A Teleconference Model with Acoustic Impairments Suitable for Speech Quality Assessment

Flávio R. Ávila, Luiz W. P. Biscainho, Leonardo de O. Nunes, Alan F. Tygel, Bowon Lee, Amir Said, Ton Kalker, and Ronald W. Schafer

Abstract— **Modern teleconference systems have set a new paradigm for speech quality, calling for a more rigorous control over potential impairments and resulting quality of service. Reliable and efficient tools for quality assessment (QA) of speech should be automatic and capable of emulating subjective tests. The present paper addresses this topic by proposing a simple teleconference model (called TMAI) intended as a framework to the design of QA tests for acoustically impaired speech. A procedure for building a database of reference and degraded signals to be employed in QA tests which uses the TMAI is described.**

Keywords— **Teleconference, Acoustic degradations, Echo, Noise, Reverberation, Quality assessment.**

Resumo— **Sistemas modernos de teleconferência determinaram um novo paradigma para a qualidade de conversação, exigindo um controle mais rigoroso sobre potenciais defeitos e a consequente qualidade de serviço. Ferramentas confiáveis e eficientes para avaliação de qualidade (AQ) de conversação devem ser automáticas e capazes de emular testes subjetivos. Este artigo trata desse tema, ao propor um modelo simples de teleconferência (chamado TMAI) para servir de arcabouço para o projeto de testes para AQ de fala degradada por fatores acústicos. É descrito um procedimento para construção de uma base de dados com sinais de referência e degradados a ser empregada em testes de AQ, o qual utiliza o TMAI.**

Palavras-Chave— **Teleconferência, Degradações acústicas, Eco, Ruído, Reverberação, Avaliação de qualidade.**

I. Introduction

The field of speech quality evaluation has increased in importance since the digitization process of telephone network has taken place. Traditionally, quality was measured mainly through subjective tests. In the so-called listening tests, subjects are required to grade the quality of a set of signals. The results are usually given in terms of a Mean Opinion Score (MOS), which averages subjects' grades for a given signal as a dimensionless number in the range from 1 (bad quality) to 5 (excellent quality).

Listening tests can be intrusive, when subjects compare the signal under test (SUT) with a reference, or non-intrusive, when the quality of the SUT is assessed in absolute terms, without reference. Detailed specifications for both can be found in ITU-P.800 recommendation [1].

Although effective for quality assessment, subjective tests are expensive and time-consuming, which makes them impractical in many applications. Objective quality assessment (QA) tries to emulate the results of subjective tests in an automated fashion. The PESQ (Perceptual Evaluation of Speech Quality) [2], [3] algorithm is a popular standard for objective QA of speech signals. It has been designed to evaluate telephone voice signals and achieves high correlation with MOS for a variety of impairments.

In the case of teleconference systems, the quality of service is limited due to a combination of network impairments (such as jitter, packet loss, and nonlinear distortions) and acoustic degradations (such as background noise, echo, and reverberation). In recent years, an increasing demand for high-quality services has taken place, which motivates the development of specific tools for QA of voice signals coded at high data rates.

Acoustic degradations are determined by environmental characteristics, which makes them less controllable than electrical ones. QA for this class of impairments becomes even more challenging if one considers their concurrent occurrence. An integrated QA tool directed to simultaneous impairments would require:

- a simple model for the teleconference system including the degradations of interest, yet allowing the definition of meaningful subjective and objective QA tests;
- an associated database of impaired speech, including the desired degradation combinations.

In this context, a conceptual model for transmission of speech signals through a teleconference system (henceforth referred as TMAI – Teleconference Model with Acoustic Impairments) is proposed. The model describes the signal paths from the mouth of each speaker to his/her own ears and to the other speaker's ears, and includes the incidence of acoustic echo, noise and reverberation. In order to control the separation and combination of these different effects and to avoid multiple feedback, the model is separated into two scenarios:

- Local: describing those impairments originated at the talker side, such as local background noise and local room reverberation, as perceived by the talker.
- Remote: describing those impairments originated at the listener side, such as acoustic echo, due to loudspeaker-microphone coupling, as perceived by the talker.

