PERFORMANCE ASSESSMENT METHOD FOR SPEECH ENHANCEMENT SYSTEMS

[1]*Bernard Grundlehner,* [1]*Johan Lecocq,* [2]*Radu Balan,* [2]*Justinian Rosca*

[1]Bernard.Grundlehner@siemens.com
[1]Siemens VDO Infotainment Solutions, P.O. Box 8807, 5605 LV, Eindhoven, The Netherlands
[2]Siemens Corporate Research, Princeton, NJ 08540, USA

ABSTRACT

A new method to assess noise reduction algorithms with respect to their ability to enhance the perceived quality of speech is presented. Such algorithms consist of both single-microphone systems and multiple microphone systems. Tests of the presented method show a higher correlation with subjective assessments than any other objective system known by the authors. It is believed that this method is suitable to improve the comparability between noise reduction algorithms. Another area of application could be the optimization of parameters in a noise reduction algorithm, as well as the optimization of the geometric microphone positioning.

1. INTRODUCTION

In a time where more than half of the conversations with cellular phones are initiated in the car and where it is prohibited to have a hand held conversation while driving, hands-free systems have become popular. Speech quality is then severely degraded due to noise and echo. To compensate for this, echo cancellation and noise reduction algorithms are used. Several suppliers of such algorithms exist nowadays, all claiming to have the best technology. Very often, they do not present comparable, quantitative data, which makes it very difficult to rank the offered algorithms. Normally, an extensive listening test has to be organized, with many people involved, to determine the relative Mean Opinion Scores. An alternative is the objective PESQ test (ITU-T P.862). However, PESQ focuses merely on speech codecs, not on noise reduction algorithms, as mentioned in [1]. Many other objective metrics, like SNR or distortion metrics, do not have much correlation with perceived quality of speech, since each of them reveals only part of the picture.

This paper will focus on the assessment of Noise Reduction algorithms. A new method will be introduced, that enables the prediction of the quality of noise reduction algorithms. The paper is organized as follows. First, some existing objective measures will be introduced. In the following section, it is shown how these measures are combined to form one overall measure. After that, the results of a comparative study is presented, where the new combined measure is compared with a subjective MOS test and with the PESQ MOS. In the last section, the conclusions are drawn.

2. EXISTING OBJECTIVE MEASURES

This section will introduce objective measures, known from literature to correlate to some extend to human perception of speech quality. Let $s(n)$ denote the noise-free speech signal at time n and $\hat{s}(n)$ the corresponding output of the noise reduction module, applied to the speech signal recorded under noisy conditions. $s(n)$ and $\hat{s}(n)$ are scaled in amplitude and aligned in time.

2.1. General SNR

The general SNR is defined (see [2]) as:

$$SNR = 10\log_{10}\frac{\sum_n s^2(n)}{\sum_n \left(s(n)-\hat{s}(n)\right)^2} \qquad (1)$$

2.2. Segmental SNR

The segmental SNR (SNR_{Seg}) is defined as averages of measurements of SNR over short, "good frames". The "good frames" are frames where the SNR is above a lower threshold (for example -10 dB), and saturated at an upper threshold (in our application +30 dB). "Short frames" have a typical length of 15-25 ms. The SNR_{Seg} is defined as:

$$SNR_{Seg} = \frac{1}{N}\sum_{k=1}^{N} 10\log_{10}\left[\frac{\sum_{n\in \mathrm{frame}_k} |s(n)|^2}{\sum_{n\in \mathrm{frame}_k} |\hat{s}(n)-s(n)|^2}\right] \qquad (2)$$

with: N the number of good frames and $n \in \mathrm{frame}_k$ the time instances n that are in the time window of the k^{th} frame.

2.3. Frequency Weighted Segmental SNR

The Frequency Weighted Segmental SNR (SNR_{FWS}) is similar to the SNR_{Seg}, with an additional averaging over frequency bands. The frequency bands are proportional to the ear's critical bands. Noise-dependent weights $w_{j,k}$ can be applied in each frequency band. The SNR_{FWS} can thus be defined as:

$$SNR_{FWS}=\frac{1}{N}\sum_{k=1}^{N}\frac{1}{W_k}\sum_{j=1}^{F}10\log_{10}\left[\frac{w_{j,k}\cdot\sum|s(n)|^2}{\sum|\hat{s}(n)-s(n)|^2}\right] \quad (3)$$

with $W_k = \sum_{j=1}^{F}(w_{j,k})$ and F the number of frequency bands.

