AN ASSOCIATIVE MEMORY FOR AUDITORY RECALL

Richard J. Reid

Computer Science Department
Michigan State University
East Lansing, Michigan U.S.A.

Abstract. A model of auditory temporal recall based on correlographic recording and Interrogation by convolution is presented. Complex tree and loop structured memory sequences can be recorded, with selective recall based on fragmentary cues. The sequences are "remembered" by proceeding through a recorded net of associations until the coherence of the retrieved signals falls to a sufficiently low level. This indicates no further significant recall is possible without an additional external cue.

A computer simulation of this model using digitised speech has produced variable-length and cyclic sound-segment recall from fragmentary cues.

Introduction. The problem of recording information in a manner so as to associate it with other recently received information and/or old information recalled by the present environment, appears to be at least a prerequisite to Intelligent behavior.

* We are concerned here with presenting a model of paired-associate recording and recall that may have the necessary potential.

The possibility of recall based upon holographic techniques was presented by Stroke (1) as a compensation or filter effect, and later by Collier and Pennington (2) as recall based upon fragments (possibly displaced) of the original scene. This verified Van Heerden's (3) earlier prediction made on a theoretical basis.

The experimental setup for displaying this phenomena is shown In Figures 1 and 2.

An interesting possibility of using this as a model of an associative memory comes from the fact that multiple exposures can be made on the same hologram and, after development, illumination by a fragment from one of the scenes will recall that original scene provided suitable circumstances are insured.

In the following discussion we shall be primarily concerned with auditory signals rather

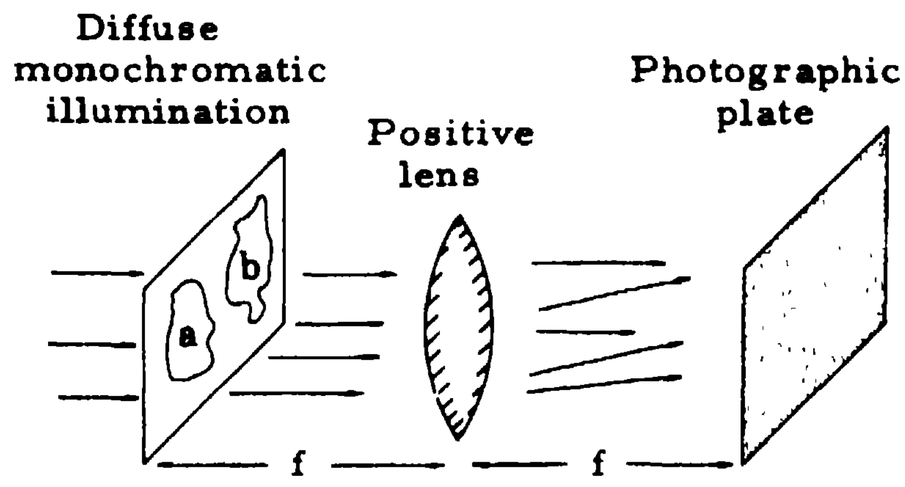

Figure 1.

Recording the paired association from a scene composed of two fragments.

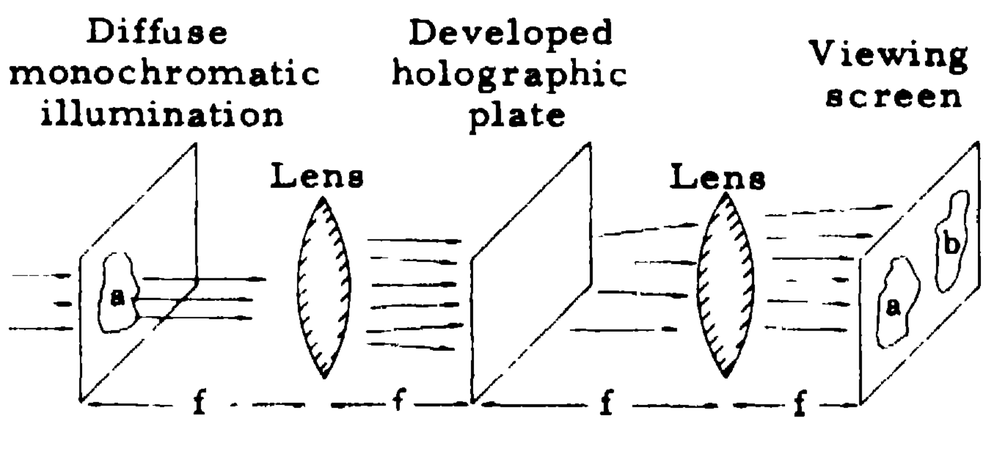

Figure 2.

