A NEW METHOD FOR OBJECTIVE ASSESSMENT OF SPEECH QUALITY

Jayme Garcia Arnal Barbedo, Amauri Lopes, Flavio Olmos Simões and Fernando Oscar Runstein

Abstract - This paper presents a new method for the objective quality assessment of speech signals, named Objective Assessment of Speech Signals (OASS). The proposal represents an improvement of a previous method (MOQV), with two new features: 1) a routine to detect and correct variable delays between the original and degraded signals; 2) a new psycho-acoustic model for the ear. The paper details the new routine to eliminate variable delays, the basic structure of the signal quality measure, the improvements introduced in the psycho-acoustical model and a new procedure to extract the cognitive difference between the signals. The paper also describes the tests performed to validate the method and respective performance parameters. OASS is compared to MOQV and also to PESQ, which is currently adopted as standard by International Telecommunication Union (ITU).

Palavras-chave: objective speech quality measures, variable delay, psycho-acoustic model, cognitive difference, polynomial mapping.

Resumo - Este artigo apresenta um novo método para a avaliação objetiva da qualidade de sinais de voz, denominado Avaliação Objetiva de Sinais de Voz (OASS). A proposta representa um aperfeiçoamento de um método anterior (MOQV), com a incorporação de duas novidades: 1) uma rotina para detectar e corrigir atrasos variáveis entre os sinais originais e degradados; 2) um novo modelo psicoacústico para a modelagem do ouvido. O artigo detalha a nova rotina para eliminar atrasos variáveis, a estrutura básica do processo de medição da qualidade dos sinais, os aperfeiçoamentos introduzidos no modelo psicoacústico e um novo procedimento para extração da diferença cognitiva entre os sinais. O artigo descreve também os testes realizados para validar o método e respectivos parâmetros de desempenho. O método OASS é comparado aos métodos MOQV e PESQ, este último adotado atualmente como padrão pela International Telecommunication Union (ITU).

Keywords: medidas objetivas de qualidade de voz, atraso variável, modelo psicoacústico, diferença cognitiva, mapeamento polinomial.

Jayme Garcia Arnal Barbedo and Amauri Lopes are with Department of Communication of the School of Electrical and Computer Engineering of the State University of Campinas (Unicamp), Caixa Postal 6101, CEP: 13.083-970, Campinas - SP - Brazil. Flávio Olmos Simões and Fernando Oscar Runstein are with Fundação Centro de Pesquisa e Desenvolvimento em Telecomunicações, Campinas - SP - Brazil. E-mails: jgab@decom.fee.unicamp.br, amauri@decom.fee.unicamp.br, simoes@cpqd.com.br, runstein@cpqd.com.br.

1. INTRODUCTION

The development of the digital signal processing techniques and technology has motivated a growing interest in more efficient voice coding/decoding methods and devices. One of the most important aspects in the development of such devices is their quality assessment. Among the features to be assessed, the perceived speech quality is particularly significant.

Classical objective measures for quality assessment of speech signals, such as error rate and signal-to-noise ratio, do not exhibit high correlations with the sensibility of telecommunication system users. Therefore, subjective quality measures are still widely employed. However, their cost, complexity and time investment has motivated the search for new objective methods to estimate the subjective quality.

In this context, a number of proposals were presented aiming an efficient model for the behavior of human listeners in a subjective test. Some of these methods obtained relative success: Perceptual Speech Quality Measure (PSQM) [1], the former International Telecommunication Union (ITU) standard [2] and still largely used; Perceptual Analysis Measurement System (PAMS) [3], the first one capable to take into account variable delays between original and degraded signals; and Perceptual Evaluation of Speech Quality (PESQ), the standard currently adopted by ITU-T [4].

Despite the great evolution observed in the last years, no method succeeded in modeling all kinds of practical situations until now, justifying the search for new techniques that allow objective measures completely replace the subjective measures. Such situation lead to the development, in 2001, of the Objective Measure of Speech Quality method (Medida Objetiva da Qualidade de Voz – MOQV) [5,6], which is based in the PSQM structure, and includes some new features making it more complete and versatile. Later, some new improvements were introduced, most of them related to the use of neural networks to perform the mapping between objective and subjective values [7,8,9]. Those techniques use artificial neural network to provide a better fitted non-linear mapping, reaching excellent results. However, the robustness of the method is reduced when faced to untrained or unknown conditions and, therefore, it is necessary wide training sets. This situation has motivated the search for a different strategy able to provide excellent performance and to keep robustness at same time.

The efforts to achieve such objectives originated a new method, the OASS, object of this paper. The solution was the adoption of a more robust perceptual model together with a polynomial mapping. This mapping is classical, but avoids the disadvantages of the specialization inherent to the artificial neural networks. Additionally, the quality of

this new model has ensured good results. Finally, OASS also incorporates a routine to detect and eliminate variable delays, which have grown in importance due to the increasing of voice transmissions through packet-based and mobile systems.

The OASS routine to compensate variable delays was inspired on those ones used in PAMS and PESQ, but its originality is guaranteed since no formulas or equations are presented in any of the publications related to such methods. The same is true with respect to the psycho-acoustic model. Therefore, an intuitive translation from words to equations was necessary. This translation was performed in such a way that the resulting routines could be incorporated in MOQV without major changes in its structure. This procedure avoided time-consuming adjustments, and good results were reached quickly.

In addition to an excellent performance, as presented in Section 5, OASS offers the option to use a structure optimized for the Portuguese spoken in Brazil, making this method an effective alternative to PESQ.

This paper is organized as follow: Section 2 presents some basic features common to all perceptual speech assessment methods; Section 3 presents some general observations for a better understanding of the rest of the text; Section 4 describes the routine to compensate variable delays; the psycho-acoustical model of OASS is presented in Section 5; Section 6 presents tests, results and a comparison of OASS with other methods; some main conclusions are presented in Section 7.

2. BASIC STRUCTURE OF PERCEPTUAL METHODS

The basic structure common to all methods for objective assessment of speech quality is shown in Figure 1. The last block, representing the mapping from objective to subjective values, is optional.

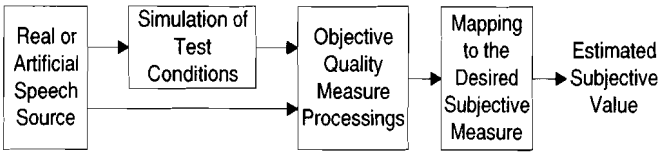


Figure 1. Objective speech quality measures: basic scheme.

The most well succeeded methods for objective speech 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 the five-stage scheme in Figure 2.

The first three stages of Figure 2 describe the translation of the outer sound field into electrical impulses in the inner ear. This translation divides the outer sound field into spectral components. The level sensitivity and the spectral selectivity are improved by active processes, which normally include some kind of loop.

The two last stages of Figure 2 describe the process of transformation of those excitation patterns (electric impulses) into sensations. The auditory nerve and neurons transmit the electrical impulses to auditory areas of brain, where they are translated into sensorial quantities. The auditory areas have several mechanisms that can influence the formation of sensorial quantities [10].

The translation of outer sound field into 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 "peripheral ear model". In the last stages of the hearing process, individual preferences cannot be clearly separated from the properties of the auditory system. Those stages, which include pattern recognition and hearing stream processes, are referred as cognitive model.

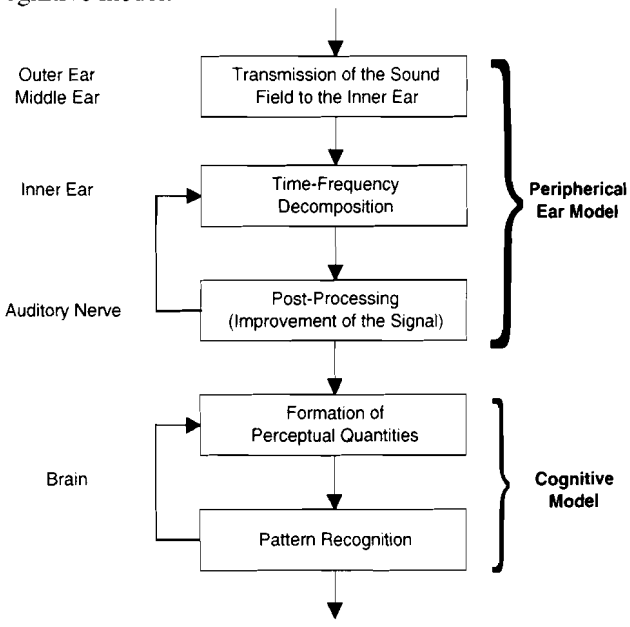


Figure 2. Stages of the hearing process.

The performance of any method for objective assessment of speech quality is closely related to the quality of the representation of such stages. The psycho-acoustical model of OASS follows as close as possible the scheme shown by Figure 2, and it is in line with the latest advances in human hearing modeling.

