New Cues in Low-Frequency of Speech for Automatic Detection of Parkinson's Disease

E.A. Belalcazar-Bolaños[1], J.R. Orozco-Arroyave[1,3], J.F. Vargas-Bonilla[1], J.D. Arias-Londoño[1], C.G. Castellanos-Domínguez[2], and Elmar Nöth[3]

[1] Universidad de Antioquia, Medellín, Colombia
[2] Universidad Nacional de Colombia, Manizales, Colombia
[3] Universität Erlangen-Nürnberg, Erlangen, Germany

Abstract. In this paper, the analysis of low-frequency zone of the speech signals from the five Spanish vowels, by means of the Teager energy operator (TEO) and the modified group delay functions (MGDF) is proposed for the automatic detection of Parkinson's disease.

According to our findings, different implementations of the TEO are suitable for tackling the problem of the automatic detection of Parkinson's disease. Additionally, the application of MGDF for improving the resolution of the speech spectrum in the low frequency zone is able for enhancing differences exhibited between the first two formants from speech of people with Parkinson's disease and healthy controls.

The best results are obtained for vowel /e/, where accuracy rates of up to 92% are achieved. This is an interesting result specially if it is considered that there are muscles that are involved in the production of the vowel /e/ but not in other vowels.

Keywords: Group Delay Functions, Parkinson's disease, Teager Energy Operator.

1 Introduction

The Parkinson's disease (PD) is characterized by the loss of dopaminergic neurons in the mid brain and its main symptoms of PD are tremor, rigidity and other movement disorders. It is demonstrated that about 90% of the people with Parkinson's disease (PPD) also develop speech impairments [1], however only from 3% to 4% of the patients receive speech therapy [2].

Acoustic studies further indicate reduction of the sound pressure level, reduced pitch variability, phonatory instability, increased noise and cycle to cycle variability during phonation of PPD [3]. Movement velocities and displacements of the velum also tend to be reduced in PPD, and acoustic studies suggest increased nasal airflow for speech in about 70% of PPD [4], [5], such increment is characterized for generating alterations in the speech spectrum [6]. Additionally, there are evidences about the alterations in the low frequency zone of the speech spectrum and the excess of nasal airflow during phonations of PPD [6], however, few works have taken advantage of this fact. In [7] the authors divide

J.M. Ferrández Vicente et al. (Eds.): IWINAC 2013, Part I, LNCS 7930, pp. 283–292, 2013.

the speech spectrum into its high and low zones for finding the voice low tone to high tone ration (VLHR) as the quotient between the low frequency power and high frequency power. The VLHR is used for the automatic evaluation of nasalization in speech signals. According to their results the VLHR is a potential quantitative index of hypernasal speech and can be applied in either basic or clinical studies. Likewise, in [8] the authors evaluate the performance of a speech recognition system based on modified group delay functions (MGDF), applied over the low frequency zones in the speech spectrum, for the automatic detection of dysarthria in continuous speech. According to their results, the application of GDF is not suitable for the automatic detection of dysarthric speech signals in continuous speech.

Additionally, the speech impairments in PPD are related to the vocal fold bowing and incomplete vocal fold closure [9], besides nonlinear behavior of vocal fold vibration has already been demonstrated in [10]. Considering these evidences, it is possible to include nonlinear energy operators for characterizing speech from PPD. In [11] the authors use the Teager energy operator (TEO) [12] for calculating signal to noise ratio measures in speech signals from PPD, however such operator has not been directly used as a measure of the speech for the automatic detection of PD.

With the aim of include the analysis of the low frequency zone of the speech spectrum and the nonlinear behavior of the vocal folds vibration, in this work the use of MGDF and TEO is proposed as a new method for the automatic classification of speech signals from PPD and HC.

The rest of the paper is organized as follows. In section 2, a methodology proposed is presented, section 3 provides details about the experimental framework with some experimental details about the followed methodology for classification and error estimation, section 4 show the graphics and tables with success classification rates. Finally, in section 5, some conclusions are provided.

2 Metodology

Figure 1 shows a schematic of the methodology used in this work. Every stage of the process will be explained in the following.

