euronoise

Influence of informational content of background noise on speech quality evaluation for VoIP application

Adrien Leman[a], Julien Faure[a] and Etienne Parizet[b]

[a]France Telecom, 2, Avenue Pierre Marzin, 22300 Lannion, France
[b]Laboratoire Vibrations Acoustique, Insa Lyon, 25 bis, av. J. Capelle, 69621 Villeurbanne Cedex, France
adrien.leman@orange-ftgroup.com

With the rise of mobile communication systems, the background noise in the speaker's environment and its interaction with VoIP network impairment affects speech quality perception. This effect should be taken into account in non-intrusive models in order to improve accuracy of end user perception measurement. The aim of this study is to determine the impact of information contained in background noise (background speech, environmental sources) on speech quality perception.
A subjective test on the speech quality perception in real network conditions has been done comparing the effect of stationary background noise mainly due to transmission equipment (electrical humming or blowing noises) with the effect of non-stationary environmental noise (public place, traffic noise, background conversation). Interactions between these different background noise condition and realistic network impairments (coders and packets loss) were also studied. The speech quality was evaluated through the Mean Opinion Score determined from an Absolute Category Rating method.

1 Introduction

During a telecommunication, the estimation of speech quality in presence of background noise can be influenced by the signal to noise ratio, the frequency and temporal masking effects [1], and the informational content of noise [2]. However, currents measurement methods of voice quality are often limited and neglect the influence of the informational content of the masking noise.

Previous works suggests that the informational masking effects can affect the speech comprehension [3], [4], [5] or the annoyance prediction from psychoacoustic metrics of the different sound [2], or the meaning in the context of sound quality assessment [6]. In this paper, the informational masking effect is analyzed to evaluate the effect of the meaning of background noise on the quality judgment for telephony application.

For the present study, six background noises were chosen with several characteristics as stationary and non–stationary sound and with or without informational content. Non-stationary background noises are commonly found in our sound environment and are representative of the background noise captured in a real communication. Stationary noises simulated line noise from a telephone line. The six selected noises were:

- A pink-noise considered as the reference (stationary noise, with a -3 dB/oct frequency content);
- A cocktail party noise which was a random noise with a frequency content similar to a standardized human's voice (stationary);
- An Electric noise, i.e. an harmonic sound with a fundamental frequency of 50Hz, simulating a circuit noise (stationary) ;
- A city's environment noise with presence of cars, horns, … (non-stationary) ;
- A restaurant's environment noise with presence of babble, glass, laughter… (non-stationary) ;
- An intelligible voice, recorded from a TV source (non-stationary).

All sounds were sampled at 8 kHz (16 bits); a band pass IRS filter (300 – 3400 Hz) was used to simulate a real telephone network.

To analyze the role of the meaning of background noise on speech quality, the noises must be presented with the same loudness level. To that effect, a preliminary experiment was realized to equalize the perceptual loudness of the five noises according to the reference pink noise and for a given sound pressure level.

Then an experiment was conducted to evaluate the effect on voice quality of the six noises, at three different loudness values, and a third experiment was devoted to interactions between different kinds of noise and classical degradations due to a VoIP communication (codec G711 / G729 and Packet loss ratio 0% / 3%). Results of this experiment were compared with existing models.

2 Preliminary experiment : Loudness equalization of the six background noises

Loudness is defined as the subjective intensity of a sound. The perceptual loudness of stationary noises can be correctly estimated by the model developed by Zwicker [7]. By contrast, no methods are really effective to evaluate the loudness of non-stationary noise. According to [8], an efficient method consists in the adjustment test, which was used in that study.

2.1 Adjustment test

Adjustment test was conducted at three perceptual loudness levels. This test consisted in asking subjects to modify the level of each sound so that its loudness was equal to the loudness of the reference sound (pink noise).

