THE LOÈVE-KARHUNEN EXPANSION AS A MEANS OF INFORMATION COMPRESSION FOR CLASSIFICATION OF CONTINUOUS SIGNALS

SATOSI WATANABE

FOREWORD

This report was prepared at the International Business Machines Corporation, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York, under Contract AF 33(657)-11347, for Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio. The work was in support of Project 7233, "Biological Information Handling Systems and Their Functional Analogs," Task 723305, "Theory of Information Handling." The research was initiated 15 July 1963 and completed 14 July 1964. The technical contract monitor was Hans L. Oestreicher, PhD, Chief, Mathematics and Analysis Branch, Biodynamics and Bionics Division, Biophysics Laboratory.

This technical report has been reviewed and is approved.

J. W. HEIM, PhD
Technical Director
Biophysics Laboratory
Aerospace Medical Research Laboratories

ABSTRACT

The classification of an ensemble of continuous signals $\{f^{(\alpha)}(x)\}$ is accomplished by expansion in terms of a set of orthogonal functions $\{\psi_i(x)\}$. The coefficients $c_i^{(\alpha)}$ of the expansion represent the desired information content of the continuous signals. A probability measure ρ_i with respect to $f^{(\alpha)}(x)$ is defined on $\{\psi_i(x)\}$. That set of functions $\{\phi_j(x)\}$ is desired which concentrates the ρ's on a few functions instead of distributing them over many. This is accomplished by minimization of the entropy function $S(\{\psi_i\}) = -\sum_{i=1}^{\infty} \rho_i \log \rho_i$ with respect to $\{\psi_i\}$. The ρ_i are shown to be the diagonal terms of the autocorrelation matrix G, whose complete set of eigenfunctions is a desired set of orthogonal functions and whose eigenvalues represent the associated ρ's. The expansion of the set $\{f^{(\alpha)}(x)\}$ in terms of the eigenfunctions of the autocorrelation matrix is the Loève-Karhunen (L-K) expansion. This expansion in addition minimizes the mean square error which results from truncating the expansion. One application of the method is the clustering of ninety five samples of the spontaneous activity potential in the ventral nerve cord of a crayfish. The first L-K expansion coefficient divides these samples into three major groups. In another application samples of the power spectrum of twelve isolated vowel sounds spoken by nineteen persons are separated into roughly nine disjoint regions. This is accomplished by using the first three L-K coefficients. Use of the fourth coefficient discriminates the most troublesome remaining vowel from the rest.

TABLE OF CONTENTS

SECTION I

INTRODUCTION

First, a few general remarks about the problem of classification. We are given a collection of objects each of which is described by a set of predicates or a set of variables. The task of classification can be divided into two kinds according to the modes of application: recognition and clustering. We usually speak of recognition when the collection consists of two subcollections, such that the objects in one of the subcollections are already placed in different classes, and the task consists of placing items of the second subcollection into these ready-made classes "in imitation" of the paradigms given by the first subcollection. We speak of clustering, when neither classes nor paradigms are given, and the task consists of introducing classes in such a way that objects within each class cohere well together.

There is no unique solution to either of these problems, basically due to the "generalized theorem of the ugly duckling" (ref 4). To apply this theorem to the case of continuous variables, we note that due to the inevitable error and inaccuracy of observation, the continuous variable can be quantized so that the information can usually be expressed without loss by a finite number of two-valued variables (predicates). What the theorem states is essentially the following. If a particular group of r non-identical objects stand mutually in a certain formal relationship with reference to s predicates, then any other group of r non-identical objects stands in exactly the same formal relationship with regard to s other predicates, albeit these predicates may not be found in the original set of predicates, but identifiable among the possible Boolean functions of the original predicates. Thus, for instance, any arbitrary pair of two objects in the collection shares a constant number of predicates. Thus, if all possible applicable predicates have the same weight, then there cannot be any classes or classification. Conversely, if we recognize some useful classes among objects, that means that some predicates are given more importance than some others. By the same token, some variables are instrumental in defining useful classes and some others are not. Classes per se do not exist in any given collection of objects. They emerge only when the purpose or usage is given for which classes are used, since some predicates and variables are more pertinent to the given usage of the classification.

The task of classification then becomes largely one of extracting more important predicates or variables. This implies that elimination of irrelevant information and compression of relevant information is the primary goal of classification problems.

Suppose, for instance, we have the following recognition problem. In the subcollection of paradigms, we have one hundred yellow roses and one hundred white foxes, and we are told that the first hundred items belong to class I and the second hundred items belong to class II. Suppose that the second subcollection consists of a yellow fox. Does it belong to class I or class II? It should belong to class I, if color description is more important and it should belong to class II, if biological species is more important. The first answer may be correct for the usage with reference to color photography and the second answer may be correct for the usage of biologists. In the first case, for the purpose of information compression, we should suppress all predicates other than color, and similarly in the second case.

In addition to this general arbitrariness inherent to any inductive process, the problem of clustering involves other types of arbitrariness, such as a desirable number of classes, a desirable number of members in each class, a suitable definition of membership and non-membership of a class, etc.

Probably, the most general starting point for classification is the assumption that each subcollection which can be taken in the entire collection is assigned a certain degree of "cohesion" (ref 4). In some cases, cohesion may originate from an "emergent" property which cannot be reduced to a bilateral relation, such as similarity, dissimilarity, proximity and distance. What is dealt with in the present paper, however, is the case where the cohesion is reducible to a bilateral relation and further the "distance" between two objects is given or calculable, in such a way that there exists a set of variables which subtend a space in which the Euclidean distance becomes precisely the distance in the above sense. In this particular case, the essential algorithm of classification becomes discovery of the optimal way of information compression which on the average loses as little information as possible with respect to the distance between pairs of objects.

The problem we are facing in this paper is one of classifying continuous signals. Since a continuous curve corresponds to continuously many variables each of which can take continuously many possible values, the first step of information compression will be to expand the continuous function with the help of an orthogonal function set and use the coefficients of the expansion as carriers of information. This reduces, at least, the number of continuously valued variables to

an enumerable infinity. The next step of information compression will be to choose an orthogonal function set in such a way that a finite number of terms in the expansion may be sufficient to convey the information as to the basic concept of distance to a good approximation. The following sections will be concerned with this problem of optimal expansion. After these mathematical sections, a few examples of application will be introduced.

Although, the above description assumes that the objects are continuous curves, actually the same principle of extracting optimal variables when the distance is well defined can be applied to the cases where the objects are vectors with an enumerable or finite number of components. The results of those cases will be published elsewhere.

The present paper is concerned mainly with the aspect of information compression, which is only part of the process of recognition. The problem of zoning (division of the space into disjoint volumes corresponding to classes) and the problem of decision making (such as the Bayesian algorithm) require, among others, careful study in connection with classification and recognition, but are left out from this paper.

