České vysoké učení technické v Praze, Fakulta elektrotechnická

Czech Technical University in Prague, Faculty of Electrical Engineering

Ing. Jan Holub, Ph.D.

Metody hodnocení kvality přenosu řeči – jejich současný stav a možnosti vývoje

Speech Transmission Quality Assessment Methods – Current Situation and Future Trends

Summary

The lecture describes state of the art in the area of speech transmission quality assessment. The first part is focused on the commonly used methods like subjective listening tests and standardized non-intrusive and intrusive measurement algorithms.

Two new algorithms developed during GACR 102/01/1355 project at Dept. of Measurement of FEE CTU Prague are described in the second part. The first algorithm substitutes short time Fourier transform with wavelet transform to achieve reduction of the required computational power. The second algorithm has been designed for measurements in low bit-rate networks where, on the top of general lower frequency resolution, the input speech is usually affected by heavy background noise. In this area standardized algorithms can not be used at all.

Souhrn

Přednáška popisuje současný stav poznatků v oblasti hodnocení kvality přenosu řečových signálů. Její první část je zaměřena na popis obvyklých metod, jako jsou poslechové testy a standardizované neintrusivní a intrusivní měřicí algoritmy.

Ve druhé části jsou popsány dva nové měřicí algoritmy, vyvinuté v rámci projektu GACR 102/01/1355 na katedře měření FEL CVUT v Praze. První algoritmus využívá náhrady Fourierovy transformace transformací vlnkovou pro dosažení úspory výpočetní obtížnosti, druhý je určen pro měření v sítích s nízkými přenosovými rychlostmi se vstupními signály zatíženými šumem, kde nemohou být použity standardní algoritmy pro měření kvality přenosu řeči.

Klíčová slova: kvalita přenosu řeči, detekce řečového signálu, MOS

Keywords: speech transmission quality, speech detection, MOS

České vysoké učení technické v Praze
Název: Metody hodnocení kvality přenosu řeči – jejich současný stav a možnosti vývoje
Autor: Ing. Jan Holub, Ph.D.
Počet stran: 19
Náklad: 150 výtisků

ISBN XX-XX-XXXXX-X

Contents

1 Introduction

Voice communication over long distances has become one of the most elementary attributes of our modern culture. In this context, speech processing and data reduction algorithms have become indispensable features allowing optimal usage of transmission media and increase the number of simultaneously transmittable conversations. Modern telephone communication networks provide a wide range of voice services using many transmission systems. In particular, the rapid development of digital technologies in the area of mobile telephony has led to an increased need for evaluating and optimizing the transmission characteristics of the devices involved. Hence, the area of speech transmission quality evaluation both by subjective and objective methods has undergone a rapid and continuous development during the past two decades.
Despite the fact that certain algorithms for speech transmission quality evaluation are approved by international standardizing bodies (ITU-T, ANSI, ETSI) for professional use, the algorithms suffer from insufficiencies that block their applicability e.g. for new transmission technologies. Due to this fact, new algorithms are being developed.

2 Methods for Speech Transmission Quality Measurements

2.1 Listening Tests

Listening and conversational tests have been standardized as the methods for subjective determination of transmission quality. These tests relate real world distortions created in a laboratory environment to the subjectively perceived quality. E.g. recommendation ITU-T P.800 [1] describes approved methods which are considered to be suitable for determining how satisfactory given telephone connections may be expected to perform. They contain recommended subjective evaluation procedures for conversational and listening-only tests. The parameter MOS (mean opinion score), ranging from 1 (worst quality) to 5 (best quality) is introduced there.

2.2 Intrusive Objective Measurements

Intrusive measurements of speech transmission quality usually require special test calls generated by the measurement system and require that the original (non-distorted) speech sample is available to the measurement algorithm. The algorithm itself then compares original and transmitted speech samples and identifies and integrates the perceptual differences between them. Known psycho-acoustical aspects of human hearing (human ear loudness and frequency resolution and sensitivity, temporal and frequency masking, etc.) are/should be modeled by the algorithm to estimate the subjectively perceived quality in terms of the MOS value as would have been obtained in a listening tests. Typical examples of an intrusive algorithm are PSQM [3], [4] and latest PESQ [5],[10],[11]. The correlation coefficient between the PESQ MOS estimate and the related MOS from a formal listening tests is in most cases above 0.9. PESQ was validated for various transmission and coding technologies including mobile networks and Voice over Internet

Protocol (VoIP) transmissions. The typical length of the analyzed speech samples is 8-12 s. The comparison between listening tests and objective intrusive approach is depicted on Fig. 1.