2.2 Stimulus presentation

All sounds were presented through a monaural headphone. According to a normalized communication restitution level at 79 db SPL, three signals-to-noise ratios were chosen at 16, 24 and 32 dB, corresponding respectively to calibrated level for the reference pink noise of 63, 55 and 47 dB SPL.

Zwicker's model [7] was also used to determine the loudness of these three levels, corresponding respectively to 62, 70.5, and 78 phones.

The stimuli were presented in a calm room with a graphics interface developed under Matlab.

2.3 Subjects

A total of twenty expert subjects at FT group between 23 and 41 years old participated to the experiment. All subjects reported normal hearing abilities (no hearing–impairment), and were used to realize this type of subjective test.

2.4 Results

The SPL values of sounds obtained from this experiment and averaged over the 20 subjects are presented on Figure 1.

It can be seen that differences up to 8 dB have to be applied to the sounds so that they appear as equally loud as the pink noise. In particular, electric noise must be lower, which can be explained by the ear functioning and the annoyance due to certain kinds of noise (e.g., alarm, buzzer, bell) [9].

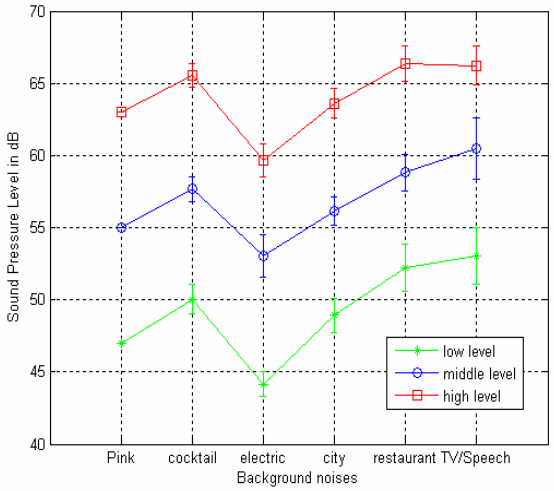

Figure 1 : SPL values of sounds with equal loudness, for three levels of the reference sound (pink noise)

3 First experiment

The first experiment was realized in order to study the interaction between the characteristics and levels of the background noises on speech quality. For that purpose, ACR methodology was used.

3.1 Absolute Category Rating

The Absolute Category Rating (ACR) [10] is a category judgment method in which the test sequences are presented one at a time and are rated independently on a category scale. Subjects have to give their answer before 10 seconds after the presentation of each sequence (in our experiment, the rating time was fixed to 5 seconds). The five-level scale for rating overall quality is called the Mean Opinion Scare (MOS) and is represented by 5 categories of estimated quality:

1→Bad 2→Poor 3→Fair 4→good 5→ Excellent

This procedure is often used for telephony application because it does not need the comparison between stimuli. Therefore, more sound samples can be included in the experiment.

3.2 Stimuli

Eight sentences were selected from a normalized list of double sentences, produced by 4 talkers (2 women and 2 men). They all had the same duration (8 seconds). In this first experiment, these sentences were mixed with the six kinds of background noise at three perceptual loudness levels, determined in the preliminary experiment; and a condition without any background noise was also included. The codec used was G711 without packet loss.

A total of 8*6*3+8 = 152 test stimuli were developed for the analysis, for a duration of 33 minutes.

3.3 Stimulus presentation

The test was realized in a calm room with normalized monaural headphones.

3.4 Subjects

24 subjects between 20 and 59 years old participated to the experiment. All subjects reported normal hearing abilities (no hearing–impairment). They had not been trained to evaluate any specific perceptual auditory characteristics and could, therefore, be considered as naive in this respect.

3.5 Results

A three way ANOVA with repeated measure was used to identify significant effects. It appeared that the three factors (loudness, background noise and sentences) were significant, as well as the interaction between the first ones:

Loudness → F=72,10 / p =0,000***
Background noise * Loudness → F=12,28 / p=0,000***
Sentences → F=11,13 / p= 0,000***
Background noise → F= 4,06 / p=0,002**

By far, the highest F-ratio is obtained by loudness, as can be seen in figure 2; the interaction between loudness and background noise also appears in this figure. The effect of sentences (F=11.13) appeared to be mainly due to one female voice, for which speech quality was always lower than for the other voices.

