List of unusual symbols: $\bar{x}$, several formulas (1) through (13)

Number of pages: 8

Number of tables: 4

Number of figures: 9, including one figure that contains 3 different images (i.e. Figure 2 contains Renata, Teeny and Car & Gate screenshots)

Footnote: 1

Keywords: image quality metrics; noise reduction evaluation; subjective evaluation.
ABSTRACT

In this paper, several state-of-the-art noise reduction algorithms are ranked using both subjective evaluation and twenty objective measures proposed in literature. The goal of the study was to find the objective quality measure that best relates to the subjective assessment for noise reduction. A measurement set-up comprising a simulation of a TV-transmission channel has been used to provide a realistic assessment of the performance in TV-applications of algorithms reducing Gaussian white noise. Ranking results are given of the subjective evaluation and all objective measures. The correlation between the subjective evaluation results and the objective measures is found to be very low. Results show that even a combination of objective measures only approximates the subjective assessment of the quality of noise reduction algorithms to a limited extent.

Key words: image quality metrics; noise reduction evaluation; subjective evaluation.

INTRODUCTION

In every part of the video chain, from the source to the display, the video may be impaired by noise. Additive Gaussian white noise is the most important analogue noise type and it primarily enters the system during the analogue transmission phase from the broadcasting of the composite video signal to the reception of the signal at the user’s premises. In literature, many studies of noise reduction on images [12,18] and, incidentally, on video [1], can be found. In these papers, many noise reduction methods are ranked using objective quality measures only and no limitation is put on algorithm complexity. In addition, no effort is taken to simulate the noise conditions that are important for consumer TV-applications. Drewery et al. [2], notably, proposed a subjective method for noise reduction performance evaluation. They used this method for the evaluation of a standard noise reduction algorithm suitable for consumer TV-applications. Subsequently, a limited comparison of the subjective results with the theoretically obtained amount of noise reduction was made. With this method it seems possible to estimate the amount of noise reduction in the case of this generic noise reduction algorithm applied to still images.

In our paper, a large number of state-of-the-art video noise reduction algorithms designed for consumer TV-applications (an application area posing severe demands on memory use and algorithm efficiency) are evaluated on their overall image quality using both a subjective ranking method and a large number of objective quality measures. The noise conditions of consumer TV-applications are carefully simulated, so that the evaluation of the performance of these algorithms can be done in a realistic way. The goal of the study was to find the objective quality measure that best relates to the subjective assessment for noise reduction.

TEST PROCEDURE

The following noise reduction algorithms were included in the test: the Spatial Weighted Aperture Noise reduction algorithm (SWAN), proposed by Ojo et al. [13], the temporal Dynamic Noise Reduction algorithm (DNR)[14], the Edge PReserving spatial noise reduction algorithm (EPRES) combined with the PERCision Adaptive Temporal noise reduction algorithm (PERCAT), proposed by Jostschulte et al. [9], and the Fuzzy Combined Spatial and
Temporal noise reduction algorithm (FCST), proposed by Mancuso et al. [11]. The noise reduction algorithms have been implemented using software simulations.

![Figure 2: Screenshots of the clean sequences used in the study](image)

- a) Renata, originated by RAI, Italy
- b) Teeny, unknown origin
- c) Car & Gate, originated by CCETT, France

As we observed a clear advantage of Motion Compensa-
tion (MC), we upgraded the referred algorithms with MC using motion vectors obtained with a three-dimensional recursive search motion-estimator [5]. DNR was included in the test with and without several added features, namely MC and High-frequency Bypass (HB) [6] (as indicated in Table 1).

The video sequences were processed using a software simulation of a PAL-transmission channel. In this channel, Gaussian white noise was added to the PAL-encoded video sequences and the result was Low-Pass Filtered (LPF) and PAL-decoded again, as shown in Figure 1. The resulting noise impaired sequences had noise levels of 26 and 32 dB PSNR.