The TMAI should allow the description of compatible subjective and objective QA tests to aid in the design, setup and everyday maintenance of a teleconference system [4].

Flávio R. Ávila, Luiz W. P. Biscainho, Leonardo de O. Nunes, and Alan F. Tygel. Programa de Engenharia Elétrica, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, E-mails: flavio@lps.ufrj.br, wagner@lps.ufrj.br, lonnes@lps.ufrj.br, alan@lps.ufrj.br. Bowon Lee, Amir Said, Ton Kalker, and Ronald W. Schafer. HP Labs, Palo Alto, U.S.A, E-mails: bowon.lee@hp.com, amir.said@hp.com, ton.kalker@hp.com, ron.schafer@hp.com.

Since most reliable QA tools are intrusive [2], [5], a reference point as well as a test point must be accessible in both local and remote models.

The proposed TMAI can serve as a framework for generating a database necessary to investigate QA of multi-degraded speech.

After this introduction, the paper is organized as follows. In Section II a complete model for a teleconference system is described, together with a simplified version of it: the TMAI. Section III shows the two separate scenarios, local and remote, in which the TMAI is divided. Section IV gives the guidelines for building a database of speech signals to be employed in intrusive listening tests. Conclusions are drawn in Section V.

II. Complete Conceptual Model

This section presents a simple model to be used as a framework for QA in teleconference systems. The purpose of this model is twofold:

- Representing the targeted impairments in a way that allows the controlled generation of a comprehensive database of degraded signals to aid in the development of automatic QA tools;
- Allowing the definition of access points to reference as well as tested signals that are meaningful for both subjective and objective tests.

In [6], p. 226, a quite complete model for VoIP connections is presented. Following similar lines but highlighting acoustically induced degradations, a modified model for a teleconference system is shown in Figure 1.

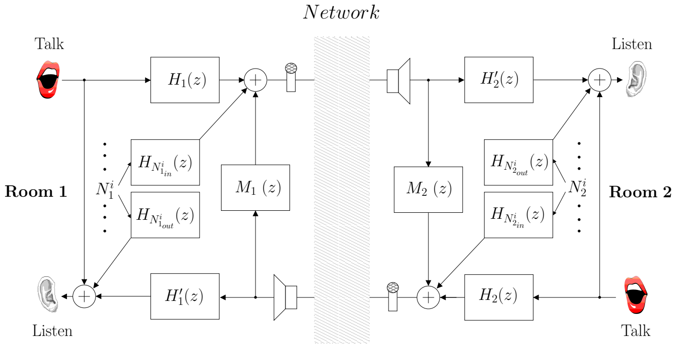

Fig. 1. Complete teleconference model. The "network" block represents network-induced impairments, such as delay, attenuation, packet loss, jitter, clipping and frequency distortions. Each N^i models i-th noise source, like air-conditioner, computer or external traffic jam, which is convolved with transfer functions $H_{N^i_{\text{in}}}$ to microphone and $H_{N^i_{\text{out}}}$ to the listeners' ears. Acoustic coupling between loudspeaker and microphone inside each room is modeled by $M(z)$.

In that figure, one can see speaker 1 in room 1 in conversation with speaker 2 in room 2. Voice coming out of mouth of speaker 1 reaches his/her own ears (through direct path only, for the sake of simplicity) and microphone 1 through room 1 impulse response (RIR) $H_1(z)$. Each noise signal N_1^i from a given source (such as air-conditioner, computer or external traffic jam) i inside the room, after being modified by $H_{N^i_{1_{\text{in}}}}$ and $H_{N^i_{1_{\text{out}}}}$ along its respective paths, reaches the speaker's ears and microphone 1 also. On the other hand, voice coming from room 2 through loudspeaker 1 is directed to the ears of speaker 1 through room response $H'_1(z)$, and is acoustically fed back to microphone 1 via $M_1(z)$. An analogous description suits room 2. Network-induced degradations such as delay, attenuation, packet loss, jitter, clipping and frequency distortions are implicitly depicted inside the grey box.

