
Self-Organizing Map for the Classification of Normal and Disordered Female Voices

Daniel E. Callan*

Ray D. Kent

Department of
Communicative Disorders
University of Wisconsin-
Madison

Nelson Roy

Department of
Communication Disorders
Minot State University
Minot, ND

Stephen M. Tasko

Department of
Communicative Disorders
University of Wisconsin-
Madison

The goal of this research was to train a self-organizing map (SOM) on various acoustic measures (amplitude perturbation quotient, degree of voice breaks, rahmonic amplitude, soft phonation index, standard deviation of the fundamental frequency, and peak amplitude variation) of the sustained vowel /a/ to enhance visualization of the multidimensional nonlinear regularities inherent in the input data space. The SOM was trained using 30 spasmodic dysphonia exemplars, 30 pretreatment functional dysphonia exemplars, 30 post-treatment functional dysphonia exemplars, and 30 normal voice exemplars. After training, the classification performance of the SOM was evaluated. The results indicated that the SOM had better classification performance than that of a stepwise discriminant analysis over the original data. Analysis of the weight values across the SOM, by means of stepwise discriminant analysis, revealed the relative importance of the acoustic measures in classification of the various groups. The SOM provided both an easy way to visualize multidimensional data, and enhanced statistical predictability at distinguishing between the various groups (over that conducted on the original data set). We regard the results of this study as a promising initial step into the use of SOMs with multiple acoustic measures to assess phonatory function.

KEY WORDS: voice, self-organizing map, classification, acoustic, dysphonia

One of the persistent problems in the study of disordered voice quality is the efficient acoustic description of clinically relevant differences. Of the various studies that have attempted to relate perceptual ratings of disturbed voice quality to acoustic variables, weak to moderately strong associations have been reported for different types of voice quality disorder (Arends, Povel, Van Os, & Speth, 1990; Dejonckere & Wieneke, 1992, 1993; Eskinazi, Childers, & Hicks, 1990; Kent et al., 1994; Murry, Singh, & Sargent, 1977; Wolfe, Fitch, & Cornell, 1995; Zwirner, Murry, & Woodson, 1991). Development of an acoustic analysis procedure that is both efficient and interpretable with respect to perceptual judgments of voice quality has been impeded by several conditions. First, acoustic variables that are the best predictors for a certain voice type may not be the best predictor for voices of a different type (Wolfe & Steinfatt, 1987). Second, expert judges may differ considerably from one another in the acoustic factors that are correlated with their perceptual ratings of pathological voices (Kreiman & Gerratt, 1996; Kreiman, Gerratt, & Precoda, 1990; Kreiman, Gerratt, Precoda, & Berke,

*Currently affiliated with ATR Human Information Processing Research Laboratories, Kyoto, Japan.

1992). That is, it appears that different judges weight differently the various acoustic properties of a given voice or voice type. Third, the dimensions used for perceptual ratings of voice quality can be highly interdependent. For example, Kreiman, Gerratt, and Burke (1994) concluded that the “breathy” and “rough” voice qualities are related multidimensional constructs and that judgments of roughness depend strongly on concomitant breathiness, but not vice versa. A similar interaction has been described for judgments of vocal pitch and roughness (Wolfe & Ratusnik, 1988). Moderately to severely dysphonic vowels were matched with a significantly lower pitch than were mildly dysphonic vowels.

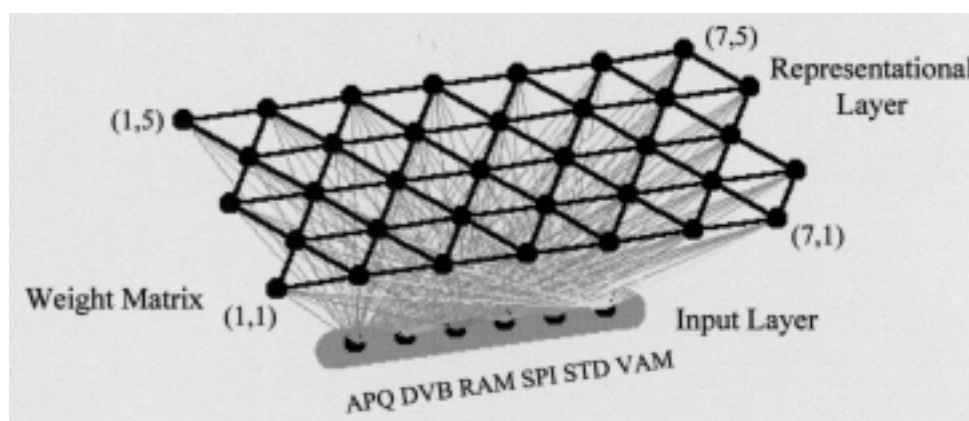
Solving the problem of acoustically characterizing disordered voice quality will likely require multiple acoustic measures that are differentially weighted. However, there are two main concerns when analyzing data with multiple dimensions: (1) it is often difficult to visualize data characterized by more than two or three dimensions and (2) it is often difficult to characterize the weightings among the multiple dimensions of the data. The latter problem is especially challenging when no gold standard exists. One method to determine the weighting of the acoustic measures for the purpose of voice classification is based on the inherent nonlinear regularities that exist among the acoustic measures. In this regard, SOMs may be used to set the weights of various acoustic measures in a manner that depends on such inherent nonlinear regularities.

There are three main characteristics of SOMs that make them desirable as a tool to classify multidimensional data: (1) SOMs can represent multidimensional data along a two dimensional surface of nodes. This provides an easy way to visualize the relative distribution of the exemplars of the various groups across the surface

of the SOM. (2) In some cases, the nonlinear nature of the SOM permits better classification performance than traditional multivariate statistical techniques. (3) The relative weighting of the input dimensions responsible for defining the distribution of exemplars into groups across the SOM can be discerned easily.

