TOWARDS AN OBJECTIVE MODEL OF THE CONVERSATIONAL SPEECH QUALITY

Marie Guéguin[1,2,3]*, Régine Le Bouquin-Jeannès*[1,2]*, Gérard Faucon*[1,2]*, Vincent Barriac*[3]

[1]INSERM, U642, Laboratoire Traitement du Signal et de l'Image, Rennes, France
[2]Université de Rennes1, LTSI, Campus de Beaulieu, Rennes 35042 Cedex, France
[3]France Télécom R&D, TECH/QVP/MOV, Lannion 22307 Cedex, France
E-mail: *marie.gueguin@univ-rennes1.fr*

ABSTRACT

This paper proposes an approach to model the conversational speech quality and an optimisation of this approach to some practical cases. A new subjective test has been designed to study the relationship between conversational speech quality and talking, listening and interaction qualities, when facing echo and delay. The results show that, in these conditions of degradation, the subjective conversational mean opinion scores (MOS) given by subjects can be estimated from the talking and listening quality scores by a multiple linear regression, which coefficients are calculated to minimize the mean squared error (MSE) between subjective and estimated conversational scores. We show the validity of the proposed method in predicting the conversational quality scores for the conditions assessed in this subjective test. For this, a comparison between subjective and estimated conversational quality scores is performed, by means of correlation coefficient and mean absolute error.

1. INTRODUCTION

Although providing new services, recent telecommunications technologies (mobile, voice over IP) have added impairments (speech distortion, longer delays, or packet loss and jitter in VoIP) to traditional telephone speech impairments (echo, delay, sidetone distortion, noise). To satisfy their customers, telecommunications operators have to assess the quality as perceived by users. Subjective methods involve human subjects testing systems in various network conditions and voting on an opinion scale. The scores obtained for each tested condition are averaged to get a mean opinion score (MOS) [1]. These subjective tests are the only way to assess perceived speech quality in telecommunications, but they are complex, cost- and time-consuming.

Consequently objective methods have been introduced to predict the speech quality as perceived by users (for a review, see [2]). They are trained on subjective test results and their performance is evaluated by comparison with subjective scores. Among them, the model known as perceptual evaluation of speech quality (PESQ) was standardized in 2001 as ITU-T Rec. P.862 [3]. PESQ models the perceived speech quality in the listening context (mainly impacted by speech distortion due to speech codecs, background noise and packet loss). Another perceptual model known as perceptual echo and sidetone quality measure (PESQM) [4] models speech quality in the talking context (mainly impacted by echo associated with delay and sidetone distortion). As the listening context, the talking context plays an important role in our perception of speech quality, since distortion or echo of our own voice can be very disturbing when talking. Due to its complexity, no objective perceptual model of speech quality in the conversational context has been developed yet. A conversation between two persons is an alternation of talking, listening and interacting phases [5]. On a speech quality point of view, it is then interesting to study whether the conversational quality can be decomposed in different components (listening, talking and interaction qualities) corresponding to these different roles, as it is suggested by several sources like [6]. We assume the validity of this decomposition in the following of this paper.

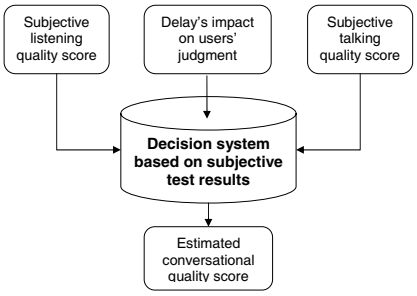

Fig. 1. Approach on a subjective level

In section 2, we propose our approach to estimate the conversational quality score from the talking, listening and interaction quality scores. A new subjective test specially designed for this issue, as well as the results of this test, 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 new subjective test, and the performance of our estimation of the conversational scores is presented.

