

ascom *Technical White Paper Series*

*Speech Quality and its
Objective Evaluation
with **PACE***

Speech Quality and its Objective Evaluation with PACE

Ascom White Paper Series Issue No. 103/98

© Copyright 1998 by Ascom Infrasy AG

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 3-9521195-1-2

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

Ascom Infrasy AG

P.O. Box
CH-4503 Solothurn
Switzerland
Phone +41 32 624 21 21
Fax +41 32 624 21 43
E-mail qvoice@infrasy.ascom.ch
<http://www.ascom.ch/qos>

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

Summary	2
Basics of Speech Quality Measurement	3
What is Speech Quality?	3
Why Measure Speech Quality?	4
Who Needs Speech Quality Measurements?	4
Methods for Speech Quality Evaluation	5
Overview	5
Subjective Methods	6
Comparison-Based Objective Methods	7
Other Objective Methods	8
PACE – A Closer Look at Comparison-Based Methods	9
Preprocessing	10
Time-Frequency Mapping	10
Critical-Band Filtering	11
Loudness Compression	11
Time Spreading	12
Frequency Spreading	13
Importance-Weighted Comparison	14
Transformation to MOS Value	14
Other Comparison-Based Objective Schemes	15
PACE Performance Results	16
How to Measure Performance	16
ITU-T Test Results of PACE	17
Concluding Remarks	19
References	20

Summary

This White Paper provides an introduction to speech quality evaluation. It explains the PACE algorithm for speech quality evaluation and describes the excellent results of PACE in a recent ITU-T comparison with other algorithms. The main points of the White Paper are summarized below.

- Speech quality is defined as a mean opinion score (MOS). This quantity denotes the average of a representative number of human opinions on speech quality.
- Subjective methods to assess speech quality derive MOS values from listening experiments with a carefully chosen listener panel. Objective methods, by contrast, compute MOS values from speech samples. Subjective methods are necessary for calibration, whereas objective methods enable the automation of speech quality evaluation.
- Most objective methods are based on a comparison between a reference sample and an impaired version of the reference (i.e. the sample at the listener side).
- PACE consists of three main parts. First, a pre-processing unit normalizes the reference and impaired sample. Second, an auditory model is used to transform both samples such that only perceptually relevant features are retained. Finally, an assessment unit evaluates the perceptual difference between reference and impaired sample and then outputs the result as a MOS value.
- In a recent comparison of several algorithms by the ITU-T, PACE outperformed all other tested algorithms. It was the only algorithm that showed excellent agreement with subjective MOS values in all tested conditions. Moreover, PACE was the only algorithm which successfully rated speech samples that were impacted by transmission errors.

Basics of Speech Quality Measurement

What is Speech Quality?

Speech quality is a complex psycho-acoustic phenomenon within the process of human perception. As such, it is necessarily subjective: Every person interprets speech quality in a different way. Even the perception of a single person varies with mood, interest and expectation.

Nevertheless, if speech quality is to be quantified and measured automatically, the dependence on individual opinions must be eliminated. Therefore, speech quality is generally expressed as a *Mean Opinion Score (MOS)*. This quantity denotes the average of a representative number of opinions on speech quality. MOS values are reproducible and reliable, and thus can be used to measure speech quality.

Figure 1 illustrates some of the many facets of speech quality in speech communication networks. The following definition can be found in [1]:

Mouth-to-ear speech quality (also "End-to-end speech quality") is defined as the degree of speech quality that a listener perceives at his terminal with a talker at the far end.

Speech quality is just one component of the overall (two-way) communication quality perceived by a user; it is concerned only with one-way speech transmission from a talker to a listener. In particular, speech quality ignores effects such as echoes at the talker side or transmission delays. Rather, it is affected by psychological factors such as shown in Figure 1:

- **Intelligibility:** Quality of perception of the meaning or information content of what the talker has said;
- **Naturalness:** Degree of fidelity to the talker's voice;
- **Loudness:** Absolute loudness level at the listener's side.

