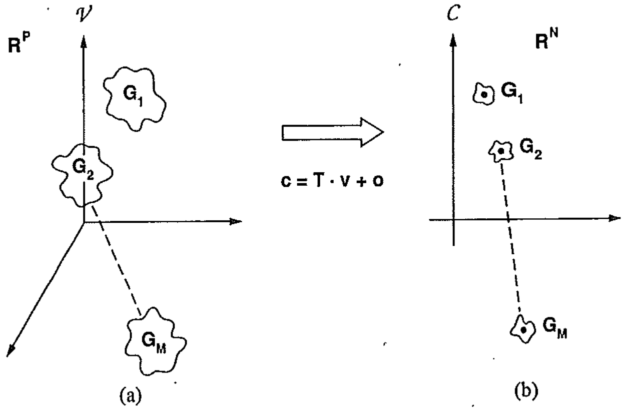
Minimum Mean-Square Error Transformations of Categorical Data to Target Positions

Stephen A. Zahorian, *Member, IEEE*, and Amir Jalali Jagharghi, *Member, IEEE*

***Abstract*—A new algorithm is described for transforming multidimensional data such that all the data points in each of several predefined categories map toward a category target position in the transformed space. The procedure is based on minimizing the mean-square error between specified category target positions and actual transformed locations of the data. Least squares estimation techniques are used to derive linear equations for computing the transformation coefficients and for determining an origin offset in the transformed space. However, for additional flexibility in the transformation, a method is presented for combining the linear transformation with a nonlinear connectionist network transformation. This procedure can, among other things, be used as a tool to evaluate the precision with which physical measurements of psychophysical stimuli correlate with the perceptual configuration of those stimuli. Potential speech science applications include automatic phoneme recognition, evaluating the power of acoustic features in predicting phonetic categories, speech processing for sensory substitution devices, and speaker normalization. Experimental results illustrate some of these applications with vowel data.**



(a) (b)

Fig. 1. Transformation of categorical data from R^P to R^N.

I. Introduction

A COMMON goal in a wide variety of disciplines, including speech science, psychology, anthropology, economics, and political science, is to correlate continuous-valued measurements of ''objects'' with perceptual or interpretable variables which also describe those objects. For example, in speech science, a fundamental objective of many research studies is to link the acoustic measurements of a speech signal to variables which describe the perception of that signal. In both speech science and other applications these objects also can often be classified in terms of predefined groups or categories, such that objects within a group are more similar to each other than to the objects in another group. Fig. 1 depicts this scenario in stylized form. As a more specific example, each object could correspond to a spoken vowel. Each data point in the original space of Fig. 1 would then correspond to a set of acoustic measurements, such as formant values, for one repetition of one vowel. A large number of data points comprise each group as measurements from many repetitions of a particular vowel from one or more speakers. Repeating this process for other phonologically distinct vowels, data points are obtained for the other groups. The transformation should then map the measurements of phonologically equivalent vowels to clustered regions in the transformed space and the measurements of phonologically distinct vowels to separate regions. Such a transformation is possible only if the phonologically distinct vowels cluster to some degree in the measurement space with respect to the specific phonetic categories.

In the general case, if data can be assigned to categories, and if the data within each category in the original space are more similar in some sense to other data in that group than to data in other groups, a transformation is to be derived such that all the data points from each group in the original space map to tightly clustered regions in the transformed space. Ideally, clusters corresponding to different groups should be widely separated. Typically, the number of dimensions in the transformed space is small (for example, two dimensions enable convenient visual interpretations of the data); conversely, the number of dimensions in the original space is generally large compared to the number of dimensions in the transformed space. Although there is no fundamental requirement that the transformation between the two spaces be linear, the procedure developed in this paper, which we refer to as minimum mean-square error coordinate transformation (MSECT), is a linear transformation, i.e., matrix multiplication, plus a translation of the origin. The restriction to a linear transformation plus an offset enables a straightforward solution for computing the transformation coefficients. However, a method is also described for combin-

Manuscript received August 12, 1989; revised October 22, 1990. This work was supported in part by NSF Grant IRI-8702649.

S. A. Zahorian is with the Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA 23529.

A. J. Jagharghi was with the Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA 23529. He is now with Vegyan Inc.

IEEE Log Number 9104024.

ing MSECT with a nonlinear connectionist (neural) network transformation.

The application of the general procedure described requires several user choices: dimensionality and selection of the original variables; assignment of data to distinct categories; dimensionality of the transformed space; and, finally, target positions for cluster centers in the transformed space. This procedure only applies if these choices can be made, especially the selection of the target space and the locations of target positions in this space. Potential solutions to this fundamental source of difficulty, that is, the specification of the target space, depend on the particular application. For example, if the procedure were to be applied to the matching of acoustic measurements of speech stimuli to a perceptual configuration for those stimuli, targets could be chosen as the configuration of stimuli derived from multidimensional scaling [11] of perceptual data. For this application, the MSECT procedure described in this paper could be used to assess the ability of different sets of measurement variables to predict perceptual variables. That is, better measurement variables will result in better matches, in the sense of lower mean-square error, to target points in the perceptual space.

The organization of the remainder of this paper is as follows. In the next section we present a mathematical derivation for computing the transformation coefficients, based on the minimization of mean-square error. The following section is a brief discussion of existing multidimensional transformation techniques which relate to the algorithm presented in this paper. The use of a connectionist network "preprocessor" for the linear transformation is presented. A section on experimental results presents several examples of applications of the transformation in speech processing. These examples also illustrate three different methods for selecting target positions in the transformed space. Finally, a summary of the advantages, disadvantages, and potential applications of the procedure is given.

II. Mathematical Derivation

In this paper script letters, such as $\mathcal{V}$, refer to a vector space; uppercase boldface letters, such as $\boldsymbol{V}$, refer to a matrix each of whose columns is a point in $\mathcal{V}$; and lowercase boldface letters, such as $\boldsymbol{v}$ are the vector coordinates of a point in $\mathcal{V}$.

Let $\mathcal{V}$ be a P-dimensional real continuous-valued vector space. Similarly, let $\mathcal{C}$ be an N-dimensional real continuous-valued vector space. Thus data points in $\mathcal{V}$ and $\mathcal{C}$ are vectors in R^P and R^N, respectively. Further assume that a collection of data can be classified or categorized in terms of M predefined groups, with data points within a group more "similar" to other data points in that group than to the data in other groups. Let each data category have a target position in $\mathcal{C}$ about which transformed data are expected to be well clustered. Denote the number of data points in each category as $NS(j)$, $j = 1$ to M. A linear transformation $\boldsymbol{T}$ and an offset $\boldsymbol{o}$ are desired which map points in $\mathcal{V}$ to points in $\mathcal{C}$ as follows:

$$c = \boldsymbol{T} \cdot \boldsymbol{v} + \boldsymbol{o}. \tag{1}$$

The transformation matrix $\boldsymbol{T}$ and the offset vector $\boldsymbol{o}$ are to be found such that the mean-square error, averaged over the entire data set, between specified target positions in $\mathcal{C}$ and projections from the measurement space, is minimized. That is, $\boldsymbol{T}$ and $\boldsymbol{o}$ will transform data from $\mathcal{V}$ to $\mathcal{C}$ such that the data in each category will cluster about its specified target position in $\mathcal{C}$ as well as possible in the mean-square error sense. This error can be formulated as follows. Choose M vectors in $\mathcal{C}$, each of which represents a distinct target position in the transformed space. From these vectors, form the N-by-M matrix $\boldsymbol{C}$, each of whose columns is one of the target vectors chosen. Thus $\boldsymbol{C}_{ij}$ is the ith component of the jth target vector. Similarly, form a triple-indexed array $\boldsymbol{V}$, such that $\boldsymbol{V}_{ijk}$ is the ith component of the kth data point in category j in the original space. With this notation, the mean-square "projection" error for a particular data point in the jth category is given by

$$D_{jk} = \sum_{i=1}^{N} \left(\boldsymbol{C}_{ij} - \left(\sum_{l=1}^{P} \boldsymbol{T}_{il} \boldsymbol{V}_{ljk} \right) - \boldsymbol{o}_i \right)^2. \tag{2}$$

