Objective Measure of Speech Quality in Channels with Variable Delay

Jayme Garcia Arnal Barbedo[1], Amauri Lopes[1]
Flávio Olmos Simões[2], Fernando Oscar Runstein[2]

[1] - Department of Communications - FEEC - UNICAMP - Campinas - SP - Brasil - Tel: (19) 3788-3703
{jgab, amauri}@decom.fee.unicamp.br
[2] - Centro de Pesquisa e Desenvolvimento em Telecomunicações - Campinas - SP - Brasil - Tel: (19) 3705-6693
{simoes, runstein}@cpqd.com.br

***Abstract* - This work deals with the objective quality assessment of speech signals in the telephone band. Particularly, presents a routine to detect, estimate and eliminate variable delay introduced during transmission and/or reception. This routine was integrated to an objective assessment method named MOQV (Objective Measure of Speech Quality), increasing its scope.**
The paper also presents the performance of the set MOQV plus routine, when applied to a database of CPqD Foundation. Such results are compared to those ones provided by MOQV without the routine and by PESQ, currently recommended by International Telecommunication Union (ITU).

***Keywords:* objective speech quality measures, variable delay, MOQV, PESQ.**

***Resumo* - Este trabalho trata da avaliação objetiva da qualidade de sinais de voz na faixa de telefonia. Em particular, apresenta-se uma rotina para detecção, estimação e eliminação de atrasos variáveis introduzidos durante a transmissão e/ou recepção. Esta rotina foi incorporada a um método objetivo de avaliação chamado MOQV (Medida Objetiva da Qualidade de Voz), aumentando sua aplicabilidade.**
Apresenta-se também o desempenho do conjunto MOQV mais rotina quando aplicado a uma base de dados da Fundação CPqD. Os resultados são comparados àqueles fornecidos pelo MOQV sem a rotina e pelo método PESQ, atualmente adotado como padrão pela International Telecommunication Union (ITU).

***Palavras-Chave:* medida objetiva de qualidade de voz, atraso variável, MOQV, PESQ.**

I. INTRODUCTION

The quality assessment of speech codecs is necessary to the development and homologation of those devices, as well as to the choice of the most adequate codec for an application. The quality assessment of telephony systems is necessary during implementation, vigilance and maintenance. Methods for this kind of assessment are in the context of telephone speech quality assessment.

Subjective quality measurement is still widely employed. However, its cost, complexity and time consumption have motivated the research into new objective methods to estimate the subjective quality of speech.

In this context, ITU-T presented two Recommendations: 1) Rec. P.861 [1] from 1996, introducing the PSQM method; 2) Rec. P.862 [2] from 2001, which presents the PESQ method. Those methods are based on psycho-acoustical models of human ear and estimate the quality score that would be reached if subjective tests were used.

PESQ represents a progress over PSQM, since it amplifies significantly the list of situations where PSQM can be applied. The most important improvement is the capability to assess the quality of signals that have been transmitted over systems that introduce variable delays. This condition has currently gotten an increasing importance due to the transmission of voice over TCP/IP, ATM networks and mobile systems. A second factor responsible for the evolution of PESQ is the employment of a more sophisticated perceptual model.

The description of PSQM in the Rec. P.861 is well enough detailed. The same does not occur with the description of PESQ in Rec. P.862. In particular, the description of the solution adopted for the identification and elimination of variable delays is rather succinct and does not provide enough information to allow an implementation of the algorithm. The same occurs with the description of PAMS, the first method capable to correct variable delays [3].

Taking into accounting this situation and the availability of MOQV [4], [5], similar to PSQM, a new routine was developed and integrated to MOQV, allowing its application to signals with variable delays. Additionally, the resulting method was optimised for Brazilian Portuguese, using a database created by CPqD Foundation. Such database was produced using the procedures described in Recommendations P.800 [6] and P.830 [7] of ITU-T.

This paper introduces the main features of the routine developed to identify and eliminate variable delays. It also presents the results reached with MOQV operating jointly with this routine. These results are compared with those achieved by original MOQV and PESQ.

II. THE MOQV ALGORITHM

The best objective methods employ a mathematical model for the human ear [4]. Figure 1 illustrates the basic characteristics common to such methods.