The author notes here that his attention was drawn to the Loève-Karhunen expansion by Eugene Wong (ref 5), and that he became interested in this method mainly due to its striking similarity with the theory of the density matrix in quantum mechanics. After the author decided to try this method on a speech recognition problem, it came to his attention that Kramer and Mathews (ref 2) tried a similar method on the coding problem with vocoders. Their method corresponds to a discrete (instead of continuous) version of the L-K method, and should become identical at the level of computation on a digital computer. A surprising fact is that Kramer and Mathews came to this idea through a consideration quite different fromthe L-K method, and that the name of L-K method is not mentioned even a single time in their paper. R. Bakis and G. Hu (ref 1) independently, but later than these two authors, also applied the method to the problem of isolated vowel recognition. The present author is inclined to believe that had their results been as encouraging as ours (see section VI of this paper) they would have followed through their initial attempts and developed a useful speech handling methodology by now.

SECTION II

OPTIMIZATION OF INFORMATION COMPRESSION

By an "ensemble", we understand here the collection of a large number N of similar objects; of which $Nw^{(\alpha)}$, $\alpha = 1, 2, \ldots, \nu$ belong to type α, whereby the relative frequency $w^{(\alpha)}$ satisfies the postulates of probabilities: $w^{(\alpha)} \geq 0$ and $\sum_{\alpha} w^{(\alpha)} = 1$. Since we usually consider the limiting case $N \to \infty$, we can make abstraction of the number N, and represent the ensemble as a set of types, $\alpha = 1$, $\alpha = 2, \ldots$, with a probability measure $w^{(\alpha)}$.

We are given an ensemble of (complex) functions $f^{(\alpha)}(x)$, $\alpha = 1, 2, \ldots$, of a (real) variable x with respective weight $w^{(\alpha)}$, where each $f^{(\alpha)}(x)$ is square-integrable in the domain (a, b). We assume in the present paragraph that each function $f^{(\alpha)}(x)$ is already normalized so that

$$\int_a^b f^{(\alpha)*}(x)\, f^{(\alpha)}(x)\, dx = 1 \tag{1}$$

where the star means the complex conjugate.

Let $\{\psi_i(x)\}$, $i = 1, 2, \ldots$ be a complete set of orthonormal functions also defined in (a, b) so that

$$\int_a^b \psi_i^*(x)\, \psi_j(x)\, dx = \delta_{ij} \tag{2}$$

If we expand $f^{(\alpha)}(x)$ in terms of the $\psi_i(x)$ as

$$f^{(\alpha)}(x) = \sum_{i=1}^{\infty} c_i^{(\alpha)} \psi_i(x) ,$$

$$c_i^{(\alpha)} = \int_a^b \psi_i^*(x) f^{(\alpha)}(x)\, dx. \qquad (3)$$

Then, the normalization (eq 1) will, in virtue of eq 2 result in

$$\int_a^b |f^{(\alpha)}(x)|^2 dx = \sum_{i=1}^{\infty} |c_i^{(\alpha)}|^2 = 1. \qquad (4)$$

It may be considered as very natural to define the "distance" between two functions $f^{(\alpha)}(x)$ and $f^{(\beta)}(x)$ by

$$\text{Dist } (f^{(\alpha)}, f^{(\beta)}) = \int_a^b |f^{(\alpha)}(x) - f^{(\beta)}(x)|^2 dx$$

$$= 2 - \int_a^b [f^{(\alpha)*}(x) f^{(\beta)}(x) + f^{(\alpha)}(x) f^{(\beta)*}(x)] dx \qquad (5)$$

or equivalently in terms of the coefficients $c_i^{(\alpha)}$,

$$\text{Dist } (f^{(\alpha)}, f^{(\beta)}) = \sum_{i=1}^{\infty} |c_i^{(\alpha)} - c_i^{(\beta)}|^2$$

$$= 2 - \sum_{i=1}^{\infty} (c_i^{(\alpha)*} c_i^{(\beta)} + c_i^{(\alpha)} c_i^{(\beta)*}) \qquad (6)$$

The second term of the right-hand sides of eq 5 or eq 6 is nothing but the "correlation" between the two functions, hence, we can say that the larger the correlation the smaller the distance and vice versa, which is intuitively satisfactory.

An important fact to note in connection with the notion of distance (hence also correlation) is that it is formally invariant both for the unitary transformation from the x-representation to the i-representation,

and for the unitary transformation from the i-representation based on $\{\psi_i\}$ to the j-representation based on another orthonormal function set $\{\phi_j\}$. The first transformation is characterized by the matrix

$$(x \mid T \mid i) = \psi_i(x) \tag{7}$$

which satisfies

$$(i \mid T^{-1} \mid x) = (i \mid \overline{T} \mid x) = \psi_i^*(x) \tag{8}$$

where the superscript (-1) means the inverse and the bar (-) means the hermitian conjugate. The second transformation is characterized by the matrix

$$(i \mid T \mid j) = \int_a^b \psi_i^*(x)\ \phi_j(x)\ dx \tag{9}$$

with

$$(j \mid T^{-1} \mid i) = (j \mid \overline{T} \mid i) = \int_a^b \phi_j^*(x)\ \psi_i(x)\ dx \tag{10}$$

From $T\,T^{-1} = T^{-1}T = 1$ follows then

$$\sum_i |(i \mid T \mid j)|^2 = \sum_j |(i \mid T \mid j)|^2 = 1 \tag{11}$$

The magnitude $|c_i^{(\alpha)}|$, or more conveniently the squared magnitude $|c_i^{(\alpha)}|^2$, of the coefficient for $\psi_i(x)$ in the expansion of $f^{(\alpha)}(x)$ in eq 3 can be regarded as a good measure of the extent to which the (normalized) function $\psi_i(x)$ is useful in representing the given function $f^{(\alpha)}(x)$. Those base functions $\psi_i(x)$ whose coefficients are small in magnitude in eq 3 can be ignored without altering the situation

appreciably. Therefore, its average in the ensemble

$$\rho_i = \sum_{\alpha} w^{(\alpha)} |c_i^{(\alpha)}|^2 \tag{12}$$

can be considered as the measure of importance of $\psi_i(x)$ when we represent the ensemble of functions by the function set $\{\psi_i(x)\}$. The convenience of using the ρ_i thus defined for this purpose stems from the fact that it is a probability measure defined on $\{\psi_i(x)\}$ satisfying

$$\rho_i \geq 0, \quad \sum_{i=1}^{\infty} \rho_i = 1 \tag{13}$$

For a given ensemble of functions $\{f^{(\alpha)}\}$, $\{w^{(\alpha)}\}$, the values of the ρ's vary depending on the choice of the orthogonal function set $\{\psi_i\}$. From the point of view of information compression, which is a necessary step in recognition problems, those orthogonal function sets are preferred for which the ρ's are concentrated on a few functions instead of being widely spread over many functions. To formulate mathematically this idea, it would be convenient to introduce the entropy function in terms of the ρ's

$$S(\{\psi_i\}) = -\sum_{i=1}^{\infty} \rho_i \log \rho_i \tag{14}$$

which depends on the set $\{\psi_i\}$ for a given ensemble $\{f^{(\alpha)}\}$ with $\{w^{(\alpha)}\}$. The desirable choice of the function set $\{\psi_i\}$ will then be obtained by minimization of S by variation of the $\{\psi_i\}$. Let this optimal function set $\{\phi_j\}$ satisfying

$$S(\{\phi_j\}) = \operatorname*{Min}_{\{\psi_i\}} S(\{\psi_i\}) \tag{15}$$

be called the optimal function set for information compression, or by abbreviation OFSIC. The next section will give the way to obtain the OFSIC.

