

Factors Impacting Objective Algorithms for Speech Quality Assessment on Mobile Networks

Technical Paper

© Ascom 2009. All rights reserved.

TEMS is a trademark of Ascom. All other trademarks are the property of their respective holders.

No part of this document may be reproduced in any form without the written permission of the copyright holder.

The contents of this document are subject to revision without notice due to continued progress in methodology, design and manufacturing. Ascom shall have no liability for any error or damage of any kind resulting from the use of this document.

Contents

1.	Abstract.....	4
2.	Introduction	4
3.	Impacting Factors on Speech-Quality Assessment.....	5
4.	Analysis Procedure and Results	7
4.1.	The Test Database	7
4.2.	Statistical Metrics Used for the Evaluation.....	7
4.3.	The Pearson Correlation Coefficient.....	8
4.4.	The Root-Mean-Squared Error	8
4.5.	Error Distribution	9
5.	Performance of the Algorithms.....	10
5.1.	Technology Dependency	10
5.2.	Outliers.....	10
5.3.	Interpretation of the Statistical Metrics	11
6.	Analysis of the Impacting Factors	13
7.	Conclusions	14

1. Abstract

Speech-processing procedures and algorithms are designed to ensure speech quality over end-to-end communication links. Subjective methods for speech-quality evaluation are less efficient and require too much time to accomplish. In an increasingly demanding real-time field-measurement environment, objective measures have turned out to be more suitable than subjective ones.

Recently, a number of algorithms have been developed which, while using almost the same building blocks, are biased towards different speech-quality assessment factors. Some of these algorithms have been analyzed using a comprehensive mobile-network database, and the results point out valuable conclusions regarding how these measures perform in the wireless domain and what main factors affect their performances.

2. Introduction

Currently, all objective measurement systems for speech-quality evaluation are perceptual-domain measures. They use two signals as their input, an original signal (reference pattern) and the corresponding output signal after its transition through the network under test. Signal processing within the perceptual measures consists of three major steps: pre-processing ([1], [2]); psycho-acoustic modeling ([1], [2]), and cognitive modeling ([1], [2], [3]) (see Table I). Detailed analysis and comments regarding the processing steps can be found in [4].

Cognitive modeling is a complex process that differentiates between *objective* methods of assessing *subjective* judgment of speech quality. The assessment is accomplished by determining a perceptual distance between the measured signal and the reference and then by creating a figure of merit that describes speech quality. The figure of merit is generally a non-linear function of the subjectively determined MOS (mean opinion score) value. In order to obtain an objective estimator for the MOS value, it is necessary to map the objective result to the MOS scale, which ranges from 1 to 5. This mapping is usually called calibration.

The best-known algorithms for objective speech-quality evaluation based on a psycho-acoustic sound-perception model are: BSD (Bark Spectral Distance) [5], WSSD (Weighted Spectral Slope Distance) [6], PSQM (Perceptual Speech Quality Measure) [1], MNB 1 & 2 (Measuring Normalizing Blocks) [2], PAMS (Perceptual Analysis Measurement System) [7], [8], TOSQA [9] and PESQ (Perceptual Evaluation of Speech Quality) [11].

3. Impacting Factors on Speech-Quality Assessment

It should be noted that these algorithms implement and combine almost the same building blocks (Table I), but in different ways and biased towards different factors affecting speech-quality evaluation. The importance of the impacting factors and the combination of the blocks determine differences in performance between algorithms.

The typical Hoth noise characteristic and the acoustical-electrical frequency response of the handset define the environmental model. A perceptual model of the human evaluation of speech quality should take account of the environmental model. However, not all algorithms do.

The algorithms use filter banks or FFT to model the psycho-acoustic perception process. It is known that the choice of any of these solutions does not affect the performance of the algorithm [10].

Neither do the algorithms bear the time-masking process. It is shown in [1] that this process has no effect on the performance of the algorithms designed for speech-quality assessment.

To deal with filtering in the system under test, a compensation for the transfer function is considered. This becomes important with regard to severe filtering that can be disturbing to the listener.

The short-term gain-variation impact factor is accounted for by using a compensation for time-varying gain variations between original and degraded signals.

While the pre-processing step and the psycho-acoustic model describe the process of human hearing, only the cognitive model considers the complexity of human judgment in perceiving speech quality. There is no point in developing a high-quality algorithm for the psycho-acoustic process without modeling some of the important cognitive effects, such as asymmetry and silence-interval processing [3]. Asymmetric distance is the only factor that accounts for complex perceived distortions such as the ones determined by time-clipping effects or codecs with noise suppressors during silence intervals.

