14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

A STEP FURTHER TO OBJECTIVE MODELING OF CONVERSATIONAL SPEECH QUALITY

M. Guéguin[1,2,3], *R. Le Bouquin-Jeannès*[2,3], *G. Faucon*[2,3], *V. Gautier-Turbin*[1], *and V. Barriac*[1]

[1]France Télécom R&D, TECH/SSTP/MOV, 22307 Lannion Cedex, France
[2]INSERM, U642, Laboratoire Traitement du Signal et de l'Image, Rennes, France
[3]Université de Rennes 1, LTSI, Campus de Beaulieu, 35042 Rennes Cedex, France
phone: +33(0)296053978, fax: +33(0)296053530, e-mail: *marie.gueguin@francetelecom.com*

ABSTRACT

A new approach to model the conversational speech quality is proposed in this paper. It has been applied to some conditions of echo and delay tested during a subjective test designed to study the relationship between conversational speech quality and talking, listening and interaction speech qualities. A multiple linear regression analysis is performed on the subjective conversational mean opinion scores (MOS) given by subjects with the talking and listening MOS as predictors. The comparison between estimated and subjective conversational scores show the validity of the proposed approach for the conditions assessed in this subjective test. The subjective talking and listening quality scores are then replaced with objective talking and listening quality scores provided by objective models. This new conversational objective model, feeded by signals recorded during the subjective test, presents a correlation of 0.938 with subjective conversational quality scores in these conditions of impairment.

1. INTRODUCTION

From classical telephony to IP or mobile networks, the world of telecommunications has greatly evolved for 15 years introducing new impairments to those already encountered. IP telephony generates packet loss or/and variable delay (jitter), mobile telephony introduces non-stationary noises or/and longer delays. Consequently telecommunication operators need to assess the speech quality of their networks to ensure the quality of service. Subjective tests involve persons testing networks in different conditions and voting on an opinion scale. The mean of their votes in a given condition, named Mean Opinion Score (MOS) [1], gives the quality of the communication link in this condition as perceived by users. Although providing reliable indication of the human perception of speech quality, subjective tests are cost and time consuming. Then objective methods are necessary for telecommunication operators to assess speech quality, being as close to human perception as possible.

Several methods have been proposed since the 1990s (intrusive, non-intrusive, parameter-based or signal-based methods) [2], the most developed being the family of intrusive signal-based models also known as perceptual models. They are based on psychoacoustics considerations and are trained on subjective databases to represent human perception at best. Among these perceptual models, the ITU-T has normalized the perceptual evaluation of speech quality (PESQ) in 2001 as ITU-T Rec. P.862 [3]. PESQ models the listening speech quality which is especially degraded by speech distortion due to codecs, background noise and packet loss. When talking on the phone, the talking quality can also be disturbing as impacted by echo or/and sidetone distortion. Then another perceptual model known as perceptual echo and sidetone quality measure (PESQM) has been proposed by Appel and Beerends [4] to model the talking speech quality. However being efficient in their respective contexts, these models are not able to predict the speech quality in the conversational context in which two persons converse. This context is impacted by the listening and the talking degradations and by the degradations affecting the interaction quality (*i.e.* delay and double-talk quality). Our aim is then to study the conversational speech quality as a combination of the listening, the talking and the interaction speech qualities.

In section 2, we propose a model of conversational quality score. A new subjective test specially designed for this issue and the obtained results are presented in section 3. In section 4, the relationship between conversational quality and talking, listening and interaction qualities is determined on a subjective level by using the results of the subjective test, and the performance of our estimation of the conversational scores is presented. In section 5 this relationship determined on a subjective level is transposed to an objective level and then applied on the signals recorded during the subjective test.

