ON THE SELECTION OF MEANINGFUL SPEECH PARAMETERS USED BY A PATHOLOGIC/NON PATHOLOGIC VOICE REGISTER CLASSIFIER

Juan I. Godino-Llorente, Santiago Aguilera-Navarro, Carlos Hernández-Espinosa, Mercedes Fernández-Redondo, Pedro Gómez-Vilda

LTR (Lab. de Tecnología de Rehabilitación), E.T.S.I. de Telecomunicación, Ciudad Universitaria, 28041 Madrid, Spain. Tel.: 34.1.5495700 Ext. 540  Fax: +34.1.336.73.23  e-mail: {godino, aguilera}@die.upm.es

Dpto. de Informática, Universidad Jaume I, Campus Riu Sec, Castellón, Spain

Facultad de Informática, UPM, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain

ABSTRACT

Most of vocal and voice diseases cause changes in the voice. These diseases have to be diagnosed and treated during an early stage. There is an increased risk for vocal and voice diseases due to the modern way of life. Acoustic voice analysis is an effective and non-invasive tool due to: a) Objective support of the diagnostics. b) Screening the vocal and voice diseases and especially their early detection. c) Objective determination of the impairment of the vocal function. d) Objective evaluation of the effect of the air pollution on the voice. e) Evaluation of surgical and pharmacological treatments. f) Evaluation of the rehabilitation.

Many algorithms to calculate acoustic parameters have been developed and it is demonstrated that there is a great correlation between deviations of parameters and pathologies. The effectiveness and importance of the acoustic analysis of pathological voices has been proven by many experimental researches demonstrating that acoustic parameters of pathologic voices are deviated from the mean.

The authors have focused their task in separation of pathologic/non pathologic voices, and evaluating the meaningful acoustic parameters by means of neural network technology and pruning methods.

1. INTRODUCCION

It is well known that most of the vocal and voice diseases cause changes in the acoustic voice signal. These diseases have to be diagnosed and treated during the early stage. Acoustic voice analysis helps us to classify voice registers into pathological/non pathological.

Usually, analysis of pathological voice signals is carried out by means of acoustic parameter analysis. Such parameters are extracted from the voice signal using digital signal processing techniques. In the bibliography there are a wide number of different parameters that may be extracted and studied but, ENT specialists and speech therapists do not use most of them because they do not provide helpful information.

The authors are involved in the task to classify speech registers in pathogenic and non-pathologic classes, and to decide which are the most significant acoustic speech parameters.

2. DATABASE USED

Kay Elemetrics has recorded to CD-ROM a database developed by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab. It contains over 1,400 voice samples of approximately 700 subjects. Included are sustained phonation and running speech samples from patients with a wide variety of organic, neurological, traumatic, and psychogenic voice disorders, as well as normal voices.

All of the speech samples were collected in a controlled environment with 25 kHz or 50 kHz sampling rate, and 16-bit resolution.

Acoustic parameters are calculated using Multi-Dimensional Voice Program (MDVP™, which calculates over 30 parameters.


The database contains sustained phonation and running speech samples, but due to the non-stationary features of the speech signal, extraction of acoustic parameters is carried out over sustained vowel phonation. Phoneme /ae/ has been studied.

3. PRUNING THE DATABASE

The first step is pruning the database because examining the different registers, it is observed that the same register appears labelled with two, three or more pathologies. In order to exclude those registers that appear more than once, the database have to be pruned manually.

Once wrong-labelled registers are pruned, there are 360 registers (from a set of 1400) left: 53 are normal voices, and the rest, pathological. All of them correspond to the phonation of the English vowel /ae/. Each register is quantified using a n-dimensional vector composed by 26 acoustic parameters extracted from the voice register.

5. THE CLASSIFIER: A NEURAL NET

The authors are using a widely used classifier in pattern recognition: a neural network. A multilayer feedforward perceptron (MLP) has been chosen. The Learning algorithm used is backpropagation with momentum. Such architecture is widely used in pattern classification.

