A non-intrusive signal-based model for speech quality evaluation using automatic classification of background noises

Adrien Leman, Julien Faure, Etienne Parizet

A non-intrusive signal-based model for speech quality evaluation using automatic classification of background noises

Adrien Leman[1], *Julien Faure*[1], *Etienne Parizet*[2]

[1] France Telecom R&D, 2, Avenue Pierre Marzin, 22300 Lannion, France
[2] Laboratoire Vibrations Acoustique, INSA de Lyon, 25 bis, av. J. Capelle, 69621 Villeurbanne Cedex, France

adrien.leman@orange-ftgroup.com, julien.faure@orange-ftgroup.com, etienne.parizet@insa-lyon.fr

Abstract

This paper describes an original method for speech quality evaluation in the presence of different types of background noises for a range of communications (mobile, VoIP, RTC). The model is obtained from subjective experiments described in [1]. These experiments show that background noise can be more or less tolerated by listeners, depending on the sources of noise that can be identified. Using a classification method, the background noises can be classified into four groups. For each one of the four groups, a relation between loudness of the noise and speech quality is proposed.

Index Terms: speech quality, background noise classification, non-intrusive model, model based on signal

1. Introduction

Previous subjective experiments [1] have shown the influence of loudness of background noise on perceived speech quality. An interesting point to note is that their relation depends on meaning associated with the sources of the background noise. For example, if the listener identifies the noise as coming from a source in the vicinity of the talker, some tolerance was noticed for the voice quality assessment. This can be seen in figure 1, for noises coming from: a city environment, inside of a restaurant, and a television. The evaluation of speech quality is higher for these noise sources than for others which do not have informational content. This result was confirmed in two experiments: the first combined different kinds of noises with three different levels of loudness (cf. figure 1), and the second also involved various voice codecs and IP degradations.

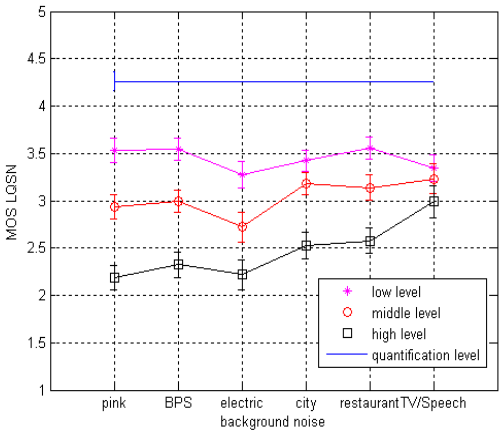

Figure 1: *Differences of speech quality for the six kinds of noises for three loudness levels (the flat line represents the situation without background noise)*

This effect of background-noise-type influencing evaluation of speech quality is not taken into account in existent models, such as ITU-T PESQ/P.862 [2] and ITU-T G.107/E model [3]. The present paper proposes an original model of speech quality perception for situations in which there are different kinds of noises present in the speech signals.
First, models relating loudness and type of the background noise to speech quality will be presented. Then, a classification scheme allowing to separation of the four groups of background noises will be described. This model will be tested using stimuli from the first experiment [1] and the results will be compared to the existing PESQ model [2].

2. Presentation of the proposed model

Subjective experiments [1] lead us to separate background noise into four classes according to the level of tolerance of noise in the assessment of speech quality:

- Class 1 → ***Intelligible noise.*** This group includes noise from music or some other speech. This class of background noise is characterized by high tolerance of noise by subjects concerning the speech quality perception, in comparison with a random noise with same loudness.
- Class 2 → ***Environmental noise.*** This noise has informational content and has given information about the location of the talker, such as city noise, restaurant noise... This class is characterised by light tolerance of noise by subjects concerning the speech quality perception.
- Class 3 → ***Breath noise.*** This noise is stationary and does not contain informational content. Examples are random pink noise, stationary wind noise or stationary speech noise.
- Class 4 → ***Crackling noise.*** This noise does not contain informational content and is stationary, such as electric noise. This class is characterised by a significant decrease in speech quality perception, as assessed by subjects, in comparison to random noise with the same loudness.