SECTION III

ENTROPY-MINIMIZING PROPERTY OF LOEVE-KARHUNEN EXPANSION

We shall first define a special function set, which we denote by $\{\phi_j\}$ and then proceed to demonstrate that this function set is indeed an OFSIC.

A hermitian matrix G called the autocorrelation function, or the density matrix is given in the x-representation by

$$(x|G|y) = \sum_{\alpha} w^{(\alpha)} f^{(\alpha)}(x) f^{(\alpha)*}(y), \quad a \le x \le b, \quad a \le y \le b$$

$$= \sum_{\alpha} w^{(\alpha)} \sum_{i} \sum_{k} c_i^{(\alpha)} c_k^{(\alpha)*} \psi_i(x) \psi_k^*(y) \tag{16}$$

The same matrix in the i-representation based on an arbitrary function system $\{\psi_i(x)\}$ is

$$(i|G|k) = \int_a^b \int_a^b (i|T^{-1}|x)\, dx\, (x|G|y)\, dy\, (y|T|k)$$

$$= \sum_{\alpha} w^{(\alpha)} c_i^{(\alpha)} c_k^{(\alpha)*} \tag{17}$$

where the unitary transformation T is given by eq 7.

We define a special function system $\{\phi_j(x)\}$ as the set of eigenfunctions of G, i.e.,

$$\int_a^b (x|G|y)\, \phi_j(y) dy = \lambda_j \phi_j(x) \tag{18}$$

When there is degeneracy, the eigenfunctions are not uniquely defined, but we assume, for the present, that within this freedom one complete orthonormal function set is arbitrarily chosen. If there are μ linearly independent functions among $\{f^{(\alpha)}\}$, there will be μ non-zero eigenvalues, allowing the definition of μ eigenfunctions. To make $\{\phi_j(x)\}$ a complete function set, we have to augment this set of μ functions with the eigenfunctions corresponding to the multiply degenerate eigenvalue zero. The expansion of a function $f^{(\alpha)}$ in terms of

such a particular set $\{\phi_j\}$,

$$f^{(\alpha)}(x) = \sum_{j=1}^{\infty} g_j^{(\alpha)} \phi_j(x) \tag{19}$$

is nothing new and is known under the name of the Loève-Karhunen expansion (ref 3) usually with an additional condition (eq 30) below, but the property we are now going to prove is something so far unnoticed. If we insert eq 19 in eq 16 and write out eq 18 we obtain

$$\sum_{\alpha} w^{(\alpha)} g_j^{(\alpha)} g_l^{(\alpha)*} = \lambda_j \delta_{jl} \tag{20}$$

In other words, if we transform eq 17 to the j-representation based on $\{\phi_j\}$ with the help of the unitary transformation T^{-1} of eq 9 it becomes diagonal, i.e.,

$$(j|T^{-1}|i)(i|G|k)(k|T|l) = \lambda_j \delta_{jl} \tag{21}$$

and conversely

$$(i|G|k) = \sum_j (i|T|j) \lambda_j (j|T^{-1}|k) \tag{22}$$

where $(i|G|k)$ is the i-representation based on $\{\psi_i\}$.

Whether or not the special function system $\{\phi_j\}$ is used, the diagonal elements $(i|G|i)$ of eq 17 are all non-negative and sum up to unity, for they are nothing but ρ_i of eq 12

$$\rho_i = (i|G|i) \tag{23}$$

From eq 22 follows then

$$\rho_i = \sum_{j=1}^{\infty} A_{ij} \lambda_j \tag{24}$$

where the matrix

$$A_{ij} = |(i|T|j)|^2 \geq 0 \tag{25}$$

is "double-stochastic", due to eq 11,

$$\sum_j A_{ij} = \sum_i A_{ij} = 1 \tag{26}$$

Now we are prepared to prove

Theorem 1. For a given function ensemble, the value of the entropy defined in eq 14 on the basis of an arbitrary function system $\{\psi_l\}$ cannot be smaller than its value based on a Loève-Karhunen function system $\{\phi_j\}$.

In other words, the Loève-Karhunen function system is an OFSIC. If there is more than one OFSIC it is only due to the degeneracy of the matrix G. To prove theorem 1 we need only introduce as a lemma the well-known H-theorem in its simplest version, to which we need not give any demonstration here.

Lemma 1 (H-Theorem). Let p_i, $i = 1, 2, \ldots$, be a probability distribution, i.e., $p_i \geq 0$ and $\sum_i p_i = 1$, and let A_{ij} be a double stochastic variable, i.e. it is a matrix satisfying eq 25 and eq 26. Then

$$-\sum_j p_j \log p_j \leq -\sum_i q_i \log q_i \tag{27}$$

where

$$q_i = \sum_j A_{ij} p_j \tag{28}$$

All we need to prove theorem 1 is to replace $\{p_j\}$ and $\{q_i\}$ by $\{\lambda_j\}$ and $\{\rho_i\}$ respectively in the lemma. Eq 27 becomes here

$$-\sum_j \lambda_j \log \lambda_j \leq -\sum_i \rho_i \log \rho_i \tag{29}$$

satisfying eq 15. The question as to when the equality sign in eq 27 becomes valid, i.e., as to when two function systems become both optimal, will be discussed in Appendix I.

SECTION IV

ERROR-MINIMIZING PROPERTY OF LOÈVE-KARHUNEN EXPANSION

In the discussion of this section, we have to agree that the labeling of the eigenvalues λ_j of G is so made that

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots . \tag{30}$$

This is often included in the definition of the Loève-Karhunen expansion but is irrelevant to the theorem of the last section. (Let us note here that relabeling can be considered a unitary transformation). The Loève-Karhunen expansion is usually characterized as the one which minimizes the mean square error committed by taking only a finite number of terms of expansion, in the following sense. The average of the squared error committed by taking only n terms in the expansion is

$$E(\{\psi_i\}, n) = \sum_{\alpha} w^{(\alpha)} \int_a^b \left| f^{(\alpha)}(x) - \sum_{i=1}^{n} c_i^{(\alpha)} \psi_i(x) \right|^2 dx \tag{31}$$

where $c_i^{(\alpha)}$ is given by the second line of eq 3. The above mentioned characterization of the Loève-Karhunen function system $\{\phi_j\}$ implies then

$$E(\{\phi_j\}, n) = \operatorname*{Min}_{\{\psi_i\}} E(\{\psi_i\}, n) \tag{32}$$

for each n. This property is, of course, also very desirable for the purpose of information compression, but it is rather interesting that the entropy-minimizing requirement and the error-minimizing requirement result in the same function system.

