A Perceptual Objective Measure for Noise Reduction systems

Valérie Gautier-Turbin, Nicolas Le Faucheur

France Télécom R&D, France

valerie.gautierturbin@francetelecom.com, nicolas.lefaucheur@francetelecom.com

Abstract

Evaluation of noise reduction systems is mandatory to ensure that speech quality is improved or at least preserved. The only ITU standardized methodology is a subjective one now defined in ITU-T Recommendation P.835 [1]. However, subjective testing being time consuming, costly and not easy to perform, objective methods are always wanted by network providers or algorithms designers. In this paper, we present a new objective measure for the evaluation of noise reduction systems. Based on the use of perceptual criteria, this new objective measure is very promising and efficient: results yield a very high correlation with the subjective procedure.

Keywords

Noise reduction, objective evaluation, psychoacoustic.

1 Introduction

Typically, noise reduction features allow to reduce the background noise level during speech calls. The noise reduction algorithm continuously estimates the background noise power spectrum based on input speech samples and voice activity detection. However, since a filtering based on a background noise estimate is applied, the noise reduction processing is never perfect and some audible degradation may then occur on the speech signal. Performance of the noise reduction is deeply dependent on the background noise estimate and varies with implemented algorithms (spectral subtraction, frequency filtering, in-band filtering, time domain filtering, etc). Noise reductions features often adversely affect the speech component as more noise is suppressed: there tends to be increasing degradation of the speech or signal component as more of the noise or background component is removed. In this situation, subjects can often become confused as to what they should be responding to in their ratings of the overall "quality" of the waveforms: while the background may have been improved because there is less noise present in the waveform, the speech signal may have been degraded in the process.

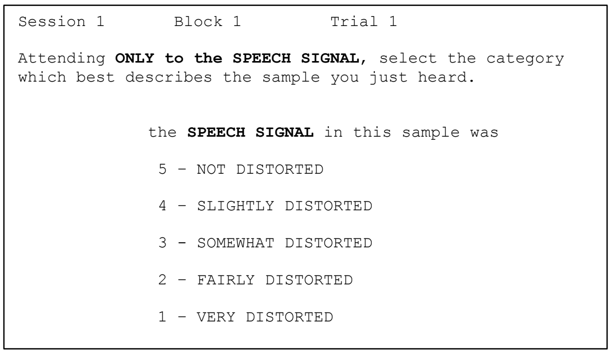
Before adding a noise reduction feature in a network or terminal, performance evaluation is therefore mandatory or highly recommended to ensure that overall quality is improved or at least preserved. However, this is not an easy task: definition of subjective and objective evaluation methodologies was under study, in particular at the ITU (International Telecommunications Union), for the past 4 years and is still under study for the objective part [2]. However, a subjective methodology for evaluating noise reduction systems was defined and is described in ITU-T Recommendation P.835.

In this paper, we first propose a presentation of the current standardized subjective evaluation methodology for noise reduction systems. Based on the approach of this subjective procedure, we designed a new objective measure described in section 3. Results presented in section 3.4 show that the use of human hearing criteria leads to a very efficient and reliable objective measure since objective and subjective results are very highly correlated.

2 ITU-T Recommendation P.835

Methods for subjective determination of transmission quality are defined in ITU-T Recommendation P.800 [3]. However, these methods are inappropriate for the evaluation of noise reduction systems because they use a single-scale rating. Studies have shown that noise reduction algorithms evaluation is a bi-dimensional problem: some subjects rate the quality depending on the amount of the reduced noise (they do not pay attention to degradation) and others rate the quality only depending on the presence of distortion on the speech signal. To solve these issues, ITU-T Recommendation P.835 "Subjective test methodology for evaluating speech communication, systems that include noise suppression algorithm" was defined and approved in November 2003 [1].

Principle of the methodology described in ITU-T P.835 is to require each listener to successively attend to and rate the waveform on: the *speech signal*, the *background noise*, and the *overall effect: speech + background noise*. A speech sample for a P.835 trial is composed of 3 sub-samples broadcasted successfully to the listeners. For the signal, subjects are instructed to attend *only* to the *speech signal* and rate the speech on the five-category distortion scale shown in Figure 1. For the background, subjects are instructed to attend *only* to the *background* and rate the background on the five-category intrusiveness scale shown in Figure 2. For the third sub-sample in each trial, subjects are instructed to listen to the speech + background and rate it on the five-category overall quality scale shown in Figure 3, the Mean Opinion Score (MOS) used with the ACR [3]. To control for the effects of rating scale order, the order of the rating scales shall be balanced across the experiment, i.e., scale order should be "Signal, Background, Overall Effect" for half of the trials, and "Background, Signal, Overall Effect" for the other half.