2. CONVERSATIONAL SPEECH QUALITY MODEL

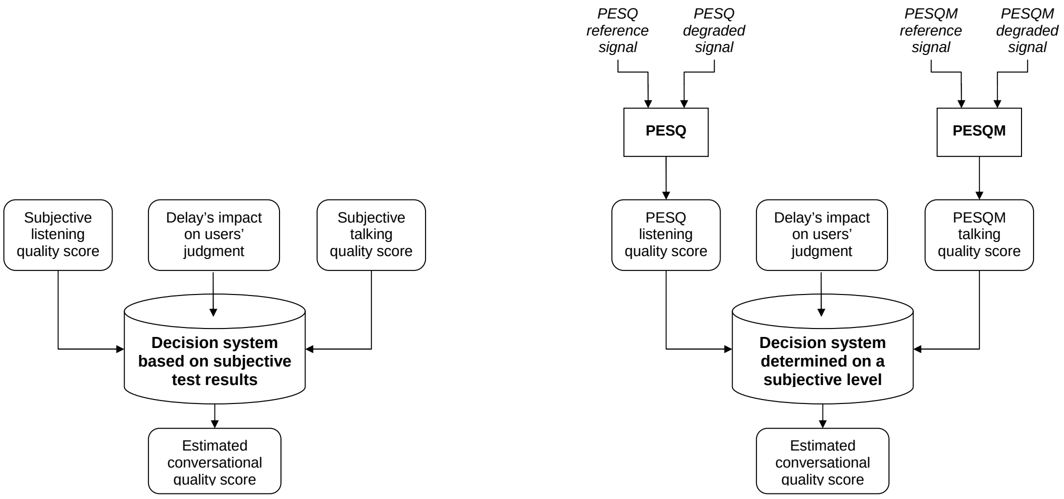
Our model consists in two steps:

Determination on a subjective level of the relationship between the conversational speech quality scores and the listening, talking and interaction speech quality scores,

Transposition on an objective level of the relationship determined on a subjective level.

Our conversational speech quality model combines three metrics: the subjective listening quality score, the subjective talking quality score and the subjective interaction quality score, from which it computes an estimated conversational quality score as close as possible to subjective conversational quality score. Contrary to listening and talking speech qualities which can be assessed during subjective tests thanks to standardized methodologies ([1] and [5], respectively), interaction speech quality is difficult to assess as it has no corresponding standardized methodology. Interaction speech quality is mainly impacted by delay, which decreases interaction between the interlocutors. Then we consider the delay value as an indicator of the interaction speech quality in our model, by using the knowledge on the impact of the delay on users' judgment assessed during subjective tests.

Depending on the impairments affecting the communication, the conversational speech quality is more or less influenced by one of the three metrics, and its relationship with listening speech quality, talking speech quality and delay value changes. To take into account this influence of the impairment on this relationship, our model comprises a decision system which weights the influence of the three metrics on the conversational quality score. Subjective tests are necessary to determine, depending on the impairments, the relationship that links conversational quality score to listening quality score, talking quality score and delay value. Once determined on a subjective level, the decision system can be applied on an objective level by replacing talking and listening subjective scores with objective scores, provided respectively by PESQM and PESQ models. The objective models are feeded by speech signals recorded during subjective tests.


Subjective listening quality score
Delay's impact on users' judgment
Subjective talking quality score
Decision system based on subjective test results
Estimated conversational quality score
PESQ reference signal
PESQ degraded signal
PESQM reference signal
PESQM degraded signal
PESQ
PESQM
PESQ listening quality score
Delay's impact on users' judgment
PESQM talking quality score
Decision system determined on a subjective level
Estimated conversational quality score

(a) Approach on a subjective level

(b) Approach on an objective level

Figure 1: Approaches on subjective and objective levels to estimate conversational quality scores

Fig. 1 presents the two steps of our model. The determination on a subjective level of the relationship between the conversational speech quality score and the listening speech quality score, the talking speech quality score and the delay value is given in Fig. 1(a). Fig. 1(b) describes the transposition on an objective level of the relationship determined on a subjective level.

3. SUBJECTIVE TEST ON ECHO AND DELAY

In order to determine the relationship that links conversational quality score to listening quality score, talking quality score and delay value, we performed a subjective test. We proposed a subjective methodology to study this relationship, which assessed the listening, talking and conversational qualities on both sides of a vocal link within a unique test session [6].

3.1 Description

The conversation-opinion test involves couples of non-expert subjects (A and B) located in two separate rooms. They communicate with analogical handsets through the switched telephone network (G.711 speech codec). For each tested condition, the test is split in three phases. During the first phase, subject A reads a text and subject B listens, to assess talking quality on side A and listening quality on side B. During the second phase, roles are inverted. During the third phase, subjects have a short free conversation to assess conversational quality on both sides. At the end of each phase, both subjects are asked to judge the overall quality on the absolute category rating (ACR) opinion scale of ITU-T P.800 [1] (5 = Excellent, 4 = Good, 3 = Fair, 2 = Poor, 1 = Bad). The test conducted here with this new methodology examined the quality in presence of delay and electric echo, using 8 test conditions, combining 4 conditions of one-way delay (0, 200, 400 and 600 ms) and 2 conditions of echo (no echo and 25 dB-attenuated echo). The delay impairment was chosen to determine its impact on users' judgment to be used in our model presented in Fig. 1. According to ITU-T G.114 [7] the upper threshold of one-way delay for an acceptable conversational quality is 400 ms. However, a recent study [8] reported that users' perception of delay may have changed, new technologies (mobile, IP) getting customers used to longer delays. So we performed this subjective test on the one-way delay with values below and above the ITU-T G.114 threshold of 400 ms. Fifteen couples of non-expert subjects (18 female and 12 male) participated in this test. Only subjects on side A (11 female and 4 male) underwent delay and echo, so only their results are presented here.

