Proceedings of The Institute of Acoustics

COMPARATIVE RESULTS FOR COVARIANCE AND AUTOCORRELATION IN LINEAR PREDICTION OF CONTINUOUS SPEECH

R.C.L. O'NEIL
DEPARTMENT OF COMPUTING, NORTH STAFFORDSHIRE POLYTECHNIC,
STAFFORD, ENGLAND.

Introduction

The Z transform form of the transfer function of the vocal tract is 1/A(Z), where

$$A(Z) = \sum_{i=0}^{M} a_i Z^{-i} \quad \ldots\ldots (1)$$

it may be shown [1] that

$$\sum_{i=1}^{M} a_i c_{ik} = -c_{ok} \quad \ldots\ldots (2)$$

where

$$c_{ik} = \sum_{n=n_o}^{n_1} s(n-i)s(n-k), \quad k=1, \ldots, M \quad \ldots\ldots (3)$$

The predictor coefficients are a_i; a_o is 1, the number of poles of the model used in this paper was M=12 and C_{ik} is an element of a(M+1)x(M+1) covariance matrix. The covariance is evaluated from sample $s(n_o)$ to $s(n_1)$ of the speech samples, i.e. $s(n_o-M)$ to $s(n_1)$; matrix C is symmetric, $C_{ik}=C_{ki}$.

The Covariance method of evaluating the a_i consists of solving eqns. (2). The Autocorrelation method of solution differs from the Covariance in that the limits for n in (3) are no longer n_o to n_1 but $-\infty$ to ∞. This may be converted to a short-term Autocorrelation by setting all samples s(n)=0 for n<0 and n≥N. This is accomplished by using a Hanning window function on 256 samples per calculation frame [2]. By putting this restriction on eqns. (3), the form of the Autocorrelation elements c_{ij} become:-

$$c_{ij} = \sum_{n=0}^{N-1-|i-j|} s(n)s(n+|i-j|) \quad \ldots\ldots (4)$$

each diagonal in the c matrix now degenerating to a single value. There are thus only M+1 values to be computed for the c_{ij} in Autocorrelation. Thus:-

$$r(\ell) = \sum_{n=0}^{N-1-\ell} s(n)s(n+\ell), \quad \ell=0, \ldots, M \quad \ldots\ldots (5)$$

and

$$\sum_{i=1}^{M} a_i r(|i-j|) = -r(j), \quad j=1, \ldots, M \quad \ldots\ldots (6)$$

COMPARATIVE RESULTS FOR COVARIANCE AND AUTOCORRELATION IN LINEAR PREDICTION OF CONTINUOUS SPEECH

Prony's method [3] models the speech wave as a linear combination of real exponentials and damped sinusoids during voiced speech only. This mathematical approach reduces exactly to that of the Covariance method indicating that zero predictor error should arise in voicing if Covariance is used.

Once the a_i have been found by either method and after solution of the 12th order polynomial A(Z)=0 (by Bairstow's method) a sonagram-like "predictogram" plot of formants, bandwidths, predictor error and frame energy may be displayed [4].

Experimental Techniques:

The sampling rate of the speech, both real and synthetic, was 10 kHz. and the quantization 12 bits. The speech was stored on a fixed head disc in a PDP11. Pre-emphasis of +6dB/octave was used and found to improve results significantly over the cases where it was not used. All software was written in PAL 11 assembly language.

Results:

FIG 1, REAL depicts plots of a male voice producing a steady oo in a quiet office with a reasonably good quality microphone. The plots are labelled such that B corresponds with B in the synthetic oo depicted below, i.e. B = Covariance with pre-emphasis, D = Covariance without pre-emphasis, A = Autocorrelation with pre-emphasis, C = Autocorrelation without pre-emphasis in both FIG 1 and FIG 2. For the synthetic oo, pitch = 120 Hz, F_1 = 250 Hz, F_2 = 880 Hz, F_3 = 2080 Hz.

In all cases investigated, both real and synthetic, the Autocorrelation with pre-emphasis was best; sometimes the Autocorrelation without pre-emphasis was superior to the Covariance with pre-emphasis but the usual order was A, B, C, D. Similar results were obtained with other male speakers, as shown in FIG 2 for speakers R and K for vowel ee. The vertical bars in the predictogram are proportional to frequency/bandwidth. The horizontal separation of each frame plotted is proportional to the constant 128 samples for Autocorrelation and proportional to the pitch pulse separation for Covariance. If real roots were found for the 12th order polynomial, no formant or bandwidth is plotted. The maximum number of formants possible in a 12 pole model is 6 and in the synthetic speech case although only 3 formant frequencies were used in the synthesis, up to 6 formants may result from the analysis calculations.

Using synthetic speech and steady vowels standard deviations were calculated for the 4 possible combinations of method and pre-emphasis and some examples are quoted below for the synthetic oo.

