International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 18 (2015) pp 39670-39674

Automatic Determination of Pathological Voice Transformation Coefficients

H. Belgacem,

Signal Processing Laboratory, Sciences Faculty of Tunis El Manar – Tunisia m.haythem@yahoo.fr

L. Salhi,

Signal Processing Laboratory, Sciences Faculty of Tunis El Manar – Tunisia lotfi.salhi@laposte.net

A.Cherif

Signal Processing Laboratory, Sciences Faculty of Tunis El Manar – Tunisia adnane.cher@fst.rnu.tn

Abstract
Different methods in speech synthesis can be implemented to effect transformation such as WSOLA, TDHS, TD-PSOLA…only MBOROLA method is automatic. Methods based on the principle of synchronization with the fundamental has been used to perform temporal or frequency changes of a speech signal. One of the biggest challenges in vocal transformation is the selection of modified parameters that will make a successful speech resynthesis. The goal of this study is to improve the performance of an automatic system of determination of the pathological voice transformation coefficients with a multilayer neural network. First, we will use the classic analysis methods usually used in the domain of the automatic speech recognition and applied here in a pathological context. Then we proposed a dynamic system with wavelets packet analysis (WPA); to extract the distinctive parameters, contrary to a technique of analysis with a stationary window.
Our own experiments and results of the new proposed method are described in this paper. These experiments were among the earliest attempts to achieve full automation of the method TD-PSOLA.

Keywords: Wavelets packet analysis, transformation coefficients, Pathological voice, Multilayer neural networks.

Introduction

In this study, we propose a system, which will allow us to put in correspondence a pathological voice with the adequate coefficients of pitch-scale (beta) and time-scale (alpha). These coefficients will be later used by the method TD-PSOLA to get a better intelligibility of the pathological voice has to needed to be transformed.
Given the complexity of the function desired and its nonlinearity, methods based on mathematical modeling have proved unsatisfactory. Hence, we will develop a system based on multilayer neural network (MNN). Practically, neural networks are non-linear statistical data modeling tools. They can be used in order to model complicated relationships between inputs and outputs. An MNN can learn by example. Once a neural network is trained on the data, it will able to make predictions by detecting similar patterns in future data. The choices of the input vector, the output vector, the number of hidden layers, the number of hidden neurons, the transfer functions, give a various combinations and can make the difference between the degrees of success of the proposed system [1]. For the MNN it is practically impossible to specify in advance a satisfactory architecture, it is only several experimental trials that we can solve this problem [2]. So far, as there is no exact formula, we had to design several models of networks and choose among them the most successful model.

Proposed System

In the input of a neural network, the pathological signal is presented as a vector of N parameters. We test the performances of several systems whose parameters of the input vector. The number of input neurons varies between 7 and 45 for the classic analysis parameters. For the dynamic system with wavelets packet analysis, it provides a two-dimensional pattern of wavelet coefficients. The middle values content of wavelet coefficients at various level of scaling is used to formulate a feature vector of speech sample. The outputs vectors will consist of a vector of two elements: pitch-scale (beta) and time-scale (alpha) coefficients (Figure 1).
Given that, we have the wished output, training will be supervised. In supervised learning, both the input and the expected output patterns (targets) are provided (training pairs). The network processes these inputs and compares the resulting outputs to the desired ones. [2]

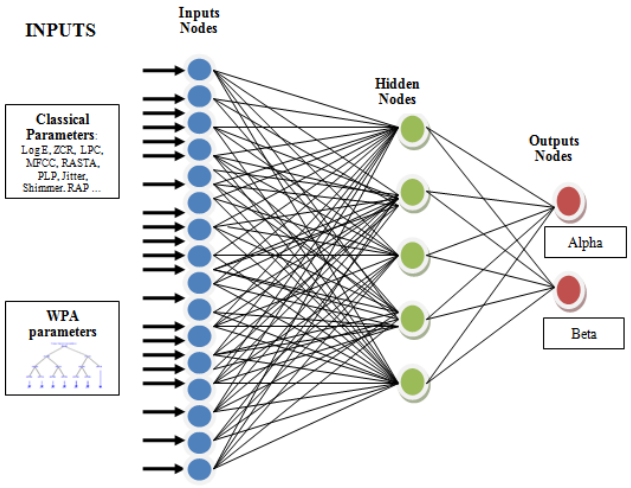

Fig.1. Architecture of the multi-layer perceptron

i. Wavelet transforms analysis

The wavelet transform can be viewed as transforming the signal from the time domain to the wavelet domain. This new domain contains more complicated basis functions called wavelets, mother wavelets or analysing wavelets.

