A Comparative Analysis of Vocoders for Communications in Portuguese

Dirceu Leite Cavalcante
Depto. de Eletrônica e Sistemas
Universidade Federal de Pernambuco
Recife, PE, Brazil
dirceu_cavalcante@hotmail.com

Rafael Dueire Lins
Depto. de Eletrônica e Sistemas
Universidade Federal de Pernambuco
Recife, PE, Brazil
rdl@ufpe.br

Abstract— **This paper analyzes the performance of voice encoders used for digital communications, especially in VoIP calls, for Brazilian Portuguese phonemes. The focus is the adaptation capacity of encoders to vocal tract changes in these intervals during the signal processing. Short sentences containing all the basic Portuguese phonemes were used. Information about fundamental frequency and the three first formants was extracted, for each interval, for a group formed by men and women of different ages. These intervals were created using the software Praat, which was also used to extract all the parameters. The results obtained showed a subtle difference in signal processing of the formant frequencies between sexes. It was observed a phenomenon of pitch correction in some intervals with a variant phoneme for both sexes. The encoders G.722, G.723.1, G.726, G.728, G.729A, iLBC and Speex were analyzed.**

Keywords - vocoders, G.722. G.723.1, G.726, G.728, G.729A, iLBC, Speex.

I. Introduction

The explosive growth of the Internet in the 1990s, allied to lower cost of broadband connection, allowed the birth and spreading of multimedia services, such as Voice over IP (VoIP). Actually, VoIP is a set of protocols for transport of digitalized human voice over packet-switched networks [1], as IP networks. The main attractive of VoIP is the low cost and quality as good as digital Public Switched Telephone Network (PSTN) calls.

Human voice encoders or *vocoders* are responsible for the voice digitalization in VoIP systems. They limit the audio quality, taking into account environmental and network factors. Parametric vocoders extract parameters of voice quantized samples to model the speaker's vocal tract to synthesize voice, allowing transmissions with better information per byte ratio.

The focus of this paper is to compare the adaptation capacity of vocoders to vocal tract changes during the spelling of intervals containing Portuguese phonemes. Possible changes in voice processing with sex and age are analyzed also.

II. Speech Sounds

From all sounds produced by human vocal tract, a phoneme is the smallest sound unit capable of distinguishing words in a language. Thus, the English words "net" and "jet" are distinguished by the initial phonemes /n/ and /j/. It is conventional to transcribe phonemes between slashes to differentiate from phonetic translations. The typical phoneme length is within 50-100 ms.

Phonemes are classified in two major groups: vowels and consonants. Vowels are phonemes produced by continuous flow of air that travel across the vocal tract. When the air flow suffers interruption, typically in the supralaryngeal cavity, a consonant is emitted.

Since there are phonetic variations of a language even within a country, the classification of phonemes varies accordingly. The foreign accent is caused by a phonetic error, a mispronunciation of a phoneme not generally accepted (some "*errors*" become a regional accent).

A. Vowels

There are 13 vowel phonemes in Portuguese spoken in Brazil: /á, â, ã, é, ê, ẽ, ó, ô, õ, í, ĩ, ú, ũ/. They can be classified by articulatory and acoustic parameters: tongue position in vertical and horizontal axes, nasality and roundedness of lips as in Table I.

B. Consonants

There are 19 consonant phonemes in Portuguese spoken in Brazil: /b, d, f, g, j, k, l, m, n, p, r, s, t, v, x, z, R, λ, ñ/. They can be also classified by articulatory and acoustic parameters: manner and place of articulation of the air flux function of vocal cords and cavities, as in Table II.

III. A Source-Filter Model of Speech Production

The human vocal tract can be considered as an acoustic tube with variable sectional area that extends from vocal cords to supralaryngeal cavity. This approximation leads to innumerous models of speech production as the model source-filter.

