CLASSIFYING SYNTHESIZED AND REAL VOICE PATTERNS USING NEURAL NETWORKS

Alexandre de Souza Brandão
Universidade Federal Fluminense
PGMEC – Programa de Pós-Graduação em Engenharia Mecânica
Rua Passo da Pátria, 156 - 24120-240 - São Domingos Niterói –RJ – Brasil
brandaoalexandre@ig.com.br

Edson Cataldo
Universidade Federal Fluminense
Departamento de Matemática Aplicada, Centro, Niterói, Brasil
PGMEC – Programa de pós-graduação em Engenharia Mecânica
Programa de pós-graduação em Engenharia de Telecomunicações
Rua Mário Santos Braga, s/ No - 24020-140 – Centro – Niterói – RJ – Brasil
ecataldo@zipmail.com

Fabiana Rodrigues Leta
Universidade Federal Fluminense
Departamento de Engenharia Mecânica
PGMEC – Programa de Pós-Graduação em Engenharia Mecânica
Rua Passo da Pátria, 156 - 24120-240 - São Domingos – Niterói – RJ- Brasil
fabiana@ic.uff.br

Jorge Carlos Lucero
Universidade de Brasília
Departamento de Matemática
DF 70910-900
lucero@mat.unb.br

Abstract. *Artificial neural networks (ANN) are computational models with particular characteristics such as the ability of learning and the ability of classifying and organizing data. One of its applications is the classification of patterns. In this paper, we will use ANN to classify patterns of voice (voiced sounds) using acoustic measures in signals originated from real voices and also from synthesized voices. The first idea was to find a way to classify sinthesized voices generated by the SINTESE software (Cataldo et al, 2005); it uses mechanical models of the vocal cords and the vocal tract to generate synthesized voices. We analyze normal voices and also voices with characteristics of patholog.y*

Keywords*: Neural networks, Mechanical models, Pattern matching, Vowel synthesis, Acoustic measurements.*

1. Introduction

One of the main motivations to study the voice production mechanism is that the human voice is one of the main means of communications.

Voice production starts with a contraction-expansion of the lungs. At this moment, an air pressure difference is created between the lungs and a point in front of the mouth, causing an airflow. This airflow passes through the larynx and, before homogeneous, it is transformed into a series of pulses (glottal signal) of air that reach the mouth and the nasal cavity. The pulses of air are modulated by the tongue, teeth and lips; that is, by the geometry of the vocal tract, to produce what we hear as voice. The glottal signal, however, has important properties which are complex to be reproduced; they are intimately related to the anatomic and physiological characteristics of the larynx.

Studying the system of voice production in a simpler way, we consider four distinct groups: the first one, called *respiration group*, is related to the production of an airflow that starts and ends in the ending of the trachea. In the larynx, we find the organs of the second group, responsible for the production of the glottal signal, which is called the *vocalization group* (the vocal cords belong to this group). The glottal signal is a signal of low intensity, which needs to be amplified and emphasized at determined harmonic components, so that the phonemes can be characterized. This group is called *resonant group.* This phenomenon occurs when the airflow passes through the vocal tract (portion that goes from the larynx up to the mouth). Finally, the pressure waves are radiated when they reach the mouth. This group is called *radiation group.*

In the production of voiced sounds, the airflow coming from the lungs is interrupted by the quasi-periodic vibration of the vocal cords, as illustrated in the Fig. 1.

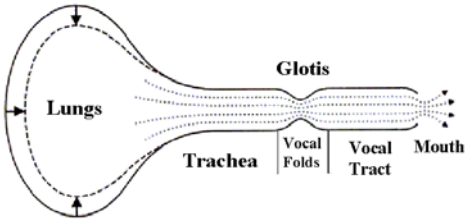


Figure 1. A representation of the voice production system (adapted from Titze (1994a)).

In the last decades, the dynamics of the vocal cords has been extensively studied and some mechanical models were developed. These models differ from each other according to the representation of the vocal cords, viewed as mechanical systems that modulate the airflow.

The SINTESE software generates voiced sounds and we looked for a way to classify these sounds.

In this work, we use signals from real voices (recorded and further digitalized) and from synthesized voices. In the last case, we used mechanical models and the software SINTESE to generate the voiced sounds, presented by Cataldo *et al.* (2005).