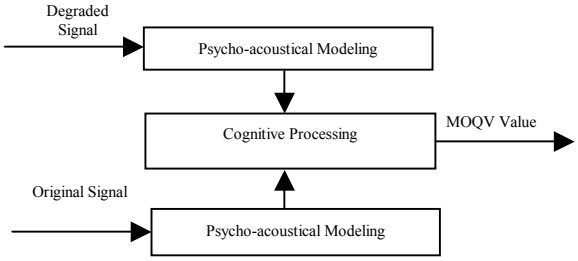

Figure 1 - Basic scheme of psycho-acoustical methods

The psycho-acoustical versions of the original and degraded signals are obtained by transforming the signals from time domain into frequency domain, mapping the spectral components into a psycho-acoustical scale and introducing the non-linearities of the ear in terms of the received acoustical loudness. Those transformations and processing are generically named psycho-acoustical modelling. The cognitive processing generates a signal corresponding to the difference between the original and the degraded psycho-acoustical versions of the speech signal. A quality measure of the degraded signal relative to the original signal is calculated from this difference.

This is the strategy adopted in MOQV [4], [5], whose development was based on the PSQM. MOQV achieves a performance comparable to PSQM, and presents several features that simplify and turn its utilization more flexible.

Analysing Figure 1, it is possible to conclude that any temporal misalignment between the input signals will produce artificial differences between them. Those differences will cause an incorrect reduction of the estimated quality. Therefore, it is necessary to eliminate any constant misalignment between the signals. Moreover, it is also necessary to eliminate the variable delays whose values are not large enough to produce subjective quality degradation. On the other hand, those variable delays that produce subjective degradation should not be eliminated because, in principle, they should be used to estimate the degradation. Nevertheless, no current objective method is able to estimate such degradation. Therefore, variable delays that produce subjectivity quality degradation are detected and eliminated.

The original MOQV is prepared to correct a constant misalignment between the signals, but is not able to eliminate variable delays.

III. ROUTINE FOR IDENTIFICATION AND ELIMINATION OF VARIABLE DELAYS

The proposed routine will be described in two steps: the first one, presented here, is a general overview; the second one, a detailed description, starts in Subsection III.1.

Starting the general overview, the first step of the proposed algorithm estimates and eliminates an average delay between the original and degraded signals. This is done calculating the cross-correlation between the envelopes of both signals as a function of the relative temporal shift between them. The delay estimate is the shift corresponding to the peak of the correlation. Once the average delay is estimated, the signals are aligned according to this estimative.

The second step aims to define segments of the signals where the delay is constant; next, such delay is eliminated. This is done through several processing stages. In the first one, the signals resulting from the correction of the average delay are divided into segments named utterances (according to a criterion that will be detailed later). Following that, each utterance is submitted to a process of delay elimination, now using a distinct procedure, which is more precise for estimating small delays. This technique employs a histogram, which provides a delay estimation and a correspondent confidence measure.

This estimate is used for the refined alignment of an utterance. After this adjustment, a test is performed. If the confidence measure is greater than 98%, the algorithm will deal with the next utterance, indicating that the previous utterance has constant delay, which was already corrected. Otherwise, the utterance is divided into two segments according to a rule that will be described later. Let A and B be those segments. Each part is treated as an individual utterance, and will be submitted to the tests described above. Taking part A first, if the test is positive, the algorithm begins to treat part B, indicating that A has constant delay, which was already corrected. Otherwise, A is divided into two segments and so on. This procedure continues with subdivisions of the divisions until positive tests end the process for a given utterance. Afterwards, the algorithm will treat the next utterance.

At the end of this processing, the signals are recomposed into one piece, and all delays present before the routine are eliminated.

The proposed routine will be now described in details. Figure 2 presents the basic functioning scheme of the method here suggested.

III.1. DETERMINATION OF ACTIVE AND SILENT SEGMENTS

The first step of this stage is the determination of the effective beginning and end of the speech signals as described in [4]. After that, the silence at the beginning and end of the speech files is eliminated.

A speech segment classifier based on neural networks is used to determine the active and silent segments of the resulting signals [8]. This routine splits the signal into 10 ms frames, extracts three different parameters from each of them (energy, number of zero crossings and autocorrelation) and combines them using a neural network, which determines if the segment is active voice or silence.