The source-filter model is used by many speech synthesis algorithms [3] and can be approximated by a series connection of three roughly independent functions:

$$X(s) = Z_r(s)T_g(s)U_g(s) + T_g(s)P_h(s) + T_f(s)P_f(s). \qquad (1)$$

$$\alpha + \beta = \chi.$$

TABLE I. PORTUGUESE VOWEL PHONEMES

Vertical Axis	Horizontal axis						Roundedness of lips
	Frontal		*Central*		*Back*		
Open							*Rounded*
	/á/						*Not-rounded*
Open-mid					/ó/		*Rounded*
	/é/		/â/	/ã/			*Not-rounded*
Close-mid					/ô/	/õ/	*Rounded*
	/ê/	/ẽ/					*Not-rounded*
Close					/ú/	/ũ/	*Rounded*
	/í/	/ĩ/					*Not-rounded*
	Oral	*Nasal*	*Oral*	*Nasal*	*Oral*	*Nasal*	
	Nasality						

The functions are: three sources, two transfer functions and one radiation characteristic.

A. Source Functions

The source of sound in vocal tract can be either a fluctuating pressure or a fluctuating flow.

The Voice Source ($U_g(s)$) consists of puffs of air released through the glottis at a frequency of F_0 called Pitch. The vocal cords vibration is periodic, smooth like a half-wave-rectified cosine and its spectrum drops as $1/f^2$ about 500 Hz. Pitch can be estimated by vocal cords parameters like and mass. In male adults the Pitch extends from 80 to 200 Hz, in female adults from 150 to 350 Hz, and in children from 200 to 500 Hz.

The Aspiration Source ($P_h(s)$) is caused by “turbulent” jets of air from the glottis striking against the false vocal cords. The sound radiated spectrum is nearly flat from 500 to 3000 Hz and can be modeled by a white noise source.

The Frication Source ($P_f(s)$) is caused by turbulent jets of air from a vocal tract constriction within the supralaryngeal cavity and only occurs in sounds with obstructed air flow (constrictive and occlusive sounds).

B. Transfer Functions

The transfer function is the ratio between flow and pressure at the lips and flow or pressure at the source.

The Glottal Transfer Function ($T_g(s)$) is active between the glottis and the lips, so it works for both voicing and aspiration sources. It can be approximated by an all-pole function, and its poles are complex:

$$s_n = -\pi B_n + j2\pi F_n. \qquad (2)$$

The pole frequencies F_n are called formant frequencies, and the B_n are called formant bandwidths. The formant frequencies are resonance frequencies of the entire vocal tract. The three lower formant frequencies (F_1, F_2 and F_3) are sufficient to distinguish speaker voices.

One may suppose that the vocal tract is modeled as a hard-walled tube of constant area, closed at one end (glottis) and open at the other end (mouth). The system will resonate at any frequency for which the standing wave pattern has flow zero at the closed end and null pressure at the open end. The resonant frequencies are such that the tube length, *L*, is an odd multiple of a quarter wavelength:

$$F_n = (2n - 1)v_{sound}/4L. \qquad (3)$$

This constant area approximation is a reasonable model for the spelling of one phoneme. For men (L ~ 17.5 cm), the three first formant frequencies are at approximately 500 Hz, 1500 Hz and 2500 Hz. For women (L ~ 15 cm), the three first formant frequencies are at approximately 580 Hz, 1750 Hz and 2900 Hz.

TABLE II. PORTUGUESE CONSONANTS PHONEMES

Function of cavity		*Oral*						*Nasal*
Manner of articulation		*Oclusive*		*Constritive*				*Oclusive*
				Fricative		*Lateral Aproximant*	*Trill*	
Function of vocal cords		*Unvoiced*	*Voiced*	*Unvoiced*	*Voiced*	*Voiced*	*Voiced*	*Voiced*
Place of articulation	Bilabial	/p/	/b/					/m/
	Labiodental			/f/	/v/			
	Linguodental	/t/	/d/					
	Alveolar			/s/	/z/	/l/	/r/	/n/
	Postalveolar			/x/	/j/			
	Palatal					/λ/		/ñ/
	Velar	/k/	/g/					
	Uvular						/R/	

