

**ascom** Technical White Paper Series

# **Interpretation of Speech Quality Values From PACE**

## **Interpretation of Speech Quality Values From PACE**

Ascom White Paper Series Issue No. 104/01

© Copyright 2001 by Ascom AG, Carrier Products

All rights reserved

The information contained herein is for the personal use of the reader

Library of Congress Catalog-in-Publication Data

Wu, Raymond

ISBN 3-9521195-3-9

All inquiries and requests for titles of other White Papers in the Networking Series should be addressed to the publisher

### **Ascom AG**

Carrier Products

Glutz-Blotzheim-Strasse 3

CH-4503 Solothurn, Switzerland

Phone +41 32 624 21 21

Fax +41 32 624 21 43

carrierproducts@ascom.ch

www.ascom.com/qvoice

### **Limits of Liability and Disclaimer of Warranty**

The author and publisher of this White Paper have used their best efforts in preparing the material presented and make no warranty with regard to its content.

Price: US\$ 20.00

1st Edition

Printed in Switzerland

## Contents

<b>Introduction</b> .....	<b>2</b>
<b>“Subjective” and “Objective” Speech</b>	
<b>Quality Measurements</b> .....	<b>3</b>
Subjective Tests .....	3
Precision of Subjective MOS Values .....	4
Systematic Errors and Cultural Variations .....	5
Measuring Speech Quality “Objectively” With a Computer .....	7
<b>How Reproducible Are The MOS Values From an Algorithm?</b> .....	<b>9</b>
<b>What MOS Values Can be Expected From a Specific Codec?</b> .....	<b>11</b>
<b>Some Recommendations</b> .....	<b>12</b>
Keep only one digit after the decimal point of a MOS value .....	12
Do not only concentrate on the average MOS for an area .....	12
For maximal reproducibility use always the same speech sequence ..	13
Exercise care when comparing the performance of different networks .....	15
<b>References</b> .....	<b>16</b>

## Introduction

The purpose of this white paper is to give some suggestions to users of QVoice on how to interpret and use the MOS Speech Quality values given by the PACE (or any other MOS based) algorithm. In particular we want to discuss to what precision the values correspond to speech quality as perceived by listeners and how reproducible the measured values are. So for example what variation can be expected just from the random changes of external circumstances, as compared to when a statistically significant change.

Readers may also wish to read this document in conjunction with some introductory explanation on speech quality and its MOS based evaluation, for example the Ascom QVoice technical white paper series "Speech quality and its objective evaluation with PACE".

## “Subjective” and “Objective” Speech Quality Measurements

Fundamentally speech quality is something subjective. For example there is no other way to decide which of two kinds of distortions is worse, than to let people listen to the two speech sequences. Therefore speech quality is really **defined** as the average opinion of many people listening to the same speech sequence. To make such “measurements” reproducible, detailed procedures for “subjective tests” with many test parameters have been defined, e.g. by the ITU in [1].

### Subjective Tests

Typically different test conditions are involved in such tests and various speech sequences with a variety of qualities are used. Groups of people listen to the speech sequences in a special audio environment and judge the speech quality according to the following standard 5-level listening quality scale:

- **Excellent**
- **Good**
- **Fair**
- **Poor**
- **Bad**

Numerical values are associated with these levels, namely 1 for “bad” till 5 for “excellent”. Then for each speech sequence the average over the opinions of all people is taken. This average is called “Mean Opinion Score” (MOS).

To get a value that only depends on the distortions, one uses several speech sequences and sends them through the same channel. Typically, male and female speakers are used and care is taken to use sentences which are phonetically typical for a given language. The average of the scores for these speech sequences is called “condition MOS” (as it should only depend on the transmission conditions).

In an “Absolute Category Rating” (ACR) test, people listen only to the degraded sequence, not the originals. Also here we described a “Listening Only Test” (LOT), as opposed to tests where people speak with each other over a distorted line where problems like long delays will also play a role.

### Precision of Subjective MOS Values

Clearly, the more people are involved in the test, the less will their individual preferences play a role, and the more precise will the average value be. To estimate this precision, let us first look at the standard deviation (= the “typical” variation) of the judgements of different people when judging speech sequences exposed to the same degrading conditions. First note that the people taking part in the test can only choose one of the 5 quality scores, thus they can’t give intermediate judgements. This alone introduces an error of typically somewhat less than 0.5 .