3. GENERAL OBSERVATIONS

Before start describing the OASS method, it is important to highlight some assumptions and procedures, in order to ease the understanding of the text. Such observations are divided into two subsections.

3.1 SAMPLING FREQUENCY

All the strategies presented in the following sections were developed assuming signals with a sampling frequency of 16 kHz. In the case of signals sampled at other rates, the OASS routine executes the proper decimation or interpolation, in order to force the signal to have a sampling frequency of 16 kHz. It is important to emphasize that the decimation of higher rates signals does not imply in loss of information, since OASS works only with narrowband speech signals (300 HZ – 3400 Hz), and then Nyquist theorem will always be respected.

3.2 EMPIRICAL VALUES

Several values used along the text were empirically optimized. They were determined using a group of reference signals, which was carefully chosen to be as representative as possible, containing most of the conditions usually found in speech signals. The empirically obtained values were then tested with speech signals not used in the original database. This procedure was used to validate or reject them. Therefore, the values in the text that have not been particularly explained were determined using the approach described in this section. Values obtained using more complex strategies are described at the point they appear for the first time.

4. VARIABLE DELAY ROUTINE

An important limitation found in most of the objective speech quality assessment methods, including MOQV, is their inability to deal with variable delay. The increasing of packet-based transmissions and mobile communications, which introduce this kind of delay, motivated the development of methods capable to deal with variable delays.

First of all, it is necessary to consider two classes of variable delays: 1) those that cause degradation of the subjective perception of the signals; 2) those that do not cause subjective degradation but that degrade the performance of objective methods. The variable delays of the first class are not eliminated in order to avoid misevaluation of the subjective quality estimate. On the other hand, those in the second class must be eliminated, because the temporal misalignment between original and degraded signals causes an artificial growth of the difference between them. Since objective methods use this difference to estimate the subjective quality, the artificial growth corrupts the estimate. An algorithm designed to deal with variable delays must perform the selection of delays that must be eliminated.

The first method able to deal with variable delays was PAMS [3]. Its delay identifier was also employed in the PESQ method [4]. The lack of information about the characteristics of this identifier has motivated the development of a new solution, which is described next. Figure 3 presents the basic scheme of the proposed procedure.

An overview of the routine is described in the following. The details are described after that.

The first step of the algorithm estimates and eliminates any average delay between original and degraded signals. This is done calculating the cross-correlation between the envelopes of both signals as a function of the relative shift between them. The delay estimate is the displacement corresponding to the peak of correlation. After that, the signals are aligned accordingly.

The second step identifies the segments of the signals where the delay is constant, estimates these constant delays and eliminates them. To do this, the algorithm divides the signals resulting from the first delay correction into segments, called utterances, accordingly to some criteria that will be discussed later. Each utterance is submitted to a new delay elimination procedure, able to perform precise estimates of slight delays. This procedure employs a histogram, which provides a delay estimate and corresponding confidence measure.

Such estimate is used in the fine alignment of an utterance. After this adjustment, a test is performed. If the confidence measure is greater than 98%, it is decided that the utterance had a constant delay and that it was already eliminated. Then, the algorithm goes to the next utterance. Otherwise, the utterance contains variable delays. Thus, it is split into two parts according to a given criterion. Each of both parts is treated in the same way as the utterance, until a stop criterion has been satisfied, as will be described in the following. After this procedure, all delays within the utterance were eliminated and the algorithm goes to the next utterance, until the whole signals are aligned.

The following Subsections details all the procedures used in the routine, according to Figure 3.

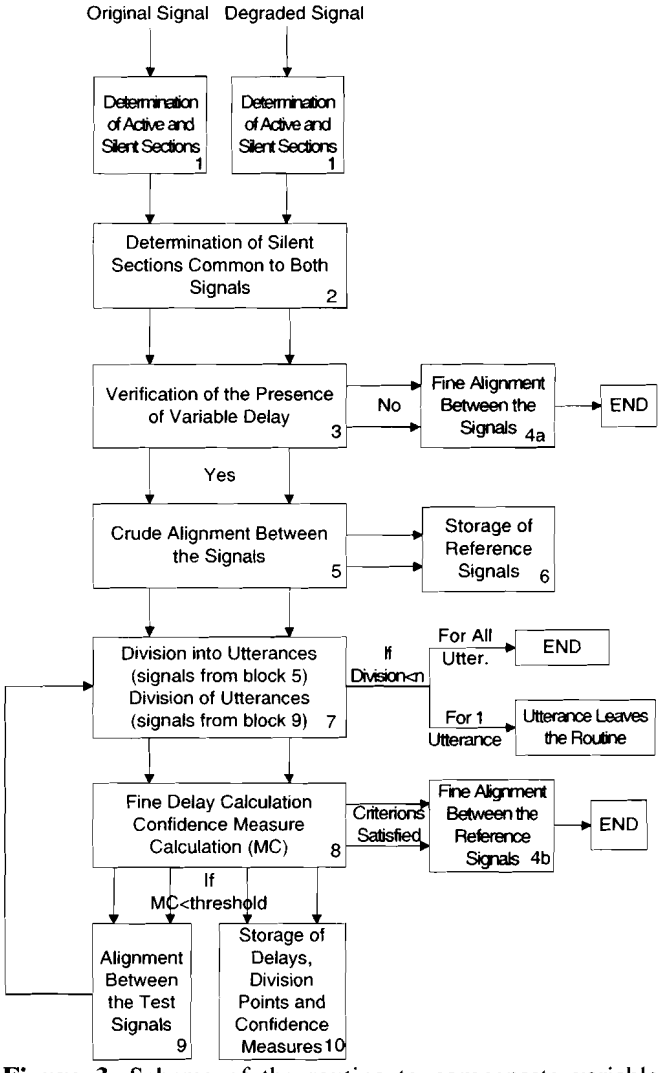


Figure 3. Scheme of the routine to compensate variable delays (n = 3,072 samples).

4.1 DETERMINATION OF ACTIVE AND SILENT SECTIONS

A speech classifier based on artificial neural networks determines the silent and active sections of the resulting signals [11]. This routine splits each signal into 10 ms

frames and, for each frame, extracts three parameters: energy, number of zero crossings and autocorrelation. These parameters are combined using a MLP neural network with one hidden layer, which determines if the frame is active or silent. The effectiveness of this technique was very high (around 99 %) for signals with low and medium levels of degradation; for very noisy signals, the rate of errors varies considerably, depending on the type and intensity of the noise. Since the signals used in subjective and objective quality assessments seldom have such levels of degradation, this neural speech classifier can be used without major concerns.

Based on those classifications, the next step determines the limits of the silent and active sections of the signals, according to the following rules:

- the first section is active speech, since the silent at the beginning of the signal must be eliminated according to the procedure described in Section 5.1;
- the beginning of a silent section is the first sample of a frame classified as silence whose prior frame was classified as active and whose 9 posterior frames were classified as silence;
- the beginning of an active section is the first sample of a frame classified as active whose 9 prior frames were classified as silence;
- the last section is active speech, since the silent section at the end of the signal must be discarded according to the procedure described in Section 5.1.

Those rules prevent very small silent sections (less than 100 ms) from being classified as silence. Otherwise they would lead to an excessive number of silent sections, prejudicing the operation of the routine. Moreover, silent periods less than 100 ms are naturally found during a spelling. Therefore, they can be considered as integral part of active speech itself.

4.2 DETERMINATION OF SILENT SECTIONS COMMON TO BOTH SIGNALS

This stage aims to determine components of silent sections that are common to both signals. As can be seen in Figure 4, the signals are compared and silent components with same index in both signals are identified and temporarily eliminated before the execution of next stage. After that, they are reintroduced at the exact point they were extracted. Such procedure aims to avoid false identification of variable delays, as will be discussed in the following Subsection.

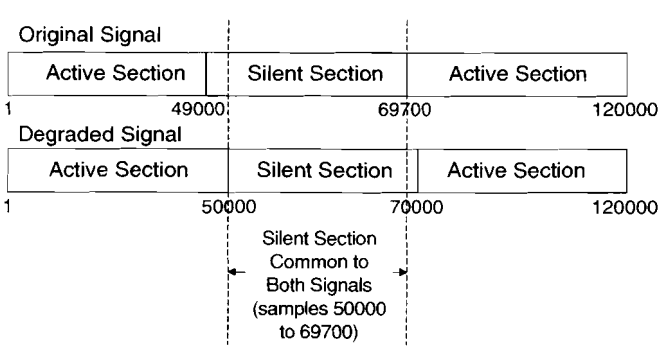


Figure 4 - Determination of silent components common to both signals.

4.3 VERIFICATION OF THE PRESENCE OF VARIABLE DELAY

Despite the increasing number of devices capable to introduce variable delays, most of the systems and codecs still produce constant delay. Therefore, most of the speech signals to be tested contain only invariable delay. Although the proposed routine is able to process such signals, this stage avoids submitting them to posterior stages, what would represent an unnecessary waste of time. On the other hand, the time added by this stage is negligible.