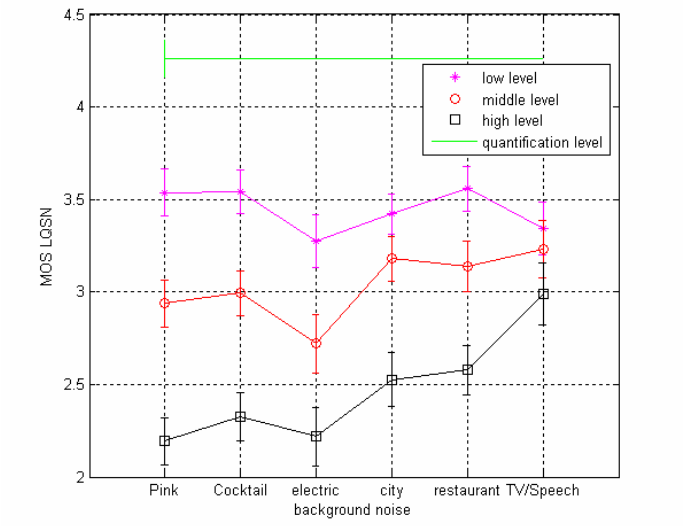

Figure 2 Differences of speech quality for the six kinds of noise for the three loudness levels (the green line represents the situation without background noise)

Figure 2 shows that, for an increase of noise loudness, users are more indulgent with environment's noises than for stationary noises. When noises have a loudness of 62

phones (low-level cf. 2.2), significant differences between the different kinds of background noise cannot be observed. Furthermore, when noises have a level of 78 phones (high level cf. 2.2), an overall difference of 0.5 MOS between the background noises is observed. For the case of TV noise, a lower interaction effect is observed between the kind of background and the noise loudness level. All types of background noise can be separated between noises with informational content (non-stationary noise) and circuit noises (stationary noise). The result is presented in Figure 3, which shows that subjects gave highly MOS values for non-stationary noises at high levels.

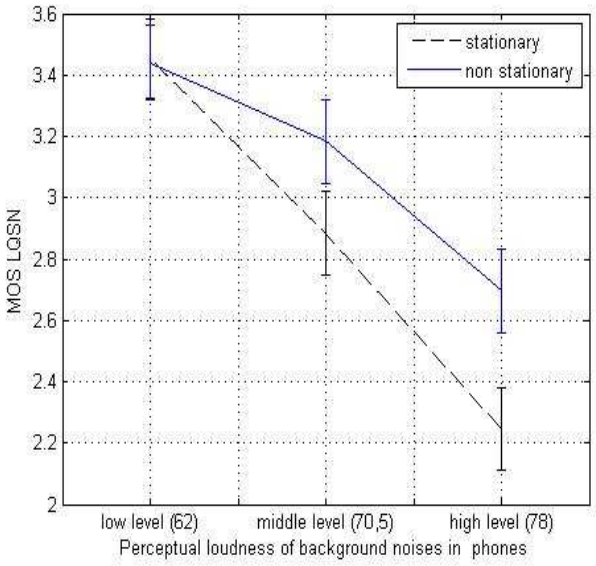

Figure 3 Differences of speech quality between stationary and non stationary noises for the three levels

4 Second experiment

The second experiment was conducted in order to study the interaction between the background noises characteristics and degradations due to VoIP transmission.

As for the first experiment, MOS scores were obtained thanks to a subjective ACR test as describe in section 3.1.