2. Acoustic measures in signals of voice

One of the main ways to characterize the signal of voice is through acoustic measures; which can reveal important physiological characteristics (Titze,1994a,1994b,Vieira,1997). Using the acoustic measures, we may build a quantitative description of the voice pattern. Once such descriptions have been built, they may be feed into a trained ANN to classify the voices. The acoustic measures used in this work were obtained using the software PRAAT (www.praat.org), which is a free software used in phonetics analysis.

There are three basic acoustic measures used in signals of voice: *pitch, jitter* and *shimmer*, described in the following paragraphs.

The pulses of air (from the glottal signal) are formed between two instants of the vocal cords closure, whose temporal distance is called a *period* (of time). Although in real voices the waveform is not exactly *periodic*, we still consider them as quasi-periodic, with slow time variations of the amplitude and period. The fundamental frequency of the signal of voice (evaluated as the inverse of the period) is called *pitch* and the variation of the *periods of time* is called *jitter* (Titze, 1994a, 1994b,Vieira, 1997).

The amplitude variation of the pulses of air in the glottal signal, evaluated using the difference between the amplitude of the voice signal of one *period* and the next one, is called *shimmer* (Titze, 1994a, 1994b;Vieira, 1997). So, *jitter* and *shimmer* are acoustic measures of the variation in the period and in the amplitude of the glottal signal, respectively.

There are many kinds of measuring *jitter* and *shimmer*. In the following, we describe those ones applied on this work.

2.1. Jitter measures

***local Jitter*:**

$$Jitter_{(local)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|P_{i+1} - P_i\right|}{\frac{1}{N}\sum_{i=1}^{N} P_i} \tag{1}$$

where N is the number of samples and P_i and P_{i+1} are consecutive periods of time in the voice signal.

Jitter (rap) and Jitter (ppq5):

$$Jitter_{(rap)} = \frac{\frac{1}{N-J+1}\sum_{i=1}^{N-J+1}\left|\frac{1}{J}\left(\sum_{j=1}^{J} P_{i+j-1}\right) - P_{i+0.5(J-1)}\right|}{\frac{1}{N}\sum_{i=1}^{N} P_i} \tag{2}$$

where N is the number of samples; P_i and P_{i+1} are consecutive periods of time in the signal of voice and J is the length of a window of time. In the *Jitter (rap)* measure, the value of J is 3 and in the *Jitter (ppq5)* measure its value is 5.

Jitter (ddp):

$$Jitter_{(ddp)} = \frac{\frac{1}{N-3}\sum_{i=1}^{N-3}\left|\left|P_{i+3}-P_{i+2}\right|-\left|P_{i+1}-P_i\right|\right|}{\frac{1}{N}\sum_{i=1}^{N}P_i} \tag{3}$$

where N is the number of samples and P_i , P_{i+1} , P_{i+2} and P_{i+3} are consecutives periods of time.

2.2. Shimmer measures

Local Shimmer **:**

$$Shimmer_{(local)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N}\left|A_{i+1}-A_i\right|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \tag{4}$$

where N is the number of samples and A_i and A_{i+1} are the amplitudes of the signals of voice corresponding to the consecutives periods of time.

Shimmer (apq3), Shimmer (apq5) e Shimmer (apq11):

$$Shimmer_{(apq3)} = \frac{\frac{1}{N-J+1}\sum_{i=1}^{N-J+1}\left|\frac{1}{J}\left(\sum_{J=1}^{i}A_{i+j-1}\right)-A_{i+0.5(J-i)}\right|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \tag{5}$$

where N is the number of samples; A_i and A_{i+1} are the amplitudes of signals of voice corresponding to the consecutives periods of time and J is the length of a window of time. In the case of *Shimmer apq3* the value of J is 3 and in the case of *Shimmer apq5* its value is 11.

Shimmer(ddp):

$$Shimmer_{(ddp)} = \frac{\frac{1}{N-3}\sum_{i=1}^{N-3}\left|\left|A_{i+3}-A_{i+2}\right|-\left|A_{i+1}-A_i\right|\right|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \tag{6}$$

where N is the number of samples and A_i , A_{i+1} , A_{i+2} and A_{i+3} are the amplitudes of the signals of voice corresponding to the consecutives periods of time.