The analysis of the delays is based on a subroutine largely used along the main routine. This subroutine is described in details in Section 4.7, where it is shown that it produces a precise estimate of the delay. The particular strategy used in this stage consists on calculating the delay of properly selected frames of 8,000 samples (one at the beginning of the signal, one at the middle and one at its end). If the differences among the delays for those frames do not exceed 1 ms, and if their confidence measures are more than 0.5 (50 %), then it is considered that the signals have constant delay, whose value is the mean calculated over the three estimates.

The elimination of silent sections common to both signals, as described in Section 4.2, aims to avoid that any of the three chosen frames has a long silent period, which could contaminate the delay estimate for that interval. This occurs because silent periods are composed basically by noise, and the cross-correlation between two independent noises does not estimate delays correctly.

When the delays and confidence measures fulfil the requirements described above, the silent sections are reintroduced and the signals are aligned according to the mean delay value for the three frames, as can be seen in the block 4a of Figure 3.

The alignment procedure is shown in Figure 5. After the correction of a positive delay of n samples, the n first samples of the degraded signal are eliminated. In order to assure that the resulting signals have the same length, the n last samples of the original signal are also discarded. The procedure is inverted when the delay is negative. After these procedures, the routine is ended.

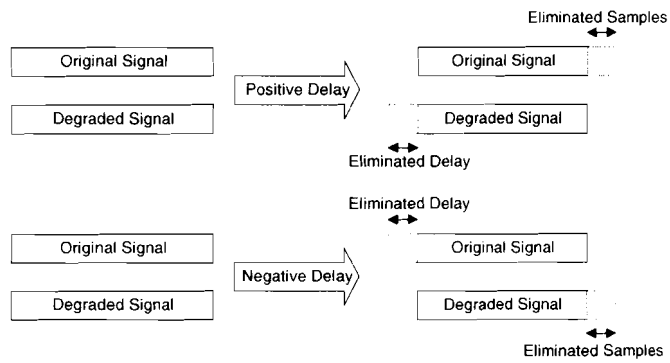


Figure 5. Scheme for signals alignment.

4.4 ROUGH ALIGNMENT

When the signals present variable delay, they are submitted to the other stages of the proposed routine. Initially, the silent sections common to both signals are reintroduced. After that, a first alignment of the signals is

performed by a subroutine whose basic scheme is shown in Figure 6. The signals are divided into 4 ms frames, without superposition. Next, the energy of each frame is calculated in order to determine the envelope of both signals. The cross-correlation between the envelopes is computed using Fast Fourier Transform, as can be seen in Equations 1 to 3.

$$X(k) = \Im\{x(n)\} \quad , \quad Y(k) = \Im\{y(n)\} \tag{1}$$

$$S(k) = Y(k) \cdot X^{*}(k) \tag{2}$$

$$s(n) = \Im^{-1}\{S(k)\} \tag{3}$$

where:
- $x(n)$ and $y(n)$ are, respectively, the original and degraded signals in the time domain;
- $\Im$ represents the Discrete Fourier Transform, and $\Im^{-1}$ represents its inverse;
- $X(k)$ and $Y(k)$ are, respectively, the original and degraded signals in the frequency domain;
- $X^{*}(k)$ represents the complex conjugate of $X(k)$;
- $s(n)$ is the vector of values for the cross-correlation.

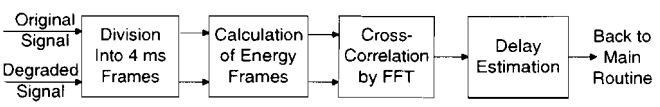


Figure 6. Scheme for crude delay estimate.

The position of the largest absolute value of the correlation is assumed as the estimate for the rough delay. After that, the signals are aligned according to this rough estimate. The precision of such estimate strongly depends on the amount of information enclosed in the envelope [12]. Due to the highly non-stationary characteristic of speech signals, normally the envelope carries enough information to provide delay estimates with a resolution of approximately 8 ms, which is enough to the purpose of a first alignment [12].

The alignment based on this rough delay estimate prevents that the misalignment between the whole signals be excessive, what could prejudice the performance of the next steps of the routine. All future delay estimates use, in some manner, the calculation of cross-correlation. This calculation is as much precise as larger is the number of samples common to both signals (or segments). Then, reducing the misalignment allows better future delay estimates. This fact becomes even more important when it is observed that the procedure for fine delay estimates will split the signals into blocks of limited size, what means that there will be few samples to work with.

4.5 STORAGE OF REFERENCE SIGNALS

After the rough alignment, the resulting signals, named reference signals, are stored. They are the signals that will be aligned after all segments with constant delay have been determined. This storage is necessary because the following processing stages will successively align the signals in order to obtain more precise delay estimates for each segment. Both signals loose a large amount of samples during this process, and they are not adequate to be used in the rest of the OASS routine. Therefore, all cumulative adjustments are stored and, at the end, after all criteria had been satisfied, the adjustments are applied to the reference signals.

4.6 DIVISION OF SIGNAL SECTIONS

The strategy to detect and measure the variable delay consists on dividing the signals into small portions, named utterances. After the rough alignment (block 5 in Figure 3), the signals are divided into utterances according to the following criteria:
- the first utterance starts at the beginning of the signal and ends at the half of the first silent section;
- the intermediary utterances are located between the middle of two consecutive silent sections;
- the last utterance is located from the middle of the last silent section to 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 eventual 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 the determination of an individual section delay. Practical tests showed that such value is enough even for cases with severe 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 3). 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.

4.7 FINE DELAY ESTIMATE AND CALCULATION OF CONFIDENCE MEASURE

This Section describes the procedure applied to an utterance with length larger than n = 3,072. This procedure aims to estimate the utterance delays and employs a precise delay estimate technique.

The fine delay estimate and the confidence measure are calculated accordingly to Figure 7, which details block 8 of Figure 3.

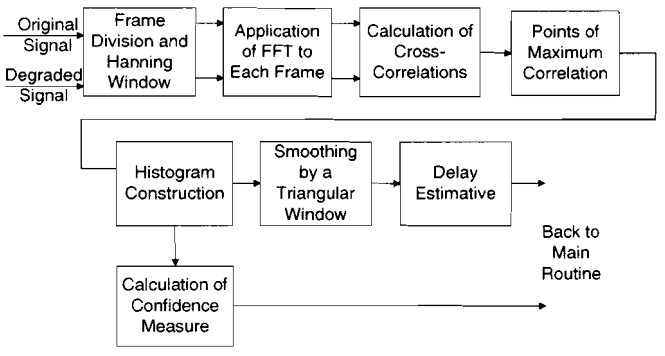


Figure 7. General scheme for the fine delay compensation.

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 present 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 6). Another important parameter responsible for this performance is the high degree of superposition between the frames, which is equal to 87.5 %. This value is superior to those traditionally adopted to such purpose (typically 50 % to 75 %). Despite the slightly greater computational effort required by this approach, it assures more points to construct the histogram and, consequently, more precise estimates. For example, a segment of 10,000 samples will be divided into frames of 2,000 samples; the number of points to construct the histogram in function of the superposition will be 9 for 50 %, 17 for 75 % and 33 for 87.5%. Such additional points are important to assure reliable estimates.

After the division into frames, the cross-correlation for each frame is calculated using the strategy described in Section 4.4. Thus, a histogram is constructed accordingly to the following criterion:

- the value v and index j of the maximum correlation is identified for each frame, according to Equations 4 and 5

$$v(m) = \max_{n=1,\cdots,N}\{s(m,n)\} \tag{4}$$

$$j(m) = n \mid s(m,n) \text{ is maximum} \tag{5}$$

where, m is the index of the frame, n is the index of the samples, s represents the values of the cross-correlation and N is the number of samples per 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. This raising aims to concentrate the values of the cross-correlation around unit, avoiding an excessive domination of high cross-correlation frames.

- consider a particular value for the index j in Equation 5. All the weights corresponding to this particular index j are summed to generate the bar at the index j of a histogram. The procedure is synthesized in the following loop:

$h = 0$;

for $m = 1 : M$

$$h[j(m)] = h[j(m)] + [v(m)]^{0.125}$$

end

where M is the number of frames.

The resulting histogram is then normalized by the sum of all weights, making its area equal to one. This procedure is shown in Equation 6.

$$h_n(p) = \frac{h(p)}{\sum_{m=1}^{M} v(m)}; \quad p = -\max[j(m)], \cdots, \max[j(m)] \tag{6}$$

In this stage, a confidence measure for the delay estimate is obtained from the histogram $h_n(p)$. This value is defined as the percentage of the normalized histogram area that is concentrated up to 1 ms (16 sample) around the histogram maximum, as shown in Equation 7.

$$cm = \frac{\sum_{p=c-16}^{c+16} h_n(p)}{\sum_{p=-\infty}^{\infty} h_n(p)} = \sum_{p=c-16}^{c+16} h_n(p) \tag{7}$$

where c is the index of the maximum of $h_n(p)$.