Here we are only interested in the rough size of errors. From the literature we get some idea of this size. From figure 4 in [4] we see that the standard deviation for the five point scale is around 0.9 , whereas from table 2 in [3] we get a typical value of about 0.75 .

Now the average over the judgements of N people is taken so that the individual deviations are “averaged out”. More precisely, when forming the average of N independent random variables with equal standard deviation, the standard deviation  $\sigma$  of the average will be smaller by a factor of  $\sqrt{N}$ :

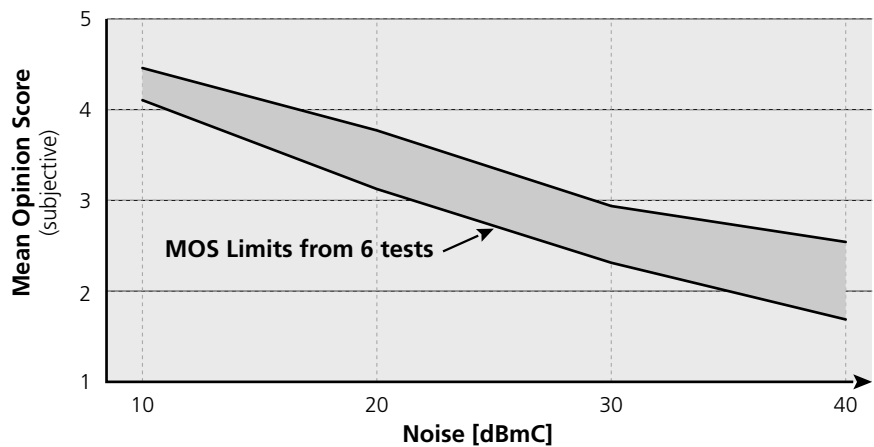
$$\sigma (\text{average}) = \sigma (\text{individual}) / \sqrt{N}$$

Thus for example a subjective test with 24 people will yield a MOS value per speech sequence that typically deviates by about  $0.8 / \sqrt{24} \approx 0.16$  from the “real” value (24 people is a typical size of such a test, see [4] before table 1). Note that because of the square root, even a subjective test with many more people will have a substantial error, so for example a test with 250 people still would have an error of about  $\Delta \text{MOS} \approx 0.8 / \sqrt{250} \approx 0.05$ .

### Systematic Errors and Cultural Variations

Even though great care is usually taken to conduct subjective tests according to a standard procedure as described in [1], there are rather large differences in the results. For example it is known that the range of qualities of the presented speech sequence influences the scores given by people. Also the different prior hearing experience (e.g. telephone conversation) of people in different countries greatly influences their judgement of audio quality.

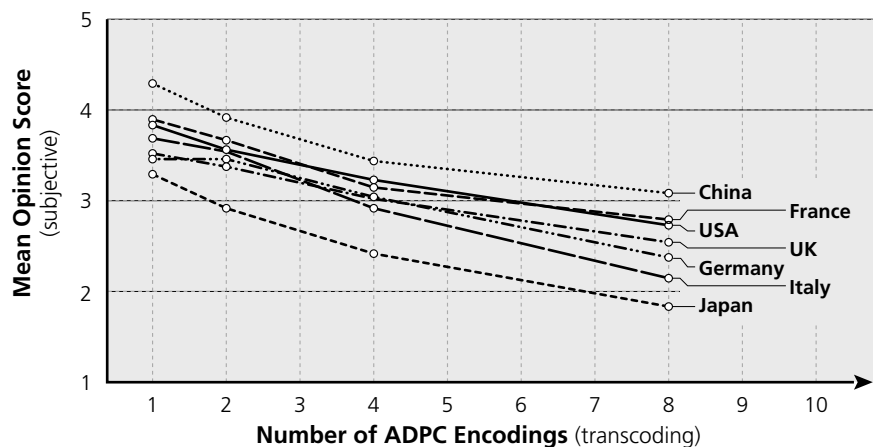
The figure below is adapted from figure 5 in [3]. It shows the range of MOS values obtained from 6 different subjective tests for different levels of added noise.



**Figure 1:**  
Range of MOS values from six subjective tests for various noise levels

Also [6] (end of page 153) states a systematic difference between subjective tests of about 0.5. The big variation of the results of tests in different countries is illustrated in the picture below, which is adapted from figure 2 in [7]. Speech sequences in the respective languages were exposed to the same degrading conditions, namely multiple encodings (transcoding) with an ADPCM (adaptive differential pulse code modulation) algorithm.