A bi-dimensional figure of merit creates a good possibility of integrating the effects of more-important factors under the single value of the speech-quality estimator.

The calibration (mapping to the MOS scale) is more a part of the evaluation process than of the algorithm itself. A comparison-evaluation analysis requires that all algorithms be exposed to the same mapping procedure.

TABLE I PROCESSING BLOCKS OF ALGORITHMS FOR OBJECTIVE SPEECH-QUALITY EVALUATION

Processing Step	Block	Ag1	Ag2	Ag3	
Pre-Processing	Adjustment unit	X	X	X	
	Environmental model (Hoith noise characteristic and IRS filtering)	X	X	X (no noise)	
Psycho-acoustic perception Mode 1	Solution 1	Time-frequency transformation		X	X
		Bark transformation		X	X
		Frequency masking		X	X
		Time masking			
		Compensation for the transfer-function equalization			X
		Compensation for time-varying gain variations between original and degraded signals			X
		Loudness transformation 1,2,3*		X/2	X/3
	Solution 2	Critical filter – bank analysis	X		
		Frequency masking	X		
		Time masking			
Loudness transformation 1,2,3*		X/1			
Cognitive Mode 1	Type of distance	Generalized/ Euclidean distance	X		
		Asymmetric distance		X	
		Symmetric and asymmetric distance with different power of integration			X
	Types of figures of merit	Uni-dim figure of merit (w s.1.p**)		X	
		Uni-dim figure of merit (w/o s.1.p.)	X		
		Bi-dim figure of merit (w s.1.p)			X
		Bi-dim figure of merit (w/o s.1.p)			
	Calibration				

* 1 defines a simple logarithmic transformation, 2 defines the complex Zwicker transformation and 3 defines the Zwicker transformation with recency effect

** s.l.p=silence interval processing

4. Analysis Procedure and Results

The analysis conditions defined here were used to evaluate some current objective methods designed to estimate subjective speech quality for mobile networks.

Three algorithms were chosen for the analysis. The goal was to track the effect of some blocks and impacting factors on algorithm performances.

4.1. The Test Database

The database contains field-collected samples for IS-136 850MHz and 1900MHz (8kb/s VSELP & 8kb/sACELP vocoders), CDMA 850MHz (13kb/sQCELP vocoder), GSMEFR, iDEN and AMPS networks. The collected samples correspond to four pairs (each of seven seconds' length) of two sentences, spoken by four talkers (two male, two female) distorted by different mobile network conditions.

The subjective ratings were gathered from a subjective listening test called a Mean Opinion Score test (also referred as Absolute Category Rating, ACR). In order to eliminate the randomness and the risk of a high standard deviation in the MOS panel, each sample (speech file) obtains an average subjective score of 44 votes.

4.2. Statistical Metrics Used for the Evaluation

In order to have a fair comparison, the outputs of the algorithms are mapped to the subjective scale 1 to 5. The mapping is accomplished using a logistic function defined by:

$$y = \frac{1}{1 + \exp(c * x + d)}$$

The comparison between the mapped output values of each algorithm is done using three statistical measures: the correlation coefficient per files, the root-mean-squared error and the error distribution. The analysis is performed per file as recommended for strong time-variant systems, such as mobile networks.

4.3. The Pearson Correlation Coefficient

The Pearson Correlation Coefficient gives a basic measure for quality of fit between objective and subjective scores, describing the extent to which data points are scattered with respect to the linear mapping $y=ax+b$

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

By y_i are denoted the subjective scores and by x_i the objective scores.

4.4. The Root-Mean-Squared Error

The root-mean-squared-error (rmse) is the standard deviation of the set of residuals with normalisation to N:

$$RMSE = \sqrt{\frac{1}{N} \sum (x_i - y_i)^2}$$

N represents the number of speech files considered in the analysis. Since each of the speech files is graded by the algorithm with a single value, N also equals the number of objective scores.

The results of the comparison are presented in Table II.

