PROPOSAL AND VALIDATION OF AN OBJECTIVE METHOD FOR QUALITY ASSESSMENT OF SPEECH CODECS AND COMMUNICATION SYSTEMS

Jayme Garcia Arnal Barbedo
jgab@decom.fee.unicamp.br

Amauri Lopes
amauri@decom.fee.unicamp.br

Resumo

Apresenta-se neste trabalho um método de avaliação objetiva de qualidade de codecs na faixa de telefonia (300-3400 Hz) que reúne as principais características dos métodos presentes na literatura, denominado Medida Objetiva de Qualidade de Voz (MOQV). Apresenta-se uma análise de seu desempenho, tendo como base os experimentos presentes em um banco de dados da *International Telecommunication Union* (ITU). A análise de desempenho oferece uma ampla caracterização do método, permitindo determinar-se em quais circunstâncias seu uso é ou não adequado, além de fornecer subsídios para sua utilização frente às diferentes situações encontradas na prática.

Palavras-chave: *avaliação de qualidade, codecs de voz, método MOQV, método PSQM, modelagem psicoacústica.*

Abstract

This work presents a method of objective quality assessment of telephony band codecs (300 – 3400 Hz) that assembles the main characteristics of the methods found in the literature, named "Medida Objetiva de Qualidade de Voz" (MOQV). An analysis of its performance is presented, based on a set of experiments available in a database of the International Telecommunication Union (ITU).Such performance analysis provides an ample characterization of the method allowing the determination of the circumstances where its use is adequate, besides providing subsidies to its use face to several practical situations.

Keywords: *quality assessment, speech codecs, MOQV method, PSQM method, psycho-acoustic modeling.*

1 Introduction

The enhancement of the digital signal processing techniques and technology has motivated a growing interest in more efficient voice coding/decoding devices (codecs). The codec quality assessment is necessary to the development of those devices and also to the new digital telecommunication network planning.

The classic objective measures for performance assessment of speech signals, such as error rate and signal-to-noise ratio, do not exhibit good correlations with the impression of the telecommunications systems users. Therefore, the subjective measures of quality are still widely employed. However, their costs, complexity and waste of time, motivated the search for new efficient methods to perform objective measures that estimate the subjective quality in a suitable way.

In that context, ITU-T stimulated the development of a new objective method capable to attend some rigorous performance exigencies. Several proposals were submitted to a battery of tests and, after a careful comparison, the Perceptual Speech Quality Measure (PSQM) (ROYAL PTT, 1991; BEERENDS & STEMERDINK, 1994) reached the best estimates of the subjective measures. This method was adopted by ITU-T, originating the Recommendation P.861 (ITU-T Rec. P.861, 1996).

This paper presents the fundamental concepts employed in the development of a new method based on the PSQM, but incorporating some proposals that improve its performance, besides additional resources that turn it more complete and versatile (BARBEDO, 2001). This package is named MOQV method. The tests performed with the MOQV are also presented, whose results characterize the behavior and scope of this method face to practical situations.

The Section 2 presents a brief description of the anatomy and mathematical modeling of the human ear, which are used to extract the main concepts used in the most of the existing objective measures. The section 3 presents the basic scheme of the MOQV method and its main stages. The Section 4 details the functioning and structure of the MOQV method. The Section 5 presents the description of the tests performed for the MOQV, whose results allows the description of the scope of such method. Finally, the Section 6 presents the conclusions.

2 Psycho-acoustic Methods

The best objective methods are based on a mathematical modeling of the human ear (BEERENDS & STEMERDINK, 1994; ATKINSON, 1997; VORAN, 1999; BERGER, 1997). The Figure 1 illustrates the basic common characteristics of those methods.

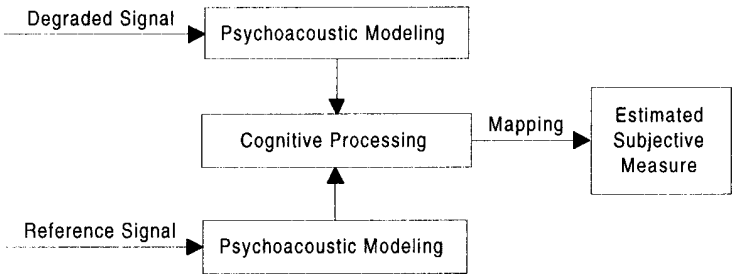

Figure 1. Basic scheme of the psycho-acoustical based measures.

2.1 Psycho-acoustic Modeling

The pre-processing accomplished by the ear over the acoustic signal is an objective activity, since it involves the transformation of the acoustic signal incoming the outer ear into electrical impulses in the neuron bundles distributed along the cochlea in the inner ear. The subjective processing will be performed by the superior functions of the brain cortex, based on that condensed signal generated by ear. In this way, a good model of the processes involved in the speech perception can be applied to a large number of people.

The Figure 2 shows a longitudinal section of the ear. The acoustic wave that arrives at the auditory duct is transformed into oscillations of the little bones in the middle ear – hammer (malleus), incus (anvil) and stapes (stirrup). The middle ear bones stimulate the cochlea through the oval window, producing the movement of the inner liquid. The cochlea can be modeled as a tube with two chambers separated by a structure named basilar membrane, as showed by the Fig. 3. Opposed to the oval window, there is an orifice that performs the junction of such chambers, named helicotrem. The basilar membrane presents a mechanical resistance that varies along its extension: it is slender and tight near the oval window, vibrating in the higher frequencies, while it is thick and flaccid in its apex, vibrating at lower frequencies. The basilar membrane is still composed by two other structures: the basilar fibers and the organ of Corti. The basilar fibers are about 20,000 slender spines whose width varies along the membrane, being shorter near the oval window and longer at the cochlea apex (GUYDON, 1985). The organ of Corti is a spiral structure within the cochlea, and its the sense organ of hearing. The waves generated by the stirrup, in response to a sinusoidal signal, travel along the cochlea, vibrating the basilar membrane at the same frequency of the input signal (FLETCHER, 1953). Therefore, the basilar fibers vibrate and stimulate the hair cells, which are little structures that transform the basilar fiber movement into neural impulses. Such impulses are then transmitted by the cochlear nerve to the specific area in the brain cortex.

2.2 Mathematical Model of the Human Ear

Each point of the basilar membrane is more sensitive to a certain frequency, called characteristic frequency. At a specific point of the basilar membrane, the plot of its response to the frequency of vibration present in oval window is equivalent to

that of a low-pass filter with nearly constant quality factor. Therefore, the basilar fibers located at the high characteristic frequency sector respond to a larger band of frequencies than the fibers at the low characteristic frequency sector (CAMPOS NETO, 1993).