Session 1 Block 1 Trial 1

Attending **ONLY to the SPEECH SIGNAL**, select the category which best describes the sample you just heard.

the **SPEECH SIGNAL** in this sample was

5 - NOT DISTORTED

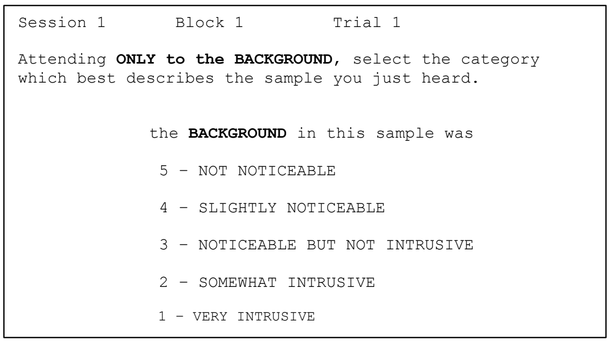
4 - SLIGHTLY DISTORTED

3 - SOMEWHAT DISTORTED

2 - FAIRLY DISTORTED

1 - VERY DISTORTED

Fig. 1 Speech signal rating scale



Session 1 Block 1 Trial 1

Attending **ONLY to the BACKGROUND**, select the category which best describes the sample you just heard.

the **BACKGROUND** in this sample was

5 - NOT NOTICEABLE

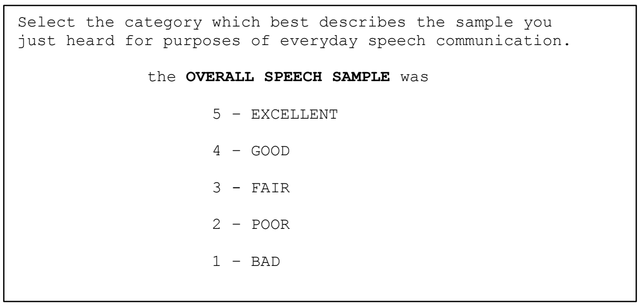
4 - SLIGHTLY NOTICEABLE

3 - NOTICEABLE BUT NOT INTRUSIVE

2 - SOMEWHAT INTRUSIVE

1 - VERY INTRUSIVE

Fig. 2 Background noise rating scale



Select the category which best describes the sample you just heard for purposes of everyday speech communication.

the **OVERALL SPEECH SAMPLE** was

5 - EXCELLENT

4 - GOOD

3 - FAIR

2 - POOR

1 - BAD

Fig. 3 Overall quality rating scale

3 POMNR: a Perceptual Objective Measure for Noise Reduction systems

Subjective testing is the most reliable methodology to evaluate the overall perceived quality of a system as perceived by the user. However, since it is costly and time consuming, objective methodologies are always preferred even if less reliable. We propose here a new Perceptual Objective Measure for the evaluation of Noise Reduction systems (POMNR): POMNR qualifies the overall perceived quality of a noise reduction system and, in particular, gives an estimate of the background noise score as specifies in ITU-T Recommendation P.835.

3.1 Test principle

As opposed to subjective testing, POMNR is an objective method which principle is described on Figure 4. The objective evaluation procedure of POMNR is based on the use of three signals:

- *x,* the clean speech signal i.e. a signal without background noise;
- *xb*, the noisy speech signal which is obtained by adding the desired amount and type of background noise to the clean signal *x*;
- *y*, the processed (or noise reduced) speech signal (i.e. *y* is the result of the processing of the signal *xb* by the noise reduction system).

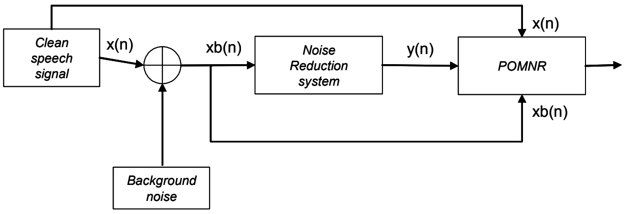


Fig. 4 Principle of POMNR

3.2 Parameters measured by POMNR

The objective of POMNR is to qualify as precisely as possible the performance of a noise reduction system. Overall perceived quality can be impacted by different factors, the most important ones being identified and measured by POMNR:

- the processing delay introduced by the noise reduction feature,
- the Adaptation Time (AT) of the noise reduction system,
- the efficiency of the noise reduction system in terms of noise reduction called IE_NR (Index of Efficiency of the Noise Reduction system),
- the annoyance due to the presence of noise,
- the presence of distortion on the speech signal.

Computation of the processing delay introduced by the noise reduction system is important since a very long delay could lead to audible echo phenomena for example. The adaptation time can be compared to the convergence time of an echo cancellation algorithm: the adaptation of a noise reduction system corresponds to the time needed by the system to provide the maximum noise reduction (or the time needed by the system to be in a stationary mode).

