Estimation of the speech quality of noise reduced signals

Valérie Gautier-Turbin, Nicolas Le Faucheur

France Télécom R&D, France

valerie.gautierturbin@orange-ftgroup.com, nicolas.lefaucheur@orange-ftgroup.com

Abstract

The evaluation of noise reduction systems is mandatory to ensure that speech quality is improved or at least preserved. Today, there is only one and only one methodology to evaluate noise suppression algorithms, which is standardized by the ITU (International Telecommunications Union): this is a subjective procedure described in ITU-T Recommendation P.835. According to this procedure, we designed an instrumental tool called POMNR (Perceptual Objective Measure for Noise Reduction systems). We focus here on the computation of the "speech score" provided by POMNR. Analysis of experimental results shows that it is a very promising approach leading to high correlation with the standardized subjective procedure.

Keywords

Noise reduction, objective evaluation, psychoacoustic.

1 Introduction

When we want to qualify performance of noise reduction features, we ideally would like to estimate the overall perceived quality as provided by users. When dealing with noise reduction systems, this is not an easy task.

A possible approach is the use of subjective tests. In subjective tests, human subjects test systems under different network conditions and judge the speech quality by voting on an opinion scale. For each condition, the scores given by the subjects are averaged to get a mean opinion score (MOS) [1]. However, it was shown that subjective tests carried out according to ITU-T Recommendation P.800 were not suitable for the evaluation of noise reduction systems [2]. The major problem comes from the use of a single-scale rating, the MOS. Indeed analysis of perceived quality of noisy files yields a bi-dimensional problem: some users focus on speech distortion whereas others focus only on background noise level. 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 [3]. 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*. 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.

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 first focus on classical objective measures used to qualify performance of noise reduction algorithms. We then present POMNR, a new objective measure we developed based on the approach of the only standardized subjective procedure. More precisely we explain the methodology used to evaluate the quality of the speech signal. Experimental results are presented in section 4 to illustrate the efficiency of POMNR.

2 Classical objective methods

As opposed to subjective testing, objective methods allow quick and repeatable tests (the exact same result is obtained if identical conditions are used). Moreover, a wide range of conditions can be tested rapidly. In case of noise reduction assessment, artificially noisy signals are often used. This is interesting since a wide range of noise type and level can be generated and tested rapidly. On the other hand, artificially noisy signals do not take into account human behaviors such as the Lombard effect [4].

The most commonly used measures to objectively evaluate performance of noise reduction algorithms rely on the computation of Signal to Noise Ratios (SNR). As an illustration, the SNRI (Signal to Noise Ratio Improvement) computes the SNR before and after noise processing [5]. The comparison in terms of SNR allows qualifying the amount of noise reduced by the noise reduction algorithm. The major drawback of SNR-based measures is that speech signal distortion is not taken into account as well as noise distortion. As a result, SNR-based measures are not reliable for speech quality assessment, especially for low SNR (i.e. SNR<20 dB).

Another approach to objectively evaluate performance of noise reduction systems consists in computing a distance between the "reference" signal and the "processed" signal. The first major difficulty here is the definition of the reference signal: should we consider as the reference signal the clean speech signal (i.e. the signal with no noise) or the noisy speech signal? It is commonly agreed to choose the clean speech signal as the reference although the goal of a noise reduction algorithm is to reduce noise level and not to completely remove the noise. An illustration is the computation of the cepstral distance [6]. It enables to detect distortion of the speech signal. However, it does not make any difference between audible and inaudible distortions. Moreover, such a method does not provide any information on the amount of reduced noise.

In order to provide information on speech distortion as close as possible to human perception, the application of PESQ (*Perceptual Evaluation of Speech Quality*) model standardized as ITU-T Recommendation P.862 [7] was proposed. Principle consists in computing two PESQ scores obtained in a given configuration without noise reduction and with noise reduction respectively [8]. This method allows evaluating the quality of speech and noise but does not give any indication on the amount of reduced noise. Moreover, PESQ is intended to measure one-way quality on narrow-band telephone signals and models the perceived quality in the listening context (mainly impacted by speech distortion due to speech codecs, background noise and packet loss). PESQ can predict the Listening Effort of noisy and/or noise reduced signals [9] but has never been validated for the prediction of P.835 quality scores.