The Frication Tranfer Function ($T_f(s)$) is also an all-pole function and its poles can be interpreted as the glottal poles:

$$s_{f,n} = -\pi B_{f,n} + j2\pi F_{nf,n}. \quad (4)$$

Using the same constant area approximation, the fricative formants can be estimated using now the length, l_f, from the open end to the constriction point. In Table III, there are the first fricative formant estimations for various articulation points. Notice in that the first fricative formant frequencies for Bilabial and Labiodental articulation points are penalized with low sampling frequency.

C. *Radiation Characteristic*

The radiation characteristic is the ratio between the sound recorded by the microphone and the flow or pressure at the mouth.

It is possible to simplify the equations by calculating the pressure at the lips.

If the source is a pressure, then:

$$R(s) = 1 \quad (5)$$

If the source is a flow, we can use the Flanagan's resistor-inductor model of the radiation impedance [4]:

$$R(s) = Z_r(s) = sL_rR_r/(R_r+L_r) \quad (6)$$

TABLE III. ESTIMATION OF FIRST FRICATIVE FORMANT

Articulation Point	**l_f (cm)**	**$F_{f,1}$ (Hz)**
Bilabial	0-0,5	17700-∞
Labiodental	1-1.5	5900-8800
Alveolar	1.5-2.5	3500-5900
Palatal	2.5-3.5	2500-3500
Velar	3.5-8	1100-2500

IV. EXPERIMENTS

To measure the adaption capacity of vocoders to represent vocal tract variations, parameters of source-filter model (Pitch and formant frequencies) were extracted during intervals containing all Portuguese consonant and vowel phonemes. Phrases were used to simulate conversation, instead of isolated phonemes, making it closer to real situations.

Samples were collected and separated in groups containing men and women with different ages, as in Table IV. All recordings and parameters extractions were made using the software Praat, which allows fast spectrogram visualization and easy separation of intervals in voiced zones. The recorded samples contain the voiced zone between silence zones. This way it is possible reject samples with background noise which can cause bad spectrogram visualization.

TABLE IV. GROUPS FORMED FOR THE ANALYSIS

	Group	**Age**	**Size of group**
Male	HG1	14-34	2
	HG2	35-59	2
	HG3	60-80	2
Female	MG1	14-34	2
	MG2	35-59	2
	MG3	60-80	2

Pulse Code Modulation (PCM) was used to record the samples and to compare the results with test versions of the vocoders G.722, G.723.1, G.726, G.728, G.729A, iLBC and Speex. The encoding algorithms are showed in Table V. Notice that all vocoders have predictive analysis and some can be used with other media like video.

Only the signal processing was analyzed, with serial encoding and decoding of samples bitstream. So delay and others factors are not considered. The total time of processing was collected with 0.2 s resolution.

V. COMPARATIVE ANALYSIS

The analysis comprehends four aspects: perceptual, time processing ratio, Pitch error and formant frequencies error.

A. *Perceptual Analysis*

Noises and distortion were perceived only in files decoded by G.722 and G.723.1. The vocoder G.722 introduced a small noise in all decoded audio files, just like a clip in the beginning of file, with no perceptablealteration in the spectrogram. The vocoder G.723.1 introduced strong noise and attenuation in some decoded files with alteration in spectrogram.

B. *Time Processing Ratio*

This ratio is the relationship between the time of signal processing (encoding plus decoding) and length of the audio file, showing the statistic percentile of the voice frame processing time for all vocoders, except Speex.