It is possible to distinguish an input layer, a hidden and an output layer (Figure 1). The output of each neurone can be calculated by means of the next expression [2]:
Let $x=(x_1,x_2,\ldots,x_n)$ be an observation of the n-dimensional Euclidean feature space $E_n$. Then, $x$ can be represented by:

$$x = \sum_{i=1}^{N} x_i \cdot e_i$$

where $(e_1, e_2, \ldots, e_N)$ is a basis for $E_n$, and $e_i$ is a vector with all zero entries except the $ith$ feature. Feature selection is equivalent to projecting an observation into a $k$-dimensional subspace denoted by $(e'_1, e'_2, \ldots, e'_k)$.

Data are normalised before giving to the net. Criterion used to normalise input features is as follows:

$$x'_i = \frac{x_i - m_i}{v_i}$$

where $m_i$ is the mean value and $v_i$ is the sample standard deviation of feature $i$.

5.1 Net size

A MLP with a single hidden layer has been used. The number of neurones in the input layer is 26 due to the fact that we dispose of 26 acoustic parameters as features to represent the voice signal. The number of hidden layers is a parameter to be adjusted during the training phase. Output layer has a single neuron that will be “1” or “0” activated depending if we are pathologic or non-pathologic voice registers.

Choosing the net size is a critical problem: the smaller net (smaller number of hidden units) with good generalisation capability should be chosen.

5.2 Training and simulation

Weights are randomly initialised.

Training is carried out using an error backpropagation algorithm that allows modifying momentum and learning rate.

The authors have divided data into two subsets: the first subset will be used to train the net (70%), the second, will be used to simulate or validate the results.

Data are normalised before giving to the net.

5.3 Feature selection

Let $x=(x_1, x_2, \ldots, x_n)$ be an observation of the n-dimensional Euclidean feature space $E_n$. Then, $x$ can be represented by:

$$x = \sum_{i=1}^{N} x_i \cdot e_i$$

where $(e_1, e_2, \ldots, e_n)$ is a basis for $E_n$, and $e_i$ is a vector with all zero entries except the $ith$ feature. Feature selection is equivalent to projecting an observation into a $k$-dimensional subspace denoted by $(e'_1, e'_2, \ldots, e'_k)$.

Data are normalised before giving to the net. Criterion used to normalise input features is as follows:

$${\text{Compute sample mean vector } m(p)= (m_1(p), m_2(p), \ldots, m_n(p)) \text{ and } m(n)= (m_1(n), m_2(n), \ldots, m_n(n)) \text{; compute sample variance vector } v(p)= (v_1(p), v_2(p), \ldots, v_n(p)) \text{ and } v(n)= (v_1(n), v_2(n), \ldots, v_n(n)), \text{ using training sample vectors of class non-pathological (n) and pathological (p)} \text{.}$$

Normalise every training sample vector:

$$x'_{j}(p)= \frac{x_{j}(p) - m_{j}(p)}{v_{j}(p)} \rightarrow \text{Class: pathologic}$$

$$x'_{j}(n)= \frac{x_{j}(n) - m_{j}(n)}{v_{j}(n)} \rightarrow \text{Class: non-pathologic}$$

Once data are normalised, the net has been trained and has been tested using the test subset of input patterns.

The authors would like to determine the best features subset $NI$ of a set containing the $N$ input features (26 acoustic parameters). The goal is to remove non-significant input features reducing the input space by discarding irrelevant features. Methods used to prune the input features space are called “pruning methods”.

Five different pruning methods have been applied. The first one is based on an analysis of the training set and the rest on an analysis of a trained multilayer feedforward network.

An extensive revision of feature selection methods for neural networks and an experimental comparison among them can be found in [6], [7].

5.3.1. 1st Criterion:

Based on statistics calculated from input patterns (1st method). It is described in [5].