SOMs allow two-dimensional representational mapping, over an array of processing units (called nodes), of the multidimensional nonlinear regularities inherent in the input data space (Kohonen, 1995). Typically, SOMs are characterized by an input layer and a representational layer as shown in Figure 1. The input layer can be thought of as containing a separate processing node for each of the various data parameters. The representational layer is composed of a two dimensional array of processing nodes, each of which receives a connection from all of the nodes in the input layer. Associated with each of these connections is a particular weight value. It is the relative value of the weights across the two-dimensional node matrix that allows the representation of features inherent in the multidimensional input data (Kohonen, 1995). This representational mapping is formed in a self-organizing unsupervised manner; that is, the values of the weights depend on characteristics of the input data, and there is no teacher signal that modifies these weights on the basis of some optimal target. Training of the SOM occurs by initially randomizing the values of the weights of the connections from the input layer to the representational layer. Data samples are iteratively presented, and the node in the representational layer is selected that has weight values closest in Euclidean distance to the input data. The weights that project to this node as well as its neighbors are adapted toward the value of the input data. The magnitude of the change is a function

Figure 1. The architecture of the SOM is shown, with the input layer composed of nodes (processing units) that encode the values of each of the input data parameters. In this case the six data parameters are amplitude perturbation quotient (APQ), degree of voice breaks (DVB), rahmonic amplitude (RAM), soft phonation index (SPI), standard deviation of the fundamental frequency (STD), and variation of peak-to-peak amplitude (VAM).



of a parameter called the learning rate (Kohonen, 1995). Each of the data samples is iteratively presented to the network several times until a stable representational mapping forms across the node array. One can easily visualize the distribution of exemplars, both those on which the SOM was trained and novel ones, across the two-dimensional node array. In addition, the weightings of the multiple input dimensions can be evaluated by observing the corresponding weight matrix of each of the input dimensions.

SOMs have been used to characterize several aspects of normal and disordered voice. In a series of studies carried out by Leinonen and colleagues (Leinonen, Hiltunen, Kangas, Juvas, & Rihkanen, 1993; Leinonen, Kangas, Torkkola, & Juvas, 1992; Rihkanen, Leinonen, Hiltunen, & Kangas, 1994), a SOM was trained to detect characteristics of dysphonic voice by recognition of spectral composition. To determine how this map would characterize disordered voice, samples were taken from individuals classified with varying degrees of dysphonia using a method derived from the GRBAS perceptual rating scale (Hirano, 1981). The voices were classified on the basis of the degree of dysphonia, roughness, breathiness, and strain. When samples of the vowel /a/ from both normal and disordered voices were presented to the SOM, the trajectory pattern across the map differed for the two types of voices. It was also found that rough and breathy voices had different trajectory patterns. However, the SOM was not able to distinguish the degree of dysphonia. In a later study conducted by Rihkanen et al. (1994), the SOM was trained using spectral vectors from both normal and pathologic voices. This method of training allowed the relative degree of dysphonia to be encoded across the SOM.

Overall, the SOMs used by Leinonen and colleagues are very successful in providing a two-dimensional visual representation of multidimensional properties of voice and speech that is easy to understand. However, one of the disadvantages of using multiple spectral dimensions as input is that it is difficult to evaluate which acoustic properties are important in the map organization that makes classification of normal and disordered voices possible.

The SOM has also been used to categorize voice disorders by using the dimensions of a perceptual rating scale as input. In the study carried out by Leinonen, Hiltunen, Laakso, Rihkanen, and Poppius (1997), the categorization of various forms and degrees of dysphonia was accomplished by using perceptual ratings of pathology, roughness, breathiness, strain, and asthenia as input dimensions to train the SOM. Five different categories of voice quality were formed across the SOM. The categories were organized from right to left across the SOM depending on the relative degree of severity.

The SOM trained on various dimensions of a perceptual rating scale was successful in categorizing both the degree of pathology and the ratio of breathiness and roughness across the map. However, there are several disadvantages to the use of perceptual rating scales, mainly due to their subjective nature. As was pointed out by Leinonen et al. (1997), it would be interesting to determine if the five categories of pathology formed in the multidimensional perceptual space had corresponding acoustic patterns.

The goal of the present research is to develop an objective means of assessing voice quality, on the basis of multiple acoustic measures, by utilizing a SOM to classify samples of normal and disordered voice in an unsupervised manner. Unsupervised self-organization of the map allows for the weightings of the input acoustic measures to be determined on the basis of the inherent nonlinear regularities of the input data space. Unsupervised learning is particularly advantageous when no gold standard exists for classification. It is hoped that by using a SOM, which is trained using various objective acoustic measures, the subjective nature inherent to perceptual rating scales will be avoided. In this study, voice samples from sustained vowel /a/ productions for four groups of subjects were used to train the SOM. These groups consist of individuals with normal voice, individuals with verified adductor spasmodic dysphonia, individuals with functional dysphonia before behavioral management, and the same individuals with functional dysphonia after behavioral management. Analysis of the two-dimensional SOM provides an easy way to visualize and evaluate the classification of the various groups. It is hoped that the SOM will provide better predictability for classification than traditional statistical analyses. Analysis of the SOM also provides an easy way to visualize and evaluate the weighting of the various acoustic measures underlying the representational map.

Method

Subjects

The voice samples of subjects included in the study consisted of individuals with verified adductor spasmodic dysphonia (SD; $n = 30$), individuals with functional dysphonia pretreatment (PR; $n = 30$), individuals with functional dysphonia post-treatment (PS; $n = 30$), and individuals categorized as having normal phonation (NR; $n = 30$). (The voice samples for the PR and PS groups were taken from the same individuals before and after treatment.) Individuals were given a diagnosis of functional dysphonia if they had no apparent structural or neurological pathology and had a positive sustained response to behavioral management. Individuals received the

diagnosis of spasmodic dysphonia following complete voice evaluation that failed to identify structural pathology sufficient to explain the nature and quality of voice symptoms. Voice symptoms were perceptually consistent with adductor spasmodic dysphonia (with and without tremor), and patients failed to respond to behavioral management. All participants were female nonsmokers who spoke English as a first language. Voice samples were acquired by asking the participants to sustain the vowel /a/, at a comfortable pitch and loudness, for at least 5 seconds. However, in some cases, individuals within the dysphonic groups (SD and PR) were unable to sustain phonation for a full 5 seconds. We chose to use a sustained phonation task because it is routinely collected as part of most clinical voice evaluations, and it can be easily submitted to acoustic analysis. If analysis procedures, such as the one used in this study, are to be clinically useful it is important that their implementation require minimal time and effort. The voice samples came from two different sources. Functional dysphonic and spasmodic dysphonic voice disordered samples were acquired at a hospital-based voice pathology clinic at the University of Wisconsin–Madison. The normal and five of the spasmodic dysphonic voice samples were taken from the voice disorders database compiled by the Voice and Speech Laboratory at the Massachusetts Eye and Ear Infirmary in Boston (Voice Disorders Database, 1994). Studio level recordings with a sampling rate of 25 kHz were used for both sources.