A. Simplification

At this point, additional simplifications yield the scheme in Figure 2, which is proposed as a reference model for teleconference systems in this paper and will be called TMAI (Teleconference Model with Acoustic Impairments).

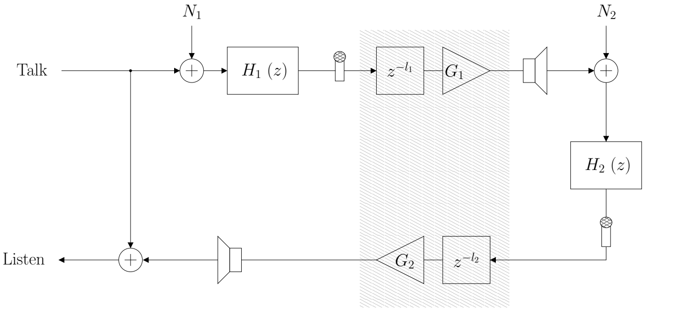

Fig. 2. Simplified teleconference model. $H_1(z)$ and $H_2(z)$ are general RIRs, and network effects are encapsulated into delay and path losses.

Only the transmission/reception cycle related to the voice of speaker 1 is depicted, since the model can be mirrored to describe the opposite paths. The following items have been modified:

- Everything related to speaker 2 has been suppressed.
- Network effects have been encapsulated into a delay of l_1 samples and a gain $G_1 < 1$ from room 1 to room 2, and a delay of l_2 samples and a gain $G_2 < 1$ from room 2 to room 1. The inclusion of those elements will be justified later on.
- All possible noise sources are combined into a single one for each room.
- The same RIR response $H_1(z)$ is employed to represent the paths from noise source and from speaker's voice to the microphone in room 1; the same response $H_2(z)$ is used to represent the paths from noise source and from loudspeaker to the microphone in room 2. This simplification implicitly assumes sufficient similarity between the involved responses in each room, and keeps the description of degradations under reasonable levels of variability. Furthermore, the overall characteristics of background noise are not sensibly affected by the RIR.
- The effects of room 1 along the path from the loudspeaker to the ears of speaker 1 has been disconsidered. Under the assumption of similarity between responses inside the same room, $H_1(z)$ would suffice to describe acoustic conditions in room 1.
- Acoustic coupling between loudspeaker and microphone in room 1 has not been considered, assuming the effect of multiple feedback can be neglected.

Now, in order to ease the isolated and/or composed description of the acoustic issues of interest in both rooms, the described model has been divided into two parts, as described in the next section.

III. Scenarios

The formerly described model can be further simplified to make possible the separate characterization of echo, noise and reverberation effects in each room. This could be done by separating the general system in Figure 2 in two subparts: local and remote scenarios[1]. Although it is possible to recombine them into the complete model, the idea is to use each scenario by itself.

A. Local scenario

The so-called local scenario, shown in Figure 3, corresponds to the upper left part of Figure 2. It can characterize degradations acoustically induced at the talker's side, i.e. background noise originated by local sources and reverberation due to the local RIR.

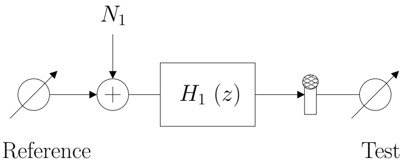

Fig. 3. Local scenario model, including impairments acoustically generated in local room.

In this framework, intrusive QA tests could compare the reference signal that comes out from the speaker's mouth with the signal captured by the microphone to be sent to the remote side, thus encompassing all the modifications suffered inside the local room. This scheme can be used *in loco* if the test signal is sent back to the speaker via headphones. Alternatively, a repeatable test can be conceived by substituting the talker by some pre-recorded signal coming from a loudspeaker, and resorting to an external listener.