After each frame has been classified, the algorithm identifies the signal segments composed exclusively by active frames, here named active sections. In the same way, the segments composed exclusively by silent frames are the silent sections. The determination of the boundary of active and silent sections obeys the following rules:

- the first section is considered active speech, since the silent segment at the beginning of the signal was eliminated;
- the beginning of a silent section is the first sample of a frame classified as silence, whose previous frame has been classified as active speech and whose 9 subsequent frames have been classified as silence;
- the beginning of an active section is the first sample of a frame classified as silence, whose 9 previous frames have been classified as silence;
- the last section is considered active speech, since the silent segment at the end of the signal was removed.

The rules above intend to avoid that silent sections smaller than 100 ms be classified as silence, since such sections are naturally found during the spell of a sentence.

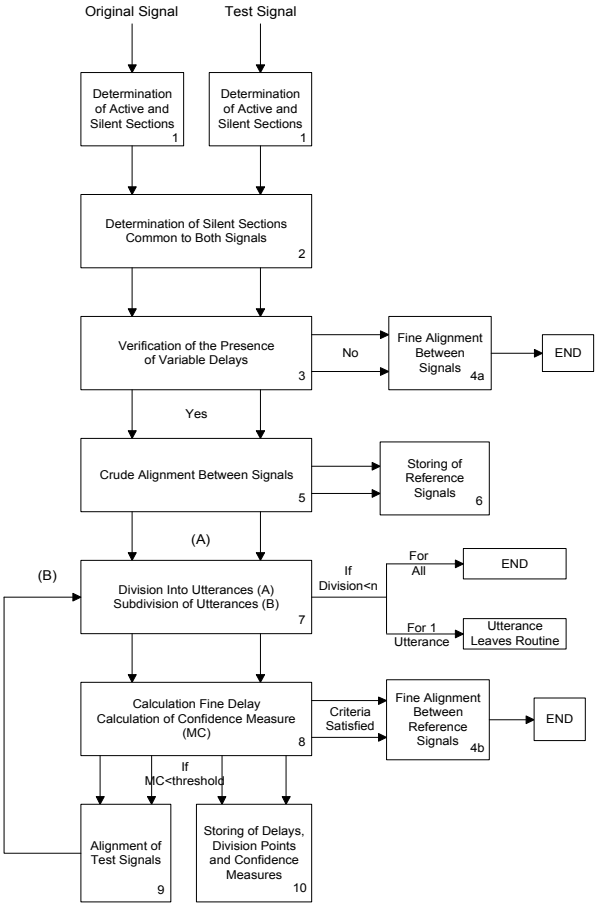

Figure 2 - General Scheme of the Routine for Elimination of Variable Delay (n = 3.072 samples).

III.2. VERIFICATION OF EXISTENCE OF VARIABLE DELAY

This stage avoids that signals containing only constant delay be submitted to posterior stages, which would represent an unnecessary waste of time.

Initially, the samples pertaining to a silent section with same index in both signals are identified and temporarily eliminated. After the execution of this stage, they are reintroduced at the exact point from where they were extracted. The objective of the exclusion is to avoid mistakes in the identification of the presence of variable delay, as will be described next.

The verification of the existence of variable delay is performed by a subroutine named *atrco*, widely used along the main routine. This subroutine is able to precisely estimate the delay between the signals or their segments, and it is detailed in Subsection III.5. The strategy employed here consists of calculating the delay between frames of 8,000 samples, selected in three specific points of both signals (the value 8,000 was chosen empirically as a compromise between estimate precision and typical length of the speech files standardized by ITU). The first frame begins 1,000 samples from the beginning of the signal, the second one is located at the middle of the signal and the last one ends 2,000 samples before the end of the signal. If the difference between any two delays obtained for those three frames is less than 1 ms, and all confidence measures (Subsection III.5) are greater than 0.5 (50%), then it is assumed that the delay is constant along the signals and its value is the average of the delays obtained for those three frames. In this case, the silent samples temporally eliminated are reintroduced, the signals are aligned using the average delay, and the algorithm is ended (stage 4a in Figure 2).

If the above conditions are not satisfied, then the signals have variable delay. In this case, they are submitted to the next stages of the routine, as follows.