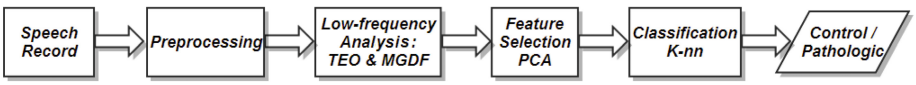

Fig. 1. Methodology

Preprocessing and low frequency analysis. Speech recordings are preprocessed by means of a short temporal analysis using windows of 40ms length with an overlap of 20ms. After, each frame is characterized using a low frequency analysis that is performed through the implementation of TEO and MGDF.

The TEO is a nonlinear operator that is used for taking advantage of the existence of multiple components on a signal $x(n)$ and it is given by [12]:

$$\Psi[x(n)] = x(n)^2 - x(n+1)x(n-1) \tag{1}$$

In the case of nasalized voice signals it is demonstrated that the speech spectrum has multiple components due to the presence of nasal formants and anti formants [6]. The voice signal $x(n)$ can be expressed as the sum of two uncorrelated components $s(n)$ and $g(n)$, when the definition 1 is applied over a multi-component signal $x(n) = s(n) + g(n)$ it is obtained the following expression [13]:

$$\Psi[x(n)] = \Psi[s(n)] + \Psi[g(n)] + \Psi_{cross}[s(n), g(n)]\Psi_{cross}[g(n), s(n)]$$

where $\Psi_{cross}[s(n), g(n)] = g(n)s(n) - g(n+1)s(n-1)$. Note that TEO does not obey the superposition principle because additional terms appear when it is applied over a multi-component signal. Considering this fact, it is possible to apply different measurements that reflect differences in the estimation of the TEO for multi-component signals and for single-component signals.

There are other two different ways for the estimation of the TEO, one is based on the fast Fourier transform (FFT) and defines the TEO as follows [14]:

$$\gamma_n = \left[\sum_{i=1}^{\Omega} i^2 S_n(i)\right]^{\frac{1}{2}}$$

where $S_n(i)$ is the power spectral density from the n speech frame, calculated using the FFT, and i is the frequency value in the discrete domain.

Another estimation is presented in [15] as the generalized version of the TEO and the expression is as follows:

$$\Psi[x(n)] = x(n)^{\frac{2}{m}} - [x(n-M)x(n+M)]^{\frac{1}{m}} \tag{2}$$

Where M and $m \in \mathbb{Z}$. Note that the expression 2 allows the variation of exponents (m) and delays (M).

The different versions of the TEO exposed above can be used for taking advantage of the existence of additional formants in the voice signals from PPD due to the excess of nasalization [6]. One healthy voice only has vocal formants, thus its spectrum can be expressed as $S_{HC} = \sum F(\omega)$, where $F(\omega)$ represents the oral formants. A speech recording from PPD, with excess of nasalization, has oral formants, anti-formants and nasal formants, thus its spectrum can be represented by $S_{PPD} = \sum F(\omega) - \sum AF(\omega) + \sum NF(\omega)$, where $AF(\omega)$ represents the anti-formants and $NF(\omega)$ the nasal formants. When a healthy voice is filtered by means of a low pass filter, it is possible to extract its first formant F_1, however, if a speech signal with excess of nasalization is filtered, the result will contain F_1 along with two additional components due to the existence of anti-formants and nasal formants. If instead a bandpass filter is used (with cutoff bands around F_1), both the result for healthy and nasalized voices will contain only F_1. The differences between both bandpass and lowpass filtered signals can

be exploited for the automatic detection of excess nasalization in speech from PPD by means of the implementation of different measures oriented to detect such differences.

The first measurement is Correlation Teager Energy Operator ($CTEO$) [13]. It is calculated as the correlation between the TEO estimated for the low-passed and band-passed filtered signals, the result of this measure is expected to be high for the case of healthy voices (both filtered signals contained only F_1), but if the correlation is calculated for pathological voices, the result is expected to be low due to the existence of additional nasal formants and anti-formants in the low passed filtered signal.

The other three measurements are the Euclidean distance (ED), the logarithmic distance (LD) and the area under the TEO estimates (IA). All of these measurements are calculated between the TEO estimated for the low-passed and band-passed filtered signals. In all cases, it is expected to obtain higher values of the measures in the case of nasalized voices due to the existence of additional components that introduce differences between both estimations of the TEO.