Spasmodic dysphonia is a complex phonatory disorder that may take different forms that have been recognized as distinct disorders (Aronson, 1973). Most research on spasmodic dysphonia distinguishes at least adductor and abductor types. Acoustic analyses of both adductor and abductor spasmodic dysphonia has shown several abnormalities, including increases in some parameters (standard deviation of f_0 , jitter and shimmer, percentage or degree of aperiodic phonation, and variability in the amplitude or fundamental frequency of sustained vowel phonation) and decreases in others (range of vocal intensity and harmonic to noise ratio) (Cannito, McSwain, & Dworkin, 1996; Hartman & Aronson, 1981; Ludlow & Conner, 1987; Ludlow, Naunton, & Bassich, 1984; Ludlow, Naunton, Terata, & Anderson, 1991; Rontal, Rontal, Rolnick, Merson, Silverman, & Truong, 1991; Zwirner, Murry, Swenson, & Woodson, 1992). It does not appear that mean f_0 is reliably different between normal voices and spasmodic dysphonia. Because spasmodic dysphonia appears to effect several different acoustic features of voice, it is an interesting disorder for multidimensional analysis.

Functional dysphonia is a condition of voice abnormality for which there is no apparent neurologic or structural pathology, or for which any obvious pathology does not seem to explain the dysphonia (Wilson, Deary, Scott,

& MacKenzie, 1995). The condition is more common in females and often follows symptoms of upper respiratory tract infection (Aronson, Peterson, & Litin, 1966; Kinzl, Biebl, & Rauchegger, 1988; Perrin, Berger-Vachon, Kauffman, & Collet, 1997; Wolfe et al., 1995). Few acoustic studies of functional dysphonia have been reported, and it is therefore uncertain how it should be described and differentiated from other dysphonias. However, it appears that useful measures include mean speaking f_0 , f_0 variation, jitter, and shimmer (Klingholz & Martin, 1985; Kotby, Titze, Saleh, & Berry, 1993; Roy & Tasko, 1993). A primary reason for the inclusion of patients with functional dysphonia in the present study was the opportunity to conduct a multidimensional analysis of voices in the same individuals before and after treatment using manual musculoskeletal tension reduction (Roy & Leeper, 1993). Therefore, it is possible to study dysphonia and normal, or nearly normal, voices in the same subjects).

Acoustic Measures

The central 1-second segment of each sustained /a/ production was extracted for acoustic analyses. The acoustic analyses included 22 measures of the Multi-Dimensional Voice Profile (MDVP; Kay Elemetrics), using the default measurement parameters. If automated voice classification procedures are to be of practical clinical utility, they require input variables that are easily and widely available. MDVP meets this requirement.

In addition to the MDVP analysis, a cepstral analysis using CSL (Model 4300B, Kay Elemetrics) to acquire the amplitude and quefreny of the dominant rahmonic was also conducted. The cepstrum is a log power spectrum of a log power spectrum. The cepstrum analysis shows the dominant energy corresponding to the harmonic peaks of the spectrum. This is called the dominant rahmonic. Many studies have indicated a high correlation between the dominant rahmonic and voice quality (Dejonckere & Wieneke, 1992, 1993, 1994; Koike, 1986). The time axis of the cepstrum is called quefreny. The quefreny of the dominant rahmonic corresponds to the fundamental period of the signal. One of the motivating factors for including the cepstrum-based measures is the finding made by Hillenbrand, Cleveland, and Erickson (1994) and Hillenbrand and Houde (1996), which suggests cepstrum-based measures are good predictors of breathiness in sustained vowels. The cepstrum measures were acquired over the central part of the down-sampled signals (12500 Hz) using a 1024-point FFT with pre-emphasis of .9 and a Hamming window weighting.

The acoustic parameters used to train the SOM were selected according to their ability to classify the four different groups as revealed by a stepwise discriminant

Table 1. Stepwise discriminant analysis of the original data for all four groups (SD, PR, PS, NR; $N = 120$). At each step, the variable that minimizes the overall Wilks's lambda is entered. Minimum significance of F to enter is .05. Maximum significance of F to remove is .1.

Step entered	Variable entered	Wilks's lambda	df	F	df1	df2	Sig.
1	RAM	0.358	116	69.4	3	116	0.001
2	VAM	0.273	116	35.0	6	231	0.001
3	STD	0.242	116	24.4	9	278	0.001
4	APQ	0.221	116	19.2	12	299	0.001
5	SPI	0.199	116	16.4	15	310	0.001
6	DVB	0.178	116	14.8	18	314	0.001

analysis. The results of the stepwise discriminant analysis are displayed in Table 1. The variables entered at each step were those that minimize the overall Wilks's lambda. When shimmer was taken out of the analysis, amplitude perturbation quotient was entered in the stepwise discriminant analysis. This resulted in better classification of the four groups (68.3% with amplitude perturbation quotient entered vs. 65.8% with shimmer entered). On the basis of apriori assumptions and the results of the stepwise discriminant analysis, the following acoustic variables were selected to train the SOM:

APQ—Amplitude perturbation quotient gives a measure in percent of the variability of the peak-to-peak amplitude within the analyzed voice sample at a smoothing factor of 11 periods, taken from the Voice Disorders Database (1994).

DVB—Degree of voice breaks (in percent) is the ratio of the total duration where the fundamental frequency cannot be tracked to the time of the complete voice sample, taken from the Voice Disorders Database (1994).

RAM—Rahmonic amplitude denotes the amplitude of the first dominant peak in the cepstral analysis, which corresponds to the harmonic peak of the spectrum of the signal, taken from CSL (1994).

SPI—Soft phonation index is an average ratio of the lower frequency to the higher frequency harmonic energy, taken from the Voice Disorders Database (1994).

STD—Standard deviation of the fundamental frequency (in hertz) of the vocalization consisted of the standard deviation of all extracted period-to-period fundamental frequency values within the 1-second sample, taken from the Voice Disorders Database (1994).

VAM—Coefficient of variation of the peak-to-peak amplitude (in percent) for the voice sample is taken from the Voice Disorders Database (1994).

It is worth noting that the acoustic measures the MDVP considers indicative of tremor (f_0 and amplitude

tremor amplitudes and frequencies) did not distinguish the SD group from the other three groups.

Network Architecture and Training Input to the SOM

The SOM was trained using 30 exemplars from each of the groups (SD, PR, PS, and NR). The six input dimensions used to train the SOM consisted of the acoustic measures of APQ, DVB, RAM, SPI, STD, and VAM. It is important to emphasize that these measures were selected over other possible measures provided by the MDVP, on the basis of how well they were able to classify the four groups according to a stepwise discriminant analysis. The entire input data space was normalized to have a mean of zero and a variance of one. Input normalization is thought to be important to ensure that the input dimensions are not disproportionately scaled. Input dimensions that vary greatly in scale may cause biases in the corresponding weight matrices of the nodes in the representational layer (Kohonen, 1995).