**Figure 2:**  
Results of subjective tests of the same conditions but carried out in different countries (and with speech sequences in the local language)



It is also possible that a given speech codec (coder/decoder) performs differently for different languages. Differences in languages include different syllables and different rate of occurrence of syllables, and also the role of tones (e.g. in Chinese).

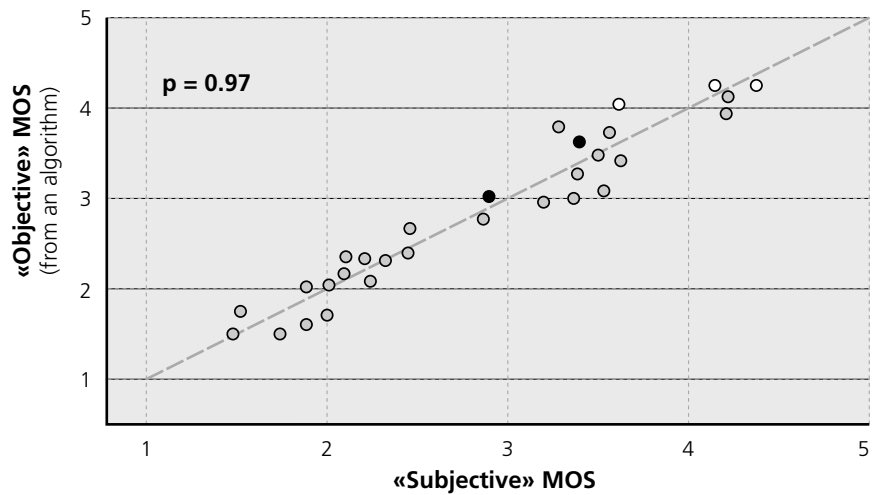
In conclusion we see that the absolute precision of MOS values as speech quality is finite. However, this does not mean that **changes** in conditions cannot be measured with better precision. Actually the uncertainties of subjective tests were one reason for developing algorithms which would give more reproducible results.

## Measuring Speech Quality "Objectively" With a Computer

Subjective tests are costly and slow, in particular they can't be used to monitor a network in real time. Therefore, computer algorithms have been developed which try to approximate the correct (subjective) MOS. As opposed to the situation in ACR subjective tests where people listen only to the degraded signal, such algorithms (e.g. PACE or PESQ) compare an original undistorted speech sequence with what's obtained by passing this sequence through a transmission mechanism, e.g. through a cellular telephone network. These algorithms are quite sophisticated because there can be many kinds of distortions and they should be weighted according to how disturbing they appear to people.

Typically such algorithms contain many free parameters which are then adjusted to give about the same results as subjective tests. In practice, there are databases of speech sequences (in pairs of original and distorted) and the results of subjective tests for these sequences. These are then used to calibrate the algorithm i.e. to adjust the parameters.

Considering that speech quality is difficult to quantify (or even define), these algorithms fulfill their functions very well. Still, users must keep in mind that the resulting values have a finite precision. The following plot, which is adapted from Fig. 4.9 of [5], plots the results of some algorithm as a function of the MOS values obtained with subjective tests for a number of conditions. One can see that the typical difference (standard deviation) is on the order of 0.25. As here, often also the correlation coefficient  $\rho$  is given, which for a given spread of the data, is related to the standard deviation.



**Figure 3:**  
Objective MOS versus subjective MOS  
for a number of conditions

In table 4.3 (page 135) of [5], somewhat larger deviations between subjective and objective MOS values are given. Also [6] (pp.154/155) indicates a standard deviation of 0.2 or more. Often the correlation coefficient  $p$  is used to indicate the agreement of objective and subjective MOS values. But note that for a spread of MOS values as in the above figure, even a correlation coefficient of  $p = 0.99$  still corresponds to a standard deviation of about 0.15.