2.3 Non-Intrusive Objective Measurements

A typical example of non-intrusive measurement is INMD's (In-service, non-intrusive monitoring devices), as specified in ITU-T rec. P.561 [6], typically use a non-intrusive approach, allowing to monitor large amounts of live traffic. Impairment-related parameters like echo, signal to psophometrically weighted noise ratio etc. are collected from the voice channel (the recommended length of the speech sample in this case is 20s of active speech per direction) and combined by means of a suitable algorithm (P.562 [7]), to a final estimation of speech quality (CCI – Call Clarity Index).
Another ITU-T non-intrusive speech quality analysis approach is known under the working title P.SEAM [9]. It has recently been approved by ITU-T after a selection procedure where two algorithms were compared [8], [9]. The P.SEAM combines three non-intrusive algorithms and achieves a correlation coefficient with listening tests of around 0.8.

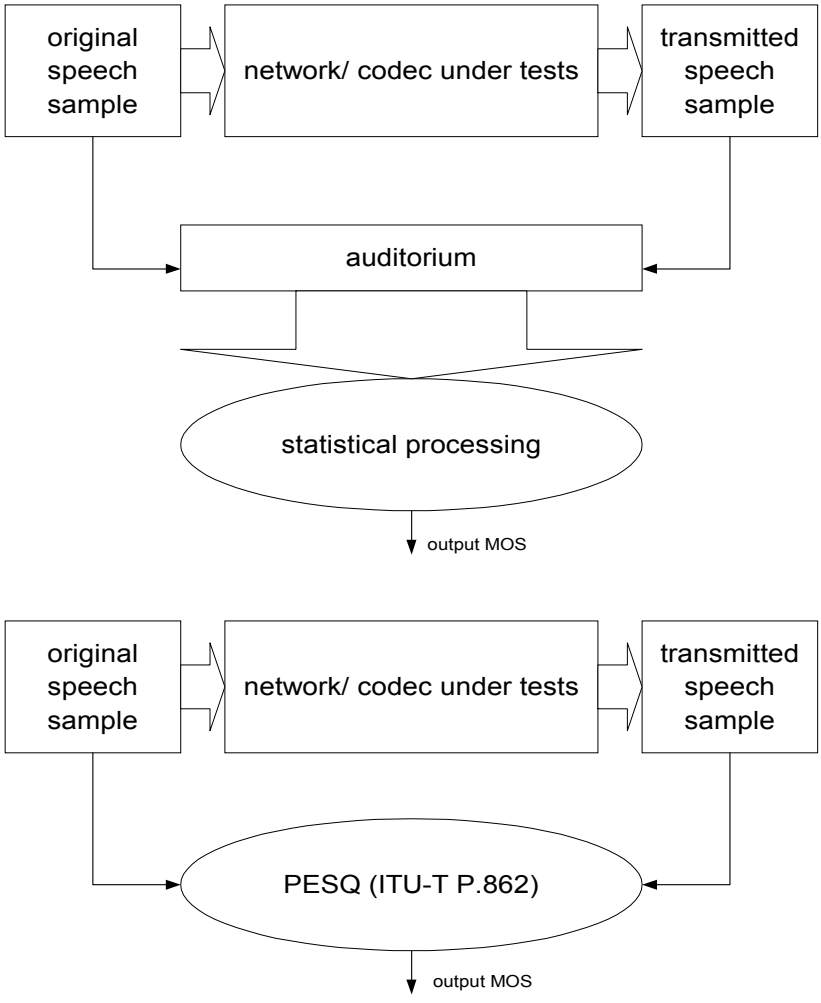


Fig. 1 Subjective (top) and objective (bottom) speech transmission quality measurements

