

On the evaluation of the conversational speech quality in telecommunications

Marie Guéguin^{*,a}, Régine Le Bouquin-Jeannès^a, Valérie Gautier-Turbin^b, Gérard Faucon^a,
and Vincent Barriac^b

Abstract

In this paper we propose an objective method to assess speech quality in the conversational context by taking into account the talking and listening speech qualities and the impact of delay. This approach is applied to the results of four subjective tests on the effects of echo, delay, packet loss and noise. The dataset is divided into training and validation sets. For the training set, a multiple linear regression is applied to determine a relationship between conversational, talking and listening speech qualities and the delay value. The multiple linear regression leads to an accurate estimation of the conversational scores with high correlation and low error between subjective and estimated scores, both on the training and validation sets. In addition, a validation is performed on the data of a subjective test found in the literature which confirms the reliability of the regression. The relationship is then applied to an objective level by replacing talking and listening subjective scores with talking and listening objective scores provided by existing objective models, fed by speech signals recorded during the subjective tests. The conversational model achieves high performance as revealed by comparison with the test results and with the existing standard methodology "E-model", presented in the ITU-T (International Telecommunication Union) Recommendation G.107.

Index Terms

speech quality assessment, objective model, conversational context, PESQ, PESQM.

* Corresponding author

^a Laboratoire Traitement du Signal et de l'Image,

INSERM U642, Université de Rennes 1,

Campus de Beaulieu, F-35042 Rennes Cedex, France.

E-mail: {marie.gueguin, regine.le-bouquin-jeannes, gerard.faucon}@univ-rennes1.fr.

^b France Telecom R&D,

TECH/SSTP/MOV,

F-22307 Lannion, France.

E-mail: {valerie.gautierturbin, vincent.barriac}@orange-ftgroup.com.

1 Introduction

Speech quality evaluation has become a huge challenge for telecommunications operators with the evolution of telephony networks. New technologies like GSM or voice over IP (Internet Protocol) have indeed generated new degradations such as packet loss, non-stationary noise, speech distortion due to low-bit rate coding or longer delays due to digital processing. After spotting on the issues encountered in speech quality assessment in Section 2, this paper presents a review of the existing subjective and objective methods for speech quality assessment in Section 3. The review emphasizes the lack of objective signal-based models for conversational speech quality. Section 4 presents an original approach to objectively model the speech quality in the conversational context with the analysis of signals. The subjective experiments conducted to construct and validate this approach are described in Section 5 and the objective model is presented in Sections 6 and 7. Finally, the results obtained for the impairments studied in the subjective experiments are provided in Section 8.

2 Speech quality

2.1 Definition

In a general way, quality is subjective as it depends on the one that judges it. Speech quality is then complex to define: each one has a personal interpretation of a given sound event. This subjectivity firstly plays a role in the perception of the sound event and next in its description. When a sound event occurs, the human auditory system analyzes the signal on its content as well as on its form [1], [2]. If the sound event corresponds to speech, the content (*i.e.* the semantic information) and the form (*i.e.* the acoustic signal) are analyzed. In telecommunications speech quality refers to the quality of the form of the speech signal, however the interpretation of speech quality is influenced by the content of the acoustic signal, in a measure that depends on each person (individual factors). The subjectivity secondly takes place in the description of the sound event, and

thus in the judgment of its speech quality. This judgment depends on the expectation and former experience of each person, that constitutes the internal reference to which each new sound event is compared. So, Jekosch [3] describes speech quality as the result of a perception and judgment process, during which the listener establishes a relationship between what he perceives (*i.e.* the sound event) and what he expects (*i.e.* the internal reference): speech quality is not absolute, but is attributed by the listener.

2.2 Quality criteria

Speech quality is a multidimensional phenomenon [4]: it can be assessed according to different quality criteria. The two principal criteria are loudness and intelligibility (*i.e.* the level and the comprehensibility of the speech signal, respectively), which allow the listener to hear and understand the message of the talker. Other quality criteria have been studied next, such as the agreement (*i.e.* the overall satisfaction of the user concerning the system used), the listening effort, the global impression, the fidelity or the naturalness of the voice. Therefore, speech quality can be studied towards several criteria, being more or less influential and being potentially interdependent. Those methods analyzing speech quality as a multidimensional phenomenon are named analytic methods [5]. Given the large number of possible quality criteria, it seems however very complex to explore all dimensions of speech quality [6]. Speech quality is then often expressed by a scalar: those methods analyzing speech quality as a unidimensional phenomenon are named utilitarian methods [5]. If analytic methods are interesting to understand how speech quality assessment is constructed by the person, utilitarian methods are now the most widely used, in particular for speech quality prediction as it is less complex to model a scalar score than several ones.

2.3 Context

Speech quality perception depends on the context in which the judging person is placed [7]. Three contexts exist: the listening context, the talking context and the conversational context.

2.3.1 Listening context

By definition, the listening context corresponds to the situation in which the participant listens to a vocal message, without speaking. In everyday life, users are placed in this situation when, for instance, they call their answering machine. Such a context can be disturbed by speech distortion due to codec, noise, information loss, signal level. These different impairments decrease speech

quality by affecting intelligibility, voice naturalness or loudness, diminishing the comprehensibility of the vocal message by the subject.

2.3.2 Talking context

In the talking context, the subject speaks, without receiving an answer in return. Users are placed in this context when they record a message on an answering machine. Impairments in this context are principally the distortion of the sidetone signal, the echo and the noise. Contrary to the listening context, it is less obvious how degradations encountered in the talking context affect speech quality, but they can be very disturbing for the talker. When we speak, we perceive our own speech signal (retroaction), transmitted from our mouth to our ear by air path and bone conduction. This retroaction signal allows us to adapt our volume, our pitch, and to control our articulation [8]. The sidetone signal refers to the sounds picked up by the handset microphone and transmitted to the loudspeaker of the same handset, with a low delay (a few milliseconds) [9]. In itself the sidetone signal is not disturbing and is even wanted by the talker: the disturbance it creates depends on its level and its distortion. The distortion of the sidetone signal will disturb the retroaction signal and then make the production of speech more difficult for the talker. Echo is produced by an acoustic coupling or by an electric coupling. The perception of echo depends mainly on two parameters: its delay and its attenuation. The echo will disturb the talker by returning her/him the signal she/he has just pronounced attenuated and delayed. With noise, the retroaction signal is degraded by noise, then the talker will raise the level of her/his voice to compensate this loss of information. This effect is known as the Lombard sign [10], [11]. These different degradations will lead the talker to articulate with more effort.

2.3.3 Conversational context

In everyday life, users are not often placed in a pure listening or talking context during a telephone communication, but more often in a conversational context. In [12], Richards presents a study and a description of the conversation. During a conversation (face-to-face and via a telecommunications system), participants exchange information in turn: they alternatively adopt the roles of listener and talker, and this alternance introduces interaction between the participants. Those roles are not mutually exclusive: in a conversation, participants can both speak at the same time (double-talk) or be both silent at the same time (mutual silence). Richards proposes a 4-state model of the conversation between two interlocutors, as perceived by one participant, provided in the Fig. 1.

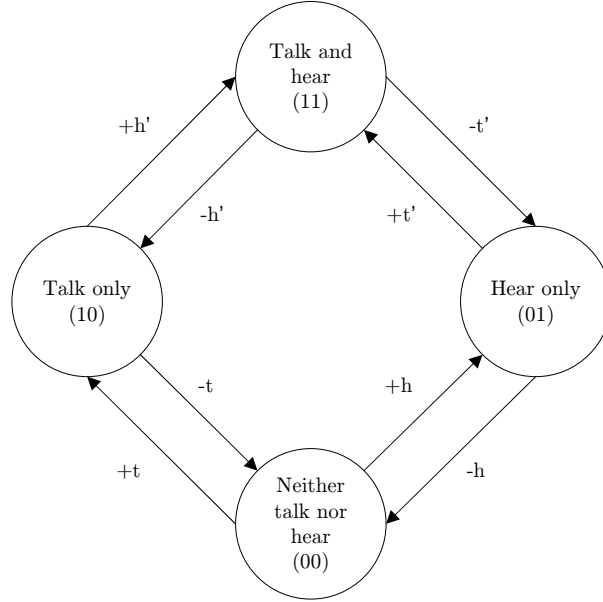


FIG. 1: States and events in conversation as perceived by one participant [12]

The 4 states are the following:

- 00 The participant is neither talking nor listening.
- 01 The participant is hearing speech from his partner but is not talking.
- 10 The participant is talking but not hearing.
- 11 The participant is talking while hearing speech from his partner.