TABLE V. VOCODERS USED IN THE ANALYSIS

Vocoder	**Algorithm**	**Transmission (kbps)**	**MOS**
G.711	PCM	64	4,4
G.722	SB-ADPCM	64, 56 and 48	4,5
G.723.1	MP-MLQ and ACELP	6,3 and 5,3	4
G.726	ADPCM	40, 32, 24 e 16	4,3
G.728	LD-CELP	16	4,2
G.729A	CS-CELP	8	4
iLBC	iLBC	13,33 and 15,2	4,1
Speex	CELP	2 to 44	3,8

Figure 1 shows the results of the Time processing Ratio for all groups. Notice that the worst results belong to G.723.1 and G.728, and the best results belong to G.722 and G.726

with time processing inferior to the resolution. There is no perceptable difference between sexes and ages.

C. *Pitch Error*

The Pitch error was obtained by the percentile difference between the minimum, mean and maximum Pitch extracted by PCM audio files with the extracted by decoded files.

Figures 1 to 7 show the mean Pitch error for each group. There is a subtle difference between results for women and men, where the pitch mean error is greater for women in all vocoders.

A correction Pitch phenomenon was observed, for both sexes and all ages, in intervals containing consonantal variants phonemes, usually in intervals containing /p, t, k, s/ phonemes.

Figures 8 and 9 present the graphics of the Mean Pitch extracted by PCM (original) and decoded files for a female 18 years old. One may notice that, in the last interval, there is no Pitch information in PCM file, but the vocoders G.723.1, G.729A, iLBC and Speex introduced Pitch Information.

From the fidelity point of view, this phenomenon represents a prediction error because it modified the original information. At same time, it represents a correct estimative about the kind of sound should be in the interval.

