A NEW METHOD FOR OBJECTIVE ASSESSMENT OF AUDIO QUALITY

Jayme G. A. Barbedo and Amauri Lopes

Abstract - **This paper presents the initial studies and the structure adopted in the development of a new method for objective assessment of audio quality, named Objective Measure of Audio Quality (Medida Objetiva da Qualidade de Audio - MOQA). New techniques are presented and their impact on the global performance of the method is analysed. The results are compared to that one reached by PEAQ method, which is currently adopted as standard by International Telecommunication Union.**

***Keywords* - Objective Methods, Audio Quality Assessment, MOQA, Neural Network**

I. INTRODUCTION

The digital transmission and storing of audio signals have been strongly based on algorithms for data reduction, which are adapted to several peculiarities of human auditory system, as the masking effects. Such algorithms do not necessarily aim the minimization of distortions. They do intend some manipulations of the audio signal in such a way that the users minimally perceive them. Therefore, the quality of the so-called perceptual coders cannot anymore be assessed by the traditional methods based on the global value of distortion, such as the signal-to-noise ratio (SNR) and total harmonic distortion (THD). In certain cases, the noisy structures are so effectively masked by the signal that they become nearly inaudible, even when the signal has a SNR as low as 13 dB.

In this way, the use of subjective tests is necessary to perform confident quality assessments of perceptual codecs. Nevertheless, such tests are expensive in terms of time and cost. So, the development of objective measures able to replace efficiently the subjective tests is highly desirable.

Some methods were proposed at the late seventies, but the first perceptual codecs (MPEG and Dolby) at the late eighties turned such measures obsolete. Then, in 1994, the ITU-R (*International Telecommunication Union - Radiocommunication*) performed an open call of proposals, in order to establish a standard for objective audio quality measurement. Six methods were proposed [1, 2, 3, 4, 5], none of them reaching the minimum acceptable performance. After that, the proponents concentrated their efforts in the development of a single method composed by the best former proposals, originating the method Perceptual Evaluation of Audio Quality (PEAQ) and a new recommendation, the ITU-R BS-1387 [6]. This method presents a clearly better performance than its predecessors. Nevertheless, it is not good enough for the most part of practical conditions. Such situation has motivated the search for new methods capable to overcome those limitations. In that context, a new method (MOQA), object of this paper, has been developed. More details about its implementation can be found in [7].

II. HUMAN HEARING

The most well succeeded methods of objective audio assessment are based on concepts extracted from psychoacoustics, which deals with the behavior of hearing. The human perception of sound can be roughly described through a five-stage scheme, as showed in Figure 1 [8]. The outer sound field is transmitted to the inner ear and separated into spectral components. The sensitivity of the ear and its spectral selectivity are improved by active processes, which normally include some kind of loop. The neural excitations in the inner ear are transmitted to the auditory areas of the brain by the auditory nerve, where they are translated into sensorial quantities. The auditory areas of the brain have several kinds of mechanisms that can influence the formation of sensorial quantities [9].

The first three stages of Figure 1 describe the translation of the outer sound field into the neural excitations (electrical impulses conducted by the neurons to the specific area of the brain cortex), and the two last describe the process of transformation of those excitation patterns into sensations. The translation of the outer sound field into the neural excitations is almost independent of personal preferences, and represents the part of hearing primarily based on the physiological structure of the auditory system. In a perceptual model, those steps are called "peripherical ear model". In the last stages of the hearing process, the individual preferences cannot

Jayme G. A. Barbedo e Amauri Lopes, Department of Communications, FEEC, UNICAMP, Campinas – SP, Tel: (19) 3788-3703; {jgab, amauri}@decom.fee.unicamp.br. This work was supported by Fapesp, Proc. n. 01/04144-0.

be clearly separated from the most common properties of the auditory system. Those stages, which include pattern recognition and hearing stream processes, are referred as cognitive model [8].