Summing over all M data categories and the number of data points in each category $NS(j)$, the total mean-square projection error is

$$D = \sum_{j=1}^{M} \sum_{k=1}^{NS(j)} \sum_{i=1}^{N} \left(\boldsymbol{C}_{ij} - \left(\sum_{l=1}^{P} \boldsymbol{T}_{il} \boldsymbol{V}_{ljk} \right) - \boldsymbol{o}_i \right)^2. \tag{3}$$

We thus need to determine all the elements in $\boldsymbol{T}$, $\boldsymbol{T}_{mn}$, $1 \leq m \leq N$, $1 \leq n \leq P$, and all elements in $\boldsymbol{o}$, $\boldsymbol{o}_m$, $1 \leq m \leq N$, such that D is minimized. This can be accomplished by taking the partial derivatives of D with respect to each $\boldsymbol{T}_{mn}$ and each $\boldsymbol{o}_m$ and setting these derivatives to zero. Thus the problem is a linear mean-square estimation problem which consists of computing the minimum of a function of $(P + 1) \cdot N$ variables. Thus a total of $(P + 1) \cdot N$ independent equations are needed to solve for all unknowns. With this approach, the following set of equations is obtained:

$$\frac{\partial D}{\partial \boldsymbol{T}_{mn}} = \sum_{j=1}^{M} \sum_{k=1}^{NS(j)} -2\left[\boldsymbol{C}_{mj} - \left(\sum_{l=1}^{P} \boldsymbol{T}_{ml} \boldsymbol{V}_{ljk} \right) - \boldsymbol{o}_m \right] \boldsymbol{V}_{njk} \quad \text{for} \quad \begin{matrix} 1 \leq m \leq N \\ 1 \leq n \leq P \end{matrix} \tag{4}$$

and

$$\frac{\partial D}{\partial \boldsymbol{o}_m} = \sum_{j=1}^{M} \sum_{k=1}^{NS(j)} -2\left[\boldsymbol{C}_{mj} - \left(\sum_{l=1}^{P} \boldsymbol{T}_{ml} \boldsymbol{V}_{ljk} \right) - \boldsymbol{o}_m \right] \quad \text{for } 1 \leq m \leq N. \tag{5}$$

Letting these derivatives equal zero, results in the required $(P + 1) \cdot N$ equations:

$$\sum_{j=1}^{M} \sum_{k=1}^{NS(j)} C_{mj} V_{njk} = \sum_{j=1}^{M} \sum_{k=1}^{NS(j)} \left(\sum_{l=1}^{P} T_{ml} V_{ljk} + o_m \right) V_{njk} \quad (6)$$

$$\text{for} \quad \begin{array}{c} 1 \leq m \leq N \\ 1 \leq n \leq P \end{array}$$

and

$$\sum_{j=1}^{M} \sum_{k=1}^{NS(j)} C_{mj} = \sum_{j=1}^{M} \sum_{k=1}^{NS(j)} \left(\sum_{l=1}^{P} T_{ml} V_{ljk} + o_m \right)$$

$$\text{for } 1 \leq m \leq N. \quad (7)$$

Equations (6) and (7) can be combined and rewritten as N sets of matrix equations, with $P + 1$ unknowns in each equation, each of the form

$$\boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{b}. \quad (8)$$

In (8), $\boldsymbol{A}$ is a $(P + 1)$ by $(P + 1)$ square matrix, $\boldsymbol{x}$ is a $(P + 1)$ column vector of unknowns, and $\boldsymbol{b}$ is a $(P + 1)$ column vector.

The mth matrix equation is used to solve for the mth row of $\boldsymbol{T}$ and the mth component of $\boldsymbol{o}$. The elements of $\boldsymbol{A}$ for the mth matrix equation are

$$A_{pq} = \sum_{j=1}^{M} \sum_{k=1}^{NS(j)} V_{qjk} V_{pjk}, \quad 1 \leq p \leq P, 1 \leq q \leq P \quad (9)$$

$$A_{pq} = \sum_{j=1}^{M} \sum_{k=1}^{NS(j)} V_{pjk}, \quad 1 \leq p \leq P, q = P + 1 \quad (10)$$

$$A_{pq} = \sum_{j=1}^{M} \sum_{k=1}^{NS(j)} V_{qjk}, \quad p = P + 1, 1 \leq q \leq P \quad (11)$$

$$A_{pq} = \sum_{j=1}^{M} \sum_{k=1}^{NS(j)} 1 = \sum_{j=1}^{M} NS(j),$$

$$p = P + 1, q = P + 1. \quad (12)$$

The unknowns in $\boldsymbol{x}$ are

$$x_p = T_{mp}, \quad 1 \leq p \leq P \quad (13)$$

and

$$x_p = o_m, \quad p = P + 1. \quad (14)$$

Finally, the elements of $\boldsymbol{b}$ are given by

$$b_p = \sum_{j=1}^{M} \sum_{k=1}^{NS(j)} C_{mj} V_{pjk}, \quad 1 \leq p \leq P \quad (15)$$

and

$$b_p = \sum_{j=1}^{M} \sum_{k=1}^{NS(j)} C_{mj}, \quad p = P + 1. \quad (16)$$

The matrix $\boldsymbol{T}$ and the offset vector $\boldsymbol{o}$ are determined by solving all N sets of matrix equations. $\boldsymbol{A}$ is the same for each of the N sets of equations and thus must only be computed once, rather than N times; $\boldsymbol{A}$ can also be computed more efficiently by taking advantage of its symmetric form. The vector $\boldsymbol{b}$ is a function of m. Thus solving for each row of $\boldsymbol{T}$ and corresponding element in $\boldsymbol{o}$ is an independent problem. In effect, each matrix equation represents a special case of multiple linear regression analysis of the original variables to a single coordinate in the transformed space. However, unlike typical applications of regression analysis, the regression is performed with a single target value for each data category. Nevertheless, since all variables are continuous and the transformation is linear, the objective is merely to transform data to the vicinity of these target values. We call this transformation minimum mean-square error coordinate transformation (MSECT). A unique solution is guaranteed if the matrix $\boldsymbol{A}$ is nonsingular and thus invertible. Note that the P-by-P submatrix consisting of the first P rows and P columns is a scaled correlation matrix. Thus the rank of $\boldsymbol{A}$ is at least P, provided no measurement variable is completely dependent (and thus supplying only redundant information) on the other variables. The last row of $\boldsymbol{A}$ consists of scaled mean values for each of the original variables, augmented by the scale factor (the number of data points) in the last position. Except for very unusual combinations of data, this row will be independent of the preceding P rows. In numerous examples conducted with speech spectral data in our lab, numerical instability problems in the inversion of $\boldsymbol{A}$ have not been encountered for values of P up to 16.