In this section, with the help of Appendix II, we shall give a proof for eq 32 through an approach a little different from the usual one. Namely, we shall introduce a theorem (presumably new) which has an interesting implication also in connection with the theory of Markov chains.

We can write eq 31 also as

$$E(\{\psi_i\}, n) = \sum_{\alpha} w^{(\alpha)} \sum_{i=n+1}^{\infty} |c_i^{(\alpha)}|^2$$

$$= 1 - \sum_{i=1}^{n} \rho_i \qquad (33)$$

where ρ_i is given by eq 12 and eq 13. Hence, eq 32 becomes simply

$$\sum_{j=1}^{n} \lambda_j \geq \sum_{i=1}^{n} \rho_i \qquad (34)$$

where the λ_j are defined by eq 18, and constitute a special case of the ρ_i in the representation diagonalizing G. Eq 32 or eq 34 asserts its validity no matter how the ρ's are labeled, provided the λ_j's are arranged as in eq 30. In view of eq 24, we can further rewrite eq 34 as

$$\sum_{j=1}^{n} \lambda_j \geq \sum_{i=1}^{n} \sum_{j=1}^{\infty} A_{ij} \lambda_j \qquad (35)$$

The theorem we need then in order to prove eq 34 or eq 35 is the following:

Theorem 2. Two probability distributions $\{\rho_i\}$ and $\{\lambda_j\}$ satisfying, $\rho_i \geq 0$, $\lambda_j \geq 0$, $\sum_{i=1}^{\infty} \rho_i = 1$, $\sum_{j=1}^{\infty} \lambda_j = 1$ are

connected by a double stochastic matrix A_{ij} as in eq 24. The labeling of the λ_j is done in a descending order as in eq 30. Then, for any arbitrary n, the sum of the first n elements of the $\{\lambda_j\}$ is not less then the sum of the first n elements of the $\{\rho_j\}$.

I conjectured without proof validity of Theorem 2 and told Professor S. Kakulani about it, who thereupon immediately provided me with

an intuitively interesting proof. I reproduced the orally transmitted proof in Appendix II in my own language. Any defect is to be attributed to me.

Eq 20 can be interpreted as showing that the L-K coefficients g_ℓ's with different ℓ's are statistically uncorrelated. This provides another argument for using the L-K coefficients as variables to represent economically the members of the ensemble. It would be desirable from the viewpoint of elimination of redundancy to use variables which are naturally statistically independent, but in the absence of such variables, statistically uncorrelated variables may be the next best. Further, if each of the g_ℓ's has a Gaussian distribution, they become statistically independent.

For the purpose of discrimination it is advisable to subtract the average function $\sum_\alpha w^{(\alpha)} f^{(\alpha)}(x)$ from each function $f^{(\alpha)}(x)$ before applying the L-K method, for if this average is not zero, the first eigenfunction tends to represent this average function which has no discriminating usefulness.

SECTION V

EXAMPLE OF CLUSTERING: ACTION POTENTIAL SPIKES IN NEURON BUNDLE

The following is an example showing that in some cases the distribution of objects along the axis representing a Loève-Karhunen coefficient already displays obviously disparate peaks so that clustering can be done without further elaborate methods such as one proposed in reference 4.

It is desirable but not practicable to measure simultaneously the potential of each of several neurons separately. We could achieve the same goal, however, if we measure the electrostatic potential of a bundle of neurons altogether, and then let a computer recognize the classes of shapes of action potentials, since each neuron in the bundle, due to its geometrical position relative to the electrode, gives rise to a different but definite shape of action potential on the gross electrode.

Dr. W. Makous of IBM Research took a digitized record of potential (quantized into 64 amplitude categories) of spontaneous activity of the ventral nerve cord of a crayfish at a sampling rate of 13,400 cycles per second. At the same time, a photographic record of the oscilloscope display of the same measured potential was taken. The results explained below are those obtained in the first attempt of this kind, and by no means represent the limitation of the method.

Figures 1 and 2 are results obtained by applying the L-K method to 95 spikes measured in one series of observations made on the ventral cord of a crayfish. As the histogram of figure 1 shows, the first L-K coefficient classifies these 95 into 3 major groups. The left group consists of dull concave-shaped mono-phasic spikes. The right group consists of dull convex-shaped mono-phasic spikes. The central group consists of tri-phasic spikes which are definitively genuine action potentials. This separation into three groups is a case of "natural" clustering. Figure 2 is the classification of the same 95 spikes, according to the value of the curve at the point farthest from zero. Since, the L-K method normalizes every curve, it is advisable to use the size of each curve as another characteristic variable. The three kinds of shaded area on figure 2 correspond respectively to their counterparts in figure 1. The tri-phasic action potentials are separated again into three sub-classes in figure 2, and this distinction corresponds very well with human discrimination made on the photographic record.

The eigenvalues of the G-matrix were showing fast descent with increasing number of the index 0.4678, 0.2849, 0.0432, 0.0335, 0.0196, 0.0173, 0.0140, 0.0122, 0.0112, 0.0098, ... The entropy was 2.586 bits.