On the other hand, the MGDF are applied to improve the resolution of the low frequency zone in the speech spectrum. The group delay function is defined as $\tau(\omega) = -\frac{\partial\theta(\omega)}{\partial\omega}$, and for discrete time signals it can be computed as [16]:

$$\tau(K) = \frac{X_R.Y_R + X_I.Y_I}{|X(K)|^2} \tag{3}$$

where $X(K)$ and $Y(K)$ are the N-point discrete Fourier transform (DFT) of the sequences $x(n)$ and $nx(n)$, respectively. The subscripts R and I denote the real and imaginary parts, respectively.

To reduce the spiky nature of the group delay function, which is due to the pitch peaks, noise, and windowing effects , the original function is modified as [17].

$$\tau(K) = \mathbf{sign}\left|\frac{X_R.Y_R + X_I.Y_I}{S(K)^{2\gamma}}\right|^{\alpha}$$

where $S(K)^2$ is a cepstrally smoothed version of $|X(K)|^2$ and the **sign** is the original sign of the group delay function given in 3. To derive the a smoothed group delay spectrum, the values of α and γ should be less than 1 [17].

In the process for the improvement of the signal's spectrum resolution the voice signal is low-pass filtered with a cutoff frequency of $800Hz$, then the MGDF is calculated for the filtered spectrum and finally the first two formants are estimated. The first formant (F_1) is found below the $500Hz$ and the second formant (F_2) is between $500HZ$ and $800Hz$. For the case of speech signals with excess of nasalization, the amplitude of F_1 will be lower than in the case of healthy voices, thus the ratio between the amplitudes of F_1 and F_2, called $\tau(F_1)$ and $\tau(F_2)$, can be used as an index of the presence of nasalization in the speech signal [17]. Such index is called group delay function-based acoustic measure ($GDAM$) and is defined as:

$$GDAM = |\tau(F_1)| \,/\, |\tau(F_2)|$$

Automatic Features Selection and Classification. The selection of features is addressed through the application of principal components analysis (PCA). It is a statistical technique applied here to find out a low-dimensional representation of the original feature space, searching for directions with greater variance to project the data. Although, PCA is commonly used as a feature extraction method, it can be useful to properly select a relevant subset of original features that better represent the studied process [18]. In this sense, given a set of features $(\boldsymbol{\xi}_k : k = 1, \ldots, p)$ corresponding to each column of the input data matrix $\mathbf{X}$, the relevance of each $\boldsymbol{\xi}_k$ can be analyzed for finding the resulting subspace $\mathbf{Y}$. More precisely, relevance of $\boldsymbol{\xi}_k$ can be identified looking at $\boldsymbol{\rho} = \begin{bmatrix} \rho_1 \ \rho_2 \cdots \ \rho_p \end{bmatrix}^\top$, where $\boldsymbol{\rho}$ is defined as $\boldsymbol{\rho} = \sum_{j=1}^{m} |\lambda_j \mathbf{v}_j|$. ($\lambda_j$ and $\mathbf{v}_j$ are the eigenvalues and eigenvectors of the initial matrix, respectively). Therefore, the main assumption is that the largest values of ρ_k point out to the best input attributes, since they exhibit higher overall correlations with principal components.

The decision of whether a voice recording is from PPD or HC is taken with a K nearest neighbor (K-nn) classifier. Considering that the aim of this work is to analyze the discrimination capability of the described features, this classifier is chosen because of its simplicity allowing us to focuses our analysis to the features and not to the classifier.

3 Experimental Framework

Database. The data for this study consists of speech recordings from 20 PPD and 20 HC sampled at $44.100Hz$ with 16 quantization bits. All of the recordings were captured in a sound proof booth. The people that participated in the recording sessions are balanced by gender and age: the ages of the men patients ranged from 56 to 70 (mean 62.9 $\pm$ 6.39) and the ages of the women patients ranged from 57 to 75 (mean 64.6 $\pm$ 5.62). For the case of the healthy people, the ages of men ranged from 51 to 68 (mean 62.6 $\pm$ 5.48) and the ages of the women ranged from 57 to 75 (mean 64.8 $\pm$ 5.65). All of the PPD have been diagnosed by neurologist experts and none of the people in the HC group has history of symptoms related to Parkinson's disease or any other kind of movement disorder syndrome.

The recordings consist of sustained utterances of the five Spanish vowels, every person repeated three times the five vowels, thus in total the database is composed of 60 recordings per vowel on each class. This database is built by *Universidad de Antioquia* in Medellín, Colombia.

