AD-770 298

THE USE OF A TWO-POLE LINEAR PREDICTION MODEL IN SPEECH RECOGNITION

John Makhoul, et al

Bolt Beranek and Newman, Incorporated

Prepared for:

Advanced Research Projects Agency

September 1973

BBN Report No. 2537
A.I. Report No. 7

THE USE OF A TWO-POLE LINEAR PREDICTION MODEL IN SPEECH RECOGNITION*

JOHN MAKHOUL
JARED WOLF

SEPTEMBER 1973

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Defense Supply Service - Washington under Contract No. DAHC-71-C-0088. Order No. 1967. 15-

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce for sale to the general public.

*An earlier version of this paper was presented at the 85th meeting of the Acoustical Society of America in Boston, Ma., April 10-13, 1973. Abstract in: J. Acoust. Soc. Am., Vol. 54, No. 1, p.339, July 1973 (A).

ABSTRACT

In speech recognition applications, it is often desirable to make a gross characterization of the shape of the spectrum of a particular sound. The autocorrelation method of linear predi tion analysis leads to an all-pole approximation to the signal spectrum. Hence an LPC analysis using two poles produces one possible gross characterization. The two poles are computed as the roots of a quadratic equation whose coefficients are the linear prediction parameters, which are simple functions of the autocorrelation coefficients R_0, R_1, and R_2. The poles are either both real or form a conjugate pair in the z plane. This fact, together with the exact positions of the poles, is particularly useful in describing certain gross characteristics of the spectrum. The spectral dynamic range of the two-pole spectrum and the normalized minimum error are suggested as more suitable substitutes for the two-pole bandwidths in interpreting the information supplied by the model for the purpose of spectral characterization.

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Bolt Beranek and Newman Inc.
5Ø Moulton Street
Cambridge, Mass. Ø2138

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
"The Use Of A Two-Pole Linear Prediction Model In Speech Recognition"

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report

5. AUTHOR(S) (First name, middle initial, last name)
John I. Makhoul
Jared J. Wolf

6. REPORT DATE
September 1973

7a. TOTAL NO. OF PAGES
21

7b. NO. OF REFS
2

8a. CONTRACT OR GRANT NO.
DAHC-71-C-ØØ88
15-

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
BBN Report No. 2537
A.I. Report No. 7

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

In speech recognition applications, it is often desirable to make a gross characterization of the shape of the spectrum of a particular sound. The autocorrelation method of linear prediction analysis leads to an all-pole approximation to the signal spectrum. Hence an LPC analysis using two poles produces one possible gross characterization. The two poles are computed as the roots of a quadratic equation whose coefficients are the linear prediction parameters, which are simple functions of the autocorrelation coefficients R_0, R_1, and R_2. The poles are either both real or form a conjugate pair in the z plane. This fact, together with the exact positions of the poles, is particularly useful in describing certain gross characteristics of the spectrum. The spectral dynamic range of the two-pole spectrum and the normalized minimum error are suggested as more suitable substitutes for the two-pole bandwidths in interpreting the information supplied by the model for the purpose of spectral characterization.

Unclassified
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech Analysis						
Speech Recognition						
Linear Prediction						
Spectral Models						
All-Pole Models						

Unclassified
Security Classification

TABLE OF CONTENTS

I. INTRODUCTION

In the analysis of speech signals it is often desirable to make gross characterizations of speech spectra. This is useful in speech recognition for the purposes of segmentation as well as the general classification of the different sounds. In the past, gross spectral characterizations have been obtained by computing parameters that depended on the energy contained in different regions of the spectrum. Other methods have employed measurements of zero crossing rates and zero crossing distances. In this paper we describe a method for the gross characterization of speech spectra using a simple linear prediction model.

It is well known that in linear prediction the signal spectrum is modeled or approximated by an all-pole spectrum [1,2]. The number of poles in the approximate spectrum is arbitrary and is set to different values depending on the sampling frequency of the signal and on the particular application. For example, for a 10 kHz sampled signal, a 14-pole model is now common for the purposes of spectral envelope estimation and formant extraction. However, if we assume that our gross characterization is to consist in the poles themselves, then a 14-pole model contains too much information, and it takes a relatively long time to compute, since it involves finding the roots of a 14th degree polynomial. We have found that a two-pole model is optimal in terms of three things:

(1) ease of computation,
(2) adequacy of representation,
(3) ease of interpretation.

These three points are discussed in the following three sections.

II. TWO-POLE MODEL

The transfer function of the two-pole model is given by

$$H(z) = \frac{A}{1-a_1z^{-1}-a_2z^{-2}} \tag{1}$$

where a_1 and a_2 are the predictor coefficients, and A is a gain factor.

The coefficients a_1 and a_2 can be computed using either the autocorrelation or covariance method of linear prediction [1]. Although much of the discussion in this paper also applies to the covariance method, we shall work exclusively with the autocorrelation method. In the latter method, a_1 and a_2 are solutions to the two equations

$$a_1R_0 + a_2R_1 = R_1 \tag{2a}$$

$$a_1R_1 + a_2R_0 = R_2 \tag{2b}$$

where R_i is the ith autocorrelation coefficient of the signal. The solution of (2) gives:

$$a_1 = \frac{r_1(1-r_2)}{1-r_1^2} \qquad (3a)$$

and

$$a_2 = \frac{r_2-r_1^2}{1-r_1^2} \qquad (3b)$$

where

$$r_i = \frac{R_i}{R_0}, \quad i=0,1,2, \qquad (4)$$

are the normalized autocorrelation coefficients with the property that $|r_i|\leq 1$. The gain factor A can be shown to be equal to

$$A = \sqrt{R_0 V} \qquad (5)$$

where

$$V = 1-a_1 r_1 - a_2 r_2 \qquad (6)$$

is the normalized minimum error [1].