Instead, a hybrid measure which combines computation of SNR and PESQ scores can be used. An illustration is the Quality Index (QI) which was proposed to predict overall perceived quality as well as to quantify the amount of reduced noise [10]. This hybrid measure seems to be interesting but it still remains to define how these parameters relate with human perception.

At the moment, several models of speech quality exist [11] [12] but none which could exhibit a high correlation with results obtained according to the subjective procedure described in [3].

3 The objective measure POMNR

We propose a Perceptual Objective Measure for the evaluation of Noise Reduction systems (POMNR). The development of POMNR was motivated by having an instrumental tool capable of predicting the quality of a noise reduction algorithm in accordance with ITU-T Recommendation P.835.

3.1 Principle of POMNR

POMNR is an objective method which principle is described on Figure 1. The objective evaluation procedure of POMNR is based on the use of three signals:

- *x*, the clean speech signal (no noise);
- *xb*, the artificially 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).

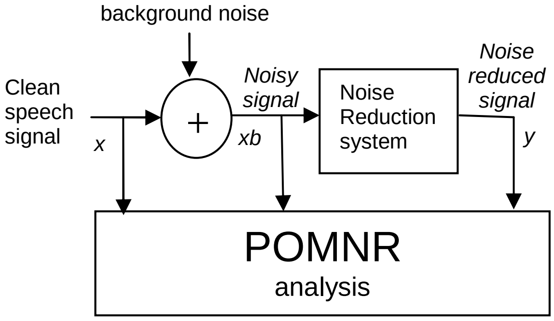

Fig. 1 Principle of POMNR

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 of the noise reduction system,
- the efficiency of the noise reduction system in terms of noise reduction called Index of Efficiency of the Noise Reduction system,
- the annoyance due to the presence of noise,
- the presence of distortion on the speech signal.

More details about the first four parameters can be found in [13]. Here, we focus on the last parameter.

3.2 Perceived speech quality

Analysis of subjective results which were collected according to ITU-T Recommendation P.835 shows that the judgement of subjects about the speech quality is influenced by several factors [2]:

- the presence of degradations on the speech signal itself;
- the presence of degradations on the speech sample (i.e. speech + silence/noise);
- the amount and the quality of the reduced noise when speech is not active also influence the final score.

The objective estimation of the speech quality of a noise reduced signal may be dependent on the parameters described above. Also, three signals are clearly involved for the speech quality estimation: the clean signal, the noisy signal, and the reduced noise signal. The comparison of those three signals exercises a major influence in the objective evaluation of the perceived speech quality, as it was proposed in [8] for example. All these remarks were taken into account when implementing the estimate of the speech signal score in POMNR.

3.3 Estimation of the speech signal score

POMNR computes a Speech Objective Score mapped to a MOS-scale (SOS_{MOS}) which is an estimate of the speech signal MOS defined in [3]. Computation procedure is described on Figure 2. Processing is first based on a frame-per-frame analysis (256 samples at a sampling frequency of 8 kHz, 50% overlap). A Voice Activity Detection (VAD) is applied to each frame in order to classify the analyzed frame: VAD(m) is equal to 1 if the frame m corresponds to the presence of speech (speech or speech+noise) and VAD(m) is equal to 0 if no speech is detected (frame corresponding to silence or presence of noise only).