The transition between the 4 states is controlled by the 8 following conversational events:

- +t The participant starts talking while not hearing speech from his partner.
- +t' The participant starts talking while hearing speech from his partner.
- t The participant stops talking while not hearing speech from his partner.
- t' The participant stops talking while hearing speech from his partner.
- +h The participant starts hearing speech from his partner while not talking himself.
- +h' The participant starts hearing speech from his partner while talking himself.
- h The participant stops hearing speech from his partner while not talking himself.
- h' The participant stops hearing speech from his partner while talking himself.

With this description, one notices that conversation, as perceived by one participant, is composed of listening and talking periods, alternating according to the interaction with the interlocutor. On

a speech quality point-of-view, the conversational context is then affected by the degradations encountered in the listening context, those encountered in the talking context, and those affecting the interactivity of the conversation (*i.e.* the delay and the speech quality during double-talk periods).

The delay decreases the interactivity of the conversation by increasing the double-talk and mutual silence periods, as shown in Fig. 2.

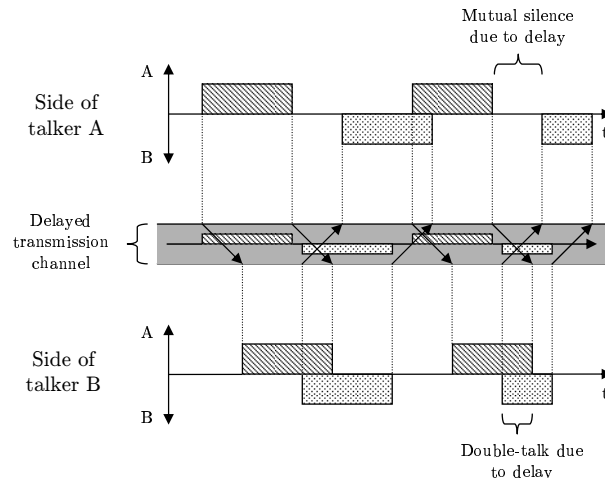


FIG. 2: Impact of the delay on conversation [13]

2.3.4 Synthesis

Speech quality assessment is influenced by several parameters which depend on: (i) individual factors of the judging person (former experience, expectation and mood), (ii) speech signal content, (iii) quality criteria considered, (iv) extra-individual factors (context and environment).

3 Speech quality evaluation review

Speech quality has to be assessed by telecommunications operators, to guarantee the satisfaction of their customers. Subjective tests have been the first method used, as they allow to achieve users' perception.

3.1 Subjective methods

During a subjective test, participants are asked to test a telecommunication system under different conditions of degradation, determined by the test designer, and to score the corresponding

speech quality on a notation scale. As mentioned in Section 2, speech quality is subjective and depends on the one that judges it. Then, participants' scores for a given test condition are averaged to get the mean opinion score (MOS), which allows to decrease the individual subjective effect on speech quality assessment. Moreover, speech quality perception depends on the context in which the person that judges it is placed: different types of subjective tests are then available.

3.1.1 Listening tests

In a listening test participants are placed in a listening situation and listen to audio signals corresponding to different conditions of degradation. Audio signals have been preliminarily recorded by several talkers, all participants listen to the same recordings. In this kind of test, one studies degradations affecting the listening context, such as speech distortion due to codec, noise and packet loss. The International Telecommunication Union (ITU) defines in the Recommendation P.800 [14] the methods for speech quality notation. The most widely used is the ACR method (Absolute Category Rating) with the categories provided in Table 1. One can also mention the DCR method (Degradation Category Rating) with the categories provided in Table 1. Several questions can be asked to the participants to assess different dimensions of speech quality, such as listening quality, listening effort and disturbance due to noise. It is also recommended by the ITU that each test contains reference conditions (*i.e.* without degradation) to provide a reference listening quality to participants.

TAB. 1: Opinion scales defined by the ACR (Absolute Category Rating) method (left) and the DCR (Degradation Category Rating) method (right)

Speech quality (ACR)	Score	Disturbance due to degradation (DCR)	Score
Excellent	5	Imperceptible	5
Good	4	Perceptible but not annoying	4
Fair	3	Perceptible and slightly annoying	3
Poor	2	Annoying	2
Bad	1	Very annoying	1

3.1.2 Talking and listening tests

In a talking and listening test, participants are placed in the talking context: they have to talk in the handset microphone and simultaneously listen to the loudspeaker. One studies impairments

impacting the talking context, such as echo, sidetone distortion and noise. Participants assess the tested conditions with one of the methods defined in Recommendations P.800 [14] and P.831 [15]. As for the listening test, each test must contain reference conditions to anchor participants' judgment. Questions asked in the talking context generally concern overall quality, degradation due to echo and degradation due to noise.

3.1.3 Conversation tests

Conversation tests are designed to evaluate quality in the most realistic situation. Two participants are installed in two separate rooms and have a conversation through the tested telecommunications system. The conditions in this kind of test concern the degradations encountered in the listening and talking contexts, as well as those affecting specifically the interaction such as the delay and the double-talk. The ITU recommends that each conversation test comprises reference conditions. Tested conditions can be either the same for the two participants (symmetric test) or different (asymmetric test). As the goal is to reproduce a realistic telephone conversation, scenarios are generally provided to participants. Short Conversation Test (SCT) scenarios have been created in this purpose [16]. They consider various situations (*e.g.* railway information, travel agency, pizza service, etc.). Each participant then scores the quality of the conversation he has just had according to one of the methods defined in Recommendations P.800 [14] and P.831 [15]. In this kind of test, participants are asked to assess overall quality, degradation due to echo, degradation due to noise and interruption effort. Contrary to listening tests which only require the recording of audio signals in different conditions of degradation and the broadcasting of these recordings to the participants, conversation tests necessitate the conception of a full duplex connection which degrades speech quality in live. Moreover, the recording of conversation speech signals has to be in real time and integrated to the system. Therefore, conversation tests are money- and time-consuming and more rare in the literature than listening tests.

3.1.4 Limits

Whatever the kind of test to be performed, numerous precautions have to be taken in order to control the different sources of variabilities, such as the choice of the participants, the choice of the tested conditions or their order of presentation, and to obtain reliable results. Those tests are then complex and expensive to design. Objective methods are an alternative to subjective methods and allow the automation of speech quality assessment. They have to present a high correlation

with subjective test results, which represent users' judgment. Subjective data are then necessary to build objective models.

3.2 Objective methods

Firstly, simple tools have been used to evaluate speech quality, such as signal-to-noise ratio (SNR), segmental signal-to-noise ratio (SNRseg), mean squared error (MSE), cepstral or spectral distances. Those simple objective measures are not well correlated with subjective data [17], since speech quality is affected by complex degradations, which can mutually mask or emphasize each other. More elaborated objective methods have been necessary. Nowadays, numerous objective models of speech quality exist. They can be classified according to three criteria:

- the measures they are based on: physical measures of the system (parametric) or speech signals (signal-based),
- the information they need: both sides of the system (end-to-end or with reference) or only one side (single-ended or without reference),
- the context they model: listening, talking or conversation.

It is furthermore important to distinguish between two methods of measure of the signals and/or parameters of the system:

- intrusive methods are used in the models with reference. They pass a reference signal through the tested system and capture the degraded signal outputting the system: they disturb the network,
- non-intrusive methods only require the degraded signal (without reference) and can be used in live networks.

Fig. 3 presents the classification of different existing objective methods for speech quality assessment according to these three criteria.