2. METHOD

Our approach, given in Fig. 1, consists in combining two scores: the talking quality score and the listening quality score given by subjects during a subjective test. It also takes into account the delay which is the main impairment impacting the interaction quality, by using the knowledge on the impact of the delay on users' judgment assessed during subjective tests. This combination of three components (talking quality, listening quality and delay) is not an obvious and simple juxtaposition. The conversational speech quality is more or less influenced by one of the three components, depending on the impairments affecting the communication. It is also not necessarily obvious that these three components have no mutual influence on each other, but we will assume it. We introduce a decision system in

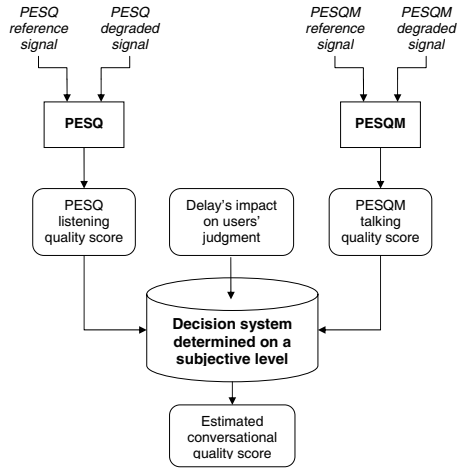

Fig. 2. Approach on an objective level

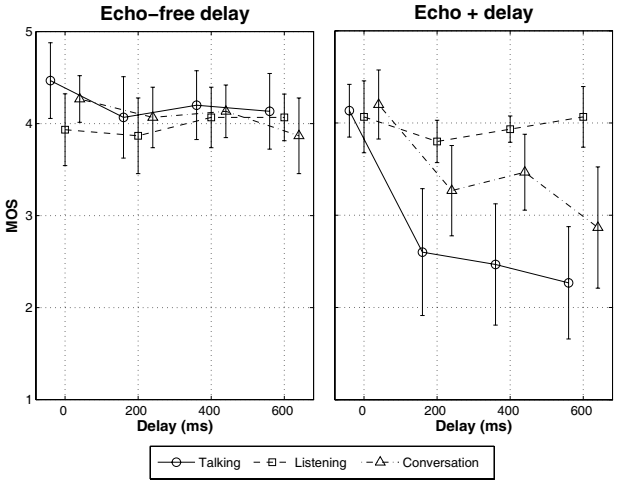

Fig. 3. Subjective test results

our approach, so that it takes into account the influence of the type of impairment on this combination. The decision system weights the influence of the three components on the conversational quality score. Consequently, subjective tests are necessary to determine, depending on the impairments, the relationship that links conversational quality score to talking quality score, listening 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 by objective scores, provided respectively by PESQM and PESQ models from speech signals recorded during subjective tests. Fig. 2 describes this transposition into the objective level.

3. SUBJECTIVE TEST AND RESULTS

This section describes the new subjective test designed to study the relationship between conversational quality and talking, listening and interaction qualities.

3.1. New subjective test and methodology

A subjective test has been conducted to study the relationship between the conversational quality score and the three components (talking quality score, listening quality score and delay value). Since no methodology exists to assess this relationship, we propose a new subjective test methodology. Our methodology determines the listening, talking and conversational qualities on both sides of a vocal link within a unique test session. According to ITU-T P.800 [1], 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 free conversation (using the short conversation scenarios developed in [7]) to assess conversational quality on both sides. At the end of each phase, both subjects are asked to assess 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 eight test conditions, combining:

- 4 conditions of one-way delay: 0, 200, 400 and 600 ms
- 2 conditions of echo: no echo and 25 dB-attenuated echo.

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. The delay impairment was chosen to determine its impact on users' judgment in order to be used in our approach presented in Fig. 1. According to ITU-T G.114 [8] the upper threshold of one-way delay for an acceptable conversational quality is 400 ms. However, a recent study [9] reported that users' perception of delay may have changed, new technologies (mobile, IP) getting customers used to longer delays. Other similar results were obtained in subjective tests by the European Telecommunications Standards Institute (ETSI) and the 3GPP [10]. So it seemed necessary to us to perform a new subjective test on the one-way delay with values below and above the ITU-T G.114 threshold of 400 ms.