Table 1: Selected noise reduction algorithms for two subjective evaluations

<table>
<thead>
<tr>
<th>Subjective test 1</th>
<th>Subjective test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWAN (26 dB PSNR only)</td>
<td>SWAN (32 dB PSNR only)</td>
</tr>
<tr>
<td>DNR (MC, HB)</td>
<td>DNR</td>
</tr>
<tr>
<td>SWAN + DNR (MC,HB)</td>
<td>DNR (MC)</td>
</tr>
<tr>
<td>EPRES + PERCAT (MC)</td>
<td>DNR (MC, HB)</td>
</tr>
<tr>
<td>FCST (MC)</td>
<td>SWAN + DNR (MC,HB)</td>
</tr>
</tbody>
</table>

The same procedure without adding noise was used to obtain (original) reference video sequences that were necessary to calculate some of the objective measures. The parameters of the noise reduction algorithms were optimised for each of the two noise impairment levels. Three different standard test sequences were used, as shown in Figure 2, Renata (with a detailed background and horizontal motion), Teeny (with fast motion) and Car & Gate (with zoom and some horizontal motion).

### EVALUATION METHODS FOR NOISE REDUCTION ALGORITHMS

Different methods to evaluate noise reduction algorithms can be distinguished: subjective, objective and hybrid evaluation methods. While subjective evaluation measures try to capture the human preference for a certain noise reduction algorithm through panel tests, objective and hybrid evaluation measures do this by using a formula describing luminance or signal characteristics. The difference between objective and hybrid evaluation measures is that hybrid evaluation measures try to incorporate certain characteristics of the human vision system.

In a subjective evaluation, there are three distinctive comparison approaches:

1. a positive comparison, i.e. assessment of the quality improvement of the noise-reduced video $G_F$ with respect to the noise-impaired video $G$;
2. a negative comparison, i.e. impairment assessment of the noise-reduced video $G_F$ with respect to the noise-free original video $F$ and,
3. multiple comparisons, i.e. assessment of noise-reduced video $G_F$ obtained by different noise reduction algorithms among themselves.
For these evaluation approaches, there are different scales recommended by ITU [7]. For a positive comparison, the quality scale is recommended, e.g. double stimulus continuous quality scale. The impairment scale, e.g. double stimulus impairment scale is recommended for a negative comparison. In case of multiple comparisons, the paired comparison method can be used, for which no recommended scale is given. Which approach to choose depends primarily on the aim of the perceptual experiment and the expected difference in quality between \( G_F \) and \( F \). Since in our study relatively high levels of noise have been added to the original \( F \), a large quality difference is expected between \( G_F \) and \( F \). Moreover, since state-of-the-art noise reduction algorithms have been applied to \( G \), also a large quality difference is expected between \( G_F \) and \( G \). Because the quality differences between the noise reduction algorithms could be small in some cases, a comparison of the individual quality performance with respect to the noise-free original \( F \) or to the noisy sequence \( G \) may not be accurate enough to distinguish the noise reduction algorithms from each other.

Three categories of objective and hybrid evaluation measures can be distinguished:

1. **luminance distortion** measures, i.e. measures that describe the errors in luminance value on pixel-basis of the noise-reduced video \( G_F \) with respect to the original video \( F \).
2. **luminance spatial variation** measures, i.e. measures that describe the change in luminance variation in a video picture, due to noise reduction, and
3. **processing-related** measures, that describe processing aspects of the algorithm, e.g. stability with respect to different noise sources, effectiveness of discriminating noise-distorted pixels from original image pixels and maximum noise removal capability.

In the category of objective **luminance distortion** measures, a further subdivision can be made into logarithmic and non-logarithmic measures. (Peak-) Signal-to-Noise Ratio (Improvement) ((P)SNR(I)) [18] belong to this first sub-category. Their definition is given in Equations (2) through (4) in terms of the Standard Logarithmic Luminance Distortion Measure (SLDM), that is introduced in Equation (1).

\[
SLDM(n) = 10 \log_{10} \left( \frac{\sum Q(\tilde{x}, n)}{\sum |F(\tilde{x}, n) - G_F(\tilde{x}, n)|^2} \right) \tag{1}
\]

\[
SNR(n) = SLDM, \text{ with } Q(\tilde{x}, n) = F(\tilde{x}, n)^2, \tag{2}
\]

\[
PSNR(n) = SLDM, \text{ with } Q(\tilde{x}, n) = Q = F_{\text{max}}^2, \tag{3}
\]

\[
(P)SNRI(n) = SLDM, \text{ with } Q(\tilde{x}, n) = |F(\tilde{x}, n) - G(\tilde{x}, n)|^2, \tag{4}
\]

where \( \tilde{x} \) is the pixel position \((x,y)\) of the image \( n \) and \( F_{\text{max}} \) is the maximum luminance value of \( F \), which value depends on the amount of bits used for the luminance representation.