Let $(m_{i}^{(p)}, \sigma_i^{(p)})$ and $(m_{i}^{(n)}, \sigma_i^{(n)})$ be two pairs of the sample mean value and sample standard deviation of feature $i$ computed from the training sets of both classes. Feature $i$ is said to have higher discriminating capability than feature $j$ if

$$\frac{|m_{i}^{(p)} - m_{i}^{(n)}|}{(\sigma_i^{(p)})^2 + (\sigma_i^{(n)})^2} > \frac{|m_{j}^{(p)} - m_{j}^{(n)}|}{(\sigma_j^{(p)})^2 + (\sigma_j^{(n)})^2}$$

5.3.2. 2nd Criterion:

Based on calculus of sensibilities of the weights of the hidden layer and the input features. Those parameters with the smallest sensibilities are pruned. Sensibility $S_i$ of input variable $i$ is introduced in ref. [4], and could be calculated by:

$$S_i = \sum_{j \in N_i} S_{ij} = \sum_{j=1}^{N} S_{ij}$$

where $S_{ij}$ is a sensitivity of a weight $w_{ij}$, and summation is over a set $N_i$ of outgoing weights of the $ith$ neuron. Or using another order of weight summation, $s_{ij}$ is a sensitivity of a weight $w_{ij}$ connecting the $ith$ neuron to the $jth$ neuron in the next layer.

Generation of colour maps (Hinton diagrams) is useful for visual determination of the most important input variables, but is rather subjective. Therefore the absolute magnitude of a weight may be used as its sensitivity:

$$S_i = |W_i|$$

5.3.3. 3rd Criterion:

The sensibility $S_i$ of the input feature $i$, is calculated according to [4]:
\[ S_j = \sum_{j=1}^{n} \left( \frac{W_{ij}}{\max_{j} W_{ij}} \right)^2 \]

Where \( \max_j \) is taken over all weights ending at neuron \( j \).

### 5.3.4. 4\(^{th}\) Criterion

Based on normalised input sensibility [3] of the output variables with respect to the input features. (2\(^{nd}\) and 3\(^{rd}\) methods). Those parameters with the smallest sensitivities are pruned. Sensibility \( \sigma_k \) of \( k \)th output variable with respect to the \( i \)th input feature is introduced in [3]

Those input features with the smallest sensitivities are pruned. Sensibility \( \sigma_k \) of \( k \)th output variable with respect to the \( i \)th input feature is introduced in [3]

Sensibility \( \sigma_{ki} \) of the input feature \( i \) with respect to output \( k \) is calculated by means of:

\[ S_{ki} = \sum_{j} W_{kj} W_{ij} h_j (1 - h_j) \]

\[ \sigma_{ki} = \sqrt{\frac{s_{ki}}{\sum_{j} s_{kj}^2}} \]

Such calculus must be considered for every input pattern. The maximum values of normalised sensitivities are selected. The complete algorithm is described in [3]

This criterion is based in the calculus of Jacobian sensitivity matrix of outputs with respect to input vector components (as explained in [3]).

### 5.3.5. 5\(^{th}\) Criterion:

Sensibility \( \sigma_{ki} \) of the input feature \( i \) with respect to output \( k \) is calculated by means of:

\[ S_{ki} = \sum_{j} W_{kj} W_{ij} h_j (1 - h_j) \]

\[ \sigma_{ki} = \sqrt{\frac{x_i \cdot s_{ki}}{\sum_j x_i \cdot s_{kj}^2}} \]

Such calculus must be considered for every input pattern. The maximum values of normalised sensitivities are selected. The complete algorithm is described in [3]

This criterion is based in the calculus of Logarithmic sensitivity matrix of outputs with respect to input vector components (as described in [3]).

### 6. RESULTS

Training begins with 1000 epochs. The number of epochs was decreased progressively. Sum of mean squared error is controlled as parameter to stop training. The net was trained, in the first stage, using the whole features set we have (26 parameters).