2.4. Itakura-Saito distance

The Itakura-Saito distance (d_{IS}) is defined as the time-averaged Itakura-Saito distance, taking only "good frames" into account. Itakura-Saito distance is a distance measure over the Linear Prediction Coefficients (LPC) of the two corresponding signal frames.
Let $\alpha(k)$ and $\beta(k)$ denote the LPC of for frame k of the signals s and $\hat{s}$, respectively:

$$\alpha^T(k) = [1\ -a_{s,1}\ \cdots\ -a_{s,M}] \quad (4)$$
$$\beta^T(k) = [1\ -a_{\hat{s},1}\ \cdots\ -a_{\hat{s},M}] \quad (5)$$

where M denotes the prediction order.
Let $\gamma(k)$ be the autocorrelation function of frame k of s, and $R_s(k)$ the Toeplitz matrix of $\gamma(k)$, then the Itakura-Saito distance is defined by:

$$d_{IS} = \frac{1}{N}\sum_{k=1}^{N}\log\frac{\beta^T(k)R_s(k)\beta(k)}{\alpha^T(k)R_s(k)\alpha(k)} \quad (6)$$

where the averaging is done over synchronized, "good frames" of the data. The "good frames" are defined here as:

$$R_s(0,0) > 1\cdot10^{-4} \quad (7)$$

It must be noted that a lower d_{IS} indicates a better quality of speech.

2.5. Weighted Spectral Slope measure

The Weighted Spectral Slope Measure ($WSSM$) is defined as time averaged Weighted Spectral Slope Measure, where only the "good frames" are averaged. The Weighted Spectral Slope Measure measures the weighted differences of spectral slope over 25 critical frequency bands between the two corresponding signal frames.
First, the energy in each of the 25 frequency bands is computed, for both $s(n)$ and $\hat{s}(n)$, resulting in $E_s(f)$ and $E_{\hat{s}}(f)$, respectively. The spectral slope at each frequency band is defined by:

$$\Delta E_s(f) = E_s(f+1) - E_s(f) \quad (8)$$
$$\Delta E_{\hat{s}}(f) = E_{\hat{s}}(f+1) - E_{\hat{s}}(f) \quad (9)$$

After that, the nearest peak $P(f)$ is located by searching upwards if $\Delta E(f) > 0$ and downwards otherwise. Then, the weight in each band is calculated as:

$$W(f) = \frac{W_s(f) + W_{\hat{s}}(f)}{2} \quad (10)$$

where

$$W_s(f) = \frac{20}{20+E_{s,max}-E_s(f)}\frac{1}{1+P_s(f)-E_s(f)} \quad (11)$$
$$W_{\hat{s}}(f) = \frac{20}{20+E_{\hat{s},max}-E_{\hat{s}}(f)}\frac{1}{1+P_{\hat{s}}(f)-E_{\hat{s}}(f)} \quad (12)$$

The magnitude of the weight reflects whether the band is near a spectral peak or valley, and whether the peak is the largest in the spectrum.
Finally, the WSSM is calculated as:

$$WSSM=\frac{1}{N}\sum_{k=1}^{N}\left[\frac{\sum_{f=1}^{24}W(f)[\Delta E_s(f)-\Delta E_{\hat{s}}(f)]^2}{\sum_{f=1}^{24}W(f)}\right] \quad (13)$$

where the averaging is done over synchronized, "good frames" of the data as defined in (7). Similar to d_{IS}, a lower $WSSM$ indicates a better speech quality.

3. COMBINATORY ALGORITHM

This section will explain how the objective measures from the previous section can be combined to form a new score. This resulting score will be referred to as SVMOS. It is assumed that the starting point is a collection of recordings, where noisy input speech samples are processed by the noise reduction algorithms under test and the outputs are recorded.
We denote by $c(DUT,k,t)$ the criterion k, for recording t (one of the utterances in a recording session), and for algorithm number DUT. We define:

$$c(DUT,1,t) = SNR \quad (14)$$
$$c(DUT,2,t) = SNR_{Seg} \quad (15)$$
$$c(DUT,3,t) = SNR_{FWS} \quad (16)$$
$$c(DUT,4,t) = -d_{IS} \quad (17)$$
$$c(DUT,5,t) = -WSSM \quad (18)$$