III.3. ROUGH ALIGNMENT BETWEEN SIGNALS AND STORAGE OF REFERENCE SIGNALS

After variable delays have been detected, a first alignment of the signals is performed, using another subroutine. The subroutine used here is different from that in the last subsection, because it is simpler and more appropriate to the estimation of a rough average delay over the complete signals.

This subroutine splits the signals into 4 ms frames without superposition. Afterwards, it calculates the energy of each frame, aiming to extract their envelopes. Then, it calculates the cross-correlation between those envelopes as a function of the shift between them. The location of the cross-correlation peak is taken as the estimate of the delay. Finally, the signals are aligned using that delay estimate.

This subroutine is able to estimate delays with an 8 ms resolution. This crude procedure permits to reduce

the misalignment between whole signals. Considering that all delay estimates use, in some way, the cross-correlation, and since this calculation is as more precise as bigger is the number of samples common to both signals (or segments), a reduction on misalignment will allow more precise delay estimates.

After this crude alignment, the resulting signals are stored and named reference signals; they are the signals that will be actually aligned when all constant delay segments and the respective delays have been determined and, therefore, they are the signals that will be used in the rest of the method to estimate the quality. This storage is necessary because the signals used to estimate the delays are divided and submitted to several alignments, where they lose a number of samples. Therefore, they would not be adequate to be used in the rest of the routine for estimation of subjective quality.

III.4. Division of Signals Segments

After the crude alignment, the signals are subdivided into utterances (stage A in Figure 2) according to the following criteria:
- the first utterance begins at the beginning of the signal and ends at the half of the following silent section;
- the intermediary utterances begin at the half of a silent section and ends at the half of the following silent section;
- the last utterance begins at the half of the last silent section and continues until the end of the signal.

If the signal does not have silent segments, it will exist only one utterance embracing the whole signal.

Each utterance is submitted to a first test to verify its length. If it is larger than n = 3,072, the utterance will be analysed by next procedures to check the presence of delays. Otherwise, it is decided that the utterance is too small and that the rough alignment has corrected any possible delay; then, the algorithm starts to analyse the next utterance. This test is the first criterion to stop the processing of an utterance, and n = 3,072 is the minimum length that will be analysed aiming at the determination of an individual section delay. Practical tests showed that such value is enough even for cases with strongly variable delays.

There are other stop criteria, as will be described in the following. If none of them is satisfied, the utterance being tested is divided at its half (signals coming from block 9 in Figure 2). Each half is then treated as an utterance. This procedure is repeated until all subdivisions fulfil at least one of the stop criteria. Then, the algorithm begins to process the next utterance.

III.5. Calculation of Refined Delays and Confidence Measure

Refined delays and their respective confidence measures are calculated by the subroutine *atrco*, which was cited before. Initially, the signals or their segments are divided into frames using a Hanning window. The frame length, which was determined empirically, is 20 times the square root of the corresponding segment size. This variable length takes into account that large segments frequently exhibit large misalignments, whereas small segments presents small misalignments. This is so because each segment is subdivided before being submitted to the alignment process. Then, small segments frequently have already been submitted to some alignment procedure. As a consequence, the delay analysis of large segments demands larger set of samples than the smaller ones, since the correction is based on the correlation. Using frames with variable length has assured the high performance reached by this routine (see Section V). Another important parameter responsible for this performance is the high degree of superposition between the frames, which is equal to 87.5%, assuring a high number of points to construct the histogram.

After the division into frames, the cross-correlation for each frame is calculated using the strategy described in Subsection III.3. Thus, a histogram is constructed accordingly to the following criteria:
- the value and index of the maximum correlation is identified for each frame;
- the maximum value of the cross-correlation of each frame, raised to the power 0.125, is taken as a weight for each frame;
- the weights are grouped and summed accordingly to the corresponding index, generating the bars of a histogram.

Raising the maximum value of each cross-correlation to the power 0.125 concentrates the values around unit, avoiding an excessive domination of high cross-correlation frames.

The resulting histogram is then normalized by the sum of all weights, making its area equal to one.

In this stage, a confidence measure for the delay estimate to be obtained from the histogram is calculated. This value is defined as the percentage of the normalized histogram area that is concentrated up to 1 ms around the histogram maximum.