2.3. ***Pitch***

$$Pitch_{(average)} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}P_i} \tag{7}$$

where N is the number of samples and P_i corresponds to the periods of time.

3. Results from measures in real voices

We used 10 (ten) signals of voice from people with normal voices, 12 (twelve) signals of voice from people with nodules in the vocal cords and 8 (eight) signals of voice from people with oneside paralysis of the vocal cords.

The utterance used was the sustained vowel /e/. Generally, this utterance is preferred because its measures are easier to be identified.

We show in the Tab. 1, 2 and 3 the set of measures obtained. We have also measured the intensity of the signals of voice and the correspondings *pitches*.

Table 1. Measurements in signals of normal voices.

Measurement	SN1	SN2	SN3	SN4	SN5	SN6	SN7	SN8	SN9	SN10
Jitter (local):	0,27%	0,32%	0,45%	0,36%	0,70%	0,27%	0,12%	0,19%	0,19%	0,26%
Jitter (rap):	0,14%	0,18%	0,27%	0,17%	0,39%	0,11%	0,05%	0,09%	0,10%	0,16%
Jitter (ppq5):	0,15%	0,19%	0,39%	0,23%	0,48%	0,15%	0,07%	0,12%	0,12%	0,15%
Jitter (ddp):	0,42%	0,54%	0,81%	0,50%	1,17%	0,34%	0,15%	0,28%	0,29%	0,47%
Shimmer (local):	1,27%	4,48%	1,59%	3,28%	3,27%	3,98%	1,78%	5,15%	4,05%	2,83%
Shimmer (apq3):	0,70%	2,51%	0,82%	1,78%	1,73%	2,21%	0,99%	3,10%	2,41%	1,62%
Shimmer (apq5):	0,79%	2,87%	0,97%	2,08%	1,94%	2,43%	1,13%	3,22%	2,36%	1,74%
Shimmer (apq11):	0,92%	3,11%	1,23%	2,69%	2,48%	2,94%	1,30%	3,57%	2,68%	1,92%
Shimmer (dda):	2,11%	7,53%	2,45%	5,34%	5,20%	6,63%	2,96%	9,29%	7,24%	4,85%
Intensity (dB):	69,08	74,44	62,87	72,73	56,76	70,6	73,18	68,4	64,13	63,32
Pitch (Hz):	188,6	123,11	168,75	101,05	181,51	102,79	162,63	134,11	123,72	199,38

Table 2. Measurements in signals of voices from people with nodules in the vocal cords.

Measurement	SPN1	SPN2	SPN3	SPN4	SPN5	SPN6	SPN7	SPN8	SPN9	SPN10	SPN11	SPN12
Jitter (local):	0,23%	0,78%	0,55%	0,47%	0,32%	0,33%	0,31%	1,28%	0,71%	0,33%	0,57%	0,41%
Jitter (rap):	0,13%	0,47%	0,32%	0,27%	0,19%	0,19%	0,18%	0,72%	0,36%	0,19%	0,34%	0,22%
Jitter (ppq5):	0,13%	0,47%	0,35%	0,34%	0,19%	0,20%	0,17%	0,94%	0,51%	0,21%	0,35%	0,30%
Jitter (ddp):	0,38%	1,41%	0,95%	0,82%	0,57%	0,56%	0,53%	2,15%	1,08%	0,57%	1,01%	0,65%
Shimmer (local):	2,17%	4,31%	6,02%	4,00%	2,95%	4,05%	3,36%	7,60%	7,09%	2,67%	6,33%	3,45%
Shimmer (apq3):	1,16%	2,38%	3,45%	2,07%	1,64%	2,24%	1,97%	4,28%	3,64%	1,51%	3,61%	1,84%
Shimmer (apq5):	1,33%	2,77%	3,76%	2,38%	1,84%	2,51%	1,96%	4,89%	4,73%	1,73%	3,97%	2,03%
Shimmer (apq11):	1,64%	2,90%	4,09%	3,31%	2,26%	2,88%	2,25%	5,63%	5,78%	1,89%	4,54%	2,64%
Shimmer (dda):	3,49%	7,13%	10,34%	6,20%	4,92%	6,73%	5,91%	12,85%	10,93%	4,52%	10,82%	5,53%
Intensity (dB):	68,34	58,83	55,35	61,43	70,08	60,33	69,59	60,55	63,52	72,32	59,82	73,74
Pitch (Hz):	198,05	163,27	217,81	146,69	214,28	173,29	161,96	171,47	180,43	200,12	179,49	207,49

Table 3. Measurements in signal of voice from people with oneside paralysis in the vocal cords.