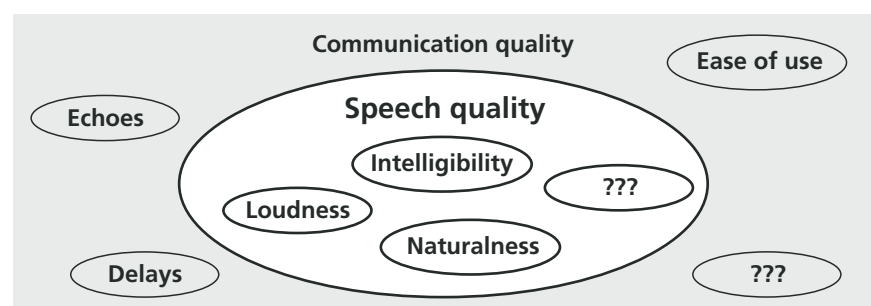


Figure 1:
Aspects of speech quality

Why Measure Speech Quality?

In both mobile and fixed speech transmission networks, there are several reasons for measuring speech quality:

- **User perception:** Users rate quality of service to a great extent by the speech quality they perceive. Therefore, speech quality is a strong indicator for customer satisfaction.
- **End-to-end measurement for any impairment:** Speech quality is measured end-to-end and thus yields a compact rating for the whole transmission link. Moreover, speech quality evaluation is – within reasonable limits – a “black-box” approach which works irrespective of the kind of impairment and the network devices causing it.
- **Reliable automated measurement:** State-of-the-art speech quality assessment algorithms (such as Ascom’s PACE algorithm) yield results very close to those of human listeners. This allows the automation of speech quality evaluation, thereby reducing costs and enabling a faster response to customer needs.

Who Needs Speech Quality Measurements?

With the liberalization of telecommunications, there is an increased interest in speech quality:

- **Network operators:** Monitoring speech quality enables problem detection (e.g. component failures) and shows possibilities for enhancements (e.g. coverage in cellular/PCS networks).
- **Service providers:** Speech quality measurements enable the comparison of different network providers based on their price/performance ratio.
- **Regulators:** Speech quality measurements provide a measurement basis in order to specify the requirements that network operators have to fulfil.

In all cases, speech quality assessment must be automated. The traditional methods for speech quality assessment based on subjective rating of speech samples are far too expensive, too slow and lack precise repeatability.

Methods for Speech Quality Evaluation

Overview

Methods for evaluating speech quality fall into two general classes:

- **Subjective methods,**
- **Objective methods.**

Subjective methods make use of a listener panel to assess speech quality. Speech quality is expressed as a *mean opinion score (MOS)*, which is the average speech quality perceived by the members of the panel.

Objective methods replace the listener panel by an algorithm to compute a MOS value from a speech sample. They aim at delivering MOS values that are as close as possible to the ratings obtained from listening experiments for arbitrarily impaired speech samples.

Both approaches have their merits, as shown in Table 1. Subjective methods are the only means to obtain real-world data about users' perception of speech quality. Objective methods, on the other hand, automate the task of speech quality assessment with moderate effort.

In practice, results from subjective listening experiments are used to calibrate objective methods for maximum agreement with subjective ratings.

Subjective Methods	Objective Methods
+ "exact" speech quality	+ good agreement with "exact" speech quality
– high effort to reproduce results	+ easily reproducible results
– no automated measurements	+ automated measurements
– high effort (time/costs)	+ moderate effort
+ applicable to any impairment	– May be sensitive to unforeseen impairments

Table 1:
Comparison of subjective vs objective methods for speech quality assessment

Subjective Methods

Although subjective assessment of speech quality requires a substantial effort, it is indispensable as a reference for objective measurement methods. This section provides a short overview on the subjective determination of speech transmission quality. More information about subjective methods can be found in [5].

The results of subjective listening-opinion tests are influenced by a wide variety of conditions, and great care must be taken to obtain reliable and reproducible results. Some of the factors to be controlled are:

- **Speech material:** Perception depends on the gender of talkers, their pronunciation, the language, length and content of samples, the recording room and equipment characteristics.
- **Experiment set-up:** Results can depend on nationality and gender of listeners, recent previous experience with listening tests, instruction of listeners about the experiment, duration of test sessions, and order of presentation of speech samples.
- **Listening conditions:** Loudness of presented speech samples and choice of equipment (headphones/telephone handsets) can influence the rating.