3.1. Step 1: Compute the score

For each criterion $1 \leq k \leq 5$, and each recording t, we find the minimum and maximum value over all DUTs:

$$minC(k,t) = \min_{1 \leq DUT \leq N} c(DUT,k,t) \quad (19)$$

$$maxC(k,t) = \max_{1 \leq DUT \leq N} c(DUT,k,t) \quad (20)$$

with k the number of algorithms under test. Next, we set a *score* to each device according to the following rule:

$$S(DUT,k,t) = \begin{cases} 1 & \text{if} c(DUT,k,t) > (2 \cdot maxC(k,t) + minC(k,t))/3 \\ -1 & \text{if} c(DUT,k,t) < (maxC(k,t) + 2 \cdot minC(k,t))/3 \\ 0 & \text{if otherwise} \end{cases} \quad (21)$$

In other words, a score of $+1$ is set to those devices that have the criterion in the upper third of the range, a score of 0 is set to those devices that have the criterion in the middle third of the range, and a score of -1 is set to those devices that have a score in the bottom third of the range.
A scoring method that is more in line with the subjective test method that is used (to be discussed in section 4), has also been applied. In that case, the scores can range from -3 up to $+3$, each score attributed to the corresponding one seventh of the range. This, however, did not lead to any significant improvement of the correlation between the final objective scores and the scores obtained with the subjective tests.

3.2. Step 2: Compute the Average Score

Next the score is averaged over time using two weights: the number of good frames, $T(DUT,k,t)$, and the other is a function of the recording SNR, $F(SNR(t))$:

$$AS(DUT,k) = \frac{\sum_t S(DUT,k,t)T(DUT,k,t)F(SNR(t))}{\sum_t T(DUT,k,t)F(SNR(t))} \quad (22)$$

The function $F()$ takes into account the importance of several regimes of functioning. Thus, if behavior in quieter environment is deemed more important than in louder conditions, then $F()$ should decay fast toward low SNR (see Figure 1, left plot); instead, if louder environment is more likely than quieter, then $F()$ should decay fast toward high SNR (see Figure 1, right plot). For equal SNR weight, the function $F(SNR) = 1$ can be used.
According to each criterion, we can now rank the devices.

3.3. Step 3: Compute the Merit Figure

The *Merit Figure* is next computed by merging the average scores over different criteria, into one single number.

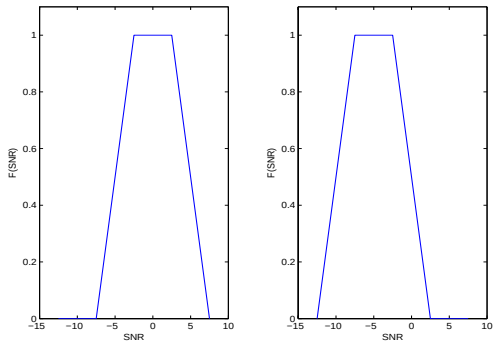

Figure 1: Example of $F()$ that weights more high SNRs (left), and more low SNRs (right).

This is done by a weighted average of $AS()$ as follows:

$$M(DUT) = (\sum_{k=1}^{5} AS(DUT,k)w_k)/(\sum_{k=1}^{5} w_k) \quad (23)$$

The weighting coefficients w_k reflect the relative contribution of each measure on the overall perceived speech quality. They are optimized with a separate set of measured data. They reflect, to some extend, the correlation between the subjective speech quality and the objective measure in question (see table 9.7 in [2]).

4. RESULTS

Several test recordings were used to optimize the parameters and verify the consistency of the system as described in the previous sections. Here we present the conclusions. The recordings were done in a driving car, under several road conditions. The output of 7 devices under test were recorded simultaneously, together with an unprocessed microphone output (used for reference). Some devices consisted of multiple microphone arrays, some had only one microphone input.
The recordings are processed as described in the previous sections, with $F(SNR) = 1$. Furthermore, the PESQ MOS is computed.
The outcomes are compared with the outcome of a subjective test, performed in conformance with the Comparison Category Rating, as defined in [3]. This test computes a *relative* Mean Opinion Score, called CMOS.
The results are plotted in figures 2, 3 and 4.
Pearson's formula is used to see how well the objective criteria compare with the subjective CMOS:

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} \quad (24)$$

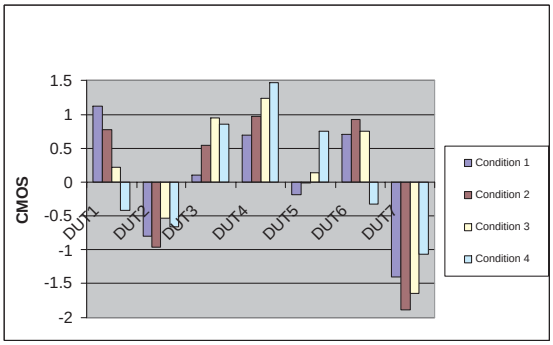

Figure 2: Results of subjective Comparison Category Rating

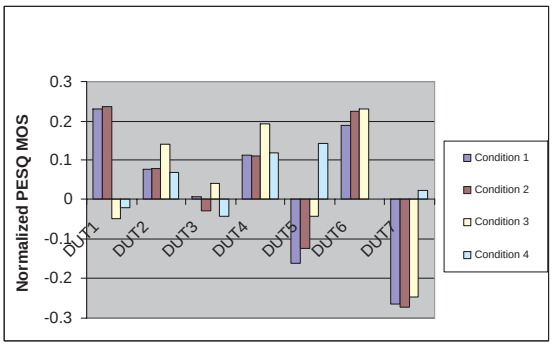

Figure 3: Results of objective PESQ test

In this formula, x_i is the CMOS for DUT i, and $\overline{x}$ is the average over the CMOS values, y_i is the predicted MOS value for DUT i and $\overline{y}$ is the average over the predicted MOS values.

In figure 2, the results of the subjective test under several conditions are shown. In the case of the first condition, a CELP codec encoded and decoded the output signal of the noise reduction module. In figure 3, the outcomes of the PESQ test are depicted, after normalization. Normalization is done for easy comparison between the several methods. It is done by subtracting the mean of the overall score from the individual scores (this is done since the PESQ MOS is positive by definition). Note that this, as well as (24) assume a linear PESQ MOS-scale, which might not be entirely correct. Figure 4, finally, shows the SVMOS.

If the outcomes of all road conditions and all the devices (except for the unprocessed microphone input) are taken into account, the correlation, computed with (24), turns out to be as follows:

- Pearson's correlation between the PESQ MOS and the CMOS is 0.69,

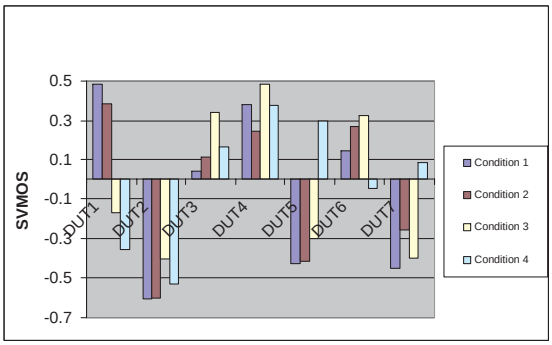

Figure 4: Results of objective SVMOS

- Pearson's correlation between the SVMOS and the CMOS is 0.81.

The SVMOS has in this case a significantly higher correlation over the PESQ MOS, whereas more extensive optimization of the weighting factors can even increase the correlation. It can therefore be concluded that the presented system is potentially a better predictor of the performance of a noise reduction system.

5. CONCLUSIONS AND RECOMMENDATIONS

A new way of assessing noise reduction modules has been presented, which combines several metrics known from literature to correlate with speech intelligibility into one relative score, called SVMOS. It has been shown to have a larger correlation with intelligibility of speech than the existing PESQ MOS. It can be useful when comparing the noise reduction modules that are available on the market, but also for optimization of the parameters in the algorithms, or optimization of placement of the microphones. Although the results are very encouraging, the authors are convinced that independent tests of the system might give rise to further optimizations.

6. REFERENCES

[1] International Telecommunication Union, "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 P.862, 2001.

[2] John R. Deller, John H.L. Hansen Jr., and John G. Proakis, *Discrete-time Processing of Speech Signals*, IEEE Press, 2000 edition, 2000.

[3] International Telecommunication Union, "Methods for subjective evaluation of transmission quality," ITU-T P.800, 1996.