Measurement	SPP1	SPP2	SPP3	SPP4	SPP5	SPP6	SPP7	SPP8
Jitter (local):	0,60%	0,45%	1,27%	0,47%	3,91%	0,64%	4,49%	2,28%
Jitter (rap):	0,34%	0,26%	0,73%	0,27%	2,33%	0,37%	2,59%	1,38%
Jitter (ppq5):	0,36%	0,27%	0,80%	0,28%	2,65%	0,42%	2,82%	1,61%
Jitter (ddp):	1,01%	0,76%	2,18%	0,81%	6,99%	1,12%	7,76%	4,13%
Shimmer (local):	4,50%	4,64%	7,99%	3,48%	11,02%	4,20%	31,50%	11,53%
Shimmer (apq3):	2,61%	2,59%	4,59%	1,96%	6,33%	2,38%	19,11%	6,61%
Shimmer (apq5):	2,91%	3,05%	4,89%	2,19%	6,74%	2,63%	17,34%	7,34%
Shimmer (apq11):	3,19%	3,16%	5,72%	2,41%	7,86%	2,96%	20,08%	8,21%
Shimmer (dda):	7,82%	7,77%	13,78%	5,89%	18,98%	7,13%	57,32%	19,84%
Intensity (dB):	75,41	63,23	65,81	62,4	68,82	77,95	58,43	64,07
Pitch (Hz):	156,78	140,33	134,84	308,15	151,45	189,88	81,43	172,6

In general, the measures are below the 5% level recommended as approximate superior limit for the reliability of acoustic measures (Titze,1994), which validates their use here.

4. Neural networks

An artificial neural network (ANN) consists of a computational model in layers of processors (called neurons), linked by weighted conexions (Haykin, 2001).

In this work we used a *feed forward network* developed in the MATLAB environment. In ANN`s like those, the data go from the units of input to the units of output directly; that is, whithout feedback.

The Fig. 2 shows the ANN designed for this work.

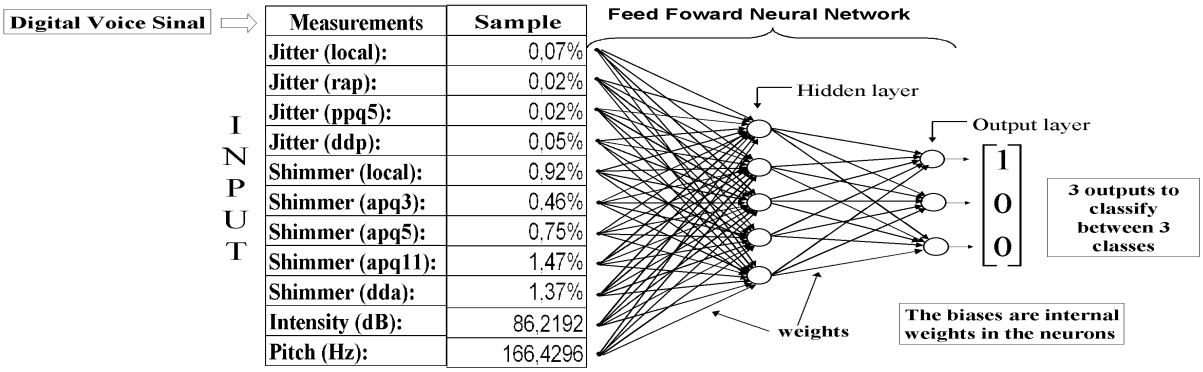


Figure 2. Neural network designed for this work.