Human hearing criteria are then used when evaluating loudness densities of the current frame m for signals x, xb and y –$S_X(m)$, $S_{Xb}(m)$, $S_Y(m)$ respectively– according to the procedure described in [7]. Step 2 consists in computing the masking threshold of the clean signal x (S_{msk}) as explained in [14]. This step is very important since the masking threshold S_{msk} will determine if identified degradations (in step 6) are audible or not. In step 3, for each critical band b (b=1,…,18 since the sampling frequency is 8 kHz), we calculate average distances between loudness densities of signals x and y ($d_{YX}(m,b)$) and between loudness densities of signals xb and y ($d_{XbY}(m,b)$), as follows:

$$\begin{aligned} d_{YX}(m,b) &= S_Y(m,b) - S_X(m,b) \\ d_{XbY}(m,b) &= S_{Xb}(m,b) - S_Y(m,b) \end{aligned} \tag{1}$$

where S_Y, S_X, and S_{Xb} stand for the loudness densities of signals y, x and xb respectively, m is the current frame and b the critical band.

Then, next step (step 4) consists in creating 3 groups of distances (G1, G2, and G3) based on the following rules:

- ▪ $d(m,b) \in$ G1 if $d(m,b)>0$ and $d(m,b)>S_{msk}(m,b)$
- ▪ $d(m,b) \in$ G2 if $d(m,b)<S_{msk}(m,b)$ & $d(m,b)>-S_{msk}(m,b)$
- ▪ $d(m,b) \in$ G3 if $d(m,b)<0$ and $d(m,b)<-S_{msk}(m,b)$

At step 5, size of analysis frame is modified: we now work with frames p, with size of frame p = q x size of frame m, q integer. Analyzing longer frames allows to better determine degradations of the speech signal (q=20 for example). This is also used in PESQ for example [7]. In step 6, degradations deg(i) (i=1,…, 4) of the processed signal are computed according to equation (2), with P=24 for example:

$$\mathrm{deg}(i) = \frac{\sum_{p=1}^{P}\left(\sum_{b=1}^{18}\left(\sum_{\substack{m\in p\,\& \\ (m,b)\in U_i\,\& \\ VAD(m)=k}} S_X(m,b) * \left|d_{YX}(m,b)\right|\right)\right)}{\sum_{p=1}^{P}\left(\sum_{b=1}^{18}\left(\sum_{m\in p} S_X(m,b)\right)\right)} \tag{2}$$

where m is the frame index, b the critical band, k is an integer (equal to 0 or 1), and U_i stands for the subset of frames to be considered for the degradation deg(i).

For i=1, U_1=G1 and k=0: deg(1) corresponds to the residual noise for silent frames.
For i=2, U_2=G3 and k=1: deg(2) corresponds to subtractive degradations due to noise during speech activity.

For i=3, U_3=G1 and k=1: deg(3) corresponds to additive degradations during speech activity.
For i=4, U_4=G1∪G2∪G3 and k=1: deg(4) corresponds to overall degradation due to noise during speech activity.

In step 7, we classify the processed signal according to the weight of each group (G1, G2, G3), i.e. we determine the frames distribution into the three different groups. This defines the "group size" (s) of the processed signal (6 different group sizes are used in our work). The last two steps correspond to mapping of objective results to subjective results and to mapping to the MOS scale. The first mapping (computation of values of ω(i,s)) relied on the use of subjective data compliant to ITU-T Recommendation P.835 and was performed as defined in equation (3):

$$\begin{aligned} SOS = \sum_{i=1}^{4} \omega(i,s)\deg(i) \\ + \omega(5,s) std\left(d_{YX}(m,b)\right) \\ + \omega(6,s) std\left(d_{XbY}(m,b)\right) \\ + \omega(7,s)\left(d_{cep}(x,y)\right)_{VAD=1} \\ + \omega(8,s) \end{aligned} \tag{3}$$

For each size s, a set of values ω(i,s) had to evaluated, which means 48 values for our chosen parameters. The Speech Objective Score (SOS) is the result of a linear combination of 7 parameters, "std" standing for the standard deviation and $d_{cep}(x,y)$ for the cepstral distance of signals x and y [6].

The estimate of the ITU-T P.835 subjective speech score SOS_{MOS} is then computed with a 3-order polynomial function.