3. Wavelet Transform Application for Speech Transmission Quality Measurements

In all the standardized speech transmission quality evaluating algorithms, the spectrum analysis is based on Fourier Transform (FT), in particular Short Time Fourier Transform (STFT) is commonly used. Let us have a detailed look to the suitability of FT and STFT for speech signal analysis. If the frequency content of the signal vary substantially from the interval to interval as in the speech signals, the standard Fourier transform

$$F(\Omega) = \int_{-\infty}^{\infty} f(t) e^{-j\Omega t} dt \quad \leftrightarrow \quad f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\Omega) e^{j\Omega t} d\Omega \tag{1}$$

sweeps evenly over the entire time axis and hide any local anomalies of the signal. It is clearly not suitable for non-stationary signals. Confronted with this challenge, Gabor in 1946 resorted to the windowed STFT, which moves a fixed-duration window over the time function and extracts the frequency content of the signal within that interval.
The STFT positions a window *g(t)* at some point τ on the time axis and calculates the Fourier transform of the signal contained within the spread of that window,

$$F(\Omega, \tau) = \int_{-\infty}^{\infty} f(t) g(t - \tau) e^{-j\Omega t} dt \tag{2}$$

When the window *g(t)* is Gaussian, the STFT is called the Gabor transform. In speech transmission quality analysis, von Hann window is traditionally used.
The difficulty with the STFT is that the fixed-duration window *g(t)* is accompanied by a fixed frequency resolution and thus allows only a fixed time-frequency resolution. This is a consequence of the classical uncertainty principle. Each element of the resolution cell is constant for any frequency and time shift as indicated by the rectangles of fixed area and shape in the left pattern of Fig. 2.

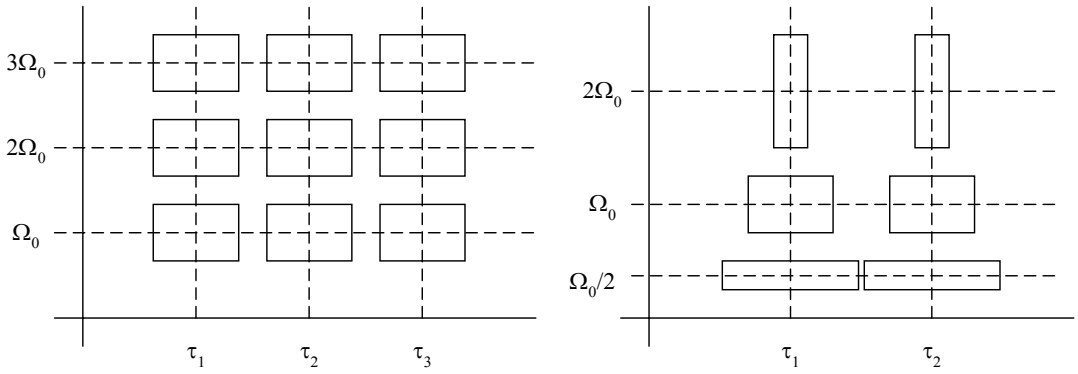


Fig.2 Short Time Fourier Transform (left) and Wavelet Transform (right) time-frequency resolution plane. In case of WT, lower frequencies are analyzed with wider time window, while higher frequencies are resolved using narrower time window

Any trade-off between time and frequency must be accepted for the entire (Ω,τ) plane. The wavelet transform, on the other hand, is founded on basis functions formed by *dilation* and *translation* of a prototype function *ψ(t)*. These basis functions are short-duration, high-frequency and long-duration, low-frequency functions. They are better suited for representing short bursts of high-frequency signals than the STFT.
This concept is suggested by the *scaling* property of Fourier transforms. If

$$\varphi(t) \leftrightarrow \Psi(\Omega)$$

constitute a Fourier transform pair, then

$$\frac{1}{\sqrt{a}}\varphi(\frac{t}{a}) \leftrightarrow \sqrt{a}\Psi(a\Omega) \tag{3}$$

where a > 0 is a continuous variable. Thus a contraction in one domain is accompanied by an expansion in the other, but in a non-uniform way over the time-frequency plane. The wavelet family is thus defined by scale and shift parameters *a,b* as

$$\varphi_{ab}(t) = \frac{1}{\sqrt{a}}\varphi\left(\frac{t-b}{a}\right) \tag{4}$$

and the Wavelet Transform is then

$$W(a,b) = \int_{-\infty}^{\infty} \varphi_{ab}(t)\, f(t)\, dt \tag{5}$$

Thus, Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT) [1] provides an alternative to the classical Short-Time Fourier Transform (STFT) for the analysis of non-stationary signals.
Proper choice of scales can set sensitivity in specific frequency domains (e.g. in Bark scale [5]). The last but not least benefit is that any of the source signals can be analyzed at once, not part by part with overlapping. This significantly contributes to computation power savings.