B. Remote scenario

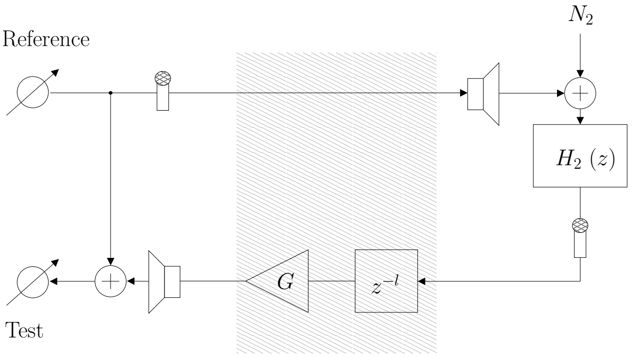

Fig. 4. Remote scenario model, including impairments acoustically generated in remote room.

The so-called remote scenario, shown in Figure 4, corresponds to the remainder of Figure 2 after suppression of the local scenario part. It can characterize degradations acoustically generated at the talker's opposite side, mainly acoustic echo resulting from loudspeaker-microphone coupling combined with transmission paths' delay.

Local effects are discarded as if $H_1(z) = 1$ and $N_1 = 0$ in Figure 3. A total delay of $l = l_1 + l_2$ samples and a total gain of $G = G_1G_2$ can be adopted without loss of generality. In practice, echo-return loss (ERL) $1/G$, usually expressed in dB, describes the attenuation from direct sound to its replica. Additive noise N_2 and RIR $H_2(z)$ make for a more realistic modeling of the remote part.

In this framework, intrusive QA tests could compare the signal delivered to the microphone to be sent to the remote side against its version contaminated by echo received via loudspeaker, as perceived by the talker. This scheme can be used *in loco* if the path between mouth-microphone can be considered ideal and the test signal is sent back to the speaker via headphones instead of a loudspeaker. Alternatively, a repeatable test can be conceived by inserting the reference signal electrically "after" the microphone, and resorting to an external listener.

IV. Building of an Auxiliary Database for Intrusive Listening Tests

As an application of the TMAI, this section gives some guidelines for building a database for voice quality assessment in the teleconference context. The proposed model is taken as a framework for generating a set of degraded signals from a set of reference signals, suitable for posterior intrusive listening tests. The following steps should be performed in this process:

1) Specify the characteristics required by the reference signals that will be prepared from high-quality pre-recorded sentences;
2) Specify the set of parameter combinations that will describe each degradation to be evaluated, constraining their number to keep the database at a manageable size;
3) Generate the set of reference signals that matches the strategy of parameters combination, and apply the respective degradation to each reference signal to generate the set of degraded signals.

The local and remote scenarios of TMAI are necessary in steps 2 and 3 to simulate the targeted degradation types. In the next sections, possible design choices required by the procedure above are discussed.

A. Reference signals

Reference signals for subjective listening tests are described in the ITU P.800 recommendation [1], which defines the so-called "sample" as each stimulus that will be corrupted by the system under test. Listeners should compare the clean sample signals against the degraded ones.

A sample is composed by short meaningful sentences, lasting between 2 and 3 seconds, and taken from non-technical literature or newspaper. Sentences should not make sense among each others. For increasing variability, it is desirable to have phonetically balanced sentences, read by male and

[1] A similar idea is adopted in E-model [7].

female speakers, so that results obtained over a given corpus can be consistent with others obtained over a wider corpus.

The composition of a sample depends on the type of listening test to be executed. For the Absolute Category Rating (ACR), P.800 recommends grouping 2 to 5 sentences to form a sample. If the Degradation Category Rating (DCR) is to be applied, samples are made of a combination of 2 sentences, with an interval of 0.5 s between them.

It is also important that sentences inside a sample have a similar loudness level, otherwise subjects can be distracted and thus evaluate the signal inappropriately. Loudness equalization is not trivial, since loudness misalignment can occur even inside a single sentence. Several strategies can be employed to tackle this issue.

1) **Normalization by mean power:** calculate the power $P_k = 1/N \sum_{n=0}^{N-1} x_k^2[n]$ for all K sentences $x_k[n]$, compute their mean power $P_m = 1/K \sum_{k=0}^{K-1} P_k$ and scale all signals by $x_{k_{\text{eq}}} = x_k \sqrt{P_m/P_k}$. All signals $x_{k_{\text{eq}}}$ will have the same power.
2) **Normalization by mean power during voice activity:** Same as 1), except for calculating P_k after having passed x_k through a Voice Activity Detector (VAD), thus equalizing only the active parts of speech. The scaling step is performed over the original x_k;
3) **Normalization by "perceived" mean power:** Same as 2), except for processing x_k by an A-weighting filter [8], which simulates human loudness as a function of frequency, prior to the calculation of P_k;
4) **Loudness equalization:** Same as 1), except for replacing power P with loudness L, directly computed by some algorithm from the literature (e.g. [9]).