3.2.1 Parametric models

Parametric models use physical measures of the system under test to provide a speech quality score. Among parametric models, the E-model is the most widely used. It has been developed as an end-to-end tool for network designers and standardized in 1998 by the ITU in the Recommendation G.107 [18]. It has been optimized thanks to numerous subjective tests. The E-model outputs a transmission rating factor R , computed from physical measures on both sides of the system under

		listening	talking	conversation
parametric	end to end	G.107 "E model" (1998)		G.107 "E model" (1998)
	single ended	P.564 (2006)		P.562 "CCI" (2000)
signal-based	with reference	P.862 "PESQ" (2001)	PESQM (2002)	
	without reference	P.563 (2004)		

FIG. 3: Classification of existing objective methods for speech quality assessment

test such as the delay, echo, attenuation, room noise, etc. It can be used to estimate conversational quality and listening quality.

The model CCI (Call Clarity Index), described in the ITU-T Recommendation P.562 [19], is the equivalent of the E-model without reference. It assesses conversational speech quality from measures of the system under test (*e.g.* speech level, noise level, echo attenuation) obtained from In-service Non-intrusive Measurement Devices (INMD), described in the ITU-T Recommendation P.561 [20]. This model interprets the measures obtained from INMD to predict conversational quality, as perceived by each user of the communication system.

The ITU-T Recommendation P.564 [21] describes a conformance test for single-ended models of the listening quality in VoIP. It fixes performance objectives that have to be reached by models such as PsyVoIP [22] and [23]. The goal of those models is to use information of IP packets without depacketizing vocal data contained in IP flow, in order to supervise IP network quality in real time. The model computes quality parameters (packet loss rate, packet type and jitter) from the information contained in the Real-Time Protocol (RTP) header. An objective listening quality score is then estimated from these quality parameters.

Parametric models are rapid: they can easily be integrated in network elements and terminals. However, they do not achieve the same performance as signal-based models in estimating users' perception of speech quality.

3.2.2 Signal-based models

Signal-based models, by definition, use the reference and degraded signals (end-to-end or with reference) or the degraded signal only (single-ended or without reference) to predict the speech

quality score of the system under test.

Models with reference pass a reference signal through the system under test, capture the degraded signal, and compare the two signals to get a quality score, which has to be well correlated with the subjective score. Among the models with reference, the most commonly used are those based on a comparison of internal representations specific of the human ear, named perceptual models. This method consists in transforming the physical representation of a signal (measured in decibels, seconds and hertz) into its psychoacoustic representation (measured in sones, seconds and barks) and is based on psychoacoustic principles detailed in Zwicker and Feldtkeller [24]. Among perceptual models, the model known as Perceptual Evaluation of Speech Quality (PESQ) was normalized in 2001 by the ITU as ITU-T Recommendation P.862 [25]. PESQ is intended to measure one-way quality on narrowband telephone signals and models the perceived speech quality in the listening context (mainly impacted by speech distortion due to speech codecs, background noise and packet loss). It leads to a correlation close to 0.935 with subjective data [26], [27].

Single-ended methods analyze signals without known reference. The single-ended equivalent of PESQ has been standardized by the ITU in the Recommendation P.563 [28]. It evaluates listening speech quality under numerous conditions of degradation (distortion due to echo cancellers or noise reduction systems, packet loss, distortion due to codec and ambient noise on send side). It detects speech frames in the degraded signal and extracts a set of parameters for the analysis of the vocal tract and of the unnaturalness of speech, the analysis of strong additional noise, the analysis of interruptions, mutes and time clipping. The final speech quality score is computed thanks to a linear combination of these different quality parameters.

The preceding models work in the listening context. The perceptual model PESQM (Perceptual Echo and Sidetone Quality Measure) [29] assesses quality in the talking context potentially affected by echo and/or sidetone distortion, and leads to a correlation close to 0.9 with subjective data. PESQM has the same principle as PESQ: it compares a degraded signal with the corresponding reference signal. In the talking context, the reference signal is the one pronounced by the participant in the microphone and the degraded signal is the one returned by the system in the loudspeaker of the same participant, potentially affected by echo and/or sidetone distortion.

3.2.3 Limits

As it can be seen from this review of existing objective models summarized in Fig. 3, there is no non-parametric model of the speech quality in the conversational context.

4 Overview of the proposed model to assess conversational speech quality

4.1 Objectives

The proposed objective model of the conversational speech quality has to achieve the following objectives:

- signal-based analysis,
- with or without reference depending on the application scenario,
- electric-electric connexion,
- narrowband connexion,
- call-by-call evaluation.

The last three objectives have been fixed by the Question 20 of the Study Group 12 of the ITU-T aiming at standardizing an objective model of conversational speech quality.

4.2 Proposed method

As mentioned in the paragraph 2.3.3, the conversation, as perceived by one participant, is composed of listening and talking periods, alternating with the interaction with the interlocutor. Based on this observation, the key idea of the proposed approach is to estimate the conversational speech quality from the listening, the talking and the interaction speech qualities. The listening and talking speech qualities are clearly defined both subjectively (standardized subjective methodologies P.800 [14] and P.831 [15]) and objectively (*e.g.* objective models PESQ [25] and PESQM [29]). Interaction speech quality is not well known, except that it is mainly impacted by delay (*cf.* Fig. 2). Then, in our model, we consider the delay value as an indicator of the interaction speech quality, by using the knowledge on the impact of the delay on users' judgment in subjective tests.

The proposed model to assess conversational speech quality, presented in Fig. 4(b), consists in two parts:

- the *integration part* combines the listening quality score, the talking quality score and the delay value to estimate the conversational speech quality score,
- the *measurement part* provides the objective quality scores to the integration part and is based on the existing objective models in the different contexts (*cf.* Fig. 3).

One advantage of this approach is that the integration part is common to all applications, only the objective models of the measurement part change according to the applications.

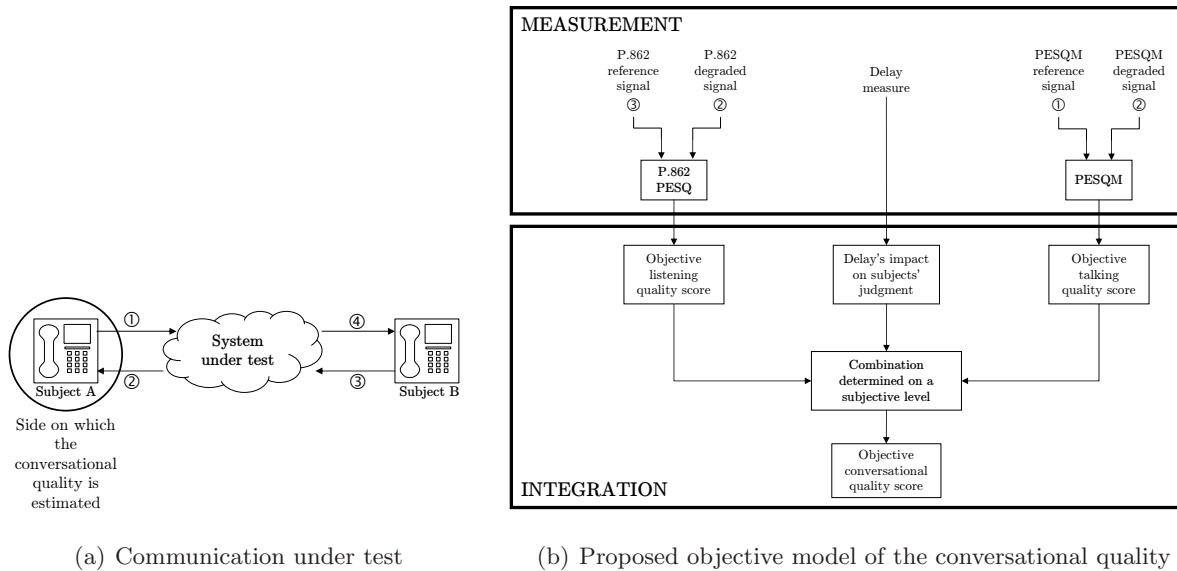


FIG. 4: Scheme of the proposed objective model of the conversational quality

4.2.1 Integration part

During the building of the model, the relationship linking the conversational quality score to the talking quality score, the listening quality score and the delay value is determined through the scores provided by subjective tests investigating different impairments.

