QUALITY MEASURES FOR SINGLE CHANNEL SPEECH ENHANCEMENT ALGORITHMS

Pia Dreiseitel

Signal Theory
Darmstadt University of Technology
Merckstr. 25, Darmstadt, Germany
pia.dreiseitel@nt.tu-darmstadt.de

ABSTRACT

We give an overview of speech quality measures and apply them to the typical shortcomings of speech enhancement algorithms. This is important for comparing different approaches to noise reduction, and it suggests ways to improve a speech enhancement algorithm. Objective quality measures are compared to the subjective quality evaluation of human listeners by means of extensive listening tests. As a result, we propose a hybrid method for measuring the quality of speech enhancement algorithms.

1. INTRODUCTION

Measuring the quality of speech enhancement algorithms is a difficult task. Many different aspects affect the overall quality of a speech enhancement algorithm and it is not easy to measure how a speech signal changes when an enhancement algorithm is applied. The term "speech enhancement is used here interchangeably with "noise reduction. We examine only single channel noise reduction systems, and the application we have in mind is the enhancement of heavily distorted speech signals, e. g. the hands-free telephone environments for cars. Therefore the optimal solution for the enhanced signal is always the plain speech signal itself. The reader should recall that speech enhancement here means restoring the speech signal from a distorted signal.

Very often, only the attenuation of the background noise is taken as a measure for the performance of a speech enhancement algorithm. Sometimes two properties of the signal are under investigation: the degradation of the speech signal and the attenuation of the background noise [10]. However, this appears not to be sufficient. We propose three different classes of properties instead:

- Variations in the pure speech signal (negative)
- Variation in the noise characteristics (negative)
- Attenuation of the background noise (positive)

All of these classes can be examined by a large number of distance measures, called symptoms. A small number of examples will be given later in this article. It appears to be obvious that different noise reduction systems affect the symptoms in different ways leading to a different performance. We now want to find out which of the symptoms above are relevant for a general quality measure.

After intensive listening tests, that are discussed later in this article, an opinion poll about speech enhancement systems was performed. More than thirty test persons were asked about their impression of what is absolutely important for a speech enhancement system. The poll results are depicted in Fig. 1. It may be a bit astonishing that the actual noise attenuation is the least crucial point of the noise reduction system. Speech degradation or an unnatural characteristic of the remaining noise is far more important to the overall judgement.

The noise reduction algorithms under investigation in this paper are the spectral subtraction rule [1] or the MMSE-estimation and its derivations [4, 5]. An overview on noise reduction algorithms is given in [3, 6].

2. MEASURING THE QUALITY OF SPEECH ENHANCEMENT ALGORITHMS

Probably right from the beginning of digital speech processing, people were also interested in objective quality measures [7, 8], most of which were distance measures between the processed signal and the original speech signal.

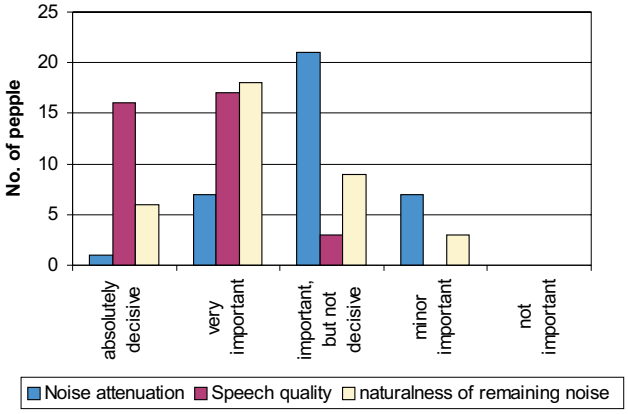

Figure 1: Results of an opinion poll after an intensive listening test. Speech quality appears to be the most crucial symptom.

