fact lead the author to the following heuristic to determine the value of E_0 at each time increment:

  $$E_k = |e_k|^2 + 7 \times G_k$$

  G_k corresponds to the standard deviation of the estimated white noise input signal. The constant multiplier of seven was determined experimentally. The value of G_k can be computed recursively over a fixed length sliding window. The choice of the window length is important. If the window is too long, the gain of the filter will vary slowly with time, and fluctuations over voiced/unvoiced regions may be missed. On the other hand, if the window is too short, peaks might occur in the gain sequence as a result of pitch pulses in the voiced regions. In our tests on real speech we chose the window length to be at least two pitch periods in length.

- A minimum value for the VFF is defined to prevent the memory of the algorithm from becoming too short:

  $$\lambda_{\min} = \frac{2(p+q)-1}{2(p+q)}$$

  If $\lambda < \lambda_{\min}$ then set $\lambda = \lambda_{\min}$. This value of $\lambda_{\min}$ corresponds to a memory length of $2(p+q)$ samples, which is the minimum that is required for convergence of the RLS algorithm [6].

- The threshold value for detecting different input signals was determined experimentally. A value of $\lambda_0 = \lambda_{\min} + .01$ is used throughout the rest of this document.

Summary of the proposed algorithm:

Innovation: $\alpha_k = y_k - \phi_k^t \hat{\theta}_{k-1}^*$

Update gain vector: $K_k = \dfrac{P_{k-1}\phi_k}{\left(\lambda + \phi_k^H P_{k-1}\phi_k\right)}$

Update forgetting factor:

$$\lambda_k = 1 - \alpha_k^2 \left(1 - \phi_k^t K_k\right)^2 \Big/ E_0(\theta)$$

Input estimation:

If $\lambda_k < \lambda_0$ (Pulse input)

$$u_k^w = 0$$

$$u_k^p = y_k - \phi_k^t \hat{\theta}_{k-1}^*$$

If $\lambda_k \geq \lambda_0$ (White noise input)

$$u_k^w = \alpha_k \left(1 - \phi_k^t K_k\right)$$

$$u_k^p = 0$$

Update parameters:

$$\hat{\theta}_k = \hat{\theta}_{k-1} + K_k\left(y_k^* - \phi_k^H \hat{\theta}_{k-1} - \hat{u}_k^{*p}\right)$$

Update covariance matrix:

$$P_k = \lambda^{-1}\left[P_{k-1} - K_k \phi_k^H P_{k-1}\right]$$

Constant Fractal limit:

$$\text{If}\,(FD_k > F_{limit})\ \&\ (FD_{k-1} \leq F_{limit})$$

Choose the fractal limit so that:

$$1 < F_{limit} < 2$$

FD_k is the fractal dimension of the speech signal at time instant k.

The Error information is variable according to:

$$E_k = \left|e_k\right|^2 + 7 \times G_k$$

The value of G_k corresponds to the standard deviation of the estimated white noise input signal.

The block diagram of the WRLS-VFF is shown in figure 3.

4. Experiments & Results

Pitch pulse cancellation

In many high pitched voices, like those of children or some women, the fundamental frequency (pitch pulse frequency) of the speech is close to the frequency at which the first formant occurs. In these cases the normal LPC method cannot determine the frequency of the first formant accurately, as shown by Miyanaga *et al.* [2]. The WRLS-VFF estimates the input signal to the vocal tract and can thus cancel its influence on the estimated parameters.

This was shown in an experiment by varying the pitch pulse frequency from 100Hz to 225Hz when the first formant lies at 250Hz. An AR-order of 6 was used for the LPC method of Marple. The ARMA order was p=6 and q=2 (poles and zeros respectively) in the proposed method. The results are shown in table 1.

F_0 (Hz)	1st formant freq. (Hz)	MARPLE p=6	Proposed Method (ARMA)
100	250	249	250
125	250	252	250
150	250	263	250
175	250	235	250
200	250	223	250
225	250	235	250
Average Error		12.2 %	0 %

Table 1 Cancelling the influence of pulse excitation.

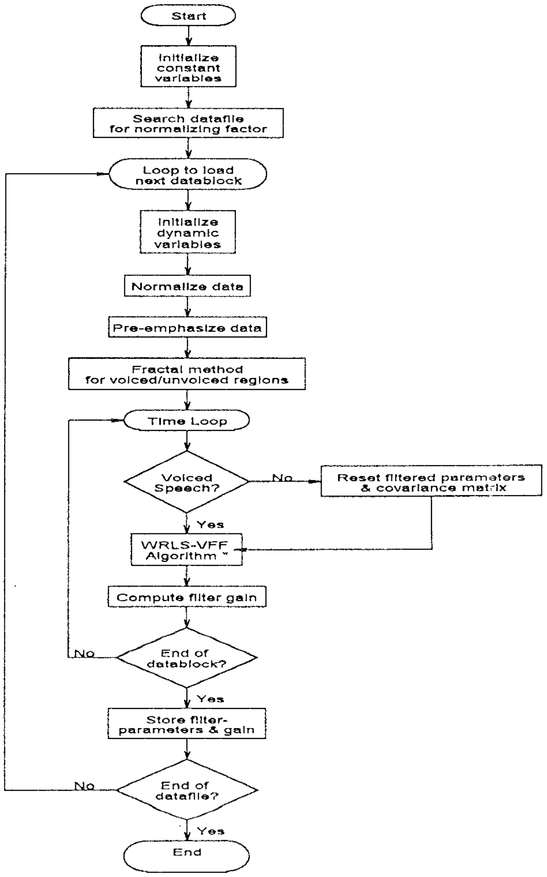
Start
Initialize constant variables
Search datafile for normalizing factor
Loop to load next datablock
Initialize dynamic variables
Normalize data
Pre-emphasize data
Fractal method for voiced/unvoiced regions
Time Loop
Voiced Speech?
No
Reset filtered parameters & covariance matrix
Yes
WRLS-VFF Algorithm
Compute filter gain
End of datablock?
No
Yes
Store filter-parameters & gain
End of datafile?
No
Yes
End

Figure 3 Blockdiagram of the proposed method.

