The 7th International Telecommunications Symposium (ITS 2010)

Classification of voice aging based on the glottal signal

Leonardo Forero[†], Marco Silva[‡], Edson Cataldo[‡], José Apolinário Jr.[§], and Marley Vellasco[†]

[†]**PUC-Rio**
Department of Electrical Engineering
Rua Marquês de São Vicente, 225
22.451-900 Rio de Janeiro, Brazil
{Mendonza,Marley}@ele.puc-rio.br

[‡]**UFF**
Program of Telecommunications Engineering
Rua Passo da Pátria, 156
24.210-240 Niterói, Brazil
ecataldo@im.uff.br,marcoground@gmail.com

[§]**IME**
Department of Electrical Engineering
Praça General Tibúrcio, 80
22290-2770 Rio de Janeiro, Brazil
apolin@ime.eb.br

***Abstract*— Classification of voice aging has many applications in health care and geriatrics. This work focuses on finding the most relevant parameters to classify voice aging. The most significant parameters extracted from the glottal signal are chosen to identify the voice aging process of men and women. After analyzing their statistics, the chosen parameters are used as entries to a neural network and to a support vector machine set to classify male and female Brazilian speakers in three different age groups: *young* (from 15 to 30 years old), *adult* (from 31 to 60 years old), and *senior* (from 61 to 90 years old). The *corpus* used for this work was composed by one hundred and twenty Brazilian speakers (both males and females) of different ages. As compared to similar works, we employ a larger *corpus* and obtain a superior classification rate.**

***Keywords*— Speech processing; voice aging; glottal source; neural network classifier.**

I. Introduction

Body aging reminds us, humans, of our vulnerable condition. Since the aging process is reflected in voice production, voice aging is of great interest in various areas of research [1].

Being aging a physiological process, continuous and irreversible, major physical changes will certainly occur in the elderly. As causes of an aging voice, we can cite the changes that occur in the body, including loss of lung capacity, changes in the epithelium of the larynx, calcification of cartilage, and changes in the vibrations of the vocal folds. From 1995 to 2005, the life expectancy of Brazilian people has increased more than 8% [1]. This scenario is a global tendency, resulting in the need for developing new treatments to provide better life quality to senior people, including care for their voice.

The most common method for extracting voice features is directly from the speech signal. An alternative approach corresponds to some characteristics extracted from the glottal signal, which is the signal obtained just after the vocal cords and before the vocal tract. Since one of the consequences of aging is the change of the vocal fold structures, the glottal signal seems to be important for it preserves characteristics from the movement of the vocal folds without the influence of the vocal tract.

Moreover, it is known that parameters extracted from the glottal signal can help the diagnosis of pathologies of the vocal folds better than those extracted from the voice signal [2]. Nowadays, obtaining this signal is easier due to the development of algorithms that can perform an inverse filtering from the voice signal, eliminating the influence of the vocal tract [2], [3]. Not long before, it used to be necessary to use equipment coupled with micro cameras to record sounds just after air has passed through the vocal folds. This was an invasive technique and very difficult to be carried out.

In a previous work aiming the classification of voice aging, tools such as Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), and neural networks have been used with Mel-Frequency Cepstral Coefficients (MFCC) and the parameters jitter and shimmer as input; adding up to 37 input parameters and using a *corpus* of 12 speech recordings of men and women, the best result obtained was 80% in terms of correct classifications [4].

The MFCC have proven effectiveness for speaker recognition [5], but their performance is not equally well in the task of voice aging classification. Comparing several classifiers for voice aging, the best results were obtained with the use of neural networks and support vector machine (SVM), while GMM and HMM provide the best results for speaker identification [5], [6]. This is mainly due to the fact that, for the classification of voice aging, the temporal feature seems to be not relevant [6]. This work unveils the most relevant parameters and use them in the automatic classification of a large set of speech signals.

This paper is organized as follows: Section II deals with the extraction of a number of features of the

glottal signal while Section III presents the simulation results. Finally, conclusions are summarized in Section IV.

II. Feature extraction from the glottal signal

A. The glottal signal

The voice signal, particularly the one related to voiced sounds, e.g. vowels, starts with the contraction-expansion of the lungs, generating a pressure difference between the air in the lungs and the air near the mouth. The airflow created passes through the vocal folds, which oscillates at the fundamental frequency (*pitch*) of the voice. This oscillation modifies the airflow coming from the lungs, changing it into air pulses. The pressure signal formed by the air pulses is quasi-periodic and it is referred to as the glottal signal.