A wavelet prototype function at a scale s and a spatial displacement u is defined as (1) : [3]

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi(\frac{t-u}{s}) \quad (u \in IR,\ s \in IR_+^*) \tag{1}$$

The Continous Wavelet Transform (CWT) is an excellent tool for mapping the changing properties of non-stationary signals. The CWT is also an ideal tool for determining whether or not a signal is stationary in a global sense. When a signal is judged non-stationary, the CWT can be used to identify stationary sections of the data stream. Specifically, a Wavelet Transform function f(t) Є L2(R) (defines space of square integrable functions) can be represented as (2):

$$W(f)(u,s) = \int_{-\infty}^{+\infty} f(t)\psi_{u,s}^*(t)dt = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\psi^*(\frac{t-u}{s})dt \tag{2}$$

The factor of scale includes an aspect transfer at a time in the time brought by the term u, but also an aspect dilation at a time in time and in amplitude brought by the terms s and s et √ s.

ii. The wavelet transforms coefficients

We use a wavelet filter bank to extract a three, five and seven wavelet coefficient then we calculate their corresponding energy in the case of energy coefficients. The Training curve with three coefficients is given in Figure 2-a, with five coefficients in Figure 2-b, and with seven coefficients in Figure 2-c.

The obtained training curve with three coefficients is given in Figure 3-a, with five coefficients in Figure 3-b, and with seven coefficients in Figure 3-c.

iii. Evaluation techniques for the networks performance: Correct identification rate (CIR)

The network performance resulted in the correct identification rate (CIR) of voice. After the learning phase, a network must be able to provide the correct values of pair of coefficients alpha, beta necessary to better voice transformation. Recall that the correct identification rate of a network is defined by (3):

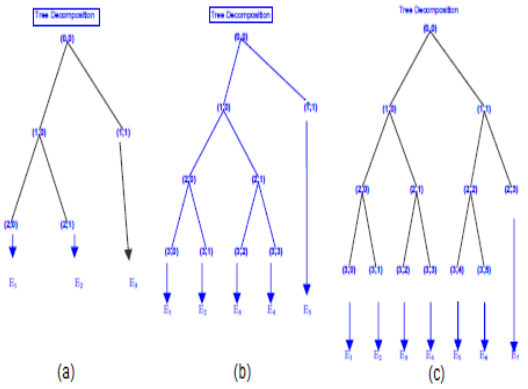

Figure 2. a- Training curve with three coefficients. - bTraining curve with five coefficients- c-Training curve with seven coefficients.