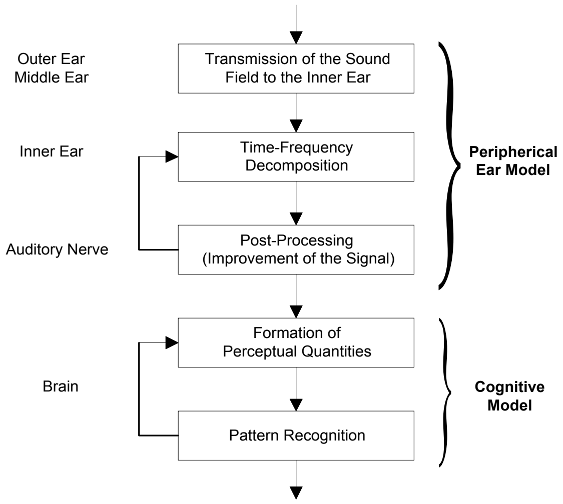
Outer Ear
Middle Ear
Transmission of the Sound Field to the Inner Ear
Inner Ear
Time-Frequency Decomposition
Peripherical Ear Model
Auditory Nerve
Post-Processing (Improvement of the Signal)
Formation of Perceptual Quantities
Brain
Cognitive Model
Pattern Recognition

Fig. 1 - Stages of the hearing process

III. Perceptual Measures

Figure 2 shows the basic structure common to all objective audio quality measures. Each block is briefly explained in the following.

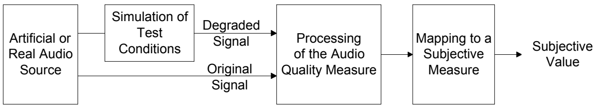
Artificial or Real Audio Source
Simulation of Test Conditions
Degraded Signal
Original Signal
Processing of the Audio Quality Measure
Mapping to a Subjective Measure
Subjective Value

Fig. 2 - Basic scheme of objective audio quality measures

- Artificial or real audio source: the test signals to be used are usually the same musical excerpts used in the subjective assessment of codecs. However, in principle any kind of audio signal, including the artificial ones, can be used.

- Simulation of test conditions: here, the test signal is submitted to conditions that may potentially introduce degradations, as several kinds of codification, bit errors, noise, or any other situation desired to be assessed; at same time, a unaltered version of the signal is kept for later comparison with the degraded version.

- Audio quality measure: this stage is the most important of any method for audio quality assessment; here are included the time-frequency decomposition, the modelling of the human hearing features (among them, the masking, briefly described in section 3.1) and the cognitive subtraction, which produces the perceptual difference among the signals. As result, a quality measure of the tested signal is obtained.

- Mapping to a subjective measure: this stage transforms the objective measure, represented in a particular objective scale, into a standard ITU subjective scale. This stage is optional and can be performed by polynomials or artificial neural networks.

A. Masking Modelling

Masking is the most important phenomenon in the quality perception of a signal. For that reason, its correct modelling is an essential factor in the performance of an objective method for audio assessment.

The masking phenomenon is due to ear limitations in terms of temporal, spectral and amplitude resolution, combined to an also limited dynamic range. When two signals are close enough to each other, in time or frequency domain, the weaker signal may become inaudible due the presence of the stronger one.

The modelling of masking effects is a feature common to all perceptual methods. The simultaneous (spectral) masking is always modelled by applying a spreading function, which corresponds to the shape of an average masking curve. Temporal masking effects are frequently implicitly modelled in the expressions of the model, but in a crude way, due to the limited temporal resolution of the time-frequency decomposition normally used.

IV. The MOQA Method

In this first version, the MOQA method borrowed several characteristics from the PEAQ method, as, for instance, its basic structure. As the research evolves, it is expected that both methods become more unrelated, since several new features should be implemented in next versions. Such novelties will include new strategies to calculate the model output parameters, the improvement of the psycho-acoustic model and some modification of the time-frequency decomposition.

Nevertheless, it is important to note that the version presented here has its own implementation, which has enough peculiarities and innovations to be considered as an original method. Furthermore, those new features represent important contributions towards a more efficient audio assessment method. Such new features will be detached in the following subsections.

A. General Structure

The general structure of MOQA method is shown in Figure 3, where the input signals correspond to the original signal, which will be taken as reference, and the degraded signal, which is the original signal submitted to some kind of condition capable to insert distortions.