In the sub-category of non-logarithmic **luminance distortion** measures one finds the Mean Distortion (MD) [5,18] and the Mean Square Error (MSE) [18], described using Equations (6) and (7), respectively with the help of Equation (5), introducing the Standard Luminance Distortion Measure (SLDM).

\[
SLDM(n) = \left( \sum Q(\tilde{x}, n) \right) \left( \sum |F(\tilde{x}, n) - G_F(\tilde{x}, n)|^2 \right) \tag{5}
\]

\[
MD(n) = \frac{1}{SLDM(n)}, \text{ with } Q(\tilde{x}, n) = |F(\tilde{x}, n) - G_F(\tilde{x}, n)| \tag{6}
\]

\[
MSE(n) = \frac{1}{SLDM(n)}, \text{ with } Q(\tilde{x}, n) = Q = 1. \tag{7}
\]

Detail-Preservation Capability (DPC) [12] is a measure that uses one of the mentioned logarithmic or non-logarithmic **luminance distortion** measures on the high-frequency content (detail) of the video.

In the category of **luminance spatial variation** measures, the Mean Busyness (MB) [5,18] describes the amount of spatial variation in an image. A sharp decrease of MB can indicate a loss of spatial detail due to noise filtering. MB is defined in Equation (8),

\[
MB(n) = \frac{1}{N} \sum_{\tilde{x}} \text{median}\{\text{diff}(\tilde{x}, n)\}_{j=0,1,2...17}, \tag{8}
\]

where \( N \) is the total amount of pixels of the image \( n \), \( \tilde{x} \) is the center pixel at position \((x,y)\) for which the busyness is calculated, \( \{\text{diff}(\tilde{x}, n)\} \) is the set of absolute horizontal and vertical luminance differences in a 3x3 window around \( \tilde{x} \) and \( \text{median}\{\cdot\} \) is the median operation.

In the category of **processing-related** measures three different measures are important. The Correct Processing Ratio (CPR) [5,18] is a measure for the precision of the noise reduction algorithm meaning that the right operations are performed on each pixel in a noise-impaired image, i.e. pixels of the original video \( F \) that are impaired by noise need to be changed (called \( N_c \)), while original pixels that have not been altered by noise need to be preserved (called \( O_p \)). In noise reduction there are two counterproductive processing operations, i.e. altering original non-noise impaired pixels and preserving noise-impaired pixels. Equation (9) defines CPR.

\[
\text{CPR}(n) = \frac{1}{N} \sum_{\tilde{x}} \text{median}\{\text{diff}(\tilde{x}, n)\}_{j=0,1,2...17}, \tag{8}
\]
\[ CPR(n) = \frac{1}{N} \sum \sum \left\{ y_n(x,n) + O_p(x,n) \right\} \]  

(9)

where \( N(x,n) \) and \( O_p(x,n) \) are defined as stated in Equation (10).

\[ N(x,n) = \begin{cases} 1 & G(x,n) = F(x,n) \land G_F(x,n) = G(x,n) \\ 0 & \text{otherwise.} \end{cases} \]  

\[ O_p(x,n) = \begin{cases} 1 & G(x,n) \neq F(x,n) \land G_F(x,n) \\ 0 & \text{otherwise.} \end{cases} \]  

(10)

A second processing-related measure is the Stability \( (ST) \) criterion [5,18] that describes the difference in processing results of a sequence \( F \) that is impaired by different noise realisations using sources of a similar type (e.g. Gaussian white noise) and strength. Formally, this is defined in Equation (11) as:

\[ ST(n) = \frac{1}{N} \sum \sum \{ I_j(x,n) \} \]  

(11)

where \( M \) is the number of pairs of images processed with different noise realisations and \( I_j(x,n) \) is an indicator function defined in Equation (12).

\[ I_j(x,n) = \begin{cases} 1 & g_1(x,n) = g_2(x,n) \\ 0 & \text{otherwise.} \end{cases} \]  

(12)

where \( g_1(x,n) \) and \( g_2(x,n) \) is a pair of consecutively filtered images, impaired by noise source 1 and noise source 2, respectively.