Table 1 shows the error as function of number of neurons of the hidden layer. The ratio of misclassification obtained is really good. So we have involved ourselves in the task of pruning the input pattern space, in order to use a shorter number of acoustic parameters. The goal is to classify using the shortest number of acoustic parameters.

#### 6.1 Performance detector matrix

The authors distinguish several kinds of error:

- **Correct Rejection**: detector found no event when indeed none was present.
- **Correct detection**: detector found an event when one was present.
- **False negative**: the classifier missed an event
- **False positive**: the detector found an event when none was present.
- **Total error**: percentage of erroneous decisions
- **Percent correct detection**: CD/T2
- **Percent correct rejections**: CR/T1
- **Percent false positives**: FP/T2
- **Percent false negatives**: FN/T1
- **Total error**: TE=(FP+FN)/(T1+T2)

<table>
<thead>
<tr>
<th>Nº NEUR.</th>
<th>CD/T1</th>
<th>CR/T1</th>
<th>FP/T2</th>
<th>FN/T1</th>
<th>TE</th>
<th>SSE</th>
<th>Nº epch.</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>94.87</td>
<td>100</td>
<td>0</td>
<td>5.9</td>
<td>3.48</td>
<td>9.8e-009</td>
<td>437</td>
</tr>
<tr>
<td>10</td>
<td>94.06</td>
<td>100</td>
<td>0</td>
<td>5.9</td>
<td>4.06</td>
<td>9.9e-009</td>
<td>446</td>
</tr>
<tr>
<td>7</td>
<td>94.06</td>
<td>100</td>
<td>0</td>
<td>5.9</td>
<td>4.06</td>
<td>9.7e-009</td>
<td>459</td>
</tr>
<tr>
<td>5</td>
<td>94.06</td>
<td>100</td>
<td>0</td>
<td>5.9</td>
<td>4.06</td>
<td>9.9e-009</td>
<td>446</td>
</tr>
<tr>
<td>4</td>
<td>94.06</td>
<td>100</td>
<td>0</td>
<td>5.9</td>
<td>4.06</td>
<td>9.9e-009</td>
<td>459</td>
</tr>
<tr>
<td>2</td>
<td>94.06</td>
<td>100</td>
<td>0</td>
<td>5.9</td>
<td>4.06</td>
<td>9.7e-009</td>
<td>510</td>
</tr>
<tr>
<td>1</td>
<td>94.06</td>
<td>100</td>
<td>0</td>
<td>5.9</td>
<td>4.06</td>
<td>9.7e-009</td>
<td>487</td>
</tr>
</tbody>
</table>

### 6.2 Meaningful features

Parameters extracted from the voice register are:


The next table shows parameters ordered by their importance according to the five criteria. Meaningful features appear at the bottom of the table.
Neural networks technology seems to be a promising tool to classify voice registers attending to their condition of pathological or non-pathological. Anyway, we have to be wise because the database stores a collection of very significant medical cases. Conclusions have to be tested with a larger database.

### 6.1.1. 1st Criterion
Applying those techniques described in [5] and neural network technology, classifying into pathologic/non-pathologic with the same error ratio is carried out using only two input features: (NHR and VTI). The neural network has only a single hidden layer with a single neuron and two input features.

### 6.1.2. 2nd Criterion
Applying techniques described in [4] and neural network technology, classifying into pathologic/non-pathologic with the same error ratio is carried out using only two input features: (NHR and VTI). The neural network has only a single hidden layer with a single neuron and two input features.

### 6.1.3. 3rd Criterion
Applying techniques described in [4] and neural network technology, classifying into pathologic/non-pathologic with the same error ratio is carried out using only two input features: (Fo and Fhi). The neural network has only a single hidden layer with a single neuron and two input features.

### 6.1.4. 4th Criterion
Applying techniques described in [3] (related with calculus of Jacobian Sensitivity) and neural network technology, classifying into pathologic/non-pathologic with the same error ratio is carried out using only two input features: (VTI and NHR). The neural network has only a single hidden layer with a single neuron and two input features.