A similar behavior is observed for the response along the basilar membrane to a specific frequency tone, as showed in the Fig. 3. For each frequency, there is a point in the basilar membrane where the vibration reaches its higher intensity. That point position, measured from helicotrem, is closely proportional to the logarithm of the tone frequency. Around this point there is a zone of 1.5 millimeter where such vibration will be present.

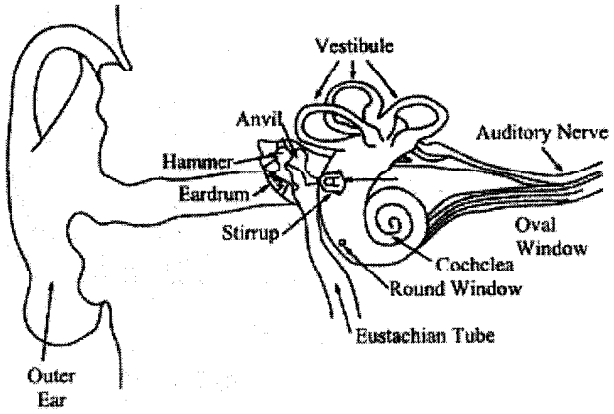

Figure 2 - Longitudinal section of the ear: main structures.

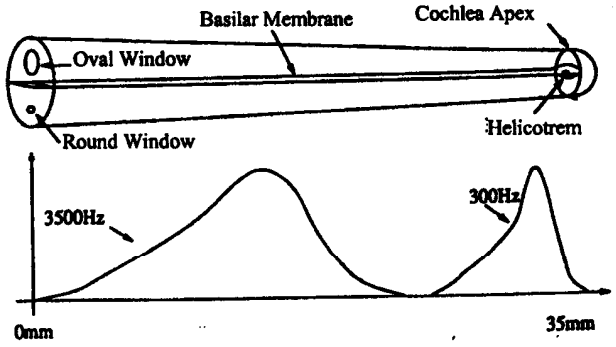

Figure 3 - Cochlea, basilar membrane and vibration of the membrane in two different frequencies.

An important aspect of the hearing is the masking phenomenon, where a tone becomes inaudible in the presence of another. Many masking phenomena can be explained in terms of frequency bands known as critical bands, which were determined by psycho-acoustic experiments. A critical band defines a region around a central frequency such that when two signals stimulate a critical band, the one with higher energy will dominate the perception and will mask the other one. So, two distinct tones are distinguishable only when they are in different critical bands. The resolution for distinguishing two frequencies varies from 100 Hz, for lower frequencies, to 6 kHz for higher frequencies.

For a better characterization of such critical bands, the frequency response showed in the Fig. 4 should be considered. This graphic specifies the relative power needed to hear a single tone as a function of the frequency. It is observed that the human hear is much more sensitive for the 2-4 kHz band, with the minimum situated at 3.3 kHz. The typical frequency range of the human speech is between 500 Hz and 2 kHz, where the lower frequencies correspond to the vowels and the higher frequencies to the consonants.

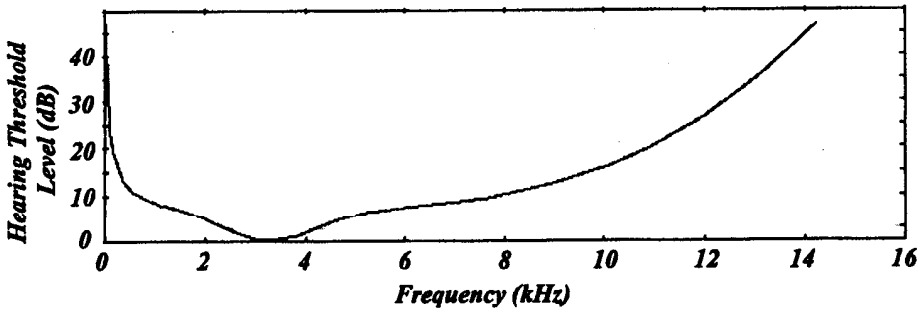

Figure 4 - Variation of the hearing threshold level.

The masking effect can be sensed if a simple experiment is performed: the ear is stimulated with a 1 kHz signal and, at same time, with another signal with variable frequency. In the Fig. 5, the dotted line represents the audibility of the variable frequency signal without the presence of the 1 kHz signal, while the straight line represents the audibility of the signal in the presence of the 1 kHz tone.

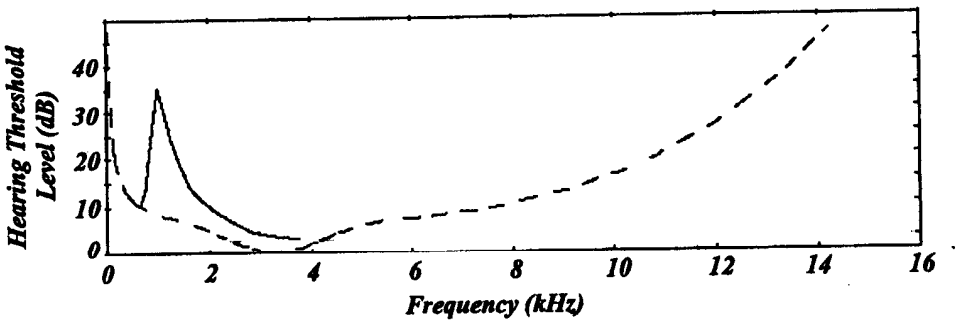

Figure 5 - Masking effects for 1 kHz.

It can be noted that near to 1 kHz, the signal is totally masked by the 1 kHz tone, since the relative power needed to hear it passed from 8 dB to 40 dB. So, it can be defined a "proximity area", where a signal masks all its neighbors, named, as commented earlier, critical bands. Nevertheless, the masking does not occur only in the frequency domain. Its effect can be also sensed in the time domain. After a high amplitude signal, it is necessary some time to be possible to hear a low amplitude one. An experiment can be performed: the ear is stimulated with a standard signal of 1 kHz and 60 dBA and, just after the end of the standard signal, the ear is stimulated by a test signal of, for example, 1.1 kHz. The time required to the recovering is about 200 ms.

Another curious phenomenon occurs in the human auditory system: a strong signal can mask a weak signal before the last one occurs. The explanation for this fact is the time that the brain needs to process the information. Nevertheless, the pre-masking time required is small (2 to 5 ms) in comparison with the post-masking time, which can reach 100 ms (BITTENCOURT, 1997).

It is important to observe that the critical bands can be defined around any central frequency. The bandwidth of each critical band corresponds to a uniform spacing of 1.5 mm along the basilar membrane, which corresponds to nearly 100 Hz for frequencies bellow 500 Hz and nearly 1/6 of the central frequency of the band for frequencies above 1000 Hz (FOURCIN, 1977). Then, the amplitude response in frequency, for each critical band, can be modeled as a low-pass filter with growing bandwidth with the frequency. Such filters have very pronounced cuts: 65 dB/octave for the critical bands around 500 Hz and 100 dB/octave around 8 kHz.