The algorithm used for the ANN learning is the backpropagation algorithm (Haykin, 2001): we give a set of initial weights and the input values - the input values are the measures described earlier - given in a vector shape. Each input vector represents the characteristics that we want to distinguish among the samples of the signals of voice. During the process of the ANN training, the correct result, for each input, is showed to the ANN; that is, the weights are fitted using the *Delta rule* (Haykin, 2001). The function of activation needs to be differentiable (because the backpropagation algorithm is based on the derivative of this function) and non-decreasing (because the signal of the derivative cannot change so that the convergence of the algorithm is guaranteed). The function of activation *(f)* used by the input neurons was defined by $f(v) = 1/(1+e^{-av})$, $a>0$, because with this function we can map the nonlinearities. The neurons in the output layer use linear functions of activation.

We trialed the ANN using input data that had not been used for training the network. The idea is that the Euclidian distance between the input vectors is short. So, input vectors that were not used for training the network, but belong to the same pattern, will be fitted between those vectors.

The ANN is trained with vectors S, which components are S_k ; $k = 1,2,3$ given by:

$S_k = 1$ – when the sample belongs to the corresponding class
$S_k = 0$ – when the sample does not belong to the corresponding class C_k

Then, the class is represented by a vector, with three components, described by: $S = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ - normal pattern; $S = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ - nodules pattern and $S = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ - paralysis pattern.

As we do not know, initially, the variability of the set of samples (we use only 30 samples), the number of neurons was defined experimentally. That is, the ANN was trained initially with three neurons in the hidden layer, further with four, five and so on, observing the number of correct responses in relation to the samples left out from the training. We achieved the better index of correct responses using five neurons in the hidden layer.

The number of samples left out from the training was based on the following relation:

$$r = 1 - \frac{\sqrt{2W-1}-1}{2(W-1)} \tag{8}$$

whereas r represents the fraction of the set of samples that will be used in the network trainning and W is the total of weights used in the ANN (Haykin, 2001). The weights used were: Input Weights – 55; Output Weights – 15; Biases – 8; Total – 78. Then, r = 0.9256. So, 92,56% of the samples were used to train the ANN and the other samples were used for testing the ANN. As we had 30 samples, we used 27 for training the ANN and 3 for testing it.

5. Comparison of real voices and synthesized voices samples

5.1. Classifying patterns of real voices

As we have already said, from the 30 samples (presented in the Tab. 1,2 and 3), 27 were used for training the network and 3 were used for testing it. We extracted the following columns: 5 from Tab. 1, 8 from Tab. 2 and 7 from Tab. 3. These columns were chosen because they presented values in some of its components higher than its own corresponding classes. It was only a criterion of choice, but we could simply extracted columns randomly. The Tab. 4 presents the correct classification for values that were not applied on the training.

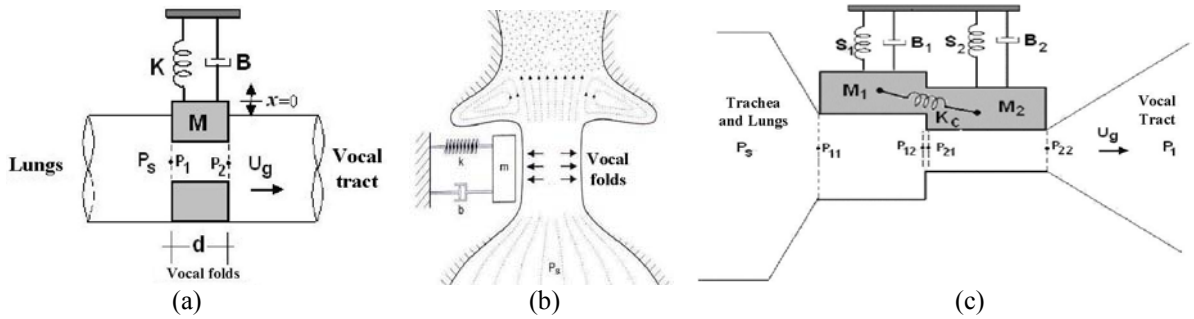
Table 4. The correct classification of values that were not used in the training.