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

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 3 and Section 4.8) and subdivided into two segments (block 7 and Section 4.6). 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 3), and the routine is terminated.

4.8 REALIGNMENT OF TEST SIGNALS

The segments that do not fulfill at least one of the stop criteria described in Section 4.7 are aligned accordingly to their estimated delays. After that, they are again submitted to the procedures indicated by blocks 7 and 8 in Figure 3. This process is repeated until some of the stop criteria have been fulfilled.

4.9 STORAGE OF DELAYS, DIVISION POINTS AND CONFIDENCE MEASURES

This stage stores the division points and the delays of all segments that satisfied a stop criterion. Both values are used in the final alignment of the reference signals.

The confidence measures are stored only for those segments that do not fulfill any stop criteria. Such measures must be compared to the values obtained after each division, as described in Subsection 4.7 (second stop condition).

5. PSYCHO-ACOUSTICAL MODEL

The reference signals, after the alignment process, are submitted to a psycho-acoustical model in order to determine an estimate for the subjective quality. The

psycho-acoustical models used by the predecessors of OASS have some limitations, especially due their excessively empirical structure and poor modelling of the masking effect. The psycho-acoustical model developed for the OASS was inspired in that one used in the PESQ method, and overcomes the first cited limitation using several enhanced features. The explicit inclusion of the masking has been systematically tested in several methods, always with disappointing results [13,14]. OASS was also not succeeded in modelling such phenomenon. A final solution to this problem is still under study [13].

It is important to emphasize that the psycho-acoustical model used in OASS is not merely a copy of that one used in PESQ. Actually, only a general description of PESQ is available in the literature. Such description was used to orientate the development of the procedures here described. Therefore, most of the strategies here adopted are original or, at least, variations from those used in PESQ.

Figure 8 shows the general structure of OASS. The actual psycho-acoustical processing starts after block 7, but an explanation of the previous processing is necessary. Therefore, each block will be detailed in the following.

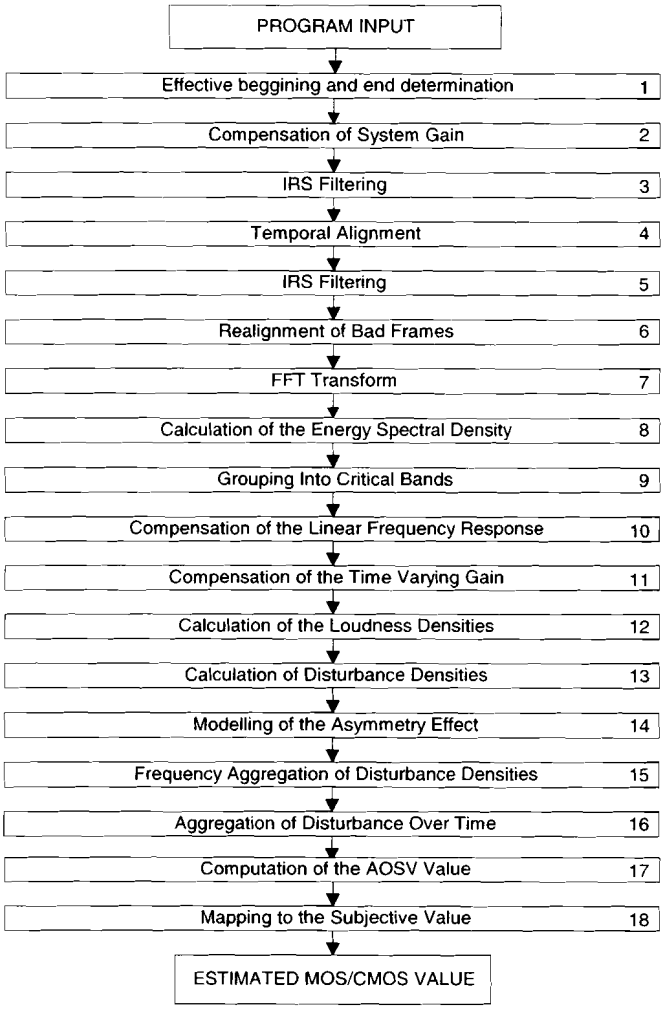


Figure 8. General scheme of the OASS method.

5.1 DETERMINATION OF THE EFFECTIVE BEGINNING AND END

The block 1 indicates the detection of the effective beginning and end of the signals, which is performed by procedures standardized by Rec. P.861 [2] and is identical to that one used in MOQV. The beginning sample is the first one whose magnitude, summed to the magnitudes of the four prior samples, is equal or greater than a determined value (200 in the case of signals sampled with 16 bits). In the same way, the final sample is that one whose magnitude, summed to magnitudes of the four following samples, is equal or greater than 200. The samples preceding the effective beginning and those ones subsequent to the effective end of the speech file are discarded.

5.2 COMPENSATION OF SYSTEM GAIN

This procedure replaces the global calibration of MOQV. After the determination of the effective beginning and end of the signals, they are scaled to a common energy level, as shown in Equations 8 and 9, in order to eliminate the gain of the transmission system.

$$gc = \sqrt{\frac{\sum_{n=1}^{N}(x(n))^2}{\sum_{n=1}^{N}(y(n))^2}} \tag{8}$$

$$yg(n) = gc \cdot y(n) \tag{9}$$

In the Equations 8 and 9, $x(n)$ and $y(n)$ are, respectively, the original and degraded signals in the time domain, N is the total number of samples of the signals, gc is the gain compensation factor and $yg(n)$ is the degraded signal after the gain compensation.

5.3 MODELLING OF HANDSET CHARACTERISTIC

This procedure, corresponding to blocks 3 and 5 in Figure 8, has two basic differences when related to that one previously used in MOQV [5]: 1) it is applied before the beginning of the psycho-acoustical model; 2) it is used twice. It is assumed that the subjective tests are performed using telephone handsets with a frequency-domain response that follows an IRS (Intermediate Reference System) receive characteristic [15]. A perceptual model must take these responses into account in order to model the signals the subjects actually heard.

Before the application of the variable delay routine (from now called simply VDR), which is described in Section 4 and represented by block 4 in Figure 8, the signals are filtered for the first time. This first filtering was not used in the first versions of this method, but tests revealed that the VDR works better if the signals are filtered. This is so because the filtering eliminates high-frequency components that actually would not be heard in a telephone transmission. The eliminated components can be quite different in both signals, since they do not belong to the standard telephone frequency bandwidth. Then, the filtered signals are more closely related. As a consequence, cross-correlation allows better delay estimates, and correct alignments are more likely to be obtained.

After the filtering, the signals are transformed back to the time domain and submitted to the VDR.

5.4 NEW FRAME DIVISION AND REALIGNMENT OF BAD FRAMES

After the alignment, the signals are again submitted to the filtering with the handset characteristics, as indicated in Block 5 of Figure 8. This new filtering aims to attenuate possible spurious components produced by the alignment process.

The next stage in Figure 8 (Block 6) consists of two actions: 1) the signals are divided into 32 ms frames with 50% superposition and 2) estimation of residual misalignments.

This new frame division aims to split the signals into segments whose length is more appropriate for the Block 7 of Figure 6.

Despite the alignment produced by VDR, some of these new frames can still contain some residual misalignment. Such frames, so-called "bad frames", must be identified and treated accordingly. Here, the main routine is split into two possible paths, as can be seen in Figure 9:

1) If the signals had constant delay before the alignment, this implies that the whole signals were aligned at once; in this case, no bad frames will occur, and then the signals are simply divided into 32 ms frames with a Hanning window. The frames are 50% superposed. After, the signals are processed by the rest of the routine.

2) If the signals had variable delays, a new alignment procedure is launched, in order to compensate possible misalignments remaining from the VDR. This strategy is performed outside VDR in order to use the division of the signals directly in the rest of the OASS routine. If this test was incorporated to VDR, the signals should be regrouped, filtered with the handset characteristics, and then divided again. Such strategy would add unnecessary computational burden to the program.