B. Obtaining the glottal signal

As before mentioned, in order to obtain the glottal signal, it used to be necessary an invasive process. Nowadays, it is possible to obtain the glottal signal using non-invasive methods, by performing an inverse filtering on the voice signal, which consists in eliminating the influence of the vocal tract and the voice radiation caused by the mouth, preserving the glottal signal characteristics [7].

Algorithms that estimate the glottal signal from the speech signal can be classified into two main categories: semi-automatic and manual. In this paper, the inverse filtering algorithm used is of the semi-automatic category, called PSIAIF (Pitch Synchronous Iterative Adaptive Inverse Filtering) [8], [9]. This tool was chosen due to its high performance and easy operation. There is a toolbox implementation in Matlab©, named Aparat [10], which was constructed based especially on the PSIAIF method to obtain the glottal signal; this software also performs the extraction of its main features or parameters.

The parameters which can be extracted from the glottal signal can be divided into three groups: time domain, frequency domain, and those that represent the variations of the fundamental frequency [8].

1) The time domain parameters: The time domain parameters that can be extracted from the glottal signal are described as follows [8], [9].

Closing phase (*Ko*): It describes the interval between the instant of the maximum opening of the vocal folds and the instant when they close [8] (see Fig. 1).

Opening phase (*Ka*): It describes the interval between the instant when the vocal folds start the oscillation up to their maximum opening [8] (see Fig. 1).

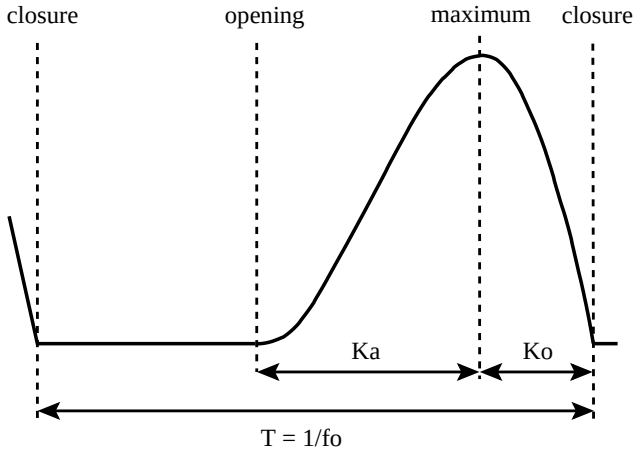

Fig. 1. Parameters Ko e Ka obtained from the glottal signal.

Opening quotient (*OQ*): The ratio between the total time of the vocal folds opening and the total time of a cycle (or period) of the glottal signal (T). It is inversely proportional to the intensity of the voice. The smaller it is, the higher the voice intensity [1], [8]. It is computed as:

$$OQ = \frac{(Ko + Ka)}{T}. \tag{1}$$

It can be divided into OQ_1 and OQ_2, calculated, respectively, from the beginning of the glottal signal formation and from the glottis opening [9].

Closing quotient (*CIQ*): The ratio between the closing phase parameter (*Ko*) and the total length of a glottal pulse (T) [8]. It is inversely proportional to voice intensity. The smaller it is, the higher the voice intensity. It is given by

$$CIQ = \frac{Ko}{T}. \tag{2}$$

Amplitude quotient (*AQ*): The ratio between the glottal signal amplitude (*Av*) and the minimum value of the glottal signal derivative (d_{min}) [11]. It is related to the speaker phonation [9] and is given by

$$AQ = \frac{Av}{d_{min}}. \tag{3}$$

Normalized amplitude quotient (*NAQ*): It is calculated by the ratio between the amplitude quotient (*AQ*) and the total time length of the glottal pulse (T) [12]. It is given by

$$NAQ = \frac{AQ}{T}. \tag{4}$$

Opening quotient defined by Liljencrants-Fant (*OQa*): This is another opening quotient but calculated

from the Liljencrants-Fant model for inverse filtering. Details about this model can be found in [13].

Quasi opening quotient (*QoQ*): It is a relationship between the glottal signal opening at the exact instant of the oscillation and the closing time [9]. It has been used to classify emotions [14].

Speed quotient (*SQ*): is defined in (5) as the ratio of the opening phase length to the closing phase length [8]:

$$SQ = \frac{Ka}{Ko}. \quad (5)$$

It can be divided into SQ_1 and SQ_2, calculated, respectively, from the beginning of the glottal signal formation and from the glottis opening [9].

2) The frequency domain parameters: The frequency domain parameters are next addressed.

Difference between harmonics ($DH12$): Also known as $H_1 - H_2$ and it is the difference between the values of the first and second harmonics of the glottal signal [15], [16]. This parameter has been used to measure vocal quality.