	A(S.D., <F>)	B(S.D., <F>)	C(S.D., <F>)	D(S.D., <F>)
F_1 = 250	11, 302	22, 254	13, 250	48, 202
F_2 = 880	22, 1000	43, 989	73, 1048	108, 1018
F_3 = 2080	37, 1850	134, 1909	221, 1964	230, 1985

COMPARATIVE RESULTS FOR COVARIANCE AND AUTOCORRELATION IN LINEAR PREDICTION OF CONTINUOUS SPEECH

References

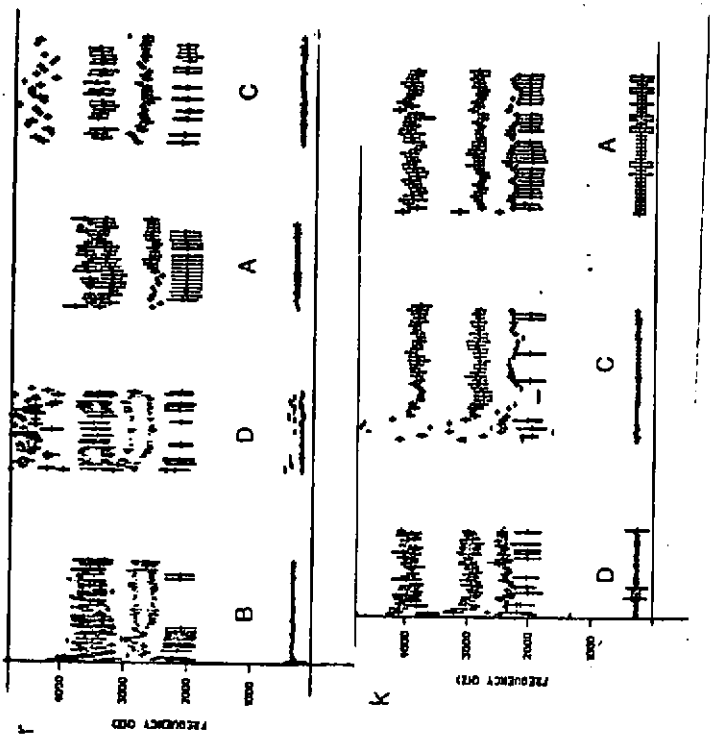
(1) J.D. MARKEL and A.H. GRAY, Linear Prediction of Speech, SPRINGER, 1976.

(2) J. MAKHOUL, Linear Prediction: A Tutorial Review, Proc. IEEE, 63, 4, 1975.

(3) R.N. McDONOUGH, Matched Exponents for the Representation of Signals, Ph.D. dissertation, Dept. of E.E., John Hopkins Univ., 1963.

(4) P.J. BRADLEY and R.C.L. O'NEIL, Linear Predictive Preprocessing for a Speech Understanding System, Proc. Inst. Acoustics, 1976, Edinburgh Meeting.

REAL ee
FIG 2



COMPARATIVE RESULTS FOR COVARIANCE AND AUTOCORRELATION IN LINEAR PREDICTION OF CONTINUOUS SPEECH

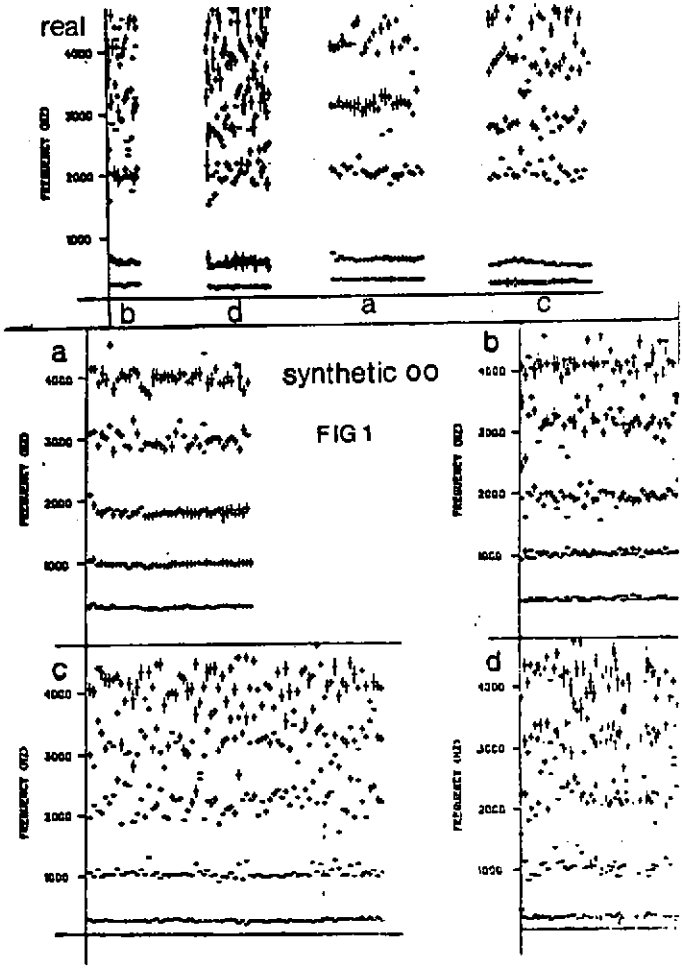


FIG 1