A quite involved procedure for speech level equalization can also be found in [1].

B. *Degraded signals*

After reference signals have been specified, the next step in the design of subjective tests is creating routines to impose specified degradations to the reference signals. Two routines are required, for local and remote scenarios; both can be straighforwadly implemented based on Figures 3 and 4, respectively.

Different types of degradation are described by particular parameters, which should be set beforehand. With respect to TMAI, the parameters are: background noise type and SNR, RIRs (which characterize reverberation), and transmission gain (G) and delay (l) (which define echo perception).

In this section, the focus is on the choice of parameters set to be evaluated in subjective tests, disregarding time/costs constraints. Firstly, some considerations on each acoustic impairment and its respective parameters are given. Then, the combination of individual parameters is addressed.

1) Background noise – type and SNR: Since the TMAI takes into account only acoustic impairments, N_1 and N_2 in Figure 2 model background noise in rooms 1 and 2, possibly generated by sources such as computer, air-conditioner or external traffic jam. In general, they are characterized as broadband noises with particular colorations.

The level of ambient noise is usually measured in dBA [8], an absolute loudness measure based on an approximation for the human loudness curve [10]. However, in listening tests, one is typically interested on the relative level of noise compared with the voice signal, for which the usual SNR (Signal-to-Noise Ratio) measure is more suitable.

Nevertheless, SNR only do not suffice to characterize the perception of noise, for neglecting psychoacoustic effects such as loudness variation with frequency and masking phenomenon [10]. For instance, for a given SNR, noise concentrated around 2 kHz is much louder than noise concentrated around 10 kHz.

For the sake of variability, it is recommended to include several types of noise in the database, both of natural (recorded in typical teleconference rooms) and artificial (created by coloring computer generated white noise) types.

The SNR range for each noise type should be set according with the desirable levels of annoyance, considering the application of interest. For example, white noise at 60-dB SNR is almost imperceptible, while at 20-dB SNR can be very annoying. In the end, one would have a set $\text{NT} = \{N_1, N_2, \ldots, N_{\#\{NT\}}\}$[2] of types of noise and a set $\text{SNR}^i = \{\text{SNR}_1^i, \text{SNR}_2^i, \ldots, \text{SNR}_{\#\{\text{SNR}^i\}}^i\}$ for each type of noise i.

2) Reverberation – Room Impulse Response (RIR): Reverberation is an acoustic phenomenon caused by multiple reflections of the sound signal within the limits of a room. It depends on the room geometry, speaker and microphone placement, as well as walls and ceiling acoustic properties. Main reverberation characteristics can be summarized by T_{60}, room volume and source-microphone distance [11]. The T_{60} is defined as the time the sound pressure level takes to decrease 60 dB from its steady-state level after the sound has abruptly stopped.

The model adopted for reverberation consists of convolving the reference signal with a representative room impulse response (RIR), which is defined as the response to an impulsive input at a certain point of the room. Thus, the degraded signal is obtained by

$$x^{\text{rev}}[n] = \sum_{\tau} x[n-\tau]\,\text{IR}[\tau], \tag{1}$$

where IR is a previously recorded RIR and $x[n]$ is the reference signal. After the convolution operation, the signal is normalized so that the energy level is not modified.

It is recommended to use artificial as well as natural (measured in real rooms) RIRs. The former case allows controlling the range of T_{60} with greater acuity. The latter case has the appeal of allowing to model real teleconference rooms.

Methods for generating artificial RIRs can be found in [11], several of them parameterized by T_{60}. The degree of annoyance associated to reverberation is not monotonic w.r.t. T_{60}; a room with low reverberation can be more disturbing than a room with medium reverberation. On the other hand, excessive reverberation can be very annoying and impair conversation. A typical studio is designed for T_{60} around

[2] The simbol $\#\{A\}$ represents the number of elements of A, that is, its cardinality.

