ON THE APPLICATION OF AR MODEL IN SEGMENTING ISOLATED-WORD SPEECH SIGNALS

Xi Xiao, D. Nandagopal & D.A.H. Johnson
Department of Electrical & Electronic Engineering
University of Melbourne
Parkville, 3052
Victoria
Australia

ABSTRACT

In this paper, we present a segmentation method which uses AR modelling of the spectrum of the fullwave rectified speech signal. The FFT of the model coefficients yields a smoothed time domain signal with well defined minima, which locate the segment boundaries. The method is robust in the presence of noise. It is a useful first step in speech processing to segment speech signals into sub-frames which can be treated as time-invariant (stationary) processes.

THE SEGMENTATION ALGORITHM

Let s(n) be the sequence of speech samples representing a word and let P(n) be a modified version of s(n) where

$$P(n) = abs[s(n)]$$

P(n) can be regarded as the real power spectrum of some 'pseudo' time domain signal. The inverse Fourier transform of P(n) yields a complex 'pseudo' time domain signal, p(n). p(n) is then AR modelled using LPC of order say 26, and the spectrum of the model, $\hat{P}(n)$, found. This 'pseudo' frequency domain corresponds to the real time domain and can therefore be related directly to the original speech signal s(n). In particular the real part of $\hat{P}(n)$ is a smoothed version of the signal and has well defined minima which can be used as the segment boundaries.

FORMULATION

Let P(n) be the finite length fullwave rectified signal ($o < n < N - 1$) and x(k) be the 'pseudo' time domain signal such that

$$x(z) = \sum_{k=0}^{w-1} x(k)\, z^{-k} \qquad (1)$$

and

$$P(n) = x(z) \mid z = e^{j\frac{2\pi n}{N}} = x\left(e^{j\frac{2\pi n}{N}}\right)$$

$$= \sum_{k=0}^{N-1} x(k)\, e^{-jnk2\pi/N}$$

$$(o \leqslant n, k \leqslant N - 1) \qquad (2)$$

It is obvious that P(n) is the Discrete Fourier Transform of x(k) by equally evaluating x(z) on unit circle at Z-plane.

Suppose that x(k) can be modelled by Linear Predictive Coding of order P, then the estimate for x(k) is given as

$$\hat{x}(k) = - \sum_{l=1}^{P} a(l)\ x\ (k - l) \qquad (3)$$

$$(p < k \leqslant N\text{-}1)$$

where a(1) ..... a(p) are LPC constants to be determined.

The estimate error is thus

$$E(k) = x(k) - \hat{x}(k) = x(k) + \sum_{l=1}^{P} a(l)\ x\ (k - l) \qquad (4)$$

$$(p < k \leqslant N\text{-}1)$$

The constants a(1), a(2) ..... a(p) are determined by minimizing the total squared estimate error

$$E_p^2 = \min_{a(1) \ldots a(p)} \{ \sum_k (x(k) - \hat{x}(k))^2 \} \quad (5)$$

Applying Z-transform to equation (4), we obtain an estimate for X(Z),

$$\hat{X}(z) = \frac{E_p}{1 + \sum_{l=1}^{p} a(l) z^{-l}} \quad (6)$$

Now, let $Z = e^{jnk2\pi/N}$, n = 0, 1, ..., N-1,

We have

$$\hat{X}(e^{jn \cdot 2\pi/N}) = \frac{E_p}{1 + \sum_{l=1}^{p} a(l) e^{-jl \cdot n \cdot 2\pi/N}} \quad (7)$$

compare equation (2) and (7), an estimate for P(n) is given as

$$\hat{P}(n) = \hat{X}(e^{jn \cdot 2\pi/N}) = \frac{E_p}{1 + \sum_{l=1}^{p} a(l) e^{-jl \cdot n \cdot 2\pi/N}} \quad (8)$$

From equation (6) and (8), we can now describe the problem as below:

Given a finite length of signal P(n) with N sample points, find a polynomial of Z with an order P much lower than the sample points such that the reciprocals of the values of the polynomial on unit circle on Z-plane are the best estimate of the whole span of P(n) under the criteria that the spectrum of P(n) can be modelled by LPC constants a(1), a(2)... a(p) as defined in equation (3).

A relatively high LPC order is used in the above procedure to ensure that the estimate error is small. As will be shown in next section, this algorithm is robust in detecting the onsets of both speech signal and the voiced parts, but oversegmentation is brought up because of high LPC order.

The next step is to measure the characteristics distance between every two adjacent segments. If the distance is smaller than an empirical threshold, these two segments are merged into one.

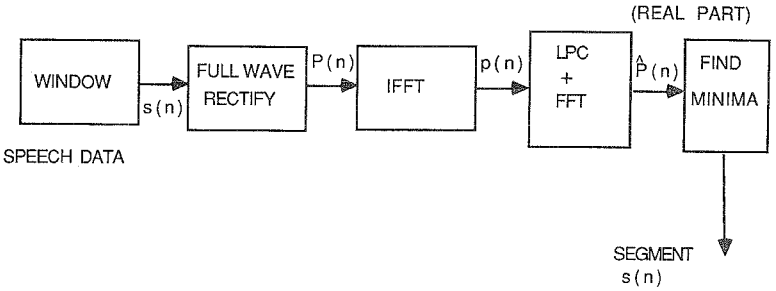
Now let's consider two adjacent segments R and T, and denote the signal within these two segments as X_R of lengths N_R and X_T of lengths N_T and, respectively. The statistic

$$\chi^2 = (\hat{\alpha}_R - \hat{\alpha}_T)\left\{\sigma_R^{*2}(N_R' D_R)^{-1} + \sigma_T^{*2}(N_T' D_T)^{-1}\right\}^{-1}(\hat{\alpha}_R - \hat{\alpha}_T)$$

has an (asymptotic) Xp^2 distribution under the null hypothesis that the two sequences emanate from series with the same LPC vector α.