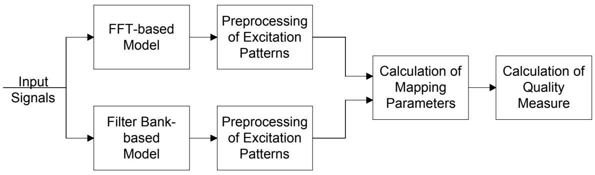
Input Signals
FFT-based Model
Filter Bank-based Model
Preprocessing of Excitation Patterns
Preprocessing of Excitation Patterns
Calculation of Mapping Parameters
Calculation of Quality Measure

Fig. 3 - General structure of MOQA method

As in PEAQ, two different models for the ear were implemented. The main distinctive characteristic of the MOQA models is the strategy adopted to perform the time-frequency decomposition (Fast Fourier Transform or Filter Bank). The models will be described with more details in the following, as well the processings indicated in Figure 3.

B. FFT-Based Model

The main feature of this model is the low computational burden. Its basic scheme is shown in Figure 4.

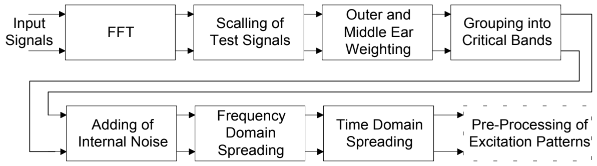
Input Signals
FFT
Scalling of Test Signals
Outer and Middle Ear Weighting
Grouping into Critical Bands
Adding of Internal Noise
Frequency Domain Spreading
Time Domain Spreading
Pre-Processing of Excitation Patterns

Fig. 4 - Basic scheme of FFT-based model

The inputs for this model, which are the original and degraded signals aligned in the time domain and sampled at a rate of 48 kHz, are divided into 42 milliseconds blocks (2048 samples), with a 50% superposition. After that, a Hanning window is applied.

Each windowed block is transformed to the frequency domain by a FFT algorithm. At last, each block is scaled to the playback level (if such level is unknown, it is recommended the adoption of 92 dB_{SPL}). A weighting function is applied to the spectral coefficients in order to model the frequency response of outer and middle ears.

The weighted spectral coefficients are grouped into critical bands and an offset is added to simulate the internal noise of the auditory system. The next step is to submit the signals to two spreading functions, the first one modeling the frequency domain masking and the second one modeling the time domain masking (see Figure 4). Such processing results in the so-called excitation patterns, which are submitted to some additional processing, as described latter.

In addition to the excitation patterns, another parameter, the error signal, is extracted at this stage. It is obtained after the weighting to model the frequency response of outer and middle ears, by calculating the difference between the power spectrums of the original and degraded signals. This difference is mapped into a perceptual scale by grouping into critical bands. This signal will play an important role in the calculation of the variables from which the objective measure is obtained through an artificial neural network.

There are some fundamental differences between the model here implemented and the one adopted by the PEAQ method. Among them, two are detached:

1- Use of a more efficient algorithm to the time-frequency decomposition, which reduces the needs for storage by 90%, making the program faster and more efficient;

2- The PEAQ method employs a normalization factor in order to keep the frame energies constant after the spreading performed in the frequency domain. However, such factor does not play its role efficiently and was replaced by a simpler procedure, where the relation between the energies before and after the spreading is computed for each frame. Then, the correspondent relation will multiply each frame submitted to the spreading. Furthermore, this procedure is computationally simpler.

C. Filter Bank-Based Model

The main feature of this model is its good temporal resolution, which allows one to obtain, theoretically, more precise results. On the other hand, the computational effort demanded is higher. Figure 5 shows the basic scheme adopted for this model.

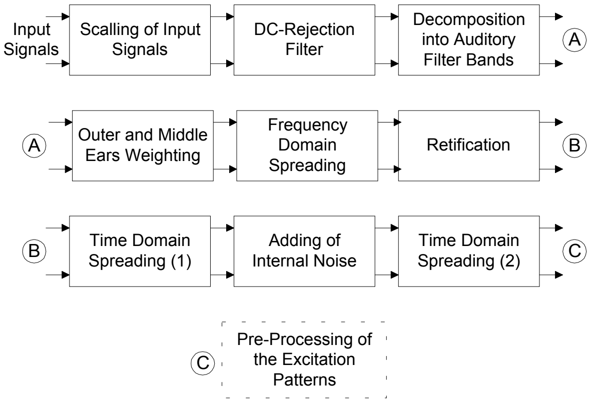
Input Signals
Scalling of Input Signals
DC-Rejection Filter
Decomposition into Auditory Filter Bands
A
A
Outer and Middle Ears Weighting
Frequency Domain Spreading
Retification
B
B
Time Domain Spreading (1)
Adding of Internal Noise
Time Domain Spreading (2)
C
C
Pre-Processing of the Excitation Patterns