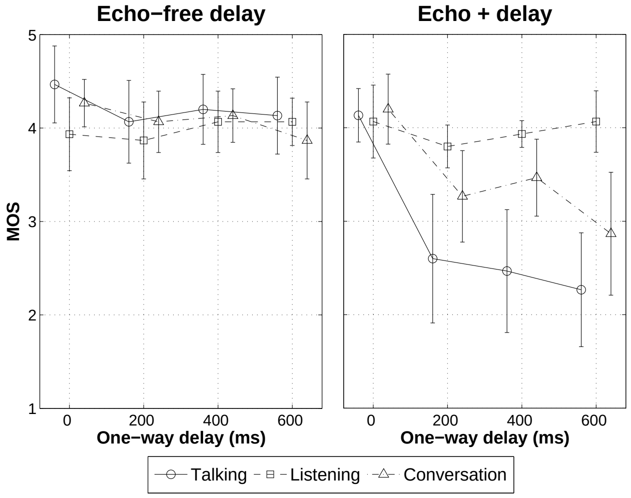


Figure 2: Subjective test results

3.2 Results

In Fig. 2, the mean opinion scores and the corresponding 95% confidence intervals are presented, according to the context (listening, talking, 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.

On Fig. 2 (left side), in the case with echo-free delay, subjects' judgment is almost constant, whatever the delay and the context. These results show that, for values between 0 and 600 ms, the one-way echo-free delay has little impact on subjects' judgment, in these conditions of interactivity. However, larger values of one-way delay (e.g. 800 ms) would probably be perceptible and disturbing for users. Given the results of our test, for these values of delay and in these conditions of interactivity, delay will not be considered in our estimation, and the conversational score will be estimated from talking and listening scores. On Fig. 2 (right side), in the case with echo and delay, the echo has an important effect on the mean overall judgment, except for a delay of 0 ms (echo not perceptible) and in the listening context which is not affected by echo. Subjects' judgment depends on the context, since there is a difference between the scores in the talking context and the scores in the conversation

Table 1: Summary of the multiple linear regression analysis

Predictor	Coef	StDev	t	Pr> |t|
Talking	0.541	0.076	7.106	.00086
Listening	-0.543	0.657	-0.826	.446
(Constant)	4.011	2.563	1.565	.178

RMSE = 0.179, R^2 = 0.911, F = 25.67, p = .0023

Table 2: Summary of the simple linear regression analysis

Predictor	Coef	StDev	t	Pr> |t|
Talking	0.525	0.072	7.314	.00033
(Constant)	1.905	0.262	7.276	.00034

RMSE = 0.175, R^2 = 0.899, F = 53.49, p = .00022

context. Subjects are more disturbed by echo in the talking context, where they are more attentive to the quality assessment than in an interactive context, where their attention is shared between the task of conversation and the task of quality judgment.

4. DETERMINATION ON A SUBJECTIVE LEVEL

4.1 Analysis of regression

The test results show that the one-way delay (echo-free delay below 600 ms) has no great impact on subjects' judgment. To estimate the conversational quality score, we perform an analysis of multiple linear regression from the talking and listening quality scores:

$$\widehat{MOS}_{conv} = \alpha \times MOS_{talk} + \beta \times MOS_{list} + \gamma$$

where MOS_{talk} and MOS_{list} are respectively the subjective talking and listening quality scores, and $\widehat{MOS}_{conv}$ is the estimated conversational quality score. Coefficients α and β, and constant γ are computed to minimize the mean squared error (MSE) between conversational subjective MOS and estimated scores.

Compared to our previous study [9] in which we separated the four conditions with echo-free delay and the four conditions with echo and delay, we choose here to perform the multiple linear regression analysis on the whole set of conditions (the 8 test conditions). Indeed, regrouping the conditions leads to a larger number of trials for the regression analysis and then to a more reliable regression.