Three basic properties of the transformation procedure can be shown readily from the formulation.

1) The match to the target positions, as measured by the error D, is independent of the scaling and offsets of the variables in $\mathcal{V}$. Equations (9)–(16), which are used to compute the solution for $\boldsymbol{T}$ and $\boldsymbol{o}$, indicate if the transformation $\boldsymbol{T}$ and offset $\boldsymbol{o}$ are obtained with a given scaling and offset for the lth variable in $\mathcal{V}$, and if this variable is then scaled by s and shifted by z, i.e., $V'_{ljk} = s \cdot V_{ljk} + z$, $1 \leq j \leq M$, $1 \leq k \leq NS(j)$, the new solution is given by

$$T'_{il} = (1/s) \cdot T_{il} \quad (17)$$

and

$$o'_i = o_i - \sum_{l=1}^{P} T_{il} \cdot z/s$$

$$\text{for } 1 \leq l \leq N. \quad (18)$$

This transformation and offset, applied to the modified variables, results in the same D value obtained with the original transformation and offset and the original variables (3).

2) The match to the target positions, as measured by a ratio of between-cluster variance to within-cluster variance, is independent of a uniform expansion, contraction, or offset of all target points in $\mathcal{C}$. Equations (9)–(16) show that if every element in the matrix $\boldsymbol{C}$ is scaled by s, then every element in the solution $\boldsymbol{T}$ and $\boldsymbol{o}$ will also be scaled by s. Therefore, every transformed data point in $\mathcal{C}$ will

be similarly scaled, thus preserving the ratio of between-cluster to within-cluster variance. Similarly, if a constant z is added to every element in $\mathcal{C}$, z is also added to every element in $\boldsymbol{o}$ and $\boldsymbol{T}$ is unchanged, thus preserving the ratio of variances.

3) The dimensionality N of the transformed space can be any desired value. However, since $\boldsymbol{T}$ has P columns, the rank of $\boldsymbol{T}$ cannot be larger than P. Similarly, since the target matrix $\boldsymbol{C}$ has M columns, at most M rows of $\boldsymbol{C}$ are linearly independent, which in turn limits the rank of $\boldsymbol{T}$ to M. Combining these two restrictions, along with the observation that the number of linearly independent variables in $\mathcal{C}$ is limited by the rank of $\boldsymbol{T}$, the maximum number of linearly independent variables in $\boldsymbol{C}$ is the minimum of P and M. Therefore, unless linear dependence of the transformed data is desired, N should not exceed the smaller of M and P.

A Fortran program has been written to implement the MSECT transformation.[1] The program computes the elements of $\boldsymbol{A}$ and $\boldsymbol{b}$ (8) using (9)–(12) and (15), (16). A standard matrix inversion routine [17] is used to solve (8) for the elements of $\boldsymbol{T}$ and $\boldsymbol{o}$. The computation times for the elements of $\boldsymbol{A}$ and $\boldsymbol{b}$, respectively, are approximately proportional to $P^2 \cdot L$ and $N \cdot P \cdot L$, where L is the total number of data points. One $(P + 1) \cdot (P + 1)$ matrix must be inverted. All arithmetic operations are performed with single precision (32-b) real arithmetic except the matix inversion which is performed with double precision (64-b) arithmetic. This program has been used to compute transformations for many examples and has been used to experimentally verify the three properties listed above.

III. Discussion

The algorithm developed in this paper bears resemblance to several multidimensional statistical techniques which can be used to examine groups of objects with respect to several variables simultaneously. In most of these techniques, a linear transformation projects data to a reduced-dimensionality subspace such that certain structural characteristics of the data are preserved in the transformed space. For example, the transformation might be based on maximizing the percentage of variance accounted for in the reduced-dimensionality space. The transformed space may be used to simplify automatic classification, to help "discover" assumed fundamental factors which underlie the measured variables, to compare the structure of data from two spaces, or, as an aid for interpreting the relationships among the groups under investigation. These techniques include principal-components analysis, discriminant analysis, multidimensional scaling, and generalized Procrustes transformations. Since these established methods are discussed in detail in many references, we merely give a few comments concerning the relationship of these techniques to the algorithm presented in this paper. Both principal-components analysis (PCA) [8], [9] and linear discriminant analysis (LDA) [2], [5], [12] can be used to derive linear transformations for dimensionality reduction. For PCA, the transformation maximizes the total data variance, or signal energy, "captured" in the reduced-dimensionality space. For LDA, the transformation maximizes the ratio of between-group to within-group data variance, where the groups or categories are defined by the user of the transformation.

The mathematical derivations of both PCA and LDA are quite different from the mathematical derivation for MSECT given in a previous section, since both of these transformations are orthogonal transformations derived solely from the data itself, with no provisions for arbitrarily specified target positions in the transformed space. Both of these procedures have been used to transform acoustic measurements of speech stimuli to low-dimensionality spaces (usually two to five dimensions) [10], [21], [27] for the purposes of data reduction, automatic speech recognition, visual speech displays for speech articulation training, or to compare the configuration of the data in a physical measurement space to a perceptual configuration of the data. For the first application mentioned, PCA is the desired technique. For the second application, LDA would appear to be best suited. The MSECT method developed in this paper is, however, best suited for the third and fourth applications.

Multidimensional scaling (MDS) can be used to derive multidimensional spaces such that experimentally determined perceptual similarities correspond to distances in the derived space [1], [11]. For typical phonetic applications in speech science, the inputs to MDS are confusion matrices obtained from a perceptual experiment. Each element in the confusion matrix is an estimate of perceptual similarity between two phonemes. The confusion matrix is converted to a proximity matrix of similarity scores—larger confusions imply greater similarity values. Using a variety of criteria, according to the details of the particular method chosen, the model forms a multidimensional configuration such that distances between the category positions correspond as well as possible to proximity values. As the dimensionality of the derived space increases, the match between proximities and distances improves. MDS has no direct relationship to MSECT, PCA, or LDA. However, MDS does provide an analytical method for deriving a perceptual configuration of speech stimuli. MSECT can then be used to evaluate the degree to which acoustic measurements of speech stimuli can be transformed to this perceptual configuration.

The class of procedures most similar to MSECT are called "Procrustes transformations" in the literature [1]. These procedures can be used to compare the structure of data configurations in two multidimensional vector spaces. Although there are many variations, in the basic formulation, an orthogonal rotation is used to best match one multidimensional "object" or configuration to another

[1] To obtain a copy of this program, and examples for testing, send a self-addressed and stamped 5-1/4-in floppy disk mailer with a 1.2 megabyte 5-1/4-in DOS formatted floppy diskette to the authors.

multidimensional "object" or configuration. In contrast to MSECT, each "object" consists of the same number of points. The basic orthogonal rotation has been enhanced in "generalized Procrustes analysis" to include translation and uniform scaling so that the two sets of data points best match in a mean square error sense [7], [22]. Procrustes analysis has been applied to problems in many fields, including the comparison of physical and perceptual spaces for speech stimuli [10], [20].