TABLE II CORRELATION COEFFICIENTS PER FILES AND ROOT MEAN SQUARE ERROR

Tech	Algorithm1		Algorithm 2		Algorithm 3	
	Corr	Rmse	Corr	Rmse	Corr	Rmse
Netw1	0.85	0.62	0.88	0.54	0.92	0.44
Netw2	0.79	0.58	0.91	0.37	0.94	0.31
Netw3	0.83	0.51	0.92	0.36	0.96	0.27
Netw4	0.82	0.40	0.92	0.29	0.91	0.31
Netw5	0.88	0.41	0.89	0.4	0.94	0.34
Netw6	0.75	0.45	0.85	0.35	0.88	0.36
Netw7	0.88	0.32	0.78	0.43	0.92	0.29

4.5. Error Distribution

The error distribution is calculated using the residual error resulting from the application of the correlation line $y=x$ to a data set

$$e_i = x_i - y_i$$

Residuals are the errors arising when the objective quality predictions are mapped to the given subjective decisions.

The error distribution is a powerful tool for comparison analysis regarding the performances of algorithms. For example, algorithm 1 and algorithm 2 perform almost the same correlation coefficient and rmse for network 5 case (Table II), but algorithm 2 would be preferred. It is better to have the majority of the files with an error up to 0.5 MOS, instead of having many with errors up to 1 MOS, as shown in Figures 1 and 2.

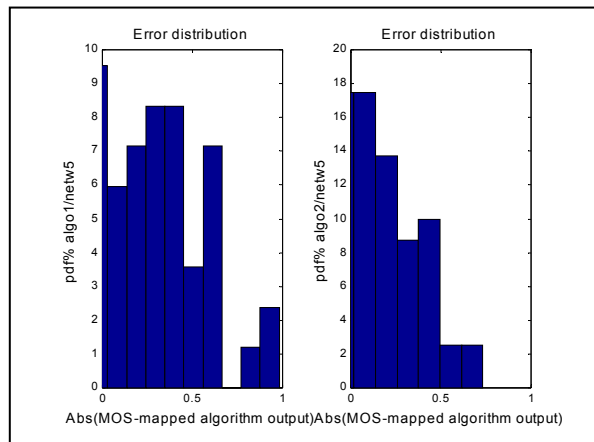


Figure 1

Figure 2

5. Performance of the Algorithms

In the conditions of this test, over all technologies, algorithm 3 performs best, as shown in Table II. Analyses of the results in Table II and of the algorithm's design shown in Table I unveil some important issues that need to be understood.

5.1. Technology Dependency

Within the same application type (mobile networks), the performances of the algorithms depend on the technology. Algorithm 3 consistently handles all technologies with almost the same performance accuracy. Algorithms 1 and 2 seem to be biased toward some technologies when it comes to accuracy. The ideal algorithm should exhibit the same performance regardless of distortion type.

5.2. Outliers

Outliers are defined as the points for which the rmse is very high in comparison with the average rmse value. The error distribution shape has a stretched tail (Figure 3), which could be artificially interpreted as errors of the algorithm. Thus, the outliers can alter the accuracy of the evaluation analysis and should be studied before estimating performance.

The outliers are generally caused by the anomalies in the MOS panel due to the randomness of the subjective results. That is pretty much removed by comparing the objective score (the output of the algorithm) given to a speech file with an average of a number of subjective scores for the same file.

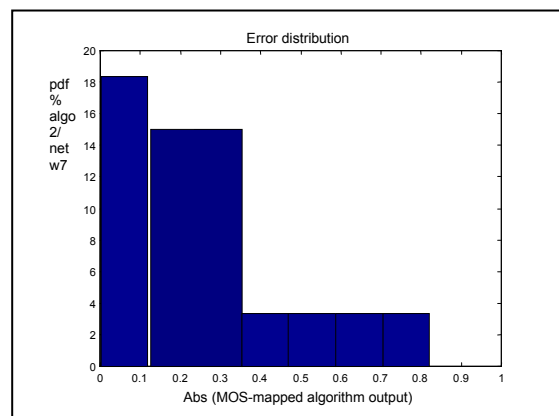


Figure 3

5.3. Interpretation of the Statistical Metrics

Statistical metrics should be viewed from three angles: the dependency of the algorithm output on the MOS scores, the mapping effects, and the error distribution.

1. The algorithm output-MOS scale dependency Widely spread and highly non-linear dependencies (Figure 4) determine low correlation coefficients (Table II). Also, the dependencies characterized by a lot of isolated points will cause low correlation coefficients. High linear dependencies between the MOS scale and the algorithm output (Figure 5) determine high correlation coefficients (Table II).



