Models and Analysis of Vocal
Emissions for Biomedical Applications
(MAVEBA 1999)
Firenze, Italy, September 1-3, 1999

DOES IT AFFECT FEATURE "SEX" ON AUTOMATIC DETECTION OF IMPAIRED VOICES?

[†]Juan I. Godino-Llorente, [†]Santiago Aguilera-Navarro,
[††]Sira E. Palazuelos-Cagigas, [††]José L. Martín-Sánchez
[†]LTR (Lab. de Tecnología de Rehabilitación), E.T.S.I. Telecomunicación,
Ciudad Universitaria, 28040 Madrid, Spain. Tlph:+34.1.5495700 Ext.540 Fax: +34.1.336 73 23
e-mail: {godino, aguilera}@die.upm.es
[††]Dpto. de Electrónica, Universidad de Alcala, Campus Universitario, N II, Km 33.6, 28871 Madrid

ABSTRACT

Voice registers are widely affected when voice diseases appear. These diseases have to be diagnosed and treated during an early stage. Detection of voice diseases may be carried out by means of acoustic analysis of voice register. Many algorithms to calculate acoustic parameters have been developed and have been demonstrated that there is a great correlation between parameter deviations and impairments presence.

The effectiveness and importance of the acoustic analysis of pathological voices have been proven by many experimental researches which demonstrate that acoustic parameters of pathological voices are deviated from the mean. So, voice registers can be vector quantified in order to classify into healthy and impaired voices.

It is well known that male and female voices have different acoustic properties. Due to this fact, we may think that feature gender has to be kept in mind as a new feature in order to detect voice impairment from the voice register alone.

The aim of this paper is to study the influence of the feature gender to carry out classification and automatic detection of voice diseases.

1. INTRODUCTION

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 detect voice disorders.

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.

It was found out that vocal and voice diseases generally cause the following changes in the voice register:

- Increase of hoarseness degree because pathological voice contains noisy components.
- Large variations in pitch periods and pitch peaks (pulses) amplitudes.
- Pitch period breaks generation during sustained vowel phonation.
- Presence of sub harmonics.
- Deformation of the pitch pulses shape.
- High frequency noisy components.
- Noisy components in spectra and cepstra.

This is tested by means of acoustic parameters, which allow us to measure the degree of "normality" of the voice signal.

The effectiveness and importance of the acoustic analysis of pathological voices have been proven by many experimental researches. A great amount of acoustic parameters are handled in the bibliography. The most important are:

1. Fundamental frequency: (measured in hertz). This parameter is especially important because other parameters are based on its accuracy [8]. In noisy conditions, due to the presence of laryngeal pathology, it is difficult to measure, therefore it is necessary to calculate it in three different domains: temporal, cepstral and frequencial. Boyanov et al. [9] propose estimation methods of the fundamental frequency.
2. Amplitude perturbation (shimmer): it measures the maximum variation of peak-to-peak amplitude of the signal. Energy, when

sustained phonation of a vowel, stays practically constant unless in pathology presence. In this case, important variations are produced [10]. Algorithm for shimmer measurement is proposed by Kasuya et al [11].

3. Pitch perturbation (jitter): it measures the variation of the fundamental frequency. The fundamental frequency in the sustained phonation of a vowel stays practically constant unless in pathology presence. In this case, important variations are produced [10]. Shimmer algorithm is proposed by Kasuya et al [11]. Ludlow et al demonstrated in [12] the possibility of detecting certain laryngeal pathology from jitter and shimmer measurements.
4. Low to high energy ratio. Pathological voices are characterised by an increase in the signal energy in high frequency components. Generally, it is taken as low frequency the 70 to1500 Hz range, and as high frequency the 1500 to 4000 Hz range. In Boyanov et al [13], a calculation method is proposed.
5. Harmonic to noise ratio: pathological voices are characterised by a smaller harmonic to noise ratio than healthy ones [14]. This is due to the non regularity of the vibration of the vocal cords and to the closure problems. De Krom proposes the HNR algorithm in [15].
6. Normalised noise energy: is a measurement of the ratio between the noise component and the harmonic component on the signal. Kasuya first proposed such measure in [16].