The results of the analysis of regression are shown in Table 1, including coefficients values (Coef), their standard deviations (StDev) and the significance tests for each predictor (t and Pr> |t|). In addition, Table 1 displays the root mean squared error (RMSE) and the results of the significance test (F statistic and its p-value) for the multiple coefficient of determination (R^2) of the regression. Although the analysis of regression is significant ($F = 25.67, p < .05$), the significance test on the regression coefficients shows that the coefficient corresponding to the Listening predictor (*i.e.* β) is not significantly different from zero ($p = .446$) and is moreover negative, which was not expected. Indeed, logically when the talking or the listening quality increases (resp. decreases) the conversational quality increases (resp. decreases). These phenomena reflect the near collinearity between the listening quality score with little variation (in this test) and the constant term γ.

The predictors corresponding to non-significant coefficients are rejected, in order to get a more reliable regression. In this test, this leads to a simple linear regression analysis with the Talking predictor, rejecting the Listening predictor (*i.e.* $\beta = 0$). The results of the analysis of the simple linear regression are shown in Table 2. The multiple coefficient of determination (R^2) of the simple linear regression is highly significant ($F = 53.49, p < .05$). The significance tests for the Talking predictor and the constant term show that they are both highly significantly non null ($p < .05$). The simple linear regression provides a lower RMSE than the multiple linear regression, and a slightly lower coefficient of determination (R^2). The adjusted coefficients of determination ($AdjR^2$) of both regressions can be compared to avoid the bias due to the removal of one predictor in the simple linear regression. For the multiple linear regression we obtain $AdjR^2 = 0.858$ and $AdjR^2 = 0.865$ for the simple linear regression, confirming that the simple linear regression is more efficient than the multiple linear regression.

Table 3: Coefficients and performance criteria of the simple linear regression (*i.e.* $\beta = 0$)

α	γ	R	MSE	MAE
0.525	1.905	0.948	0.023	0.112

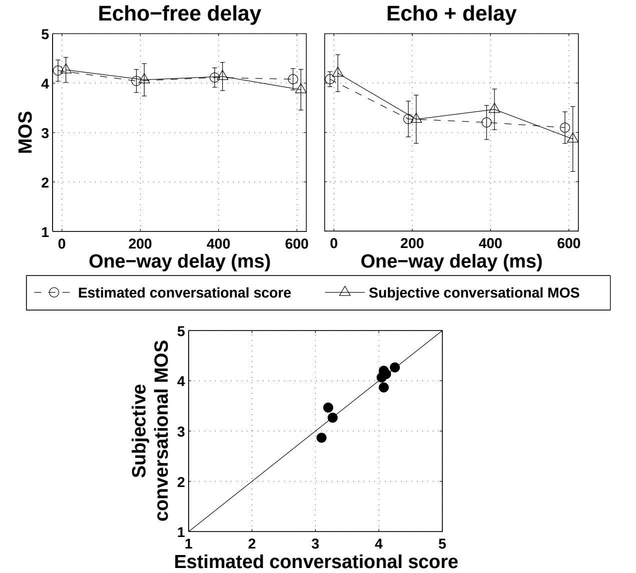
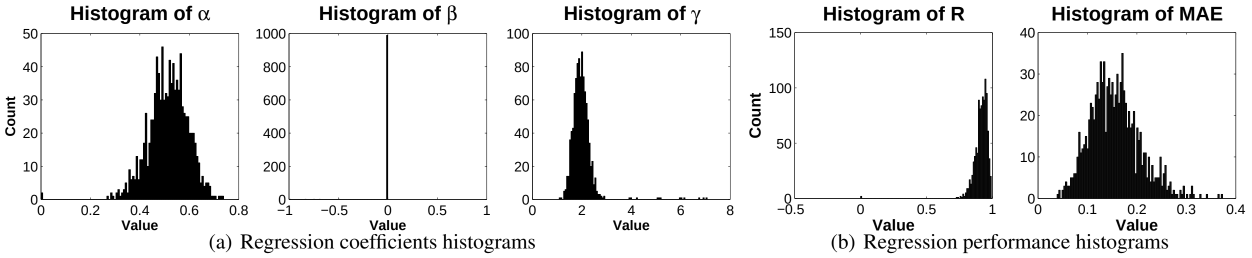


Figure 3: Performance of our conversational model on a subjective level

The obtained regression coefficients are recalled in Table 3. In the same table, the correlation coefficient (R), mean squared error (MSE) and mean absolute error (MAE, expressed in MOS) between subjective and estimated conversational scores are given. The relationship between the subjective conversational scores and the subjective talking and listening scores on a subjective level leads to high performance (high correlation coefficient and low mean absolute error). The estimated conversational scores obtained with the regression coefficients given in Table 3 and the subjective conversational MOS are given in Fig. 3 (above) with the corresponding 95% confidence intervals. The curves have been offset horizontally for clarity. Fig. 3 (below) represents the corresponding mapping between subjective and estimated conversational scores.