There are several ways to assess the subjective quality of speech samples. Some common methods are outlined in Table 2.

Subjective Methods	Objective Methods
Absolute Category Rating	Speech samples are rated without a reference, using a 5-point scale from "excellent" to "bad" (see Table 3).
Degradation Category Rating	Quality degradation of speech samples is rated relative to a reference sample of "best" quality (for the specific application), using a 5-point scale from "degradation inaudible" to "degradation very annoying".
Comparison Category Rating	Quality of speech sample is rated relative to a reference sample of better or worse quality, using a 5-point scale from "much better" to "much worse".

Table 2:
Some subjective speech quality
assessment methods

Absolute category rating can be used if there is no reference sample of known quality. On the other hand, both degradation and comparison category rating are more sensitive to small quality differences, and thus can distinguish better between speech samples of similar quality.

If a “best quality” reference for the considered application is available, the degradation category rating is the method of choice. For example, the best achievable speech quality in a GSM system is determined by the speech quality of the speech compression algorithm.

In any case, the resulting speech quality scale can be transformed – with some caveats – by appropriate statistical methods to a scale similar to the one in Table 3.

Quality of speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 3:
ITU-T scale for absolute
category rating

Comparison-Based Objective Methods

Virtually all objective speech assessment schemes currently in use fall into the category of comparison-based methods, which are akin to the comparison category rating. Generally, these schemes are based on the comparison of a known reference speech sample with an impaired version of that sample, as sketched in Figure 2.

Comparison-based methods require carefully chosen reference samples for representative and reliable results. Among other factors, the gender of talkers and the phonetical content of the samples play an important role.

To assess the speech quality of a transmission link, the reference speech sample is transmitted. Both the reference and the received (impaired) sample are processed in order to ease comparison. The comparison result is transformed to a MOS value for the impaired sample.

Among the objective methods in use today, comparison-based schemes yield, by far, the most accurate results. However, the fact that they are inherently intrusive – a reference sample must be transmitted – could be a disadvantage in some applications.

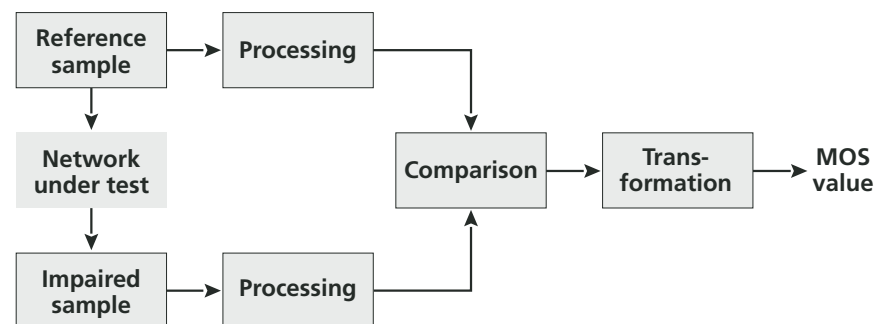


Figure 2:
Comparison-based objective methods

Other Objective Methods

Absolute estimation methods do not use a reference sample. Rather, they attempt to assess speech quality from measuring the impaired sample only. Hence, these methods are non-intrusive. However, the known schemes cannot cope with the complex non-linear impairments encountered in modern communication systems, and they are generally less accurate than comparison-based schemes. A typical example for the class of absolute estimation methods is INMD (In-service Non-intrusive Measuring Device) proposed by the ITU [4].

Transmission rating models are primarily intended for planning purposes. This class of schemes attempts to predict speech quality from knowledge about network devices and parameters. Necessarily, these schemes make simplistic assumptions about the overall effect of cascaded devices on speech quality. A typical example is the E-model ([2], [3]), where the impairments attributed to individual devices are essentially added together.

PACE – A Closer Look at Comparison-Based Methods

This section explains comparison-based methods in more detail. The discussion is focused on Ascom’s new PACE algorithm. Its structure is similar to that of most other algorithms in use and thus can be viewed as representative for its class. (This is not to say that all methods are equal – expert knowledge and experience is needed to obtain a high-performance algorithm.)