For each of the four classes of noise, a relation between loudness of the noise and estimated speech quality with MOS-LQSN score has been determined (figure 2) from the 152 stimuli used in the first experiment [1]. The classes are empirically labelled for each of the 152 sounds. The optimum relation between loudness of the noise and MOS-LQSN is characterized by a logarithmic regression.

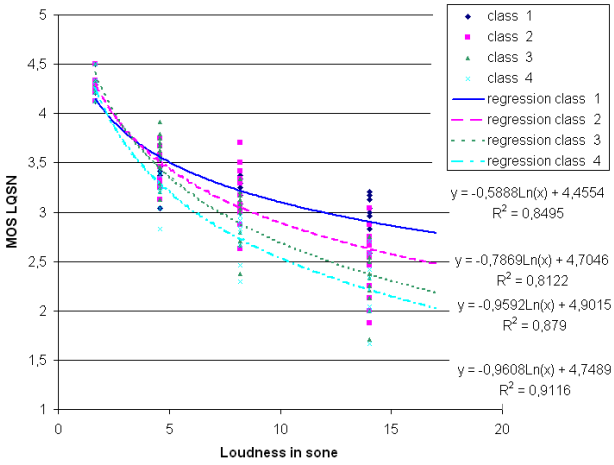

Figure 2: *relation between noise level and speech quality for the four classes of background noises.*

Four steps are necessary in applying the speech quality model. First, voice activity detection (VAD) is used to detect the presence of background noise in the audio signal. This technique is described in [4]. Second, the overall loudness of the background noise is computed using Zwicker's model [5]. Third, the class of background noise is calculated from signal analysis. Finally, a logarithmic relation is applied according to the class of background noise in order to obtain the MOS Score.
The next part of the paper describes the way this classification model is constructed.

3. Classification of background noises

3.1. Methodology

First of all, a set of 632 stimuli was selected, composed of various types of background noises from two experiments explained in [1], and from a public sound database.
Eight indicators were then computed on the 632 background noises of this set; some of them are well-known objective measures for the discrimination between different kinds of sound [6] or recognition of real-life sound and speech [7].

1. ***The signal correlation:*** This is the correlation between the signal and its one sample shifted version (Bravais-Pearson coefficient)
2. **The zero-crossing rate (ZCR) of noise**
3. ***The variation of acoustic power of noise***
4. ***The spectral centroid of noise***
5. ***The spectral roughness of noise***
6. ***The spectral flux of noise***
7. ***The spectral rolloff point of noise***
8. ***The harmonic coefficient***

The classes are empirically labelled according to the tolerance level in the assessment of speech quality, which was obtained after listening of each sound.
The classification tree algorithm [8] was then used to obtain a full decision tree. The entry parameters to the classification tree algorithm were made up of the 8x632 indicators and the class label of each of the 632 sounds. Post-analysis produced an optimum decision tree using only necessary indicators presented above, and keeping a low classification error.

3.2. Stimuli

The set of 632 stimuli was a combination of 344 stimuli used in experiments 1 and 2 (see [1] for detail) and of 288 other stimuli issued from a public sound database.
These 288 new stimuli were composed of 48 new sounds, such as circuit noise, wind noise, car noise, vacuum cleaner noise, hairdryer noise, babble noise, natural noise or music noise, presented with six conditions of degradation.
To simulate noise for narrow band transmission, each noise signal was sampled at 8 KHz and filtered with a band pass IRS filter (300 – 3400 Hz), then encoded and decoded either with G.711 or G.729.
To simulate noise for wideband transmission, each sound was sampled to 16 KHz and filtered with a band pass filter using ITU-T P.341 (50 – 7000 Hz), then encoded and decoded with G.722.
The three coding conditions were presented at two loudness levels (N=63 and 47 dB SPL in the case of random pink noise). Each noise signal was eight seconds long.