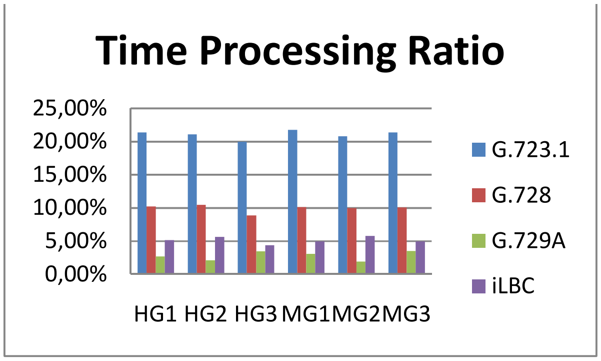

Figure 1. Time Processing Ratio Graphic for all groups

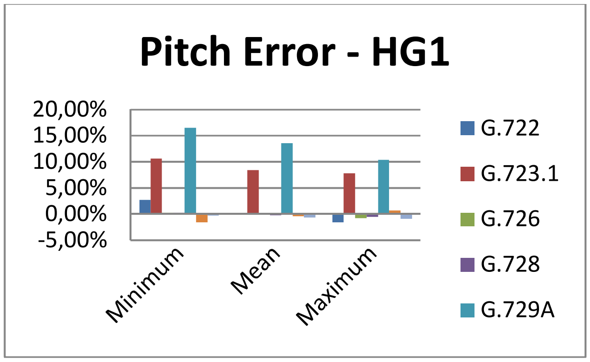

Figure 2. Mean Pitch error for males with 14-34 years old

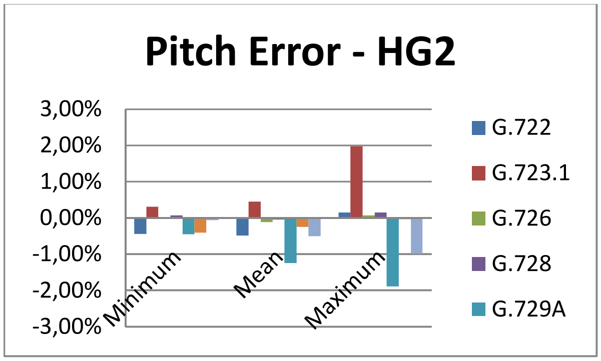

Figure 3. Mean Pitch error for males with 35-59 years old

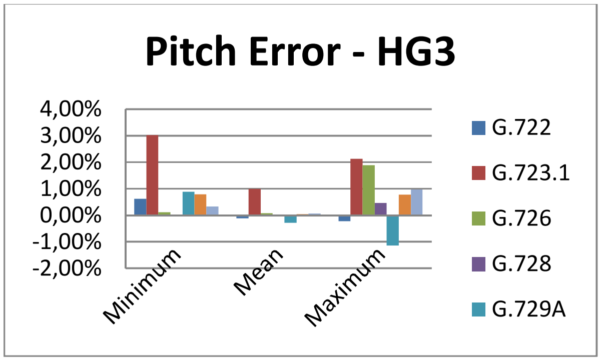

Figure 4. Mean Pitch error for males with 60-80 years old

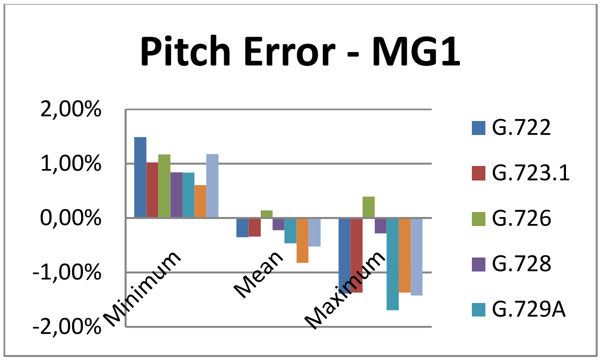

Figure 5. Mean Pitch error for females with 14-34 years old

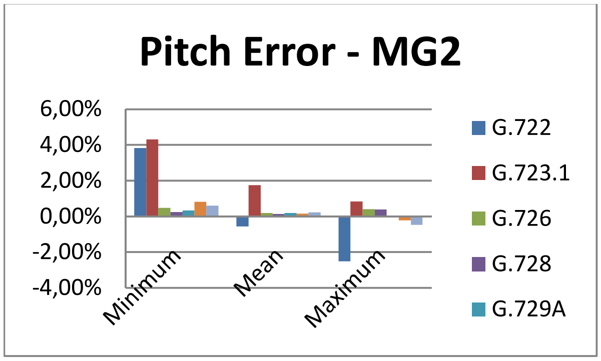

Figure 6. Mean Pitch error for females with 35-59 years old

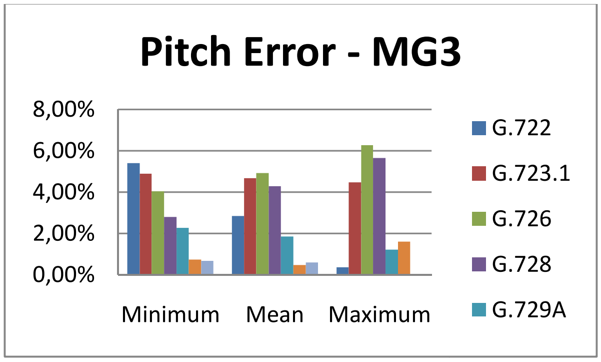

Figure 7. Mean Pitch error for females with 60-80 years old

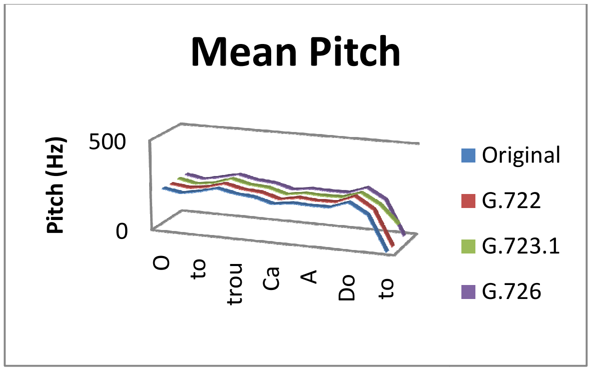