3.2. Subjective test results

Fig. 3 represents the mean opinion scores (MOS) and the corresponding 95% confidence intervals obtained for the overall quality, according to the situation, to the one-way delay and to the presence of echo. The curves have been offset horizontally for clarity. In the case with echo-free delay (Fig. 3, left side), subjects' judgment is almost constant, whatever the delay and the situation. These results show that, for values between 0 and 600 ms, 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. The conversational score will then be estimated from talking and listening scores. In the case with echo and delay (Fig. 3, right side), the echo has a strong effect on the mean overall judgment, except for a delay of 0 ms (echo not perceptible) and in the listening situation which is not affected by echo. Subjects' judgment depends on the situation, which indicates that there is a difference between the talking situation and the conversation situation. Subjects are more disturbed by echo in the talking situation, where they are more attentive

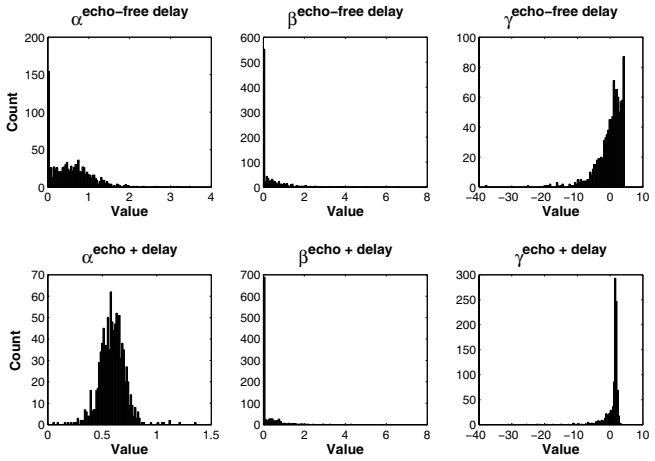

Fig. 4. Histograms of 1000 bootstrap values of regression coefficients in both impairment cases

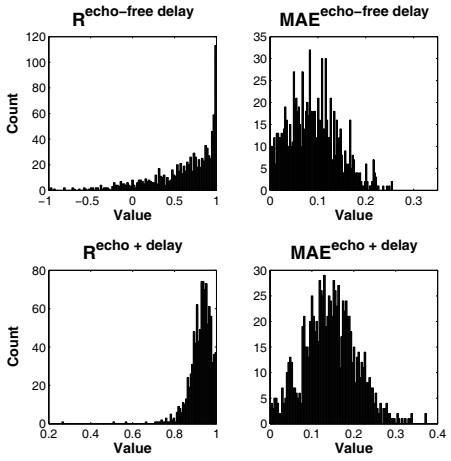

Fig. 5. Histograms of 1000 bootstrap values of regression performances in both impairment cases

to the quality assessment, than in an interactive situation, where their attention is shared between the task of conversation and the task of quality evaluation.

4. ESTIMATION OF THE CONVERSATIONAL QUALITY SCORE

The analysis of these results mainly shows that the echo-free delay has little impact on subjects' judgment in these conditions of interactivity and for these values of delay. So conversational quality is estimated from subjective talking and listening quality scores. We choose to apply a multiple linear regression, for its simplicity:

$$\widehat{MOS}_{conv} = \alpha \times MOS_{talk} + \beta \times MOS_{list} + \gamma \qquad (1)$$

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 calculated to minimize the mean squared error (MSE) between conversational subjective MOS and estimated scores. Logically, when the talking or the listening quality increases (resp. decreases) the conversational quality increases (resp. decreases). Consequently coefficients α and β are constrained to be positive, constant γ can be either positive or negative.