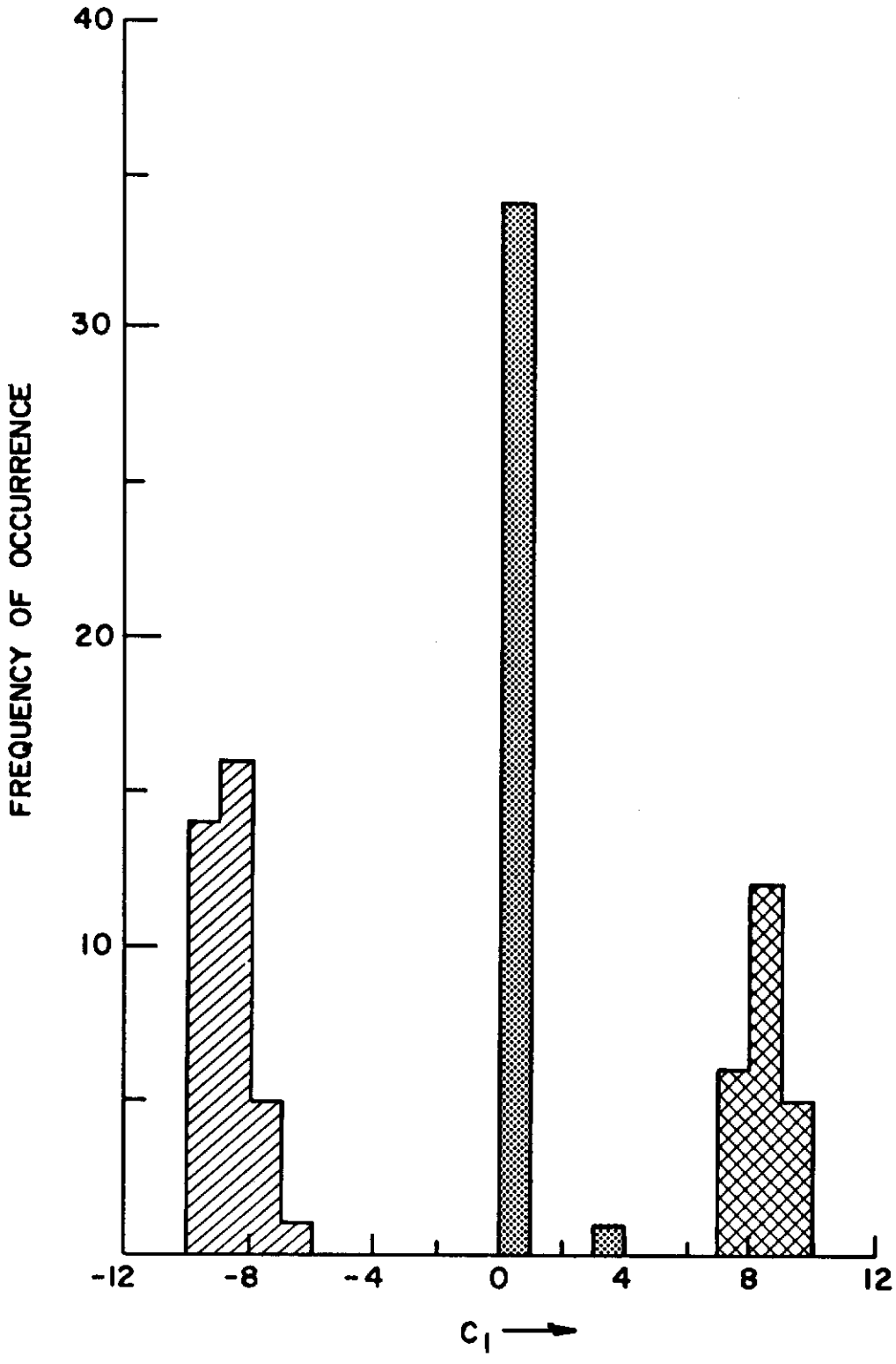

Figure 1 Histogram of first L-K coefficient for 95 samples of spikes of spontaneous activity in the ventral nerve cord of a crayfish.

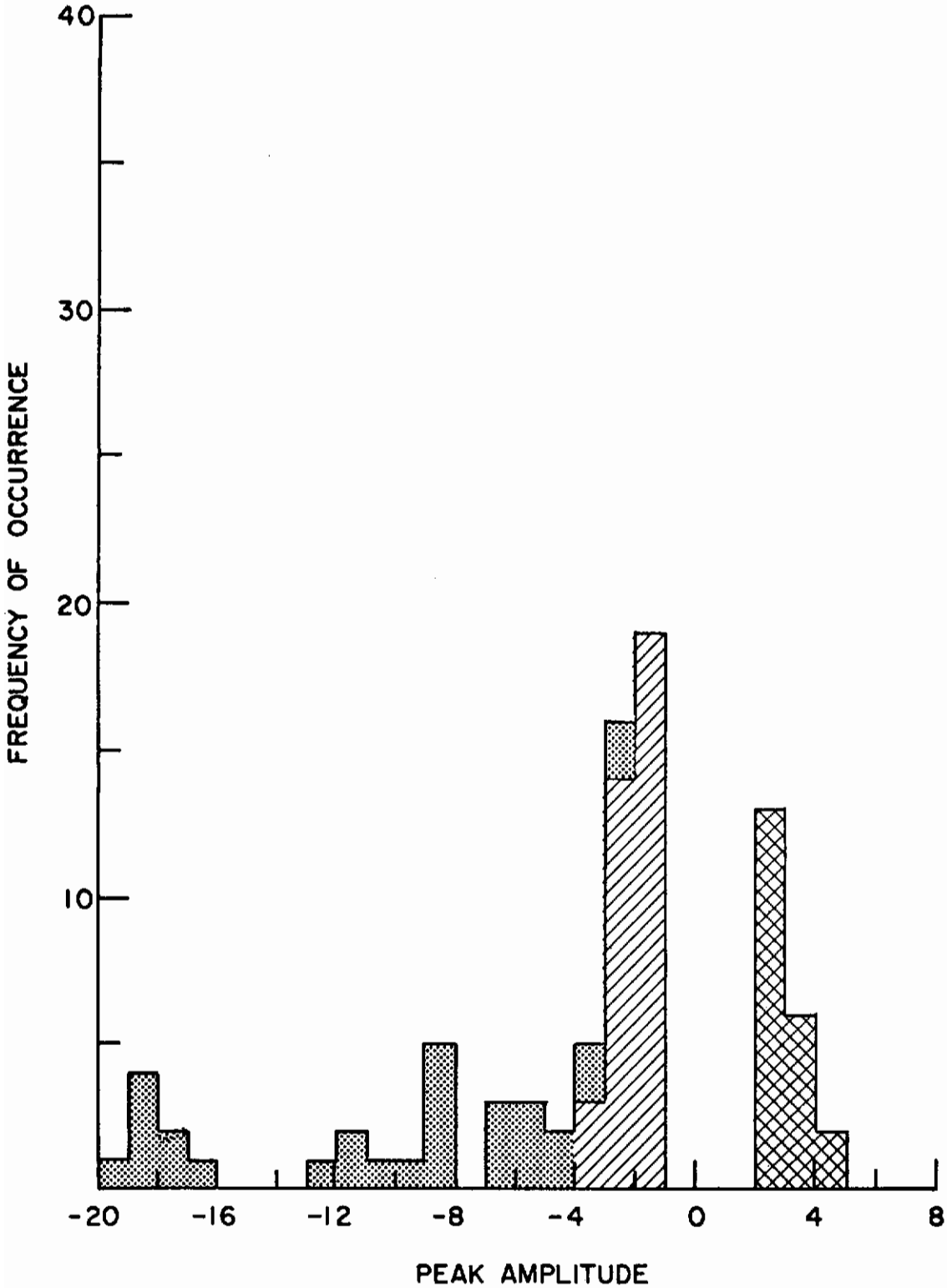

Figure 2 Histogram of the peak amplitude of 95 samples of spikes of spontaneous activity in the ventral nerve cord of a crayfish.