Reconstruction of original scene by fragmentary cue from the original exposure.

than visual, but thay are generally Just a restriction to one dimension rather than two.

The possibility of recording the auditory analog to Figures 1 and 2 was suggested by Longuet-HIggins (4) and extended to a neural model by Chopping (5). In the auditory model a given time signal is considered to be made up of two successive fragments, a(t) and b(t). Then the magnitude of the power spectral density is recorded as

$$|Ffa(t)\} + F\{b(t)\}| \ll (A + B)(A + B)$$

where

$$A = F\{a(t)\}$$

is the Fourier transform of a(t) and ( ) is the complex conjugate.

Then upon recall the illumination by A would give

$$A(A + B)(A + B) = AAA^* + AAB^* + ABA^* + ABB^* \text{ . fl]}$$

If we examine a single one of these four sum terms in Equation [1], say ABA , it represents the convolution of the signal b(t) with the autocorrelation of the signal a(t). Assuming the autocorrelation peak at t - 0 is much greater than at any of the other registrations, then the convolution of this with b(t) will reproduce b(t). Further assume that for a term such as AAB , there are no peaks in the cross correlation represented by AB that are comparable to the autocorrelation peaks. In this case, the convolution with a(t) would act to severely attenuate a(t).

For the above assumptions then, the four terms in Equation [1] give the approximation

$$AAA + AAB + ABA + ABB = a(t) + n + b(t) + a(t)$$

where n represents noise as an attenuated and multiply-echoed a(t).

The reconstruction sequence above does contain the recall of b(t) and by a property of the Fourier transform, it will have the same relative time displacement from a(t) it had in the original exposure.

Thus, by recording the power spectrum of the original sequence and illuminating that recording with the spectrum of a fragment and inverse transforming the transmitted spectrum, and finally, playing the resulting sequence, one hears the recall fragment followed by the remainder of the original sequence.

It is similarly possible to use the cueb(t) to recall the entire sequence. If the simulation is performed using the discrete Fourier transform, then the multiplications of the spectra correspond to circular correlations or convolutions. Thus, in the replay, the recalled segment a(t) would follow after the cue b(t).

Multiple exposures can be made and each exposure can be composed of more than two fragments. However, there is some difficulty in overcoming the noise that is added from these additional terms. For example, suppose we have made two exposures, the first using the fragments a_1 and a_2 and the second using a_3 and a_4. Then upon attempted recall using a_1 we obtain

$$F^{-1}\{A_4[(A_1+A_2)(A_1+A_2)^* + (A_3+A_4)(A_3+A_4)^*]\}$$
$$\approx a_4 + n_1 + n_2 + a_4 + a_3 + n_3 + a_4 . [2]$$

where the n's represent the noise contributions, and, of course, the a_4'a and a_3 have the imperfections caused by their recalling autocorrelation terms. But, in any case, the recalled segment a_3 is present and experimentally is discerobable.

A study of the signal-to-noise ratios that can be expected based upon various numbers of exposures and fragment length to total signal length has been reported by WIllshaw and Longuet-Higgins (6).

Proposals have been made by Gabor (7) and Willshaw, Buneman and Longuet-Higgins (8) to eliminate a significant amount of noise by recording only the necessary correlographic terms rather than the power spectral density. This enhancement is included in the sections below.

The very peaked nature of the autocorrelation assumed above does not occur in practical signals, especially speech signals. And, although some have assumed that visual scenes will have this property if they are composed of very narrow line segments, this will insure that property in only one direction. The correlation can still be quite broad for translations along the direction of the lines and may give severe undulating ghosts in those directions.

The next sections also include a method for sharpening this autocorrelation peak without having to restrict the nature of the primary input signals.

A form of sequential recall was proposed by Longuet-Higgins (9), but it is not possible in

holographic recording without receding the information between successive cues.

The model to be discussed will allow the sequences to be tied back upon themselves to give continuing circular recall. In the simulation, an estimate is made of the signal-to-noise ratio, and as this falls to a small value, the recall is terminated pending the next external cue.