### 6.1.5. 5th Criterion
Applying techniques described in [3] (related with calculus of Logarithmic Sensitivity), and neural network technology, classifying into pathologic/non-pathologic with the same error ratio is carried out using only three input features features: (PER, Jita and Fhi). The neural network has only a single hidden layer with a single neuron and three input features.

### 7. CONCLUSIONS
Neural networks technology seems to be a promising tool to classify voice registers attending to their condition of pathological or non-pathological. Anyway, we have to be wise because the database stores a collection of very significant medical cases. Conclusions have to be tested with a larger database.

### 7. CONCLUSIONS
Neural networks technology seems to be a promising tool to classify voice registers attending to their condition of pathological or non-pathological. Anyway, we have to be wise because the database stores a collection of very significant medical cases. Conclusions have to be tested with a larger database.

### 8. FUTURE WORK
Due to the fact that it seems to be easy to distinguish between pathologic and non-pathologic voices by means of acoustic parameters and neural networks, the next step will be to distinguish between a set of pathologies. For this purpose we may use a similar scheme to the one proposed, trying to identify which are the most significant acoustic parameters for each pathology. Anyway, a new well-labelled database of pathological voices is needed.

### 9. ACKNOWLEDGEMENTS
This project has been financed and supported by the Health Ministerial of Spain (IMSERSO) : TER 96-1938-C02-01.

### 10. REFERENCES


### Table 3 Features ordered by their importance. Features at the bottom are the most important features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFR</td>
<td>ATRI</td>
</tr>
<tr>
<td>sAPQ</td>
<td>Jita</td>
</tr>
<tr>
<td>Shim</td>
<td>FTRI</td>
</tr>
<tr>
<td>Fatr</td>
<td>Fo</td>
</tr>
<tr>
<td>ATRI</td>
<td>FTRI</td>
</tr>
<tr>
<td>APQ</td>
<td>Fhi</td>
</tr>
<tr>
<td>sPPQ</td>
<td>Ffr</td>
</tr>
<tr>
<td>Jitt</td>
<td>APQ</td>
</tr>
<tr>
<td>To</td>
<td>sPPQ</td>
</tr>
<tr>
<td>PPQ</td>
<td>PER</td>
</tr>
<tr>
<td>RAP</td>
<td>RAP</td>
</tr>
<tr>
<td>FTRI</td>
<td>To</td>
</tr>
<tr>
<td>SEG</td>
<td>FTRI</td>
</tr>
<tr>
<td>ShdB</td>
<td>Fhi</td>
</tr>
<tr>
<td>NHR</td>
<td>VTI</td>
</tr>
<tr>
<td>VTI</td>
<td>NHR</td>
</tr>
</tbody>
</table>

Classifying can be done using one single hidden layer with one neuron. The number of input features is two. Meaningful acoustic parameters to diagnose voice diseases, depend on the pruning method used. But, taking a look to the last five rows in table 3, we can conclude that the meaningful features are (ShdB, NHR and VTI): selecting those features allow us to classify introducing some redundancy in the input pattern. A short description of those features is provided:

“NHR”: Noise-to-Harmonic Ratio is an average ratio of energy of the in-harmonic components in the range 1500-4500 Hz to the harmonic components energy in the range 70-4500 Hz.

“VTI”: Voice Turbulence Index is an average ratio of the spectral in-harmonic high-frequency energy to the spectral harmonic energy in stable phonation areas.

“ShdB”: Shimmer in dB gives an evaluation of the period-to-period variability of the peak-to-peak amplitude within the analyzed voice sample.

### 7. CONCLUSIONS
Neural networks technology seems to be a promising tool to classify voice registers attending to their condition of pathological or non-pathological. Anyway, we have to be wise because the database stores a collection of very significant medical cases. Conclusions have to be tested with a larger database.