4.2 Bootstrap analysis

Given the few data available (8 conditions and 15 subjects), we perform a bootstrap analysis (described in [10]) on the 15 subjects in order to validate our model. At each iteration, a random sample of 15 subjects, with replacement, is drawn. For each condition, scores of the random sample are averaged to get a conversational, a talking and a listening MOS. The analysis of multiple linear regression is performed from these scores and coefficients α, β and γ are deter-



(a) Regression coefficients histograms

(b) Regression performance histograms

Figure 4: Histograms of regression coefficients and performance obtained by bootstrap on subjects

mined. The predictors corresponding to non-significant coefficients are then rejected. 1000 iterations are performed to obtain the distribution of each coefficient. The corresponding histograms are given in Fig. 4(a) and the histograms of the corresponding performance (correlation coefficient R and mean absolute error MAE expressed in MOS) are provided in Fig. 4(b). The histograms of the regression coefficients show that their distributions are quite sharp and centered on the coefficient values obtained with the regression on the whole set of subjects (*cf.* Table 3). The distributions of the regression performance are sharp too and centered around 0.9 for the correlation coefficient and around 0.15 MOS for the mean absolute error. These histograms confirm that whatever the set of subjects considered, the regression is reliable and close to the regression obtained with the whole set of subjects.

5. TRANSPOSITION ON AN OBJECTIVE LEVEL

The regression determined on a subjective level is transposed on an objective level by replacing the subjective talking and listening quality scores with objective talking and listening quality scores, *i.e.* with PESQM and PESQ scores respectively. As PESQM is not an ITU-T standard, no source code is available and we had to implement and optimize it on the basis of the information given in [4] and of a talking subjective test. Our version of PESQM lead to high correlation with subjective talking scores.

5.1 Recorded speech signals

PESQ and PESQM models are feeded by the speech signals recorded during the subjective test presented in section 3. For each phase (described in section 3) 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. Our model on an objective level (*cf.* Fig. 1(b)) has four inputs: the reference and degraded signals of PESQ, and the reference and degraded signals of PESQM. For PESQ the reference and degraded signals are those recorded during the listening phase of each subject, and for PESQM the reference and degraded signals are those recorded during the talking phase of each subject.

5.2 Description

Our algorithm consists in three successive steps:

Computation of PESQ score The reference and degraded signals of PESQ are pre-processed to fit PESQ constraints [11]. The PESQ score is computed for each couple of reference and degraded signals and for each subject.

Computation of PESQM score The PESQM score is computed for each couple of reference and degraded signals and for each subject.

Computation of estimated conversational score The estimated conversational score for each condition and for each subject is computed with the PESQ score and the PESQM score obtained in the corresponding condition and for the corresponding subject, thanks to the coefficients α, β and γ determined in section 4. The final estimated conversational score for each condition is the average of the conversational scores obtained in this condition over all subjects.

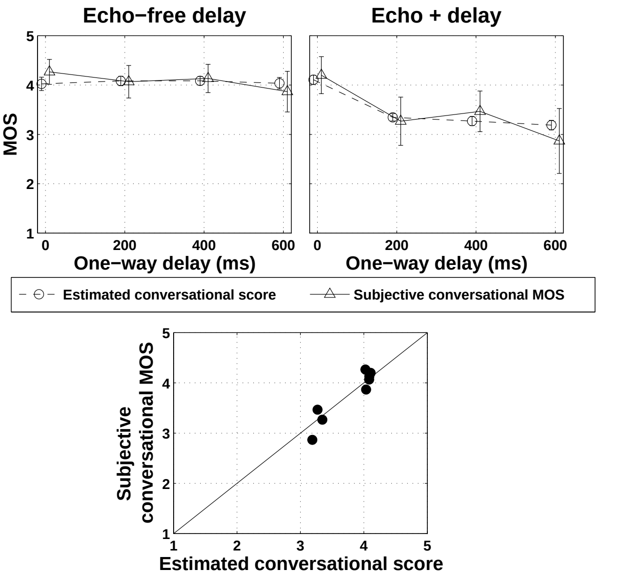


Figure 5: Performance of our conversational model on an objective level

5.3 Performance

The subjective and estimated conversational scores and the corresponding 95% confidence intervals for each condition are given in Fig. 5 (above). The curves have been offset horizontally for clarity. The mapping between subjective and estimated conversational scores is represented in Fig. 5 (below).