A perceptual scale, named Bark scale, was defined by Zwicker (ZWICKER, 1991). Such scale relates the acoustic frequencies to the perceptual frequency resolution, in such a way that 1 Bark covers one critical band. Then, a relation between the frequencies in Hertz (f) and the values in the Bark scale (b) was defined as:

$$b = 6\ln\left(\frac{f}{600} + \sqrt{\frac{f^2}{600^2} + 1}\right) \quad \text{and} \quad f = 600\,\text{senh}(b/6) \tag{1}$$

The shape of the critical bands, whose peaks are, by definition, spaced by 1 Bark, is named spreading function of the basilar membrane, and its formulation varies significantly from authors to authors: some of them approximate the curves by line segments, while others use analytical functions. Equation 2 shows an example of such formulation (HERMANSKY, 1990):

$$F(b) = \begin{cases} 0 & b - b_l < -1{,}3 \\ 10^{2{,}5(b-b_l+0{,}5)} & -1{,}3 < b - b_l < -0{,}5 \\ 1 & -0{,}5 < b - b_l < 0{,}5 \\ 10^{1{,}0[0{,}5-(b-b_l)]} & 0{,}5 < b - b_l < 2{,}5 \\ 0 & b - b_l > 2{,}5 \end{cases} \tag{2}$$

where l refers to the l-th critical band and b_l is the central frequency of such band, in Bark. As defined by this function, the crossing point of two adjacent critical bands occurs at the extremities of the plain regions. In some formulations, such crossing occurs at 3 dB below the peak value (NOLL, 1974).

The total number of critical bands depends on the band of the signals considered in the model. If a signal has a f_0 Hz bandwidth, then the number of critical bands will be given by:

$$B = 6\ln\left(\frac{f_0}{600} + \sqrt{\frac{(f_0)^2}{600^2} + 1}\right) \tag{3}$$

For bandwidths from 0 to 5 kHz, the Bark scale varies from 0 to 16.9 Bark. For the whole audible band, the considered band is between 0.5 and 24.5 Bark.

The pattern D(b), generated along the cochlea by a narrow-band signal, can be modeled as the convolution of the signal energy in the Bark domain, Y(b), with the spreading function of the basilar membrane, F(b), (ZWICKER, 1991):

$$D(b) = F(b) * Y(b) \tag{4}$$

The signal D(b) is called the excitation pattern and can be understood as the energy distribution along the basilar membrane.

The pattern generation for the voice signals is usually approximated by the same linear model of the Eq. (4), despite the complexity and non-linear characteristics of the voice.

Yet, it is necessary to consider that the critical bands have different gains, that is, the sensibility is not uniform along the basilar membrane, since the human auditory response, for frequencies higher than 5 kHz, falls down in a rate of 18 dB/ octave. The weighting of the excitation pattern by the sensibility of the basilar membrane is equivalent to the conversion from an intensity to an excitation (or loudness) scale, expressed in phonons. By definition, the loudness in phonons of a signal, with a certain level and frequency, is the intensity in dB of a 1 kHz tone that sounds like the signal.

That last conversion is based in experimental data, and it has been modeled as the filtering of the excitation pattern by the loudness correction function, E(b), given by the Eq. (5) (HERMANSKY, 1990):

$$P(b) = E(b) \cdot D(b) \tag{5}$$

where P(b) is the signal in phonons. An example of formulation for the function E(b), by convenience expressed in Hz, is given by the Eq. (6) (HERMANSKY, 1990):

$$E(f) = (2\pi)^2 \frac{(f^2 + 1200^2) f^4}{(f^2 + 400^2)(f^2 + 3100^2)} \tag{6}$$

Nevertheless, it was verified that the loudness scale is not linear when related to the loudness perceived by the human ear. That is, if a signal has loudness around 40 phonons, when 10 phonons are added to the signal, the perceived loudness will duplicate; however, if the signal is close to the hearing threshold, the perceived loudness will be 10 times stronger. So, it is necessary to transform this non-linear scale (phonons) into a linear scale (sonons). By definition, 1 sonon is the power increase that duplicates the perceived loudness. The conversion function from phonons to sonons is a non-linear warping function, which is approximated as (HERMANSKY, 1990):

$$L(b) = [P(b)]^{0,33} \tag{7}$$

where L(b) is the subjective loudness (perceived by the listener), contrasting to the objective loudness P(b). The subjective loudness L(b) represents the perceptual energy spectral density sent to the brain cortex. Then, some processing may be applied to L(b), such as linear prediction, perceptual spectral distance, perceptual cepstral density, etc., in order to get a distortion measure of a signal related to another one.

The perceptual modeling presented here is used in several objective methods, including the PSQM and the MOQV, as will be shown in the next sections.

2.2 Cognitive Processing

Referring again to the Fig. 1, the cognitive processing consists of all the processing performed after the calculation of the difference signal between the inner representations of the original and degraded signals. The quality of the degraded signal is calculated using that processed difference signal. For PSQM and MOQV methods, in particular, the mean level of that signal, calculated over the time, is directly related to the speech quality given by the codec.

The inaudible differences between the original and degraded signals, within the precision of representation and judgment, have no influence in the resulting score of the objective measure. Moreover, if the signals are identical, a perfect quality will be estimated, independently of the characteristics of the input signals.

3 Proposed Method

The proposed method is based on the structure of the PSQM method. The Figure 6 shows the basic scheme used in the MOQV method. The dotted lines indicate the main stages:

1 - Pre-processing.
2 - Frequency warping and mapping into critical bands.
3 - Telephone-band filtering and environmental noise.
4 - Intensity warping.
5 - Cognitive processing.