Fig. 5 - Basic scheme of the filter bank-based model

The original and degraded signals at the input of this model are adjusted to the playback level and are sent through a high-pass filter to remove DC and subsonic components. Then, the signals are decomposed into 40 bands by linear-phase FIR filters, which are equally distributed across the perceptual scale. A frequency-dependent weighting is applied to the decomposed signal, in order to model the spectral features of outer and middle ears. The level-dependent spectral resolution of the input components to the auditory filters is modeled by a frequency-domain convolution of the outputs with a level-dependent spreading function.

The envelopes of the signals are calculated using the Hilbert-transform of the band pass signals (rectification) and a time domain convolution with a window function is computed in order to model backward masking. Then, a frequency dependent offset is added to take into account the internal noise in the auditory system and to model the threshold in silence. Finally, a second time-domain convolution is carried out using an exponential spreading function that take into account the forward masking.

The implementation of this model in the MOQA method also presents two major differences when related to that one used in the PEAQ. The first of them is the use of more efficient routines, in order to reduce the time required to execute the program. The second one, more important, is related to the implementation of the auditory filters. In PEAQ, the FIR filters are implemented recursively. This approach inserts a pole in the equations of the filters that must be canceled by a correct allocation of zeros. Therefore, although the filters still present a finite impulse response, their implementation is quite related to those ones used in IIR filters and reduces considerably the computational burden required.

Such approach led to very slow runs in Matlab®. Thus, the development of a structure able to limit the use of loops and better adapted to the peculiarities of the simulation environment was strongly recommended. After several attempts, a very efficient structure, which uses only matrix operations, was created [10]. This technique was at least five times faster than any other strategy tried here.

D. Pre-Processing of Excitation Patterns

This stage consists of four procedures aiming to prepare the excitation patterns for an adequate extraction of the output parameters:

1- Level and pattern adaptation: the average levels of the original and degraded signals are adapted to each other by filters and correction factors, in order to compensate level disparities and linear distortions.

2- Modulation: filters and weighting factors are applied in order to calculate a measure for the modulation of the envelope at each filter output. The resulting patterns are used to calculate the output parameters 1 and 2 in Section 4.5.

3- Loudness: this processing aims to determine the loudness of the resulting excitation patterns, in agreement to Zwicker's expression for the specific loudness [11]. The resulting patterns are also used in the calculation of some output parameters.

4- Masking threshold: it is obtained by the appropriate weighting of the excitation patterns, and it is used in the calculation of one output parameter.

E. Model Output Parameters

The model output parameters are submitted to an artificial neural network that produces a quality measure. Such parameters are described next, divided into groups in agreement to their purpose. Some of them were inspired in the PEAQ method, while others are completely new.

1- Modulation difference: it is calculated from the temporal envelopes of original and degraded signals. This group is composed by four parameters, three related to the FFT-based model and one related to the filter bank-based model.

2- Noise loudness: the parameters belonging to this group estimate the partial loudness of distortions added to the original signal. This group is composed by three output parameters, two from the filter bank-based model and one from the FFT-based model.

3- Bandwidth: the two parameters resulting from this stage provide an estimation of the average bandwidth of the original and degraded signals, in terms of FFT lines.

4- Noise-to-mask ratio: this group is composed by two parameters, one from each model, consisting on the relationship between the noise and masking patterns levels, in dB.

5- Relative number of disturbed frames: it is composed by only one output parameter deriving from the FFT-based model, and is given by the number of frames whose mask-to-noise ratio exceeds determined value in dB.

6- Detection probability: this group estimates the probability that a listener will detect a given disturbance. In PEAQ, it is composed by two parameters, both related to the FFT-based model. One of them was eliminated because its results are very poor. Furthermore, the other parameter was modified, leading to much better results than those ones obtained by the corresponding variable in PESQ.