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. Up to 10336 subjective data were collected. English native speakers were used as testers for tests with English samples.

4.1 About the noise reduction features

Six noise reduction systems (NR1, NR2,… NR6) were used during our subjective tests. NR1 and NR2 are two noise reduction algorithms developed by our laboratory, NR1 being "smooth" and NR2 aggressive. More speech signal distortion is to be expected from NR2 compared to NR1. On the other hand, less reduced noise is provided by NR1 compared to NR2. Noise reduction systems NR3 to NR6 are features integrated to two different network echo cancellers (1 and 2). The two echo cancellers allow to parameter the aggressiveness of the noise reduction: NR3 and NR4 correspond to a smooth and to an aggressive noise reduction of echo canceller 1 respectively and, NR5 and NR6 to a smooth and to an aggressive noise reduction of echo canceller 2 respectively.

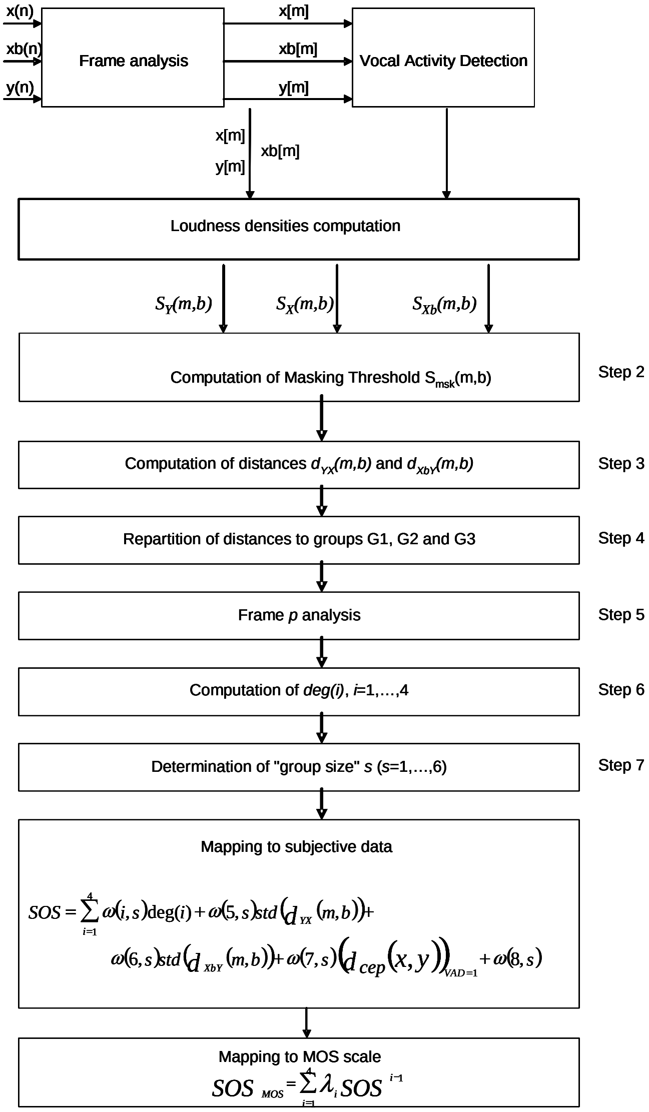
x(n)
xb(n)
y(n)
Frame analysis
x[m]
xb[m]
y[m]
Vocal Activity Detection
x[m]
y[m]
xb[m]
Loudness densities computation
$S_Y(m,b)$
$S_X(m,b)$
$S_{Xb}(m,b)$
Computation of Masking Threshold $S_{msk}(m,b)$
Step 2
Computation of distances $d_{YX}(m,b)$ and $d_{XbY}(m,b)$
Step 3
Repartition of distances to groups G1, G2 and G3
Step 4
Frame p analysis
Step 5
Computation of $deg(i)$, i=1,…,4
Step 6
Determination of "group size" s (s=1,…,6)
Step 7
Mapping to subjective data
$SOS = \sum_{i=1}^{4} \omega(i,s)\deg(i) + \omega(5,s) std\left(d_{YX}(m,b)\right) + \omega(6,s) std\left(d_{XbY}(m,b)\right) + \omega(7,s)\left(d_{cep}(x,y)\right)_{VAD=1} + \omega(8,s)$
Mapping to MOS scale
$SOS_{MOS} = \sum_{i=1}^{4} \lambda_i SOS^{i-1}$