3.1 Algorithm design

The developed algorithm follows similar scheme like standardized approaches like P.862 (PESQ) or PAMS. It consists of the following steps:

1. Raw time alignment
2. Amplitude alignment
3. or 4.Variable delay compensation
4. or 3. Samples DWT comparison
5. Psycho-acoustic model application

The order of steps 3 and 4 may vary according to the method used as will be explained further.
The input and output speech samples are aligned based on cross-correlation of absolute values of the samples. An eventual portion at the beginning and end of each sample that does not match any part of the second one is cancelled. Both samples are of the same length at the end of this step.
Two ways of eventual delay jitter compensation have been tested. The first one requires to have DWT coefficients already calculated, see Tab. 1 and also an example at Fig. 2. In all of our experiments, we have used "dmey" wavelet that is a FIR based approximation of the Meyer Wavelet. Meyer wavelet ensures orthogonal analysis.

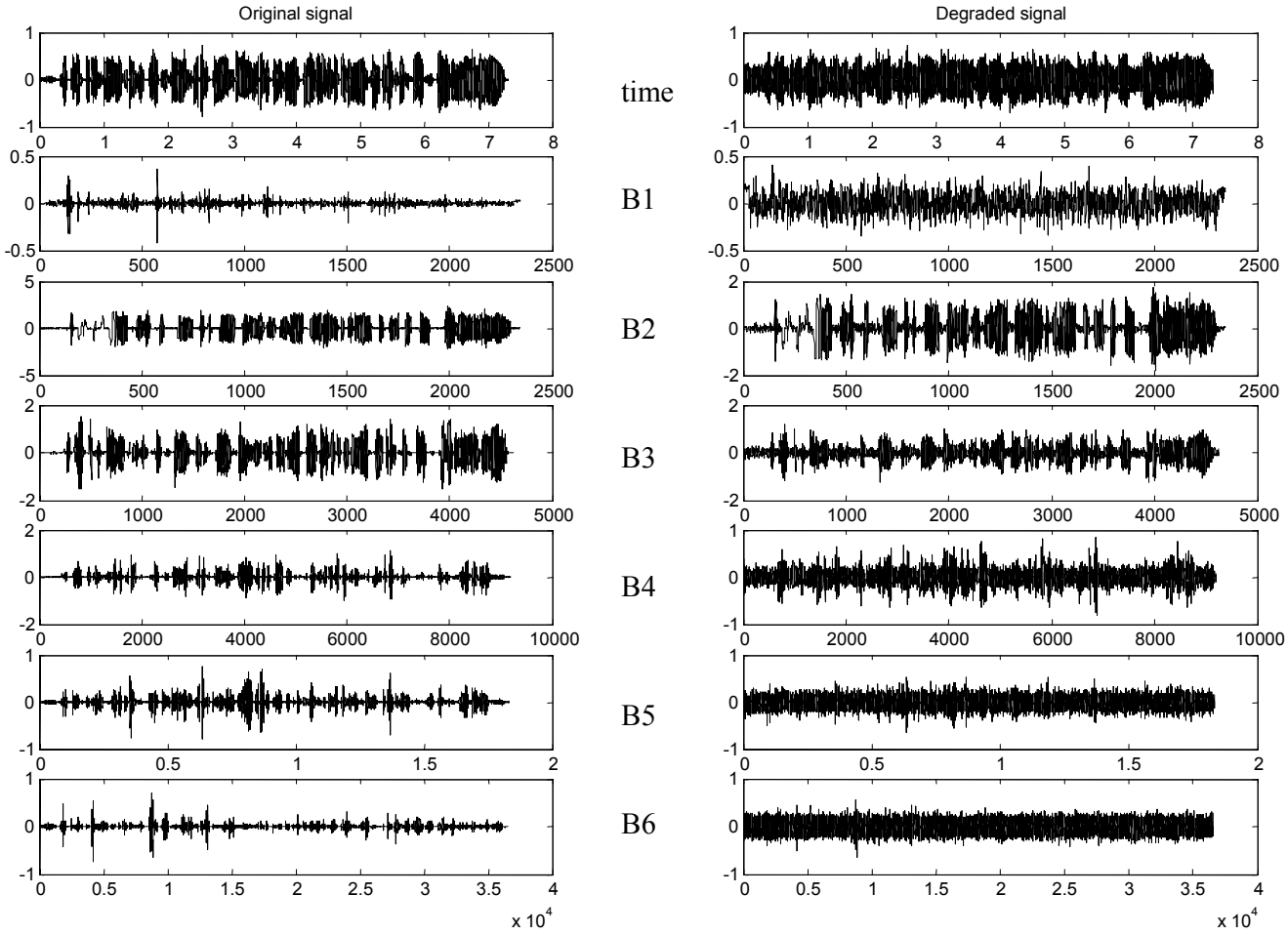


Fig. 3 DWT of the original (left) and transmitted (right) speech samples

Due to the average speech frequency occupation, the best scales for the delay examination are B3, B4, B5 that means the frequency range 250…2000 Hz.
Delay is estimated on blocks at each scale B3.. B5 separately by means of segmented cross-correlation and combined using median function applied on the corresponding time points of all the relevant scales. The fatal problem is low time resolution at lower scales caused by decimation

(see Tab. 1). Using Continuous Wavelet Transform (CWT) instead of DWT could solve this trouble but the computation would take more time in that case.
The second alternative is to compute segmented cross-correlation in the time domain (on absolute values of the samples). This is generally not so robust as previous approach but no decimation problem occurs. Another alternative would be the recursive approach as given in PESQ or PAMS.

Table 1 Scales of DWT and corresponding number of samples (Y is number of samples of the speech sample) for 8 kSa/s sampling frequency

Scale	Frequency range [Hz]	Number of Samples
B1	0...125	Y/32
B2	125...250	Y/32
B3	250...500	Y/16
B4	500...1000	Y/8
B5	1000…2000	Y/4
B6	2000…4000	Y/2

Since contemporary coders used especially in mobile networks does not necessarily keep the waveform but only spectrum amplitude information (phase information is lost), the correlation has to be performed not directly on waveforms but on their envelopes (calculated either by means of amplitude demodulation or by means of Hilbert transform).