The authors are involved in the task of detecting voice impairment from the voice register. Classifying is carried out vector quantifying voice registers using acoustic parameters. We are studying the possibility of adding a new feature to quantify the voice register: *gender* (male or female).

In order to carry out classification it seems very interesting to include *gender* as a new feature to parameterise the voice register.

2. THEORETICAL ISSUES

The problem we want to solve can be framed as a classification problem with only two different classes: normal and impaired voices.

The voice register will be quantified by means of acoustic parameters plus features such as subject's gender.

The input to the classifier is a vector containing the whole feature set that represents the voice signal.

The classifier is based on neural networks. Once the classifier has been developed, we are involved in the task of selecting meaningful features. Our goal is to evaluate the discrimination capability of the feature *gender.*

First of all, we will describe the fundamentals of the classification techniques we have used: Neural Nets. In the following sections, feature selection techniques will be also described.

3.1 The classifier: a Neural Network

Neural networks are widely used as classifiers in pattern recognition. A multilayer feedforward perceptron (MLP) has been chosen. The learning algorithm used *is backpropagation with momentum* [2]. Such architecture is frecuently used in pattern classification.

It is possible to distinguish an input, hidden and an output layer (Figure 1). The output of each neurone can be calculated by means of the next expression [2]:

$$h_j = f(\sum_{i=1}^{N} w_{ji} \bullet x_i + \xi_i) \qquad y_k = f(\sum_{j=1}^{N_j} W_{kj} \bullet h_j + \theta_j)$$

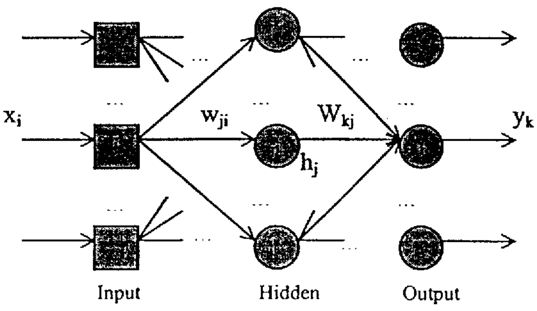

Figure 1: MLP with a single hidden layer

Where: x_i are the input features; θ_j and ξ_i are the thresholds, w_{ji} are the weights associated to the hidden layer; W_{kj} are the weights associated to the output layer; y_k are the net outputs; N is the number of neurones in the input layer; N_j is the number of neurones in the hidden layer; N_k is the number of neurones in the output layer; and, $f(\cdot)$ is the sigmoidal function [2]:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Weights w_{ji} and W_{kj} associated to the hidden and output layer are adjusted in the training phase to minimise the error of misclassification as described in [2].

It is demonstrated that a multilayer net with sigmoidal function at each neuron allows classification of non-linear patterns.

3.2 Meaningful features

Once the classifier has been developed, we would like to determine the best features subset n_l of the whole set containing n input features ($n_l>n$, $n=26$ acoustic parameters). Our objective 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" or "feature selection methods".

Pruning methods used are neural nets based. They are based on calculus of sensibilities of different parameters with respect to input features. An extensive revision of feature selection methods for neural networks and an experimental comparison among them can be found in [6], [7].

3.2.1. 1st Criterion:

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

$$S_i = \sum_{k\in\Omega_i} s_k = \sum_{j=1}^{n_j} s_{ij}$$

Where s_k is a sensitivity of a weight w_k, and summation is over a set Ω_k of outgoing weights of the *ith* neuron. Or using another order of weight summation, s_{ji} is a sensitivity of a weight w_{ji} 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:

$$S_i = |w_k|$$

3.2.2 2nd Criterion:

The sensibility S_i of the input feature i, is calculated according to [4]:

$$S_i = \sum_{j=1}^{n_j} \left(\frac{w_{ji}}{max_a |w_{ja}|} \right)^2$$

Where max_a is taken over all weights ending at neuron *j*.