4.2.2 Measurement part

The existing objective models in the listening and talking contexts are listed in Section 3.2. In the listening context, the following standardized models are available: P.862 (PESQ) [25] for an intrusive signal-based measure, P.563 [28] for a non-intrusive signal-based measure and P.564 [21] for a non-intrusive parametric measure. In the talking context, no standardized model exists, only the intrusive signal-based PESQM model [29] is available. The measure of delay can be obtained by different manners according to the aimed application, for example from the comparison of reference and degraded signals (if synchronized) by PESQ or from physical measures of the system under test (*e.g.* INMD [20]). In this study, we choose to replace talking and listening subjective scores with objective scores provided respectively by PESQM and PESQ models, which are both intrusive.

4.2.3 Processing steps

The proposed approach comprises the following steps:

- 1) Building of the model: subjective scores of talking, listening and conversation are provided by subjective tests under different conditions of degradation. From these scores and the delay value, a relationship F is determined to provide an estimation of the conversation quality score \widehat{MOS}_{conv} such as:

$$\widehat{MOS}_{conv} = F(MOS_{talk}, MOS_{list}, delay). \quad (1)$$

The building of the model and the determination of the relation F from new subjective tests are described in Section 5.

- 2) Application of the model to a tested communication:
 - a) computation, by the measurement part, of the objective listening and talking scores, by applying objective models to the signals of the tested communication or to the physical parameters of the tested system,
 - b) computation of the value of delay in the tested system,
 - c) computation of the estimated conversational quality score from the objective listening and talking scores and the delay value using the combination F .

5 Subjective tests

The goals of these tests are:

- to check the validity of the hypothesis formulated in Section 4, on the decomposition of the conversational quality into three dimensions (talking, listening and interaction qualities),
- to study and determine the relation F , under different conditions of degradation.

5.1 New subjective test methodology

Given the original approach combining conversational, listening and talking quality scores, a new subjective test methodology is needed. The objective is to assess within a unique subjective test listening, talking and conversational qualities (for both A and B sides of the link) in order to study their relationship. The test is a conversation test involving two non-expert subjects (subject A and subject B) and is split into three successive steps:

- 1) A and B have a free conversation, based on a short conversational scenario [16] (*e.g.* information on flights, hotel room booking, etc.): conversational quality on sides A and B.
- 2) A reads a text (*e.g.* sentences), B listens: talking quality on side A and listening quality on side B.

TAB. 2: Tests conditions

Test 1 - 8 conditions	
Echo level attenuation (dB)	25, >60
One-way delay (ms)	0, 200, 400, 600
Test 2 - 9 conditions	
Packet loss rate (%)	0, 5, 10
Noise level at 'send' side (dB(A))	0, 49, 59
Test 3 - 7 conditions	
Noise type	None, Hoth, Restaurant
Noise level at	48, 53, 59 (Hoth)
'receive' side (dB(A))	51, 57, 63 (Rest.)
Test 4 - 21 conditions	
Echo level attenuation (dB)	20, 30, >60
One-way delay (ms)	0, 200, 400
Packet loss rate (%)	0, 5, 10

TAB. 3: Experimental conditions

Number of subjects	13 to 20 (depending on test)
Network	PSTN (Tests 1, 3), VoIP (Tests 2, 4)
Codec	G.711
Conversation task	Free conversation

- 3) B reads a text (*e.g.* sentences), A listens: talking quality on side B and listening quality on side A.

At the end of each phase, both subjects judge the overall quality of the communication according to the ACR opinion scale provided in [14], presented in Table 1.

Four subjective conversation tests have been performed to develop the objective model. Their test conditions and experimental conditions are summarized in Tables 2 and 3, respectively.

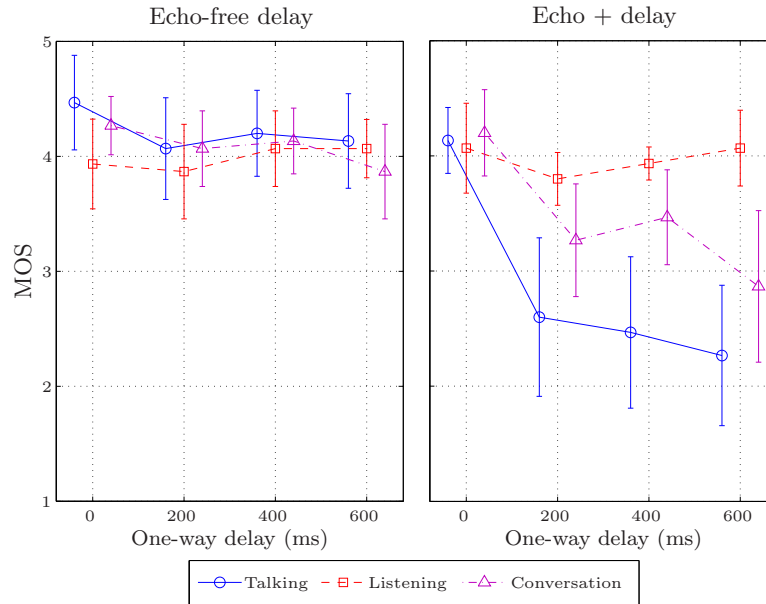


FIG. 5: Echo and delay test results

5.2 Echo and delay

The first test examined the quality in presence of delay and electrical echo, using eight test conditions (given in Table 2) and fifteen couples of non-expert subjects (18 females and 12 males). They communicated with analogue handsets through the switched telephone network (G.711 speech codec). As subjects on side A (11 females and 4 males) were the only ones to undergo delay and echo, only their results are presented here.

The delay impairment was chosen to determine its impact on users' judgment in order to be used in our model. The delay has a special status: despite a numerous literature on the delay, there is no consensus on its effect [30]. According to ITU-T G.114 [31], the upper threshold of one-way delay for an acceptable conversational quality is 400 ms. Moreover, the effect of delay greatly depends on the task of conversation, as it has been demonstrated in [32]. They compared the effect of one-way delay (from 0 to 2 s) on speech quality for 6 different conversational tasks, from very interactive one consisting in reading random numbers in turn to free conversation. The delay was clearly more disturbing and detectable in the very interactive task than in the free conversation.

Therefore, in this test, we choose to study the effect of one-way delay with values below and above the ITU-T G.114 critical threshold of 400 ms (up to 600 ms) and with a free conversation task. The echo impairment was chosen because combined with delay it degrades the talking quality.

Fig. 5 presents the mean opinion scores (MOS) and the corresponding 95% confidence intervals obtained for the overall quality, according to the context (talking, listening, conversation), to the one-way delay value (0, 200, 400 and 600 ms) and to the echo value (no echo and 25 dB-attenuated echo). The curves have been offset horizontally for clarity.

Fig. 5 (left side) shows that, in the case of echo-free delay, the one-way echo-free delay has little impact on subjects' judgment for values between 0 and 400 ms, for this task of conversation (free conversation). The quality decreases slightly between 400 and 600 ms, which is consistent with ITU-T G.114 and test results found in the literature [33]. Regarding these results, we will consider that the delay has an effect only above the threshold of 400 ms, for these conditions of interactivity. The conversational score will then be estimated from talking and listening scores and from the delay value (if it is above 400 ms). Our model could be extended to other conversation tasks and larger values of delay by conducting the appropriate subjective tests.

In the case of echo combined with delay presented in Fig. 5 (right side), echo strongly impacts subjects' judgment, except for a delay of 0 ms (echo not perceptible) and in the listening situation which is not affected by echo. Subjects are more disturbed by echo in the talking context than in the conversational context. Indeed, in the talking context subjects are more attentive to the quality and to its judgment, whereas in the conversational context their attention is shared between the task of conversation and the task of quality assessment, as studied by Gros and Chateau [34].