The structure of PACE is outlined in Figure 3. Basically, the algorithm can be subdivided into a preprocessing step, a component that implements a psycho-acoustic model of human perception, and an assessment unit. (Except for the details in parentheses, other objective schemes possess a similar structure.)

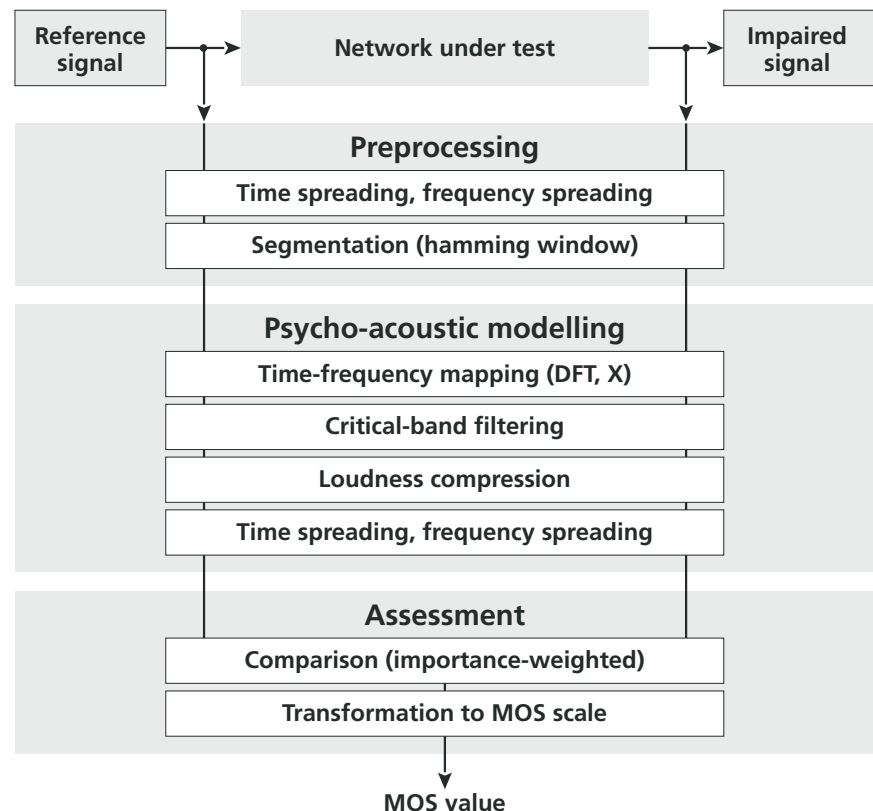


Figure 3:
Structure of PACE algorithm

Preprocessing

PACE and most other objective measurement methods in use require that reference and impaired signal be adjusted properly to each other. This includes several tasks:

- **Delay adjustment:** Since the signal is delayed in the network, it is necessary to align the impaired signal properly with the reference. This is usually done by correlating reference with impaired signal and searching for the correlation peak. Additionally, if samples are transmitted over packet-switched networks, lost packets must be padded.
- **Loudness adjustment:** Most measurement systems require that reference and impaired signal have the same loudness. This is achieved by making the average signal power equal (after removing any DC component). Loudness can also be adjusted in the perceptual domain.
- **Segmentation:** PACE divides the samples into overlapping segments to model the short-time characteristic of human perception. Each segment is passed through a Hamming window, which attenuates beginning and end of the segment. This latter operation reduces the impact of spurious frequency components, which are introduced by the segmentation process.

Time-Frequency Mapping

The human auditory system operates mainly in the frequency domain, with a non-linear relation between measured frequency (in Hz) and perceived frequency (in Bark). As a first step, each signal segment is transformed to the frequency domain using the Fourier transform. For further processing, only the magnitude of frequency components is retained because the human ear is insensitive to the phase of frequency components.