300 ms; a teleconference room should exhibit $T_{60} \approx 150$ ms; reverberation times in a concert hall must be quite longer.

These considerations would yield a set of RIR $= \{\text{RIR}_1, \ldots, \text{RIR}_{\#\text{RIR}}\}$ including natural (measured at different points of a same room) and artificial RIRs with different T_{60} values, suitable for degrading reference signals with reverberation.

3) Echo – gain and delay: Referring to Figure 2, the perception of echo occurs when the talker listens to his own voice added to a delayed and attenuated version of it. A simple characterization of this effect can be done via two parameters: delay L and gain (inverse of path loss) G.

A minimum interval of 30 ms between direct source and its replica is required to give the impression of echo. For shorter delays, the resulting perception is just of an exaggerated "side-tone" effect, while delays longer than 200 ms can affect sensibly the naturalness of conversation [12].

Path loss (attenuation) also affects echo perception. A proper range for this parameter must take the application of interest in account.

The remote scenario model (Section III-B) is recommended to evaluate echo effects. As one can see in the model, the replica signal is associated with reverberation from room 2. As a consequence, when evaluating echo, an RIR should also be assigned to room 2. The same considerations of the previous section apply here.

Following the guidelines above, one would generate a set $G = \{G_1, \ldots, G_{\#G}\}$ of gains and a set $L = \{L_1, \ldots, L_{\#L}\}$ of delays, suitable for degrading the reference signals.

4) Full-factorial combination: For a certain scenario, a degraded signal specification is done by assigning a n-tuple $P = \{p_1, p_2, \ldots, p_n\}$ containing the values of the parameters described above. The full-factorial combination of parameters is the set of all possible P, which has cardinality equal to the product of the cardinality of all sets P_i, $i = \{1, \ldots, n\}$ from which the p_i values are chosen.

For local scenario and remote scenario, the number of full-factorial combinations are, respectively:

$$N_{local} = \#\{\text{SNR}\}\#\{\text{NT}\}\#\{\text{RIR}\}, \qquad (2)$$

$$N_{remote} = \#\{\text{SNR}\}\#\{\text{NT}\}\#\{\text{RIR}\}\#\{G\}\#\{L\}. \qquad (3)$$

supposing equal amount of SNR values for each noise type.

This number could easily reach impractical values for subjective tests if each parameter set is reasonably dense. In most practical cases, combinations must be carefully limited. This is the subject of the next section.

C. Practical issues

Given the considerations made in the previous two sections, there are still some practical issues that should be dealt with when designing a database for listening tests. They are:

1) Number of evaluations to be received by a given degraded signal.
2) Number of parameter combinations.
3) Distribution of parameter combinations among the reference signals.

The last two issues arise because the number of signals in the database must be limited. Each degraded signal should be evaluated a given number of times by different listeners, so that MOS values be statistically meaningful. On the other hand, it is recommended that a given listening session should not last more than 30 minutes [1], [13]. This consideration limits the number of signals per test and increase the number of listening sessions, which can become impractical in some situations.

Referring to item 1 above, the number of times a given degraded signal should be evaluated in a listening test should be sufficiently large to keep variance low. If the group of listeners assessing the database is composed of experts, then the number of times a given signal is evaluated can be greatly reduced. An ITU definition of experts can be found in [14]. In [13], a thorough discussion of this topic can be found.

Regarding item 2, after the parameters needed for database creation have been chosen, it is necessary to evaluate how they will be combined in order to degrade the signals. Firstly, the designer should be aware of what is to be assessed by the listening tests. A combination of every possible parameter level is only necessary when the interactions between any two degradation types have to be evaluated. Otherwise, some parameter combinations can be pruned; however, special attention should be given to which of them can be discarded. In [13], several methods for removing combinations so to avoid the statistical *confounding* among the influences of each parameter in the test are presented. If the focus is on a particular degradation, then its parameters should be extensively varied, while the parameters of remaining degradations can be randomly sampled.