5.3 Packet loss and transmitted noise

The second test dealt with random packet loss and transmitted noise, using nine test conditions presented in Table 2. The noise was a Hoth noise. It was broadcast in one test room at a time with loudspeakers, and thus was picked up and transmitted to the other room by the microphone of the communication system. Moreover random packet losses (packet length = 32 ms) were introduced in the system, thus degrading the transmitted signal (*i.e.* "speech + noise"). Ten couples of non-expert subjects (10 females and 10 males) participated in this test. Only subjects at 'receive' side of the transmitted noise were asked to judge the overall quality. Subjects communicated with monaural PC headsets and Microsoft NetMeeting software (G.711 speech codec).

The segmental signal-to-noise ratio (SNRseg) is computed at 'receive' side of the transmitted noise for each condition with the following equation:

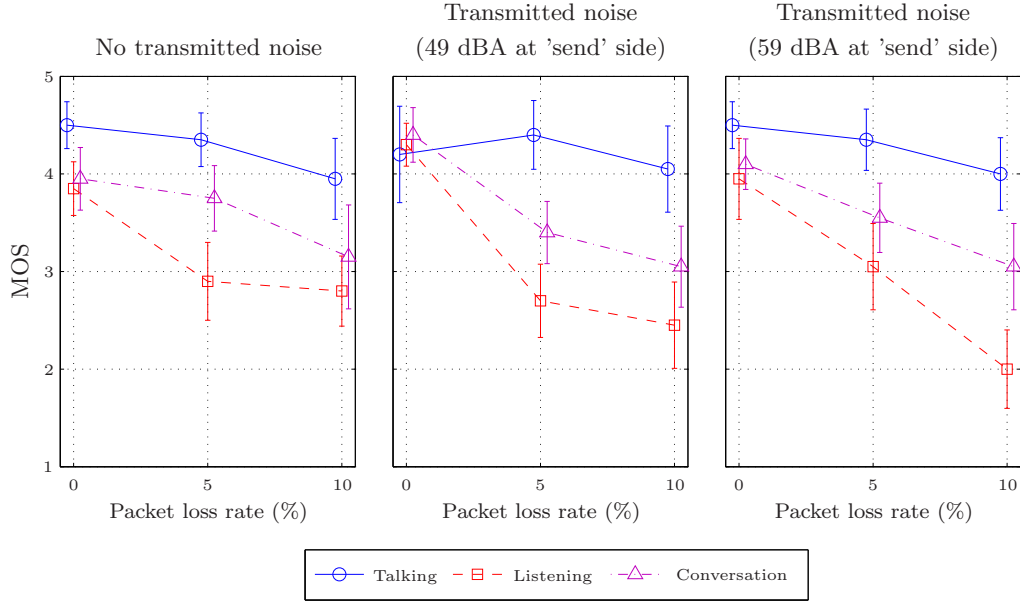


FIG. 6: Packet loss and transmitted noise test results

$$SNR_{seg} = \frac{10}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \log \frac{\sum_{n \in \mathcal{N}} x(n, i)^2}{\sigma_b^2} \quad (2)$$

where \mathcal{S} represents the set of frames that contain speech (active frames) in the noisy signal x , and $|\mathcal{S}|$ its cardinality. \mathcal{N} represents the set of samples in the i th active frame, $x(n, i)$ is the n th sample in the i th active frame and σ_b^2 is the average noise power estimated over non-active frames of the noisy signal x . This analysis is performed with Hamming windows of 256 samples length (32 ms) and 128 samples frame update step.

High signal-to-noise ratios ($SNR_{seg} > 35$ dB) are obtained, because headsets were used. Microphones were very close to subjects' mouth and were also selective, therefore they mostly picked up the speech signal. Moreover the subject at 'send' side raised his voice to compensate the loss of feedback induced by ambient noise (Lombard sign [10], [11]), which explains that the signal-to-noise ratios at 'receive' side are almost identical for the three noise levels. The presence of packet loss affects the transmitted noise, thus increasing the signal-to-noise ratios.

Fig. 6 presents the MOS and the corresponding 95% confidence intervals obtained for the overall quality (judged by subjects at 'receive' side of the transmitted noise), according to the context (talking, listening, conversation), to the packet loss rate (0, 5 and 10%) and to the level of the noise at 'send' side (0, 49 and 59 dBA). The curves have been offset horizontally for clarity.

On Fig. 6, the subjective talking score is almost constant. The subjective conversational score decreases in the same manner with the increase of the packet loss rate, whatever the transmitted noise level. The higher the transmitted noise level, the more the subjective listening score decreases with the increase of the packet loss rate. This shows that, even if the signal-to-noise ratio is high, the transmitted noise is affected by packet losses as well as the speech, and consequently packet losses are audible during the whole communication even during speech silences. So the presence of transmitted noise emphasizes the perception of packet losses. In the conversational context, subjects were probably less attentive to speech quality than in the listening context and thus less disturbed by packet loss and transmitted noise.

5.4 Noise

The third test dealt with noise, using seven test conditions presented in Table 2. The noise was introduced symmetrically in the system and was thus at the same level (electric and acoustic) on both sides. The first noise type was a Hoth noise (stationary noise) and the second one was recorded in a restaurant with people talking (non stationary noise). Seven couples of non-expert subjects (7 females and 7 males) participated in this test. They communicated with analogue handsets through the switched telephone network (G.711 speech codec). The scores of one subject were not exploitable, so the mean opinion scores were computed on thirteen subjects (6 females and 7 males).

The segmental signal-to-noise ratio (SNR_{seg}) is computed at 'receive' side of the noise for each condition according to equation 2. The obtained SNR_{seg} decreases highly as the acoustic noise level increases. The noise is introduced electrically in the circuit and then transmitted without attenuation along the system, which leads to low signal-to-noise ratios in presence of noise.

Fig. 7 presents the MOS and the corresponding 95% confidence intervals obtained for the overall quality, according to the context (talking, listening, conversation), to the noise level and to the type of noise (Hoth, restaurant). In this figure, the condition 'without noise' has been represented for both noise types (Fig. 7 left and right). The curves have been offset horizontally for clarity.

On Fig. 7 (right and left), the subjective talking, listening and conversational scores decrease as the noise level increases. Contrary to the previous two tests, the conversational score is as affected by noise as the listening and talking scores.

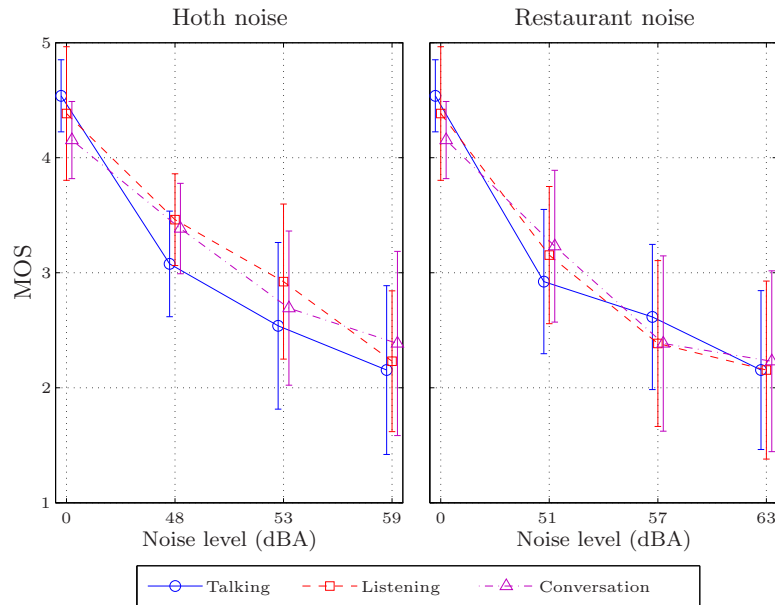


FIG. 7: Noise test results

5.5 Echo, delay and packet loss

The fourth test investigated echo, delay and random packet loss with 21 conditions presented in Table 2. The aim of this test was to study new conditions as well as conditions already tested in previous tests. Nine pairs of non-expert subjects (9 females and 9 males) participated in this test. They communicated with ISDN handsets (G.711 codec).