A third processing-related measure, is the Flat Field Noise Smoothing Ability \( (FFNSA) \) [12]. It is a measure for the maximum noise removal capability of a noise reduction algorithm. This maximum performance can be reached in areas where all noise rate reduction measures, such as detail-preservation or motion detection are switched off. In general, this happens in areas of an image without detail (flat field area) and without motion. The following measuring procedure can be used: construct a video sequence consisting of images having a constant (preferably \( F_{\text{max}}/2 \) luminance value, degrade this video sequence with (in our case: 26 or 32 PSNR Gaussian white) noise, filter the sequence with the optimal filter-settings obtained for each noise-strength and, consequently, assess the resulting sequences using, for instance, one of the luminance distortion measures, e.g. PSNRI.

Hybrid measures can be divided into two categories. The first is the category that makes certain additions to objective measures to include human vision characteristics. An important method is the addition of a frequency weighting curve to luminance distortion measures. Examples of these frequency weighting curves are the Fujio [4], CCIR 4211 and CCIR 5672 frequency weighting curves. The second category consists of measures that try to emulate the human vision characteristics directly by translating them into an objective measure. Examples here are the Spatial, Time- and Maximum Distortion Measures \( (SDM, TDM, Mdis) \) [16].

### SUBJECTIVE TEST SET-UP

The goal of the subjective tests was to evaluate both the quality performance and the noise removal capability of the noise reduction algorithms. The latter was included to detect the relation between the amount of noise that was removed and the resulting quality. Two subjective tests were necessary to limit the experimental time in each test to 40 minutes. In each subjective test, a set of 17 algorithms (see Table 1) was evaluated by using a paired comparison method, i.e. each algorithm was compared with all other algorithms for each sequence. The algorithms were tuned carefully for maximum noise removal and minimum distortion. Each pair of stimuli was shown twice on a split-screen to the test subjects. In the repetition measurement, the left-right position was switched to average out left-right preferences of the subjects. In the first subjective test 17 subjects participated, while in the second test 15 subjects gave their judgement. All subjects had a (corrected to) normal vision, as determined with the Landolt C-scale.

The sequences were displayed on a professional test monitor (Sony BVM-D24E1WE), which was set up using luminance and contrast controls by means of a PLUGE-variant test image [8]. Peak luminance level of the monitor was 180 \( \text{cd/m}^2 \), while the black level (measured in the viewing room with no illumination) was 0.2 \( \text{cd/m}^2 \). The monitor has been tested for homogeneity of the luminance and white-level over the CRT-screen. Results showed that the variation of these two parameters introduced by the
monitor was such, that an influence of the monitor on the test results could be neglected. The tests were conducted in a professional viewing room under controlled illumination conditions that complied with the ITU recommended viewing conditions [7]. Background illumination was set to 13.1 cd/m². Viewing distance was 6 times the height of the screen (i.e., 1.80 m). Each experiment started with a verbal instruction and a training session containing minimum and maximum quality difference pairs of stimuli that have been repeated later in the actual experiment. Subjects were asked to give their judgement on the stimuli with respect to quality and amount of noise on a comparison scale as presented in Figure 4. The underlining of SWAN+DNR(MC,HP) and DNR (MC,HP) indicates that these noise reduction algorithms cannot be distinguished in preference from each other, as determined with Tukey’s algorithm [15]. The overall quality results from the second test are given in Figure 5 separately for each of the three sequences, again combining the results for both noise levels. The ANOVA [10] combining the three sequences and two noise types, resulted in the following ranking of the noise reduction algorithms:

\[
\text{SWAN+DNR(MC,HP) > DNR(MC) > DNR(MC,HP) > SWAN > DNR.}
\]