(Refer to Peter De Souza and Peter J. Thomson for symbol definition and the proof) [1].

When X^2 smaller than a threshold, the null hypothesis is accepted, or, the two segments are merged.



RESULTS

Figure 1 shows a speech data segment (Windowed-Hamming), with white noise added resulting in a signal to noise ration of 12 db. Figure 2 is the 'pseudo' time domain signal obtained by calculating the IFFT of the fullwave rectified signal in Figure 1. An LPC analysis of order 26 on this 'pseudo' time domain signal gives tis 'pseudo' LPC spectrum estimate. In particular, the real part of this spectrum, depicted in Figure 3, is a smoothed version of the signal s(n). The vertical lines superimposed on Figure 3 indicate the minima between peaks. These peaks are LPC estimates of the dominant peaks in Figure 1 and thus, the minima correspond approximately to the zero crossings and are used as segment boundaries. The final segmentation is shown in Figure 4. Table 1 lists the segment boundaries extracted from the minima positions.

As shown in Figure 4, the algorithm detects the onset of both the plosives and the voiced portions quite well at sample 213 and 523 respectively (the segment from sample 0 to 213 corresponds to intentionally introduced white noise). Similar segments from sample 785 to 1631 result from the periodical structures of the voiced portions.

Figure 5 is a clear speech signal of word/TWO/ sampled at 10 kHZ. The number of sample points is 4096 including some silence on the front and end of the speech. An LPC of order 26 was used to estimate the contour shown in Figure 6. 16 valleys are found in the contour and are used as segment boundaries represented by the vertical lines in Figure 5. Table 2 lists the LPC distance between every two adjacent segments. The distances measured at boundary 1, 3 and 15 are significantly larger than others and thus, these boundaries are finally chosen.

CONCLUSION

The AR model used in this segmentation method provides stable estimates of the segment boundaries in the presence of noise. The boundaries occur at, or very near the zero crossings of the signal. The number of segments is primarily determined by the order chosen for the LPC process and the oversegmentation can be reduced by LPC distance measures.

REFERENCES

[1] Peter De Souza and Peter J. Thomson, "LPC Distance Measures and Statistical Tests with Particular Reference to the Likelihood Ratio", IEEE Trans. Acconst., Speech, Signal Processing, Vol. ASSP - 30, pp 304-315, April 1982.

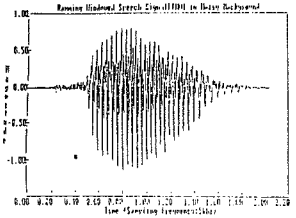


Fig. 1

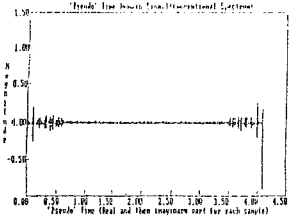


Fig. 2

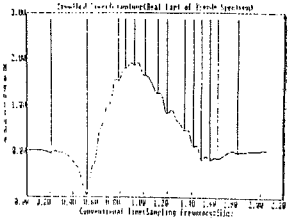


Fig. 3

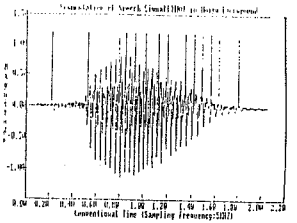


Fig. 4

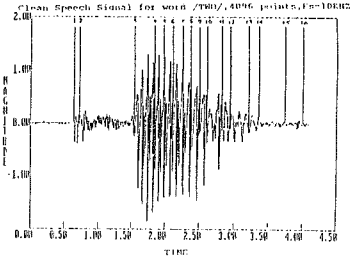


Figure 5

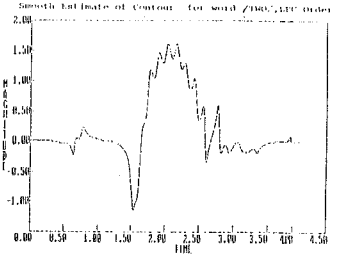


Figure 6.

Table.1. Segment Boundary List

213	523	785	844	922	1015	1122
1179	1343	1417	1484	1561	1631	1806

Table 2. Distance Measures Between Adjacent Segments

I	Boundary	Distance	I	Boundary	Distance	I
I	1	89.48	I	9	15.04	I
I	2	18.97	I	10	10.26	I
I	3	492.34	I	11	34.53	I
I	4	9.16	I	12	11.38	I
I	5	3.11	I	13	45.88	I
I	6	12.15	I	14	25.10	I
I	7	12.69	I	15	94.10	I
I	8	13.43	I	16	38.67	I