Fig. 8 presents the MOS and the corresponding 95% confidence intervals obtained for the overall quality, according to the context (talking, listening, conversation), to the packet loss rate (0, 5 and 10%), to the one-way delay value (0, 200 and 400 ms) and to the echo attenuation (no echo, 20 dB, 30 dB). The curves have been offset horizontally for clarity.

On Fig. 8, the subjective listening and conversational scores decrease as the packet loss rate increases. The subjective talking scores decrease as the echo level attenuation decreases. The subjective conversational scores seem less influenced by the delay and echo attenuation than by the packet loss rate.

6 Integration part

The data obtained during the four subjective tests aim at determining the relationship between the conversational speech quality and the listening quality, the talking quality and the delay.

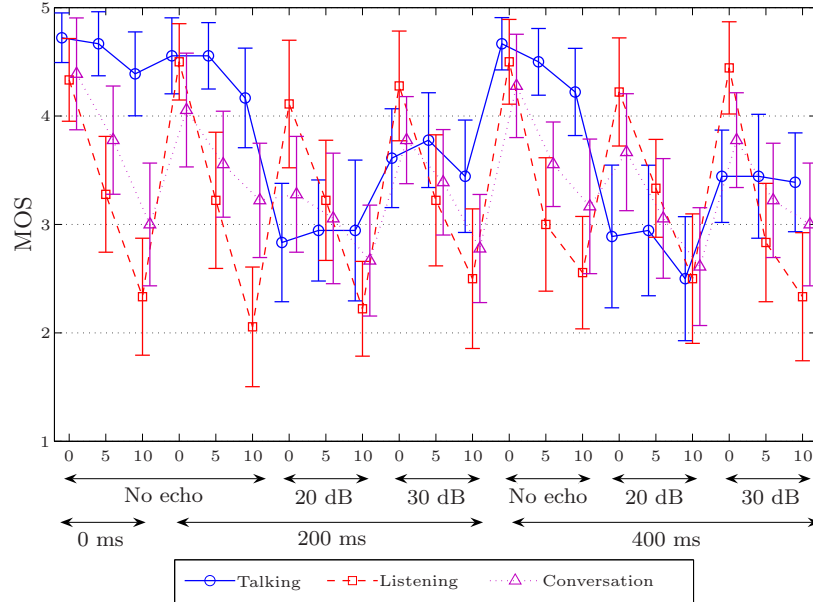


FIG. 8: Echo, delay and packet loss test results

As mentioned above, the results of the test on echo and delay show that the echo-free delay has little impact on subjects' judgment for the task of free conversation and for values of delay below 400 ms. The interaction quality is then taken into account for values of delay above 400 ms.

In this paper, based on the preceding subjective results, conversational quality score is thus estimated from subjective talking and listening quality scores, and from the value of delay above 400 ms. We choose to apply a multiple linear regression, firstly for its simplicity and secondly by analogy with the objective assessment of audiovisual quality. In the literature, audiovisual quality is generally decomposed into two dimensions (audio and visual qualities) and estimated by a multiple linear regression [35]–[37].

Consequently, the conversational speech quality is estimated according to the following regression equation:

$$\widehat{MOS}_{conv} = \alpha \times MOS_{talk} + \beta \times MOS_{list} + \delta \times \max(0, delay - delay_{threshold}) + \gamma \quad (3)$$

where \widehat{MOS}_{conv} is the estimated conversational quality score, MOS_{talk} and MOS_{list} are the subjective talking and listening quality scores respectively, and $delay_{threshold}$ is the threshold above which the delay has an effect (here $delay_{threshold} = 400$ ms). The coefficients α , β , δ and γ are calculated to minimize the mean squared error (MSE) between subjective and estimated conversational scores.

TAB. 4: Multiple linear regression analysis - Training set (Tests 1, 2 and 3)

Predictor	Coef	StDev	t-stat	$Pr > t $
Talking	0.4059	0.037	10.97	.0000
Listening	0.5519	0.045	12.37	.0000
Delay	-1.7376	0.574	-3.03	.0067
(Constant)	0.1710	0.159	1.072	.2964

RMSE = 0.144, $\bar{R}^2 = 0.948$, $F = 148.2$, $p = .000$

The training set consists of the first three tests previously detailed and comprises 24 conditions. The validation set consists of the fourth test detailed above (21 conditions) and of a subjective test provided by the literature (42 conditions) [38] detailed in the following. The coefficients of the regression are computed on the training set and applied to the validation set.

6.1 Training set

The analysis of regression is presented in Table 4, including coefficients values (Coef), their standard deviations (StDev), the significance tests for each predictor (t-stat and $Pr > |t|$), the root mean squared error (RMSE) and the results of the significance test (F statistic and its p -value) for the adjusted multiple coefficient of determination (\bar{R}^2) of the regression. The regression is significant ($F = 148.2$, $p < .05$) and the significance tests for the Talking predictor, the Listening predictor and the Delay predictor indicate that they are all significantly non null ($p < .05$).

The coefficients determined on the training set are applied to the MOS scores of the training set. The means of the correlation coefficient and of the mean absolute error are $\bar{R}_{training} = 0.978$ and $\overline{MAE}_{training} = 0.117$ MOS, respectively.

6.2 Validation set

6.2.1 Validation on the fourth test

These coefficients are then applied to the MOS scores of the fourth test detailed above. The means of the correlation coefficient and of the mean absolute error are $\bar{R}_{validation}^{test4} = 0.969$ and $\overline{MAE}_{validation}^{test4} = 0.167$ MOS, respectively. The regression coefficients computed on the training set lead to high performance on the validation set.

TAB. 5: Conditions of the test found in the literature [38]

Network	RTC (G.711 codec)	VoIP (GSM 6.10 codec)
Background noise	without or 56 dB SPL (Hoth)	without or 56 dB SPL (Hoth)
One-way delay (ms)	0, 150, 300, 500, 600	150, 300, 500, 600, 900
Echo attenuation (dB)	5, 15, >60	5, 15, >60

6.2.2 Validation on the test found in the literature

These coefficients are now applied to the MOS scores of a subjective test described in the literature [38]. This test, whose conditions are detailed in Table 5, comprises 42 conditions (network/codec, noise, delay, echo attenuation), according to the following methodology:

- a non-expert participant communicates with an experimented interlocutor,
- 16 non-expert participants, each one testing 11 conditions out of the 42,
- 4 conversation scenarios [16],
- 2 to 3 minutes of conversation (in czech),
- the non-expert participant judges the listening, talking, interaction and conversation qualities at the end of each conversation.

These conditions comprise some degradations already tested in our own tests and other ones that have not been tested (network/codec). The means of the correlation coefficient and of the mean absolute error are $\overline{R}_{validation}^{literaturetest} = 0.910$ and $\overline{MAE}_{validation}^{literaturetest} = 0.205$ MOS, respectively. The performance of the estimation, provided in Table 6, can be analyzed according to the network. The performance of the estimation is higher with the PSTN than with the VoIP network. It can be explained by the one-way delay of 900 ms tested with the VoIP network, as the proposed model has been trained on one-way delay values below 600 ms. If the conditions with 900 ms-delay are not taken into account, the performance for the VoIP network is better ($R = 0.940$ et $MAE = 0.216$ MOS). These results show that the estimation is efficient for degradations already tested in our own tests, in a different language and for a different codec.

6.3 Synthesis

The results on the training / validation datasets show the feasibility of the proposed approach combining talking and listening quality scores to estimate conversational quality score on a sub-

TAB. 6: Performance of the estimation on the test found in the literature according to the network

Performance criteria	PSTN and VoIP	PSTN	VoIP
R	0.910	0.952	0.831
MAE	0.205	0.154	0.272

jective level and under different impairment conditions. Consequently, the coefficients chosen for the application on an objective level are $\alpha = 0.4059$, $\beta = 0.5519$, $\delta = -1.7376$ and $\gamma = 0.1710$ (*cf.* Table 4). One set of coefficients is then sufficient to estimate the conversational score, under these conditions of degradation.