3.2.3. 3rd Criterion

Based on normalised input sensibility [3] of the output variables with respect to the input features. (2nd and 3rd methods). Those parameters with the smallest sensitivities are pruned. Sensibility σ_{ki} of *kth* output variable with respect to the *ith* input feature is introduced in [3]

Those input features with the smallest sensitivities are pruned. Sensibility σ_{ki} of *kth* output variable with respect to the *ith* input feature is introduced in [3]

Sensibility σ_{kl} of the input feature i with respect to output k is calculated by means of:

$$s_{ki} = \sum_j W_{kj}\cdot w_{ji}\cdot h_j\cdot(1-h_j)$$

$$\sigma_{ki} = \frac{|s_{ki}|}{\sqrt{\sum_j s_{kj}^2}}$$

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

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

3.2.4 4th Criterion:

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

Sensibility σ_{kl} of the input feature l with respect to output k is calculated by means of:

$$s_{ki} = \sum_j W_{kj}\cdot w_{ji}\cdot h_j\cdot(1-h_j)$$

$$\sigma_{ki} = \frac{|x_i\cdot s_{ki}|}{\sqrt{\sum_j x_i^2\cdot s_{kj}^2}}$$

Such calculus must be considered for every input pattern. Maximum values of normalised sensibilities are selected. Complete algorithm is described in [3]

4. PERFORMANCE DETECTOR MATRIX

To evaluate detector goodness the performance detector matrix will be filled distinguishing 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.
- **Percent correct detection: CD/T2**
- **Percent correct rejection: CR/T1**
- **Percent false positives: FP/T2**
- **Percent false negatives: FN/T1**
- **Total error: TE=(FP+FN)/(T1+T2)** percentage of erroneous decisions.

		EVENT	
		ABSENT	*PRESENT*
DECISION	*ABSENT*	CR	FN
	PRESENT	FP	CD
	TOTAL	T1=CR+FP	T2=FN+CD
	TOTAL ERRORS=FP+FN		

Table 1: performance matrix

5. 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.

Acoustic parameters used are: "Fo", "To", "Fhi", "Flo", "STD", "PFR", "Fftr", "Fatr", "Jita", "Jitt", "RAP", "PPQ", "sPPQ", "vFo", "ShdB", "Shim", "APQ", "sAPQ", "vAm", "NHR", "VTI", "SPI", "FTRI", "ATRI", "SEG", "PER". A brief description about these parameters can be found in [1]. So, a voice register can be viewed as an n-dimensional vector: $\underline{x}=(x_1, x_2, ..., x_n)$.

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 /ah/ has been studied.

5.1 Pruning 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 has 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 /ah/.

6. RESULTS

We have developed two different classifiers: the first containing feature *gender* in the input pattern, and the second without this feature.

When interpreting results, we will have to keep in mind that the detector has to maximise percentage of correct detection. The greater the correct detection ratio is, the better the detector is. In principle, we are not so worried about correct rejection ratio. This is due to the fact that is better to forecast a disease when none exist, than forecast no disease when it exists.

A MLP with a single hidden layer has been used as classifier. The number of neurones of the input layer is 26 for the first detector, and 27 for the second (26 acoustic parameters plus feature gender). One hidden layer is used. Output layer has a single neuron that will be "1" or "0" activated depending of if we are processing features extracted from normal or impaired voices.

Choosing the net size is a critical problem: the smallest net (short number of hidden units) has to be selected. It has to be done preserving the classification and generalisation capabilities.

6.1 Training and simulation

Data have been divided into two subsets: the first used to train the net (70%), the second to simulate or validate the results.