Figure 8. Pitch correction phenomenon for female 18 years old

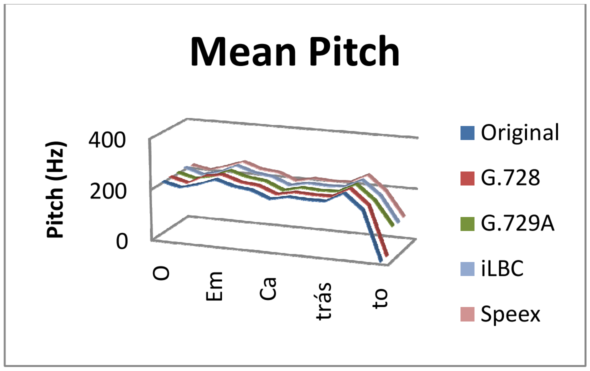

Figure 9. Pitch correction phenomenon for female 18 years old

D. *Formant Frequency Error*

The formant frequency errors were also obtained by the percentile difference between the frequencies extracted by PCM audio files with the extracted by decoded files.

Figures 10 to 16 show the mean formant frequencies errors for each group. Notice that the mean error is greater in the results for male than in female spaekers. There are greater mean errors localized in HG2 and MG3 groups.

Variation with age was not noticed for both sexes. However, it was observed some difficult estimating the third formant frequency for all vocoders.