The scores provided by PESQ, PESQM and our conversational model are compared to the corresponding subjective MOS given by subjects during the subjective test, in terms of correlation coefficient (R), mean squared error (MSE) and mean absolute error (MAE). These performance criteria are presented in Table 4. For PESQ, the correlation coefficient R is almost null as both subjective and objective listening scores are almost constant and the mean absolute error is relatively high (MAE = 0.374 MOS). For PESQM, the correlation coefficient R is very high and the mean absolute error low, indicating that PESQM is efficient in these conditions of echo and delay. Given the values of the regression coefficients (*cf.* Table 3) in these conditions of impairment, the performance of our conversational model mainly depends on the reliability of the regression determined on a subjective level and on the performance of PESQM. It is then not surprising, given the performance of both the regression analysis (*cf.* section 4) and PESQM, that our conversational model presents a high correlation coefficient and a low mean absolute error between subjective and estimated conversational scores.

Table 4: Final performance of PESQ, PESQM and our conversational model with delay and echo impairments

Performance criterion	PESQ	PESQM	Conversation model
R	-0.076	0.984	0.938
MSE	0.155	0.034	0.031
MAE	0.374	0.144	0.146

6. CONCLUSION AND PERSPECTIVES

In this paper, we propose an approach to model the conversational speech quality from talking and listening speech qualities and delay value (affecting interaction speech quality). This approach is applied to the results of a subjective test dealing with delay and echo. The results of the subjective test show that for values below 600 ms the one-way echo-free delay has only minor effect on subjects' judgment. Then we perform an analysis of multiple linear regression on subjective conversational score with subjective talking and listening scores as predictors. It appears that the subjective conversational score can be estimated from subjective talking score only, thanks to a simple linear regression. This regression results in an accurate estimation of the conversational scores with high correlation coefficient and low error between subjective and estimated scores for the tested conditions. Moreover, a bootstrap analysis on the subjects tends to confirm that this regression is efficient whatever the considered set of subjects. This relationship determined on a subjective level 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, feeded by speech signals recorded during the subjective test. Given the high performance of both the regression analysis and PESQM, our conversational objective model presents a high correlation coefficient and a low mean absolute error between subjective and estimated conversational scores for the tested conditions.

In the future, further subjective tests will be performed to extend the impairment conditions covered by our conversational model and to determine the corresponding relationship (not necessary linear) between conversational, talking and listening speech qualities. As the regression coefficients and equation may change in other impairment conditions, an impairment detector based on physical properties of the recorded signals will be necessary to choose the appropriate regression equation and coefficients.

REFERENCES

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

[2] A. W. Rix, "Perceptual speech quality assessment - A review," in *Proc. ICASSP 2004*, Montreal, Canada, May 17-21. 2004, pp. 1056–1059.

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

[4] R. Appel and J. G. Beerends, "On the quality of hearing one's own voice," *J Audio Eng Soc*, vol. 50(4), pp. 237–248, April 2002.

[5] ITU-T Recommendation P.831, *Subjective performance evaluation of network echo cancellers*, 1998.

[6] ITU-T COM 12-D.45, *Report on a new subjective test on the relationships between listening, talking and conversational qualities when facing delay and echo*, 2005.

[7] ITU-T Recommendation G.114, *One-way Transmission Time*, 2003.

[8] ITU-T COM 12-D.214, *Echo-free delay, VoIP speech quality and the E-model*, 2004.

[9] M. Guéguin, R. Le Bouquin-Jeannès, G. Faucon, and V. Barriac, "Towards an objective model of the conversational speech quality," *ICASSP 2006* (to be published).

[10] A. M. Zoubir and B. Boashash, "The bootstrap and its application in signal processing," *IEEE Signal Processing Magazine*, pp. 56–76, Jan. 1998.

[11] ITU-T Recommendation P.862.3, *Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2*, 2005.