Frame powers on corresponding scales and time shifts are compared and positive and negative differences are collected separately. Also “speech” and “silence” time periods comparisons are stored separately (that gives 2x2 sums and 2x2 counters). As voice activity detector (VAD), a simple energy threshold approach is used. Differences at different scales are weighted according to the simplified human ear sensitivity (see Table 2). The masking effect can be eventually considered at this step by highlighting the most powerful scale at each time position.

Table 2 Scales of DWT and corresponding Bark scales and averaged gains

DWT Scale	Bark Scale	Gain
B1	0-4	1e-5
B2	5-7	1e-3
B3	8-15	0.3
B4	16-27	0.9
B5	28-41	1
B6	42-55	0.8

The four calculated differences (speech-positive, speech-negative, silence-positive and silence-negative), normalized by relevant number of frames, serves as an input to psycho-acoustic model. It contains proper weighting of differences identified in silence periods against speech periods

and also positive versus negative differences. The final result is recalculated to MOS-like scale covering range 1…5. The recalculation formula has been found by least square fit to listening test results that were available for the speech samples used.

3.2 Computation Power Required

There are two independent principles that may contribute to computation savings: DWT implementation of time alignment procedure and avoiding time overlaps during Short Time Fourier Transform (STFT) as performed in contemporary standards.
The maximal computation power is saved in our approach in case that the same wavelet outputs that are further used for quality estimations are correlated without any mediation operations (like enveloping).
The second principle of saving (avoiding overlapping that is necessary for STFT procedures) is obvious. In standard methods, both original and received files are segmented into (usually) 16ms long packets with 50% overlap that are then (after windowing moreover) processed by FFT. This overlapping is necessary not to miss any short time effect that can potentially occur just on the border between two neighboring packets (in case of non-overlapped packetisation). This means that each time-domain sample is processed twice by STFT. In the wavelet-based approach no such overlapping is necessary.