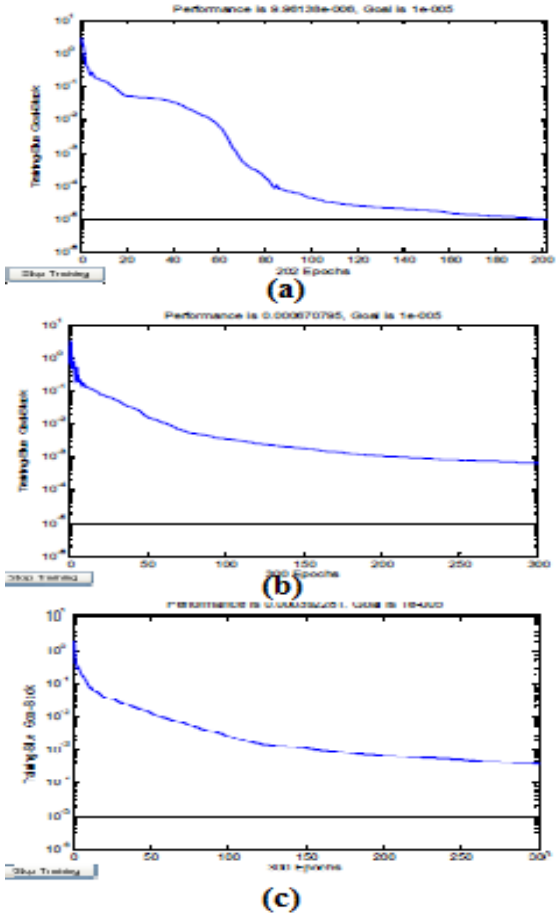

Figure 3. Training with a-3 coefficients. b-5 coefficients c-7 coefficients

$$CIR = \frac{\text{Number of pairs correctly identified}}{\text{Total number of pairs}} \% \quad (3)$$

If an estimated value, it is more than ± 5% of the reference value, then it is considered false.

Network architecture

The battery of sensor nodes in the wireless sensor network become rechargeable, the network design changes fundamentally. Designing the network that reduces the cost for power recharging is an important one.

A. Training and transfer functions

The training has always been the major problem that has defined the limits of the researches in connectionism. For our network training, we use the algorithm back propagation of gradient which turns out the most successful on this structure [3]. The initial values of the weights are randomly allocated. Thus the problem of a weighty symmetry is resolved.

In our application we tended to use transfer functions that are implemented with their derivatives in MATLAB.

Three transfer functions can be used: a linear function ('purelin') and two sïgmoïdes functions, the first one is at positive and negative outputs ('tansig'), the other one is only at positive outputs ('logsig'). We chose to use the logsig function which provides positive outputs.

B. Choice of number of neurons in hidden layers

In order to unnecessarily increase the network complexity, we choose a single hidden layer. The network using the RASTA coefficients with order 7, Jitter, Shimmer and RAP is tested on 4, 5, 8, 10, 15, 24, 35 and 50 neurons in hidden layer. All tests lead us to conclude that the network size beyond 10 neurons did not affect network performance. The Figure4 presents the variation of the CIR in function of the number of the neurons in hidden layer for the vectors V4, V11, V20, and V27, which representing the four configuration groups.

Network convergence is no longer linear; however, we could stop the algorithm much more early while maintaining more acceptable performances. This latter finding leads us to consider definitively 10 neurons in hidden layer network.

For a number of hidden layers equal to 10, we have to vary the number of epochs. We not noticed a sensitive improvement in the system performances (Figure5). In order to optimize the system, we opted for a number of epochs equal to 400.

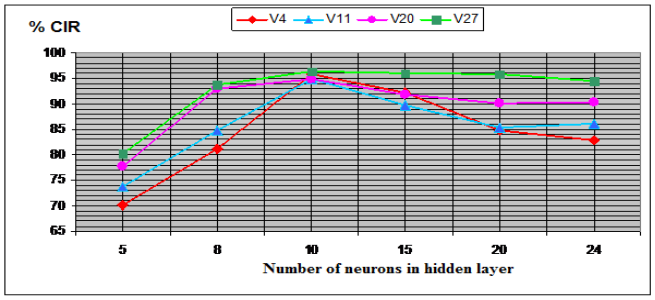

Figure4. Percentage of CIR based on the number of neurons in hidden layer

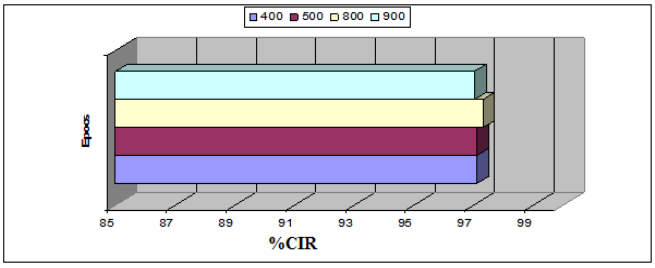

Figure5. Percentage of correct rate based on epochs

C. Shaping the network inputs

To describe the spectral envelope of the speech signal, several presentation models have been used: the coefficients of linear prediction (LPC), Mel scale cepstral coefficients (MFCC), perceptual linear predictive (PLP), the relative spectral transform with the perceptual linear prediction (RASTA-PLP). [6][7]

They are usually the emotional state, or neurological disease of the patient. Disturbances in the vocal tract, irregularities pitch, intonations, are phenomena that are often found in pathological speech processing. Jitter (local or RAP) and Shimmer's equations allow to identify and characterize with certainty the percentage of variation of a pathological signal.[5]

The classic input vectors must be trained the signal parameters coefficients associated with Jitter, Shimmer and RAP.