Critical-Band Filtering

The transformation from the frequency domain to the perceptual domain (from Hz to Bark) is known as critical-band filtering. The perceptual domain represents a linear scale for the human perception of frequencies. In particular, high frequencies are compressed logarithmically; for instance, frequencies f , $2f$ and $4f$ are perceived as “equally-spaced”. The mapping used in PACE is sketched in Figure 4; other schemes use slightly different mappings.

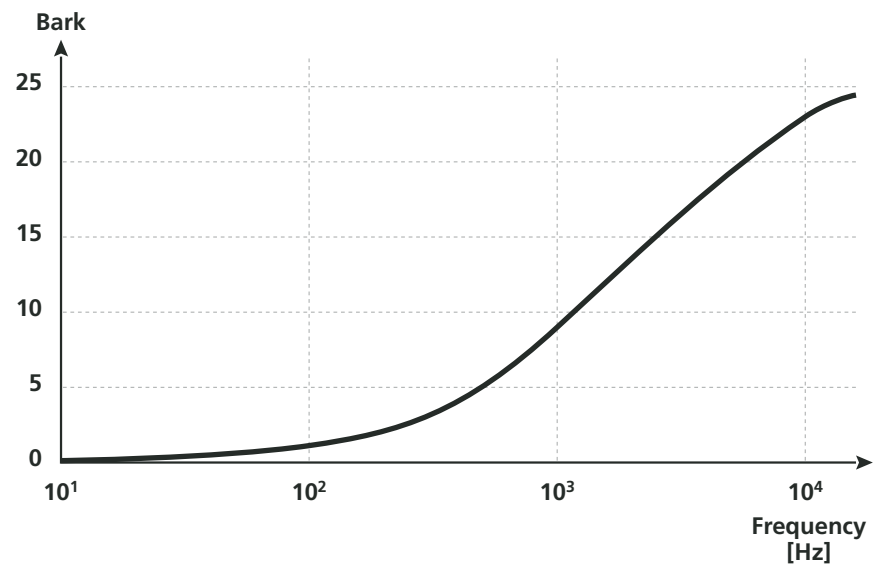


Figure 4:
Mapping from frequency domain
to perceptual domain

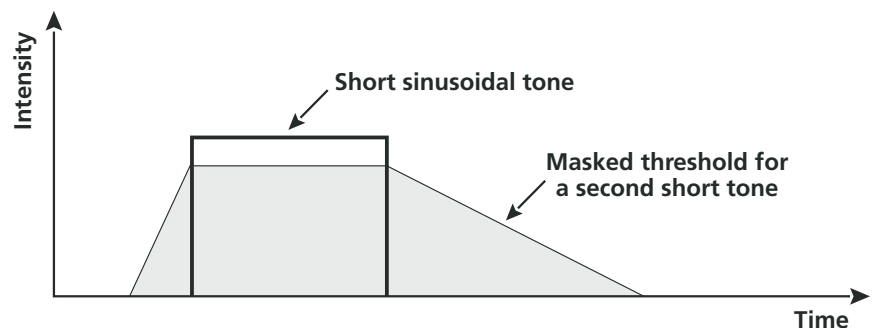
Loudness Compression

The perceived loudness of a sound is a non-linear function of the signal level. Therefore, PACE applies a compression function to obtain a linear loudness scale in the perceptual domain.

Time Spreading

The auditory system is incapable of discriminating two short pulses separated by a small time interval, as sketched in Figure 5. A coarse approximation of time-domain spreading is implicitly achieved for high frequencies by the segmentation of the speech signal because time resolution is essentially determined by the time shift from one segment to the next one. This shift is usually large enough for approximating the relatively short (<10 ms) time masking at medium to high frequencies. However, low frequencies experience much longer (>100 ms) time masking. PACE therefore uses a time-spreading algorithm to model time masking at all frequencies.