Figure 4

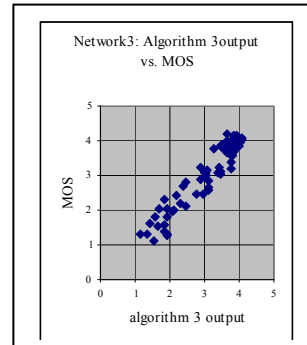


Figure 5

2. The mapping effect The mapped algorithm outputs could be more or less correlated with the MOS scores (compare Figures 6 and 7) than the raw outputs (compare Figures 8 with 9). So, despite its necessity, the mapping function might slightly alter the performance of the algorithms. Generally, very widespread algorithm outputs are brought together through the mapping function. The mapping function mainly decreases the performance for algorithms characterized by isolated points.

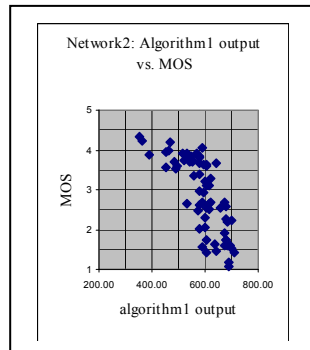


Figure 6

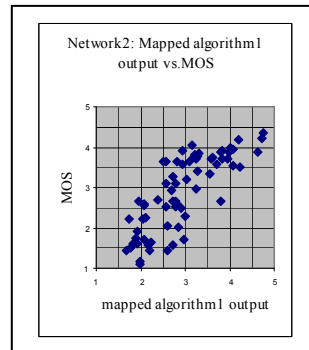


Figure 7

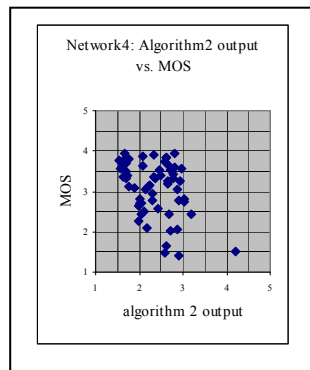


Figure 8

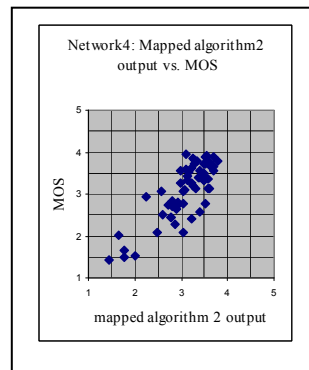


Figure 9

3. Error distribution Low correlation coefficients and high rms errors determine a flat, spread error-distribution shape (Figure 11). The speech files, which perform errors up to 0.5 MOS, tend to be as frequent as the ones which perform errors up to 1 MOS and higher. High correlation coefficients and low rmse correspond to narrow and peaked error-distribution shapes, concentrated at up to 0.5 MOS or less (Figure 10).

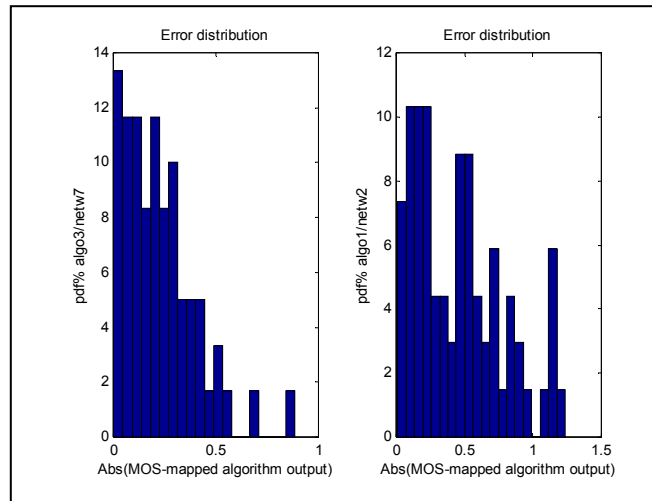


Figure 10

Figure 11

6. Analysis of the Impacting Factors

The main impacting factors are defined by the environmental model, the psycho-acoustic model, and the cognitive model.