Architecture of the SOM

The SOM contained a six-node input layer and a seven-by-five node representational layer (see Figure 1). The dimensions of the representational layer were selected on the basis of the assumption that the 120 exemplars would be distributed across the map in regions defining the four groups. A much larger representational layer is needed if a greater number of exemplars is used for training. Each of the nodes in the input layer encoded one of the six input dimensions. The seven-by-five node representational layer was organized in a hexagonal neighborhood lattice as recommended by Kohonen (1995) for better visual inspection of the resultant map. Each node in the representational layer has associated with it a neighborhood radius that defines the number of rings of surrounding nodes that can be affected by activation of that particular node. The size of the neighborhood radius of each of the nodes in the representational layer was altered throughout

training. Each of the nodes in the representational layer had weighted connections to all six input nodes.

Training of the SOM

The SOM_PAK software package developed by Kohonen et al. (1996) was used to train the SOM. The SOM forms by means of projecting the probability distribution inherent in the multidimensional input space onto the two-dimensional grid of nodes composing the representational layer. Map formation is accomplished by iteratively presenting input and allowing for the weighted connections from the representational layer to be corrected toward the input value. For more details regarding the SOM training algorithm see Kohonen (1995) and Kohonen et al. (1996).

Ten maps were trained using the same procedure but different initial random values for the connection weights. The map with the lowest overall quantization error was selected for analysis. The maps were trained in two stages: The first stage was for 2000 steps through the input data and included an initial neighborhood radius decreased from seven to zero and a learning rate of .07; the second stage was for 12000 steps through the input data and included a neighborhood radius decreased from two to zero and a learning rate of .02. The first stage of training was carried out with a larger initial neighborhood radius and learning rate to allow proper ordering of weight vectors of the nodes in the representational layer (Kohonen et al., 1996). The second stage of training was carried out to fine tune the map (Kohonen et al., 1996). After training, a Sammon mapping (Sammon, 1969), across the six weight values associated with each of the nodes in the seven-by-five representational layer, was conducted in order to determine the approximate Euclidean separation of the nodes in two-dimensional space. The separation and position of the nodes in Figures 2 and 3 is determined by the Sammon mapping. All figures were constructed using Matlab (1994). All statistical analyses were conducted using SPSS (1996).

Results and Discussion

Figure 2 displays the number of exemplars from each of the groups that are coded by the nodes of the SOM. The node positions are determined by the x and y Euclidean coordinates defined by the Sammon mapping. The lines connecting the nodes in Figure 2 represent the underlying hexagonal neighborhood lattice of the SOM. Nodes that are closer together on the lattice have more similar weight values than nodes that are spread farther apart.

The map shown in Figure 2 indicates that there is a dense clustering of NR exemplars in the region of node

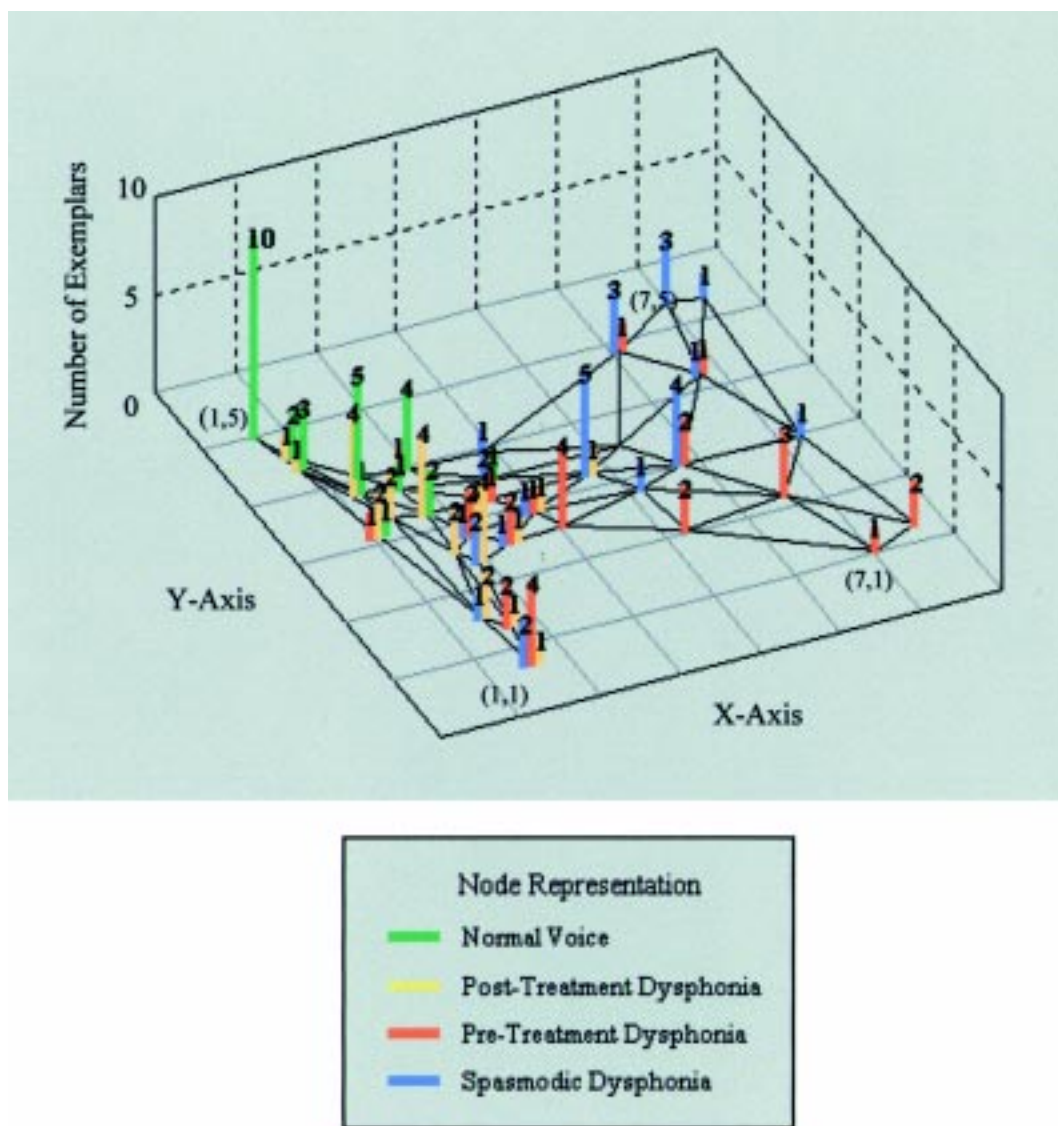
(1, 5). One third of the normal voice exemplars ($n = 10$) are represented by a single node, node (1, 5). The PS exemplars are clustered around nodes that code for NR exemplars. There is a fair degree of overlap between the nodes that code for post-treatment functional dysphonia PS exemplars and the nodes that code for NR and pretreatment functional dysphonia PR exemplars. The PR exemplars are sparsely distributed throughout a large region of the SOM, as are the SD exemplars. Nodes representing both the PR exemplars and the SD exemplars appear to have some degree of overlap with each other as well as with nodes that code for PS. One can also see in Figure 2 that the hexagonal neighborhood lattice is much less densely concentrated in the region of the map representing SD exemplars and PR exemplars than the region of the map representing NR exemplars and PS exemplars. The greater degree of spread (variability) and overlap for the SD and PR groups is consistent with perceptual studies of SD and non-organic dysphonias (Ludlow et al., 1991).