Experimental Setup. The voice recordings were segmented and windowed using frames of $40ms$ with an overlap of $20ms$. The characterization of speech recordings is made considering two versions of the TEO, one is the generalized form (TEO_1) as indicated in the equation 2 and the other one is such based on the FFT (TEO_2). In the case of TEO_1, different values of M and N are tested. In the filtering process (band-pass: BW_1 and low-pass: BW_2), different values of

bandwidth were also tested. In table 1 the results after an exhaustive search of the optimal values for M, N, BW_1 and BW_2, are indicated for each Spanish vowel. The same optimal values found for the bandwidths of the low-pass and band-pass filters are also used in the estimation of the measurements based on the TEO_2.

Table 1. Optimal values of bandwidths, exponent m and delay M for the estimation of TEO_1

Vowel	m	M	BW_1	BW_2
/a/	4	3	25	450
/e/	5	5	25	450
/i/	2	3	300	550
/o/	5	5	25	450
/u/	2	5	300	550

The full set of features is formed by 13 measures: the four measurements taken from both estimations of the TEO, the frequency values of the first and second formants (F_1 and F_2) calculated using the MGDF, the amplitude values of the first and the second formants ($\tau(F_1)$ and $\tau(F_2)$) also calculated using the MGDF and the $GDAM$ index. Each measure is obtained for every frame in the voice signal and after, four statistics are estimated per measure (mean value, standard deviation, kurtosis and Sweness), forming a total of 52 measures for representing each voice recording. Table 2 summarizes the set of features considered in this work and the indexes assigned for each one.

Table 2. Index allocation for features

	IA		$CTEO$		ED		LD		$\tau(F_1)$	$\tau(F_2)$	F_1	F_2	$GDAM$
	TEO_1	TEO_2	TEO_1	TEO_2	TEO_1	TEO_2	TEO_1	TEO_2					
Mean	1	2	3	4	5	6	7	8	9	10	11	12	13
Std	14	15	16	17	18	19	20	21	22	23	24	25	26
Kurtosis	27	28	29	30	31	32	33	34	35	36	37	38	39
Skewness	40	41	42	43	44	45	46	47	48	49	50	51	52

* IA :Area under the TEO curves, ED : Euclidean Distance, $CTEO$: Correlation Teager energy operator LD : Logarithmic Distance, $GDAM$: Group Delay Acoustic Measure

The tests performed over the proposed system have been made following the strategy indicated in [19]. The 70% of the data are used for the feature selection and for training the classifier and the remaining 30% is for testing; ten different subsets for training and testing are randomly formed in ten repetitions of the randomization of the data, in order to perform a total of ten independent experiments, each one with its results, allowing the calculation of confidence intervals for the general performance analysis and robustness of the proposed system.

4 Results and Discussion

The automatic features selection process based on PCA gives the resulting sub set of features that better represent the phenomena and also gives the order

Table 3. Indexes of selected features after PCA transformation and correlation analysis

Vowels	Feature Index															
/a/	32	12	3	20	15	26	11	46	13	2	51	29	31	1	42	14
	10	40	49	48	38	50	36	52	9	35	22	23	37			
/e/	2	31	43	12	7	15	16	27	13	24	44	14	11	1	49	37
	50	8	21	51	39	10	9	22	48	35	23					
/i/	19	2	5	40	11	31	42	12	10	39	51	26	14	23	50	49
	9	37	22	48	46	35	33	20	7							
/o/	17	5	20	30	11	25	13	33	14	27	1	42	26	40	36	35
	23	38	37	48	51	52	9	50	49	10	22					
/u/	5	34	40	42	22	14	12	18	27	25	26	52	29	49	37	11
	31	13	9	23	36	10	51	50	48	46	35	33	20	7		