After the computation of the confidence measure, the histogram is convolved with a triangular window with duration of 1 ms and peak value 1. The convolution smoothes the histogram, attenuating isolated peaks and reinforcing the closely spaced ones. The delay estimate is given by the index of the resulting point of maximum.

The final delay estimate of a segment is obtained when one of the following criteria is satisfied:
- the confidence measure is greater than 0.98;
- the division of the analysed segment does not produce better confidence measure;
- the subdivisions have delay variations up to 5 samples when compared to the delay of the corresponding original section.

On the other hand, if a segment does not satisfy any of the above criteria, it is aligned according to the fine delay estimate (see block 9 in Figure 2) and subdivided into two segments (block 7). When all

segments of the signal satisfy at least one of those criteria, the reference signals are aligned according to the delays, the division points are stored (blocks 4b and 10 in Figure 2), and the routine is terminated.

III.6. ALIGNMENT OF TEST SIGNALS AND STORAGE OF DELAYS AND DIVISION POINTS

After each division and determination of corresponding delays, the test signals are repeatedly aligned. After this alignment, the corresponding delays are accumulated and stored, as illustrated in Figure 2. After all criteria have been satisfied, such delays, together with the division points that indicate the limits of each segment, are used in the final alignment of the reference signals (see 4b in Figure 2).

IV. TESTS RESULTS

The routine was tested using a database created by CPqD Foundation. This database is composed by the material used in the quality assessment of a commercial codec and by the corresponding subjective results. Twelve test conditions were analysed and twelve listeners were employed. Each listener heard two sentences per condition, assigning a grade for each sentence. Therefore, there are 24 grades for each condition, resulting in a total of 288 grades.

Some of the degraded signals of this database have a delay variation due to a continuous loss of samples along time. Such loss produces an advance of the degraded signal related to the original one. In some cases, it was observed a variation of 1,600 samples between the beginning and the end of the files (for files with 8 seconds and frequency sampling of 16 kHz). This phenomenon was not introduced by the tested codec, but by an equipment used in the test assembly. This kind of variable delay corresponds to the worst situation that the routine could face.

Although the phenomenon of continuous loss of samples is not significant in the context of subjective tests, the same does not occur in the context of objective assessment.

The files of the database were submitted to MOQV with the routine for variable delay elimination. Afterwards, the averages of the subjective assessments were determined, in order to generate only one value for each condition. The objective and subjective values were used to optimise a polynomial mapping from MOQV values to MOS (Mean Opinion Score – a subjective scale) values. The optimisation process resulted in a third-order monotonic curve with the coefficients [0.1520; -1.0333; 1.6242; 3.3717], where the first coefficient corresponds to the third-order term. Polynomial mappings with higher orders provide better correlations, but do not present monotonic behavior.

Table 1 shows the subjective quality values and the MOQV/MOS values provided by the optimised mapping. It also shows the MOQV values without the variable delay routine. The comparison between those two kinds of results clearly shows the influence of the routine.

Table 1: MOS results of PESQ and MOQV

Condition	Subj. MOS value	MOQV MOS with routine	MOQV MOS without routine	PESQ MOS
G711	4.104	4.097	2.137	4.200
G726	4.125	4.107	3.752	4.182
MNRU (Q=25dB)	4.021	4.109	3.764	4.058
MNRU (Q=15 dB)	3.021	3.111	2.541	3.218
LD-CELP (-26dBoV)	3.750	3.760	2.094	3.743
LD-CELP (-14 dBoV)	4.229	3.819	2.120	3.942
LD-CELP (-38 dBoV)	3.375	3.438	1.705	3.396
Street noise (*)	3.083	3.404	1.684	3.155
Street noise (**)	3.125	3.209	1.689	3.169
Office noise (*)	2.979	2.992	1.355	2.956
Office noise (**)	2.979	2.952	1.378	2.981
3 cascaded codecs	3.292	2.922	1.277	3.083

(*) Without Codec. (**) With Codec.

This table also allows comparing the previous results with those ones obtained by PESQ. It is concluded that the results obtained by MOQV with the variable delay compensation routine are competitive with those ones achieved by PESQ for the tested conditions.