The poles of H(z) in the z-plane are simply the roots of the quadratic polynomial $1-a_1 z^{-1} - a_2 z^{-2}$ in the denominator of (1):

$$z_{1,2} = \frac{a_1}{2} \pm \sqrt{\frac{a_1^2}{4} + a_2}\ . \qquad (7)$$

Depending on the values of a_1 and a_2, the poles z_1 and z_2 are either both real or form a complex conjugate pair. Conversion of the poles to the s-plane is accomplished by setting

$$z = e^{sT} = e^{(\sigma+j\omega)T} = e^{2\pi T(h+jf)} \qquad (8)$$

where T is the sampling interval, f is the frequency of the pole,

and h is defined to be the half-bandwidth of the pole.

If a pole is at $z = z_r + jz_i$, then:

$$f = \frac{f_s}{2\pi} \arctan \frac{z_i}{z_r} \qquad (9)$$

$$h = \frac{f_s}{4\pi} \log (z_r^2 + z_i^2) \qquad (10)$$

where $f_s = \frac{1}{T}$ is the sampling frequency.

This completes the specification of the two poles. As can be seen from the above, the computations are straightforward. Note that if the model had more than two poles, one would have to find the roots of a polynomial of degree greater than 2, which is not a straightforward task.

III. ADEQUACY OF REPRESENTATION

In this section we show that the two-pole model is adequate for representing gross characterizations of speech spectra.

The possible positions for the two poles z_1 and z_2 form four distinct cases. Figure 1 shows the four possible prototype amplitude responses for the two-pole model. Each amplitude response is computed along the unit circle from z=1 to

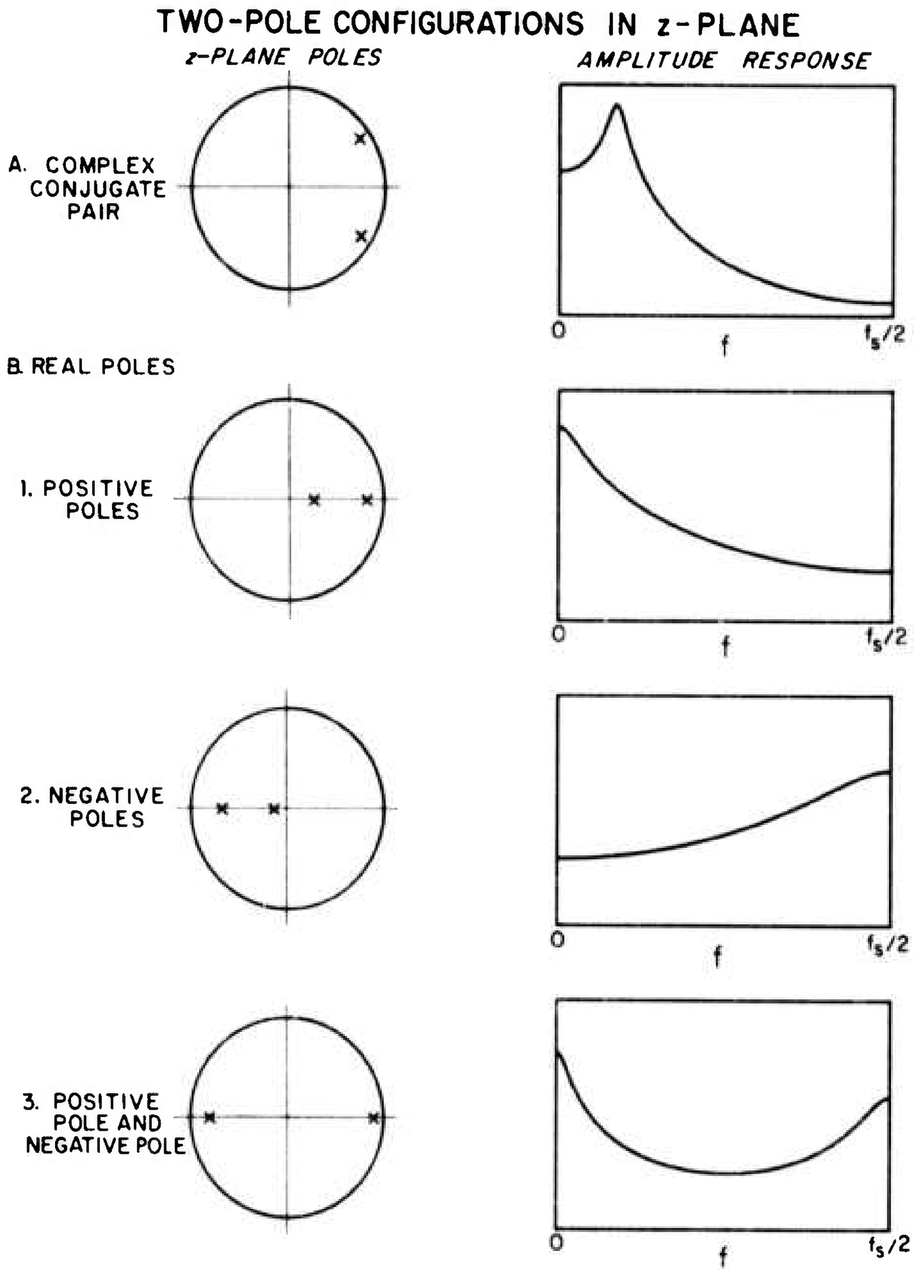
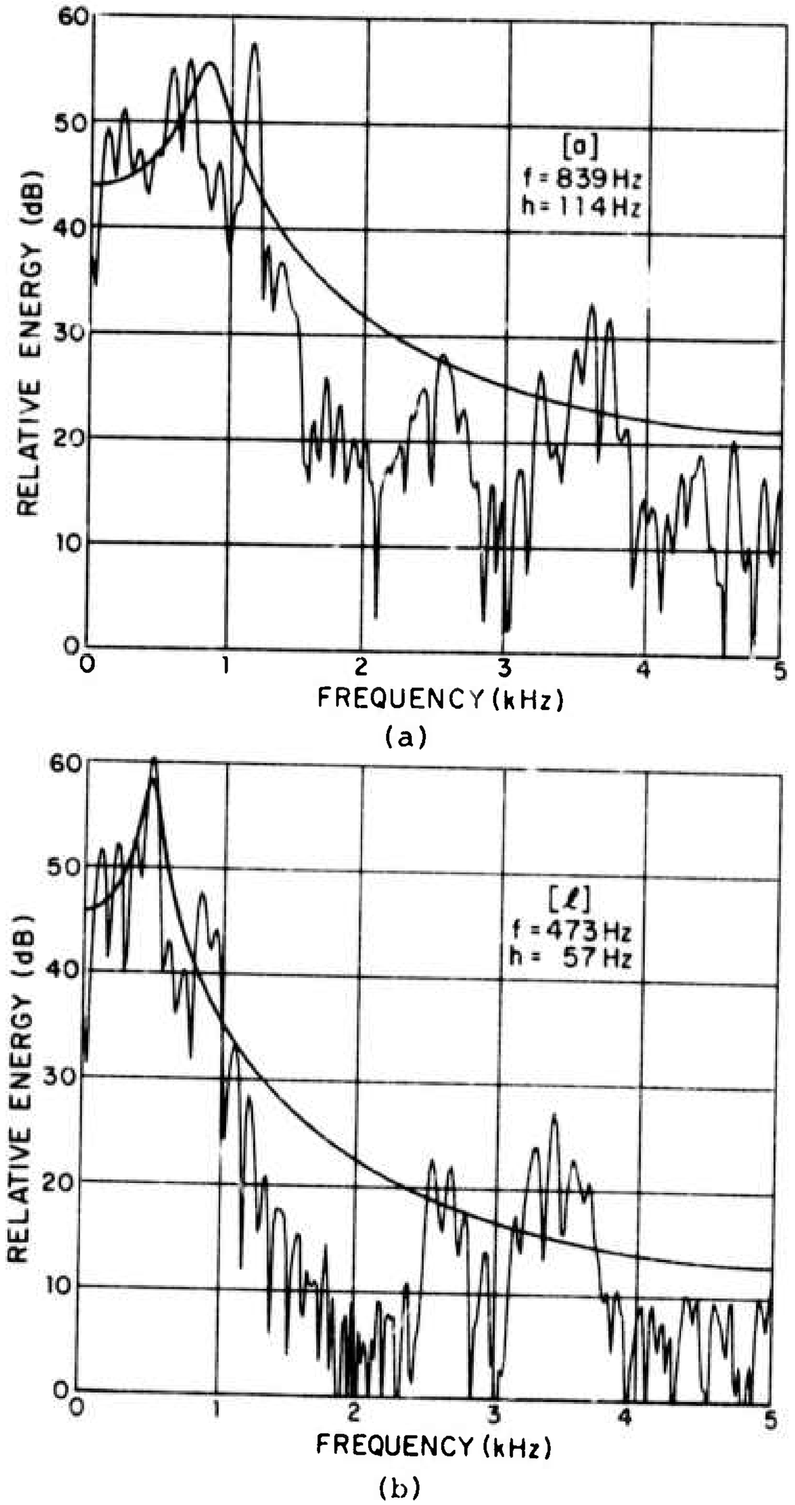