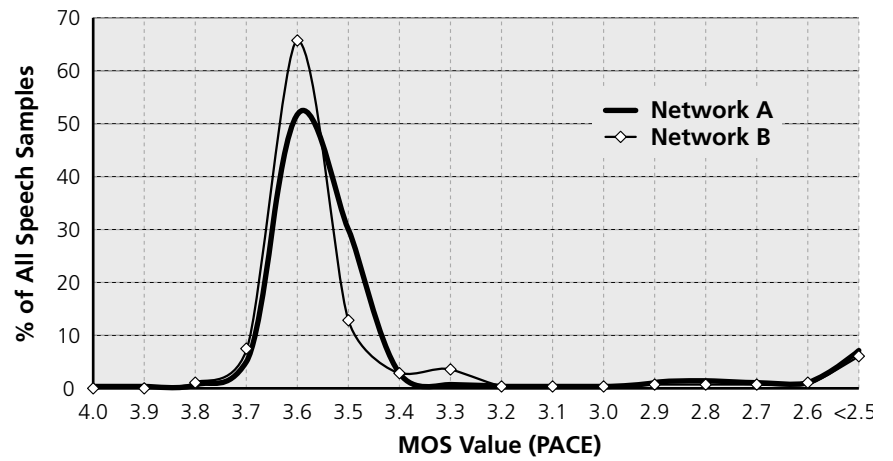
## How Reproducible Are The MOS Values From an Algorithm?

Given the same pair of original and degraded speech sequence, an algorithm will of course always give the same MOS value. On the other hand when monitoring the speech quality of a telephone network in some area, the values will naturally vary to a certain amount. This is because of random variations in conditions and more fundamentally simply because of the random nature e.g. of bit errors. For example for independent random error events, their actual number will typically vary with the square root of the average number (This is in the limit of small bit error rate, when the binomial distribution approaches a poisson distribution):

$$\text{actual} = \text{average} \pm \sqrt{\text{average}}$$

In reality the variations are rather larger because of correlations like burst errors. Randomly varying conditions include the exact number of users of a cellular phone network, interference and the precise location, as e.g. due to Raleigh fading the signal strength can vary on a short distance scale. All these variations together lead to a natural fluctuation of the measured MOS values. These errors may be considered as "statistical" as opposed to the "systematic" error discussed above which comes from the limitations of objective speech quality measurements and their calibration.

Below are some data from a measuring campaign with QVoice in a given area. From the plot one can see that the standard deviation is on the order of 0.1.



**Figure 4:** Distribution of MOS speech quality values for two cellular phone networks measured in some area. (It is, in fact, a histogram with bins of widths 0.1, but a smooth curve has been laid through them. Note that all calls with values smaller or equal to 2.5 are put together in one bin on the right)

## **What MOS Values Can be Expected From a Specific Codec?**

Depending on the speech coder/decoder (codec) there will be a maximal possible speech quality, namely when the transmission works flawlessly and so there are no bit errors. So the 60 kbit/sec PCM encoding used in most fixed telephone networks reaches at most a MOS of 4.3, while a typical 13 kbit/sec codec used in cellular phone networks reaches 3.7. Similarly, so called "enhanced full rate" and "half rate" codecs will have a somewhat higher or somewhat lower maximal MOS.

## Some Recommendations

### **Keep only one digit after the decimal point of a MOS value**

In order not to give a false impression of precision, we recommend to round MOS values to the nearest tenth, thus e.g. MOS=3.5 instead of MOS=3.465. This is because an algorithm can't be expected to measure the "real" speech quality to a higher precision and because of the natural variations of MOS measurements. Also from the above discussion of subjective tests it is clear that it takes a rather large change in MOS to be clearly noticed by individual customers.

### **Do not only concentrate on the average MOS for an area**

When summarizing measurement results, it can be misleading to concentrate on the average MOS measured e.g. in some area. Often there is only a small fraction of very bad values, which don't have much influence on the average. But it is these calls with very bad speech quality, which will disturb subscribers, while most people don't hear a big quality difference in the majority of calls, which fall into a rather narrow range of MOS values.