4.1 Stimuli

In the second test, the normalized sentences used in ACR test were mixed with the six kinds of background noise for the middle-loudness level. Then, four kinds of degradations due to the VoIP transmission were introduced:

- Codec G711 with 0% of packet loss
- Codec G711 with 3% of packet losses
- Codec G729 with 0% of packet loss
- Codec G729 with 3% of packet losses

The percentage of packet losses was generated randomly. The degradation pattern was the same for all stimuli. It should be noted that the codec G729 use a packet loss concealment algorithm.

A total of 8*6*4 = 192 conditions were presented to the subjects, for a duration of 42 minutes.

4.2 Subjects

24 naive subjects between 19 and 46 years old realized the second test (they had not participated to any of the previous experiments). All of them reported normal hearing abilities (no hearing–impairment).

4.3 Results

In a first step, the four degradations were ranked from the best to the worse averaged evaluation:

G711 0% → MOS = 4
G729 0% → MOS = 3.6
G729 3% → MOS = 3.2
G711 3% → MOS = 1.9

The best condition was obtained using G711 with 0% of packet loss, but this codec was strongly dependent is very sens on the rate of packet losses (as the worst condition was obtained for G711, 3%).

A three way ANOVA with repeated measure was used for notice significant effects. The highest F-ratio was obtained for percentage of packet losses (F=1223 / p=0.000***), expressing that the time continuity effect plays a predominant role in speech quality. Another important factor was the interaction between codec and percentage of packet losses (F=623.035 / p < 0.000***), which can be explained by the strong effect of packet loss on G711, as mentioned above.

Another significant effect was observed between codec use and different kinds of noise (F=10.88 / p<0.000***). Furthermore, no significant effect between the percentage packet loss and the different kinds of background noise were observed (F=1.66 / p=0.15).

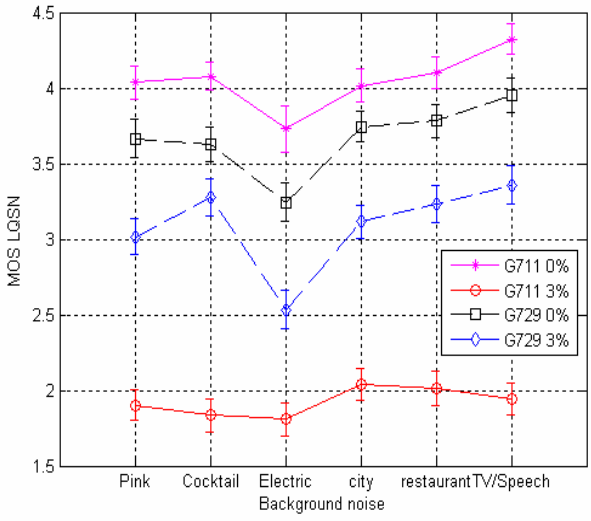

Figure 4 Differences of speech quality for the six kinds of noise for the four types of VoIP degradation

As in the previous experiment, it could be noticed that subjects were more indulgent for noises which could be identified. However, in the case of critics condition (G711; 3% packet loss), the difference between circuit noises and environment noises were lower than in any other conditions.

5 Comparisons between experimental results and existing models

The most commonly used model of voice quality are PESQ [11] and E model [12]. PESQ evaluate speech quality from comparison between original and degraded acoustical signal. E model is based on parametric factor.

Figures 5, 6, 7, and 8 show the speech quality obtained by the two experiments and the corresponding values predicted from the models (PESQ and E). The numbers correspond to the kind of background noise:

1→pink noise 2→cocktail party noise 3→Electric noise

4→city environment noise 5→restaurant environment

6→ intelligible TV noise

The two first figures (5 and 6) show a better correlation for PESQ than E model, when the background noises are presented with different loudness levels (test 1).

In the case of PESQ estimation, we generally observe an overestimation of speech quality especially for the sound mixed with electrics and pink noises (1 and 3).

For the case of E model estimation, all results were overestimated, but particularly in the case of electric and pink noises (1 and 3). PESQ model gave better results in that case (cf. Figure 5, 6) because this model uses the difference between the reference signal and degraded signal for evaluating the speech quality.