Harmonics relation factor (*HRF*): It relates the first harmonic (H_1) with the sum of the energy of the other harmonics (H_k) [17]. It has also been used to measure the vocal quality and is given by (6).

$$HRF = \frac{\sum_{k>2} H_k}{H_1} \quad (6)$$

3) Pitch parameters: The parameters that represent variations and perturbations in the fundamental frequency are addressed in the following.

Jitter: Variations in fundamental frequency between successive vibratory cycles [18], [19]. Changes in jitter may be indicative of neurological or psychological difficulties [1].

Shimmer: Variations in amplitude of the glottal flow between successive vibratory cycles [18], [19]. Changes in the shimmer is found mainly in the presence of mass lesions in the vocal folds, such as polyps, edema or carcinomas [1].

Harmonics-to-Noise-Ratio (*HNR*): is an acoustic measure that determines the ratio between the amount of harmonics (periodic) and the amount of noise (aperiodic) present in a given voice. This measurement has proved to be a sensitive and strong indicator of vocal productions in adults [1].

III. Experimental Results

A. The corpus used

The voice database used for this work is composed of 60 male and 60 female voice registers for the sustained vowel /e/, in Portuguese as spoken in the city of Rio de Janeiro, Brazil. This database was divided in three different age groups: from 15 to 30 years old ("young"), from 31 to 60 years old ("adult") and from 61 to 90 years old ("senior"). There were 20 voice registers for each age group.

TABLE I
Parameters for female groups

Parameters	Measures	Groups		
		15–30	31–60	61–90
Shimmer	Median	5.6	7.8	8.9
	Mean value	7.37	10.53	10.40
	Dispersion	36.76	83.13	37.44
AQ	Median	0.32	0.56	0.65
	Mean value	0.6	0.7	0.8
	Dispersion	0.059	0	0.018
OQ_1	Median	0.426	0.626	0.822
	Mean value	0.587	0.614	0.799
	Dispersion	0.082	0.062	0.0339
OQ_2	Median	0.28	0.448	0.598
	Mean value	0.409	0.477	0.617
	Dispersion	0.039	0.051	0.0258

B. Inverse Filtering

For each vocal register, the corresponding glottal signal was obtained by inverse filtering (using PSIAIF) and the parameters were extracted using the softwares Aparat [10] and Praat [20]. The following parameters were obtained: fundamental frequency (fo), jitter, shimmer, *HNR*, *Ko*, *Ka*, *NAQ*, *AQ*, *CIQ*, OQ_1, OQ_2, *OQa*, *QoQ*, SQ_1, SQ_2, $DH12$ and *HRF*. The parameters were separated according to the groups to which they belonged. *OQ* was divided into OQ_1 and OQ_2, the open quotients calculated from the so-called primary and secondary openings of the glottal flow. *SQ* was divided into speed quotients also computed from the primary and the secondary openings of glottal signal [8].

C. Choosing the parameters

For each parameter of the glottal signal obtained from the database, we have computed the median, the mean value, and the dispersion in each age group. The parameters that showed the greatest differences between the individual age groups are, for women: Shimmer, *AQ*, OQ_1, and OQ_2; the results are shown in Table I. Following the same approach for men, the parameters were: *Ko*, OQ_1, *OQa*, and SQ_2; the results are shown in Table II.

For the female groups, Shimmer and *AQ* has mean value and median very similar to the range of 31-60 and 61-90, but the variance is lower in the range of 61-90; that is most probably related to the fact that women after menopause, between 45 and 55, having a sudden change in hormones, produce changes in voice. The

TABLE II
PARAMETERS FOR MALE GROUPS

Parameters	Measures	Groups		
		15–30	31–60	61–90
Ko	Median	0.00061	0.00081	0.0017
	Mean value	0.0011	0.001	0.002
	Dispersion	7.2×10^{-6}	3×10^{-6}	6.6×10^{-6}
OQ_1	Median	0.4	0.41	0.83
	Mean value	0.5	0.4	0.8
	Dispersion	0.135	0.015	0.798
OQa	Median	0.165	0.18	0.35
	Mean value	0.161	0.1705	0.3045
	Dispersion	0.0025	0.0023	0.0144
SQ_2	Median	1.95	1.7	1.3
	Mean value	1.92	1.73	1.32
	Dispersion	0.4198	0.419	0.3338

OQ_1 and OQ_2 parameters are good age discriminators: their values are greater for the age group 61-90.

For the male groups, the mean value and median for *Ko* and OQ_1 are high in the 61-90 group when compared to other groups. It may occur due to the vocal folds relaxing, during the aging process, because the closing of the vocal folds takes more time.