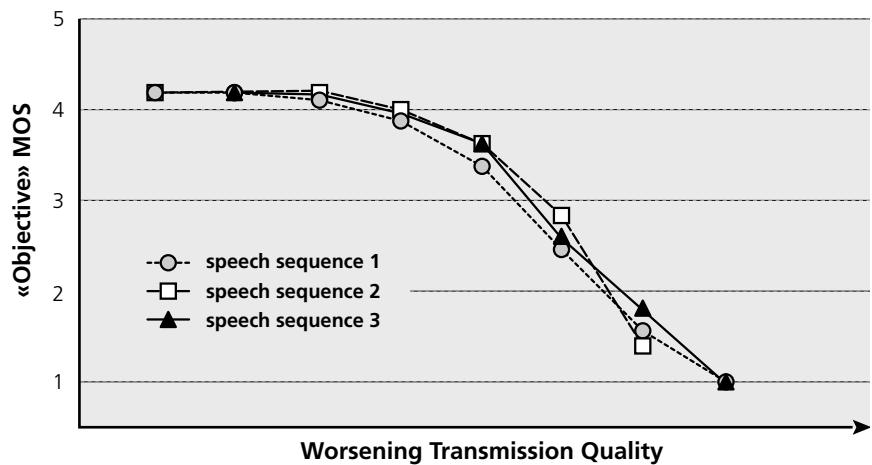
In general, cellular network tuning often takes the following priorities: ensure successful call set up and no drop calls, then improve areas with poor speech quality so that the subscribers can conduct a meaningful conversation.

So a possible way to summarize measurement results could be to give besides the fraction of successful calls and average speech quality of those calls, also the fraction of calls with speech quality below some threshold, e.g. below MOS=3.0.

**For maximal reproducibility  
use always the same  
speech sequence**

Measurement systems like QVoice often include several speech sequences (e.g. of different languages), which can be used for measurements. Experience has shown that different speech sequences will give different MOS results even under the same conditions. For example in one measurement series the German speech sequence on average gave a MOS value that was 0.2 lower than for the English sequence.

Similar differences can be seen in the figure below, which is adapted from figure 4.12 in [5]. It gives the results of an algorithm for three different speech sequences for a number of conditions.



**Figure 5:**  
MOS values from an algorithm  
for three different speech sequences  
for a number of conditions

Such differences are natural, as even subjective tests would give different results. But when, for example, monitoring the performance of a network over time, we are more interested in detecting improvements or worsening of network performance than to measure absolute values that accurately represent speech quality. So to get maximal reproducibility of measurement values, we recommend to always use the same speech sequence. For example it would make sense to use a sequence in the local language, if this is available.

**Exercise care when  
comparing the performance  
of different networks**

Of course network operators are interested in comparing the performance of competing networks. Clearly this is not so easy as there are various aspects of performance, like which area of a country is covered or speech quality comparisons. It is clear that not too much should be made of a small difference in average MOS value, like 0.1 which will anyway hardly be noticed by individual customers. On the other hand there can be significant speech quality differences like between operators, e.g. because they use speech codecs with different amounts of compression. Such a difference may well give a competitive advantage to the network with the better speech quality. Also experience has shown that especially new operators with relatively few base stations tend to keep calls or initiate calls even if the transmission quality is bad in order to get a larger covering and more customers. To see such differences between networks, it is necessary to monitor speech quality.

## References

- [1] *Methods for Subjective Determination of Transmission Quality*  
ITU-T Recommendation P.800
- [2] A.Rix, J. Beerends, M.Hollier and A.Hekstra:  
*PESQ – the new ITU standard for end-to-end speech quality assessment*  
AES 109<sup>th</sup> Convention, Los Angeles, 2000 September 22-25
- [3] William R. Daumer:  
*Subjective Evaluation of Several Efficient Speech Coders*  
IEEE Transactions on Communications, Vol. Com-30, No. 4 April 1982
- [4] A. Dehnel, H.Klaus:  
*Ein Vergleich von Urteilsskalen zur Bestimmung der Sprachqualität mit Kategorie-Einschätzungstests*  
ITG-Fachtagung Sprachkommunikation, Frankfurt am Main, 17. und 18. September 1996
- [5] J. Berger:  
*Instrumentelle Verfahren zur Sprachqualitätsschätzung*  
Shaker Verlag 1998
- [6] Markus Hauenstein:  
*Psychoakustisch motivierte Masse zur instrumentellen Sprachgütebeurteilung*  
Shaker Verlag 1997
- [7] G. Williams and H.Suyderhoud:  
*Subjective Performance Evaluation of the 32-kbit/s ADPCM Algorithm*  
IEEE Globecom '84 Conference Proceedings



Please write for titles of other White Papers in the Networking Series or for additional information and consulting services to:

**Ascom AG**

Carrier Products  
Glutz-Blotzheim-Strasse 3  
CH-4503 Solothurn, Switzerland  
Tel. +41 32 624 2121  
Fax +41 32 624 2143  
carrierproducts@ascom.ch  
www.ascom.com/qvoice