The index of efficiency of the noise reduction system (IE_NR) computed by POMNR estimates the efficiency of the system in terms of noise reduction by taking into account speech signal attenuation: a system which leads to a high noise reduction and a high speech signal level attenuation will have an index of efficiency smaller than that of a system providing a smaller noise reduction but with no speech level attenuation. Experiments performed at France Telecom R&D labs, have shown that the IE_NR, which can vary from 0 (not efficient) to more than 12 (very efficient), is a better subjective performance indicator than the classical computation of the Signal to Noise Ratio Improvement.

The two last parameters are estimates of the background noise score and the speech signal score as specified in ITU-T Recommendation P.835 (see figures 2 and 1 respectively). Computation of such estimates is based on the use of human hearing characteristics. Details are provided in next section for the background noise, the speech signal and the overall quality being still under study.

3.3 Focus on the annoyance due to noise

POMNR allows to determine objectively, by taking into account perceptual criteria, an objective score (NOS_{MOS}) equivalent to the background noise MOS score as defined in ITU-T P.835. Computation procedure is described on figure 5. Processing is based on a frame-per-frame analysis (256 samples at a sampling frequency of 8 kHz, 50% overlap). A Vocal Activity Detection (VAD) allows to determine if the current frame corresponds to the presence of noise only or to the presence of the speech (speech or speech + noise): this step is important since a specific processing is applied depending on the classification of the frame (noise or speech). Human hearing criteria are used when evaluating loudness densities [4] and coefficient of tonality [5]. The two last steps correspond to mapping of objective results to subjective results and to mapping to the MOS scale (values between 1 and 5) respectively. The first mapping (computation of values of ω_i) relied on the use of subjective data compliant to ITU-T Recommendation P.835. The Noise Objective Score (NOS) is the result of a linear combination of the 5 factors F(i) and $\omega_{i.}$ The estimate of the ITU-T P.835 subjective background noise score – NOS_{MOS} – is then computed with a 3-order polynomial function. As a result, a very reliable objective estimate of the background noise subjective score is obtained as illustrated in next section.

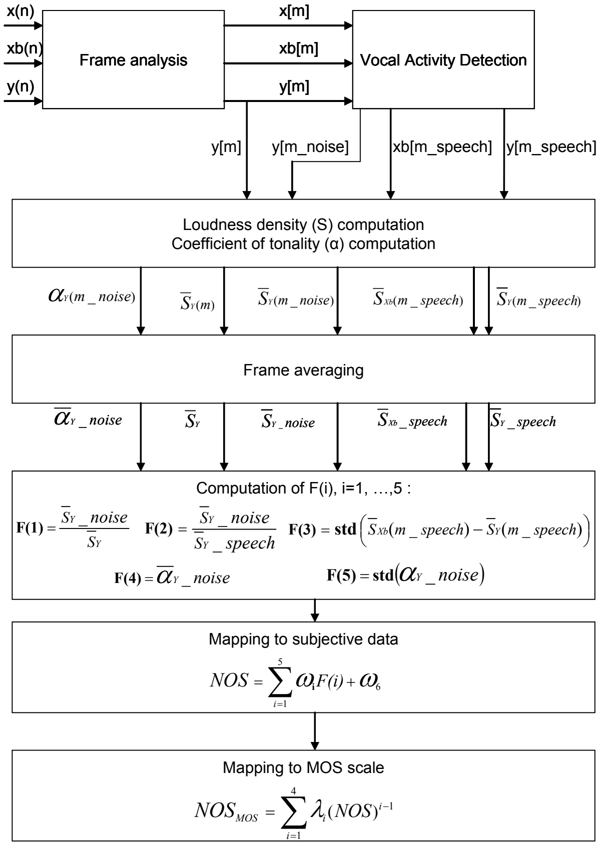

x(n)
xb(n)
y(n)
Frame analysis
x[m]
xb[m]
y[m]
Vocal Activity Detection
y[m]
y[m_noise]
xb[m_speech]
y[m_speech]
Loudness density (S) computation
Coefficient of tonality (α) computation
$\alpha_Y(m\_noise)$
$\bar{S}_Y(m)$
$\bar{S}_Y(m\_noise)$
$\bar{S}_{Xb}(m\_speech)$
$\bar{S}_Y(m\_speech)$
Frame averaging
$\bar{\alpha}_Y\_noise$
$\bar{S}_Y$
$\bar{S}_Y\_noise$
$\bar{S}_{Xb}\_speech$
$\bar{S}_Y\_speech$
Computation of F(i), i=1, …,5 :
$\mathbf{F(1)} = \frac{\bar{S}_Y\_noise}{\bar{S}_Y}$
$\mathbf{F(2)} = \frac{\bar{S}_Y\_noise}{\bar{S}_Y\_speech}$
$\mathbf{F(3)} = \mathbf{std}\left(\bar{S}_{Xb}(m\_speech) - \bar{S}_Y(m\_speech)\right)$
$\mathbf{F(4)} = \bar{\alpha}_Y\_noise$
$\mathbf{F(5)} = \mathbf{std}(\alpha_Y\_noise)$
Mapping to subjective data
$NOS = \sum_{i=1}^{5} \omega_i F(i) + \omega_6$
Mapping to MOS scale
$NOS_{MOS} = \sum_{i=1}^{4} \lambda_i (NOS)^{i-1}$