Sound Segments. Experimental facilities have been established for obtaining digitized speech. This is accomplished by low-pass filtering the speech input to less than half the sampling frequency (to prevent aliasing). This filtered input is then sampled periodically and converted via an analog-to-digital converter and stored in digitized form.

To minimize the amount of digital storage required, various sampling rates and numbers of quantization levels have been considered. After some experimentation, a sampling frequency of 4000Hz. and three bits of quantization were adopted as nearly the minimum quality suitable.

After the speech signals were in digital form they were disected into elements for a catalog of sound segments. These segments included spoken digits, phonemes and some complete words. These sound segments can be referenced by a dictionary process and concatenated for subsequent processing or digital-to-analog conversion.

Using this standard set of auditory signals obviates some of the severe problems dealing with the variability of actual speech such as changes in amplitude, pitch and time registration. Although a neuronal model has been proposed by Roy (9) for accommodating these variations, it has not been included in the present work.

Correlographic Recording. The paired-associate storage mechanism makes use of the discrete Fourier transform

$$X(j) = \frac{1}{N} \sum_{K=0}^{N-1} x(k)\, e^{-i2\pi jk/N}, \quad j=0,1,\ldots N-1$$

to transform the digitized speech segments x(k) to the frequency domain X(J). Each segment is padded to be 4096 samples long (one second of speech) and the fast Fourier transform is used.

Associates come in pairs, but they each represent a separate one-second sound interval and they are transformed separately. Their association is recorded as the product of the conjugate of the first transform and the second transform, and this product is added to any previously accumulated correlographic sum as

$$C_1(j) = C_0(j) + X_a^*(j)\, X_b(j) \qquad [3]$$

If the above second sound segment were to be followed by a segment c(t), then c would be separately transformed, associated with b and accumulated as

$$C_2(j) = C_1(j) + X_b^*(j)\, X_c(j)$$

This sequence of paired-associations can be continued down a chain of successive speech segments and/or new paired-associate sequences can be added. For example, the correlogram

$$C = X_a^* X_b + X_b^* X_c + X_d^* X_e \qquad [4]$$

records the associations a-b, b-c, and d-e.

The individual paired-associates are equivalent to the discrete Fourier transform of the circular cross-correlation between the sound segment pairs, i.e.

$$X_a^*(j)\, X_b(j) = \mathrm{DFT}\left\{ \frac{1}{N} \sum_{l=0}^{N-1} x_a(l)\, x_b(l+k)_N \right\}$$

where k is the index of the transform summation, DFT is the discrete Fourier transform, and the Index (1+k) is evaluated modulo N.

Recall* Unlike the bidirectional recall properties of the holographic recording, the cues in correlographic recording will recall only their successors. This directional characteristic has been likened to human recall which can go through

the alphabet, weekday names, etc., readily in the forward direction but poorly in reverse.

This recall is accomplished from a cue by taking its transform, passing the transform through the correlogram (multiplication), and inverse transforming the result. For instance, assume the previous recording given in Equation [4], then using b as a cue gives

$$DFT^{-1}\left\{X_b\left[X_a^* X_b + X_b^* X_c + X_d^* X_e\right]\right\} \approx n_1 + c + n_2 \quad [5]$$

subject to the previous assumptions about the auto and cross-correlations. Here then, the desired recalled element c may be obtained in a relatively unencumbered manner as opposed to the recall of a_3 given in Equation [2].

The recalled element in [5] was obtained as the convolution of the segment c with the auto-correlation of b, i.e.

$$X_c X_b^* X_b = DFT\left\{\frac{1}{N}\sum_{1=0}^{N-1} x_c(1)\left[x_b(k-1)_N * x_b(k-1)_N\right]\right\} \quad [6]$$

where k is the summation index of the transform, * is the correlation operator, and the indices (k-1) are evaluated module N.

It should be noted that the correlations and convolutions could be carried out in the time domain, but it is about an order of magnitude faster to carry it out in the transform plane in digital simulation when N = 4096.

In dealing with speech signals it is true that the autocorrelation is not a single sharp peak at t=0. Because of the periodic components of speech signals (the periods vary, but relatively slowly), the autocorrelation has a multitude of relatively high and somewhat broad peaks. In the convolution then, each of these peaks contributes a copy of the convolved signal, with each copy delayed by the interval between the autocorrelation peaks. The reconstructed signal is thus highly reverberative and is not easily identified.