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

1. Compute sample mean vector $\underline{\mu}=(\mu_1, \mu_2, ..., \mu_n)$ and sample standard deviation vector $\underline{\sigma}=(\sigma_1, \sigma_2, ..., \sigma_n)$ using training sample vectors.
2. Normalise every training sample vector:

$$x'_i=(x_i-\mu_i)/\sigma_i$$

Feature gender is not normalised. It has been "1" or "0" codified (male/female subject). The reason is that gender is a qualitative feature, while remainder features are quantitative.

Sum of mean squared error is controlled as parameter to stop training. Both nets were trained, in the first stage, using the whole feature set we have (26 and 27 features).

Finally, the net was tested using the test subset.

Table 2 shows for the first detector (not including feature gender), the error as function of number of neurones of the hidden layer. The ratio of misclassification obtained is really good.

N° NEUR.	CD/T %	CR/T %	FP/T2 %	FN/T1 %	TE %	SSE	N° epch.
4	96.52	100	0	3.47	2.36	9.9e-009	455
3	96.52	100	0	3.47	2.36	9.7e-009	434
2	96.52	100	0	3.47	2.36	9.6e-009	420
1	96.52	100	0	3.47	2.36	9.7e-009	436

Table 2 NNet Misclassification error (not including gender feature)

Table 3 shows, for the second detector (including feature gender), the error as function of number of neurones of the hidden layer. The ratio of misclassification obtained is more or less the same than in the previous detector. Convergence is similar for twice so, we may think that we can reduce dimensionality of the input vector without loosing generalisation capability.

N° NEUR	CD/T %	CR/T %	FP/T2 %	FN/T1 %	TE %	SSE	N° epch.
4	96.52	100	0	3.47	2.36	9.9e-009	424
3	96.52	100	0	3.47	2.36	9.7e-009	427
2	96.52	100	0	3.47	2.36	9.6e-009	510
1	96.52	100	0	3.47	2.36	9.7e-009	403

Table 3 NNet Misclassification error (including feature gender)

The misclassification ratio is the same for both detectors (including and non including feature gender). So, in principle, we may think that feature gender does not affect classification.

Table 4 shows parameters ordered by their importance according to the four criteria. Meaningful features appear at the bottom of the table.

1st crit.	2nd crit.	3rd crit.	4th crit.
Sex	Flo	Sex	To
To	Fo	To	Fftr
Flo	Fftr	SEG	vFo
sAPQ	FTRI	Fhi	Sex
SEG	sAPQ	sAPQ	sAPQ
Fo	Sex	vFo	FTRI
Fhi	VTI	SPI	SEG
Jita	vFo	Flo	sPPQ
PER	SEG	Fo	VTI
vFo	PER	STD	ShdB
STD	sPPQ	Jita	APQ
Fftr	PFR	PER	SPI
SPI	To	Fftr	ATRI
vAm	Fatr	vAm	PPQ
Fatr	ATRI	sPPQ	Shim
Shim	STD	Shim	vAm
ATRI	Jita	APQ	Fatr
PFR	Fhi	Fatr	STD
APQ	NHR	PFR	NHR
sPPQ	SPI	ATRI	PFR
FTRI	vAm	FTRI	Jitt
Jitt	PPQ	PPQ	RAP
PPQ	ShdB	Jitt	Flo
RAP	Jitt	ShdB	Fo
ShdB	Shim	RAP	Fhi
VTI	RAP	VTI	Jita
NHR	APQ	NHR	PER

Table 4 Features ordered by their importance. Features at the bottom are the most important features

6.5 Feature selection

Classification seems to be easy using Neural Nets. 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 paying attention to the influence of feature gender.

6.5.1. 1st Criterion

Applying techniques described in [4] and neural networks, classifying into normal/impaired voices without loosing performance may be carried out using only two input features: (NHR and VTI).

6.5.2. 2nd Criterion

Applying techniques described in [4] and neural network technology, classifying into normal/impaired voices without loosing performance may be carried out using only two input features: (RAP and APQ).

6.5.3. 3rd Criterion

Applying techniques described in [3], classifying into normal/impaired voices is carried out using only two input features: (VTI and NHR).