Although both Procrustes transformations and the MSECT algorithm are similar in that each provides a systematic method for comparing the configurations of data in two multidimensional vector spaces, the mechanism and potential applications for the two techniques are quite different. The intent of a Procrustes transformation is to compare the structure of two objects, each defined by the same number of points, without distorting the shape of the objects. Therefore, for the class of problems alluded to in this paper, the Procrustes method can only be used if group centroids in the first space are orthogonally transformed to target positions in the second space. In contrast, the MSECT algorithm makes use of an arbitrary (i.e., no requirement for orthogonality) linear transformation to best map all data in the first space to the target positions in the second space. The flexibility of MSECT is thus constrained by the requirement that a large number of data points be transformed by a single transformation, rather than only group centroids. The Procrustes method requires an orthogonal transformation since the primary objective is to compare the structure of two configurations. Property 1, invariance of matching to target positions with respect to scaling and offsets of the original variables (Section II), does not apply to Procrustes transformations, since arbitrary independent scalings of the original variables may distort the structures under investigation. With MSECT, a less restrictive transformation is allowable since the structure of the (centroid) configuration is of secondary importance to clustering of data within each category.

IV. Nonlinear Transformations

There are naturally fundamental limitations of any linear transformation in projecting measurement data to arbitrary points in a reduced-dimensionality target space such that the data will be well clustered within each category and widely separated from other categories. The success of the transformation depends on how well the data cluster for each category in the measurement space and also the configuration of these clusters relative to the configuration of the specified target points in the transformed space. At least for some conditions, the increased flexibility of a nonlinear transformation will result in better matches to target positions in the transformed space. One approach would be to augment the measurement variables with squared values, cross products of variables, and higher ordered terms, and to perform the linear transformation on the augmented variable set. This approach has resemblance to a generalized principal-components analysis [6]. We have not investigated this method because of the very high-dimensionality spaces that would be required. As an alternative, we have investigated a nonlinear transformation based on a connectionist (neural) network to derive an intermediate coordinate system with highly clustered data within each category such that the data can be projected with much greater accuracy to the final space.

The theory and application of multilayer feedforward connectionist networks to perform recognition are described in many articles (for example, the tutorial [13]). From the viewpoint of this paper, the network can be modeled as a series of linear transformations, separated by nonlinearities. Typically, there is one input to the network for each original dimension in the data space and one output node for each data category. A sigmoid nonlinearity can be used to bound each output between 0 and 1. The network can be trained as a classifier, i.e., so that each output node represents a particular category of data. For the ideal classifier, one output node of the neural network will have value 1.0 for all the data in a given category and all other output nodes will have value 0.0. Thus, for this ideal case, the neural network transforms data for M categories to an M-dimensional space such that all data within each category map to a single point; each category also corresponds to a different point in the M-dimensional output space. These M points are located at positions of unity along the M axes in the transformed space.

Using a matrix representation for this ideal case, the data transformed by the connectionist network maps exactly to an M-by-M identity target matrix. Each column of the identity matrix is the target position in M-dimensional space for one of the M categories. Since the M-by-M identity target matrix can be linearly transformed to M arbitrarily chosen positions in any-dimensionality space, a combination nonlinear/linear transformation can be obtained such that all data from each category transform to a single point in the target space, using *M arbitrarily* chosen target positions for the categories. For real data, of course, the output of the neural network will deviate from the ideal case and the data will therefore disperse in the final space. However, in experiments conducted to date with spectral measurements of both vowels and stop consonants, the use of a neural network as a "preprocessor" for the linear transformation previously described greatly improved the clustering properties of the speech data in the final space. Representative results for vowels are given in a later section of the paper.

V. Experimental Data

To illustrate potential applications of the transformation procedure discussed in this paper, we present several examples of vowel spectral measurements transformed with the MSECT procedure. For the first two examples, the vowel measurements are transformed to two-dimensional spaces and cluster plots illustrate the effects of the transformation. The examples include different methods for selecting target spaces and target positions, as well as a comparison with alternative transformation techniques. The third example points out a potential application of MSECT where the dimensionality of the transformed space equals the dimensionality of the original space.

The stimuli for these examples were the vowel portions of 99 CVC syllables spoken as isolated words by 30 speakers: 10 men, 10 women, and 10 children between the ages of 7 and 11. For each syllable, the initial consonant was one of /b, d, g, p, t, k, h, l, w/, the vowel was one of /a, i, u, ae, ɝ, I, ε, ɔ, ʌ, U, o/, and the final consonant was one of /b, d, g, p, t, k, v, s/. The syllables were selected so that each of the 6 stop consonants was paired at least once with each of the 11 vowels in both initial and final position (66 CV combinations and 66 VC combinations). Since single-syllable meaningful words were used to the extent possible, the syllable list was longer than the minimum number of 66, and the additional consonants /h, l, w, m, s/ were used. About 2/3 of the syllables were meaningful words and the other 1/3 were nonsense syllables. This vowel data presumably has more variability than the Peterson and Barney vowel data [19] (hVd only) because of the large number of consonantal contexts. The stimuli were low-pass filtered at 7.5 kHz and digitized at 16 kHz for computer analysis. A digital waveform editor was used to manually label the "steady state" portion of each vowel token for further analysis. Two types of spectral measurements were then made on each vowel stimulus. For the first case the measurements consisted of the first three formants and the fundamental frequency of voicing (*F*0). For the second case, the measurements consisted of a form of cepstral coefficients (CC's) (2 through 9) and *F*0.[2]

The formants were computed automatically after low-pass filtering the speech signal at 3.8 kHz, decimation to an 8-kHz sampling rate, and high-frequency preemphasis with $H(z) = 1 - 0.75 z^{-1}$. For each 25-ms Hanning window of the speech signal, formant candidates were computed from the roots of a 10th-order LP model polynomial. A tracking algorithm, similar to the one described in [16] was used to select three formants from the candidates in the entire interval of each steady-state vowel signal. Finally the outputs of the tracker for the center frame of each stimulus were used as the formants for that stimulus.[3] *F*0 was computed using a version of the SIFT fundamental frequency algorithm [15], for the center 30-ms Hamming-windowed frame of each vowel segment. Note that a longer window was used for *F*0 analysis so that the analysis window typically spans three or more pitch periods. In order to approximate a perceptual frequency scale of mels (*m*), the formants and *F*0 were converted to a mel frequency scale using the following equation [14]:

$$m = 2595 \log_{10} (1 + f/700). \tag{19}$$

The cepstral coefficients were computed from the original speech signal (i.e., no low-pass filtering at 3.8 kHz and no decimation) over a frequency range of 150 to 5000 Hz, with a preemphasis of $H(z) = 1 - 0.95z^{-1}$, after first warping the complete-range log magnitude spectrum with a bilinear warping factor of 0.6 [18]. The CC's were computed from the center 25-ms Hamming-windowed frame of each vowel segment. The examples are given primarily to illustrate the basic transformation procedure, with "real" data, rather than as an exposition of vowel properties. A detailed discussion of the many sources of variability for vowel data, and of the "best" measurement and perceptual spaces for vowels is beyond the scope of the present paper. There are, of course, many examples in the literature of studies of both physical and perceptual measurements of vowel data [10], [19], [20], [24], [25]. The two measurement sets described above were chosen because they are considerably different in the dimensionality of the physical space (4 versus 9), they represent somewhat different points of view regarding the physical measurements that most correlate with the perception of vowel quality, and they do suffice to illustrate potential realistic speech science applications of MSECT.