Figure 5:
Time masking

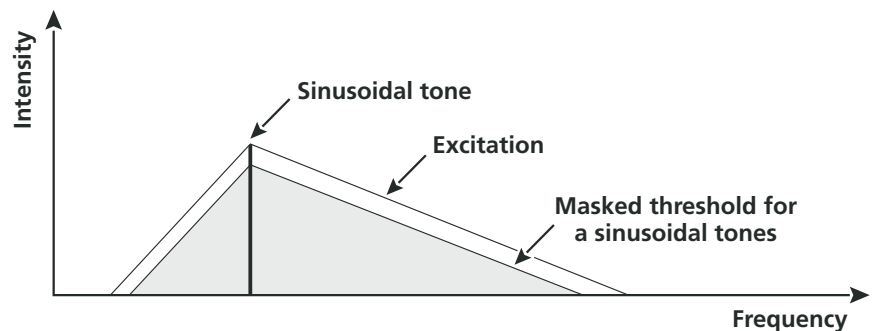


Frequency Spreading

The frequency analysis capabilities of the auditory system are not perfect, as sketched in Figure 6. If two frequency components in a sound are sufficiently close to each other, the weaker one cannot be perceived. The extent of this frequency masking effect depends on the involved frequencies, their loudness levels, amongst other factors. The slopes (the extent of exciting/masking adjacent frequencies) of the "masking triangle" depend on both centre frequency and intensity.

In its psycho-acoustic modelling part, PACE also implements frequency masking. Generally, this can be accomplished by a convolution in either the frequency domain or the perceptual domain.

Figure 6:
Frequency masking



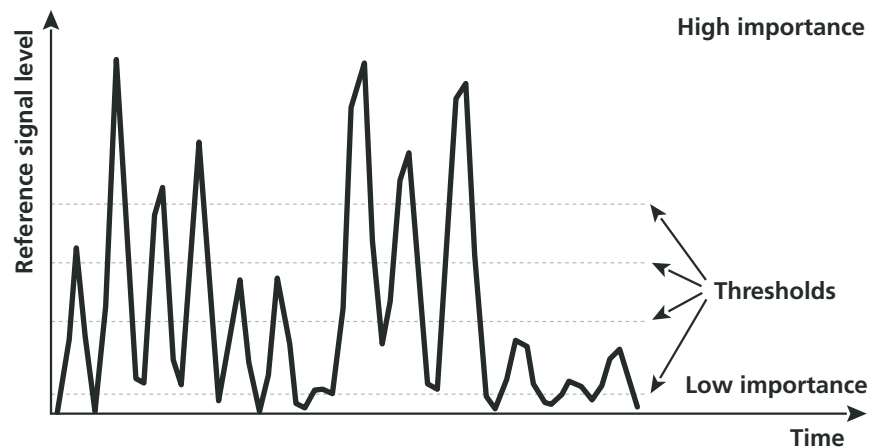
Importance-Weighted Comparison

A key element of comparison-based schemes is how they implement the comparison between reference and impaired sample. The comparison method of PACE is unique among the existing schemes. It is based on the following principle:

Signal parts with high energy are more important for the perceived speech quality

The concept is visualized in Figure 7. A similarity coefficient of reference and impaired signal is computed for four different energy thresholds. Only parts of the signal that exceed the respective threshold are considered. This can be viewed as a multi-resolution analysis with respect to signal energy. The "overall similarity" (auditory distance) is then computed using the coefficients from all thresholds; in this latter calculation, coefficients corresponding to higher energy thresholds are given more importance.

Figure 7:
Similarity calculation for different thresholds and importance weighting in PACE



Transformation to MOS Value

In general, the result of the comparison process has a non-linear relation with the desired MOS scale. PACE uses a polynomial to transform the comparison result to the ITU-T MOS scale of Table 3.

Other Comparison-Based Objective Schemes

This section briefly describes some other algorithms currently investigated within ETSI and ITU. All of them possess the general structure of Figure 3, except for the details shown in parentheses in Figure 3. Also, all of these algorithms participated in the recent ITU comparison of objective schemes (see page 17).

PSQM (Perceptual Speech Quality Measurement): PSQM was one of the first algorithms based entirely on a model of auditory perception. Originally, it was designed to assess the performance of speech codecs. It has since been revised and extended to deal with impairments encountered in networks. In contrast to most other schemes, PSQM also tries to model listening (environmental) conditions. Its comparison block attempts to find the “audible difference” between reference and impaired signal.