7- Correlation between the stereo channels: the two output parameters of this group are new, and they are not found in any other objective quality assessment method. Their calculation is performed only for the filter bank-based model. The motivation derives from the observation that eventual phase shifts between the channels can be very annoying to the listener. Those shifts can be revealed by low correlation values.

8- Perceptual streaming and informational masking: these two concepts were published in [12] and were not used in the PEAQ algorithm. Here, they were combined to result in a new output parameter, which is derived from the filter bank-based model. The perceptual streaming is a central cognitive feature of the human auditory system that separates distinct auditory events and groups them into distinct streams. If the codec distorts the input signal in such a way that the output signal is separated into two pieces by the auditory system, the original signal and the distortion, then the disturbance caused by such distortion is more intense than if both parts (signal and distortion) are integrated in only one perception. The informational masking is a central cognitive feature of the human auditory system in which distortions that should be audible, become inaudible due the informational content (complexity) of the masker signal.

9- Loudness of the difference signal using L_p norms: this parameter uses a strategy adopted by the method PESQ [13] for quality assessment of speech signals. Some averages are calculated applying different norms, in order to emphasize determined features of the difference between the signals. Firstly, an averaging at the frequency domain using a L_3 norm is performed, what means that the spectral components are raised to 3, summed, and then a cube root is extracted. The same procedure is conducted at the time domain using a L_6 norm. As result, a value representing the loudness of the difference signal is obtained. This is also a new original parameter.

The mapping of all those parameters to a subjective quality estimation was performed using a multi-layer perceptron neural networks (MLPNN) with one hidden layer. The activation functions used for the hidden layer were hyperbolic tangents. For the output layer, the activation function was linear. The training was carried out using a Levenberg-Marquardt second-order optimization method [14], with an optimization criterion based on the least squares.

V. Tests and Results

The features of the tests and the results obtained are presented in the following.

A. Databases

Among the ten databases used in the validation of the PEAQ method [6], only three were available for this research, totalizing 239 pairs of files.

The files available contain a large number of distinct distortion patterns and a wide range content. Therefore, despite this is not a large set of files, it is representative enough to allow the extraction of consistent results and conclusions.

B. Tests Description

Firstly, the parameters were individually tested, and those ones judged not appropriated to be used as inputs of the artificial neural network were eliminated. From this selection, seven parameters from the FFT-based model and four parameters from the filter bank-based model remained.

Several configurations for the neural network were tested. The configurations were obtained changing two parameters: number of inputs for the neural network and number of neurons in the hidden layer, as described next.

- Parameters used as inputs to the neural network: the strategy to test the importance and contribution of each parameter consisted, initially, of tests using all the eleven remaining parameters as input to the net; then, they were gradually eliminated and, after each removal, the performance was computed. The parameters with lower individual correlation with the subjective scores were eliminated first. Tests showed that for four inputs or less, the performance of the method drops quickly. The best results were reached using 7 parameters. More studies will be necessary to turn the rest of parameters useful to train the neural network.

- Number of neurons in the hidden layer: the number of neurons was varied from 2 to 25; such tests revealed that, above six neurons, the correlations do not present a significant improvement.

Finally, two-thirds of the files were used in the trainings and one-third in the tests.

C. Results

The criterion used to validate the method was the correlation between the objective and subjective measures. The average correlation obtained for the three databases was 0.86. This can be considered an excellent result, especially if one considers that the best mean correlations reached by the PEAQ did not exceed 0.84 [6]. Figure 6 illustrates the performance of the MOQA method. A high concentration of points around the mapping line indicates good results.

Although the training and test sets for PEAQ and MOQA are different, the above comparison are meaningful because the databases used with MOQA have a range of conditions almost as wide as that one found in the ten databases used in PEAQ tests. Therefore, MOQA was tested in circumstances similar to that faced by PEAQ.

On the other hand, although the set used in this work contains a wide range of conditions, the set of data available to train the artificial neural network of the MOQA method was significantly smaller than that one available for the PEAQ tests. Therefore, the PEAQ method was much better trained and, consequently, was able to generate better mapping surfaces. In other words, the PEAQ had available more representatives examples of each condition were available to PEAQ, in such a way to provide more information and, consequently, more precise results.