1. The environmental model is not part of algorithm 1. Since algorithm 1 performs poorly, it can be concluded that the environmental model should be considered for speech-quality evaluation. On the other hand, comparing the results for algorithms 2 and 3 (Table I), it can be seen that the Hoth noise characteristic negatively affects algorithm performance. A simple IRS filter modeling the handset acoustical-electrical frequency response should be sufficient.

2. The psycho-acoustic model conveys the following factors: time-frequency and loudness transformation types, and the compensation procedure for the tested system's characteristics.

Unlike algorithms 2 and 3, algorithm 1 uses critical filter-bank analysis for the time-frequency transformation. The comparison between algorithm 1 versus 2 and/or 3 on one side, and between algorithm 2 versus 3 on the other side, shows that the solution chosen for the time-frequency model does not affect the performances of the algorithm ([10]).

The results show that a simple logarithm, loudness transform (algorithm 1), is not sufficient to convert the speech-signal degradation into human-perceived deterioration. The complex Zwicker [10] model comprising the recency effect (algorithm 3) would be recommended. The recency effect is a powerful subjective factor, weighting perceivably more the last-heard distortion than the one at the beginning of the speech. It is required for an objective measure to model it, in order to achieve an accurate estimator for the subjective opinion. The recency effect becomes more compelling when it comes to long speech segments (telephone conversations).

The compensation procedure for the tested system's characteristics (e.g. frequency response of the handset, gain, other filtering effects) is one of the most important factors that affect speech-quality assessment. Subjective opinion is very sensitive to any type of speech processing. As a simple exercise, an evaluation of different phones for the same wireless network will show a slight difference of performance for each phone. An objective measure compensating for a system's characteristics (e.g. algorithm 3) will be able to track subjective sensitivity, conferring a better estimate of the subjective opinion.

3. The cognitive model The use of the combined asymmetric and symmetric distance with different powers of integration (unlike the Euclidean power of 2) ensures the weighting of the degradation in accordance with human perception (algorithm 3). The consideration of the silence intervals becomes salient, especially when background noise exists (algorithm 3). It is shown in [3] that during unvoiced intervals, the distortions, which require asymmetric distance (e.g. time clipping) are more significant.

7. Conclusions

A thorough and comprehensive analysis regarding some current objective measures for speech quality on mobile networks has been accomplished.

Based on analysis of the impacting factors that affect speech-quality assessment, it is recommended for an algorithm to embody the environmental model, the complex loudness transformation (Zwicker) with recency effect, the compensation procedure for a system's characteristics, and the cognitive model comprising a combined symmetric and asymmetric distance with different powers of integration.

REFERENCES

- [1] J. Beerends, J. Stemmerdink, "A perceptual speech quality measure based on a psychoacoustic sound representation", J. Audio Eng. Soc., 1992, Vol. 40, pp. 963–978.
- [2] S. Voran, "Objective Speech Quality Measurement", NTIA, Feb. 1998.

- [3] J. Beerends, "Modeling Cognitive Effects that Play a Role in the Perception of Speech Quality", Workshop on Speech Quality Assessment, Nov. 10–11, Ruhr University, Bochum, Germany, 1994.
- [4] I. Cotanis, "Comparison of different objective measurement algorithms for speech quality evaluation", LCC Int. Internal Report, Sept. 1999.
- [5] S. Wang, A. Sekey, A. Gersho, "An objective measure for predicting subjective quality of speech coders", IEEE J. on Sel. Areas in Comm., Vol. 10, No. 5, June 1992.
- [6] S. R. Quackenbush, T. P. Barnwell, M. A. Clements, "Objective Measures of Speech Quality", PrenticeHall, NJ, 1988.
- [7] M. Hollier, M. Hawksford, D. Guard, "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain", IEE Proc.-Vis. Image Signal Process., Vol. 141, No. 3, June 1994.
- [8] M. Hollier, M. Hawksford, D. Guard, "Characterization of communications systems using speech-like test stimulus", J. Audio Eng. Soc., 1993, Vol. 41, pp. 1008–1021.
- [9] J. Berger, "Ein Ansatz zur instrumentalen Sprachqualitätsabschätzung im Festnetz der Deutsche Telekom", Workshop on quality assessment in speech, audio, and image communication, ITG, EURASIP, Darmstadt, Germany.
- [10] E. Zwicker, "Psychoacoustics: Facts and Models", Springer Verlag, 1990.
- [11] ITU-T P.862, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of 3.1kHz Handset Telephony (Narrow-Band) Networks and Speech Codecs", Feb. 2001.