We present the different input vectors formed by combinations of the parameters mentioned above. These vectors are used for training and validation of the system. We choose among them the one that will be the most adapted technique for the pathological speech processing.

We propose for the classic parameters 4 groups and 30 vectors, and 180 networks from R1 to R180 (see Table 1)

For the dynamic system with wavelets packet analysis (WPA); we propose one group with 4 vectors: V31 based on norm X, V32 the energy E, V33 the Entropy H and V34 the theory dimension. For each vector we have 3 under-vectors with N=3, 5 and 7 coefficients. The number of neurons is 5, 10, 15, 20, and 24, so we get 72 networks: from R181 to R252. (See Table 2).

TABLE.1. The different classic input vectors

Group	1	2	3	4
Parameters	MFCC, Jitter, RAP, Shimmer	LPC, Jitter, RAP, Shimmer	PLP, Jitter, RAP, Shimmer	RASTA PLP, Jitter, RAP, Shimmer
N	from 8+ 3 to 14+3	from 8+3 to 14+3	from 4+3 to 11+3	from 4+3 to 11+3
Order 4	-	-	V15	V23
Order 5	-	-	V16	V24
Order 6	-	-	V17	V25
Order 7	-	-	V18	V26
Order 8	V1	V8	V19	V27
Order 9	V2	V9	V20	V28
Order 10	V3	V10	V21	V29

Order 11	V4	V11	V22	V30
Order 12	V5	V12	-	
Order 13	V6	V13	-	-
Order 14	V7	V14	-	-

TABLE.2. the WPA input vectors

vector	**V31**	**V32**	**V33**	**V34**
	norm X	energy E	Entropy H	Theoretical D
N	**3 5 7**	**3 5 7**	**3 5 7**	**3 5 7**
M=5	R_{181} R_{182} R_{183}	R_{199} R_{200} R_{201}	R_{217} R_{218} R_{219}	R_{235} R_{236} R_{237}
M=8	R_{184} R_{185} R_{186}	R_{202} R_{203} R_{204}	R_{220} R_{221} R_{222}	R_{238} R_{239} R_{240}
M=10	R_{187} R_{188} R_{189}	R_{205} R_{206} R_{207}	R_{223} R_{224} R_{225}	R_{241} R_{242} R_{243}
M=15	R_{190} R_{191} R_{192}	R_{208} R_{209} R_{210}	R_{226} R_{227} R_{228}	R_{244} R_{245} R_{246}
M=20	R_{193} R_{194} R_{195}	R_{211} R_{212} R_{213}	R_{229} R_{230} R_{231}	R_{247} R_{248} R_{249}
M=24	R_{196} R_{197} R_{198}	R_{214} R_{215} R_{216}	R_{232} R_{233} R_{234}	R_{250} R_{251} R_{252}

Results

After having presented in details the proposed algorithm for the automatic system of determination of the pathological voice transformation coefficients, we present the results of implementation of this algorithm.

First, we have built four systems with classics parameters (groups 1, 2, 3, 4) with 10 neurons in hidden layer. For each system, the network input vector is composed of the signal classic parameters, associated with Jitter, Shimmer and RAP coefficients. The total number of proposed systems configurations is 180. Each configuration undergoes a learning phase followed by a testing and validation phase. We conducted learning network using 180 sound extracts from a database of pathological sounds of OSEE research unit. We use the cross-validation protocol. 75% of the corpus is used for learning while 25% is used for the validation of the system. Networks performances resulted in the correct identification rate (CIR) of the voices. After the learning phase, a network should be able to provide the correct values of the request coefficients alpha, beta to a better voice transformation.