Given the few data (4 points for each impairment case and 15 subjects), we perform a bootstrap to determine the distribution of each value α, β and γ. At each iteration, a random sample of 15 subjects, with replacement, is drawn. For each condition, the scores of the random sample are averaged to get a conversational, a talking and a listening MOS. Values α, β and γ are determined from these scores to minimize the MSE between the conversational MOS and the estimated conversational scores given by (1). 1000 iterations are performed to obtain the distribution of each coefficient, in both cases ("echo-free delay" and "echo + delay"). The corresponding histograms are given in Fig. 4, and the histograms of the corresponding performances (correlation coefficient R and mean absolute error MAE expressed in MOS) are given in Fig. 5.

Fig. 4 shows that the distributions of regression coefficients differ from one case to another. In the "echo + delay" case, coefficients have quite sharp distributions, showing the validity of this regression whatever the sample of subjects considered. α roughly follows a normal distribution with mean of 0.56, β is null at 70%, and γ is located between 1 and 3 at 75%. The corresponding performances are very good, with a correlation coefficient around 0.9 and a mean absolute error around 0.15 MOS. In the "echo-free delay" case, coefficients' distributions are larger. β is null at 55%. α and γ have less clear distributions, presenting peaks at 0 and 4 respectively. The corresponding correlation coefficient has a distribution between -1 and 1, and the mean absolute error is low (around 0.1 MOS). This poor correlation coefficient in the "echo-free delay" case can be explained by the fact that the conversational quality score in this case is almost constant, leading to a correlation coefficient very sensitive to variations.

Table 1. Coefficients and performance criteria of the multiple linear regression with coefficients $\alpha \geq 0$, $\beta \geq 0$ in both impairment cases

Impairment case	α	β	γ	R	MSE	MAE
Echo-free delay	0.00	0.00	4.15	-	0.02	0.12
Echo + delay	0.65	0.00	1.60	0.94	0.03	0.14

We can choose the regression coefficients in both impairment cases, considering their distributions. As coefficient β in both cases is null at least at 55%, we choose β = 0 in both cases. For coefficients α and γ, we compute the 2D-histogram of the couple (α,γ) in both cases, given in Fig. 6. These 2D-histograms both present a maximum peak, corresponding to the most frequent couple (α,γ). We choose α and γ in both cases as the most frequent couple. The obtained coefficients are given in Table 1. The obtained coefficients $\alpha \geq 0$, $\beta \geq 0$ and γ are applied to the entire database (15 subjects). The correlation coefficient (R), mean squared error (MSE) and mean absolute error (MAE, expressed in MOS) between subjective and estimated conversational scores are given in Table 1, for both impairment cases "echo-free delay" and "echo + delay".

In the "echo-free delay" case, coefficients α and β are null and γ = 4.15. Indeed in this case, the three quality scores are almost constant and equal to 4.15, due to the little impact of delay on subjects' judgment. In this particular case, γ = 4.15 actually corresponds to the intrinsic value of the speech quality. The performance of our approach in this case is then difficult to estimate, as quality scores are almost constant. In the "echo + delay" case, the conversational quality score is exclusively estimated from the talking quality score

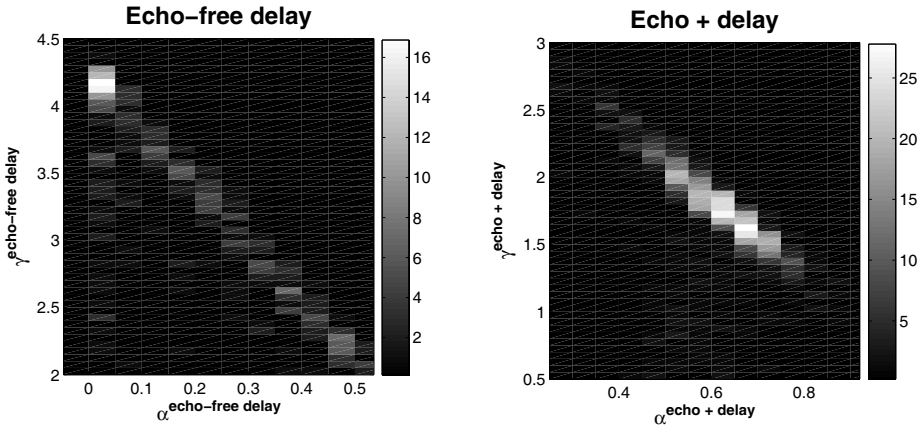

Fig. 6. 2D-histograms of regression coefficients couple (α,γ) in both impairment cases