TWO-POLE CONFIGURATIONS IN z-PLANE
z-PLANE POLES
AMPLITUDE RESPONSE
A. COMPLEX CONJUGATE PAIR
B. REAL POLES
1. POSITIVE POLES
2. NEGATIVE POLES
3. POSITIVE POLE AND NEGATIVE POLE
0
f
f_s/2

FIGURE 1: The four possible confiqurations of the two-pole model and representative spectra.

z=-1, which corresponds to a plot from zero frequency to half the sampling frequency. The first case is that of the familiar complex conjugate pair. The amplitude response is completely specified by the frequency and bandwidth of one pole. For the case of real poles, there are three possibilities. The poles can be either both positive, both negative, or one positive and one negative. A positive real pole (in the z-plane) corresponds to a pole at zero frequency and indicates energy concentration at low frequencies. A negative real pole corresponds to a pole at half the sampling frequency and indicates energy concentration at high frequencies.

All four prototype cases shown in Figure 1 do occur when modeling speech spectra. In order to give a flavor of how and when these four cases occur, we present a few examples in Figures 2-6. In each of the examples, the speech signal was low-pass filtered at 4.5 kHz and sampled at 10 kHz. The two-pole spectrum (i.e. $|H(f)|^2$) is shown superimposed over the actual speech spectrum being modeled. In each case, the corresponding speech sound is shown along with the pole parameters: f represents the pole frequency and h the pole half-bandwidth in Hz. For example, Figure 2b shows the two-pole model for an example of the sound [ℓ]. The model has a pair of conjugate poles at 473 Hz with a half-bandwidth of 57 Hz. Figures 2 and 3 show



(a)

(b)

FIGURE 2: Examples of sonorant spectra modelled by complex conjugate pairs of poles. f is the pole frequency and h is the corresponding half-bandwidth.