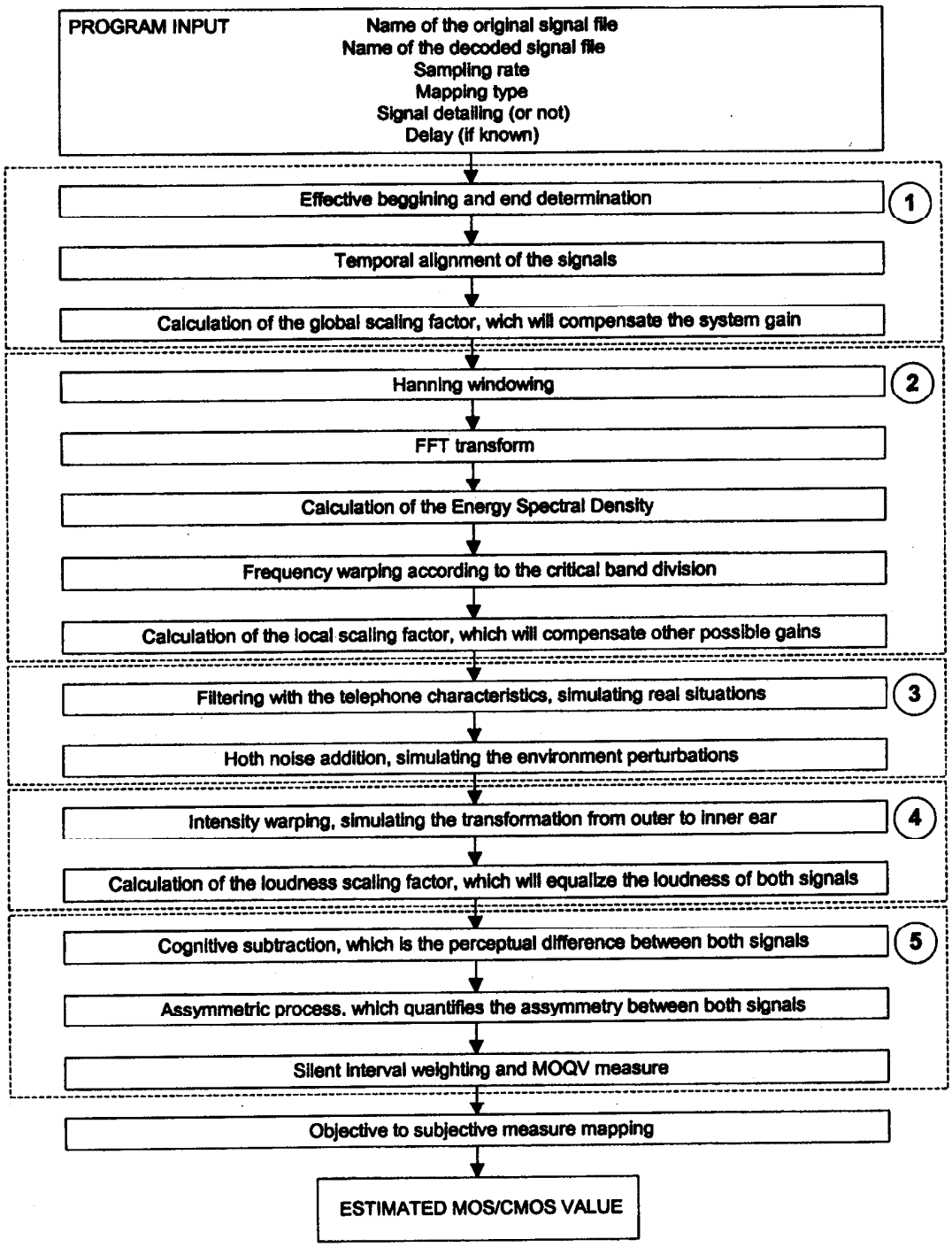

Figure 6. Scheme of the implemented program.

In the program beginning, the user must provide the name of the files corresponding to the original and degraded signals, the sampling rate (in Hertz), the option to get or not more details of the signals and, at last, the mapping options.

If the user chooses to have more details, the program will execute the original speech signal using the audio resources found in the computer; besides, both signals are plotted in the time domain, in order to be compared. At last, the signals levels in dB are presented. These levels can be changed using a third sub-routine, implemented separately from the main routine.

3.1 Pre-Processing

The detection of the effective beginning and end of the signals is performed by procedures standardized by the Rec. P.861. The beginning sample is the first whose magnitude, summed to the magnitudes of the four prior samples, is equal or superior to a determined value. In the same way, the final sample is that whose magnitude, summed to magnitudes of the four following samples, is equal or superior to a same determined value. The samples preceding the beginning and those following the end of the speech file are discarded.

The MOQV processing can be applied only to aligned signals. If the theoretic delay is not known, the temporal alignment between the signals is performed by cross-correlation implemented through a Fast Fourier Transform algorithm. The index of the maximum cross-correlation value represents the delay between both signals, and the alignment is automatically performed.

The global scaling for system gain compensation aims to adjust the energy level of the degraded signal. This is performed multiplying the degraded signal by a factor defined as the square root of the relation between the average energies of the original and degraded signals (BARBEDO, 2001).

3.2 Frequency Warping and Mapping into Critical Bands

The windowing divides the signals in frames of 256 or 512 samples, for a 8 kHz or 16 kHz sampling rates, respectively. There is a superposition of 50% between consecutive frames. Then, the FFT is calculated individually for each frame, by using a modified sub-routine, which is up to 10 times faster than the Matlab's conventional "fft" routine, since it considers only the relevant spectral components for the calculation of the MOQV measures. Then, the energy spectral density (ESD) is determined.

The ESD in the Hertz scale is warped into another scale, which is more appropriate to the inner representation, named *pitch energy density*. Such warping is performed in a short-term basis, and the new scale is based in the definition of critical bands. This mapping is slightly different from that described in the Section 2, since here each critical band is divided into 56 intervals or sub-bands corresponding to the division derived from the third-octave definition (INDUSTRIAL ACOUSTIC COMPANY, 1989). The mapping produces a sample of the ESD to each sub-band.

The last task in this stage of the processing is the local scaling, which aims to compensate some possible gain fluctuation. Only the audible time-frequency components are taken into account, that is, only the components that are above the absolute hearing threshold are taken into account in the calculation of the local scaling factor (ITU-T Rec. P.861, 1996).

3.3 Telephone-Band Filtering and Environmental Noise

The telephone-band filtering simulates the reception characteristics of the telephone handset (ITU-T Rec. P.830, 1996; ITU-T Rec. P.861, 1996).

The environmental noise is simulated by the addition of Hoth noise to the samples resulted from the telephone-band filtering. Such noise simulates the disturbance produced by sounds present in the reception environment.

3.4. Intensity Warping

The main objective of this stage is to simulate the transformation from the outer to the inner ear and also the subjective loudness generation. The intensity scale is mapped into a subjective loudness scale, originating a compressed loudness density function. The warping is performed using a non-linear compressing function proposed by Zwicker (ZWICKER, 1991). This function is generated plotting the typical pressure levels of each critical band and calculating the respective areas bellow the curve (MOORE, 1989; QUINLAN, 1992). This is not the only method for the calculation of the loudness density; several other approaches were proposed, as the A-method (STEVENS, 1972) or that described by the Eq. (1) to (7). The choice for the Zwicker's method was done after several comparative tests, which indicated this one as the most adequate to the desired application (ROYAL PTT, 1991; BEERENDS & STEMERDINK, 1994).

The loudness scaling is the last processing performed over each signal individually. Again, such scaling aims to compensate possible differences between both signals, now in the perceptual domain.