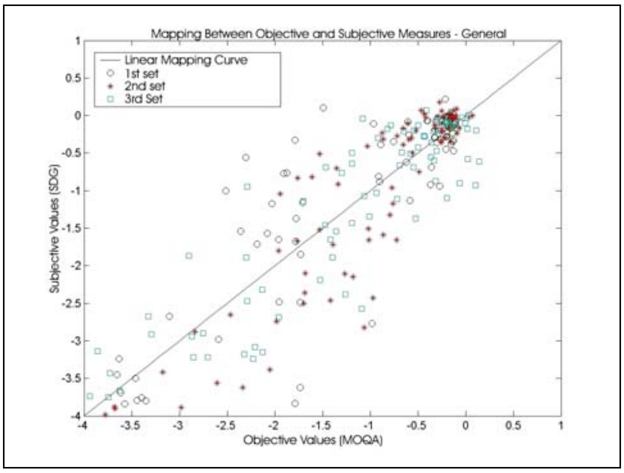

Fig. 6 - Mapping from objective to subjective values

Therefore, it is possible to say with a high degree of confidence that the MOQA reached a better performance than PEAQ. Additionally, if the complete set of data was available to the MOQA, it is very likely that it would reach even better results, since it could be better trained.

Finally, it is important to emphasize that the improvement reached by the MOQA is very significant, despite the little difference between the correlations of both methods. Most of the effort spent in the last years resulted in modest improvements [6]. Moreover, as the correlation gets closer to 1, it is more difficult is to get better results. In this context, even the slightest improvements are relevant.

VI. CONCLUSIONS

The paper presented a new method for objective audio quality assessment. This method is a first result of a research project aiming to improve the performance of the PEAQ method, currently the ITU standard for such kind of assessment.

This first proposal performs better than the PEAQ. However, although the achieved improvement be significant, its almost sure that it could be better if it was possible to train and test the new method with the same databases employed to adjust the PEAQ method.

New proposals to get further improvements are under test and efforts have been spent to acquire the PEAQ databases.

VII. BIBLIOGRAPHY

[1] T. V. Thiede, E. Kabot, "A New Perceptual Quality Measure for Bit Rate Reduced Audio", *Contribution to the 100th AES Conv.*, preprint 4280, Copenhagen, 1996.

[2] K. Brandenburg, "Evaluation of Quality for Audio Encoding at Low Bit Rates", *Contribution to the 82nd AES Convention*, preprint 2433, London, 1987.

[3] J. G. Beerends, J. A. Stemerdink, "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation", *Journal Audio Eng. Soc.*, v. 40, pp. 963-978, Dec. 1992.

[4] B. Paillard, P. Mabilleau, S. Morisette, J. Soumagne, "Perceval: Perceptual Evaluation of the Quality of Audio Signals", *J. Audio Eng. Soc.*, v. 40, pp. 21-31, Jan. 1992.

[5] C. Colomes, M. Lever, J. B. Rault, Y. F. Dehery, "A Perceptual Model Applied to Audio Bit-Rate Reduction", *J. Audio Eng. Soc.*, v. 43, pp. 233-240, April 1995.

[6] ITU-R Recommendation BS-1387, *Method for Objective Measurements of Perceived Audio Quality*, 1998.

[7] J. G. A. Barbedo, A. Lopes, "Innovations on the Objective Assessment of Audio Quality", *Contribution to the VII National AES Convention*, May 2003.

[8] T. V. Thiede, *Perceptual Audio Quality Assessment Using a Non-Linear Filter Bank*, Ph.D. Thesis, Berlin, 1999.

[9] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, Y. Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding", *Journal of the AES*, v. 45, pp. 789-814, October 1997.

[10] J. G. A. Barbedo, *1st Technical Report, Fapesp - Process no. 01/04144-0*, Campinas, July 2002.

[11] E. Zwicker, H. Fastl, *Psychoacoustics, Facts and Models*, Springer Verlag, Berlin, 1990.

[12] J. G. Beerends, W. A. C. van den Brink, "The Role of Informational Masking and Perceptual Streaming in the Measurement of Music Codec Quality", *Contribution to the 100th Convention of the Audio Engineering Society*, Preprint 4176, Copenhagen, May 1996.

[13] ITU-T Recommendation P.862, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2001.

[14] M. S. Bazaraa, H. D. Sherali, C. M. Shetty, *Nonlinear programming*, John Wiley & Sons, New York, 1993.