If this phenomena is viewed in the frequency domain, then a sharp single peak in the autocorrelation produces a relatively uniform spectrum which then fully illuminates the recall element spectrum. However, the multiple-peaked autocorrelation produces more of an irregular "comb-like" spectrum, which only partially (at certain frequencies) illuminates or filters the recall element spectrum. This gives the recalled sound the characteristic that it has been passed through a drain-pipe acoustic filter.

The desired autocorrelation properties can be produced by the following techniques.

In order to avoid the successive reinforcing registrations of the periodic components during autocorrelation, their periodicities must be removed during the recording process. This can be done by permuting the digital samples within a sound segment, and if the permutation gives the signal a sufficiently noise-like character, the autocorrelation will produce the single peak at t-0, and the associated spectrum will be relatively uniform. The reverse of this permutation can be carried out on the recalled element to restore its periodic properties.

This permutation coding seems rather similar to the first layer transformation in a multi-layer perceptrem as discussed by Minsky and Pappert (10), and the cortical layered calculations proposed by Kabrisky (12).

The following permutation is easy to visualize and implement and does not require a permutation table. Given an original sample of N - 2 points, and suppose we are going to take points from this sample but that our selector runs at linearly increasing tempo. Then the i'th original sample goes into the b'th cell as

$$b_j = b_{j-1} + i \quad , i = 0, 1, \ldots, N-1$$

and $b_0 = 0$. Now if c_j equals b_j modulo N, then this redistribution of the data is a permutation if

$$c_k \neq c_1 \text{ if } k \neq 1 \text{ for all } 0 \leq 1 \leq 2^m - 1 .$$

This can be shown to be true.

While the above, or a random permutation, does give the desired autocorrelation result, it is at the expense of a rather desirable feature of the original recorded transforms. Previously, if a cue had been displaced from the registration it had during recording, the recall phenomena would still be intact although the recalled fragment would have been similarly displaced in time. For a cue displacement of 1 sampling intervals

$$X'(j) = DFT\{x'(k-1)_N\} = DFT\{x(k)\}\, e^{-i2\pi j1/N}$$

resulting in the autocorrelation spectrum

$$X(j)\, X^*(j)\, e^{-i2\pi j1/N}$$

i.e., the same spectral magnitude for illuminating the fragment to be recalled, with the phase change causing a uniform time delay in each of the spectral components--thus the delay in the recall.

With an arbitrary permutation encoding, it is not sufficient to be quite close to the registration, the registration must be exact.

A further desirable property of possible improvements to this first level encoding should be mentioned here. In order for the various noise terms to be minimal, it is necessary that the cross-correlation between recorded signals have quite low peaks, by comparison. In a particular sample, over the range of spoken digits, phonemes, and several short words, the ratio of autocorrelation peaks to maximum cross-correlation peaks varies from three or four-to-one to 50-to-one. Of course, this depends greatly upon the samples, and the time and amplitude quantisation. The important point to be made here is that the permutation coding discussed above was done solely to enhance the autocorrelation properties, while the cross-correlation properties are hardly affected. It may be true that both properties cannot be further improved unless the permutation encoding is dependent upon the signal itself.

Sequential Recall. The possibility of recalling a list of items occurs when one considers using a recall production as a cue for a further interrogation. This is a possibility in correlographic recording whereas it is not with the hologram or holophone because they reproduce the cue as part of the recall which could not then be directly recycled for further recall. Time division selection also would not be satisfactory in a discrete system because of the periodicity requirement.

Now consider the recording

$$C = X_a^* X_b + X_b^* X_c$$

and an interrogation with the sequence a which produces

$$DFT^{-1}\{X_a X_a^* X_b + X_a X_b^* X_c\} \approx b + n_1$$

If this result is used as a cue for an additional recall we obtain

$$DFT^{-1}\{X_a (X_a^* X_b + X_b^* X_c)^2\}$$

or

$$DFT^{-1}\{X_a [(X_a^* X_b)^2 + 2X_a^* X_b X_b^* X_c + (X_b^* X_c)^2]\}$$

$$\approx n_2 + 2c + n_3 .$$

There are several important considerations in this second result. First, consider the production of the desired sequence c from the term $X_a X_a^* X_b X_b^* X_c$. We depend upon the cross-correlation of a⋆a and b⋆b (or the convolution of a⋆a and b⋆b, but they're the same in this case) to still produce a relatively uniform spectrum for illuminating X_c. That is, holes in either spectra, that of $X_a^* X_a$ or $X_b^* X_b$, are missing from the illumination of X_c.