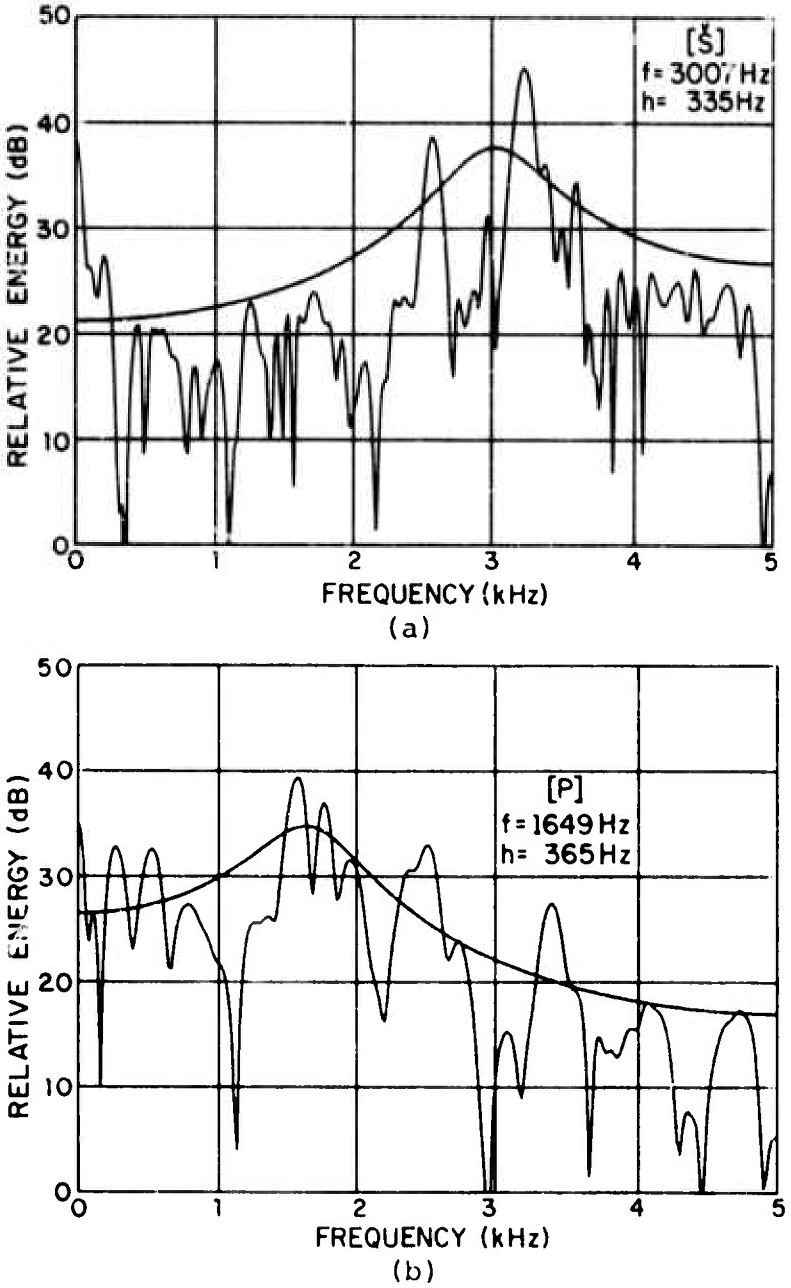


FIGURE 3: Examples of consonant spectra modelled by complex conjugate pairs of poles.

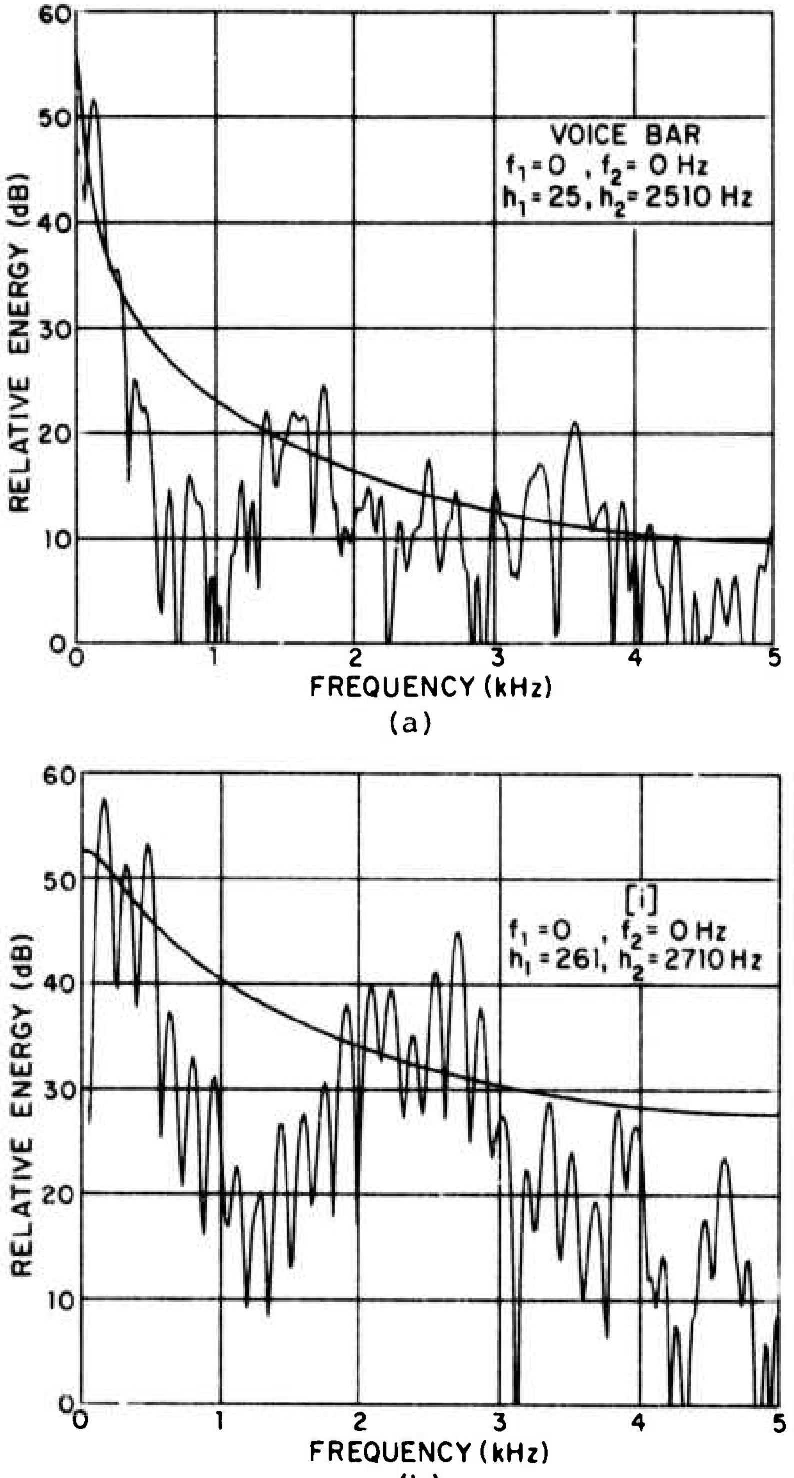
examples where the spectrum is modeled by complex conjugate pole pairs. Figure 3b corresponds to the spectrum of the burst in the plosive [p]. Note that by simply changing the frequency and bandwidth of the conjugate pair of poles, many different spectral shapes can be accomodated.