3.3. Determination of the different classes

Each used background noise for each degradation was attributed to one of the four classes (by the first author) because some that are, perfectly recognisable without degradation may no longer be as such with the insertion of degradation (encoding-decoding / packet loss). Thus, a better source recognition is observed when noise is encoded and decoded with a wideband codec (G.722) than with a narrowband codec (G.711 or G.729).

3.4. Presentation of the classification model

This section presents the results of the tree classification method explained in section 3.1. In a construction step, 500 stimuli out of 632 were randomly chosen to compute model parameters. The last 132 stimuli were used to validate the model.

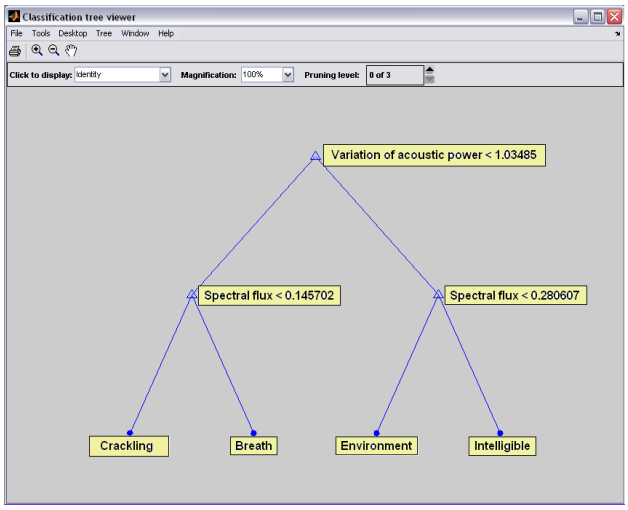

Figure 3: *Decision tree classification of the four classes of noises.*

The optimum decision tree is presented in figure 3. It shows that only two indicators were necessary to classify the different kinds of noises into four classes. The first is a temporal indicator named "**the variation of acoustic power".** The second is a spectral indicator named "**spectral flux"** according to [6].

The temporal indicator represents the time variation of the power of the noise signal. It is defined as the standard deviation of power of all frames of the signal. Power is computed for every frame, each consisting of 512 samples, with an overlap between successive frames of 256 samples (50%), corresponding to a time period of 64 ms per frame, with an overlap of 32 ms with a sample rate of 8 KHz. The acoustics power for frame ***i***, P_i is given by:

$$P_i = 10.\log\left(\frac{1}{N}\sum_{n=1}^{N} x_n^2\right). \quad (1)$$

Where ***n=1…N*** represents the samples on frame ***i*** with N=512, and x_n is the amplitude of sample ***n***. When the background noise is longer than one frame, the value ***IND_TMP*** of temporal indicator is calculated as the standard deviation of acoustic power ***P*** of all frames.

For the noises used, this indicator increases as the noise becomes more and more non-stationary.

The spectral indicator is designated by "IND_FRQ", and is calculated from the power spectral density (PSD) of the background noise. This indicator is determined per frame using 256 samples, corresponding to a time period of 32 ms with a sample rate of 8 KHz. Unlike the temporal indicator, there is no overlap between successive frames. The spectral flux ***SF*** represents how quickly the power spectrum of a signal is changing. ***SF*** for frame ***i***, SF_i is given by:

$$SF_i = 1 - \frac{\sum_k a_k(i-1).a_k(i)}{\sqrt{\sum_k a_k(i-1)^2}\sqrt{\sum_k a_k(i)^2}}. \quad (2)$$

Where a_k represents the PSD value of the frequency components ***k*** of the frame ***i*** or ***i-1***. ***IND_FRQ*** is defined as the mean of SF coefficient for all frames.