Next, consider the noise term n_2 which is produced by $X_a X_a^* X_b X_a^* X_b$. Alternatively, if we introduce the convolution operator, ⊛, the time domain equivalent is

$$n_2 = a \otimes \{(a * b) \otimes (a * b)\}$$

Unfortunately, the convolution of the two same cross-correlations tends to enhance the peaked nature of their convolution. This adds more coherence to their (the peaks') selections of time displaced copies of the signal a . Fortunately, the multiple peaks are not conglomerated, although frequently one is somewhat larger than the rest at the outset.

It is interesting to note that frequently the cross-correlation between any a and b as above, enhanced by the first convolution, is sufficient to produce an intelligible n' « a , if no other components are present. This is true of the n_3 term also, and it is only the relatively strong recall of c that masks this weaker secondary effect.

If variable length sequences are recorded, then it is important to be able to recognize when the recycling of productions as cues is to be terminated, so one can return for an external cue for the next interrogation. This will be considered below.

Cyclic Recall. If sequences can be recorded and retrieved as consecutive elements, then the sequence could also produce an item that was used as a cue somewhere previously in the list. This will then allow a loop through this sub-sequence As the simplest example consider

$$C = X_a^* X_b + X_b^* X_a$$

which is cyclic if initiated by either the cue a or b . If furnished the cue a, the r'th recall will produce

$$DFT^{-1} \left\{ Xa \left[X_a^* X_b + X_b^* X_a \right]^r \right\}$$

$$\approx \begin{cases} n_1(r) + a , & \text{if } r \text{ is even} \\ n_2(r) + b , & \text{if } r \text{ is odd} \end{cases}$$

Again, some detection as to when the recycling is to be terminated is implied.

These cyclic sequences can be longer and they may be preceded by a linear (non-cyclic) preamble.

Composite Recording. If the various possibilities of the preceding sections are combined, the composite correlogram corresponding to the sequential patterns of Figure 3 is possible.

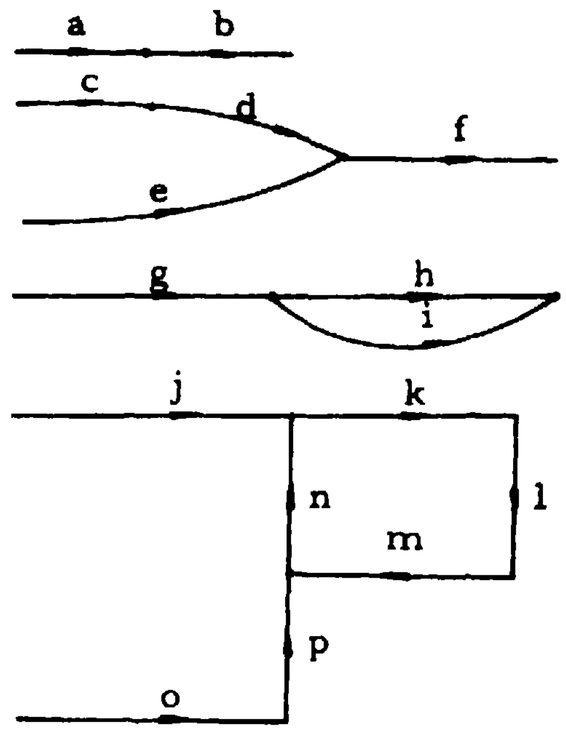

Figure 3.

A composite of sequential paired-associates recorded on a single correlogram.

Each of these paired associations is recorded individually, and the order is immaterial; although, in the simulation it is convenient to give sequences rather than individual pairs.

An interesting property of how a single new association can modify the overall behavior is illustrated by considering the recording shown in Figure 3 before the association n-k was recorded (N K added to the correlographic accumulation) versus the behavior after this single addition.

Termination Condition. With the provision for variable-length and cyclic sequences, it becomes necessary to determine an ending point for the recycling of productions as cues. For non-cyc-

lie sequences this should happen after the last element Is recalled. Last elements could possibly be especially marked for recognition as such, but that would not solve the cyclic termination problem. Instead, we attempt to use every production as a cue and notice that Its production may be totally noise-like. This approach has been somewhat successful In simulation and proceeds as follows.