Figure 4 shows two examples where the spectrum is modeled by two positive poles in the z-plane, i.e. both poles are at zero frequency. In Figure 4b, the relatively high energy at both low and mid frequencies resulted in a model with two positive real poles, instead of a complex conjugate pair, which is more common for vowels.

Figure 5 shows two examples where the spectrum is modeled by one positive and one negative pole in the z-plane. Figure 5a corresponds to a vowel-fricative transition while Figure 5b corresponds to a voiced fricative. In both cases there is energy concentration both at low and high frequencies.

Finally, Figure 6 shows an example where the spectrum is modeled by two negative poles, i.e. both poles are at half the sampling frequency (5 kHz).

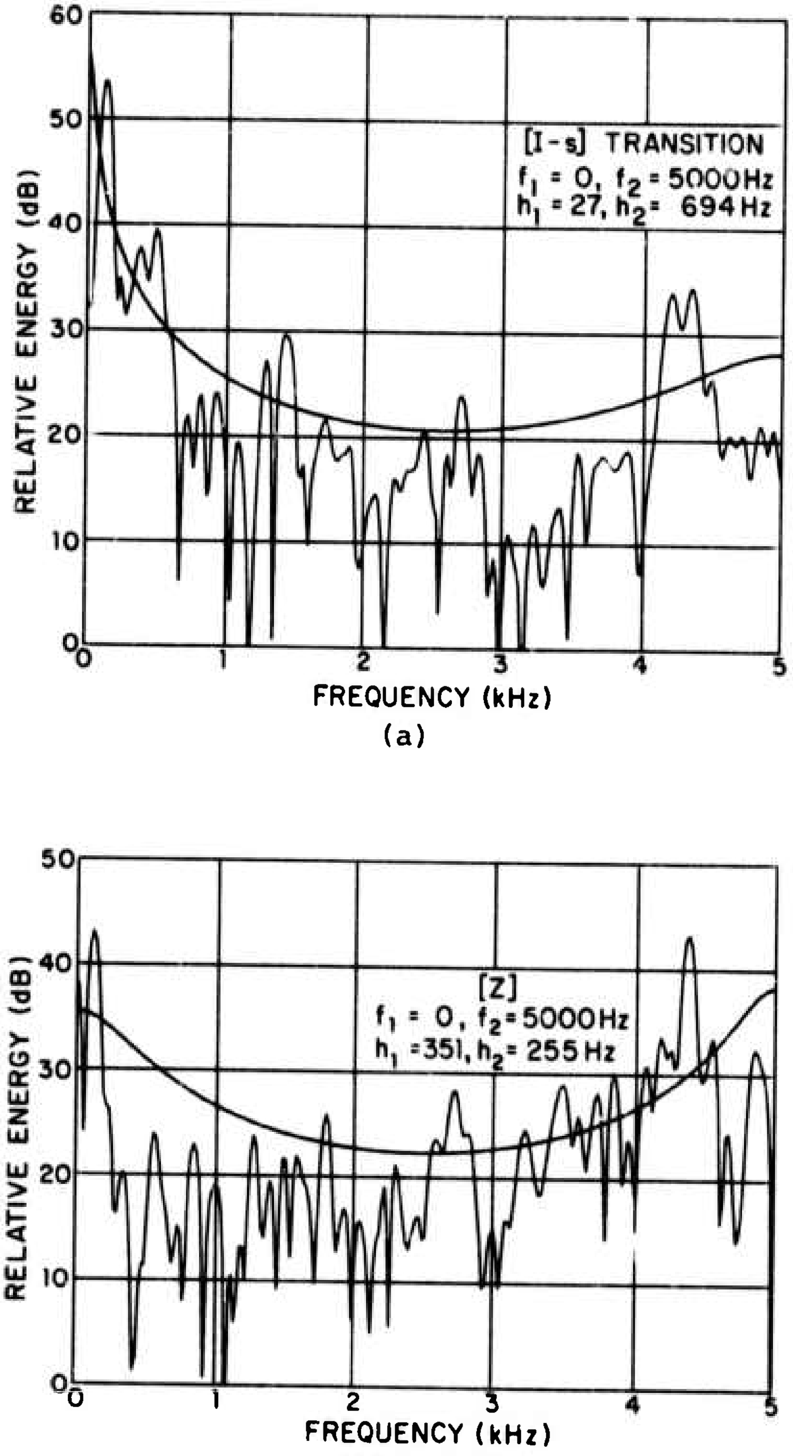
The above examples give a good indication of the adequacy of representation of the two-pole model for the gross characterization of speech spectra. Below we discuss how one interprets results of a two-pole model for segmentation and broad classification.



(a)

(b)

FIGURE 4: Examples of speech spectra modelled by two positive real poles (i.e. both poles are at zero frequency).



(a)

(b)

FIGURE 5: Examples of speech spectra modelled by one positive (0 Hz) and one negative (5 kHz) pole.