For each proposed network, we provide the learning performance shown by the correct identification rate (CIR). In TABLE3 we select for each group, the best architectures associated with each of the input vectors based on their CIR. The best identification and learning rate is obtained with V26 vector that is the vector formed by the RASTA-PLP coefficients.

Second, we compare the performance of group 4 opposite to two new groups, one is based on the RASTAPLP parameters and its first derivatives (Group 5), the other on the RASTAPLP parameters and its first derivatives and its derivatives seconds (Group 6). The results are shown in TABLE4.

TABLE.3. CIR percentage for groups 1, 2, 3, 4

Group	**1**	**2**	**3**	**4**
Parameters	MFCC, Jitter, RAP, Shimmer	LPC, Jitter, RAP, Shimmer	PLP, Jitter, RAP, Shimmer	RASTA PLP, Jitter, RAP, Shimmer
Network reference	V4 R21	V11 R63	V22 R129	V26 R153
%CIR	95,80	94,90	95,00	97,10

TABLE.4. Percentage of CIR for groups 4, 5, 6

Order	**Group 4**	**Group 5**	**Group 6**
4	93,50	93,83	94,22
5	94,21	94,92	95,10
6	96,30	97,12	97,70
7	97,10	97,80	98,20
8	96,35	96,81	97,33
9	96,15	96,91	97,42
10	96,10	96,40	96,90
11	96,00	96,80	97,00

Finally to improve the performance of the proposed automatic system of determination of the pathological voice transformation coefficients, we have built a system with wavelets packet analysis parameters, we shows that the best identification and learning rate is obtained with V34 R243 that is the vector formed by the 7 theory dimension coefficients with a CIR = 98,92%. (Figure8)

Discussions

We have tested 7 systems of neural network; we can notice that the percentage went from 74,30% to 98,92% by using the final system.

For a good correspondence between efficiency and duration of network response, we have opted for a system of 10 hidden neurons layer, and a number of epochs equal to 400.

Figure6 shows that the percentage of a successfully method goes from 90,90 % to 97,10 % using the RASTAPLP orders 7.

Figure7 shows that the percentage of a successfully method goes from 97,10 % to 98,20 %, in association with the RASTAPLP, its first and second derivatives, percentage of a successfully method by using the WPA parameters is 98,92%.

The method of prediction linear perceptive (PLP) consists on packaging a critic bands of specter in short term curried out with a correction of intensity. The signal amplitude is also pressed. Finally, the study of linear prediction RASTA is followed.

For the classic parameters, the RASTA-PLP it is based on auditory models, which provides a better representation of acoustic vectors to recognize. The introduction of other variables in the analysis phase such as Jitter, Shimmer and RAP as relevant parameters greatly increases the success rate method.