TOSQA (Telecommunication Objective Speech Quality Assessment): TOSQA uses dynamical control of several parameters. Speech quality is based on a similarity measurement between reference and impaired signal (which is essentially equivalent to a comparison). This step is based on modified short-term loudness spectra; it also reduces the influence of signal parts with low loudness.

MNB (Measuring Normalizing Blocks): The specific feature of MNB is a multi-resolution analysis in the frequency domain: After evaluating the difference between reference and impaired signal in a broad frequency band, the difference is removed and the analysis is repeated with narrower frequency bands. Speech quality is essentially determined from a linear combination of the various frequency band differences.

PAMS (Perceptual Analysis/Masurement System): This scheme provides a measurement of listening effort in addition to speech quality. It contains similar building blocks as the other schemes, but uses a perceptual filterbank to implement time-frequency mapping, critical-band filtering and frequency masking. It is claimed that this yields better temporal resolution in the signal analysis.

PACE Performance Results

How to Measure Performance

The performance of a speech quality measurement algorithm is evaluated using samples of known subjective speech quality, which is obtained from subjective listening tests. For these samples, the objective MOS value from the algorithm is compared to the subjective MOS value. The degree of agreement between subjective and objective MOS over all speech samples is expressed as a correlation coefficient: A correlation value of one implies complete agreement, whereas a value of zero indicates that the algorithm cannot predict speech quality at all.

In order to obtain a clear picture of the power of an algorithm, the speech samples used to test it should represent the intended range of applications and speech qualities. Additionally, the samples must not be used in the design of the algorithm. For speech communication networks, at least the following impairments should be tested:

- **Tandeming and transcoding:** Often, a speech signal is decoded and re-encoded several times on its way from talker to listener. For example, when a GSM user talks to a DECT user, the speech signal passes through at least two different speech codecs. Generally, every re-encoding impacts speech quality to some degree.
- **Background noise:** In a typical communication situation, a certain amount of environmental noise (vehicles, background conversation, etc.) is transmitted together with the speech signal. Therefore, speech samples with this type of impairment are routinely used to assess network performance.
- **Transmission errors:** In many cases, data arriving at the listener side is erroneous or incomplete. Typical examples are bit errors on GSM links and lost cells in ATM networks. Such errors can impact speech quality severely and in rather unpredictable ways.

For broad applicability of the results, a test set-up must also take into account many other factors: Test speech samples should cover several languages, be phonetically balanced and contain male and female speakers. Moreover, the impaired samples should cover a wide range of speech qualities and impairment sources.

ITU-T has compiled a collection of impaired speech samples along with their subjective MOS values. Using this data base has the additional benefit of enabling performance comparisons between different objective measurement algorithms.

ITU-T Test Results of PACE

In ITU-T Study Group 12, Question 13, different algorithms are under consideration for a new recommendation on objective measuring methods. During a Rapporteurs' Meeting from 9-11 September 1998 in Martlesham, UK, test results based on the ITU-T (Supplement 23) test data were presented. Table 4 shows correlation coefficients of various algorithms – including the ones described in page 15 – with subjective MOS values.

The following comments can be made on the test results:

- PACE [6] [7] was the only algorithm that showed excellent performance in all experiments. Just a few years ago, correlations of 0.9 could not even be achieved for a single application.
- In most cases, PACE outperformed all other algorithms. In the other cases, its performance was very close to that of the best algorithm.
- PACE was the only algorithm that accurately predicted speech quality in the presence of transmission errors (experiment 3, see page 18).
- PACE performed equally well for all tested languages. All other algorithms were less robust and exhibited substantial variations among the different languages.