SECTION VI

EXAMPLE OF RECOGNITION: POWER-SPECTRUM OF VOWELS

Figure 3 represents results obtained by applying the L-K method to the deviation from the average of the power-spectra (obtained by 36 bandpass filters) of twelve different vowels (i, I, ɛ, a, e, o, ʌ, u, U, ɔ, ɝ, ɚ) spoken by 19 different persons (male and female). The center-frequencies of the filters range from 100 to 10,000 cycles per second and their bandwidths from 50 to 1,200 cycles per second.

Figure 3 represents 12×19 points in the space defined by $\tan^{-1}(c_2/c_1)$ and $\tan^{-1}(c_3/\sqrt{c_1^2 + c_2^2})$, where c_1, c_2, and c_3 are the first, second and third L-K coefficients. Figure 3 shows a rough determination of 9 disjoint regions, purposely tolerating intrusion of some "alien" elements in each territory in order to avoid excessive gerrymandering. This zoning can be considered as an aid for recognition. The o and the ɔ occupy a common region, and this may lie in the nature of things. The distinction between I and e is so small that a confusion of these two may be forgiven at this stage of the game. The most troublesome one is ɝ (er, ir, or ur in the usual spelling) which intrudes different regions. However, figure 4, which represents the distribution along c_4, shows that the ɝ forms almost a separate island in the c_4 space. It is to be expected from the beginning that considerable overlapping of regions is inevitable in speech recognition. We should probably be surprised that the extent of overlapping is not more than seen on Figure 3. Location of a newly arriving signal in one

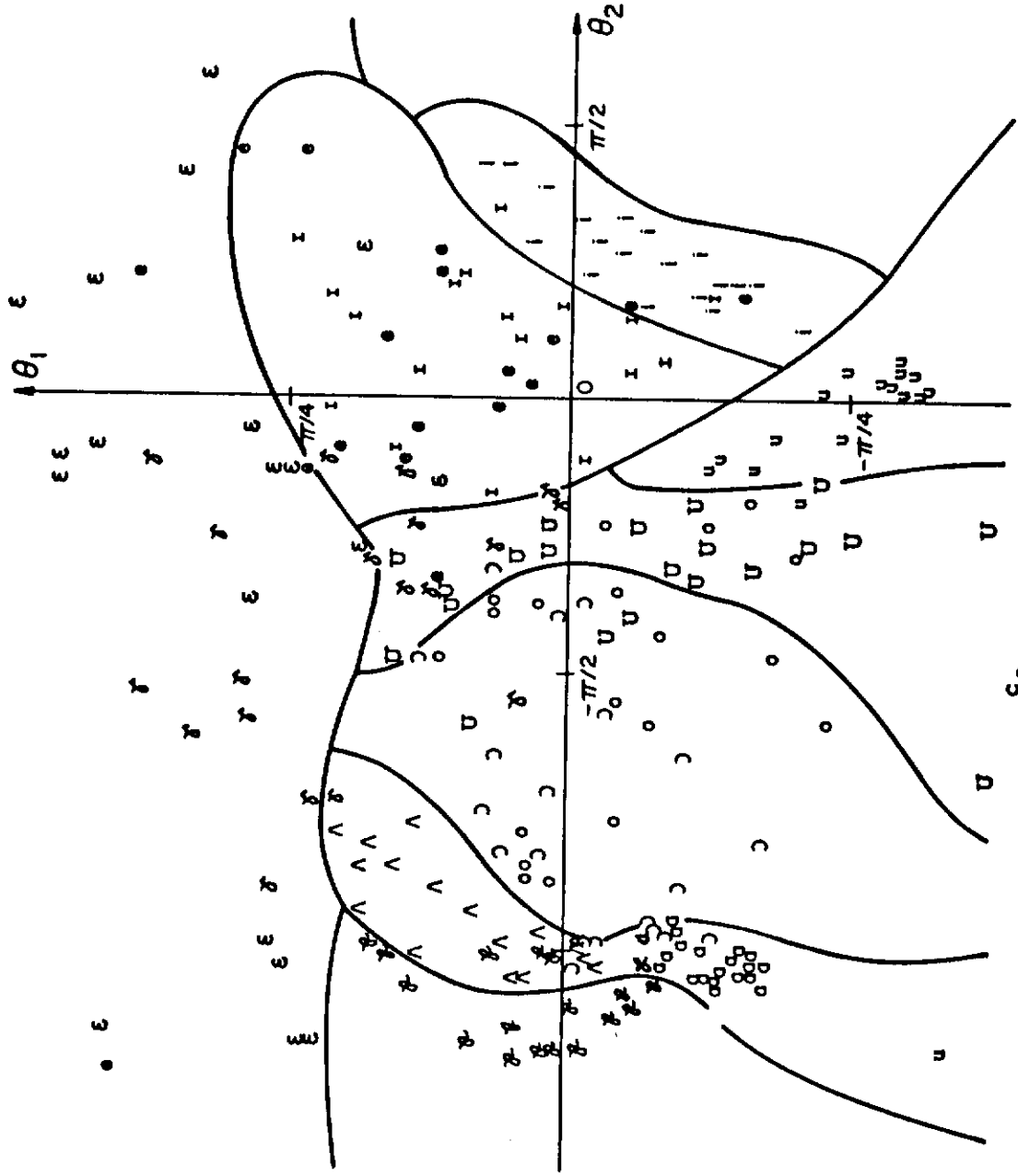

Figure 3 $\theta_1 = \tan^{-1} \frac{c_2}{c_1}$ vs. $\theta_2 = \tan^{-1} \frac{c_3}{\sqrt{c_1^2 c_2^2}}$ representing the separation of samples of twelve vowel sounds by the first three L-K coefficients.