Firstly, the proposed tree separates background noise into two categories related to the stationarity of the noise. If the variation of acoustic power is less than IND_TMP=1.03485, then the background noise is considered as stationary, otherwise it is considered as non-stationary.

Secondly, these two main categories can be subdivided using the spectral flux indicator. In the first case of stationary noise, if the value of spectral flux is lower than 0.145, the noise belongs to the class "crackling" otherwise it belongs to the class "breath". In the case of non-stationary noise, if the value of spectral flux is lower than 0.280, the noise belongs to the class "environment" otherwise it belongs to the class "intelligible".

The predictive potential of the tree can be assessed by calculating the number of background noises that are correctly classified.

This proposed tree presents a global percentage of correct classifications of 87.3 %. More precisely, the percentage of correctly classified for each class is as follows:

- 100% for the class "crackling"
- 96.4% for the class "breath"
- 79.2% for the class "environment"
- 95.9% for the class "intelligible"

It appears that "environment" class obtains a proportion of correctly classified noises lower than the other classes. It is caused by the similarity of certain noises that can be classified into two classes, for example, wind noise or hair drier sounds which are between environment and breath noises.

3.5. Validation of the classification model

The 132 stimuli not taken into account during the learning phase were used to verify the robustness of the classification model. The percentage of correctly classified background noises was 91.47%. Table 1 presents the percentages of the correctly classified noises for three combinations of stimuli: all 632 stimuli, the 500 stimuli used during the learning phase of the model of classification, and the 132 stimuli used in the validation phase.

Table 1.*Percentage of correctly classified background noises for three subsets of the stimuli corpus.*

Number of stimuli considered	Percentage correctly classified
632 (total)	87.28 %
500 (learning)	86.20 %
132 (test)	91.47 %

3.6. Advantages and applications

The advantage of the proposed classification model is mainly the low number of indicators used to classify several kinds of background noises into four classes. This model can also be used in real time, for example in telephony applications with an implementation directly at the end of the transmission close to the listener. Moreover, the proposed model of classification is valid with different conditions of degradations like packet loss, or different wideband and narrow band codecs.

The classification model of background noises can be used in many applications. For example, according to the classification result a noise cancellation tool can be used or not. If the noise is judged to be helpful, noise cancellation is not performed, as opposed to situations when the noise is considered to be disturbing. The classification model can also be used to identify the type of noise present in speech, helping to find the origin of the degradation, thus enabling improvement of the quality of service. Furthermore, the classification model can be used with existing speech quality models like G.107 [3] or PESQ [2], to take into account the different kinds of noises present in speech, in order to improve the performance of existing model.

4. Speech quality model using classification model vs PESQ model

Finally, the overall speech quality model based on loudness and on classification, as presented in section 2, was evaluated in two steps.

- Firstly, the model was applied using a perceptual loudness determined by subjective experiment described in [1] in order to evaluate the performance of the proposed model with potential errors due only to classification algorithm.
- Secondly, the model was applied using an estimation of loudness provided by the Zwicker's overall loudness model [5] in order to evaluate the performance of the model in its application (without subjective experiment needs).

The 152 stimuli of experiment 1 described in [1] were used to compare objectives scores with scores from subjective experiment. PESQ model obtained a correlation of **R=0.91 (p<0.001)**(see [1]). The PESQ evaluation performance was

taken as reference score and was compared with performance of the proposed speech quality model.

4.1. Speech quality model using perceived loudness

In this step, the speech quality model was evaluated with a classification model and subjective loudness.