Because of the large amount of test data, an ellipse depicts each vowel in each cluster plot, rather than the individual data points. For all plots, 15 "training" speakers (5 men, 5 women, and 5 children) were used to derive a particular transformation. All figures depict the data of the remaining 15 test speakers. Thus the figures represent speaker-independent test data. Each ellipse is placed at the centroid of a vowel and oriented so that its major axis coincides with the direction of maximum data variation for that vowel. The length of the major axis is equal to two standard deviations of the data in the direction of the major axis; the minor axis also equals two standard deviations of the data in the direction of the minor axis. Therefore, assuming Gaussian distributions, each ellipse encompasses about 50% of the data for each vowel. Thus, by visually inspecting the figures, the approximate degree of clustering of the data for each vowel and the separation of the vowels is readily apparent.

To give objective ratings of clustering of vowel data with respect to the 11 phonetic categories, three figures of merit were also computed for each case. First, all the vowel data for the test speakers were classified according to minimum Euclidean distance to the 11 target positions specified for the vowels. The percentage of vowels correctly identified, PCT, was computed. Recall that the target positions are not derived from the training data but rather they are defined *a priori* before the transformation is computed. The classification was repeated except that vowel centroids, computed from the training speakers, were used as reference points for each vowel rather than the specified target positions. The percentage of vowels

[2] The CC's were numbered beginning with 1 for the coefficient of the constant basis vector. Therefore, the lowest ordered CC used, number 2, the coefficient of a half cycle of a cosine over frequency, is a measure of spectral tilt.

[3] The "center frame" is one frame located at the midpoint of the manually labeled steady-state portion of each vowel.

correctly identified for the test speakers, PCC, is given for each figure. Third, the ratio of between-group to within-group variance, RV, was also computed for each example. In general, as clustering about category centroids improves, both PCC and RV increase. As the match of the transformed data to the target position improves, PCT increases.

The two measurement sets were also compared with a Bayes maximum likelihood classifier [3] in terms of their ability to distinguish the 11 vowel categories. The percentage of test data (from the 15 test speakers) correctly identified was 75.2% for the formants $+F0$ and 75.0% for the cepstral coefficients $+F0$. Thus the two parameters sets are nearly identical insofar as automatic classification of the vowel data. Since these classification results were obtained with the full-dimensionality spaces (4 and 9), using a sophisticated classifier, both PCT and PCC computed as described above for two-dimensional spaces would be expected to be less than these "upper bound" values.

Example 1: Comparison of MSECT and LDA.

In the first example, both MSECT and linear discriminant analysis (LDA) were used to transform vowel data (formants $+F0$, $P = 4$; and CC's 2 through 9 + $F0$, $P = 9$) to two-dimensional spaces. The LDA transformations were first computed for 15 training speakers (5 adult males, 5 adult females, and 5 children) for each of the feature sets. The centroids for each of the 11 vowel categories were computed in the space of the discriminant scores. These centroids, drawn as the large black dots in Figs. 2 and 3, were used as the target positions for vowels for use with MSECT. The two transformations, LDA and MSECT, were then each used to transform the vowel data from the other 15 speakers in the data base with results shown in Figs. 2 and 3. Thus all transformations were computed with training speakers whereas ellipses were drawn with results from test speakers. Because of variability between the two speaker groups, even for Figs. 2(a) and 3(a), the ellipses are not exactly centered at the target positions. The figures, as well as the computed PCT, PCC, and RV numbers given in the various panels of the figures, show that the MSECT method gives similar results to those obtained with LDA. However, this is to be expected since the targets for MSECT were obtained as a by-product of the LDA computations. This example does indicate that MSECT results in nearly the same transformation as LDA, albeit with a completely different computational procedure, with the appropriate choice of targets. MSECT is not intended to replace or compete with LDA. For an application such as automatic pattern recognition in a low-dimensionality feature space, MSECT could be used but suffers from the obvious disadvantage of requiring the user to determine "good" target positions. LDA does not suffer from this problem. The next two examples illustrate applications where the benefits of MSECT are more apparent.

Example 2: Transformations from measurement spaces to perceptual spaces for vowels.

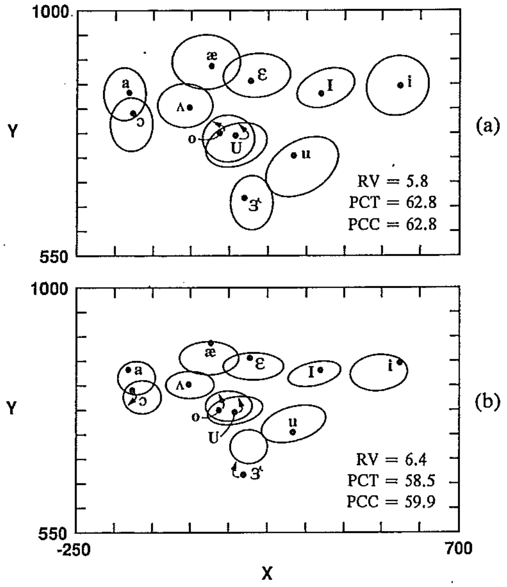


Fig. 2. Cluster plots for eleven vowels resulting from (a) LDA or (b) MSECT, using target positions obtained from LDA. The original parameters were 3 formants + $F0$.

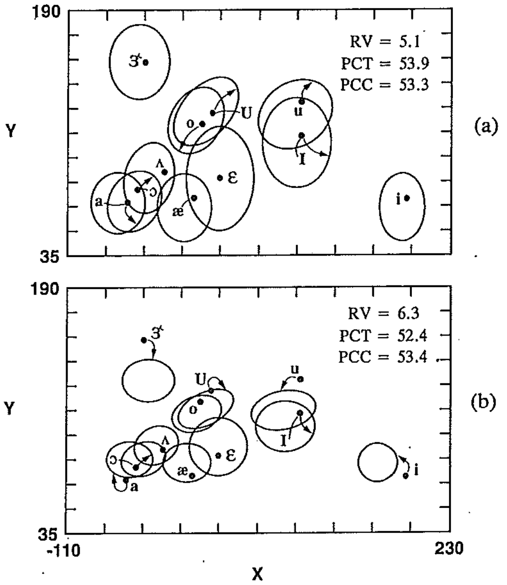


Fig. 3. Cluster plots for eleven vowels resulting from (a) LDA or (b) MSECT, using target positions obtained from LDA. The original parameters were 8 CC's + $F0$.

In this example we illustrate the use of MSECT to evaluate the two different spectral measurement sets—formants + $F0$ versus cepstral coefficients + $F0$— in terms of their ability to predict the perceptual locations of the 11 vowels. For this example spectral measurements of the vowel data, as described above, and the perceptual configuration of the vowels are required. The perceptual data were obtained using multidimensional scaling as applied to a confusion matrix obtained from a listening experiment. To obtain these data, five subjects listened to the steady-state vowel portions (as defined by the manual labeling mentioned previously) of the vowel stimuli from nine of the talkers[4] (three men, three women, and three children) in the data base. Each talker repeated each of the 11 vowels listed previously approximately 9 times, in different consonantal contexts, for an overall total of 890 stimuli. The subjects listened to the randomized stimuli in a computer-controlled experiment as many times as desired but were required to identify each token as one of the 11 vowels in a forced-choice format. The multiple-request listening format insured that performance was not degraded due to demands for an immediate response from the listeners.