3.5 Cognitive Processing

The cognitive operations are those performed over the signal corresponding to the difference between the loudness densities of both signals. They are composed by the calculation of the difference signal, the subtraction of the inner cognitive noise, the calculation of the asymmetry factors and the weighting of the silent intervals.

The asymmetry is defined as the difference of quality degradation impression between the cases where the codec introduces strange components (where a major impact is produced) and the cases where the codec suppresses components (minor impact). After that, the silent frames are identified, through a silent threshold, and properly weighted. Finally, a mean value of degradation is calculated, which represents the MOQV value.

In this stage, one additional processing was incorporated in response to the fact that the method does not work well with speech signals containing temporal clipping and/or loud distortions. Such processing consists of an additional weighting factor that improves the performance of the method for those cases (KPN, 1997).

3.6 Mapping from Objective to Subjective Values

The method offers the following options for the mapping of the MOQV value:

1) mapping to MOS (Mean Opinion Score) or CMOS (Comparative Mean Opinion Score), which are standardized subjective measures (ITU-T Rec. P.830, 1996), the first one representing a score system that defines the quality of small groups of uncorrelated sentences and the second one representing a score system that provides a relative performance of a signal related to another reference signal;

2) mapping for the French, English and Japanese languages, or a generic mapping;

3) polynomial mapping whose order can be chosen from the 1st to the 6th, determined using a database containing several speech files.

The question of mapping will be treated with more details in the next section.

The next section will present further details about the mapping problem.

4 Detailing the MOQV Program

The original PSQM computer program was implemented in ANSI C language. This one was used as base to the development of the MOQV program, which was implemented in Matlab™ (version 5.3), in order to have more flexibility and versatility for the realization of modifications and tests. Such program offers several resources that are not found in PSQM: automatic temporal alignment of the original and degraded signals by cross correlation; several sampling rates options; observation of the speech signals characteristics, such as its temporal visualization, listening of its content and its level; signal level adjustment; several mapping options; and some additional techniques aiming to improve the performance (KPN, 1997; BARBEDO, 2001). Such characteristics produced an improved interactivity, turning the program application adequate either to experts in the area or to beginners.

4.1 Data Input

At the start of the program, the user must provide some parameters, as shown in the Tab. (1).

Table 1 - Options for the input of the program.

Parameters	Options	Oblig.	Default
Original File Name	-	Yes	-
Decoded File Name	-	Yes	-
Mapping Options	Subjective Measure	No	MOS
	Language		Generic
	Mapping		3rd Order
File Detailing	Yes (option 1)	No	No
	No (option 2)		
Sample Frequency	8,000 Hz	No	16.000 Hz
	16,000 Hz		
	32,000 Hz		
	64,000 Hz		
Delay	1 to 1000	No	Calculation by correlation

4.2 Mapping

There are 48 different mapping options in the program, which were determined empirically through the tests described in the next section. The user may choose one of those options or leave the program performs the default mapping, as shown in the Tab. (1). The choice obeys the structure map = *xyz*, as described in the Tab. (2). As an example, the default mapping is given by the map = 143, that is, it is a generic third-order mapping to the subjective measure MOS. This is the suggested option for speech files in Portuguese and other languages not present in the performed tests.

Table 2 - Structure for the choice of the desired mapping.

Parameter	Input	Meaning
Subjective Measure	x = 1	MOS
	x = 2	CMOS
Speech Files Language	y = 1	French
	y = 2	Japanese
	y = 3	English
	y = 4	Another Language (generic mapping)
Order of the Polynomial Mapping	z = 1	1st Order
	z = 2	2nd Order
	z = 3	3rd Order
	z = 4	4th Order
	z = 5	5th Order
	z = 6	6th Order

4.3 Values Returned by the Program Output

After all processing, the program will return the following values:

- the delay between the signals;
- objective measure MOQV1, corresponding to the processing without the additional weighting factor described in the Section 3.5;

- objective measure MOQV2, whose processing includes the additional weighting factor;
- subjective measure 1, obtained after the mapping of the MOQV1;
- subjective measure 2, after the mapping of the MOQV2;

These two last values may be considered or not, depending on the choice of the user, since the values MOQV1 and MOQV2 may be enough to the assessment of the codec.

4.4 Time of Processing

The program was tested in a Personal Computer with Intel Pentium™ III 650 MHz processor, with a RAM memory of 128 Mbytes and WindowsME™ operational system. Five different versions of the program were developed, each one containing modifications that improved their time of processing. The main modifications were the use of faster FFT routines, reduction of data storage and the replacement of the loops by matrix and vector operations. In this way, for signals containing 120,000 samples, the time of processing is 4.78 seconds, very close to that found for the PSQM program implemented in C language and tested in the Sun UltraSparc™ stations and Unix environment. To speech signals containing 30,000 samples, such time was less than 1 second.

5 Performance Analysis

5.1 Database Used in the Tests

All tests performed used the speech files available in the ITU-T S-23 database, which is composed by speech files in English, French and Japanese, associated to a number of codecs submitted to some test conditions (ITU-T, 1995). Each file was recorded in three different versions: original, pre-processed (filtered and equalized files) and degraded. For each test, the corresponding MOS or CMOS value is provided, which may be estimated my MOQV method. Such material is divided in three experiment groups:

- 1st Experiment: the speech files present in this experiment were submitted to several of ITU and mobile-telephony standard codecs. For each one of the three languages, two talkers of each sex were used, each one enunciating a sentence. There are 44 test conditions for each talker, totalizing 528 files to the experiment (176 by language).

- 2nd Experiment: the speech files were submitted to a number of environment noise types – office, vehicle, street, white and music, with a Signal-to-Noise relation of 10 or 20 dB. Two talkers of each sex were used for the first 28 conditions, and one of each sex for the last 12. Then, there are 40 conditions, totalizing 136 test files to each one of the 3 languages (totalizing 408 files to the whole experiment).

- 3rd Experiment: this experiment simulates the effects of the transmission of the coded signal through a communications channel that introduces random and burst frame errors. Two talkers of each sex were used, with 50 conditions for each one of them, totalizing 200 test files by language. Besides the previously cited languages, this experiment includes the Italian, resulting in a total of 800 test files.

5.2 Mapping

The mapping process from objective to subjective measures is based on Eq. (8) (GUTHRIE, 1999):

$$y_i = \beta_1 + \beta_2 \cdot x_i + \ldots + \beta_p \cdot x_i^{\,p} \qquad (8)$$

where x_i represents the objective measures and y_i represents the subjective measures. The b_i parameters are obtained minimizing:

$$Q = \sum_{i=1}^{n} \left[y_i - \left(\beta_1 + \beta_2 x_i + \ldots + \beta_p x_i^{\,p} \right) \right]^2 \qquad (9)$$

The calculation of the cross-correlation from the estimated to the corresponding actual subjective measures is realized by Eq. (10) (RIX & HOLLIER, 1998):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (10)$$

where $\bar{x}$ and $\bar{y}$ represent, respectively, the means of the estimated and actual subjective measures.