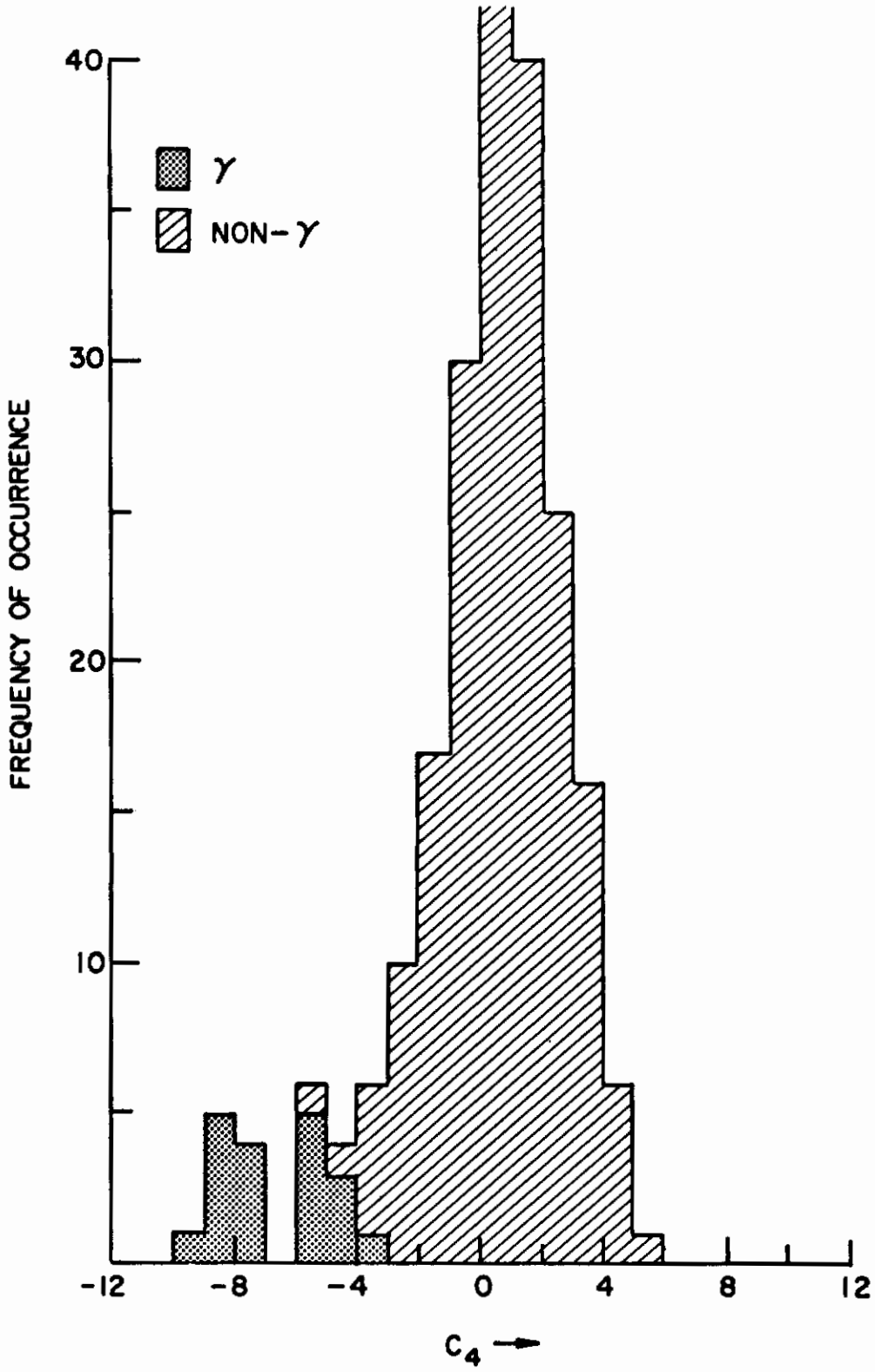

Figure 4 Histogram of the fourth L-K coefficient for γ and non γ sounds.

of the nine zones can be done by a machine. A better result is to be expected if we make smaller meshes than these nine zones and apply the Bayesian decision method. Automatic determination of zones by a computing machine may be done by a method like that of M. A. Aizerman[1], but the reward may not be as great in the present problem in comparison with the complicated computer calculation required.

---

1. See, e.g., Aizerman, M. A., "Experiments on Teaching Machines to Recognize Visual Images" in Biological Aspects of Cybernetics USSR Publishing House, Moscow 1962.

APPENDIX I

DISCUSSION OF EQUALITY CONDITION FOR THE H-THEOREM

To discuss the case where there are more than one OFSIC, we have to complete lemma 1.

Lemma 1 - continuation. Equality in eq 27 holds if and only if the p_i's having indices belonging to each "terminally connected" family have the same value.

Definition. Two indices i and k are said to be "terminally connected" (a term borrowed from a consideration of Markov chains, hence it may sound a little awkward here) if there exists a chain of indices $(i, \ldots, m, n, \ldots, k)$ in which each pair of consecutive indices, say, m and n, are such that there is, at least, one index r which satisfies $A_{rm} \neq 0$ and $A_{rn} \neq 0$. A family of terminally connected indices is such that any two members of the family are terminally connected and a member and a non-member of the family are not terminally connected.

Applying this idea to our problem, let us assume that $\{\psi_i\}$ and $\{\phi_j\}$ are such that equality in eq 27 holds. The lemma then tells us that each non-degenerate eigenfunction ϕ_j constitutes a family with only one member, and that more than one eigenfunction corresponding to a multiple eigenvalue can constitute a terminally connected family. Now, $A_{ij} = 0$ means that ϕ_j and ψ_i are orthogonal and $A_{ij} \neq 0$ means that they have mutually non-vanishing projections. Hence, if a subspace subtended by a certain number, say, d, of ϕ's is also subtended by d ψ's in such a way that no subspace of lesser dimension within it is shared by ϕ's and ψ's, then the d ϕ's are "terminally connected". From this we can conclude that $\{\phi_j\}$ and $\{\psi_i\}$ yield the same entropy value if and only if they are systems of eigenfunctions of the same matrix G and they differ only through the arbitrariness allowed by the degeneracy of the eigenvalues.

APPENDIX II

PROOF OF THEOREM 2

The proof given below of theorem 2 stems from Professor S. Kakutani.

Lemma 2. For any given integer n and for a given double-stochastic matrix $A_{ij}(i, j = 1, 2, \ldots, \infty)$, there exists another double-stochastic matrix B_{ij}, such that $B_{ij} = 0$ for $i > n$ or $j > n$ and $B_{ij} \geq A_{ij}$ for $1 \leq i \leq n$ and $1 \leq j \leq n$.