As illustrated in Figure 9, this last alignment of signals with variable delays is performed as follow:

a) The signals are divided into frames of 32 ms plus 50 additional samples, with a 16 ms superposition and no windowing, as shown in Figure 10.

b) The cross-correlation is calculated for each frame as shown in Equations 1 to 3, and the index of its peak determines the remaining delay.

c) If the remaining delay is less or equal to 50, the frames are realigned. This implies in a loss of samples equal to the compensated delay, justifying the adoption of 50 additional samples. This procedure is illustrated in Figure 11 (part 1) for a frame with a delay of 20 samples.

d) The extra samples at the end of the frames are eliminated. For the example of item *c*, the delay is 20 samples, hence the number of extra samples is 30. This procedure is shown in Figure 11, part 2. After the process, all frames must have 512 samples.

e) If the remaining delay is larger than 50 samples, the frame must be classified as silence or active speech. A frame is considered active if the sum of the absolute values of its components is greater than 100,000 in the original signal or the cross-correlation between the corresponding frames in both signals is greater than 0.5. These decision rules were determined by means of empirical tests; such values were chosen because they led to reasonable robustness for a wide variety of signals, including the noisy ones. The neural classifier was not used here because it is not necessary a high-precision classification. Therefore, its use would produce an unnecessary increase of the computational effort. After the classification, if the frame is considered active, its misalignment is considered excessive and it is eliminated from the rest of the routine. If the frame is classified as silence, no further alignment is possible and the frame is used without any modification.

At last, a Hanning window is applied to the 32 ms frames and the rest of the routine is applied.

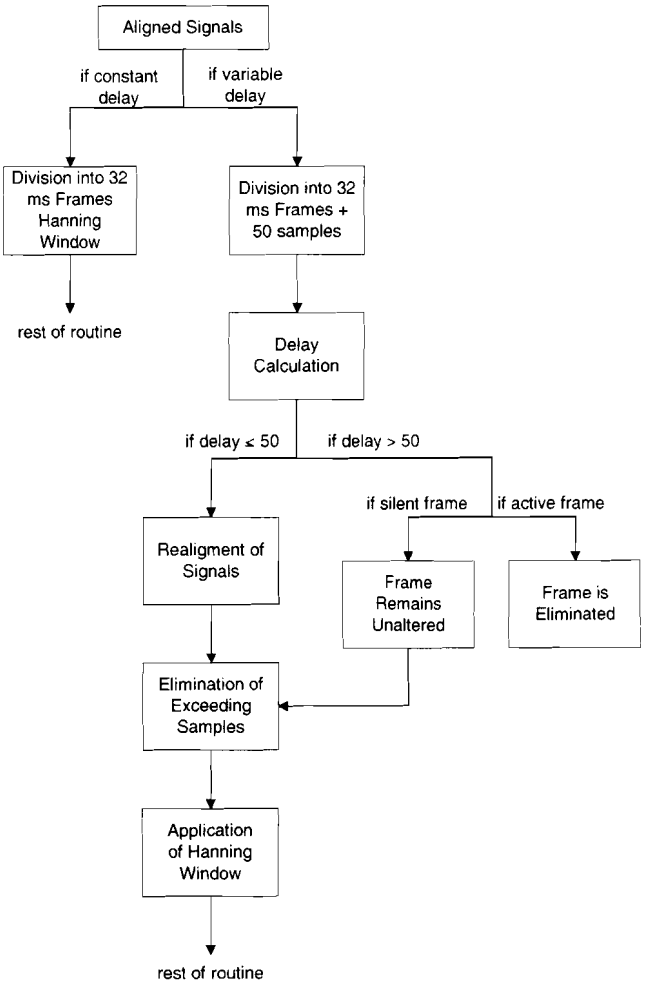


Figure 9. Strategy to perform the additional alignment between the signals.

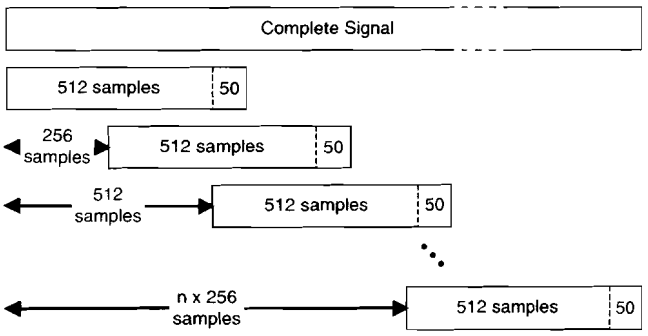


Figure 10. Division of the signal into frames.

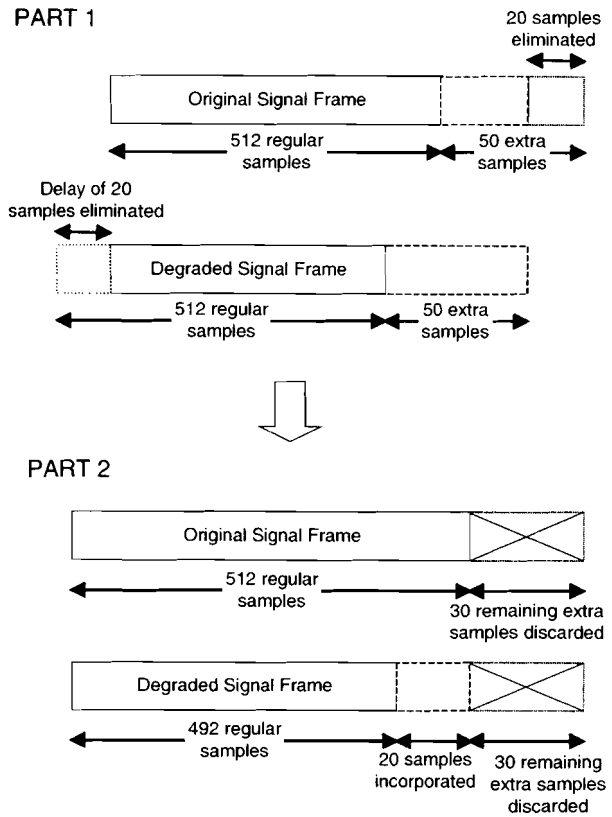


Figure 11. Example of frame realignment.

5.5 FRAME ENERGIES AND GROUPING INTO CRITICAL BANDS

This stage is similar to the corresponding one found in MOQV [5], and it is represented by blocks 7, 8 and 9 in Figure 8. The spectral energy density is calculated via FFT. After that, such energies are grouped into sub-bands, whose limits are based in the concept of critical bands [16]. Table 1 shows how the components are grouped into bands, where k represents the sub-bands, and i and u represent, respectively, the first and the last spectral line to be summed for each sub-band.

k	i	u	k	i	u	k	i	u
1	1	1	18	20	21	34	64	67
2	2	2	19	22	23	35	68	72
3	3	3	20	24	25	36	73	78
4	4	4	21	26	27	37	79	84
5	5	5	22	28	29	38	85	91
6	6	6	23	30	31	39	92	99
7	7	7	24	32	33	40	100	108
8	8	8	25	34	35	41	109	117
9	9	10	26	36	38	42	118	129
10	11	11	27	39	41	43	130	141
11	12	12	28	42	44	44	142	156
12	13	13	29	45	47	45	157	172
13	14	14	30	48	51	46	173	190
14	15	15	31	52	54	47	191	211
15	16	17	32	55	58	48	212	236
16	18	18	33	59	63	49	237	256
17	19	19						

Table 1. Grouping into critical bands

The energy of a given sub-band is determined by the sum of the energies of all spectral lines located inside its boundary (determined by the columns i and u in Table 1), as shown in Equation (10):

$$E_x(n,k)=\sum_{s=i(k)}^{u(k)} X(n,s) \quad , \quad E_y(n,k)=\sum_{s=i(k)}^{u(k)} Y(n,s) \qquad (10)$$

where $X(n,s)$ and $Y(n,s)$ are the spectral energy densities of original and degraded signals, respectively; n is the time index; and s and k are the frequency indexes before and after the grouping into sub-bands.

There are some little differences between MOQV and OASS relative to the values of some parameters and correction factors. However, their number of sub-bands is quite different: 67 for MOQV and 49 for OASS. The ensemble of 49 sub-bands is better correlated with the actual sound perception in the human ear. The patterns resulting from this stage, $E_x(n,k)$ and $E_y(n,k)$, are named pitch energy densities.