Therefore, values of r close to 1 indicate a good performance of the method and its mapping process, which is, to the case here treated, the estimated subjective values from the objective values are confident.

5.3 Kinds of Tests

A large number of situations was considered in the tests, in order to trace a profile of the performance of MOQV method. For that, it was considered:

- the characteristics of each experiment, as described in section 4.1;

- mapping for each language and the mean generic mapping related to all languages present in the database. The individual mapping results in a better correlation from objective to subjective measures, since this approach allows a better adaptation to the peculiarities of the corresponding language and culture. The results were composed with those obtained by Ericsson (FOLKESSON & KARLSSON, 1998); the second criterion produces a generic mapping, taking into account all the samples of all available languages, causing a decrease in the correlation value. Such approach is adopted by KPN Research (BEERENDS & HEKSTRA, 1998);

- the influence of the polynomial order p in the Eq. (9), allowing the determination of the best polynomial order to the mapping process.

5.4 Illustration of the Test Results

The tests produced a total of 156 plots, which describe graphically the performance of the MOQV for each considered experiment (BARBEDO, 2001; BARBEDO & LOPES, 2001). Some typical results are presented here to exemplify such results.

The Figure 7 illustrates the results obtained for the correlation between the MOQV values and the subjective measures for all languages available.

This figure was generated for the first experiment, using the second approach, that is, the generic third-order polynomial. In the horizontal axis are the MOQV values, while in the vertical axis are the actual subjective measures. The mapping function coefficients, obtained from Eq. (9), are shown in the inferior left corner. The correlation values for each language, obtained from Eq. (10), are presented in the superior right corner. The remaining plots present the same characteristics.

The Figure 8 illustrates the results obtained in the study of the most suitable value of the polynomial order p. Typically, a significant monotonicity break is observed for most of the situations when the polynomial order is greater than 3, which is not desirable. Besides, the gains obtained for the correlations become too little as p is made higher.

The Figure 9 illustrates a typical performance for each one of the experiments with generic third order polynomial mapping. As can be observed, the best performances were obtained for the first and second experiments. The correlations obtained for the third experiment were poor, as illustrated by the great spreading of the mapped values. Such observation indicates that the method is not appropriate for situations where the signals are corrupted by transmission errors. It is also important to note that the CMOS measure was used for the second experiment, since this is the measure available in the database for such experiment.

Furthermore, it was performed a statistical analysis, including histograms plotting, in order to describe the behavior of the deviation between the estimated and actual subjective measures. Additionally, a table with the maximum deviations was elaborated for some confidence intervals (BARBEDO, 2001).

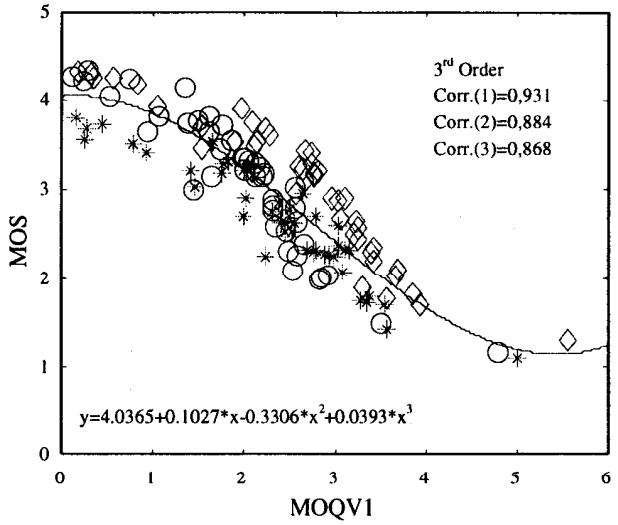

Figure 7. MOQV/MOS mapping.

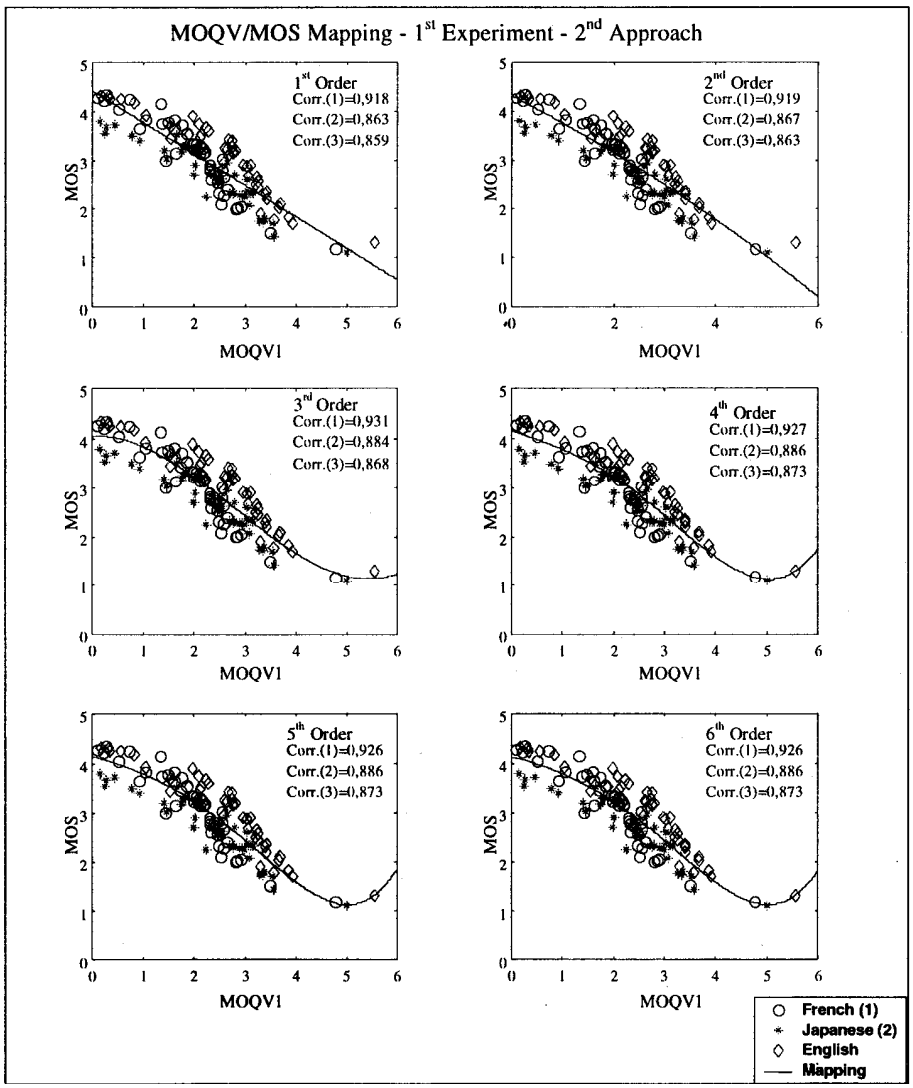

Figure 8 - Parameter *p* study.