Fig. 2 Computation procedure of SOS_{MOS}

4.2 About the background noise

Three types of noise (street, office, crowd), and four different speakers (2 males, 2 females) were used. Two different SNR were used for each noise type and each speaker: 10 and 15 dB for street noise, 15 and 20 dB for office and crowd noise. For each noise, the lowest SNR value was chosen so as to be close to real life conditions and the highest to check the validity of POMNR even for more faintly noisy signals. Lowest and highest SNR values will be considered in the near future.

4.3 Analysis of SOS_{MOS}

The design of an objective model requires the training of the objective model on a known subjective database and the validation on an unknown subjective database. In our case, the training of the model is necessary to compute the values $\omega(i,s)$ (i.e. 48 values) and λ_j (4 values). Training could then ideally be done on 75% of our subjective database and validation on the rest of the database. However, it is not possible to do it this way if we use step 7 (see Figure 2) for the computation of SOS_{MOS}: our subjective database is not big enough to have enough data for the different groups of distances (G1, G2, and G3). So, in this section, two kinds of results are presented:

- results without step 7: in this case, our database is suitable for training on 75% of the database and validation on the rest of the database (25%). The results (correlation and absolute estimation error) come from the unknown data – 25% – for the model POMNR.
- results with step 7: the database is not large enough to randomly choose 75% of the data for the training and the rest for the validation. So, in this case, results correspond to the performance of the model trained on the entire database.

When the computation of SOS_{MOS} is performed without step 7, only 8 values $\omega(i)$ and still 4 values λ_j need to be computed. Then, the training of the model on 75% of the subjective database and the validation on the rest of the subjective database lead to a correlation of 0.88 between the ITU-T P.835 subjective speech score and the corresponding objective estimate SOS_{MOS}, with an absolute estimation error (absolute value of the difference between the two scores or RMSE) of 0.35.

Now, if we use step 7, the correlation raises to 0.93 (see Figure 3) and the data dispersion (RMSE) decreases to 0.28, which demonstrates the interest for step 7. A deeper analysis shows that the RMSE is less than 0.375 for 84% of the data and less than 0.5 for 92% of the data (see Figure 4). Comparison of subjective scores and corresponding estimated objective scores obtained when averaging over conditions (i.e. an average score per noise reduction feature) clearly shows that SOS_{MOS} is a very good estimate of the P.835 subjective speech score (see Table 1). Similar conclusions are obtained with English.