An 11-by-11 confusion matrix, Table I, was obtained from the pooled responses of the five listeners with columns representing identified vowel and the rows the vowel intended by the speaker. Averaging over all vowels and all listeners, 85% of the vowel segments were identified as the intended vowel.[5] The confusion matrix was converted to symmetric form by averaging the matrix and its transpose. This pooled symmetric 11-by-11 matrix was then used as the input to the SAS ALSCAL multidimensional scaling routine [26] to compute a two-dimensional perceptual configuration for the vowels. ALSCAL was used with a Euclidean model, interval measurements, and a fourth-degree polynomial fit to the data.

TABLE I
CONFUSION MATRIX OBTAINED FROM LISTENING TO STEADY-STATE VOWEL SEGMENTS. THE ROWS ARE THE VOWELS INTENDED BY THE SPEAKERS. THE COLUMNS CORRESPOND TO LISTENER RESPONSES

	/a/	/i/	/u/	/ae/	/ʒ/	/I/	/ɛ/	/ɔ/	/ʌ/	/U/	/o/
/a/	73.6			1.6				17.7	7.0		
/i/		98.8	0.5			0.2	0.5				
/u/			96.4						0.7	2.4	0.5
/ae/	8.3			81.8			9.5	0.5			
/ʒ/					98.6		0.3		1.2	0.3	
/I/		0.3	1.1	0.6		91.7	5.1		0.6	0.6	
/ɛ/	0.3	0.5		7.6	0.3	5.6	83.8		2.0		
/ɔ/	24.2							73.6	2.2		
/ʌ/	2.0				0.3	0.3	0.8	0.3	89.1	4.6	2.8
/U/			4.8				0.4	0.4	24.4	67.8	2.2
/o/			4.3		0.9	0.2	0.7		3.8	6.1	84.0

The confusion matrix depicted in Table I shows that the main confusions are between the vowel pairs /a/ and /ɔ/, /ae/ and /ɛ/, /U/ and /ʌ/, and /o/ and /U/. Of these confusions the largest asymmetry is for the /U/ – /ʌ/ pair; /U/ is recognized as /ʌ/ in 24.4% of cases whereas /ʌ/ is recognized as /U/ in only 4.6% of cases. A more sophisticated procedure for converting the matrix to symmetric form would have changed only the details of the MDS results, since there is little consistent variation from symmetry in the original matrix. We also note that the stress value obtained for the two-dimensional MDS solution was 0.38 versus 0.24 for 3 dimensions, thus indicating that the two-dimensional solution is not a particularly good fit to the data. It must be emphasized that MDS transformations involve several subjective choices, and that real data seldom match the model assumptions precisely. Nevertheless, MDS can be used to approximate a perceptual space for stimuli such as speech sounds. A two-dimensional model was used in results for this paper for ease of comparison with the other examples.

Fig. 4 depicts the results of transforming the two sets of vowel measurements, using target positions as obtained from the MDS results, using the MSECT procedure. The MDS targets are drawn as the large black dots in Fig. 4 (and also in Figs. 5 and 6). Again the transformation coefficients were computed from 15 training speakers whereas the ellipses plotted in Fig. 4 represent the 15 test speakers. As can be seen, for Fig. 4(a) obtained with formants + $F0$, the vowel clusters overlap by approximately the same degree as for Fig. 2(b). Although RV = 3.6, versus 6.4 for Fig. 2(b), the PCC value is somewhat higher for Fig. 4(a) versus Fig. 2(b). (62.8% versus 59.9%). However, the low value for PCT (29%) indicates that the vowel data is generally not closest to the specified target position for each vowel. Nevertheless, visual inspection of Fig. 4(a) indicates that the vowel ellipses are in the vicinity of the specified target position.

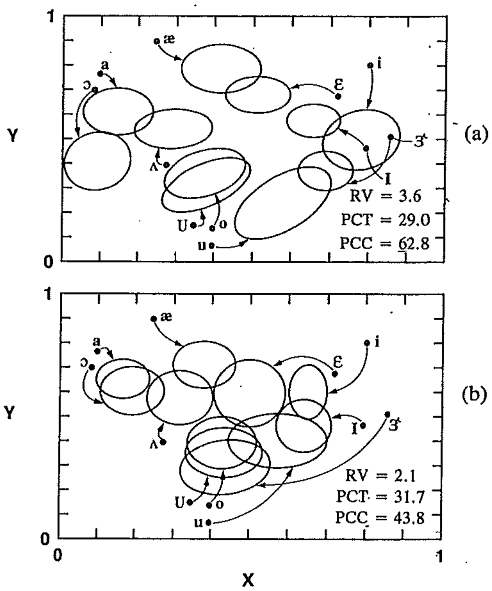


Fig. 4. Projection of vowel data to MDS target positions in a two-dimensional space using MSECT for (a) 3 formants + $F0$, and (b) 8 CC's + $F0$.

The data depicted in Fig. 4(b), derived from the cepstral coefficients + $F0$, are somewhat different from that depicted in Fig. 4(a). By inspection of the two figures, the clusters overlap much more in Fig. 4(b) than in Fig. 4(a). The RV and PCC values are also much lower for

[4] These nine talkers were selected on the basis of an automatic recognition experiment for vowels. For each of the three speaker types, a talker was selected with a high recognition rate, a low recognition rate, and an average recognition rate. Thus the talkers represented "good," "bad," and "average" speakers from the viewpoint of automatic vowel recognition.

[5] In contrast to the segments used for automatic classification, the vowel segments presented to the listeners did contain some dynamic and coarticulation information. The length of the steady-state vowels identified by the manual labeling ranged from 46 to 542 ms with an average of 160 ms and standard deviation of 74 ms.

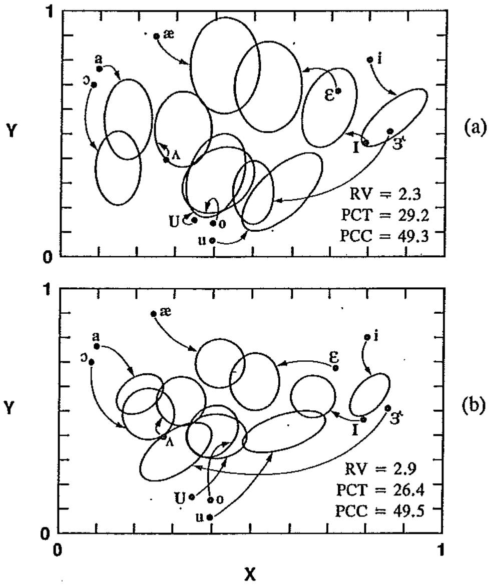


Fig. 5. Projection of vowel data to MDS target positions in a two-dimensional space using generalized Procrustes analysis for (a) 3 formants + $F0$, and (b) 8 CC's + $F0$.

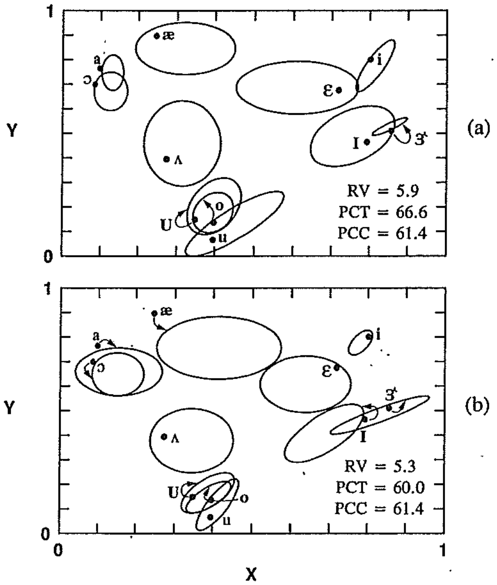


Fig. 6. Projection of vowel data to MDS target positions in a two-dimensional space using a neural network + MSECT for (a) 3 formants + $F0$, and (b) 8 CC's + $F0$.