Classification Performance

In order to determine how well the SOM classified the various groups, each node was given a representation corresponding to one of the four groups. Node representation is determined by the maximum number of exemplars from a group that fall on a particular node. In the case in which the maximum number of exemplars that is coded by a particular node is equal between two or more groups, or the case in which a node does not code for any exemplars, node representation is determined by means of majority voting of the group membership of the five nearest exemplars (in Euclidean space) to the node in question. The resulting node representations are displayed as large circles over the node matrix in Figure 3 (SD—blue, PR—red, PS—yellow, NR—green). Comparison of the node representations given in Figure 3 with the number of exemplars from each of the groups that are coded by the nodes of the SOM given in Figure 2 provides an easy way to visualize and evaluate the classification of the exemplars from each of the four groups.

The classification performance of a stepwise discriminant analysis over the original data and the classification performance of the SOM using the same input are given in Tables 2 and 3. The results indicate that the classification performance for the SOM is higher than that for the stepwise discriminant analysis. The SOM classified 75.8% of the total exemplars ($N = 120$) correctly (see Table 3), whereas the stepwise discriminant analysis only classified 68.3% of the total exemplars correctly (see Table 2). The greatest difference in classification performance between the SOM and the stepwise discriminant analysis is for the PR exemplars. The SOM classified 73.3% of the 30 PR exemplars

Figure 2. The number of exemplars is shown for each group across the nodes of the SOM: blue = spasmodic dysphonia (SD), red = pretreatment functional dysphonia (PR), yellow = post-treatment functional dysphonia (PS), green = normal voice (NR). Numeric values are given at the top of each bar. The x and y axes represent the Euclidean coordinates for each of the nodes defined by the Sammon mapping (Sammon, 1969). The node identifiers for each of the corners of the SOM are given (1, 1), (1, 5), (7, 5), and (7, 1).



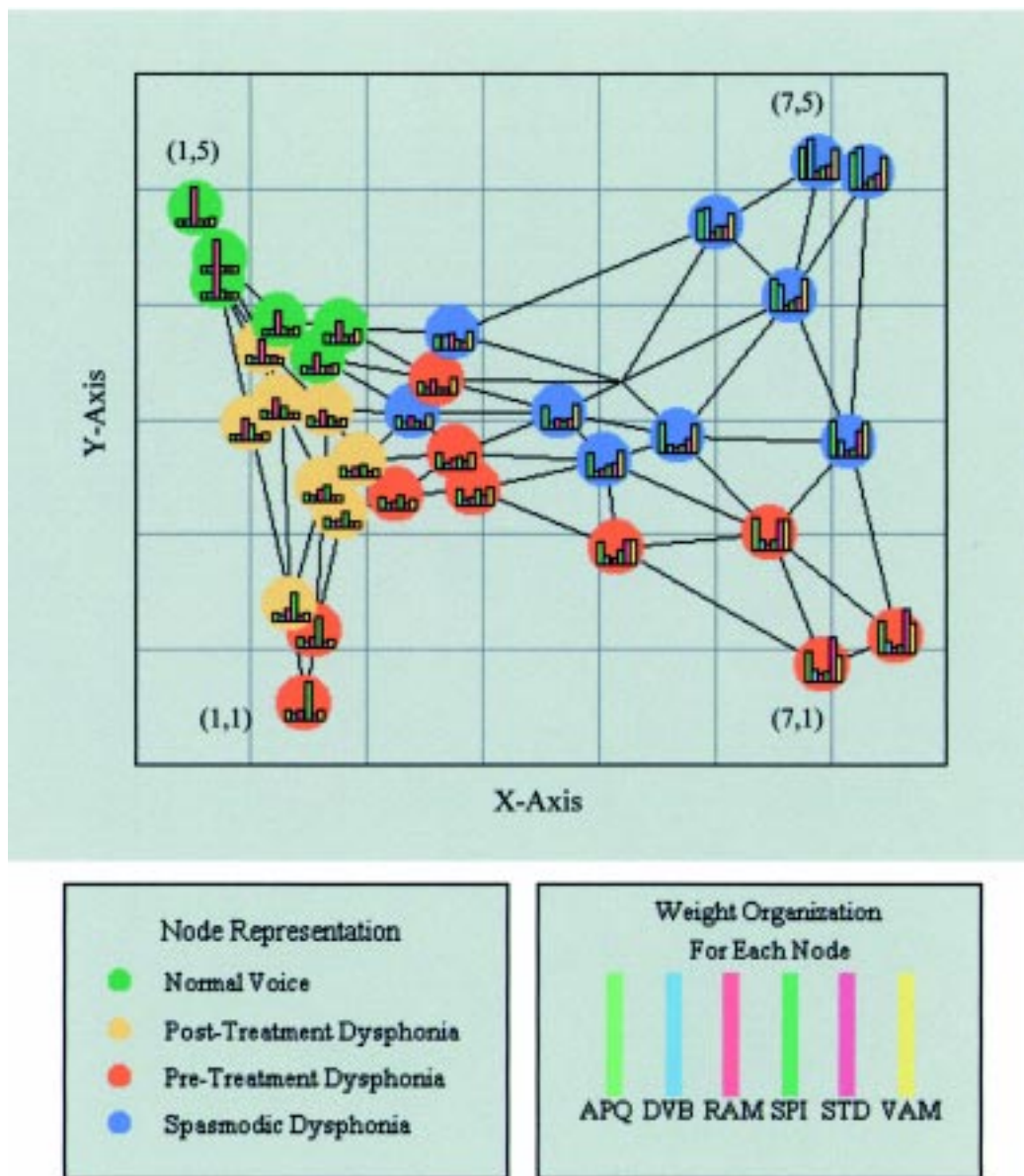
correctly (see Table 3), whereas the stepwise discriminant analysis only classified 50% of the 30 PR exemplars correctly (see Table 2). The SOM also demonstrates better classification performance for SD and PS exemplars than that of the stepwise discriminant analysis. The SOM classified 73.3% of both the 30 SD and the 30 PS exemplars correctly (see Table 2), whereas the stepwise discriminant analysis only classified 66.7% of the 30 SD exemplars and 70% of the PS exemplars correctly (see Table 2). In the case of classification of NR exemplars, the stepwise discriminant analysis is slightly better than that of the SOM. The stepwise discriminant analysis classified 86.7% of the NR exemplars correctly

(see Table 2), whereas the SOM classified 83.3% of the NR exemplars correctly (see Table 3).