After a particular production Is available, that Is, the Inverse transform of the convolution has been performed, the recalled sequence must be permuted In an Inverse manner to the encoding permutation. Then the spectrum of this correctly sequenced recall segment is examined for concentrations of power which would indicate the presence of a coherent signal. If this spectrum is given as

$$X(j)\ ,\ j = 0,\ 1,\ \ldots\ ,\ N-1$$

then define

$$\overline{P} = \frac{1}{N} \sum_{j=0}^{N-1} X(j)\, X(j)^*$$

as the average spectral power, and

$$\sigma = \sqrt{\frac{\sum_{j=0}^{N-1} \left[X(j)\, X^*(j) - \overline{P} \right]^2}{N}}$$

as the standard deviation.

Then the original signal is judged to be other than noise if at least several of the $X(j)X^*(j)$ terms are more than a few standard deviations greater than the mean. In the present experiments with 4096 spectral components, six were required to be greater than five standard deviations from the mean.

As might be expected, this detection is rather imperfect, and is set to accept possibly several recalls of noise rather than interrupt a legitimate sequence. An improvement would undoubtedly result if small clusters of exceptional power were identified.

For cyclic sequences the same termination mechanism is employed. This works because there are more and more convolutions of autocorrelations convolved with the desired sequence in one particular term of the recall sum and the number of noise terms is increasing. So, as contrasted to non-cyclic recall dropping off the end of the initiated sequence, cyclic recall fades into the noise and is ultimately aborted.

Typical Correlogram Composition. The entries in Table 1 indicate the present range of sequences that can be recorded on a single correlogram and give intelligible recall.

Length of individual sequences	Number of individual sequences
2	10
3	4
4	2
Cycle length (only a single cycle recorded)	**Recall cycles (until unintelligibility)**
2	5
3	3
4	2

Table 1.

Typical lengths and numbers of sequences and cycles that can be recorded on a single correlogram and give intelligible recall.

The sound elements are each of one second duration and correspond to 4096 samples of 3 bits each. The correlogram is recorded as two real values and 2047 complex values where the exponents are 11 bits and the mantissas 37 bits long.

Summary. It has been shown that it is possible to record several sequences of aural information in a single correlogram and retrieve the indi-

vidual sequences by furnishing the appropriate cues.

The variable-length sequences are "re-membered" by cycling a production as the next cue until the output contains no further signals.

As can be seen from Table 1, the present simulation is rather limited in terms of the total amount of information that can be recorded before the noise produced masks the desired recall elements. A practical improvement in this aspect can probably be obtained by filtering the output to consist only of the same exceptional spectral components that are used to identify the fact that there is actually a signal present.

Further improvement may be made by making the permutation (or other) encoding dependent upon some well-preserved--preserved through recording and recall--property of each individual signal.

Of great interest is how the memory mechanism can be made more dynamic. That is, successive recordings should be made as associations with recent external inputs and the internal recall they may have triggered. Also, recall itself should be done so as to modify the information retained.

References

1. G. W. Stroke, R. Restrick, A. Funkhouser and D. Brumm, Resolution-retrieving compensation of source effects by correlative reconstruction in high-resolution holography, Physics Letters, *18,* 3 (1965) 274-275.
2. R. J. Collier and K. S. Pennington, Ghost imaging by holograms formed in the near field, Appl. Phys. Lett. *8,* 2 (1966) 44-66.
3. P. J. Van Heerden, A new optical method of storing and retrieving information, Appl. Optics, 2_t 4 (1963) 387-393.
4. H. C. Longuet-Higgins, Holographic model of temporal recall, Nature, 217 (1968) 781-782.
5. P. T. Chopping, Holographic model of temporal recall, Nature, 211 (1968) 781-782.
6. D. J. Willshaw and H. C. Longuet-Higgins, The holophone - recent developments, Machine Intelligence IV, 349-357.
7. D. Gabor, Associative holographic memories, IBM Jour. Res. Dev., 13 (1969) 156-159.
8. D. J. Willshaw, 0. P. Buneman and H. C. Longuet-Hlggsin, Non-holographic associative memory, Nature, 222 (1969) 960-962.
9. op. cit. - 4.
10. A. E. Roy, Certain pattern recognition problems, private communications, (1970).
11. M. Mlnsky and S. Papert, Perceptrons, MIT Press (1969) 228-232.
12. M. Kabrisky, A proposed model for visual information processing in the human brain, University of Illinois Press, 1966.