Fig. 4(b) than Fig. 4(a). However, the PCT value of 31.7% for Fig. 4(b) versus 29.0% for Fig. 4(a) indicates that the vowel data map closer, on the average, to the MDS target positions, for the cepstral data than the formant data. Although the cepstral coefficients have a slight advantage in terms of PCT, the MSECT conversion of vowel cepstral coefficients to a two-dimensional perceptual space results in much less distinct vowel clusters than for the transformation of the formant data.

As a control experiment, the vowel data were also transformed to two-dimensional target positions using Procrustes analysis. The results for the formant and cepstral data are shown in Fig. 5. The transformations used for the data in this figure were computed using parameter centroids for the training speakers. The centroids were shifted, uniformly scaled, and orthogonally rotated to best match the MDS target positions [7]. The transformations computed in this manner were then applied to all the data of the test speakers to generate the cluster plots and to compute the various figures of merit. For the case of formants, the clustering of the vowel data is quite degraded for the Procrustes transformation relative to the MSECT transform. The RV and PCC numbers are considerably smaller for Fig. 5(a) than Fig. 4(a). However, the data in Fig. 5(b) and the figures of merit, are similar to the data in Fig. 4(b) suggesting that a similar transformation is obtained with either the Procrustes method or the MSECT procedure, for the CC data. Thus, for some cases, MSECT, which minimizes mean-square error over the entire training data set, gives superior results to a transformation which minimizes mean-square error between the centroid and target positions. For other cases, the two methods yield similar results. Without a proof, it is impossible to conclude that the Procrustes method would not give superior results to MSECT for some data. However, we have not found such a case in any of the examples investigated in the preparation of this paper.

A fundamental advantage of MSECT is that performance is not affected by linear scalings of the input variables (property 1) whereas the results obtained with the Procrustes method are highly dependent on the scaling of each input variable. To exemplify this point experimentally, the formant data were linearly scaled (arbitrarily) such that $F0$ was multiplied by 100, $F1$ by 10, $F2$ by 1, and $F3$ by 0.01. The MSECT and Procrustes transformations, with MDS targets, were again computed for these two cases. As expected, the MSECT results were identical to those obtained with the original formant values (Fig. 4(a)). However, the Procrustes results were greatly degraded—the RV, PCC, and PCT values, respectively, were 0.3, 19.0, and 21.0, as opposed to the corresponding values of 2.3, 29.2, and 49.3 (Fig. 5(a)) for the original formants. Thus the MSECT approach, as opposed to the Procrustes method, does not rely on the experimenter to make a "good" choice of the relative scaling of the input variables.[6]

As yet another approach for transforming vowel data to the MDS target positions, the combined nonlinear/linear transform, as presented in Section III, was implemented. To illustrate this method, the data were preprocessed with a two-layer feedforward perceptron network, with one input node for each speech parameter (i.e., 4 input nodes for the case of formants + $F0$ and 9 input nodes for the case of 8 cepstral coefficients + $F0$), 20 hidden nodes,

[6] Some formulations of Procrustes transformations first normalize the input variables. Linear scalings of variables would then have no effect on the results. However, normalization itself is arbitrary and would in general change the results relative to those obtained with nonnormalized variables.

and 11 output nodes.[7] The 15 training speakers were used to train the network as a classifier for the 11 vowels (thus 11 output nodes) using the backpropagation training method. Also recall that the neural network can be viewed as a nonlinear transformation of the original speech features to an 11-dimensional space. MSECT was then used to transform the 11-dimensional neural network outputs to a 2-dimensional space with the MDS target positions. Fig. 6 depicts the results as applied to the 15 test speakers. As can be seen by comparing the (a) and (b) in Fig. 6, with (a) and (b), respectively, in Figs. 4 and 5, much better data clustering and matches to specified target positions are achieved for the combination nonlinear/linear transformation than for either of the linear transformations.

The results for the nonlinear/linear transform are easily explained if the neural network is considered as a classifier that transforms all data in each category to a distinct binary code which can then be transformed to arbitrary positions in a multidimensional space. Naturally, no claim can be made that the combined nonlinear/linear transformation is optimum, since there are many, many variables which were not investigated. There are also many other forms of nonlinear transformations, besides one based on a neural network, which might be used.

The logical general conclusion from this example is that the formants are better perceptual indicators of vowels in a two-dimensional space than are the cepstral coefficients. However, such an inference must be made with great caution. In particular, the MDS procedure only gives an estimate of the perceptual configuration. This estimated configuration depends very much on the details of the original experiment, the options used in MDS, and the dimensionality chosen for the perceptual space. For example, the two-dimensional vowel perceptual configuration depicted as the targets in Figs. 4–6 is globally similar to perceptual configurations for vowels obtained in other studies [23], [24], but the details are quite different.[8] There is no real substitute for a perceptual experiment to verify the relative significance of acoustic features in speech perception. However, such experiments are very time consuming to conduct and often result in ambiguous conclusions. Thus the MSECT procedure, with or without the nonlinear preprocessing, can be used as a tool, along with other established methods, to give insight into the perceptual significance of a particular feature set.

Example 3: Speaker normalization of vowel data.

Another potential application of linear transformation procedures is for speaker normalization of vowel data [3]. Target positions for each vowel (the columns of matrix C) can easily (and seemingly logically) be obtained as the centroid of all the measurement data for that vowel from all speakers. The vowel data of each speaker can be transformed to best match the speaker-average positions. A separate transformation must be computed for each speaker, based on training data for that speaker. The transformed data for each speaker matches the speaker-average vowel positions as closely as possible, and thus can be considered as speaker-normalized data. In contrast to the previous examples, the dimensionality of the original and transformed spaces are the same and also the choice of the target positions in the transformed space is much less arbitrary.

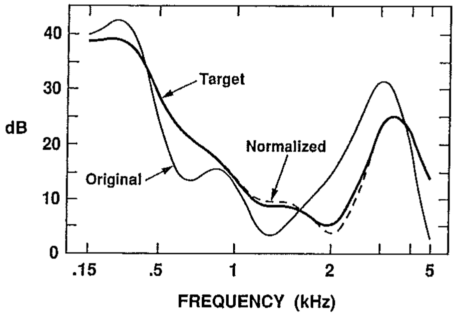


Fig. 7. Spectral plots of target talker /i/, nonnormalized /i/, and speaker-normalized /i/.

As a test of MSECT for speaker normalization using the general method outlined above, MSECT was used to normalize data from 30 speakers (10 adult males, 10 adult females, and 10 children) for 11 vowel categories with cepstral coefficients 2–9 as parameters. As a control, the data from each speaker was also normalized using a Procrustes transformation to optimally align the average vowel configuration for each speaker (averaged over multiple repetitions of each vowel) to the overall speaker average configuration. The percentages of test vowels correctly classified with the previously mentioned Bayes maximum likelihood classifier are 74.2%, 83.6%, and 80.0% for the nonnormalized data, the MSECT normalization, and the Procrustes normalization.[9] The error rates for automatic vowel classification of test data (100% − 74.2% = 25.8%) were reduced by 36% for the MSECT method versus 22% for the Procrustes method. Thus, for this example, vowel normalization computed from all the training data of each speaker outperforms normalization based on speaker centroids.