6.5.4. 4th Criterion

Applying techniques described in [3], classifying into normal/impaired voices is carried out using only three input features: (PER, Jita and Fhi).

7. CONCLUSIONS

Neural networks technology seems to be a promisable tool to detect meaningful acoustic parameters, allowing us to reduce input space dimension without loosing performance.

The number of input features can be deeply reduced. Meaningful acoustic parameters to diagnose voice diseases depend on the pruning method used.

Anyway, we have to be wise because the database used stores a collection of very significant medical cases. Conclusions have to be tested with a larger database. Also the number of normal voice registers is quite small.

Gender (sex) is not a very important feature in order to detect voice diseases (this is demonstrated due to the fact that feature sex appears at the middle-top of the table 4, where less significant features stay).

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, the next step will be to distinguish among a set of voice disorders. 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 disorder. Anyway, a wider well-labelled database of pathological voices is needed.

9. ACKNOWLEDGEMENTS

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

10. REFERENCES

[1] "Disordered Voice Database", Version 1.03, Kay Elemetrics Corp, 1994

[2] "An Introduction to computing with neural nets" Richard P. Lipmann; IEEE ASSP Magazine April 1987

[3] "On the normalised input sensitivities in neural networks" Ion Ciuca; Studies in Informatics and Control, Vol 5 No 4. 1996. Pp 409-413

[4] "Neural Network Studies. Variable Selection" Igor V. Tetko; Journal of Chemical & Informatics Computing Science. Vol 36 No 4. 1996. Pp 794-803

[5] L.L. Lee, "On two pattern classification and feature selection using neural networks". 1994 IEEE International conference on acoustic, speech and signal processing, pp 617-620

[6] M. Fernandez, C. Hernandez, "Optimal use of a trained neural network for input selection"; Proceedings of the 1999 International Work-Conference on Artificial and Natural Neural Networks (IWANN'99), June 1999, (in press).

[7] M. Fernandez, C. Hernandez, "How to select the inputs for a multilayer feedforward network by using the training set"; Proceedings of the 1999 International Work-Conference on Artificial and Natural Neural Networks (IWANN'99), June 1999, (in press).

[8] "Pitch period determination of aperiodic signals" Hedelin, P., D. Huber.; Proceedings of ICASSP'90, 1990, 361 -364

[9] "Robust hybrid pitch detector". Boyanov B, Ivanov T, Cholet G.; Electronics letters Vol 29 No 22 pp 1924-1926

[10] "Short-term stability measures for the evaluation of voice quality". Feijoo S , Hernandez C; J. of Speech & Hearing Res., vol. 33, pp. 324-334, 1990.

[11] "Novel acoustic measurements of jitter and shimmer characteristics from pathologic voice". Kasuya H , Endo And , Saliu S; Proc. EUROSPEECH'93 Berlin, pp. 1973-1976, 1993

[12] "The validity of using phonatory jitter and shimmer to detect laryngeal pathology." Ludlow C, Bassich C, Connor N, Coulter D, Lee Y; Laryngeal function in phonation and respiration, Brown & Co., Boston, pp. 492-508, 1987

[13] "Method for evaluation of the noise-to-harmonic- component ratios in pathological and normal voices". Yunik M and Boyanov B; Acustica, vol. 70, pp. 89-91, 1990

[14] "Harmonics to noise ratio as an index of the degree of hoarseness" Eiji Yumoto, Wilbur J. Gould; Journal of the acoustic Society of America, June 1982

[15] "A cepstrum-based technique for determining a harmonics to noise ratio in speech signals." De Krom G; J. of Speech & Hearing Research vol. 36, pp. 254-266, 1993

[16] "Normalized noise energy as an acoustic measure to evaluate pathologic voice". Kasuya H, Ogawa Sh, Mashie K, Ebihara S; J. Acoust. Soc. Am. vol. 80, No. 5, pp. 1329-1334, 1986