5.6 COMPENSATION OF THE LINEAR FREQUENCY RESPONSE

Under certain conditions, the degraded signal can be filtered by the system under test, modifying its spectrum. Such filtering is often barely perceived by the listener, but can cause a severe degradation on the subjective quality estimated by the routine. Therefore, the spectra of both signals must be adjusted again, in order to minimize such effect (block 10 in Figure 8). The first step is to calculate the total pitch energy density over time for both signals, as shown in Equation 11. This sum is calculated using only time-frequency components whose energy is more than 20 dB above the absolute hearing threshold [16].

$$\overline{E}_x(k)=\sum_{n=1}^{N}\tilde{E}_x(n,k) \quad , \quad \overline{E}_y(k)=\sum_{n=1}^{N}\tilde{E}_y(n,k) \qquad (11)$$

where N is the number of samples in time domain and

$$\begin{cases}\tilde{E}_x(n,k)=0 & if \quad E_x(n,k)<20dB\\ \tilde{E}_x(n,k)=E_x(n,k) & if \quad E_x(n,k)\geq 20dB\end{cases} \qquad (12)$$

and

$$\begin{cases}\tilde{E}_y(n,k)=0 & if \quad E_x(n,k)<20dB\\ \tilde{E}_y(n,k)=E_y(n,k) & if \quad E_x(n,k)\geq 20dB\end{cases} \qquad (13)$$

Next, the ratio between the averaged energy spectrum of the degraded and original signals is calculated, according to Equation 14:

$$C(k)=\frac{\overline{E}_x(k)}{\overline{E}_y(k)+\sigma} \qquad (14)$$

where σ is an arbitrarily small value used to avoid division by zero; this value is used in other Equations along the paper. After limiting its peak values to 20 dB, the ratio spectrum is used as a compensation for the pitch energy density of the original signal in order to equalize both signals, as shown in Equations 15:

$$P_x(n,k)=\tilde{C}(k)\cdot E_x(n,k) \quad , \quad P_y(n,k)=E_y(n,k) \qquad (15)$$

where $\tilde{C}$ is the ratio spectrum limited to 20 dB and E_x and E_y are the pitch energy densities of the original and

degraded signals.

It is important to note that the limit of 20 dB imposed to the ratio between the average energy spectra aims to model the fact that severe filtering can be disturbing to listeners. This limit guarantees that differences greater than 20 dB will be only partially compensated. In other words, if the ratio is less than 20 dB, it is assumed that listeners will not perceive any degradation, and then such difference is fully compensated. On the other hand, values greater than 20 dB are considered annoying, and then the routine must consider such degradation.

5.7 COMPENSATION OF THE TIME VARYING GAIN

In certain cases, the gain can fluctuate along the time, causing differences between both signals. As for the compensation described in Section 5.6, such variations must be partially or fully compensated, depending on their severity. For each frame, all samples that exceed the absolute hearing threshold are used to compute the frame energies in the original and degraded signals, as shown in Equation 16.

$$F_x(n) = \sum_{k=1}^{K} \tilde{P}_x(n,k) \quad , \quad F_y(n) = \sum_{k=1}^{K} \tilde{P}_y(n,k) \qquad (16)$$

In Equation 13, $\tilde{P}_x$ and $\tilde{P}_y$ are the values of P_x and P_y after discarding the samples whose energy is below the absolute hearing threshold, and K is the number of spectral bands. Then, the ratio between the energies of each frame is calculated and bounded from 0.0003 to 5 (the ratios located inside such boundary are fully compensated), according to Equations 17 and 18.

$$R_f(n) = \frac{F_x(n)}{F_y(n) + \sigma} \qquad (17)$$

$$\begin{cases} R(n) = \max\left(0.0003, R_f(n)\right) \\ R(n) = \min\left(5, R_f(n)\right) \end{cases} \qquad (18)$$

After that, a first-order low-pass filter is applied to those ratios, in order to slightly smooth possible sharp peaks, as shown in Equation 19.

$$R_{lp}(n) = 0.2 \cdot R(n-1) + 0.8 \cdot R(n) \qquad (19)$$

Finally, the pitch energy densities of the degraded signal are multiplied by the corresponding smoothed ratio, according to Equation 20. The pitch energy densities of the original signal are kept unchanged.

$$S_y(n,k) = R_{lp}(n) \cdot P_y(n,k) \quad , \quad S_x(n,k) = P_x(n,k) \qquad (20)$$

5.8 CALCULATION OF LOUDNESS DENSITIES

This stage is also similar to that one found in the MOQV [5]. This procedure transforms the pitch energy densities into a Sone loudness scale [16], according to Equations 21 and 22.

$$L_x(n,k) = E_q \cdot \left(\frac{S_0}{0.5}\right)^{\gamma} \cdot \left[\left(0.5 + 0.5 \cdot \frac{S_x(n,k)}{S_0(n)}\right)^{\gamma} - 1\right] \qquad (21)$$

$$L_y(n,k) = E_q \cdot \left(\frac{S_0}{0.5}\right)^{\gamma} \cdot \left[\left(0.5 + 0.5 \cdot \frac{S_y(n,k)}{S_0(n)}\right)^{\gamma} - 1\right] \qquad (22)$$

where $E_q = 0.1866055$ is a scaling factor, S_0 is the absolute hearing threshold given by Equation 23 and γ is the Zwicker power, which varies from 2 (low frequencies) to 1 (high frequencies). The resulting patterns are named loudness densities.

$$S_0(k) = 3{,}64 \cdot f(k)^{-0.8} - 6{,}5 \cdot e^{-0{,}6\cdot(f(k)-3{,}3)^2} + 10^{-3} \cdot f(k)^4 \qquad (23)$$

In Equation 23, $f(k)$ indicates de central frequency of the k^{th} spectral band.

5.9 CALCULATION OF DISTURBANCE DENSITIES

The first procedure of this stage is the calculation of the difference between the loudness densities ($L[n,k]$) of original and degraded signals (Equation 24), corresponding to the perceived disturbance between the signals. If such difference is positive, noise-type components were added to the degraded signal. When the difference is negative, components of the original signal were suppressed. The resulting pattern is named raw disturbance density (*rdd*).

$$rdd(n,k) = L_x(n,k) - L_y(n,k) \qquad (24)$$

The masking effect is modelled by applying some rules to each component of the time-frequency plane. Firstly, the energy densities of the original and degraded signals are compared. A masking value is defined for each pair of corresponding components in both signals: it is the smallest value between the energy densities of the pair, as shown in Equation 25.

$$mv(n,k) = \min\left(L_x(n,k), L_y(n,k)\right) \qquad (25)$$

Then, the disturbance densities are determined according to Equation 26.

$$\begin{aligned} &\text{if } rdd(n,k) \geq mv(n,k) \rightarrow D(n,k) = rdd(n,k) - mv(n,k) \\ &\text{if } -mv(n,k) < rdd(n,k) < mv(n,k) \rightarrow D(n,k) = 0 \\ &\text{if } rdd(n,k) \leq mv(n,k) \rightarrow D(n,k) = rdd(n,k) + mv(n,k) \end{aligned} \qquad (26)$$

The effect of this set of instructions is to pull the disturbance densities towards zero. Additionally, such procedure determines a zone where a distortion is not perceived. This zone occurs when the raw disturbance is smaller than the absolute mask value (second instruction above). Therefore, the routine considers that small distortions are inaudible when a louder component is present in the same frequency cell. The resulting patterns are named conventional disturbance densities (D).

It is important to note that this strategy approximates, in a relatively crude way, the masking effects. As commented before, the explicit modelling of the masking still needs further studies.

5.10 MODELLING THE ASYMMETRY EFFECT

The addition of spurious components to a speech signal is more annoying than the subtraction of some original components. To model this, firstly it is calculated an asymmetric factor defined as the ratio between the pitch energy densities of the degraded and original signals, as can

be seen in Equation 27. If this ratio is less than 1, it is set to zero; if it exceeds 12, it is clipped at that value. This procedure is shown in Equation 28. Therefore, the asymmetric factor will be different from zero only when a component of the degraded signal exceeds the value of its correspondent in the original signal.

$$AF(n,k)=\frac{S_y(n,k)}{S_y(n,k)+\sigma} \tag{27}$$

$$\begin{aligned} &if \quad AF(n,k)<1 \quad \rightarrow \quad AF_c(n,k)=0 \\ &else \quad AF_c(n,k)=\max(12,AF(n,k)) \end{aligned} \tag{28}$$

The value 12 adopted for the upper bound aims to avoid huge values for the asymmetric disturbance factor when components of the original signal have very low pitch energy densities. Then, the asymmetric factor (AF_c) multiplies the conventional disturbance densities (D), resulting in the asymmetric disturbance densities (DA), as shown in Equation 29.

$$DA(n,k)=AF_c(n,k)\cdot D(n,k) \tag{29}$$

Both conventional and asymmetric disturbance densities are applied to the procedures described in Sections 5.11 and 5.12. After that, they are combined according to the strategy described in Section 5.13.

5.11 FREQUENCY AGGREGATION OF DISTURBANCE DENSITIES

The conventional and asymmetric disturbance densities are aggregated (summed) along the frequency axis using two different norms and a weighting for low energy frames, as showed in Equations 30 and 31.

$$D(n)=M_n(n)\cdot\sqrt[3]{\sum_{k=1}^{K}\left(|D(n,k)|\cdot W_f(k)\right)^3} \tag{30}$$

$$DA(n)=M_n(n)\cdot\sqrt[3]{\sum_{k=1}^{K}\left(|DA(n,k)|\cdot W_f(k)\right)^3} \tag{31}$$

where M_n, given by Equations 32 and 33, is used to emphasize the disturbances that occurs during silent intervals; W_f is a constant proportional to the width of sub-band k, and is calculated according Equation 34.

$$M_n(n)=\left[\frac{\hat{E}_x(n)+10^5}{10^7}\right]^{-0.04} \tag{32}$$

$$\hat{E}_x(n)=\sum_{k=1}^{K}E_x(n,k) \tag{33}$$

$$W_f(k)=0.15734\cdot\frac{bw(k)}{bw(1)} \tag{34}$$

In Equation 31, bw is the bandwidth of the respective band.