Measurement	SN5	SPN8	SPP7
Jitter (local):	0,7	1,28	4,49
Jitter (rap):	0,39	0,72	2,59
Jitter (ppq5):	0,48	0,94	2,82
Jitter (ddp):	1,17	2,15	7,76
Shimmer (local):	3,27	7,6	31,5
Shimmer (apq3):	1,73	4,28	19,11
Shimmer (apq5):	1,94	4,89	17,34
Shimmer (apq11):	2,48	5,63	20,08
Shimmer (dda):	5,2	12,85	57,32
Intensity (dB):	56,76	60,55	58,43
Pitch (Hz):	181,51	171,47	81,43
RNA			
Output 1	3,846963059	-9,26E-07	-5,58E-07
Output 2	-2,846966271	0,838263978	4,17E-09
Output 3	3,21E-06	0,161736948	1,000000554
Result	**Normal**	**Nodule**	**Paralysis**

Once the network is trained, its results should be near the vectors corresponding to its patterns; that is, each component of an output vector should be approximately 0 or 1, and the decision threshold is 0.5. For example, an output vector corresponding to the normal pattern should have the first component greater than 0.5 and the other two components less than 0.5.

5.2. Patterns of sinthesized voices

We generate sinthesized voices (vowel /e/ sustained) using the software SINTESE, varying the subglottal pressure (Cataldo *et al*., 2005). We considered two mechanical models for the vocal cords, ilustrated in the Fig. 3 and we considered the vocal tract formed by cylindrical tubes concatenated.



(a) (b) (c)

Figure 3. (a) Mechanical model for the vocal cords created by Flangan and Landgraf (1968). (b) Vocal system used (adapted from Titze (1994a). (c) Mechanical model for the vocal cords created by Ishizaka and Flanagan (1972).

The mechanical models discussed above were used for simulating signals of normal voices and with characteristics of oneside paralysis of the vocal cord.

We sinthesized two samples for the synthesized normal voices, one for each model, varying the subglottal pressure. In the Tab. 5, we show the results achieved.

Table 5 – Synthesized signals of voice: SiN1 – FL68 model, SiN2 – IF72 model

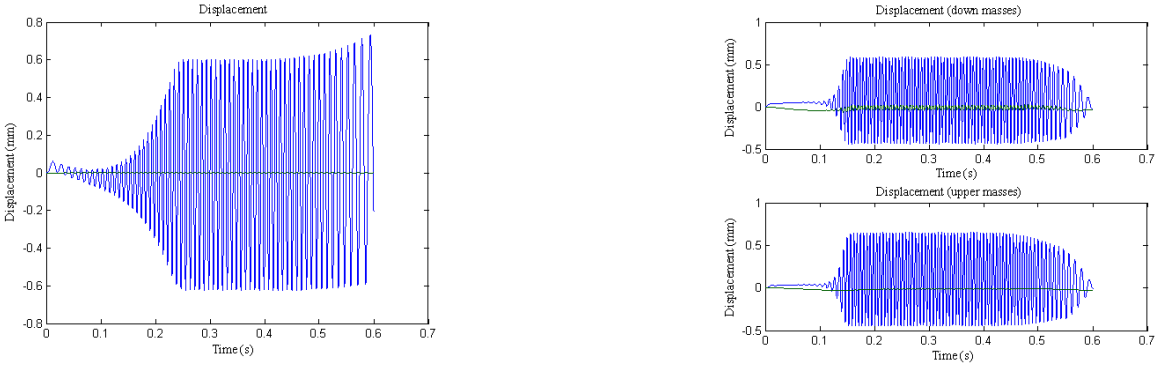
Measurement	SiN1	SiN2
Jitter (local):	0,07%	0,02%
Jitter (rap):	0,02%	0,01%
Jitter (ppq5):	0,02%	0,02%
Jitter (ddp):	0,05%	0,03%
Shimmer (local):	0,92%	0,72%
Shimmer (apq3):	0,46%	0,44%
Shimmer (apq5):	0,75%	0,72%
Shimmer (apq11):	1,47%	1,44%
Shimmer (dda):	1,37%	1,31%
Intensity (dB):	86,2192	86,43
Pitch (Hz):	166,4296	166,95
RNA		
Output 1	1,0000001	1,00E+00
Output 2	-1,94E-06	-3,22E-06
Output 3	9,04E-07	2,19E-06

5.3. Methodoloy applied on the signals of sinthesized voices

The software SINTESE allows the asymmetric case for the vocal cords. Then, we can simulate cases of oneside paralysis modifying the parameters of one vocal cord; we can also simulate other cases of patology. But, in this work, we will consider only the case of oneside paralysis of the vocal cords.