Weighting of the Acoustic Measures

The two-dimensional nature of the SOM provides an easy way to visualize and evaluate the weighting of the various acoustic measures underlying the representational map. It is important to note that a very low weight value may be just as important as a very high one in discriminating the various groups. The small bar plot on top of each node in Figure 3 represents the strength of the corresponding weight values for each of

Figure 3. The strength of the weight values is shown across the group node representations: blue = spasmodic dysphonia (SD), red = pretreatment functional dysphonia (PR), yellow = post-treatment functional dysphonia (PS), green = normal voice (NR) of the SOM for each of the six underlying acoustic parameters: amplitude perturbation quotient (APQ), degree of voice breaks (DVB), rahmonic amplitude (RAM), soft phonation index (SPI), standard deviation of the fundamental frequency (STD), and variation of peak-to-peak amplitude (VAM). The x and y axes represent the Euclidean coordinates for each of the nodes defined by the Sammon mapping (Sammon, 1969). The node identifiers for each of the corners of the SOM are given (1, 1), (1, 5), (7, 5), and (7, 1). The small bar plot on top of each node represents the strength of the corresponding weight values for each of the underlying acoustic parameters (from left to right: APQ, DVB, RAM, SPI, STD, and VAM). The relative importance of the underlying acoustic parameters in classifying the various groups across the SOM can be determined by comparing the pattern of the corresponding weight strengths to the regions of the SOM that represent the four groups.



the underlying acoustic parameters. The relative importance of the underlying acoustic parameters in classifying the various groups across the SOM can be determined by comparing the pattern of the corresponding weight values to the regions of the SOM that represent the four groups. In order to determine which acoustic parameters are responsible for classification of the four

groups across the SOM, stepwise discriminant analyses were carried out over the weights for each of the pairwise group node representations. (The results for each of the analyses are given in Table 4.) The results of the stepwise discriminant analysis for SD node representations (blue circles in Figure 3) and PR node representations (red circles in Figure 3) indicate that when

Table 2. Classification results—stepwise discriminant analysis over original data.

Total Correct 82 N = 120		Predicted							
		SD		PR		PS		NR	
		#	%	#	%	#	%	#	%
Actual	SD (N = 30)	<u>20</u>	<u>66.7</u>	3	10.0	7	23.3	0	0.0
	PR (N = 30)	5	16.7	<u>15</u>	<u>50.0</u>	9	30.0	1	3.3
	PS (N = 30)	1	3.3	2	6.7	<u>21</u>	<u>70.0</u>	6	20.0
	NR (N = 30)	0	0.0	0	0.0	4	13.3	<u>26</u>	<u>86.7</u>

Table 3. Classification results—self-organizing map.

Total Correct 91 N = 120		Predicted							
		SD		PR		PS		NR	
		#	%	#	%	#	%	#	%
Actual	SD (N = 30)	<u>22</u>	<u>73.3</u>	4	13.3	4	13.3	0	0.0
	PR (N = 30)	5	16.7	<u>22</u>	<u>73.3</u>	3	10.0	0	0.0
	PS (N = 30)	1	3.3	4	13.3	<u>22</u>	<u>73.3</u>	3	10.0
	NR (N = 30)	0	0.0	1	3.3	4	13.3	<u>25</u>	<u>83.3</u>

DVB is withheld from the analysis, VAM and STD account for 90.9% of SD node classification and 90% of PR node classification (see Table 4). Figure 3 indicates that SD node representations have higher VAM (sixth bar) weight values relative to STD (fifth bar) weight values than PR node representations.

The results of the stepwise discriminant analysis for SD node representations (blue circles in Figure 3) and PS node representations (yellow circles in Figure 3)

indicate that VAM and APQ account for 90.9% of SD node classification and 100% of PS node classification (see Table 4). Figure 3 indicates that SD node representations are characterized by high VAM (sixth bar) and high APQ (first bar) weight values; whereas PS node representations are characterized by low VAM and low APQ weight values. The results of the stepwise discriminant analysis for PR node representations (red circles in Figure 3) and PS node representations (yellow circles

Table 4. Stepwise discriminant analysis of the node weights for each of the pairwise node classifications across the SOM. At each step, the variable that minimizes the overall Wilks's lambda is entered. Minimum significance of *F* to enter is .05. Maximum significance of *F* to remove is .1.

	SD (N = 10)					PR (N = 8)					PS (N = 6)				
	Loaded vars. step	Wilks's lambda	<i>F</i>	sig	Class. percent	Loaded vars. step	Wilks's lambda	<i>F</i>	sig	Class. percent	Loaded vars. step	Wilks's lambda	<i>F</i>	sig	Class. percent
PR (N = 10)															
1. DVB	.736	6.8	.001	SD = 54.5											
				PR = 90.0											
w/o DVB															
1. VAM	.774	5.6	.029	SD = 90.9											
2. STD	.365	15.6	.001	PR = 90.0											
PS (N = 8)															
1. VAM	.190	72.5	.001	SD = 90.9	1. VAM	.509	15.4	.001	PR = 90.0						
2. APQ	.112	63.2	.001	PS = 100	2. SPI	.381	12.2	.001	PS = 87.5						
NR (N = 6)															
1. RAM	.189	64.2	.001	SD = 100	1. RAM	.179	65.7	.001	PR = 100	1. RAM	.455	14.4	.003	PS = 100	
				NR = 100					NR = 100	2. VAM	.234	18.0	.001	NR = 100	

in Figure 3) indicate that VAM and SPI account for 90% of PR node classification and 87.5% of PS node classification (see Table 4). Figure 3 indicates that PR node representations have higher VAM (sixth bar) weight values or higher SPI (fourth bar) weight values than PS node representations.