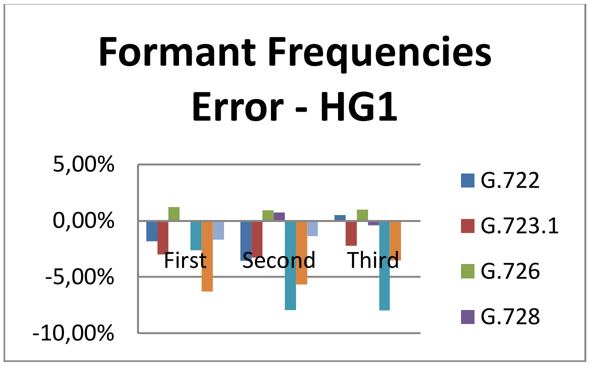

Figure 10. Pitch correction phenomenon for female 18 years old

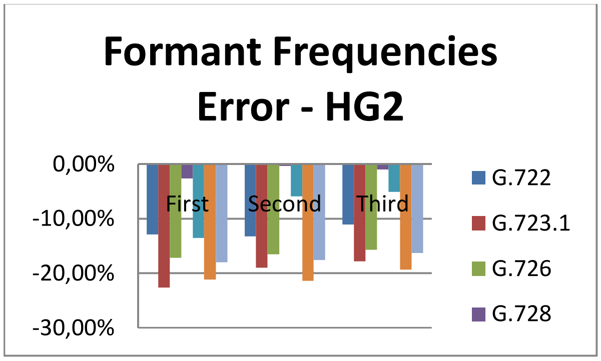

Figure 11. Pitch correction phenomenon for female 18 years old

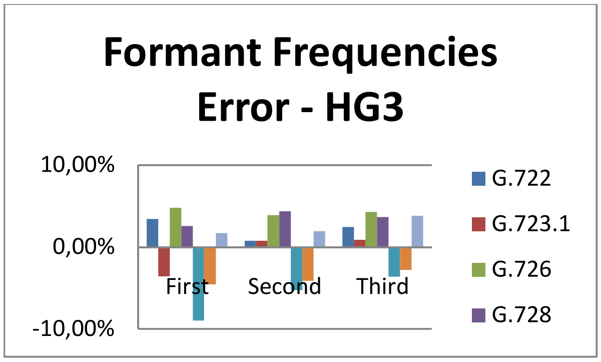

Figure 12. Pitch correction phenomenon for female 18 years old

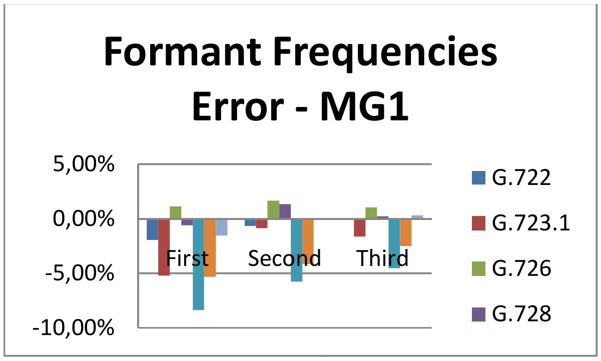

Figure 13. Pitch correction phenomenon for female 18 years old

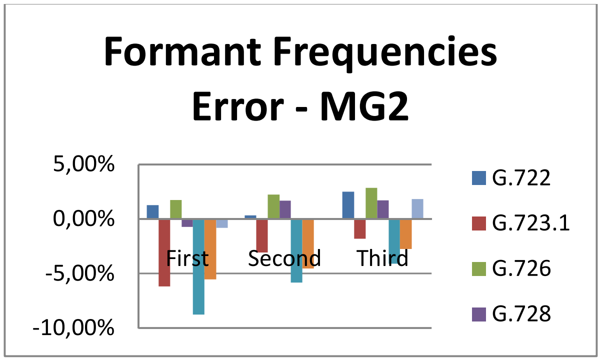

Figure 14. Pitch correction phenomenon for female 18 years old

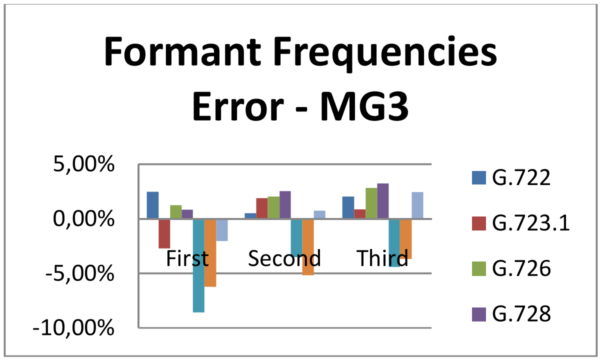

Figure 15. Pitch correction phenomenon for female 18 years old

CONCLUSIONS

The use of intervals containing all Portuguese basic phonemes allowed observing the adaptation capacity of vocoders to changes of the vocal tract, considering only the quality if voice signal point of view.

It was possible to observe that, from the degradation of voice quality point of view, the worst results belonged to G.723.1 and G.729A, while the best results to G.728 and G.726.

Results showed a little difference between men and women regarding Pitch and formant frequencies. However, it was not possible to observe differences with the increase of age. It was noticed a concentrated high level error in HG2 and MG3 groups. Using a larger group with some repeatability, it will be possible to identify which phonemes are better estimated. With this information, it would be possible to change the vocoders during the communication to improve the quality versus transmission relationship.

The Pitch correction phenomenon was observed, for both sexes and all ages, in intervals containing variant phonetics, especially in syllable with /s, p, t, k/ phonemes. It represents a correct estimative of the phoneme, increasing the clarity of sound. However, from the fidelity point of view represents a processing error by changing of the phoneme information.

Regarding time processing ratio, G.723.1 and G.728 presented the worst results with mean ratios of 20% and 10% respectively. This ratio means that the algorithm takes a percent of frame size to process the voice signal. So a ratio of 10% means that, for a frame with 10 ms length, the algorithm takes 2 ms to process the frame.

REFERENCES

[1] Voice over Internet Protocol, disponível em http://en.wikipedia.org/wiki/VoIP.

[2] MANOSSO, R., *Gramática Descritiva*, disponível em http://www.radames.manosso.nom.br/gramatica/fonemas.htm.

[3] HASEGAWA-JOHNSON, M. *Lectures Notes in Speech Production*, University of Ilinois, Feb. 17, 2000.

[4] FLANAGAN, J. L. *Spech Analysis Syntesis and Perception*, 2 ed. New Jersey, Springer-Verlag, 1972.