The spectral plots depicted in Fig. 7 also show an example of the effect of the speaker normalization. For this figure MSECT was used to compute a normalization transformation for cepstral coefficients 2 to 9 of one speaker's training data. The results of applying this transformation to a stimulus outside the training set (/i/) are shown in the figure in terms of the target spectrum for

[7] The number of hidden nodes was determined by a pilot experiment using a neural network classifier for the vowel data. For both the formant and cepstral parameters, approximately 1% higher automatic vowel classification results (test speakers) were obtained with 20 hidden nodes compared to the results for 10 or 30 hidden nodes.

[8] For example, in all cases /a/, /i/, and /u/ are well separated, /i/ is close to /I/, /a/ is close to /Λ/, etc.

[9] Unlike the other examples, for the speaker normalization experiment, half of the tokens for each vowel for each speaker were used to train the classifier and the speaker normalization transformations, and the other half of the tokens from each of the 30 speakers were used for testing. Thus the test classification rate of 74.2% without speaker normalization is slightly different than the 75% test rate obtained if half the speakers were used for training and half for testing.

/i/, the original spectrum of the /i/ for this speaker, and the spectrum of this speaker's /i/ after normalization. Note that although the transformation was based on the CC's, the CC's were converted to a spectral plot for ease of interpretation. Clearly, for this example, the normalized spectrum matches the target spectrum much more closely than does the original spectrum. The reduction in error rates for automatic classification due to the speaker normalization also indicates the viability of the approach.

VI. Summary

In conclusion, we have described a multidimensional transformation technique, which we call minimum mean-square error coordinate transformation (MSECT), for projecting data from one multidimensional space to another multidimensional space. The mathematical deviation of MSECT is based on minimum mean-square error estimation theory. Although statistical in nature, the derivation does not depend on any particular form of multivariate probability density functions in the original space. The transformation is very flexible in that there are no restrictions on the dimensionality of the transformed space, no requirements that the transformation be orthogonal; arbitrary scalings and shifts in the origin are also allowed. For some uses, MSECT results in a transformation very similar to that obtained with existing transformation procedures such as discriminant analysis. For applications where there is no desire nor natural method for choosing target positions, the requirement that target positions must be specified in the transformed space is a fundamental disadvantage of MSECT. However, for other applications, the ability to freely specify target positions for each category in the target space can be used to advantage. These applications include checking the perceptual relevance of various sets of measurable features, signal processing for speech articulation training aids, and speaker normalization of vowel data.

References

[1] I. Borg and J. Lingoes, *Multidimensional Similarity Structure Analysis.* New York: Springer, 1987.

[2] W. D. Cooley and P. R. Lohnes, *Multivariate Data Analysis.* New York: Wiley, 1971.

[3] S. F. Disner, "Evaluation of vowel normalization procedures," *J. Acoust. Soc. Amer.*, vol. 67, pp. 253–261, 1980.

[4] R. O. Duda and P. E. Hart, *Pattern Analysis and Scene Classification.* New York: Wiley, 1973.

[5] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen.*, vol. 7, pp. 179–188, 1936.

[6] R. Gnanadesikan, *Methods for Statistical Data Analysis of Multivariate Observations.* New York: Wiley, 1977.

[7] J. C. Gower, "Generalized Procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[8] H. H. Harman, *Modern Factor Analysis*, second ed., revised. Chicago, IL: University of Chicago Press, 1976.

[9] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, pp. 417–441, 1933.

[10] W. Klein, R. Plomp, and L. Pols, "Vowel spectra, vowel spaces, and vowel identification," *J. Acoust. Soc. Amer.*, vol. 48, pp. 999–1009, 1970.

[11] J. B. Kruskal and M. Wish, *Multidimensional Scaling* (Quantitative Applications in the Social Sciences Series). Beverly Hills, CA: Sage Publications, 1978.

[12] P. A. Lachenbruch, *Discriminant Analysis.* New York: Hafner, 1975.

[13] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Mag.*, pp. 4–22, Apr. 1987.

[14] J. Makhoul and L. Cosell, "LPCW: An LPC vocoder with linear predictive spectral warping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1976, pp. 466–469.

[15] J. D. Markel, "The Sift algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. 20, pp. 367–377, 1972.

[16] S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 135–141, 1974.

[17] D. D. McCracken, *Fortran with Engineering Applications.* New York: Wiley, 1967.

[18] A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," *Proc. IEEE*, vol. 60, no. 6, pp. 681-691, 1972.

[19] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, pp. 175–184, 1952.

[20] L. C. W. Pols, L. J. T. van der Kamp, and R. Plomp, "Perceptual and physical space of vowel sounds," *J. Acoust. Soc. Amer.*, vol. 46, pp. 458–467, 1969.

[21] D. J. Povel and M. Wansink, "A computer-controlled vowel corrector for the hearing impaired," *J. Speech Hearing Res.*, vol. 29, pp. 99–105, 1986.

[22] P. H. Shonemann and R. M. Carroll, "Fitting one matrix to another under choice of central dilation and rigid motion," *Psychometrika*, vol. 35, pp. 245–255, 1970.

[23] R. N. Shepard, "Psychological representation of speech sounds," in *Human Communication: A Unified View*, E. E. David and P. B. Denes, Eds. New York: McGraw-Hill, 1972.

[24] S. Singh and D. R. Woods, "Perceptual structure of 12 American English vowels," *J. Acoust. Soc. Amer.*, vol. 49, pp. 1861–1866, 1970.

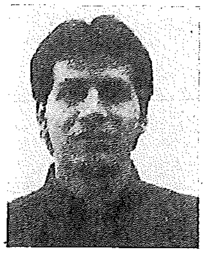
[25] A. K. Syrdal and H. S. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *J. Acoust. Soc. Amer.*, vol. 79, pp. 1086–1100, 1986.

[26] F. W. Young and R. Lewyckyj, The Alscal procedure," in *SUGI Supplemental Library User's Guide*, version 5 edition, R. P. Hastings, Ed. Cary, NC: SAS Institute, 1986, ch. 1, pp. 1–16.

[27] S. A. Zahorian and M. Rothenberg, "Principal-components analysis for low-redundancy encoding of speech spectra," *J. Acoust. Soc. Amer.*, vol. 69, pp. 832–845, 1981.

Stephen A. Zahorian (S'76–M'78) was born on November 20, 1947. He received the B.S. degree in electrical engineering from the University of Rochester in 1969, and the M.S. and Ph.D. degrees from Syracuse University in 1973 and 1978, respectively, both in electrical and computer engineering.

He joined the faculty of the Department of Electrical and Computer Engineering of Old Dominion University, Norfolk, VA, in 1979. He is currently an Associate Professor in that department. His research interests are in acoustic/phonetic automatic speech recognition, the development of speech training aids for the deaf, and digital signal processing.



Amir Jalali Jagharghi (S'90–M'90) was born in Tehran, Iran, on November 10, 1960. He received the B.S.E.E. degree from the West Virginia Institute of Technology, in 1982, and the M.S. and Ph.D. degrees in electrical engineering from Old Dominion University in 1985 and 1990.

He is currently a Research Engineer at Vegyan Inc. performing digital image and signal processing at NASA Langley Research Center in Hampton, VA. His research interests are acoustic/phonetic automatic speech recognition, digital signal and image processing, and the development of speech training aids for the hearing impaired.