The speed ratio (SQ_2) is lower for the group 61-90 years; this result was expected considering that the opening speed of the glottis decreases with age. Figs. 2 and 3 show the box plots of the parameters for female and male groups. These graphs show the median parameters for each age group and the behavior of the data.

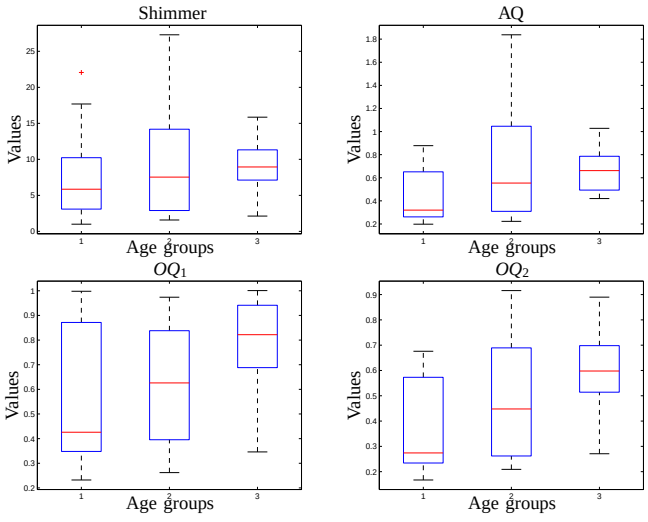

Fig. 2. Box plot of relevant parameters for female groups 1 (15-30), 2 (31-60), and 3 (61-90).

D. Classification of voice aging using NN and SVM

With the parameters of the glottal signal, the age groups were classified by means of an artificial neural

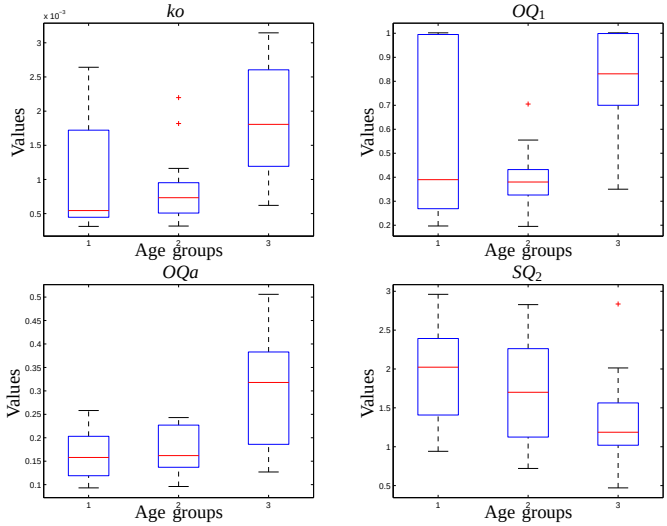

Fig. 3. Box plot of relevant parameters for male groups 1 (15-30), 2 (31-60), and 3 (61-90).

TABLE III
CONFUSION MATRICES OF THE CLASSIFICATION VIA ANN

Female groups	Age 15-30	Age 31-60	Age 61-90
Age 15-30	4	0	0
Age 31-60	0	3	1
Age 61-90	0	1	3

Male groups	Age 15-30	Age 31-60	Age 61-90
Age 15-30	4	1	0
Age 31-60	0	3	0
Age 61-90	0	0	4

network (ANN) and a support vector machine (SVM).

The ANN utilizes an MLP and is set to classify the speakers such that each output corresponds to one age group. For the training of the ANN, we have employed 65% of the speech database while 15% was used as validation set (to avoid the network overtraining and to choose the number of processors and the hidden layers), and 20% for testing. In both cases, male and female *corpora*, the best result was achieved using 3 neurons with only one hidden layer. For the female database, 10 speakers were classified correctly while 2 speakers incorrectly; this corresponds to a guess rate of 83.33%. For the male database, 11 out of 12 speakers were classified correctly, guess rate of 91.66%. Confusion matrices are presented in Tab. III.

The support vector machine (SVM) implemented to classify the speakers into the three groups achieved the best result when using RBF (Gaussian functions) kernel function with the regularization constant set to $C = 10$, and a Gaussian standard deviation of $\sigma = 1$. For the female *corpus*, the result was 10 out of 12 speakers classified correctly, that is, 83.33%. For the male set, the SVM classified correctly 10 speakers

TABLE IV
CONFUSION MATRICES OF THE CLASSIFICATION VIA SVM

Female groups	Age 15-30	Age 31-60	Age 61-90
Age 15-30	4	2	0
Age 31-60	0	2	0
Age 61-90	0	0	4
Male groups	Age 15-30	Age 31-60	Age 61-90
Age 15-30	4	0	0
Age 31-60	0	3	0
Age 61-90	0	1	4

and incorrectly 2 speakers, that is, 91.66%. Confusion matrices are presented in Tab. IV.