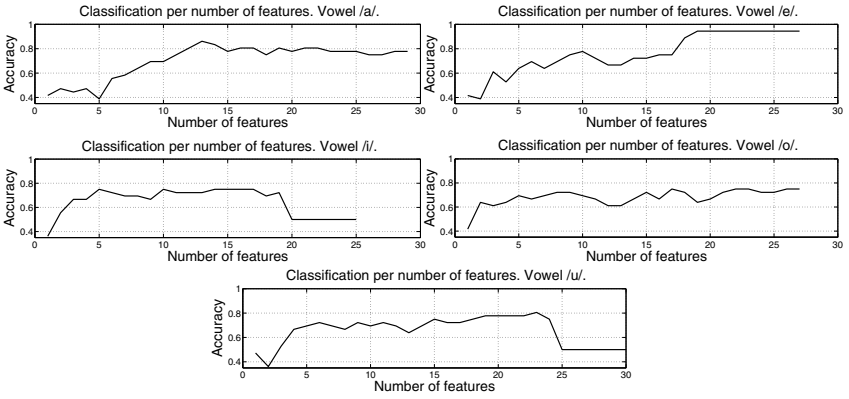

Fig. 2. Success rates obtained per vowel

of features according to their contribution in terms of the cumulative variance. Table 3 indicates which of the features remain after the features selection process per vowel. Considering that the dimensionality of the original representation space is 52, the reductions achieved with selection process ranged from 40% to 54%. Figure 2 shows the accuracy results obtained per vowel when each of the selected feature is progressively added to the classification process. The order in which each feature is added is presented in table 3.

Note that the results obtained with the vowel /e/ are better than those obtained with the other vowels. This is an interesting result specially if it is considered that the risorius muscles are involved in the phonation of vowels /e/ and /i/, but not in other vowels [20].

For the case of vowels /i/ and /u/ the accuracy rates fall when more than 20 and 25 features are considered. The results with the vowels /a/ and /o/ are more stable, with accuracies around 80%. With the aim of indicating the results in a more compact way, in table 4 the results obtained per vowel are presented in

terms of accuracy, specificity and sensitivity. Note that the performance achieved with the vowel /e/ is around 12 percentage points better than those obtained with the other vowels and in such case. This result can also be noted in the receiver operating characteristic (ROC) curves that are shown in figure 3. These kind of curves are widely used in clinical applications and the area under such curves (AUC) is considered as a good statistic for representing the general performance of the system [19]. The AUC obtained per vowel are: AUC /a/: 0.7872, **AUC /e/: 0.9214**, AUC /i/: 0.7954, AUC /o/: 0.7799 and AUC /u/: 0.8111.

Table 4. Performance measures

Vowel	Accuracy	Specificity	Sensitivity
/a/	0.7667±0.0631	0.7905 ±0.0821	0.7523±0.0639
/e/	**0.9250±0.0541**	**0.9092±0.0852**	**0.9559±0.0474**
/i/	0.7778±0.0434	0.7648±0.0516	0.8056±0.0715
/o/	0.7778±0.0393	0.8079±0.0672	0.7616±0.0517
/u/	0.7972±0.0586	0.7732±0.0592	0.8293±0.0640

* The results are presented in terms of mean value ± standard deviation.

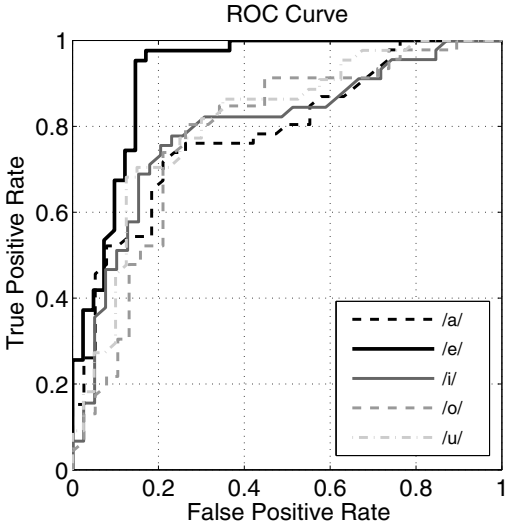

Fig. 3. ROC curves with the best results obtained per vowel

5 Conclusions

A new methodology for the automatic detection of Parkinson's disease, based on the analysis of the low frequency zone of the speech spectrum of the fiver Spanish vowels, is presented.

Different implementations of the Teager Energy Operator are suitable for tackling the problem of the automatic detection of PD. Additionally, the application of Group Delay Functions for improving the resolution of the speech spectrum in the low frequency zone is able for enhancing differences exhibited between the first two formants from speech of PPD and HC.

The stage of automatic features selection allowed the reduction of the dimensionality of the spaces in up to 52%, indicating that there were a lot of redundant information in the original representation space.