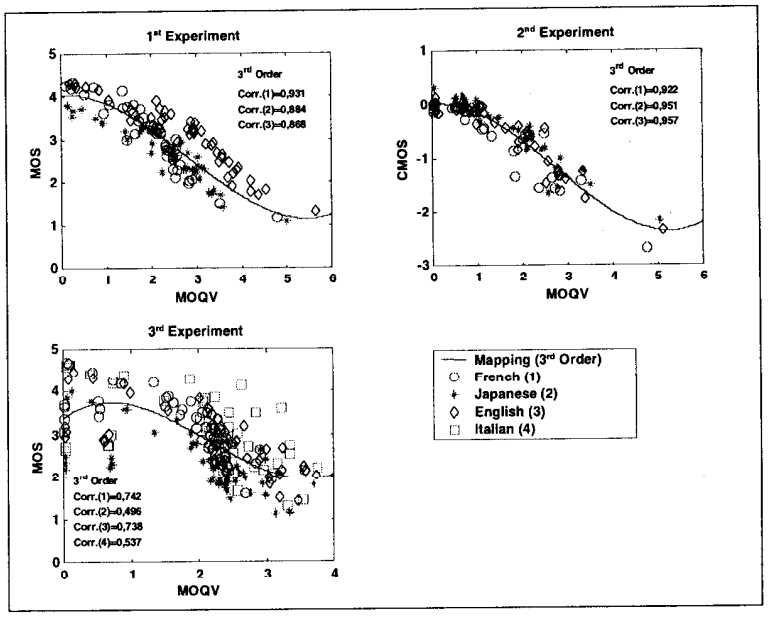

Figure 9 - Performance of MOQV method face to each experiment.

5.5 Scope

It is presented a resume of the conclusions resulting from the analysis of all the performed tests, setting the necessary information about the suitable use of MOQV method face to the several practical situations:

1 - The first approach (specific mapping) is the most adequate, since it is better adapted to the peculiarities of each language and culture. Ideally, each language should have its own mapping values. However, due to the database limitations, with files only in English, French and Japanese, a generic mapping was determined, using the all three languages, which can be used to the other languages, including Portuguese. Nevertheless, it is important to know that such mapping does not offer the best results. Then, it is suggested the realization of additional experiments in order to obtain the best mapping for the Portuguese. Alternatively, the analyses might be performed using the non-mapped MOQV values. Such option is available in the developed program and it is conditioned to the user expertise degree.

2 - The monotonicity is a desirable characteristic for this kind of mapping. The results show that the utilization of mapping orders greater than third it is not appropriate because they lead, in certain cases, to the lack of the monotonicity. The most appropriate mapping order for most of the cases is, therefore, the third. Such fact occurs due the capability of the third-order polynomials to model in a satisfactory manner, the approximate behavior of the listeners in subjective tests, that is, it represents well the tendency of loosing the linearity in the quality extremes (very clear or very degraded signal), since the listeners tend to saturate the assessment in such points. Minor monotonicity breaks were observed in such case, but the effect in the final result was not signifficant, since the rare tested cases that eventually are located in the small mapping zone where the monotonicity break occurs, will present mapped values very close to desired. The correlation gain obtained by passing from the first-order to the third-order reached 0.030 for the first and second experiments and 0.065 for the third. The gains obtained in the confidence interval analysis were also very expressive, reducing the bounds of deviation from correct values in more than 0.1 in many cases.

3 - It can be affirmed that MOQV worked well for the first and second experiments, showing that this method adapts well to cases where the signal is submitted to one or more kinds of codecs, as well as to situations where there is the presence of background noise. However, the obtained results for the third experiment were very poor, denoting a clear inadequacy to situations where the signals are subject to severe error conditions. The observed correlations were always insufficient, especially for the Japanese and Italian languages. However, taking into account all the cautions and restrictions inherent to such process, this method can be used, in such conditions, for the French and English. If the error conditions are not too severe (less than 3%), MOQV can also be used for the other languages, observing the limitations of such procedure, since the results will reveal only a performance tendency to the device under test, and not precise results. However, this alternative is conditioned to preliminary studies and also to the desired application to the assessed device.

4 - As commented before, it would be desirable an individual mapping for each language. An important consideration is the fact that the variation of mapping values from one language to another follows a certain pattern, which can be explained by the cultural characteristics of each country. Among the analysed languages, it can be clearly perceived that the Japanese listeners are the most exigent, followed by the French ones and, with a very lower degree of exigency, the North American ones.

5 - To determine an adequate mapping face for the Portuguese language peculiarities, in particular to that spoken in Brazil, it would be necessary a database with speech files and the corresponding subjective measures. Such data are not available yet due to the absence of a laboratorial structure to this purpose. The generic mapping, however, seems to be enough for most of the desired applications. Alternatively, one can use the MOQV value directly.

6 - In general, the constraints in the MOQV utilization depend strongly on the desired application for the assessed device, as well as the desired assessment quality. In less exigent applications, as military communications or VoIP (Voice under Internet Protocol), this program can be used without significant restrictions. Conversely, for more exigent applications, where the quality is a very important and a decisive factor and it is needed a more precise assessment, the use of subjective assessments may be necessary. In such cases, however, the MOQV method can be used to provide a tendency of the behavior of the tested devices, which can be important in situations where there are no enough time and/or resources to perform subjective tests.

6 Conclusion

This work presented a method named MOQV for the objective quality assessment of telephony systems and speech codecs. The method is based on the PSQM procedure (BEERENDS & STEMERDINK, 1994), and incorporates characteristics that improve its performance. The corresponding computer program incorporates signal manipulation resources, forming a complete product that makes the application of this method easier and quicker. It was described its basic characteristics and it was presented some illustrations of the results achieved through exhaustive tests using the S-23 database (ITU-T, 1995), which allowed the definition of the MOQV scope.

The general test results indicate that:

- the MOQV method works well when the signal is submitted to one or more kinds of codecs, as well as to situations where there is background noise. However it is inadequate in presence of transmission errors;

- the most appropriate polynomial mapping order is the third;

- when it is possible, an individual mapping must be performed for the specific language, in order to respect its peculiarities; if it is not possible, a generic mapping or the non-mapped MOQV value may be used;

- due the scarceness of data, it was not possible to determine a specific mapping for the Portuguese spoken in Brazil; for the most of the applications, however, the generic mapping is enough.