Experiment	Language	C #1	C #2	C #3	C #4	C #5	PACE
Experiment 1	French	0.94	0.92	0.82	0.93	0.96	0.94
Experiment 1	Japanese	0.94	0.86	0.87	0.94	0.93	0.94
Experiment 1	English	0.96	0.83	0.89	0.92	0.94	0.97
Experiment 2	French	0.95		0.37		0.94	0.97
Experiment 2	Japanese	0.93	0.96	0.41		0.91	0.95
Experiment 2	English	0.94	0.96	0.47		0.90	0.97
Experiment 3	French	0.74		0.74	0.74	0.75	0.94
Experiment 3	Italian	0.95		0.71	0.86	0.46	0.94
Experiment 3	Japanese	0.91		0.85	0.88	0.55	0.92
Experiment 3	English	0.82		0.85	0.82	0.65	0.93

Table 4:
Performance (correlation coefficients)
of algorithms assessing ITU speech database
(C #1 = Candidate no. 1)

The performance of the algorithms was assessed in three experiments, each being performed with real-world speech samples in French, Japanese, American English and Italian (experiment 3 only):

- **Experiment 1** tested a variety of tandeming conditions. In particular, it examined the ability of the algorithms to assess multiple G.729 encodings, tandeming G.729 with ITU-T speech coding standards G.726 and G.728, and interoperability of G.729 with regional standards for codecs deployed in digital mobile radio systems. Other test samples included multiple and cross tandems of G.711, G.726, G.728, G.729, IS-54, GSM-FR and JDC-HR, MNRU.
- **Experiment 2** evaluated performance under conditions of environmental background noise, including vehicle noise, office babble, street noise, room noise, and background music. Test samples included no encoding, and G.726 and G.729 encoding, under conditions of clear channel and background noise (office, vehicle, street, white and music), and MNRU under clear channel conditions.
- **Experiment 3** evaluated conditions where the communications channel was degraded by errors. Of particular interest were random and burst frame erasure (FER) conditions and conditions in which the channel provided error concealment techniques to protect some bits in the encoded stream, but provided no such protection for other bits. Multiple transcodings of G.729 under conditions of clear channel and background noise (vehicle, street, both) under error-free channels and channels with errors (random bit errors at 1, 3, 5 and 10 percent, random and bursty frame erasures at 3 and 5 percent) were used as test samples.

Concluding Remarks

This White Paper can give only a short overview of speech quality evaluation. Nevertheless, it should leave the reader with an impression of the power and versatility that objective evaluation methods have reached today. Progress has been fast and will likely keep its pace as new challenges for speech quality evaluation arise.

The PACE algorithm, which outperformed all other candidates in a recent ITU-T test, is now being introduced in the QVoice and QNet product line from Ascom. The test results indicate that PACE is robust and well suited for end-to-end measurements in both mobile and fixed networks.

Another point is worth mentioning. MOS values provide a global view on speech quality ("good", "bad"), but do not identify the source of an impairment. Special effects such as robotic voice, ping pong and echo, however, can easily be identified by listening to the impaired sample. In QVoice, the human listener is replaced by an algorithm that analyzes the speech sample and detects these effects. In many cases, this additional information is of great diagnostic value and usefully complements the PACE output.

References

- [1] ETSI D EG/STQ-00 001: Specification and Measurement of Speech Transmission Quality; Part 1: Introduction to Objective Comparison Measurement Methods for One-Way Speech Quality across Networks, 1998.
- [2] ETR 250 (July 1996): Transmission and Multiplexing (TM): Speech communication quality from mouth to ear for 3.1 kHz handset telephony across networks.
- [3] ITU-T G.107: E-Model Recommendation (to be published); see ITU-T SG12 Report 8 of the Meeting 18 - 25 Feb 1998 (Responsibility: ITU-T SG12 WP2 Q20).
- [4] ITU-T P.561: In-Service, Non-Intrusive Measurement Device – Voice Service Measurements. 02/96.
- [5] ITU-T P.800: Methods for subjective determination of transmission quality. 08/96.
- [6] ITU-T COM 12-62: Results of Processing ITU Speech Database Supplement 23 with the End-to-End Quality Assessment Algorithm 'PACE', 09/98.
- [7] P. Juric: An Objective Speech Quality Measurement in the QVoice. Proc. of IEEE 5th International Workshop on Systems, Signals and Image Processing IWSSIP'98, pp. 156-163.

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

Ascom Infrasy AG

P.O. Box

CH-4503 Solothurn

Switzerland

Phone +41 32 624 21 21

Fax +41 32 624 21 43

E-mail qvoice@infrasy.ascom.ch