According to the obtained results, it is possible to achieve accuracy rates of up to 92% when the vowel /e/ is evaluated. For the other vowels the accuracy results are near to 80%. This difference suggest a more detailed analysis of the set of muscles and/or tissues such as the risorious one, which are involved for the production of the vowel /e/.

Acknowledgments. Juan Rafael Orozco Arroyave is under grants of "Convocatoria 528 para estudios de doctorado en Colombia, generación del bicentenario, 2011" financed by COLCIENCIAS. The authors give a special thanks to all of the patients and collaborators in the foundation "Fundalianza Parkinson-Colombia". Without their valuable support it would be impossible to address this research.

References

1. Ho, A., Iansek, R., Marigliani, C., Bradshaw, J., Gates, S.: Speech impairment in a large sample of patients with parkinson's disease. Behavioral Neurology 11, 131–137 (1998)
2. Ramig, L., Fox, C., Shimon, S.: Speech treatment for parkinson's disease. Expert Review Neurotherapeutics 8(2), 297–309 (2008)
3. McNeil, M.: Clinical Management of Sensorimotor Speech Disorders, 2nd edn. Thieme, New York (2009)
4. Tjaden, K.: Speech and swallowing in parkinson's disease. Top Geriatr Rehabilitation 24(2), 115–126 (2008)
5. Aronson, A., Bless, D.: Clinical voice disorders, 4th edn. Thieme, New York (2009)
6. Kent, R., Weismer, G., Kent, J., Vorperian, H., Duffy, J.: Acoustic studies of dysarthric speech: methods, progress, and potential. Journal of Communication Disorders 32(3), 141–180 (1999)
7. Lee, G., Wang, C., Yang, C., Kuo, B.: Voice low tone to high tone ratio: A potential quantitative index for vowel [a:] and its nasalization 53(7), 1437–1439 (2006)
8. Vijayalakshmi, P., Reddy, M.: Assessment of dysarthric speech and an analysis on velopharyngeal incompetence. In: Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBS), pp. 3759–3762 (2006)
9. Perez, K., Ramig, L., Smith, M., Dromery, C.: The parkinson larynx: tremor and videostroboscopic findings. Journal of Voice 10(4), 353–361 (1996)

10. Giovanni, A., Ouaknine, M., Guelfucci, R., Yu, T., Zanaret, M., Triglia, J.: Nonlinear behavior of vocal fold vibration: the role of coupling between the vocal folds. Journal of Voice 13(4), 456–476 (1999)
11. Tsnas, A., Little, M., McSharry, P., Spielman, J., Ramig, L.: Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease 59(5), 1264–1271 (2012)
12. Kaiser, J.F.: On a simple algorithm to calculate the "energy" of a signal. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 381–384 (1990)
13. Cairns, D., Hansen, J., Riski, J.: A noninvasive technique for detecting hypernasal speech using a nonlinear operator. IEEE Transactions on Biomedical Engineering 43(1), 35 (1996)
14. Ying, G., Mitchell, C., Jamieson, L.: Endpoint detection of isolated utterances based on a modified teager energy measurement. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 732–735 (1993)
15. Eivind, K.: Signal processing using the teager energy operator and other nonlinear operators (2003)
16. Yegnanaryana, B., Duncan, G., Murthy, H.: Formant extraction from group delay function. IEE Colloquium on Speech Processing, 2/1 –2/4 (1988)
17. Vijayalakshmi, P., Reddy, M., O'Shaughnessy, D.: Acoustic analysis and detection of hypernasality using a group delay function. Transactions on Biomedical Engineering 54(4), 621–629 (2007)
18. Daza-Santacoloma, G., Arias-Londoño, J., Godino-Llorente, J., Sáenz-Lechón, N., Osma-Ruiz, V., Castellanos-Domínguez, C.G.: Dynamic feature extraction: an application to voice pathology detection. Intelligent Automation and Soft Computing 15(4), 665–680 (2009)
19. Sáenz-Lechón, N., Godino-Llorente, J., Osma-Ruiz, V., Gómez-Vilda, P.: Methodological issues in the development of automatic systems for voice pathology detection. Biomedical Signal Processing and Control 1, 120–128 (2006)
20. Phonetics, D.: Dissection of the speech production mechanism. Working Papers in Phonetics, UCLA (102), 1–89 (2002)