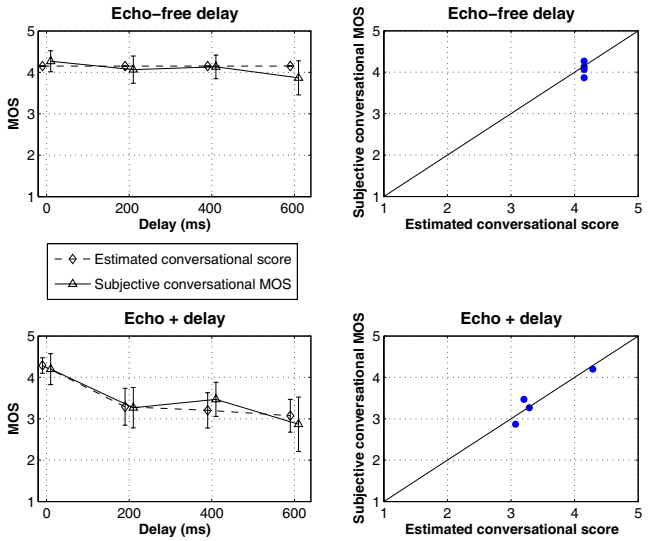

Fig. 7. Subjective and estimated conversational scores in both impairment cases

with a positive offset γ, which is logical given the impairments considered (echo + delay). The corresponding correlation coefficient is very high. The mean absolute error is very low in both cases (MAE < 0.15 MOS). It shows that the regression is efficient and leads to a good estimation of the conversational quality score. The estimated conversational scores obtained with the regression coefficients given in Table 1 and the subjective conversational MOS computed from the entire database (15 subjects) are given in Fig. 7 (left side) in both impairment cases, with the corresponding 95% confidence intervals. The curves have been offset horizontally for clarity. Fig. 7 (right side) presents the corresponding mappings between subjective and estimated conversational scores.

5. CONCLUSION

In this paper, we propose an approach to model the conversational speech quality from talking and listening speech qualities. This approach has been applied to the results of a new subjective test investigating only the effect of delay and echo on subjects without any other impairment. The analysis of the results leads to a relationship between conversational, talking and listening speech qualities by means of a multiple linear regression, which results in an accurate estimation of the conversational scores with high correlation coefficient and low error between subjective and estimated scores, in the conditions with echo. Further subjective tests will be performed to extend the decision system to other impairments and to determine the corresponding relationship (not necessary linear) between conversational, talking and listening speech qualities. Now, the decision system can then be applied on an objective level by replacing talking and listening subjective scores with talking and listening objective scores provided by PESQM and PESQ objective models. An echo detector is also necessary to differentiate the two impairment cases from speech signals recorded during the test.

6. REFERENCES

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

[2] A. W. Rix, "Perceptual speech quality assessment - A review," in *Proc. IEEE ICASSP'04*, 2004, pp. 1056–1059.

[3] ITU-T Rec. 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, pp. 237–248, 2002.

[5] D. L. Richards, *Telecommunication by speech: The transmission performance of telephone networks*, Butterworths, London, 1973.

[6] J. G. Beerends, "A subjective/objective test protocol for determining the conversational quality of a voice link," ITU-T COM12-55, 2003.

[7] S. Möller, "Development of scenarios for a short conversation test," ITU-T COM12-35, 1997.

[8] ITU-T Rec. G.114, "One-way transmission time," 2003.

[9] C. Dvorak and J. James, "Echo-free delay, VoIP speech quality and the E-model," ITU-T COM12-D.214, 2004.

[10] "Packet-switched conversational multimedia applications; Performance characterisation of default codecs (Release 6)," 3GPP TR 26.935, 2004.