5.12 AGGREGATION OF DISTURBANCE OVER TIME

The disturbance densities resulting from the frequency aggregation are divided into intervals of 20 frames. Such intervals overlap 50 %. The disturbance densities are then aggregated in each interval using a L_6 norm, as shown in Equations 35 and 36.

$$D_1(i)=\left(\frac{1}{J}\cdot\sum_{j=1}^{J}D(i,j)^6\right)^{\frac{1}{6}};\ 1\le i\le I \tag{35}$$

$$DA_1(i)=\left(\frac{1}{J}\cdot\sum_{j=1}^{J}DA(i,j)^6\right)^{\frac{1}{6}};\ 1\le i\le I \tag{36}$$

where I is the number of intervals resulting from the division, J is the number of frames inside each interval, the index i represents the intervals and the index j represents the frames inside the intervals. The value $p = 6$ adopted for the interval aggregation strongly detaches louder distortions, because when small segments of an interval are distorted, the whole interval may loose its meaning.

After that, the results are aggregated along the whole signal using a L_2 norm, as can be seen in Equations 37 and 38.

$$D_2=\left(\frac{1}{I}\cdot\sum_{i=1}^{I}D_1(i)^2\right)^{\frac{1}{2}} \tag{37}$$

$$DA_2=\left(\frac{1}{I}\cdot\sum_{i=1}^{I}DA_1(i)^2\right)^{\frac{1}{2}} \tag{38}$$

In the case of the whole signal, the presence of a distorted sentence does not imply in a meaningless signal. Then, a lower order norm ($p = 2$) was applied to reduce the emphasis for loud distortions.

5.13 COMPUTATION OF OASS VALUE

The OASS value results from a linear combination of the asymmetric and conventional disturbances (block 17 in Figure 8). In OASS, there are two possible combinations, while in PESQ there is only one [14]. One of the combinations takes into account the asymmetry effect, whereas the other does not, as showed in Equation 39. This is due to important differences observed in the behavior pattern of the listeners depending on the type of subjective measure that is to be estimated. This discussion will be retaken later, during the discussion of results.

$$\begin{aligned} &\text{if subjective value is MOS:} \\ &\quad OASS=4.5-0.9\cdot D_2-0.06\cdot DA_2 \\ &\text{if subjective value is CMOS:} \\ &\quad OASS=4.5-2\cdot D_2 \end{aligned} \tag{39}$$

5.14 MAPPING TO THE SUBJECTIVE VALUE

The mapping process from objective to subjective measures is based on Equation 40:

$$y_i=\beta_1+\beta_2\cdot x_i+\ldots+\beta_p\cdot x_i^{\,p} \tag{40}$$

where x_i represents the objective measures and y_i represents the estimated subjective measures. The β_i parameters are obtained minimizing Q in:

$$Q=\sum_{i=1}^{n}\left[s_i-\left(\hat{\beta}_1+\hat{\beta}_2 x_i+\ldots+\hat{\beta}_p x_i^{\,p}\right)\right]^2 \tag{41}$$

where n is the number of experiments whose subjective

measures s_i are known. The $\hat{\beta}_i$ parameters are the values to be adjusted in order to minimize Q: the resulting values are attributed to β_i. The order p adopted for this polynomial mapping in the following tests was 3.

6. TESTS AND RESULTS

6.1 DATABASES USED IN THE TESTS

The first database used in the tests was ITU-T S-23, which is composed by speech files in English, French and Japanese, associated to a number of codecs submitted to some test conditions [17]. This database does not include speech signals with variable delay. Each file contains the original and degraded signals, as well the corresponding subjective scores. Such material is divided in three experiment groups:

- 1st Experiment: the speech files in this experiment were submitted to several 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 for the experiment (176 by language).
- 2nd Experiment: the speech files were submitted to a number of environment noises – office, vehicle, street, white and music. The Signal-to-Noise ratio was 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. Therefore, there are 40 conditions, totalizing 136 test files for each one of the 3 languages (totalizing 408 files for the whole experiment).
- 3rd Experiment: this experiment simulates the effects of transmitting a coded signal through a communication 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 Italian, resulting in a total of 800 test files.

The second database, provided by the Centro de Pesquisa e Desenvolvimento em Telecomunicações (CPqD Foundation), is composed by 12 speech files in Portuguese, each of them submitted to 12 different conditions, resulting in 144 pairs of files. Six talkers of each sex were used in the composition of the 12 files. The subjective values found in this database correspond to the average score assigned to each one of the 12 files used in each condition. Therefore, there are only 12 points to calculate correlation and to plot mapping curves between objective and subjective values.

This set of signals contains severe delay variations. At the beginning of the files, each degraded signal presents is delayed relative to the corresponding original one. However, the degraded signals systematically loose samples with time, in such a way that, at the end, the original signals are delayed relative to degraded ones. There are cases where the delay variation between the beginning and the end of the signals exceeded 100 ms. This is a critical situation to test the routine, since there is no segment with constant delay and, therefore, this is one of the worst conditions that could be faced. Besides, the mapping curves obtained for this database are adapted to the Portuguese spoken in Brazil.

6.2 OASS FACE TO SIGNALS WITH CONSTANT DELAY

This Subsection presents the results and a plot illustrating the performance of OASS for each experiment of the S-23 database. The mapping from OASS values to estimated subjective values is done through a third-order polynomial, optimised to each experiment and language. Such optimised polynomials are included in the program, allowing the user to choose the desired order of mapping. The results are presented in terms of the correlation between the estimated subjective values and the actual subjective scores. Correlation values close to 1 indicate good performance.

- *Experiment 1*: the correlation obtained for this experiment is very high (0.9634). The performance of OASS was also tested without the modelling of the asymmetry effect, resulting in correlations below 0.85. This occurs due the subjective test applied to this experiment, the Mean Opinion Score (MOS). In this kind of test only the degraded signal is presented to the listeners. Therefore, they do not have a reference to identify precisely the suppressed components, while the additive distortions are easily perceived. Therefore, the asymmetry effect plays a fundamental rule in this experiment.
- *Experiment 2*: the results observed were also very good, with an average correlation of 0.9647, almost the same obtained for the first experiment. In this case, it was observed that the inclusion of the asymmetry effect caused a drop in the correlation value to around 0.91. This phenomenon can be explained by the fact that the subjective measure used in this experiment was the Comparative Mean Opinion Score (CMOS). In this case, the original signal is also presented to listeners to allow a comparison between the signals [18]. In this manner, listeners are able to detect the suppression and addition of components with approximately the same precision. Consequently, the asymmetric effect does not make sense anymore.
- *Experiment 3*: the correlation for this experiment reached 0.9232. This result can be considered very good, especially because this set of speech files contains the hardest conditions among all three experiments. The subjective test used here was the MOS, so the asymmetry effect must be modelled.

Figure 12 exemplifies the results obtained in the tests. The plot was generated for English files from the first experiment, and the curve was determined through a third-order polynomial mapping.

Figure 12. Example of the results achieved by OASS for signals without variable delay.

6.3 OASS FACE TO SIGNALS WITH VARIABLE DELAY

The OASS was applied to the files in the database of CPqD. The correlation is 0.9440, contrasting to the value around 0.45 obtained without VDR. It is important to highlight that even better correlations would be possible if the number of points available was larger, allowing to a better fitted mapping curve. Besides the correlation test, a careful visual inspection of the delay compensation was performed. Only slight deviations were observed,

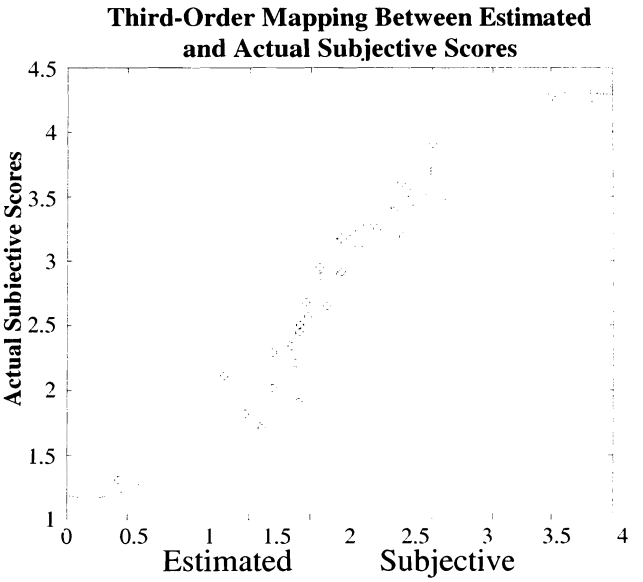
confirming that the strategy used to eliminate variable delays was really effective. Figure 13 shows the results.