The results of the stepwise discriminant analysis for SD node representations (blue circles in Figure 3) and NR node representations (green circles in Figure 3) indicate that RAM accounts for 100% of SD node classification and 100% of NR node classification (see Table 4). Figure 3 indicates that SD node representations have much lower RAM (third bar) weight values than NR node representations. The same pattern of results exists for PR and NR node classification (see Table 4 and Figure 3). The results of the stepwise discriminant analysis for PS node representations (yellow circles in Figure 3) and NR node representations (green circles in Figure 3) indicate that RAM and VAM account for 100% of PS node classification and 100% of NR node classification. Figure 3 indicates that PS node representations have combined lower RAM (third bar) and VAM (sixth bar) weight values than PR node representations.

On the basis of the results of the pairwise stepwise discriminant analyses (Table 4) as well as visual inspection of the weight values in Figure 3, the following generalizations can be made regarding the relative importance of the acoustic parameters in determining the group node representations: SD node representations are characterized by relatively high values of VAM, APQ, and DVB, as well as low values of RAM. Most of the nodes with PR representations are characterized by relatively high values of STD and VAM, with low values of RAM. Two nodes with PR representations in the region of node (1, 1) have high SPI levels. PS node representations are characterized by mid range values of RAM and SPI, as well as low values of APQ and VAM. NR node representations are characterized by high RAM values and relatively low values of all other acoustic parameters. Overall, it appears that disordered voice (SD and PR) is characterized by greater variability in amplitude—both long term (APQ, smoothing factor 11 periods) and short term (VAM, period-to-period)—and greater variability in fundamental frequency (STD). Voice that is considered nonpathologic (NR and PS) is characterized by a low degree of variability in amplitude (APQ and VAM) and fundamental frequency (STD), as well as a high degree of harmonic energy (RAM).

Conclusion

The SOM provided an objective means of assessing voice quality, on the basis of multiple acoustic measures. Analysis of the SOM allowed for an easy way to visualize

and evaluate the classification of exemplars of the various groups. The SOM demonstrated better classification performance over that given by a stepwise discriminant analysis of the original data. Analysis of the SOM also provided for a description of the relative contribution of the various acoustic measures underlying the representational map that are relevant in classification of the various groups. Further research needs to be conducted using a larger sample size to train the SOM. A larger sample size will generate maps that are able to classify novel exemplars accurately. The SOM obtained by this method may be a valuable complement to classifications that are based on auditory-perceptual ratings. It would be interesting to compare how perceptual ratings of voice quality matched up with the location, on the basis of acoustic properties, of an exemplar on the SOM.

We regard the results of this study as a promising initial step into the use of SOMs with multiple acoustic measures to assess phonatory function. Certain limitations of this present work should be noted in future applications. First, sustained vowel phonation is a convenient sample for analysis, but it probably is not sufficient to reveal the complexity of the phonatory disorders addressed here. For example, spasmodic dysphonia can be highly task sensitive, and it is possible that somewhat different features of voice disturbance should be used with tasks such as conversation or passage reading. Still, the successful classification in this study indicates that even with the relatively simple task of sustained vowel phonation, it is possible to delineate differences in clinical populations. Second, the small number of acoustic measures used in this study may not be the best ones to use in more extensive studies of voice disorders in various clinical populations. Selection of acoustic measures to be used as input to the SOM is matter that requires further work. With a larger number of exemplars, it would be possible to increase the number of acoustic measures used as input to the SOM. Third, the SOM offers interesting possibilities for the comparison of different types of analysis, including the use of input data from derived acoustic measures (as in the case of the present study), spectral vectors (Leinonen, et al., 1992), and perceptual ratings (Leinonen, et al., 1997). It would be valuable to compare the SOM results obtained with different types of input data. Such a comparison would be especially pertinent to understanding the relationships between perceptual ratings and acoustic analyses of voice qualities.

Acknowledgments

This work was supported in part by NIH Research Grant 5 R01 DC 00319-11 as well as by the National Center for Voice and Speech Grant P60 00976 from the National Institute on Deafness and Other Communication Disorders.