In a previous similar work [4], using only parameters jitter and shimmer with a similar database, a 50% accuracy was obtained. Therefore, the use of other parameters such as *Ko*, *SQ*, and *OQ*, usually employed to identify pathologies and emotions related to the voice, improved the aging classification rate.

IV. CONCLUSIONS

The results obtained here in were satisfactory and better than those achieved in previous similar work. The main contribution of this paper is the use of relevant parameters of the glottal signal in classifying aging voices. Both automatic classifiers used in this work are efficient in combining the parameters to provide an improved overall performance.

ACKNOWLEDGMENTS

The authors thank CAPES, CNPq, and FAPERJ for partial funding of this work.

REFERENCES

[1] I. R. dos Santos, "Análise acústica da voz de indivíduos na terceira idade" (*in Portuguese*), M. Sc. dissertation, Programa de Pós-graduação Interunidades em Bioengenharia (EESC/IQSC/FMRP), University of São Paulo, São Carlos, Brazil, 2005.

[2] P. Gómez-Vilda, et al., "Glottal Source biometrical signature for voice pathology detection," *Speech Communication*, vol. 51, no. 9, pp. 759-781, Sept. 2009.

[3] I. M. Verdonck-de Leeuw and H. F. Mahieu, "Vocal aging and the impact on daily life: a longitudinal study," *Journal of Voice*, vol. 18, no. 2, pp. 193–202, June 2004.

[4] A. Sadeghi Naini and M. M. Homayounpour, "Speaker age interval and sex identification based on jitters, shimmers and mean MFCC using supervised and unsupervised discriminative classification methods," in Proc. 8th International Conference on Signal Processing (ICSP), Beijing, China, Nov. 2006.

[5] J. P. Campbell, Jr., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437-1462, Sept. 1997.

[6] M. H. Sedaaghi, "A comparative study of gender and age classification in speech signals," *Iranian Journal of Electrical & Electronic Engineering*, vol. 5, no. 1, pp. 1-12, March 2009.

[7] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109-118, June 1992.

[8] H. Pulakka, "Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography," M. Sc. dissertation, Department of Computer Science and Engineering, Helsinnki University of Technology, Espoo, Finland, 2005.

[9] J. S. de Mattos, "Um estudo comparativo entre o sinal electroglotográfico e o sinal de voz," M.Sc. dissertation (*in Portuguese*), Programa de Mestrado em Engenharia de Telecomunicações, Universidade Federal Fluminense (UFF), Niterói, Brazil, 2008.

[10] Software Aparat. Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing. [Online] Available: http://aparat.sourceforge.net/index.php/Main_Page [Accessed: Feb. 2009].

[11] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," *Speech Communication*, vol. 18, no. 2, pp. 131-138, April 1996.

[12] P. Alku, T. and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, Aug. 2002.

[13] C. Gobl and A. N. Chasaide, "Amplitude-based source parameters for measuring voice quality," in Proc. Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03), Geneva, Switzerland, Aug. 2003.

[14] A. M. Laukkanen, E. Vilkman, P. Alku, and H. Oksanen, "Physical variations related to stress and emotional state: a preliminary study," *Journal of Phonetics*, vol. 24, no. 3, pp. 313-335, July 1996.

[15] I. R. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *Journal of the Acoustical Society of America*, vol. 91, no. 5, pp. 2936-2946, May 1992.

[16] M. Airas, "Methods and studies of laryngeal voice quality analysis in speech production," D. Sc. thesis, Department of Signal Processing and Acoustics, Helsinki University of Technology, Espoo, Finland, 2008.

[17] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 2394-2410, May 1990.

[18] A. S. Brandão, E. Cataldo, F. R. Leta, and J. C. Lucero, "Usando Redes Neurais para Classificação de padrões de voz" (*in Portuguese*), in Anais do XXVIII Congresso Nacional de Matemática Aplicada e Computacional (XXVIII CNMAC), São Paulo, Brazil, Sept. 2005.

[19] M. N. Vieira, "Automated measures of dysphonias and the phonatory effects of asymmetries in the posterior larynx," Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland, UK, 1997.

[20] Software Praat. University of Amsterdam. [Online] Available: http://www.fon.hum.uva.nl/praat/ [Accessed: Feb. 2009].