6.4 OASS COMPARED TO OTHER METHODS

Table 2 presents the results from MOQV, PESQ and OASS. As can be observed, the OASS outperformed the other ones in 8 of 11 situations. Even in the cases where OASS was outperformed, its performance was very good. Its superiority over MOQV for the experiments based on the database S-23 demonstrates the evolution of the psycho-acoustical model. The superiority in the experiments with signals in Portuguese is due to the improvement of the psycho-acoustical model and also to VDR.

Database	Language	MOQV	PESQ	OASS
1°	French	0.947	0.920	***0.956***
	Japanese	0.946	0.939	***0.968***
	English	0.962	0.943	***0.967***
2°	French	0.937	0.942	***0.974***
	Japanese	***0.957***	0.925	0.954
	English	0.959	0.929	***0.966***
3°	French	0.898	0.871	***0.908***
	Italian	0.899	0.929	***0.932***
	Japanese	0.901	***0.942***	0.929
	English	0.887	0.916	***0.925***
Variable Delay	Portuguese	0.476	***0.964***	0.944
Average		0.888	0.929	***0.948***

Table 2. Comparison of results from MOQV, PESQ and OASS



It is important to observe that the seeming superiority of PESQ face to OASS for the case of signals with variable delay is not conclusive. First of all, it must be noted that VDR presents a very good performance, attested by computational tests and visual inspections. Additionally, OASS was superior to PESQ in most of the fixed delay conditions, confirming that its psycho-acoustical model is, at least, as good as the one used in PESQ. Finally, the relatively small size of the database employed implies that even a little mistake in the estimate of the subjective quality can produce a significant difference in the correlation value. It happened that PESQ had a little advantage, but the opposite could easily occur. In the case of a wider database, such little mistakes could be diluted in the final calculation of correlation, what probably would imply in a similar performance for both methods.

OASS was not the first attempt to replace MOQV by a more effective method, as commented before. Some routines were developed replacing the polynomial mappings in MOQV by artificial neural networks [7,8,9]. Despite the excellent correlations, even better than those ones from OASS, they were not included in Table 2 because such approaches are not robust when faced to unknown conditions. Nevertheless, a method including both approaches will be proposed soon. In this case, the user will choose which approach is more interesting for his application.

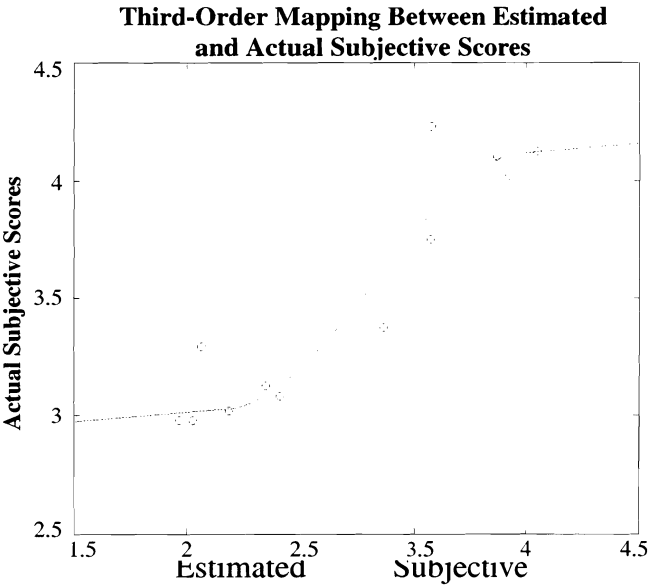


Figure 13. Results achieved by OASS for signals

containing variable delay.

7. CONCLUSIONS

A new method for objective assessment of speech quality, the OASS, was proposed as an improvement of former proposed methods. Its main contributions are a new routine to eliminate variable delay and an improved psycho-acoustical model.

The technique to eliminate variable delay showed a good performance, and no further improvements are expected. On the other hand, the psycho-acoustical model yet presents some shortcomings common to all other methods, such as the absence of an explicit modelling of the masking effect. Therefore, more studies must be performed and new improvements are possible.

The new routine fully achieved its objective, outperforming its predecessor, the MOQV. Additionally, it reached a slightly better performance than PESQ, which is the currently ITU's standard method. Therefore, OASS eliminates the lack of technical knowledge about the development of PESQ. Additionally, it eliminates the costs to use a commercial method as PESQ.

ACKNOWLEDGEMENTS

This work was supported by Fapesp, process number 01/04144-0, and by Centro de Pesquisa e Desenvolvimento em Telecomunicações (Fundação CPqD).

REFERENCES

[1] J. G. Beerends, J. A. Stemerdink, "A Perceptual Speech-Quality Measure Based on a Psycho-acoustic Sound Representation", *J. Audio Eng. Soc.*, Vol. 42, No. 3, pp. 115-123, March 1994.

[2] "Objective Quality Measurement of Telephone-Band (300 - 3400 Hz) speech codecs", *ITU-T Recommendation P.861*, 1996.

[3] A. W. Rix, M. P. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment", *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 3, 1515-1518, Istanbul, June 2000.

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

[5] J. G. A. Barbedo, "Objective Quality Assessment of Telephone-Band Speech Codecs" (in Portuguese), *Master's Thesis*, Unicamp, Campinas, July 2001.

[6] J. G. A. Barbedo, A. Lopes, "Proposition and Valuation of a Objective Measure for Assessment of the Speech Quality of Codecs" (in Portuguese), *Proceedings of the XIX Simpósio Brasileiro de Telecomunicações*, Fortaleza, Brazil, paper n. 001000000002200007, September 2001.

[7] J. G. A. Barbedo, M. V. Ribeiro, F. J. Von Zuben, A. Lopes, J. M. T. Romano, "Application of Kohonen Self-Organizing Maps to Improve the Performance of Objective Methods for Speech Quality Assessment", *Proceedings of the XI European Signal Processing Conference (EUSIPCO2002)*, vol. I, pp. 519-522, Toulouse, France, September 2002.

[8] J. G. A. Barbedo, M. V. Ribeiro, A. Lopes, J. M. T. Romano, "Estimate of the Subjective Quality of Speech Signals using the Kohonen Self-Organizing Maps", *Proceedings of the IV International Telecommunication Symposium (ITS)*, Natal, Brazil, pp. 834-839, September 2002.

[9] M. V. Ribeiro, J. G. A. Barbedo, J. M. T. Romano, A. Lopes, "Fourier-Lapped-Multilayer Perceptron (FLMLP) Method for Speech Quality Assessment", submitted to the *Special Issue on Anthropomorphic Processing of Audio and Speech*, November 2003.

[10] 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*, Vol. 45, No. 10, pp. 789-814, October 1997.

[11] M. V. Ribeiro, "Packet Reconstruction Techniques Based on Wavelet Transform and Neural Networks Applied to Waveform Coders in IP Telephony" (in Portuguese), *Master's Thesis*, Unicamp, Campinas, October 2001.

[12] A. W. Rix, M. P. Hollier, A. P. Hekstra, J. G. Beerends, "Perceptual Evaluation of Speech Quality (PESQ), the New ITU Standard for End-to-End Speech Quality Assessment: Part I – Time Alignment", *Journal of the AES*, Vol. 50, No. 10, pp. 755-764, October 2002.

[13] J. G. Beerends, A. P. Hekstra, A. W. Rix, M. P. Hollier, "Perceptual Evaluation of Speech Quality (PESQ), the New ITU Standard for End-to-End Speech Quality Assessment: Part II - Psychoacoustic Model", *Journal of the AES*, Vol. 50, No. 10, pp. 765-778, October 2002.

[14] J. G. Beerends, J. A. Stemerdink, "The Optimal Time-Frequency Smearing and Amplitude Compression in Measuring the Quality of Audio Devices", *Proceedings of the 94th Convention of the Audio Engineering Society, J. Audio Eng. Soc. (Abstracts)*, vol. 41, p. 409, May 1993, preprint 3604.

[15] "Specification for an Intermediate Reference System", *ITU-T Recommendation P.48*, 1989.

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

[17] "ITU-T Coded Speech Database", *Series P Supplement 23*, February 1998.

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

Jayme Garcia Arnal Barbedo received the B.S. degree in Electrical Engineering from the Federal University of Mato Grosso do Sul in 1998, and the M.Sc. and Ph.D. degrees in Electrical Engineering from the State University of Campinas in 2001 and 2004, respectively. Since 2004, he has been with CPqD Foundation, working at the Digital Television Division. His interest areas include speech and audio signal assessment and codification, audio signal classification and digital television.

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

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

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