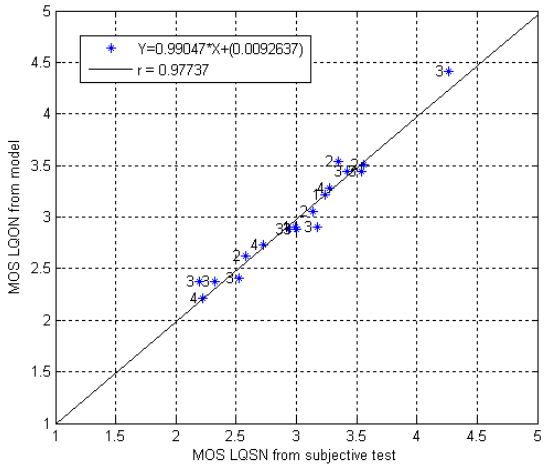

Figure 4: *Comparison of MOS between subjective experiment and speech quality model using classification model*

The labels represent the background noise classes from 1 to 4, obtain by the classification model, as defined in section 2. The performance of both methods of regression functions and classification model is measured by the correlation coefficient between the MOS scores issues from the subjective experiment and the scores estimated by the model. The correlation score was **R=O.98, (p<0.001).**

4.2. Speech quality model using calculated Zwicker's loudness model

In this step, the speech quality model was evaluated with a classification model and Zwicker's loudness model.

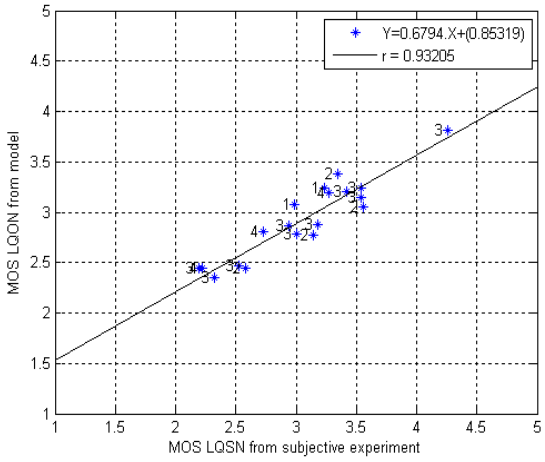

Figure 5: *Comparison of MOS scores between subjective experiment and the speech quality model using classification model and* Zwicker's *loudness model*

The labels represent the background noise classes, as defined section 2. When Zwicker's loudness model is used, the correlation coefficient was **R=0.93, (p<0.001).** This model is not as effective as using subjective loudness, but it is still more accurate than the reference PESQ model (**r = 0.91**). In the future, Moore's loudness model will be tested to compare the accuracy of the two loudness models.

5. Conclusions

The present article demonstrates that the different types of background noises present in speech signal should be taken into account in speech quality models. The developed non-intrusive model achieves a very good performance in comparison with existing intrusive models like PESQ model. It uses only two indicators issued from signal analysis and the calculated loudness. In this sense, it could be employed in real time to evaluate speech quality on a telephony network. An interesting point to note is that the proposed speech quality model, using classification, as well as giving a MOS score, allows identifying the type of noise present in speech. This can be helpful in supervision of tasks in communication networks, thus improving the quality of service.

6. References

[1] A. Leman, J. Faure, and E. Parizet, "Influence of informational content of background noise on speech quality evaluation for VoIP application," presented at Acoustics 08, Paris, 2008.

[2] R. IUT-T, "P.862 Perceptual Evaluation of Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," 2001.

[3] R. IUT-T, "G.107; The E-model, a computational model for use in transmission planning," 2003.

[4] R. ITU-T, "P.56 Mesure objective du niveau vocal actif," 1993.

[5] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*: Springer; 2nd updated ed. edition (April 14, 1999), 1999.

[6] E. Didiot, "Segmentation parole/musique pour la transcription automatique de la parole continue." Nancy: Henri Poincaré, 2007, pp. 131.

[7] D. Istrate, M. Vacher, and J. f. Serignat, "Détection et classification des sons : application aux sons de la vie courante et à la parole," 2005.

[8] L. Breiman, J. Frieman, R. Olshen, and C. Stone, *Classification and regression trees*: Chapman, 1993.