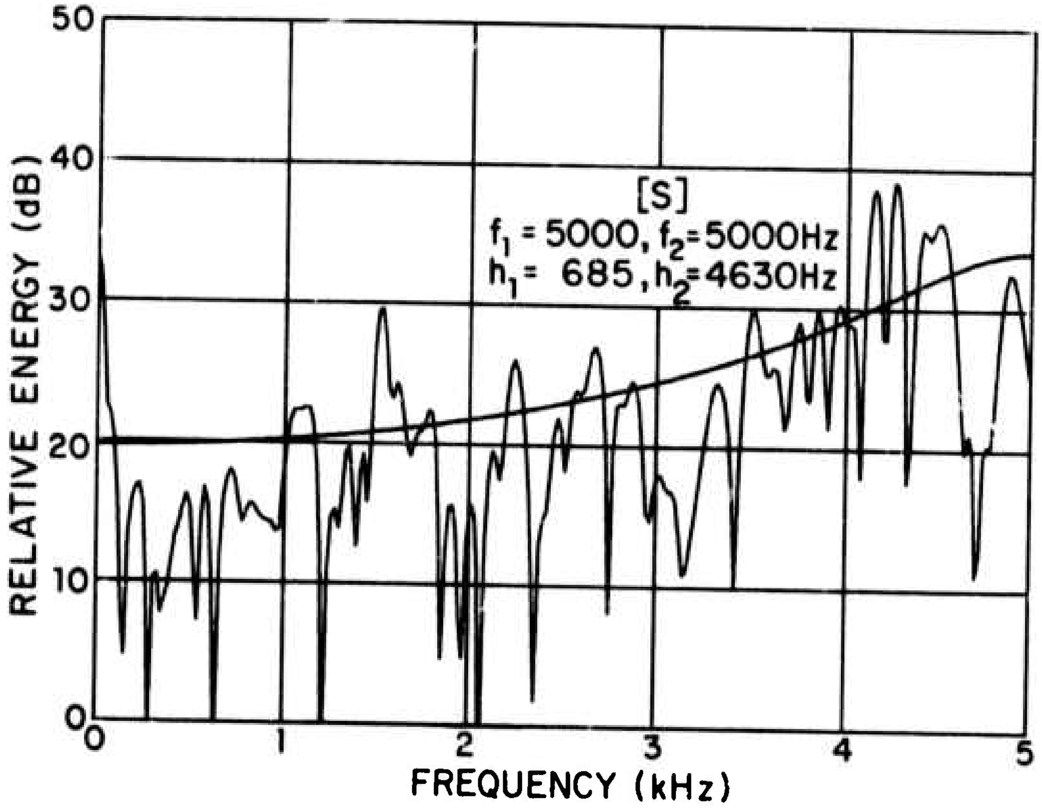


FIGURE 6: Example of a speech spectrum modelled by two negative poles (both at 5 kHz).

IV. SEGMENTATION AND BROAD CLASSIFICATION

Using the two-pole model in the recognition of continuous speech suggests that the spectral characterization described above be performed at regular closely spaced points throughout the utterance, producing a multi-parametric description of the signal. The two-pole model for each point can be represented by the frequencies and bandwidths of the two poles. This type of representation is reasonable for complex conjugate poles since there is only one frequency and one bandwidth to interpret. The frequency indicates the position of the main region of energy concentration, and the bandwidth indicates the spread of energy in that region. However, in the case of real poles, we have to deal with two possibly distinct frequencies and two bandwidths. The frequencies are always either zero or equal to half the sampling frequency, and are easily interpretable, as shown below. On the other hand, interpretation of two distinct bandwidths is far from straightforward, especially when the two frequencies are identical.

We have found that the bandwidth information can be represented in a more helpful manner in terms of the dynamic range of the two-pole spectrum and the direction or sign of its "slope". We define the spectral dynamic range to be the difference in decibels between the highest and lowest amplitude points on the two-pole spectrum. The slope of the two-pole spectrum is either positive or negative: It is positive if the energy is concentrated above the midpoint of the spectrum (2.5 kHz in our case) and negative otherwise.

From (1) and (7), it is simple to derive formulas for the two-pole dynamic range D and the sign S of the two-pole slope. There are four distinct cases.

<u>Complex conjugate poles:</u> $z_1, z_2 = a_r \pm j a_i$

Let $a = |z| = \sqrt{a_r^2 + a_i^2}$,

Then: $$S = -\,\mathrm{sign}(a_r)\ . \tag{11a}$$

If $|a_r| < \dfrac{2a^2}{1+a^2}$,

$$D = 10 \log_{10} \left[\frac{a(1+a^2+2|a_r|)}{a_i(1-a^2)}\right]^2 . \tag{11b}$$

If $|a_r| \geq \dfrac{2a^2}{1+a^2}$,

$$D = 10 \log_{10} \left[\frac{1+a^2+2|a_r|}{1+a^2-2|a_r|}\right]^2 . \tag{11c}$$

<u>Real poles:</u> $z_1 = a,\quad z_2 = b$

If sign(a) = sign(b),

$$S = -\,\mathrm{sign}(a) \tag{11d}$$

$$D = 10 \log_{10} \left[\frac{(1+a)(1+b)}{(1-a)(1-b)}\right]^2 . \tag{11e}$$

If sign(a) ≠ sign(b), let $A = |a| \geq B = |b|$,

then

$$S = -\ \mathrm{sign}(a)$$

$$D = 10 \log_{10} \frac{1}{4AB} \left[\frac{A(1+B^2) + B(1+A^2)}{(1-A)(1+B)} \right]^2 . \qquad (11g)$$

It should be clear from the above that, in the case of complex conjugate poles, the spectral dynamic range D uniquely specifies the bandwidth of the poles. For real poles, the spectral dynamic range is an intuitive substitute for bandwidth information but does not specify it uniquely. The sign of the spectral slope gives <u>additional</u> useful information only when the two poles are real with one pole at zero frequency and the other at half the sampling frequency.

The behavior of the two-pole model when applied at regular intervals to an utterance is shown in Figure 7. The utterance is "Has anyone measured nickel concentrations..." The two-pole analysis was performed at 10 msec intervals over 20 msec Hamming-windowed segments of the waveform. The pole frequencies are plotted as a single point where they are identical, and as two points for those frames where one pole is at zero and one is at 5000 Hz. Note that the scale of the frequency plot is linear only from 0 to 500 Hz, then logarithmic to 5000 Hz. Between the pole frequency and dynamic range plots, the frames in which the two-pole slope is positive are indicated.