7 Measurement part

In Section 6 the aim was to optimize the linear combination between the different subjective quality scores. In this section, the aim is to validate the practical setup of the model, using signals acquired under conversation. Consequently, once the relationship between conversational quality score and its different components has been determined on a subjective level in Section 6, talking and listening subjective quality scores are replaced by existing talking and listening objective scores, provided by PESQM [29] and PESQ [25] respectively. Contrary to PESQ, PESQM is not an ITU-T standard so neither source code nor source speech material are available. We optimized PESQM with our own subjective talking test database, with echo, packet losses and noise impairments. We obtained our own mapping function to transform PESQM scores into MOS scores, leading to an average correlation with subjective talking scores of 0.9.

7.1 Source speech material

Speech signals have been recorded during the subjective tests related in Section 5. Test signals are then representative of both male and female talkers. For each phase (described in Section 5) of each condition and for each couple of subjects, four signals are available (A to B, and B to A, on each side of the communication). Each signal is sampled at 8 kHz. As it can be seen in Fig. 4, our approach has four inputs: the reference and degraded signals for PESQ, and the reference and degraded signals for PESQM. For both objective models the reference and degraded signals are those recorded during the conversation phase of each subject. An algorithm, based on vocal activity

detection (VAD), is used to pre-process the conversation signals before application of PESQ and PESQM. Indeed, signals recorded during the conversation phase of subjective tests contain both listening and talking periods, which have to be separated for use with PESQ and PESQM. From the signals recorded during the conversation phase, this VAD-based algorithm detects sequences of conversation where only A speaks (talking phase of A, for PESQM) and sequences where only B speaks (listening phase of A, for PESQ).

7.2 Description of the algorithm

Our algorithm is constituted of four successive steps, described in the following.

7.2.1 Computation of PESQ score

The reference and degraded signals of PESQ are pre-processed, with our VAD-based algorithm, to fit PESQ constraints [39], namely a (reference and degraded) signal length between 8 and 12 seconds, a (reference) signal speech activity (measured according to ITU-T Recommendation P.56 [40]) between 40% and 80%, and a (reference and degraded) signal level of -30 dBov.

For each condition and each subject, we obtained about 4 couples of reference and degraded signals to process with PESQ (depending on the text to listen to). The PESQ score is computed for each couple of reference and degraded signals. For a given condition, corresponding PESQ scores are then averaged to get a unique PESQ score per condition and per subject.

7.2.2 Computation of PESQM score

The reference and degraded signals of PESQM are pre-processed, with our VAD-based algorithm, to obtain a (reference and degraded) signal length between 8 and 12 seconds. For each condition and each subject, we obtained about 6 couples of reference and degraded signals to process with PESQM (depending on the subject's reading speed and on the text to read). The PESQM score is computed for each couple of reference and degraded signals. For a given condition, corresponding PESQM scores are then averaged to get a unique PESQM score per condition and per subject.

7.2.3 Determination of the delay value

In this paper, the delay value is supposed to be known for a given condition. If signals on both sides of the system under test are synchronized, the delay can be determined from an intercorrelation measure.

7.2.4 Computation of estimated conversational score

Once we have a PESQ score, a PESQM score and a known delay value for each condition and each subject, we apply the coefficients α , β , δ and γ as determined in Section 6. We obtain an estimated conversational score per subject and per condition. The final estimated conversational score for each condition is the average of the conversational scores obtained in this condition.

8 Performance and validity of the proposed model

The performance evaluation procedure we chose consists in comparing, for each step of our algorithm, subjective MOS given by subjects and corresponding estimated objective scores. This procedure was performed on our four tests (echo and delay, packet loss and transmitted noise, noise, echo and delay and packet loss). Scores provided by PESQ, PESQM and our objective conversational model are compared to corresponding subjective MOS given by subjects, with the correlation coefficient (R) and mean absolute error (MAE , expressed in MOS). They are presented in Table 7. This performance has been obtained knowing the delay value. In addition, Table 7 presents the performance of our conversational model on a subjective level (*cf.* Section 6) and the performance of the ITU-T standard G.107 [18] known as the “E-model”. The E-model uses 21 measures of the system under test as input parameters, such as the rate of random packet loss or the room noise on ‘receive’ side. The output of the model is named the R-factor, which is a combination of the 21 input parameters. Here, the parameters are set to their default values defined in [18], except for the five parameters varying in the three subjective tests (one-way absolute delay (T_a) = 0, 200, 400 or 600 ms; talker echo loudness rating (TELR) = 20, 25, 30 or 65 dB; random packet-loss probability (Ppl) = 0, 5 or 10%; room noise at the ‘send’ side (P_s) = room noise at the ‘receive’ side (P_r) = 35 (default value), 48, 53, 59, 51, 57 or 63 dB(A)). We assumed the “mean one-way delay of the echo path (T)” to be equal to T_a and the “round-trip delay in a four wire loop (T_r)” to be $2 \cdot T_a$. For an easier comparison, the output R-factor of the E-model is transformed to an objective score on the MOS scale with the formulae defined in the Annex B of the ITU-T Recommendation G.107 [18].

For the whole dataset, the performance for PESQ and PESQM is high both in terms of correlation coefficient and mean absolute error. The accuracy of the proposed conversational model mainly depends on the reliability of the regression determined on a subjective level and on the performance of the objective models PESQ and PESQM. It is then not surprising, given the performance of both the proposed model on a subjective level and the objective models PESQ and PESQM, that

TAB. 7: Final performance of PESQ, of PESQM, of our conversational model (on an objective level and on a subjective level) and of the E-model for each test. R = correlation coefficient, MAE = Mean Absolute Error

Test	Criterion	PESQ	PESQM	Conversational	Conversational	E-model
				model on an objective level	model on a subjective level	
All tests	R	0.883	0.917	0.914	0.966	0.444
	MAE	0.322	0.260	0.198	0.140	1.001
Echo and delay	R	-0.061	0.978	0.927	0.990	0.926
	MAE	0.269	0.139	0.146	0.096	0.960
Packet loss and transmitted noise	R	0.943	0.379	0.929	0.953	0.951
	MAE	0.226	0.182	0.176	0.130	0.315
Noise	R	0.913	0.829	0.951	0.982	0.867
	MAE	0.269	0.462	0.177	0.123	1.062
Echo, delay and packet loss	R	0.926	0.937	0.919	0.969	0.619
	MAE	0.400	0.273	0.235	0.167	1.291

our conversational model on an objective level presents a high correlation coefficient and a low mean absolute error between subjective and estimated conversational scores on the whole dataset.

The Table 7 also presents the analysis of each test.

For the test on echo and delay, the correlation coefficient R corresponding to PESQ is almost null as both subjective and objective listening scores in this test are almost constant (*cf.* Fig. 5). For PESQM and the proposed model, the correlation coefficient R is very high and the mean absolute error low.

For the test on packet loss and transmitted noise, both PESQ and the proposed model achieve high correlation and low error. The correlation coefficient R between PESQM scores and subjective talking scores is almost null as both subjective and objective talking scores are almost constant (*cf.* Fig. 6) and the mean absolute error is low.

For the test on noise, PESQ leads to high correlation and low error. The accuracy of PESQM is bad in terms of mean absolute error. Despite this low performance, our conversational model provides a high correlation and a low mean absolute error with subjective conversational scores.

Generally, the accuracy of the proposed model on an objective level is very close to the one of the proposed model on a subjective model, which represents the performance of the conversational model used with “ideal” listening and talking models. The proposed conversational model clearly outperforms the existing E-model in terms of mean absolute error and is slightly better regarding the correlation coefficient, especially in the conditions on echo, delay and noise. Among the outliers on the mapping between subjective conversational MOS and objective E-model scores, three correspond to conditions with echo (*i.e.* TELR = 25 dB and $T_a = 200, 400$ or 600 ms), which decrease the performance of the E-model. Without these outliers, the E-model achieves $R = 0.768$ and $MAE = 0.542$ MOS, and is still less accurate than our model.