RANKING USING SUBJECTIVE TESTS

The paired comparison data were translated into a z-scale ranking order of the algorithms using Thurstone’s model [3]. Since it was found that the scores on quality and noise removal correlated highly, namely with correlation coefficient \( r = 0.99 \), rankings are presented using the quality scores only. The overall quality results from the first test, combining the results for the two noise levels, are given in Figure 3, for each of the three sequences separately. An ANalysis Of VAriance (ANOVA) [10] revealed the following ranking of the noise reduction algorithms:

\[
\text{EPRES+PERCAT(MC) > SWAN+DNR(MC,HP) > DNR (MC,HP) > SWAN > FCST(MC).}
\]

The paired comparison data were translated into a z-scale ranking order of the algorithms using Thurstone’s model [3]. Since it was found that the scores on quality and noise removal correlated highly, namely with correlation coefficient \( r = 0.99 \), rankings are presented using the quality scores only. The overall quality results from the first test, combining the results for the two noise levels, are given in Figure 3, for each of the three sequences separately. An ANalysis Of VAriance (ANOVA) [10] revealed the following ranking of the noise reduction algorithms:

\[
\text{EPRES+PERCAT(MC) > SWAN+DNR(MC,HP) > DNR (MC,HP) > SWAN > FCST(MC).}
\]

Table 2: Performance results of remaining quality measures that show different ranking results for the sequences with 26 dB noise (Between brackets the ranking order of the noise reduction algorithms deducted from the quality measures.)

<table>
<thead>
<tr>
<th></th>
<th>SWAN+DNR</th>
<th>EPRES+PERCAT</th>
<th>DNR(MC)</th>
<th>SWAN</th>
<th>FCST</th>
<th>DNR</th>
<th>DNR(MC,HP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPR [%]</td>
<td>91.63(3)</td>
<td>91.39(4)</td>
<td>92.39(1)</td>
<td>89.18(5)</td>
<td>71.70(7)</td>
<td>78.91(6)</td>
<td>92.05(2)</td>
</tr>
<tr>
<td>DPC [dB]</td>
<td>36.77(1)</td>
<td>36.53(2)</td>
<td>36.2(3)</td>
<td>35.71(4)</td>
<td>34.82(6)</td>
<td>34.95(5)</td>
<td>34.38(7)</td>
</tr>
<tr>
<td>FFNSA [dB]</td>
<td>28.76(2)</td>
<td>28.62(3)</td>
<td>30.5(1)</td>
<td>27.48(7)</td>
<td>28.0(6)</td>
<td>28.60(4)</td>
<td>28.13(5)</td>
</tr>
<tr>
<td>Mito [%]</td>
<td>1.39(3)</td>
<td>1.22(2)</td>
<td>1.21(1)</td>
<td>1.80(5)</td>
<td>2.26(7)</td>
<td>1.75(6)</td>
<td>1.95(4)</td>
</tr>
<tr>
<td>SOM [-]</td>
<td>16.30(1)</td>
<td>12.81(2)</td>
<td>11.60(3)</td>
<td>13.39(4)</td>
<td>8.07(5)</td>
<td>7.77(6)</td>
<td>6.29(7)</td>
</tr>
<tr>
<td>ST [%]</td>
<td>4.62(1)</td>
<td>4.17(2)</td>
<td>3.34(3)</td>
<td>3.69(4)</td>
<td>3.34(5)</td>
<td>4.17(3)</td>
<td>3.13(7)</td>
</tr>
<tr>
<td>TM [-]</td>
<td>0.24(4)</td>
<td>0.25(1)</td>
<td>0.19(5)</td>
<td>0.19(6)</td>
<td>0.24(3)</td>
<td>0.17(7)</td>
<td>0.25(2)</td>
</tr>
</tbody>
</table>

Figure 4: Sample of the test form on which the subjects were asked to fill in their judgements on the displayed pair of stimuli.

Figure 5: Overall quality ranking result from subjective test 2. The subjective ranking values are displayed on an arbitrary z-scale. The error bars represent the 95% confidence interval of the mean.

For the sequences with 32 dB noise from the first test, the next ranking resulted:

\[
\text{EPRES+PERCAT(MC) > DNR(MC) > DNR(MC,HP) > SWAN > FCST(MC).}
\]

While EPRES+PERCAT(MC) was always the best, the difference between the noise reduction algorithms was less clear for the sequences with 32 dB noise than for the sequences with 26 dB noise. The latter was also the case for test 2, where the following quality ranking for the sequences with 26 dB noise was found:

\[
\text{SWAN+DNR(MC,HP) > DNR(MC) > DNR(MC,HP) > SWAN > DNR.}
\]

Another ranking resulted for the sequences of the same test with 32 dB noise:

\[
\text{DNR(MC) > DNR(MC,HP) > SWAN+DNR(MC,HP) > SWAN > DNR.}
\]

This suggests that differences between the noise reduction algorithms are more visible to the observers for the 26 dB noise level than for the lower noise level of 32 dB.