Proof. Let D denote the class of $n \times n$ matrices with non-negative elements $C_{ij}(i, j = 1, 2, \ldots, n)$ which satisfy

$\sigma_j = \sum_{i=1}^{n} C_{ij} \leq 1$ and $\sigma_i = \sum_{j=1}^{n} C_{ij} \leq 1$. A row (column) whose row-sum σ_i (column-sum σ_j) is less than unity it is said to be deficient. The total number δ of deficient rows and columns is called the degree of deficiency. The degree zero, $\delta = 0$, means that the $n \times n$ matrix is double stochastic. A simple fact to note about the deficient rows and columns is that if there is one or more deficient rows (columns) then there is, at least, one deficient column (row). This can be seen easily if one notices that the sum of all row-sums is equal to the sum of all column-sums $\sum_{i=1}^{n} \sigma_i = \sum_{j=1}^{n} \sigma_j$. We shall now prove the following statement by induction with respect to δ. If $||C_{ij}||$ is a matrix belonging to D and has degree of deficiency δ, then there exists another matrix $||C_{ij}^{*}||$ belonging to D with degree $\delta - 1$ or $\delta - 2$ such that $C_{ij}^{*} \geq C_{ij}$. Among δ deficient row-sums and column-sums of $||C_{ij}||$, let the row-sum σ_k be (one of) the largest. (The argument goes exactly the same way when a column-sum is the largest). Due to the above mentioned fact, there is, at least, one column, say, the ℓ^{th}, which is deficient since there is a deficient row $\sigma_k < 1$. Increase the element $C_{k\ell}$ until its k^{th} row becomes non-deficient. The matrix thus obtained will be denoted $||C_{ij}^{*}||$. This matrix $||C_{ij}^{*}||$

differs from $||C_{ij}||$ only by one element at the intersection of the k^{th} row and the ℓ^{th} column, and $C^*_{k\ell} > C_{k\ell}$, hence we can write in general $C^*_{ij} \geq C_{ij} (i, j = 1, 2, \ldots, n)$. The degree of deficiency of $||C^*_{ij}||$ will be either $\delta - 1$ or $\delta - 2$ according as $\sigma_k > \sigma_\ell$ or $\sigma_k = \sigma_\ell$. The case $\sigma_k < \sigma_\ell$ is excluded since σ_k is the largest non-unity sum. We can continue this process until the degree of deficiency is two, and then these two deficient rows will be eliminated by one stroke. (The case $\delta = 1$ cannot happen). Coming back to the statement of the lemma, let the matrix $||A^*_{ij}||$ be defined as an $n \times n$ matrix such that $A^*_{ij} = A_{ij}$ for $1 \leq i \leq n$ and $1 \leq j \leq n$. Then, this $||A^*_{ij}||$ is a member of D, and the above proof shows that there exists a non-deficient (double-stochastic) matrix $||B^*_{ij}||$ belonging to D such that $B^*_{ij} \geq A^*_{ij} = A_{ij}$ for $1 \leq i \leq n$, $1 \leq j \leq n$. The B_{ij} of the lemma can be defined by $B_{ij} = B^*_{ij}$ for $1 \leq i \leq n$, $i \leq j \leq n$ and $B_{ij} = 0$ otherwise. This completes the proof of the lemma.

Proof to Theorem 2. Let $\epsilon_{ij} (i, j = 1, 2, \ldots, n)$ be defined by

$$B_{ij} = A_{ij} + \epsilon_{ij} \tag{36}$$

where A_{ij} and B_{ij}, $(i, j = 1, 2, \ldots, \infty)$ are related as in lemma 2. Since the matrices A_{ij} and B_{ij} are double-stochastic, we have

$$\sum_{j=1}^{n} \epsilon_{ij} = \sum_{j=n+1}^{\infty} A_{ij} \tag{37}$$

We have then

$$\sum_{j=1}^{n} \lambda_j = \sum_{i=1}^{n} \sum_{j=1}^{n} B_{ij} \lambda_j \qquad \left(\text{since } \sum_{i=1}^{n} B_{ij} = 1\right)$$

$$\geq \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \lambda_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_{ij} \lambda_{n+1} \quad \text{(due to eq 30 and eq 36)}$$

$$\geq \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \lambda_j + \sum_{i=1}^{n} \sum_{j=n+1}^{\infty} A_{ij} \lambda_j$$

(due to eq 30 and eq 37)

Therefore

$$\sum_{j=1}^{n} \lambda_j \geq \sum_{i=1}^{n} \rho_i \qquad \text{(due to eq 24)} \qquad (38)$$

This completes the proof.

REFERENCES

1. R. Bakis and G. Hu, private communication.
2. H. Kramer and M. Mathews, "A Linear Coding for Transmitting a Set of Correlated Signals", IRE Transactions on Information Theory, 2, pp 41-46, September 1956.
3. M. Loève, Probability Theory, 3rd Edition, Van Nostrand, Princeton, 1963.
4. S. Watanabe, "Mathematical Explication of Classification of Objects", Proceedings of International Symposium on Information and Prediction in Science, Brussels, 1962, ed. Dockx, Bernays, Academic Press, New York, in press.
5. E. Wong, "Vector Stochastic Processes in Problems of Communication Theory", Ph.D. Thesis, Princeton University, May, 1959.

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)
Thomas J. Watson Research Center, International Business Machines Corp., P.O. Box 218, Yorktown Heights, New York

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP
N/A

3. REPORT TITLE
THE LOÈVE-KARHUNEN EXPANSION AS A MEANS OF INFORMATION COMPRESSION FOR CLASSIFICATION OF CONTINUOUS SIGNALS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Final report, 15 July 1963 - 14 July 1964

5. AUTHOR(S) (Last name, first name, initial)
Watanabe, Satosi

6. REPORT DATE
October 1965

7a. TOTAL NO. OF PAGES
26

7b. NO. OF REFS
5

8a. CONTRACT OR GRANT NO. AF 33(657)-11347

b. PROJECT NO 7233

c. Task No. 723305

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
AMRL-TR-65-114

10. AVAILABILITY/LIMITATION NOTICES
Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson AFB, Ohio

13. ABSTRACT
The classification of an ensemble of continuous signals $[f^{(\alpha)}(x)]$ is accomplished by expansion in terms of a set of orthogonal functions $[\psi_i(x)]$. The coefficients $c_i^{(\alpha)}$ of the expansion represent the desired information content of the continuous signals. A probability measure ρ_i with respect to $f^{(\alpha)}(x)$ is defined on $[\psi_i(x)]$. That set of functions $[\phi_j(x)]$ is desired which concentrates the ρ's on a few functions instead of distributing them over many. This is accomplished by minimization of the entropy function $S([\psi_i]) = -\sum_{i=1}^{\infty} \rho_i \log \rho_i$ with respect to $[\psi_i]$. The ρ_i are shown to be the diagonal terms of the autocorrelation matrix G, whose complete set of eigenfunctions is a desired set of orthogonal functions and whose eigenvalues represent the associated ρ's. The expansion of the set $[f^{(\alpha)}(x)]$ in terms of the eigenfunctions of the autocorrelation matrix is the Loève-Karhunen (L-K) expansion. This expansion in addition minimizes the mean square error which results from truncating the expansion. One application of the method is the clustering of ninety five samples of the spontaneous activity potential in the ventral nerve cord of a crayfish. The first L-K expansion coefficient divides these samples into three major groups. In another application samples of the power spectrum of twelve isolated vowel sounds spoken by nineteen persons are separated into roughly nine disjoint regions. This is accomplished by using the first three L-K coefficients. Use of the fourth coefficient discriminates the most troublesome remaining vowel from the rest.

DD FORM 1 JAN 64 1473

AF-WP-B-AUG 64 400

Security Classification

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Signal classification Information compression Automatic data processing						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

Security Classification