2.1. Speech quality

As we learn from Fig. 1, speech quality is the most crucial part of a noise reduction system. Therefore we start with investigating the speech quality alone. Speech quality degradation again can occur in different ways. While phase distortions are usually neglected, nonlinear distortions or attenuation of parts of the speech signal change the speech quality significantly. Typical measures for speech quality, known from speech coding, were used for the speech quality evaluation.

1. Cepstral distance:

$$d_C = \frac{\sum_{k=1}^{K-1}(c_s(k) - c_{\hat{s}}(k))^2}{\sum_{k=1}^{K} c_s^2(k)}, \qquad (1)$$

$$c_s(k) = \mathcal{F}^{-1} \log |\mathcal{F}(s(k))|, \qquad (2)$$

where $\mathcal{F}(\cdot)$ denotes the Fourier transform of the input signal and $\hat{s}(k)$ denotes the estimated value of the speech signal $s(k)$.

2. Itakura measure:

$$d_I(\mathbf{a}, \mathbf{b}) = \log\left(\frac{E'}{E}\right) = \log\left(\frac{\mathbf{b}^T \mathbf{R}_{ss} \mathbf{b}}{\mathbf{a}^T \mathbf{R}_{ss} \mathbf{a}}\right) \qquad (3)$$

with $\mathbf{a}$ and $\mathbf{b}$ being the coefficients of a predictor error filter trained with the two signals to be compared. Both predictors, however, are used with the same input signal, so E' is the predictor output of the distorted signal that is computed with a predictor adapted to the clean speech signal.

3. Itakura-Saito measure:

$$d_{IS}(\mathbf{a}, \mathbf{b}) = \log\left(\frac{(\mathbf{a}-\mathbf{b})^T \mathbf{R}_{ss} (\mathbf{a}-\mathbf{b})}{\mathbf{a}^T \mathbf{R}_{ss} \mathbf{a}}\right), \qquad (4)$$

The Itakura-Saito measure shows similar results as the Itakura measure but is symmetric in $\mathbf{a}$ and $\mathbf{b}$.

The speech quality measures are also discussed in [9, 12, 11, 13].

2.2. Noise characteristics

Early listening tests and also various publications of speech enhancement algorithms point out a typical deficiency of noise reduction or speech enhancement algorithms. They tend to change the characteristics of the background noise. Tonal sounds, often called musical noise, are often inherent to speech enhancement. However, if human listeners are asked for their preferences they want the remaining noise to sound natural. If a hands-free telephone is installed in a car, the remaining noise should sound like car noise.

Noise attenuation

The actual goal of a noise reduction algorithm is the attenuation of the background noise without attenuation of the speech signal. A simple measure for the performance of a speech enhancement system is therefore the average attenuation of the background noise or in other words the enhancement of the signal-to-noise ratio which can be defined by the following equation:

$$d_{att} = \frac{\overline{n_{out}^2}}{\overline{n^2}}. \qquad (5)$$

Both the noise power before and after processing are averaged over time.

Musical noise

Especially the tonal parts in the remaining noise disqualify a noise reduction system. The measure denoted in the following equation gives a hint about the tonal distortions present in the outgoing signal. Since tonal distortions are visible as short-term variations in the periodogram, we compare the periodogram to a modell-based spectrum estimation.

$$d_{ton} = \frac{\sum_{n=0}^{N-1} |\hat{\Phi}_{nn-lpc}(n) - \hat{\Phi}_{nn-per}(n)|^2}{\sum_{n=0}^{N-1} |\hat{\Phi}_{\tilde{n}\tilde{n}-lpc}(n) - \hat{\Phi}_{\tilde{n}\tilde{n}-per}(n)|^2}. \qquad (6)$$

Difference in power level

Not really surprisingly, the difference in the noise power if compared in sequences with or without speech activ-