Formant tracking

Two spectrograms of real speech, first analysed with the block technique of Marple and then by using the proposed method are shown in figures 4 and 5 respectively.

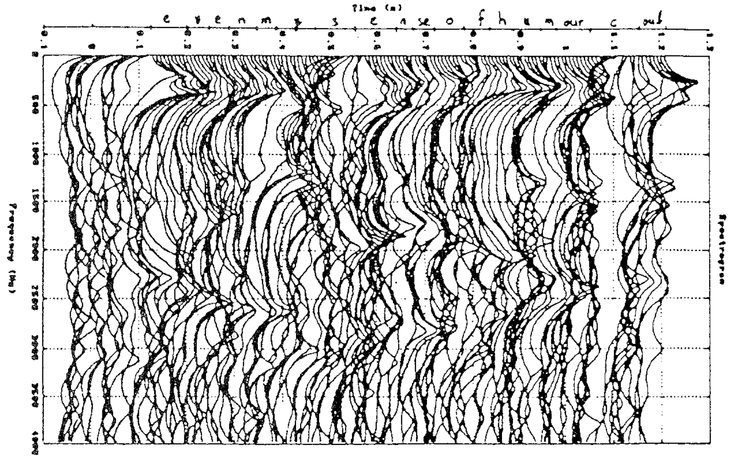

Figure 4 Estimated spectrogram for real speech signal (MARPLE technique).

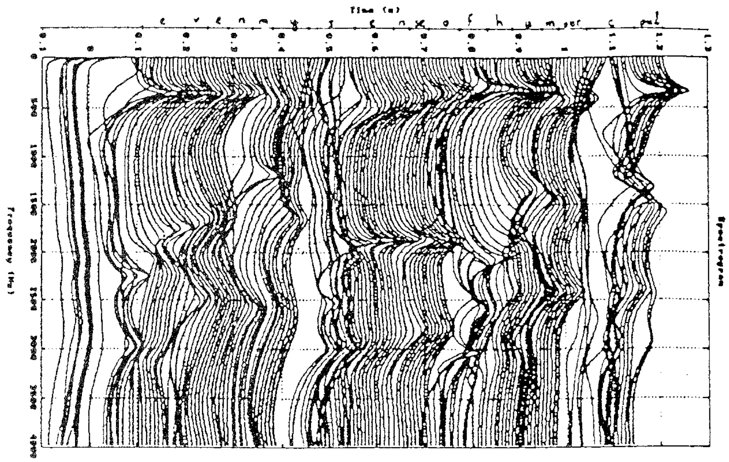

Figure 5. Estimated spectrogram for real speech signal (Proposed technique).

5. Conclusion

An adaptive method for the estimation of speech parameters was implemented.

- An RLS-based algorithm in a system identification situation was used. An effective input estimation algorithm [4] laid the foundation for eliminating the effect of pitch pulses on the estimated parameters.
- Non-stationary signals can be followed faster and with greater accuracy than with previously available techniques. This is achieved by employing a variable forgetting factor which will automatically increase or reduce the effective memory of the algorithm.
- A fractal dimension estimator will find the discontinuities jumps associated with voiced to unvoiced transitions. This dramatically increases tracking in these areas.

The proposed method can, without any modifications, be applied to accurate formant/anti-formant tracking as showed in section 4. Another immediate application is when it is used as an accurate alternative to residual-based pitch extraction [5].

References

[1] Hendrik F.V. Boshoff, A fast box counting algorithm for determining the fractal dimension of sampled continuous functions, *Proceedings IEEE COMSIG*, Cape Town, 11 Sept 1992.

[2] Yoshikazu Miyanaga ,Nobuhiro Miki, Nobuo Nagai and Kozo Hatori. A speech analysis algorithm which eliminates the influence of pitch using the model reference adaptive system. In *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-30, No. 1, pp. 88-96, Feb 1982.

[3] Hiroyoshi Morikawa and Hiroya Fujisaki. Adaptive Analysis of Speech Based on a Pole-Zero Representation. In *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-30, No. 1, pp: 77-88, Feb 1982.

[4] Y.T. Ting and D.G. Childers. Speech Analysis Using the Weighted Recursive Least Squares Algorithm with a Variable Forgetting Factor. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, N.Mex , pp. 389-392, 1990.

[5] Nancy Hubing and Kyung Yoo. Exploiting Recursive Parameter Trajectories in Speech Analysis. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I-125-28, 1992.

[6] Simon Haykin. *Adaptive Filter Theory.* Second Edition, Prentice-Hall International Inc., Englewood Cliffs, New Jersey, 07632.

[7] Marple S.L. Jr., High resolution autoregressive spectrum analysis usong noise power cancelation. *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp 345-348, 1978