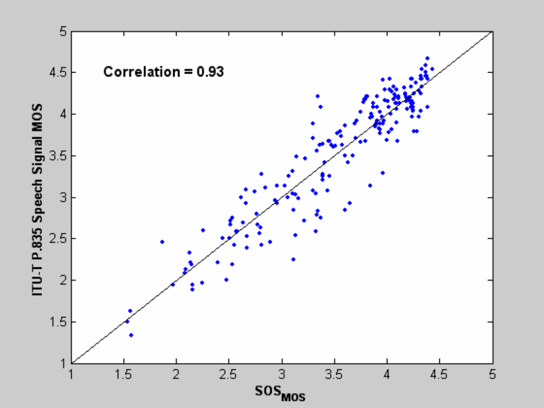

Fig. 3 Relation between subjective speech scores and SOS_{MOS}

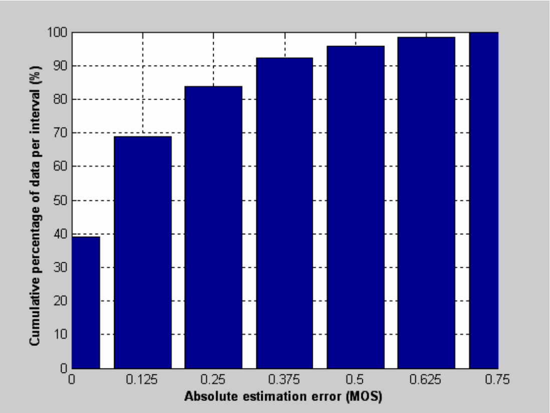

Fig. 4 Absolute estimation error cumulative distribution

Noise Reduction	*P.835 MOS*	*SOS_{MOS}*
NR1	4.04	4.04
NR2	2.91	2.95
NR3	4.03	4.03
NR4	4.00	3.92
NR5	3.69	3.71
NR6	2.70	2.78

Table 1. Comparison of P.835 MOS and SOS_{MOS}

5 Conclusion

Subjective evaluation of noise reduction systems as specified in ITU-T Recommendation P.835 is the only standardized way to determine as precisely as possible the overall perceived quality of noise reduction features. Current work at the ITU-T shows that the objective evaluation of noise reduction systems is also of great interest: Question 9 of Study Group 12 has just begun a new work item on the drafting of a new recommendation P.ONRA (Objective Noise Reduction Assessment). Based on the very promising results we obtained, the objective measure POMNR we propose could suit the requirements of Recommendation P.ONRA. The next step in designing POMNR consists in enlarging the subjective database and in checking (and improving if necessary) the reliability of our model on new (unknown) subjective data. A further step will be the study of the third score (overall effect) specified in ITU-T Recommendation P.835 and more precisely the determination of how this score correlates with a PESQ score. Note that POMNR was validated and works with narrow-band signals and restricted to G.711 coder. Work on a wideband extension is actually going on. Other coders also should be considered soon.

6 References

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

[2] A. Battistello, "Study on the methodology for evaluating the subjective quality of noise suppression algorithm", ITU-T SG 12 COM12-D.116, January 2003.

[3] "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm", ITU-T Recommendation P.835, 2003.

[4] E. Lombard, "Le signe de l'élévation de la voix", Annales des maladies de l'oreille et du laryn*x*, vol. 37(2), pp. 101-119, 1911.

[5] V. Mattila, "Objective measures for the characterization of the basic functioning of noise suppression algorithms", *Measurement of Speech and Audio Transmission Quality in Telecommunications Networks (MESAQIN)*, 2003.

[6] R. Boite and M. Kunt, "Traitement de la parole", *Complément au Traité d'Electricité,* Presses Polytechniques Romandes, 1987.

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

[8] A. Rix, "Application of P.862 to evaluation of quality of noise reduction and noise processing algorithms", ITU-T SG 12 COM12-D.45, October 2001.

[9] P. Juric, "SQuad, an objective method for evaluation of noise reduction systems", ITU-T SG 12 COM12-D.71, May 2002.

[10] J. G. Beerends, J. M. van Vugt, J. Jensen, "Predicting Listening Effort with noise suppressors on the basis of PESQ", ITU-T SG 12 COM12-C.25, March 2006.

[11] A. W. Rix, M. P. Hollier, A. P. Hekstra, J. G. Beerends, "PESQ, the new ITU standard for objective measurement of perceived speech quality, Part I – Time alignment", Journal of the Audio Engineering Society, vol. 50, pp. 755-764, October 2002.

[12] J. G. Beerends, A. P. Hekstra, A. W. Rix, M. P. Hollier, "PESQ, the new ITU standard for objective measurement of perceived speech quality, Part II – Perceptual model", Journal of the Audio Engineering Society, vol. 50, pp. 765-778, October 2002.

[13] V. Gautier-Turbin, N. Le Faucheur, "A Perceptual Objective Measure for Noise Reduction systems", *Measurement of Speech and Audio Transmission Quality in Telecommunications Networks (MESAQIN)*, June 2005.

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