	Ceps.	Itak.	Hybrid.
Spektral Subtr., VAD	0.38	0.05	1.77
MMSE (Ephraim/Malah)	0.47	0.10	1.55
MMSE-log (Ephraim/Malah)	0.40	0.06	2.10
Spektr. Subtr., VAD, SNR=0 dB, VA: $\alpha = 1.0,\ \beta_f = 0.5$, P: $\alpha = 4.0,\ \beta_f = 0,1$	0.37	0.06	1.91

ity also gives a hint of the noise reduction quality.

$$d_P = \frac{(\tilde{n}_a^2 - \overline{\tilde{n}^2})^2}{(\overline{\tilde{n}^2})^2} + \frac{(\tilde{n}_p^2 - \overline{\tilde{n}^2})^2}{(\overline{\tilde{n}^2})^2}, \qquad (7)$$

with $\tilde{n}_p^2$ and $\tilde{n}_a^2$ being the current values of background noise in pauses or speech activity, respectively and $\overline{\tilde{n}^2}$ the average value of the background noise.

2.3. Psycho-acoustic methods

There are of course many other quality measures available. A very common approach is to evaluate the psychoacoustical masking properties of the human hearing system [2]. Typically psychoacoustic methods outperform simple signal-to-noise ratio enhancement measures. Since the discussion of psychoacoustic methods opens the completely new field, we leave this class of measures out in this paper.

3. THE HYBRID METHOD

Putting together a number of different quality measures covers a wider range of possible influences and therefore gives a better idea of what happened to the speech signal [2, 11, 13]. The linear combination of different quality measures for speech or noise

$$D = a_1 \frac{1}{d_d} + a_2 d_t + a_3 d_P \qquad (8)$$

is carried out with an optimization for the least mean squared error between the objective judgement D and the judgement quotient Q_Σ which is explained later in this paper.

4. TESTING THE QUALITY MEASURES

The objective quality measures were tested with a reference noise reduction algorithm where well-known parameter modifications were performed. The spectral

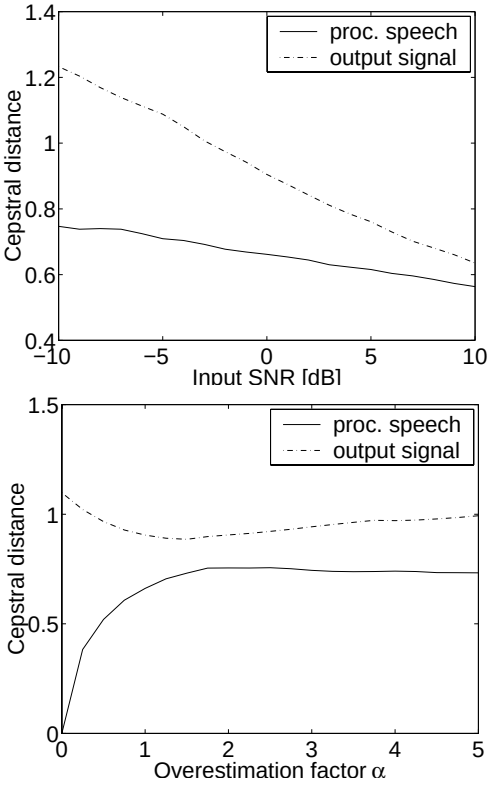

Figure 2: Cepstral distance versus overestimation factor or input signal to noise ratio.

subtraction rule

$$G(\Omega) = \max\left(\beta, 1 - \alpha \frac{\Phi_{nn}(\Omega)}{\Phi_{xx}(\Omega)}\right) \qquad (9)$$

was taken with $\beta = 0$ and a variable overestimation factor α or a variable input signal-to-noise ratio, respectively. Here $\Phi_{xx}(\Omega)$ and $\Phi_{nn}(\Omega)$ are the power spectral densities of the distorted input signal and the background noise, respectively. It is known that increasing the overestimation factor lowers the amount of musical noise but also distorts the speech signal [3]. Increasing the input signal-to-noise ratio also delivers a known behaviour: The speech signal becomes less distorted and also the background noise characteristic becomes more natural. In Fig. 2 we see that the speech quality is related to the overestimation factor. The dash-dotted line shows the cepstral distance between the original speech signal and the estimated speech signal at the output of the system. The two signals are most similar for the overestimation factor equals 1.25.