Lastly, as to the distribution of parameter combinations among the reference signals, once more the aim of the database is the main guideline to the best procedure. It is important that each parameters combination should be enforced on the same number of reference signals per speaker. This way, hopefully, differences in speaker prosody and intonation are averaged out when calculating mean opinion scores. Then, for each speaker it is necessary to establish how many times a given degradation and reference signal will be used. Ideally, each parameter combination would be used in every reference signal for a given speaker. This would allow the assessment of the impact of each degradation over MOS for a given signal. However, this scheme can produce an unmanageable number of degraded signals. Moreover, special care should be taken when choosing the number of times the same reference signal appears in a listening test. Too many repetitions may stress the listener, possibly leading to a bias in the tests' results [13]. A more pragmatic approach might be to limit the number of times a given combination of parameters is enforced on the signals uttered by a given speaker. Then, each combination can be randomly distributed among the reference sentences for each speaker. If the sentences are very similar among themselves and phonetically balanced, this randomization approach should have little overall impact in the test.

V. Conclusions

This paper proposed the TMAI, a simplified model for teleconference sytems with focus on acoustic impairments, namely: noise, acoustic echo and reverberation. The division of the model into local and remote scenarios allows separation and combination of individual impairments in a simple way. TMAI is shown to be a useful tool for QA assessment of speech signals, in particular aiding on the generation of databases of degraded speech signals. Regarding this application, the paper gives general guidelines for creation of a database with acoustically impaired signals, suitable for posterior listening tests. The main topics addressed are reference signals specification, parameters selection and its distribution among reference signals.

Acknowledgments

Flávio R. Ávila, Luiz W. P. Biscainho, Leonardo de O. Nunes, and Alan F. Tygel would like to thank CAPES, CNPq, and FAPERJ for partially funding this research.

Referências

[1] ITU-T Rec. P.800, "Methods of subjective determination of transmission quality," International Telecommunication Union, Geneva, Switzerland, 1996.

[2] J. G. Beerends, A. Hekstra, A. Rix, and M. 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 Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, Oct. 2002.

[3] A. Rix, M. Hollier, A. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part I – Time-Delay Compensation," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, Oct. 2002.

[4] S. Möller and A. Raake, "Telephone speech quality prediction: Towards network planning and monitoring models for modern network scenarios," *Speech Commun.*, vol. 38, no. 1, pp. 47–75, 2002.

[5] C. Tan, B. C. J. Moore, N. Zacharov, and V. Mattila, "Predicting the perceived quality of nonlinearly distorted music and speech signals," *Journal of the Audio Engineering Society*, vol. 52, no. 7/8, pp. 699–711, July/Aug. 2004.

[6] A. Raake, *Speech Quality of VoIP – Assessment and Prediction*. New Jersey, USA: Wiley, 2006.

[7] ITU-T Rec. G.107, "The e-model, a computational model for use in transmission planning," International Telecommunication Union, Geneva, Switzerland, 2005.

[8] H. Fastl and E. Zwicker, *Psycho-Acoustics – Facts and Models*. Berlin, Germany: Springer, 2007.

[9] J. Lysaght, L. Schoenwiesner, and L. McManus, "Implementing loudness models in matlab," in *Proc. of the 7th Int. Conference on Digital Audio Effects (DAFX)*, Naples, Italy, Oct. 2004, pp. 177–180, matlab scripts available at: http://www.brams.umontreal.ca/marcs/#Resources.

[10] M. Bosi and R. Goldberg, *Introduction to Digital Audio Coding and Standards*. Massachusetts, USA: Kluwer, 2003.

[11] W. G. Gardner, "Reverberation algorithms," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. New York, USA: Kluwer, 1998, pp. 85–131.

[12] R. Appel and J. G. Beerends, "On the quality of hearing one's own voice," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 237–248, Apr. 2002.

[13] S. Bech and N. Zacharov, *Perceptual Audio Evaluation – Theory, Method and Application*. West Sussex, England: Wiley, 2007.

[14] ITU-R Rec. BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," International Telecommunication Union, Geneva, Switzerland, 1997.