9 Conclusions and further work

In this paper, we propose an approach to objectively model the conversational speech quality from talking and listening speech qualities and the impact of the delay on subjects’ judgment. This approach is applied to the results of three new subjective tests investigating the effects of echo and delay, packet loss and transmitted noise, and noise respectively. We apply a multiple linear regression to determine a relationship between conversational, talking and listening speech qualities, and the delay value. While being simple, the multiple linear regression leads to an accurate estimation of the conversational scores with high correlation coefficient and low error between subjective and estimated scores, for each of the three tests. In addition, a cross-validation is performed on subjects’ scores which confirms the reliability of the regression. This relationship is then applied on an objective level by replacing talking and listening subjective scores with talking and listening objective scores provided by PESQM and PESQ, fed by speech signals recorded during the subjective tests. This objective combination also leads to high performance as revealed by comparison with the test results and with the existing standard methodology E-model [18], then proving the validity of the proposed conversational model.

The conversational model will be improved in the future with the integration of other modules. A first change consists in measuring the value of the delay from the recorded signals, or from the information available on the system under test. Note that this change towards practical application of the model will probably decrease the performance of the model comparing to the present performance. Improvements will also be considered concerning the interaction speech quality for other levels of interactivity and for larger values of delay. New appropriate subjective tests are necessary to determine a critical threshold of delay for each level of interactivity. Then, knowing the level of

interactivity and the measure of delay for a given conversation, the model will determine whether for this level of interactivity the measured delay is disturbing or not. The level of interactivity can be estimated for example by the “conversational temperature” proposed by Hammer *et al.* [13] which is an indicator of conversation interactivity based on the temporal characteristics of a conversation.

In the future, further subjective tests will be performed to: (i) validate our model on a larger database, (ii) extend our model to other impairments or combinations of impairments, such as echo and noise or echo and packet loss, and to determine the (not necessarily linear) corresponding regression equation.

Acknowledgment

The authors would like to thank L. Gros for her help on subjective tests’ design, and the technicians of France Telecom R&D for technical support before and during subjective tests.

References

- [1] J. Blauert, *Spatial Hearing: The psychophysics of human sound localization*. Cambridge MA, USA: The MIT Press, 1997.
- [2] S. Möller, *Assessment and prediction of speech quality in telecommunications*. Boston: Kluwer Academic Publishers, 2000.
- [3] U. Jekosch, “Sprache hören und beurteilen. Qualitätsbeurteilung von Sprechtechnologien als Forschungs- und Dienstleistungsaufgabe,” Habilitation thesis, Essen University, Germany, 2000.
- [4] J. E. Preminger and D. J. van Tasell, “Quantifying the relation between speech quality and speech intelligibility,” *Journal of Speech and Hearing Research*, vol. 38, pp. 714–725, 1995.
- [5] IEEE, “Recommended practice for speech quality measurements,” *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.
- [6] U. Jekosch, “Speech quality assessment and evaluation,” in *Proc. Eurospeech’93*, Berlin, Germany, 1993.
- [7] L. Gros, “Evaluation subjective de la qualité vocale fluctuante,” Ph.D. dissertation, Université d’Aix-Marseille II, France, 2001.
- [8] J. A. Jones and K. G. Munhall, “The role of auditory feedback during phonation: studies of Mandarin tone production,” *Journal of Phonetics*, vol. 30, pp. 303–320, 2002.
- [9] ETSI ETR 250, *Transmission and Multiplexing (TM); Speech communication quality from mouth to ear for 3.1 kHz handset telephony across networks*, Std., 1996.
- [10] E. Lombard, “Le signe de l’élévation de la voix,” *Annales des maladies de l’oreille et du larynx*, vol. 37, no. 2, pp. 101–119, 1911.
- [11] H. Lane and B. Tranel, “The Lombard sign and the role of hearing in speech,” *Journal of Speech and Hearing Research*, vol. 14, pp. 677–709, 1971.

- [12] D. L. Richards, *Telecommunication by speech: The transmission performance of telephone networks*. Butterworths, London, 1973.
- [13] F. Hammer, P. Reichl, and A. Raake, “The well-tempered conversation: Interactivity, delay and perceptual VoIP quality,” in *Proc. IEEE ICC’05*, Seoul, Korea, May 2005.
- [14] *Methods for subjective determination of transmission quality*, ITU-T Recommendation P.800, 1996.
- [15] *Subjective performance evaluation of network echo cancellers*, ITU-T Recommendation P.831, 1998.
- [16] S. Möller, “Development of scenarios for a short conversation test,” ITU-T COM12-35, 1997.
- [17] T. P. Barnwell, “Correlation analysis of subjective and objective measures for speech quality,” *Proc. IEEE ICASSP’80*, vol. 5, pp. 706–709, 1980.
- [18] *The E-model, a computational model for use in transmission planning*, ITU-T Recommendation G.107, 2005.
- [19] *Analysis and interpretation of INMD voice-service measurements*, ITU-T Recommendation P.562, 2000.
- [20] *In-service non-intrusive measurement device - Voice service measurements*, ITU-T Recommendation P.561, 2002.
- [21] *Conformance testing for narrowband voice over IP transmission quality assessment models*, ITU-T Recommendation P.564, 2006.
- [22] A. Rix, S. Broom, and R. Reynolds, “Non-intrusive monitoring of speech quality in voice over IP networks,” ITU-T COM12-D.49, 2001.
- [23] A. D. Clark, “Modeling the effects of burst packet loss and recency on subjective voice quality,” in *Proc. Internet Telephony Workshop (IPTTEL’01)*, New York City, USA, 2001.
- [24] E. Zwicker and R. Feldtkeller, *Das Ohr als Nachrichtenempfänger*. Stuttgart: Hirzel-Verlag, 1967.
- [25] *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.
- [26] A. W. Rix, M. P. Hollier, A. P. 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, 2002.
- [27] J. G. Beerends, A. P. Hekstra, A. W. Rix, and 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 Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, 2002.
- [28] *Single-ended method for objective speech quality assessment in narrow-band telephony applications*, ITU-T Recommendation P.563, 2004.
- [29] R. Appel and J. G. Beerends, “On the quality of hearing one’s own voice,” *Journal of the Audio Engineering Society*, vol. 50, pp. 237–248, 2002.
- [30] G. W. Cermak, “Subjective quality of speech over packet networks as a function of packet loss, delay and delay variation,” *International Journal of Speech Technology*, vol. 5, pp. 65–84, 2002.
- [31] *One-way transmission time*, ITU-T Recommendation G.114, 2003.
- [32] N. Kitawaki and K. Itoh, “Pure delay effects on speech quality in telecommunications,” *IEEE Journal on selected areas in communications*, vol. 9, no. 4, pp. 586–593, 1991.
- [33] S. Möller and A. Raake, “Telephone speech quality prediction: Towards network planning and monitoring models for modern network scenarios,” *Speech Communication*, vol. 38, pp. 47–75, 2002.
- [34] L. Gros and N. Chateau, “The impact of listening and conversational situations on speech perceived quality for time-varying impairments,” in *Proc. MESAQIN’02*, Prague, Czech Republic, 2002.

- [35] C. Jones and D. Atkinson, "Development of opinion-based audiovisual quality models for desktop videoconferencing," in *Proc. 6th IEEE International Workshop on Quality of Service*, 1998.
- [36] T. Tebaldi, "Influence of audio and video quality on subjective audiovisual quality - MPEG-4 and AAC coding," Master's thesis, Technische Universität Wien - Politecnico Di Milano - Institut für Nachrichtentechnik und Hochfrequenztechnik, 2005.
- [37] N. Kitawaki, Y. Arayama, and T. Yamada, "Multimedia opinion model based on media interaction of audiovisual communications," in *Proc. MESAQIN'05*, Prague, Czech Republic, 2005.
- [38] M. Kastner and C. Schmidmer, "Subjective conversational test method and results," ITU-T COM12-58, 2007.
- [39] *Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2*, ITU-T Recommendation P.862.3, 2005.
- [40] *Objective measurement of active speech level*, ITU-T Recommendation P.56, 1993.