On either side, the correlation for interaction between noises and VoIP degradations (Figure 7, Figure 8) was better with E model than PESQ.

However, a slight overestimation can be observed for electrics noise, as well as a slight underestimation for cocktail party and TV noises.

E model is an effective tool to estimate the speech quality when the sound presented packet loss and different use of codec.

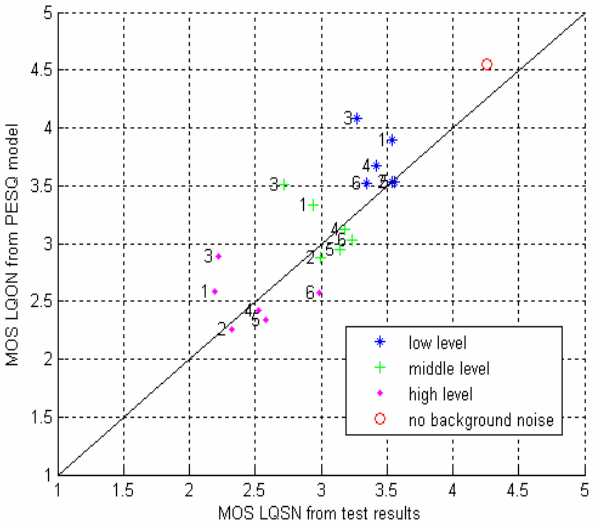

Figure 5 Comparison between test results and PESQ model for three loudness levels. [r=0.91; p<0.001]

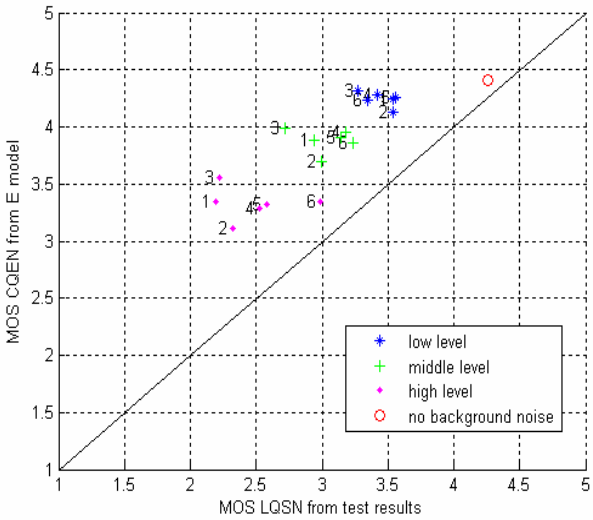

Figure 6 Comparison between test results and E model for three loudness levels. [r = 0.88; p<0.001]

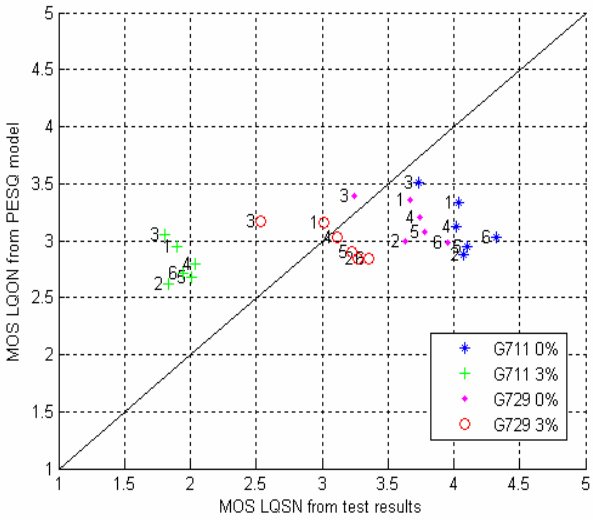

Figure 7 Comparison between test results and PESQ model for four VoIP degradations. [r = 0.48; p<0. 02]