Table 2: Performance results of remaining quality measures that show different ranking results for the sequences with 26 dB noise (Between brackets the ranking order of the noise reduction algorithms deducted from the quality measures.)
Table 3: Performance results of remaining quality measures that show different ranking results for the sequences with 32 dB noise (Between brackets the ranking order of the noise reduction algorithms deducted from the quality measures.)

<table>
<thead>
<tr>
<th>Measure</th>
<th>SWAN+DNR</th>
<th>SWAN</th>
<th>DNR (MC)</th>
<th>FCST</th>
<th>DNR</th>
<th>EPRES+PERCAT</th>
<th>DNR (MC, HP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPR [%]</td>
<td>82.49 (4)</td>
<td>76.67 (5)</td>
<td>85.91 (3)</td>
<td>67.05 (6)</td>
<td>64.82 (7)</td>
<td>86.60 (2)</td>
<td>87.17 (1)</td>
</tr>
<tr>
<td>DPC [dB]</td>
<td>40.77 (2)</td>
<td>40.13 (6)</td>
<td>40.35 (3)</td>
<td>40.30 (4)</td>
<td>40.16 (5)</td>
<td>40.87 (1)</td>
<td>39.04 (7)</td>
</tr>
<tr>
<td>FFNSA [dB]</td>
<td>34.61 (1)</td>
<td>32.02 (5)</td>
<td>32.69 (4)</td>
<td>33.69 (2)</td>
<td>30.05 (7)</td>
<td>30.92 (6)</td>
<td>32.88 (3)</td>
</tr>
<tr>
<td>Fujio PSNR [dB]</td>
<td>40.09 (1)</td>
<td>39.65 (2)</td>
<td>38.91 (5)</td>
<td>39.30 (4)</td>
<td>39.58 (3)</td>
<td>38.01 (6)</td>
<td>37.21 (7)</td>
</tr>
<tr>
<td>MB [+]</td>
<td>10.41 (4)</td>
<td>10.68 (5)</td>
<td>9.46 (2)</td>
<td>10.37 (3)</td>
<td>10.92 (7)</td>
<td>8.51 (1)</td>
<td>10.87 (6)</td>
</tr>
<tr>
<td>Mdis [+]</td>
<td>0.79 (4)</td>
<td>0.97 (6)</td>
<td>0.82 (5)</td>
<td>0.76 (3)</td>
<td>0.68 (1)</td>
<td>0.74 (2)</td>
<td>0.98 (7)</td>
</tr>
<tr>
<td>MSE [-]</td>
<td>21.19 (1)</td>
<td>24.26 (4)</td>
<td>24.50 (5)</td>
<td>23.55 (2)</td>
<td>24.18 (3)</td>
<td>26.72 (6)</td>
<td>31.29 (7)</td>
</tr>
<tr>
<td>SDM [-]</td>
<td>15.76 (3)</td>
<td>15.09 (4)</td>
<td>19.34 (2)</td>
<td>14.99 (5)</td>
<td>11.99 (6)</td>
<td>25.36 (1)</td>
<td>11.03 (7)</td>
</tr>
<tr>
<td>ST [%]</td>
<td>7.01 (3)</td>
<td>5.69 (7)</td>
<td>6.61 (5)</td>
<td>6.61 (4)</td>
<td>8.81 (2)</td>
<td>8.81 (1)</td>
<td>5.97 (6)</td>
</tr>
<tr>
<td>TDM [-]</td>
<td>0.99 (5)</td>
<td>0.08 (6)</td>
<td>0.12 (3)</td>
<td>0.12 (4)</td>
<td>0.07 (7)</td>
<td>0.17 (1)</td>
<td>0.15 (2)</td>
</tr>
</tbody>
</table>