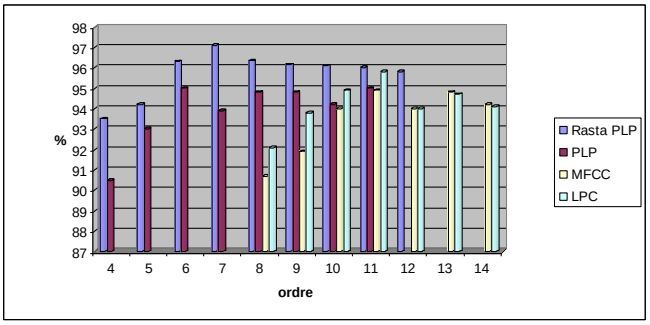

Figure 6. Percentage of correct rate for LPC, MFCC, PLP and RASTAPLP

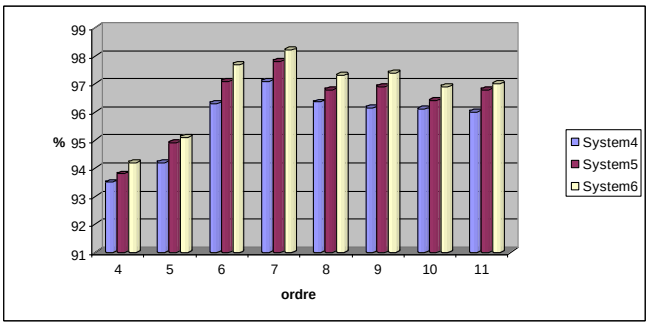

Figure 7. Percentage of correct rate for RASTA-PLP, ΔRASTA-PLP, Δ2 RASTA-PLP

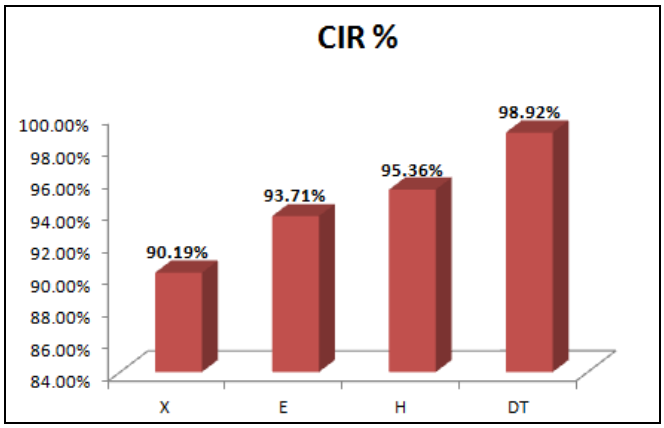

Figure 8. The best CIR for each group of WP vectors

The WPA methods with the seven theory dimension coefficients give the better results compared to the all other methods such as WPA with energy and WPA with entropy coefficients, and the RASTA-PLP analysis.

In fact, this parameter is a combination of two parameters "energy" and "entropy". It is considered as criteria, which measures the energy concentration of the signal decomposition on the orthogonal wavelet basis.

Conclusion

This study aimed at determining the best system to automatically the transformation of pathological voice coefficients using neural network for an important method commonly used in speech synthesis: TD-PSOLA.

We have discussed the choice signal classic parameters and a feature vector based on wavelet coefficients, the number of their coefficients, and the number of hidden layers and the transfer functions to be applied by different neurones to consider the best results.

The wavelet technique employs seven theory dimension coefficients is better than all other classic methods such as the RASTAPLP analysis.

The proposed system of determining transformation coefficients during the synthesis step automates the conversion task has proved to be an excellent way to made automatically speech transformation.

A neural network has the ability to detect any causal connection between the inputs vectors and outputs coefficients, even when a particular input has never been seen before. When the input data is pathologic, this ability permits excellent interpolation capabilities.

References

[1] M.A.W Saduf, (2013) " Comparative study of back propagation learning algorithms for neural networks. " Int J. Adv. Res. Comp. Sci. Softw. Eng. 3(12), 1151–1156

[2] K Sreenivasa RAO (2011), "Role of neural network models for developing speech systems" Sadhana October 2011, Volume 36, Issue 5, pp 783-836.

[3] V. Majidnezhad, (2015) " A novel hybrid of genetic algorithm and ANN for developing a high efficient method for vocal fold pathology diagnosis. " EURASIP Journal on Audio, Speech, and Music Processing (2015) 2015:3 DOI 10.1186/s13636-014-0046 1

[4] I Codello, W. Kuniszyk-JóÎkowiak (2007), "Wavelet analysis of speech signal", Annales UMCS Informatica AI 6 : pp 103-115

[5] L. Salhi, M. Talbi, A. Cherif,(2011)"Performance of wavelet analysis and neural networks for pathological voices identification," International Journal of Electronics, vol. 98, no. 7–9, pp. 1129–1140.

[6] M. Brockmann, C. Storck, M. Drinnan, (2008).Voice loudness and gender effects on jitter and shimmer in healthy adults. J Speech Lang Hear Res., 51 (5):1152–1160.

[7] T.Ratanpara, N. Patel (2015) "Singer identification using perceptual features and cepstral coefficients of an audio signal from Indian video songs. "EURASIP Journal on Audio, Speech, and Music Processing 2015:16

[8] H. Hermansky, N Morgan (1994), "RASTA processing of speech." Speech and Audio Processing, IEEE Transactions on, 2(4):578-589

[9] Y. Stylianou, (2010) Harmonic plus Noise Models for Speech combined with Statistical Methods for Speech and Speaker Modification. Thèse de doctorat, Telecom Paris.