The correlation between the actual and estimated subjective values was also calculated for each method. The correlation achieved by MOQV with the variable delay routine was 0.920, contrasting with the value of 0.450 obtained without such routine. The correlation achieved by PESQ was 0.966.

It was performed a detailed visual analysis of the alignment of the files after their submission to the routine. The results were excellent, with minimum deviations between both signals.

The new program was also tested with the constant delay files used in the development of original MOQV [4]. The results were identical to those ones obtained before the inclusion of the routine.

These last two observations allow to state that the proposed routine presents a performance at least as good as that one used in PESQ, because it provides an almost perfect alignment between the signals and does not produce side-effects when the signals do not present variable delay.

As a consequence of this observation, one can conclude that the better PESQ results are due solely to two factors: 1- the relative small number of subjective values of the CPqD database does not allow robust results, since even a little mistake in the estimate of the subjective quality can represent a significant difference in the correlation value; 2- the differences between PESQ and MOQV perceptual models.

It is important to highlight that the perceptual model used in MOQV is not the best available. Therefore, it is possible to improve the performance of both versions of MOQV (with or without the variable

delay compensation routine), what is currently being studied.

V. CONCLUSIONS

The proposed routine is an original and effective solution for the task of dividing the original and degraded signals into constant delay segments, and also for estimating such delays. Therefore, the objective to make MOQV robust for a wider range of practical situations was fully accomplished.

The MOQV with the variable delay compensation routine was applied to signals of a database from CPqD Foundation. The resulting performance is competitive when related to that one achieved by PESQ, which is currently the standard adopted by ITU-T. The quality of the alignment provided by the proposed routine allows one to state that the difference between the performances achieved by MOQV plus this routine and PESQ is only due to the difference between their perceptual models and to the limitations of the database used in the tests. Therefore, it is possible to improve the performance of MOQV plus the proposed routine through the improvement of its psycho-acoustical model.

VI. ACKNOWLEDGEMENTS

This work was supported by Centro de Pesquisa e Desenvolvimento em Telecomunicações (CPqD).

VII. BIBLIOGRAPHY

[1] ITU-T Recommendation P.861 (1996), *Objective quality measurement of telephone-band (300 – 3400 Hz) speech codecs.*

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

[3] Barbedo, J.G.A. "Objective Assessment of Speech Codecs in the Telephone Band" (in Portuguese), Master Dissertation, Unicamp, Campinas, July 2001.

[4] Barbedo, J.G.A., Lopes, A. "Proposal and valuation of an method for objective quality assessment of speech codecs" (in Portuguese), XIX Simpósio Brasileiro de Telecomunicações, SBrT 2001, Fortaleza, Brazil, artigo n. 001000000002200007, September 2001.

[5] ITU-T Recommendation P.800 (1996), *Methods for subjective determination of transmission quality.*

[6] ITU-T Recommendation P.830 (1996), *Subjective performance assessment of telephone-band and wideband digital codecs.*

[7] Rix, A.W.; Hollier, M.P.; Hekstra, A.P.; Beerends, J.G. "Perceptual Evaluation of Speech Quality (PESQ), the New ITU Standard for End-to-End Speech Quality Assessment: Part I – Time Alignment", submitted to Journal of Audio Engineering Society - June 2001.

[8] Ribeiro, M.V. "Techniques of Reconstruction of Packs Based on Wavelets and Neural Networks Applied to Waveform Coders in IP Telephony" (in Portuguese), Master Dissertation, Unicamp, Campinas, October 2001.

ABOUT THE AUTHORS

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.

Flávio Olmos Simões
He received the B.S. degree in Computer Engineering in 1996 and the M.S degree in Electrical Engineering in 1999, both from the State University of Campinas (UNICAMP). Since 1999 he is a researcher at CPqD (Centro de Pesquisa e Desenvolvimento em Telecomunicações). He works at the Audiovisual Comunication Technologies Division, in the area of speech and audio processing.

Fernando Oscar Runstein
He received the B.S. degree in Electrical Engineering from the National University of Córdoba, Argentina, in 1985, and the M.Sc. and Ph.D. degree in Electrical Engineering both from the State University of Campinas in 1990 and 1998, respectively. In 1994 he joined Telebrás (now CPqD Telecom & IT Solutions) where he is currently the coordinator of the speech processing group.