We simulated the paralysis using different cases and testing these cases in the ANN created. We varied the values of the mass or the damping or the stiffness in the mechanical models.

We show in the Fig. 4 the plots corresponding to the displacements of the masses in two cases of paralysis simulation.



(a) (b)

Figure 4. Displacement of the vocal cord, considering paralysis. (a) FL68 model, (b) IF72 model

The Tab. 6 shows the results achieved when classifying the synthesized voices, considering the case of paralysis.

Table 6. Classification for the 2 synthesized voices.

Measurement	e_FL68 paralysis	e_IF72 paralysis

Jitter (local):	1,56	1,31
Jitter (rap):	0,16	0,17
Jitter (ppq5):	0,33	0,25
Jitter (ddp):	0,47	0,52
Shimmer (local):	5,67	3,17
Shimmer (apq3):	0,8	0,56
Shimmer (apq5):	1,82	1,12
Shimmer (apq11):	5,96	2,52
Shimmer (dda):	2,4	1,67
Intensity (dB):	80,83	85,26
Pitch (Hz):	113,74	157,57
RNA		
Output 1	-6,46E-07	-6,54E-07
Output 2	-3,69E-07	-3,69E-07
Output 3	1	1
Result	**Paralysis**	**Paralysis**

6. Conclusions

We used an artificial neural network to classify patterns of real and synthesized voices, considering cases of normal and pathological voices.

The results achieved were satisfactory, once the designed ANN could distinguish signals of voices from different patterns. We used an ANN with only one hidden layer and we achieved good results; ROSA (1998) had used one with more than one hidden layer to achieve the results.

We could also test synthesized signals of voices, created by the SINTESE software, using mechanical models.

The ANN exactness concerning its classification task depends upon the variability of the sample values used for its training, as well as the number of samples, which should be sufficiently distributed. This justifies how successful the designed ANN was, although we have used a small number of samples. More sophisticated methods of error analysis using cross-validation techniques may also be applied; however, our intention here has been to introduce our algorithm and show its potencial, leaving deeper analyses for future efforts.

We also should notice that the amount of different measures in the input vector increase the ANN classifying capability. Although there are methods to select the characteristics that should be used to eliminate some components in the input vector, we did not apply them here; that is, we considered all of the information (measures) we had.

Moreover, we developed a tool that enables us to classify synthesized voices achieved through the SINTESE software.

7. References

Cataldo, E., Leta, F.R., Lucero, J., Nicolato, L., "Synthesis of voiced sounds using low-dimensional models of the vocal cords and time-varying subglottal pressure", Mechanics Research Communications, 2005 (to appear).

Flanagan, K. and Landgraf, L.,"Self-oscillating source for vocal-tract synthesizers". IEEE Trans. On Audio and Eletroacoustics, Vol. 16., pp. 57-64, 1968.

Haykin, S. "Neural Networks: Principles and Practice", 2nd edition,Bookman, 2001.

Ishizaka, K. and Flanagan, J. "Synthesis of voiced sounds from two-mass model of the vocal cords", Bell Syst. Tech. Journal, Vol. 51, pp. 1233-1268, 1972.

Rosa, M.O., "Análise Acústica da Voz para Pré-diagnóstico de Patologias da Laringe". 219p. Master degree monography. Escola de Engenharia de São Carlos, USP, São Paulo, 1998.

Titze, I. R., "Principles of voice production", PrenticeHall, Englewood Cliff New Jersey, 1994.

Titze, I. R., "Workshop on Acoustic Voice Analysis. Summary Statement", National Center for Voice and Speech, Iowa City, 1994.

Vieira, M. N., "Automated Measures of Dysphonias and the Phonatory Effects of Asymmetries in the Posterior Larynx". PhD Thesis. Edinburg University, 1997.

8. Responsibility notice

The authors are the only responsible for the printed material included in this paper.