**RANKING THROUGH OBJECTIVE MEASURES**

Twenty objective and hybrid quality measures have been used to rank the performance of the noise reduction algorithms. These quality measures have been selected, because they seemed the most promising for describing the quality of noise reduction performance. The objective measures used were: \(\text{SNR(I)}\) [18], \(MD\) [5,18], \(MSE\) [18], \(MB\) [5,18], \(ST\) [5,18], \(CPR\) [5,18], \(FFNSA\) [12] and \(DPC\) [12]. The included hybrid measures were: Weighted \(\text{SNR}\) and Weighted \(\text{PSNR}\) using three different weighting characteristics, namely the Fujio [4], CCIR 4211 and CCIR 5672 weighting curves, and \(SDM\), \(TDM\), \(Mdis\) [16]. The noise reduction algorithms were ranked using all these measures to all the noise-reduced sequences of the two subjective tests. As an example, the \(\text{PSNR}\) performance results for sequences with 26 dB noise are given in Figure 6. The ranking of the noise reduction algorithms based on the \(\text{SNR(I)}\), \(MD\), \(MSE\) and \(MB\) objective measures is the same as that determined by \(\text{PSNR}\) and have, therefore, been left out. The CCIR 4211 Weighted \(\text{PSNR}\) performance results for the sequences with 26 dB noise are given in Figure 7. The ranking for the other hybrid measures using weighting curves did not differ from this one and, thus, these measures are not shown here. The rest of the measures is shown in Table 2 for the sequences with 26 dB noise.
dB noise. These measures did not give a ranking of the noise reduction algorithms equal to that of the ones already presented. For the sequences with 32 dB noise, a similar procedure was followed. The PSNR performance results are given in Figure 8. The same ranking is obtained by using the SNR, (P)SNRI and MD objective measures. Again, these measures have been omitted. The CCIR 4211 PSNR performance results for the sequences with 32 dB noise are given in Figure 9. This ranking is the same as the ranking obtained by the two CCIR 5672 (P)SNR measures. The ranking results of the remaining objective and hybrid measures are given in Table 3. This overview of the objective quality measures illustrates that all measures are able to distinguish between the different noise reduction algorithms. According to some objective quality measures, the performance of the noise reduction algorithms differ that much that some noise reduced 26 dB sequences result in a lower noise level than some noise reduced 32 dB sequences. This gives some indication of the large spread of actual noise levels used in the study.

CORRELATION BETWEEN SUBJECTIVE AND OBJECTIVE RANKING RESULTS

To determine which objective measure is best able to predict the subjective judgements on noise reduction algorithms, the correlation between the subjective quality scores and the values of the various objective quality measures have been calculated. Remind that the subjective quality scores were highly correlated to the subjective noise scores, and hence, the same results would have been found when using the latter ones.

Table 4: Correlation coefficients for each quality measure indicating its correlation with the subjective quality scores

<table>
<thead>
<tr>
<th>Quality Measure</th>
<th>Subjective test 1</th>
<th>Subjective test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPR</td>
<td>0.610</td>
<td>0.789</td>
</tr>
<tr>
<td>MB</td>
<td>-0.509</td>
<td>-0.375</td>
</tr>
<tr>
<td>SDM</td>
<td>0.453</td>
<td>0.498</td>
</tr>
<tr>
<td>Mdis</td>
<td>-0.432</td>
<td>0.001</td>
</tr>
<tr>
<td>ST</td>
<td>0.27</td>
<td>0.296</td>
</tr>
<tr>
<td>PSNRI</td>
<td>0.258</td>
<td>0.152</td>
</tr>
<tr>
<td>DPC</td>
<td>-0.254</td>
<td>0.068</td>
</tr>
<tr>
<td>TDM</td>
<td>0.239</td>
<td>0.275</td>
</tr>
<tr>
<td>MSE</td>
<td>-0.161</td>
<td>-0.09</td>
</tr>
<tr>
<td>FFNSA</td>
<td>-0.15</td>
<td>0.086</td>
</tr>
<tr>
<td>MD</td>
<td>-0.133</td>
<td>-0.095</td>
</tr>
<tr>
<td>PSNR</td>
<td>0.103</td>
<td>-0.082</td>
</tr>
<tr>
<td>SNR</td>
<td>0.091</td>
<td>0.048</td>
</tr>
<tr>
<td>CCIR 4211 SNR</td>
<td>-0.082</td>
<td>-0.078</td>
</tr>
<tr>
<td>CCIR 4211 PSNR</td>
<td>-0.062</td>
<td>-0.082</td>
</tr>
<tr>
<td>Fujio SNR</td>
<td>-0.044</td>
<td>-0.046</td>
</tr>
<tr>
<td>CCIR 5672 SNR</td>
<td>-0.044</td>
<td>-0.052</td>
</tr>
<tr>
<td>CCIR 5672 PSNR</td>
<td>-0.027</td>
<td>-0.053</td>
</tr>
<tr>
<td>Fujio PSNR</td>
<td>-0.022</td>
<td>-0.046</td>
</tr>
</tbody>
</table>