Fig. 5 Computation procedure of the Noise Objective Score mapped to a MOS-scale (NOS_{MOS})

3.4 Experimental results

Five sessions (4 with French samples, 1 with English samples) of tests according to ITU-T Recommendation

P.835 were carried out by our laboratory between autumn 2003 and winter 2004: 6 different noise reduction systems were tested and up to 10336 subjective data were collected. Training of the objective model was performed on 75% of the resulting subjective database. Therefore validation of the objective modelling NOS_{MOS} was achieved on the other part of the subjective data. As illustrated by Figure 6, a very good correlation (0,92) is obtained between the ITU-T P.835 subjective background noise score and the objective estimate we propose, NOS_{MOS}. Moreover, data dispersion is quite small: a deeper analysis shows that the absolute estimation error (absolute value of the difference between the ITU-T P.835 subjective background noise score and the corresponding objective estimate NOS_{MOS}) is less than 0.375 for 81% of the data and less than 0.5 for 97% of the data (see Figure 7).

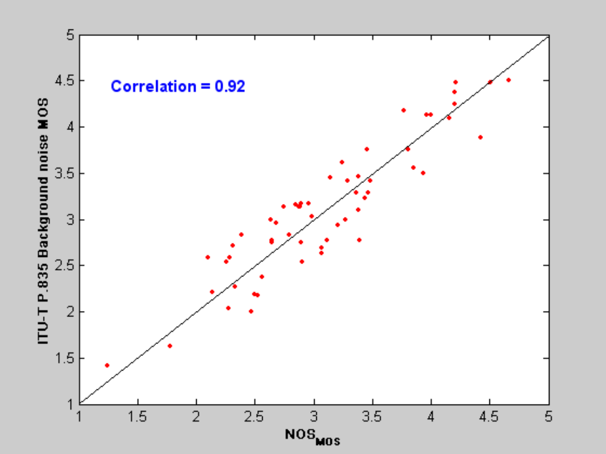


Fig. 6 Relation between the subjective noise score and our proposed objective score ((NOS_{MOS}*)*

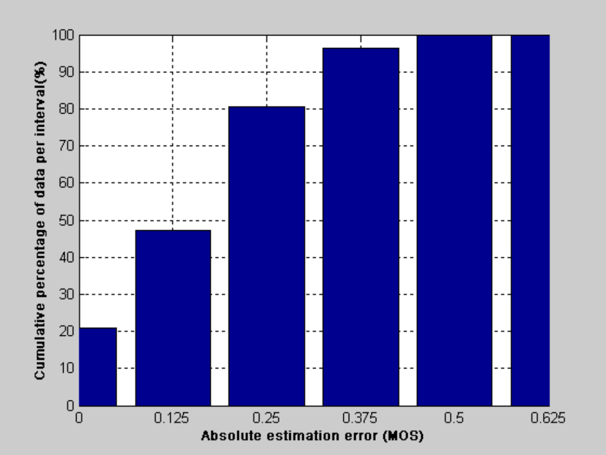


Fig. 7 Absolute estimation error cumulative distribution

4 Conclusion

Today, subjective evaluation of noise reduction systems is the only way to determine as precisely as possible the overall perceived quality. However, current work at the ITU-T shows that the objective evaluation of noise reduction systems is probably as important as the subjective approach. The Perceptual Objective Measure for Noise Reduction systems (POMNR) we propose is both an innovative methodology and a very promising and useful tool. The noise objective score NOS_{MOS} we have presented is a very reliable estimate of the subjective background noise score specified in ITU-T Recommendation P.835. The next step in designing POMNR consists in improving the reliability of our current estimate of the ITU-T P.835 speech signal score. A further, and final, step will then be the development of an estimate for the overall quality rating of P.835.

5 References

[1] ITU-T Recommendation P.835 (11/03), "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm".

[2] Draft ITU-T Recommendation G.160, "Voice Enhancement Devices".

[3] ITU-T Recommendation P.800 (08/96), "Methods for subjective determination of transmission quality".

[4] ITU-T Recommendation P.862 (02/01), "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs".

[5] Turbin V., Gilloire A., Scalart P., Beaugeant C., "Using psychoacoustic criteria in acoustic echo cancellation algorithms", Proceedings of the International Workshop on Acoustic Echo and Noise Control, pp. 53-56, London, UK, September 1997.