References

- Arends, N., Povel, D. J., Van Os, E., & Speth, L.** (1990). Predicting voice quality of deaf speakers on the basis of glottal characteristics. *Journal of Speech and Hearing Disorders, 29*, 156–170.
- Aronson, A. E.** (1973). *Psychogenic voice disorders: An interdisciplinary approach to detection, diagnosis, and therapy*. Philadelphia: W. B. Saunders.
- Aronson, A. E., Peterson, H. W., & Litin, E. M.** (1966). Psychiatric symptomatology in function dysphonia and aphonia. *Journal of Speech and Hearing Disorders, 31*, 115–127.
- Cannito, M. P., McSwain, L. S., & Dworkin, J. P.** (1996). Abductor spasmodic dysphonia: Acoustic influence of voicing on connected speech. In D. A. Robin, K. M. Yorkston, & D. R. Beukelman (Eds.), *Disorders of motor speech: Assessment, treatment, and clinical characterization*. Baltimore: Paul H. Brookes, 311–328.
- CSL [Computer program].** (1994). Lincoln Park, NJ: Kay Elemetrics.
- Dejonckere, P. H., & Wieneke, G. H.** (1992). Cepstrum analysis of pathological voices: Correlations with other acoustic parameters and with perceptual scaling. In *Proceedings of the 2nd European Congress of Oto-rhinolaryngology and Cervico-facial Surgery: Vol. 4. The new frontiers of oto-rhino-laryngology in Europe* (pp. 75–78). Bologna, Italy: Monduzzi.
- Dejonckere, P. H., & Wieneke, G. H.** (1993). Cepstra of normal and pathological voices, in correlation with acoustic, aerodynamic, and perceptual data. In M. J. Ball & M. Duckworth (Eds.), *Advances in clinical phonetics* (pp. 217–226). Amsterdam: John Benjamins.
- Dejonckere, P. H., & Wieneke, G. H.** (1994). Spectral, cepstral, and aperiodicity characteristics of pathological voices before and after phonosurgical treatment. *Clinical Linguistics and Phonetics, 8*, 161–169.
- Eskenazi, L., Childers, D. G., & Hicks, D. M.** (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research, 33*, 298–306.
- Hillenbrand, J., Cleveland, R. A., & Erickson, R. L.** (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech & Hearing Research, 37*, 769–78.
- Hartman, D. E., & Aronson, A. E.** (1981). Clinical investigations of intermittent breathy dysphonia. *Journal of Speech and Hearing Disorders, 46*, 428–432.
- Hillenbrand, J., & Houde, R. A.** (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech & Hearing Research, 39*, 311–321.
- Hirano, M.** (1981). *Clinical examination of voice*. Vienna: Springer, 83–84.
- Kent, R. D., Kim, H., Weismer, G., Kent, J. F., Rosenbek, J. C., Brooks, G. R., & Workinger, M.** (1994). Laryngeal dysfunction in neurological disease: Amyotrophic lateral sclerosis, Parkinson's disease, and stroke. *Journal of Medical Speech-Language Pathology, 2*, 157–175.
- Kohonen, T.** (1995). *Self-organizing maps*. Berlin, Springer, 77–127.
- Kohonen, T., Hynninen, J., Kangas, J., & Laaksonen, J.** (1996). *SOM_PAK: The Self-Organizing Map Program Package* (Report A31). Helsinki University of Technology, Laboratory of Computer and Information Science.
- Klingholz, F., & Martin, F.** (1985). Quantitative spectral evaluation of shimmer and jitter. *Journal of Speech and Hearing Research, 28*, 169–174.
- Koike, Y.** (1986). Cepstrum analysis of pathologic voices. *Journal of Phonetics, 14*, 501–507.
- Kotby, M. N., Titze, I. R., Saleh, M. M., & Berry, D. A.** (1993). Fundamental frequency stability in functional dysphonia. *NCVS Status and Progress Report, 4*, 201–207.
- Kreiman, J., & Gerratt, B. R.** (1996). The perceptual structure of pathologic voice quality. *Journal of the Acoustical Society of America, 100*, 1787–1795.
- Kreiman, J., Gerratt, B. R., & Burke, G. S.** (1994). The multidimensional nature of pathologic vocal quality. *Journal of the Acoustical Society of America, 96*, 1291–1313.
- Kreiman, J., Gerratt, B. R., & Precoda, K.** (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research, 33*, 103–115.
- Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S.** (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35*, 512–520.
- Leinonen, I., Hiltunen, T., Laakso, M., Rihkanen, H., & Poppius, H.** (1997). Categorization of voice disorders with six perceptual dimensions. *Folia Phoniatrica et Logopaedica, 49*, 9–20.
- Leinonen, I., Kangas, J., Torkkola, K., & Juvas, A.** (1992). Dysphonia detected by pattern recognition of spectral composition. *Journal of Speech and Hearing Research, 35*, 287–295.
- Leinonen, I., Hiltunen, T., Kangas, J., Juvas, A., & Rihkanen, H.** (1993). Detection of dysphonia by pattern recognition of speech spectra. *Scandinavian Journal of Logopaedica Phoniatrica, 18*, 159–167.
- Ludlow, C. L., & Conner, N. P.** (1987). Dynamic aspects of phonatory control in spasmodic dysphonia. *Journal of Speech and Hearing Research, 30*, 197–206.
- Ludlow, C. L., Naunton, R. F., & Bassich, C. J.** (1984). Procedures for the selection of spastic dysphonia patients for recurrent laryngeal nerve section. *Otolaryngology Head and Neck Surgery, 92*, 24–31.
- Ludlow, C. L., Naunton, R. F., Terata, S., & Anderson, B. J.** (1991). Successful treatment of selected cases of abductor spasmodic dysphonia using botulinum toxin injection. *Otolaryngology, Head and Neck Surgery, 104*, 849–855.
- Matlab** (Version 4.2c.1) [Computer software]. (1994). Natick, MA: MathWorks.
- Murry, T., Singh, S., & Sargent, M.** (1977). Multidimensional classification of abnormal voice qualities. *Journal of the Acoustical Society of America, 61*, 1630–1635.
- Perrin, E., Berger-Vachon, C., Kauffman, I., & Collect, L.** (1997). Acoustical recognition of laryngeal pathology using the fundamental frequency and the first three formants of vowels. *Medical and Biological Engineering and Computing, 35*, 361–368.
- Rihkanen, H., Leinonen, I., Hiltunen, T., & Kangas, J.**

- (1994). Spectral pattern recognition of improved voice quality. *Journal of Voice*, 8, 320–326.
- Rontal, M., Rontal, E., Rolnick, M., Merson, R., Silverman, B., & Truong, D. D.** (1991). A method for the treatment of abductor spasmodic dysphonia with botulinum toxin injections: A preliminary report. *Laryngoscope*, 84, 662–671.
- Roy, N., & Leeper, N.** (1993). Effects of the manual musculoskeletal tension reduction technique as a treatment for functional voice disorders: Perceptual and acoustic measures. *Journal of Voice*, 7, 242–249.
- Roy, N., & Tasko, S.** (1994). Speaking fundamental frequency (SFF) changes following successful management of functional dysphonia. *Journal of Speech Language Pathology and Audiology*, 18 (2), 115–120.
- Sammon, J.** (1969). A non-linear mapping for data structure analysis. *IEEE Transactions on Computers*, 18, 401–409.
- SPSS** (Version 7.5.1) [Computer program]. (1996). Chicago: SPSS.
- Voice Disorders Database** (Version 1.03) [Computer software]. (1994). Lincoln Park, NJ: Kay Elemetrics.
- Wilson, J. A., Fitch, J., & Cornell, R.** (1995). Acoustic prediction of severity in commonly occurring voice problems. *Journal of Speech and Hearing Research*, 38, 273–279.
- Wolfe, V., Fitch, J., & Cornell, R.** (1995). Acoustic prediction of severity in commonly occurring voice problems. *Journal of Speech and Hearing Research*, 38, 273–279.
- Wolfe, V. I., & Ratusnik, D. L.** (1988). Acoustic and perceptual measurements of roughness influencing judgments of pitch. *Journal of Speech and Hearing Disorders*, 53, 15–22.
- Wolfe, V. I., & Steinfatt, T. M.** (1987). Prediction of vocal severity within and across voice types. *Journal of Speech and Hearing Research*, 30, 230–240.
- Zwirner, P., Murry, T., Swenson, M., & Woodson, G. E.** (1992). Effects of botulinum toxin therapy in patients with adductor spasmodic dysphonia: Acoustic, aerodynamic, and videoendoscopic findings. *Laryngoscope*, 102, 400–406.
- Zwirner, P., Murry, T., & Woodson, G. E.** (1991). Phonatory function of neurologically impaired patients. *Journal of Communication Disorders*, 24, 287–300.

Received June 1, 1998

Accepted October 30, 1998

Contact author:

Daniel E. Callan, ATR Human Information Processing Research Laboratories, 2-2 Hikoridai, Seika-cho, Sorakugun, Kyoto 619-0288. Email: dcallan@hip.atr.co.jp