5. LISTENING TESTS

Since the recipient for speech enhancement systems is (in our case) the human listener, there is a need of comparing the test results of objective measures to listening tests performed with human listeners. Therefore a listening test with 35 persons was carried out. The test persons were asked to mark the speech signal and the background noise separately, and to say if in their impression the speech enhancement is a general improvement. Various speech enhancement schemes were used as test sequences such as the MMSE estimator or controlled spectral subtraction rules which can be found in [3].

5.1. Mean Opinion Score

A very popular score for speech quality is the Mean Opinion Score (MOS). However, the typical averaging of the outcomes of the listening tests was not performed here since the scale on which the marks were given does not have a specific quantization. So it remains very questionable if any conclusions can be drawn out of an arithmetic average. However, for the listening test the marks of the Mean Opinion Score were taken.

5.2. Judging quotient

In contrast to the normally used mean opinion score, we do not average the outcomings of an opinion poll since there is no warranty that the judgements are on a linear scale. However, the listening tests were carried out using the levels of the MOS. To achieve a scaled outcome of the listening tests we use a judgement quotient similar to the quantiles known in statistics:

$$Q(i) = \frac{\text{No. of judgements above level } (i-1)}{\text{Total number of judgements}} \tag{10}$$

We usually use the sum of all levels for this evaluation for a better averaging.

$$Q_\Sigma = \sum_{i=2}^{N} Q(i) \tag{11}$$

6. COHERENCE BETWEEN INSTRUMENTAL MEASURES AND LISTENING TESTS

If some information is required about the coherence between the subjective and the objective quality measures, usually the correlation coefficient is used. Since objective quality measures are typically not on the same scale as the subjective measure, usually a nonlinear fitting is performed. Since the amount of training data is not very large, due to the enormous effort that listening tests require, the results are heavily related to the fitting curves. We avoid this problem by using the so-called rank correlation.

6.1. Rank correlation

The data under investigation (subjective as well as objective) is put in a rising order. If for example

$$\begin{aligned} O_3 < O_2 < O_1 < O_4 \qquad & \text{(objective)} \\ \text{and} \quad S_1 < S_2 < S_4 < S_3 \qquad & \text{(subjective).} \end{aligned}$$

We therefore get the following ranks:

$$\begin{aligned} \mathcal{R}(O_3) = \mathcal{R}(S_1) = 1, \qquad & \mathcal{R}(O_2) = \mathcal{R}(S_2) = 2, \\ \mathcal{R}(O_1) = \mathcal{R}(S_4) = 3, \qquad & \mathcal{R}(O_4) = \mathcal{R}(S_3) = 4, \end{aligned}$$

where $\mathcal{R}(S_j)$ denotes the rank of a subjective quality measure, $\mathcal{R}(O_j)$ that of an objective quality measure, respectively. Analogous to the correlation coefficient, we calculate the rank correlation coefficient:

$$\rho_s = \frac{\sum_{i=1}^{U} \left(\mathcal{R}(S_i) - \overline{\mathcal{R}(S)}\right)\left(\mathcal{R}(O_i) - \overline{\mathcal{R}(O)}\right)}{\sqrt{\sum_{i=1}^{U} \left(\mathcal{R}(S_i) - \overline{\mathcal{R}(S)}\right)^2} \sqrt{\sum_{i=1}^{U} \left(\mathcal{R}(O_i) - \overline{\mathcal{R}(O)}\right)^2}}, \tag{12}$$

where $\overline{\mathcal{R}(O)}$ and $\overline{\mathcal{R}(S)}$ denote the respective average rank and have the same value, namely $U/2$.

6.2. Results

Tab. 1 shows the rank correlation between the distance measures for the speech signal and the subjective quality of the speech signal. If the quality of the speech signal alone is required, the cepstral distance shows the best correlation.