All the results are also valid for the PSQM method, indicating the necessity of additional studies and research, in order to amplify the scope and to improve the performance of both methods.

7 Acknowledgements

Thanks are due to the "Fundação de Apoio à Pesquisa do Estado de São Paulo" (Fapesp) for supporting this project (process 99/01702-0).

References

ATKINSON, D.J. *Proposed Annex to Recommendation P.861*, National Telecommunications and Information Administration (Washington, D.C.)/ International Telecommunication Union, Study Group 12 - Contribution COM 12-24-E, Dec. 1997.

BARBEDO, J.G.A. *Avaliação objetiva de qualidade de codecs de voz na faixa de telefonia. 2001. 117p.* Master's Thesis, (In Engineering) - State University of Campinas, Campinas.

BARBEDO, J.G.A.; LOPES, A. Proposta e avaliação de um método de medida objetiva de qualidade de codecs de voz. In: SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES, 19., 2001, Fortaleza. *Anais...* Fortaleza: SBrT / UNIFOR - UFC, 2001. 1 CD.

BEERENDS, J.G.; STEMERDINK, J. A. A perceptual speech-quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc.* New York, vol. 42, n. 3, p. 115 123, Mar. 1994.

BEERENDS, J.G.; HEKSTRA, A.P. *Comparison of the ITU-T P.861 PSQM, PSQM+ and MNB Objective Speech Quality Measurement Methods*, International Telecommunication Union, Study Group 12, Contribution COM 12-58-E, Sept. 1998.

BERGER, J. *TOSQA - Telecommunication Objective Speech Quality Assessment*, International Telecommunication Union, Study Group 12, Contribution COM12 34 E, Dec. 1997.

BITTENCOURT, R. *MPEG-Audio.* Disponível em: <http://www.1si.usp.br/~ricardo/mpeg/>. Acesso em 28 de setembro de 2001.

CAMPOS NETO, S. F. *Metodologias de avaliação de algoritmos de codificação de voz.* 1993. Master's Thesis (In Engineering) - State University of Campinas, Campinas.

FLETCHER, H. *Speech and hearing in communication.* Toronto: D. Van Nostrand Co., 1953, 487p.

FOLKESSON, M.; KARLSSON, A. *Results of processing the supplement 23 speech database for development of the extended P.861*, Ericsson, International Telecommunication Union, Study Group 12 – Delayed Contribution D.89, Nov. 1998.

FOURCIN, A.J. et al. Speech processing by man and machine: Group report. *Life Sciences Research Report.* Berlin, n. 5, p. 307-351, 1977.

GUTHRIE, W. F. *Engineering statistics handbook*, data analysis for process modeling. Disponível em: <http://www.itl.nist.gov/div898/handbook/>. Acesso em 10 de outubro de 2001.

GUYDON, A.C. *Fisiologia Humana.* 6. ed. Rio de Janeiro: Guanabara, 1988, 576p.

HERMANSKY, H. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America.* Melville, v. 87, n. 4, p. 88-94, Apr. 1990.

INDUSTRIAL ACOUSTIC COMPANY. *Noise control reference handbook.* New York: Industrial Acoustic Company, 1989.

ITU-T, International Telecommunication Union Telecommunication Standardization Sector. *Subjective test plan for characterization of an 8 kbit/s speech codec*, ITU-T Study Group 12 – Speech Quality Experts Group – Issue 2.0, 1995.

ITU-T, International Telecommunication Union Telecommunication Standardization Sector. *Recommendation P.830, Subjective performance assessment of telephone-band and wideband digital codecs*, 1996.

ITU-T, International Telecommunication Union Telecommunication Standardization Sector. *Recommendation P.861, Objective quality measurement of telephone band (300 - 3400 Hz) speech codecs*, 1996.

KPN RESEARCH. *Improvement of the P.861 Perceptual speech quality measure.* Netherlands, December 1997.

MACKAY, I.R.A. *Phonetics: the science of speech production.* Boston: College-Hill Publication Little, 1987.

MOORE, B. C. J. An *introduction to the psychology of hearing. 3. ed.* London: Academic Press, London, 1989.

NOLL, P. Adaptative quantizing in speech coding systems. In: INTERNATIONAL ZURICH SEMINAR ON DIGITAL COMMUNICATIONS. 1974, Zurich. *Anals...* Zurich: IEEE, 1974, p. B3.1 - B3.6.

QUINLAN, D. Subjective experiments on ISO 532 B. In: INTER - NOISE 92. 1992, Toronto. *Anals...* Toronto: The Inter-Noise series of International Congresses on Noise Control Engineering, 1992, v.1 xxxii+636 technical pages, v. 2 xxxii+628 technical pages.

RIX, A.; HOLLIER, M. *Performance metrics for objective quality assessment systems in telephony.* ITU Study Group 12 - Delayed Contribution D.79, Nov. 1998

ROYAL PTT. *Measuring the quality of audio devices.* Comité Consultatif International Téléphonique et Télégraphique. Contribution COM XII 114, Geneva, Dec. 1991.

STEVENS, S. S. Perceived level of noise by mark VII and decibels (E). *Journal of the Acoustical Society of America* - JASA, Melville, v. 51, n. 2, p. 575, 1972.

VORAN, S. Objective estimation of perceived speech quality - development of the measuring normalizing block technique. *IEEE Transactions on Speech and Audio Processing,* Piscataway, v. 7, n. 4, p. 371-390, July 1999.

WANG, S.; SEKEY, A.; GERSHO, A. An objective measure for predicting subjective quality of speech coders. *IEEE Communications Society Journal on Selected Areas in Communications.* Piscataway, v. 10, n. 5, p. 819-829, June 1992.

ZWICKER, E.; ZWICKER, U.T. Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system. *Journal of the Audio Engineering Society.* New York, v. 39, n. 3, p. 115-126, Mar. 1991.

Jayme Garcia Arnal Barbedo

He received the B.S. degree in Electrical Engineering from the Federal University of Mato Grosso do Sul in 1998, and the M.Sc. degree in Electrical Engineering from the State University of Campinas in 2001. Since 2001, he is Ph.D. student in the Department of Communications of the Electrical and Computer Engineering School of the State University of Campinas.

Amauri Lopes

Amauri Lopes received the B.S., the M. Sc. and the Ph.D. degrees in Electrical Engineering from the University of Campinas in 1972, 1974 and 1982, respectively. Since 1973 he has been with the Electrical and Computer Engineering School, University of Campinas, where he is currently an associate professor. His research areas are digital signal processing, circuit theory and digital communications.