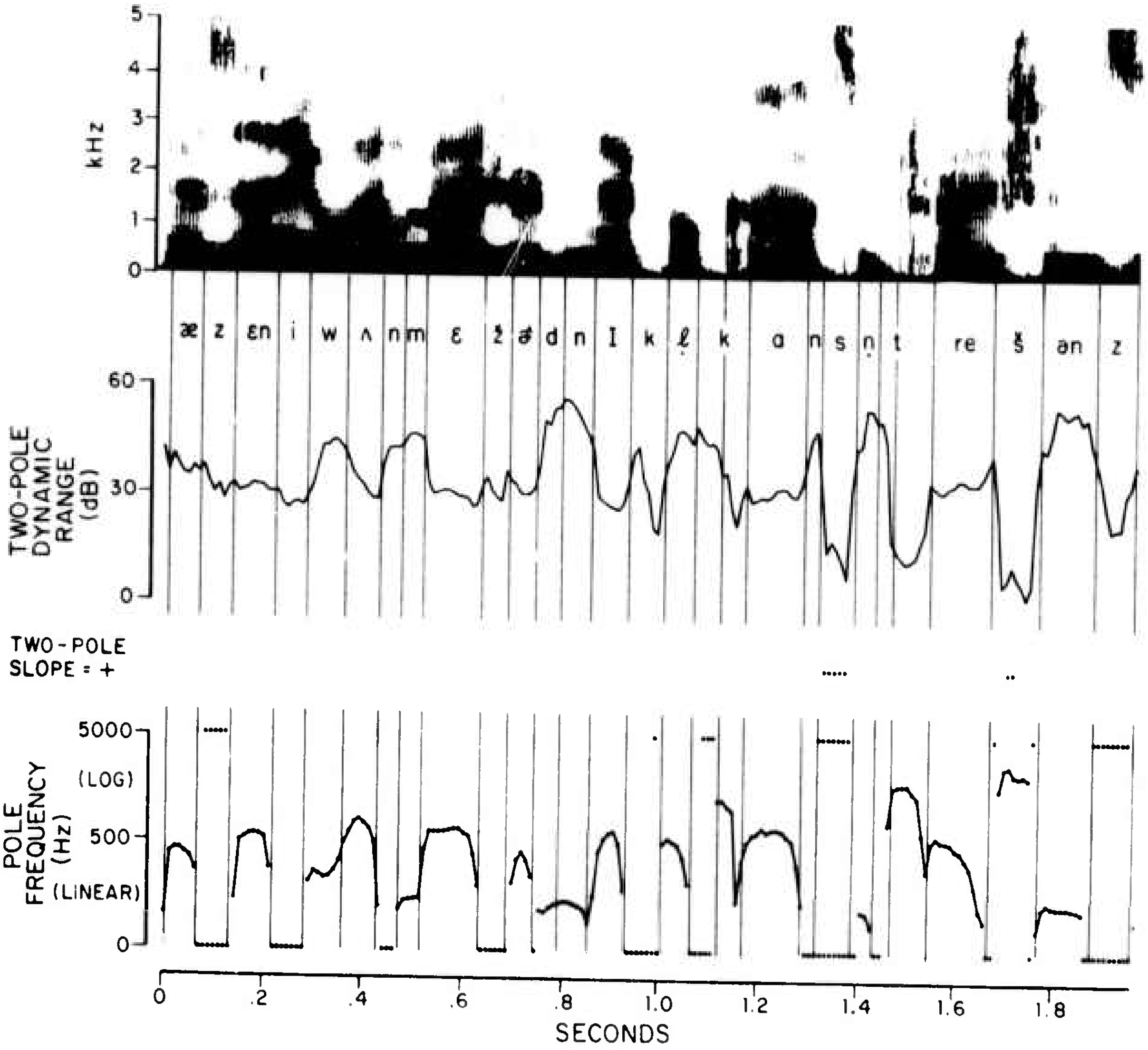


FIGURE 7: Two-pole frequency, "slope", and spectral dynamic range at 10 msec intervals in the utterance "Has anyone measured nickel concentrations..."

Many segment boundaries, particularly those where a change of manner of articulation takes place, are clearly marked by abrupt or rapid changes in the two-pole frequencies, which often switch from one of the four prototype models to another. Every such switch is a marker of spectral change, but not every one will mark a segment boundary. For example, at both the beginning and end of the [z] at t=0.08 to 0.14 seconds, there is a frame in which both poles are at 0, while the middle of the [z] has a string of frames with one pole at 0 and one at 5000 Hz. This is a common pattern of transition to and from a fricative. Of course, plosives, particularly unvoiced ones, usually show transitions corresponding to the burst-aspiration-phonation sequence, as would be expected. See the two examples of [k] around t=0.95 and t=1.10 and the [t] example around t=1.50.

Many, but not all, sonorant sequence transitions exhibit a rapid change in the two-pole frequency, which tends to follow the first formant if it is dominant, or lies between F1 and F2 if F2 is close enough to F1. See for example t=0.35, t=0.50, and t=0.85.

Not all segments with conjugate poles are sonorants, and vice versa. For example, [i] around t=0.25 is modeled with two real poles, in the way also illustrated by Figure 4b. It is also common for nasals to be modeled by two real poles, as the [n] around t=0.45, or very low-frequency conjugate poles, as the [m] immediately following.

Any occurrence of a pole at 5000 Hz indicates strident frication, or an equivalent burst, as do conjugate poles above about 1 kHz. Most examples of [s], [$\hat{s}$], and [z] will show this during at least some of their extent.

The two-pole dynamic range is quite high during nasals, because of the dominance of the low first formant. This is quite a reliable indication. Conversely the dynamic range is usually quite low during unvoiced fricatives. The measure is not quite as reliable during voiced fricatives. A positive two-pole slope is, of course, a strong indication of strident frication.