Table 1: Rank correlation ρ_s of subjective judgement and speech quality measures.

Speech measure	Q_Σ
Itakura	-0.48
Itakura-Saito	-0.48
Cepstral distance	-0.62

For the analysis of the change in the noise characteristics, we also compare the objective measures for the background noise with the results of the listening test (see Tab. 2). The measure for the tonal distortions delivers the highest agreement with the human listeners, while the actual noise attenuation does not give a prediction of the noise quality.

Table 2: Rank correlation ρ_s of subjective judgement and noise quality measures.

Noise measure	Q_Σ
Noise attenuation	0.35
Tonal distortion	-0.67

Finally, we compare the quality measures for the global performance with the overall mark of the test persons.

Table 3: Rank correlation ρ_s of subjective judgement and general quality measures.

Global measure	Q_Σ
Hybrid method	0.81
Psycho SNRE	0.17
SNRE	-0.10

It is obvious from Tab. 3 that a hybrid method including the various speech and noise measures outperforms a simple signal-to-noise ratio. We also tried a psychoacoustically-motivated signal-to-noise ratio with only a slight improvement over the normal one.

7. CONCLUSIONS

As we see from the results a perfect correspondence between subjective and objective quality measures is not possible. However, for the design of a new noise reduction algorithm the proposed quality measures help to yield information for tuning and comparing noise reduction systems.

8. REFERENCES

[1] Boll, S.F., Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 27(2):113–120, April 1979.

[2] Dreiseitel, P., Quality evaluation of noise reduction algorithms. In *Proc. 6th Int. Workshop on Acoustic Echo and Noise Control*, Pocono Manor, USA, September 1999.

[3] Dreiseitel, P., Hänsler, E., and Puder, H., Acoustic echo and noise control - a long lasting challenge. In *Proc. EUSIPCO-98, 9th European Conference On Signal Processing*, Island of Rhodes, Greece, September 1998.

[4] Ephraim, Y. and Malah, D., Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Speech and Audio Processing*, 33(2):443–445, April 1984.

[5] Ephraim, Y. and Malah, D., Speech enhancement using a minimum mean-square error short-time amplitude estimator. *IEEE Transactions on Speech and Audio Processing*, 32(6):1109–1121, December 1984.

[6] Gold, B. and Morgan, N., *Speech and Audio Signal Processing*. John Wiley & Sons, Inc., New York, 1 edition, 1999.

[7] Gray, A.H. and Markel, J.D., Distance measures for speech processing. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 24(5):380–391, 1976.

[8] Gray, R.M., Buzo, A., Gray, A.H., and Matsuyama, Y., Distortion measures for speech processing. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 28(4):367–376, 1980.

[9] Hansen, J.H.L. and Clements, M.A., Use of objective speech quality measures in selecting effective spectral estimation techniques for speech enhancement. In *Proceedings of the 32nd Midwest Symposium on Circuits and Systems*, pages 105–108, 1990.

[10] Le Bouquin, R., Faucon, G., and Akbari Azirani, A., Proposal of a composite measure for the evaluation of noise cancelling methods in speech processing. In *Proceedings of the EUROSPEECH'93*, pages 227–230, Berlin, September 1993.

[11] Magotra, N., Kirstein, M., Sirivara, S., and Hamill, T., Quantitative and qualitative (subjective) perceptual measures for speech processing applications. In *Conference Record of the Thirtieth Asilomar Conference on Signals, Systems and Computers, 1996.*, pages 766 –769, Asilomar, USA, 1997.

[12] Quackenbusch, S.R., Barnwell, T.P., and Clements, M.A., *Objective measures of speech quality*. Prentice Hall Signal Processing Series. Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[13] Wang, S., Sekey, A., and Gersho, A., An objective measure for predicting subjective quality of speech coders. *IEEE Journal on Selected Areas in Communications*, pages 819–829, June 1992.