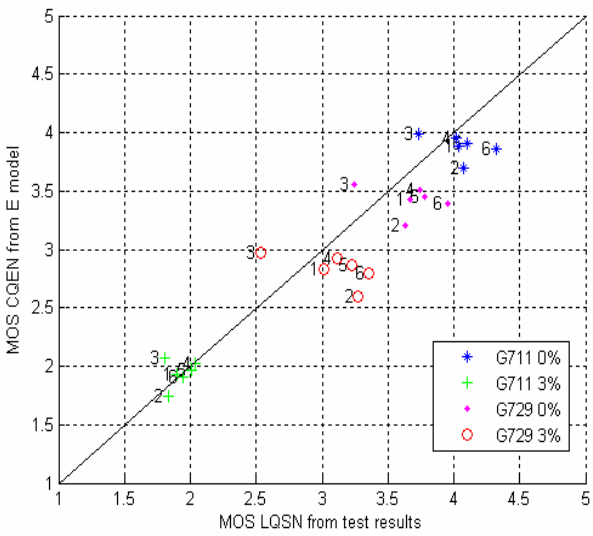

Figure 8 Comparison between test result and E model for four VoIP degradations. [r = 0.94; p<0.001].

6 Conclusion

In this paper, it was demonstrated that voice quality can be influenced by the meaning of noise in telephony context. If users identify the noise to a source present in the environment of the talker, some indulgence effect is noticed on voice quality assessment. Two tests verify this point at first between the different kinds of noise with the different loudness, and subsequently between the different kinds of noise and VoIP degradations. The results specified interactions of different kinds of noise with the loudness and codec. On the other hand, interaction between the percent of packet loss ratio and different noises does not influence Speech quality.

The effect of the difference in speech quality in presence of different kinds of background noise must be taken into account in the objective models for more accuracy. The next step is to find an indicator on the signal to measure and quantify the meaning of background noises, and then to construct an algorithm from the test results.

References

[1] Zwicker, E. and B.Scharf, *A model of loudness summation.* Psychological Review, 1965. **72**: p. 3-26.

[2] Ellermeier, W., Zeitler, A., and Fastl, H. *Predicting annoyance judgments from psychoacoustic metrics: Identifiable versus neutralized sounds*. in *The 33rd International Congress and Exposition on Noise Control Engineering*. 2004. Prague.

[3] Hoen, M., Grataloup, C.-L., et al. *Tomber le masque de l'information : effet cocktail party, masque informationnel et interférences psycholinguistiques en situation de compréhension de a parole dans la parole. .* in *Actes des XXVI journées d'études sur la parole*. 2006. Dinard.

[4] Grataloup, C., Hoen, M., et al. *Influence des paramétres psycholinguistiques du cocktail party sur la compréhension d'un signal de parole cible.* in *Actes des XXVI journée d'études sur la parole*. 2006. Dinard.

[5] Hoen, M., Meunier, F., et al., *Phonetic and lexical interferences in informational masking during speech-in-speech comprehension.* Speech communication, 2007. **49**: p. 905-916.

[6] Jekosch, U., *Meaning in the Context of Sound Quality Assessment.* Acustica, 1999. **85**: p. 681-684.

[7] Zwiker, E., *Übe psychologieshe und methodishe grundlagen der lautheit.* Acustica, 1958. **8**: p. 237-258.

[8] Boullet, I., *La sonie des sons impulsionnels: Perception, Mesures et Modèles.* 2005, Mediterrannée Aix-Marseille 2: Marseille.

[9] Ellermeier, W., Zeitler, A., and Fastl, H. *Impact of Source Identifiability on Perceived Loudness*. in *The 18th International Congress on Acoustics*. 2004. Kyoto, Japan.

[10] Calliope, *La parole et son traitement automatique.* 1989.

[11] ITU-T, R., *P.862 Perceptual Evaluation of Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs.* 2001: Genève.

[12] ITU-T, R., *G.107; The E-model, a computational model for use in transmission planning*. 2003.