The gross characterizations of speech spectra given by the two-pole model are certainly not sufficient in and of themselves to segment and roughly label continuous speech, but they do point to a large proportion of segment boundaries. Together with other obvious measurements such as energy and voicing, they form a powerful combination for the initial stages of speech recognition.

V. AN ALTERNATIVE MEASURE TO THE SPECTRAL DYNAMIC RANGE

The two-pole dynamic range is a rather intuitive measure of one aspect of spectral shape, that is, it is easily visualized from a graph of the spectrum. A clearly related (but easier to compute) measure is the normalized minimum error V, given by (6). It can be shown that the measure V is equal to the ratio of the geometric mean of the two-pole spectrum to its arithmetic mean (see [1], pp. 109-115). It has been known for some time that the ratio of the geometric mean to the arithmetic mean is a good measure of the spread of the data. For smooth spectra (as is the case for a two-pole spectrum) the spectral dynamic range is also a good measure of the spread of the spectrum. It is not surprising, therefore, that the two measures should behave in a similar fashion. This similarity is illustrated in Figure 8, which shows 200 values of V versus D for the two seconds of continuous speech shown in Figure 7. The continuous curve also plotted in Figure 8 is that of V_m, the absolute lower bound on V for each value of D (see [1], pp. 116-120). The data points themselves fall within a very well defined region, suggesting for two-pole spectra a tighter lower bound (and also an upper bound) than the V_m versus D curve shown.

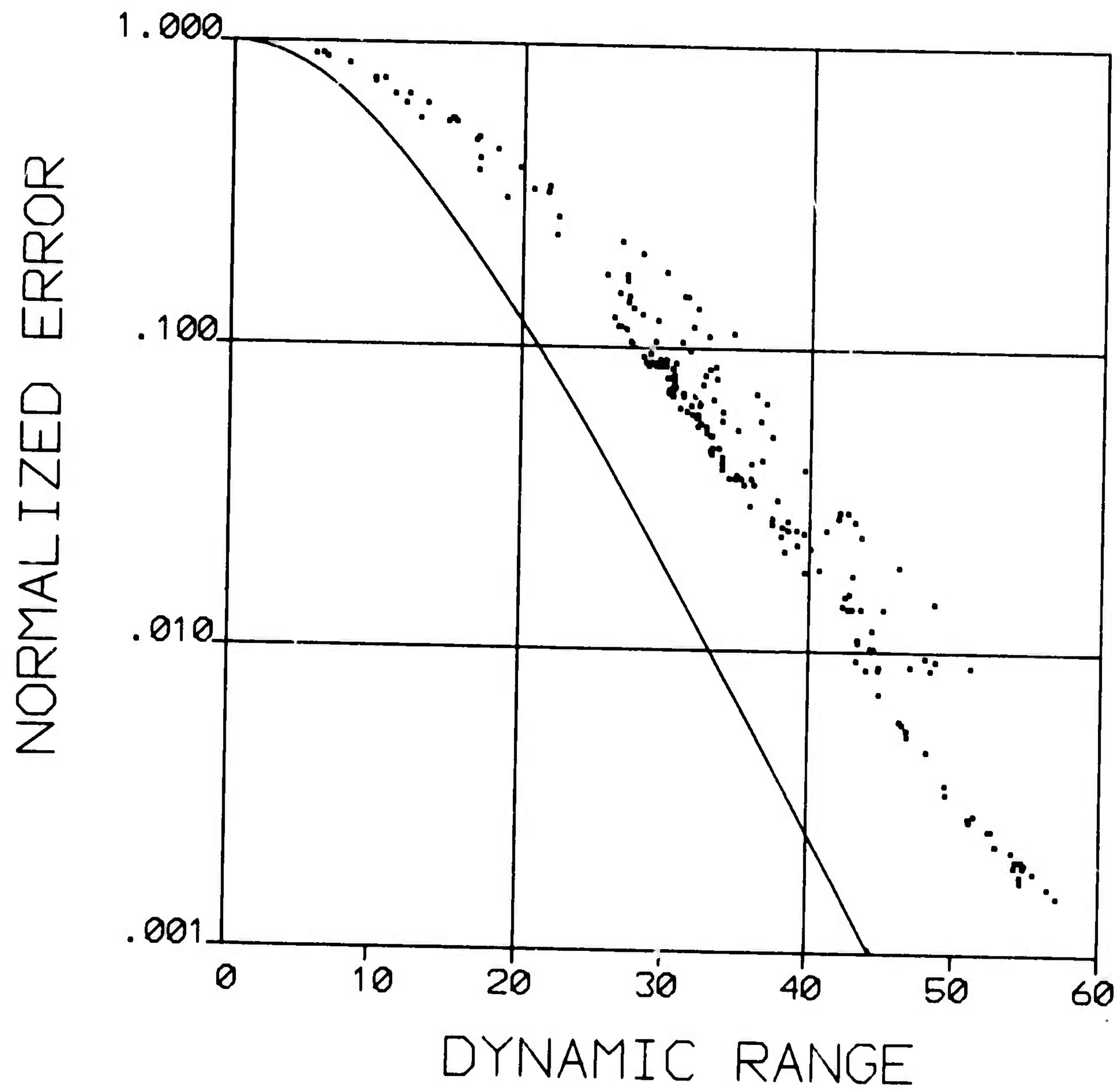


FIGURE 8: Two-pole normalized error vs. spectral dynamic range for the 200 data points in the utterance in Fiqure 7. The solid curve is V_m, the absolute lower bound on the normalized error.

Since the normalized error V is somewhat easier to compute than the spectral dynamic range D, and since it leads to very similar results, our suggestion here is that it might be preferable to use V in actual implementations.

REFERENCES

[1] Makhoul, John I., and Jared J. Wolf, Linear Prediction and the Spectral Analysis of Speech, BBN Report No. 2304, 31 August 1972.

[2] Makhoul, John, "Spectral Analysis of Speech by Linear Prediction," IEEE Transactions on Audio and Electroacoustics, AU-21, 140-148, June 1973.