The correlation between the subjective quality and objective quality measures was first investigated by performing a two-tailed Pearson’s correlation analysis for each sequence and each noise level separately. The correlation results found, however, showed a large deviation, and there was no consistent effect of video content or noise level on the ranking of the noise reduction algorithms. Therefore, it was decided to repeat the correlation analysis on the averaged (over the sequences and noise levels) subjective quality score and the objective quality measure per noise reduction algorithm. In Table 4, the calculated correlation coefficients are given. Only four objective measures (CPR, MB, Mdis and SDM) correlated to some degree (|r| > 0.4) with the subjective quality scores of the two tests. The highest correlation coefficient, r, was 0.79 for the CPR measure in subjective test 2. Because of the poor correlation results of the individual objective measures with the subjective quality, it seemed interesting to see whether a combination of objective measures could perform better. The Mdis measure, however, has been left out of this further analysis, due to its extreme low correlation with the results from the second subjective test (0.001)\(^1\). SDM and MB correlated highly with each other and, as a consequence, SDM was removed from the subsequent analysis. Using the two remaining objective measures (CPR and MB), a regression model was found (Equation 13) to describe the subjective quality ranking results.

\[
\text{Quality} = 0.664 \cdot \text{CPR}_{\text{norm}} - 0.497 \cdot \text{MB}_{\text{norm}}
\]  

Note that the CPR and MB objective data have been normalized to have a zero mean and unity variance distribution, so that the regression model’s coefficients are meaningful. This model could predict the perceived quality scores of the two subjective tests only with an overall descriptive power of 59%, which is quite low. If using the test results of each subjective test separately to calculate a predictive model with the same two objective measures, the model describing the second subjective test results gave a much higher descriptive power of 84%. For the first test the descriptive power was much lower (50%). Probably, this difference is caused by the fact that the algorithms used in the second test are more similar than the algorithms used in the first test. It should be remarked, that other combinations of objective measures have been tried in the regression analysis. None of these attempts, however, resulted in a higher descriptive power for the subjective quality scores. From the results of the correlation

---

\(^1\) In [16], an improved version of the Mdis quality measure has been suggested, which was not available at the time that this study was conducted.
analysis it is clear that none of the objective measures nor combinations of them can be used to reliably approximate the subjective ranking of the performance of noise reduction algorithms. In a more limited application, however, it could be possible that one of the objective quality measures is able to select reliably the best performing noise reduction algorithm (without bothering about the rest of the ranking). Since only twice from a group of noise reduction algorithms, a selection could be made, the reliability of the attempt is limited. But even for this limited sample size, none of the measures could select the noise reduction algorithm with the best perceived performance for both tests.

CONCLUSIONS

Several state-of-the-art noise reduction algorithms have been ranked on quality and on noise removal in a subjective evaluation and by using twenty objective quality measures. A measurement set-up comprising a simulation of a TV-transmission channel has been used to provide a realistic assessment of the performance of the noise reduction algorithms commonly applied in TV-applications for reducing Gaussian white noise. From the analysis of the objective quality measures and subjective quality results, it is shown that the correlation of most of the objective quality measures with perceived quality or perceived noise removal is low. Even a combination of objective measures only approximates the subjective assessment of the quality of noise reduction algorithms with a descriptive power as low as 59%. Since this study is to our knowledge the only one assessing noise reduction performance on video both objectively and subjectively, a comparison with the results of other studies is not possible.

REFERENCES