3.3 Results

The final version has been tested on 132 speech samples fulfilling the P.80 requirements. Those samples were obtained partly on real transmissions in GSM networks, partly by artificial distortion (noise, amplitude and temporal clipping, echo, harmonic and non-harmonic distortion introduced by means of Matlab Toolbox, including changes in transmission delay. The listening tests of the used samples covered the MOS range 1.6 … 4.2 . The correlation between results of our calculation and listening tests were for all sample subsets (noisy samples, clipped samples etc.) higher than 0.92. The maximum absolute difference of MOS results was 0.6. The calculation according our approach took about 40% time on the same HW platform (common PC, PIII, 1GHz, 256 MB RAM).
The new developed algorithm enables implementation of speech quality evaluating algorithms to the low-power devices like PDA or even mobile terminals. It opens an area to on-line quality measurements in mobile networks (e.g. for “Quality on Demand” purposes), thus contributing to efficient usage of network resources.

4. Speech Transmission Quality Measurements for Low Bit-Rate Coded Audio Signals

In the environment of low bit-rate networks (LBRN) as used in satellite, beyond line of sight (BLOS) radio and military communications (0.6, 0.8, 1.2 and 2.4 kbit/s) the automated speech transmission quality measurements are not used at all yet.
The algorithm PESQ has been validated purely on coders using bit rates higher than 4 kbits/s as given in the Chapter IV. Moreover, PESQ requires the availability of clean, noise-free input samples, which may not be available if the speech sample is affected by input e.g. environmental noise. Blind application of the PESQ standard to signals transmitted via LBRN leads to useless results. PESQ results differ too widely from those of conventional MOS tests to consider PESQ as a meaningful quantitative assessment of LBRN quality (correlation coefficient under 0.70 for speech database as described further) [18].
In cooperation with C3A NATO, we have developed an algorithm that works reliably in LBRN networks even in cases when noise-free original sample is not available (see Fig. 4).

4.1 Algorithm design

The solution requires a noise analysis algorithm that performs an analysis of both input (= received) and output (= transmitted) speech samples. The output of the noise analysis consists of

1. Speech-to-noise ratio estimation,
2. Noise spectral ripple estimation.

Input and output samples are compared using the PESQ-like algorithm described in the Chapter 3. Finally, the noise description estimates of both input and output speech samples and their evaluation by the algorithm are fed to combining block that calculates the output MOS estimation.
To verify our algorithm, we have used a database of speech samples gathered during a project [12], [13], [14] and made available for us by NATO C3 Agency. The database contains speech samples affected by various types and levels of background noise (speech babble, field shelter, civil car, HMMWV and P4 military wheeled vehicles, Bradley and LeClerc tracked vehicles, Black Hawk helicopter, and F-16 and Mirage 2000 fixed wing aircraft). The samples are coded by several audio coders, including the currently wide spread CVSD and also future STANAG 4591 coder MELPe (Mixed Excitation with Linear Prediction / enhanced).
The noise evaluation algorithm is based on a stochastic approach. The noise is expected to be quasi-stationary, stable from a spectral point of view during the speech sample duration (8-12 s). The statistically oriented approach (performed on 16 ms packets of the speech sample) is then based on the idea that spectral noise estimation in each frequency bin (e.g. in each FFT point) can be obtained as its amplitude with the most frequent occurrence. Such an approach does not require voice activity detection (VAD) which may be unreliable for samples affected by heavy background noise.
The final combining block that generates output MOS estimation is in the described algorithm, based on the 3rd-order polynomial fit. 300 ACR MOS (Absolute Category Rating, Mean Opinion Score) listening test results were used as reference data. 30 representative samples were used to

develop the algorithm. Consequently, the remaining 270 samples have been used to validate the algorithm. The separation of training and validation samples contributes to the objectivity of the algorithm testing.

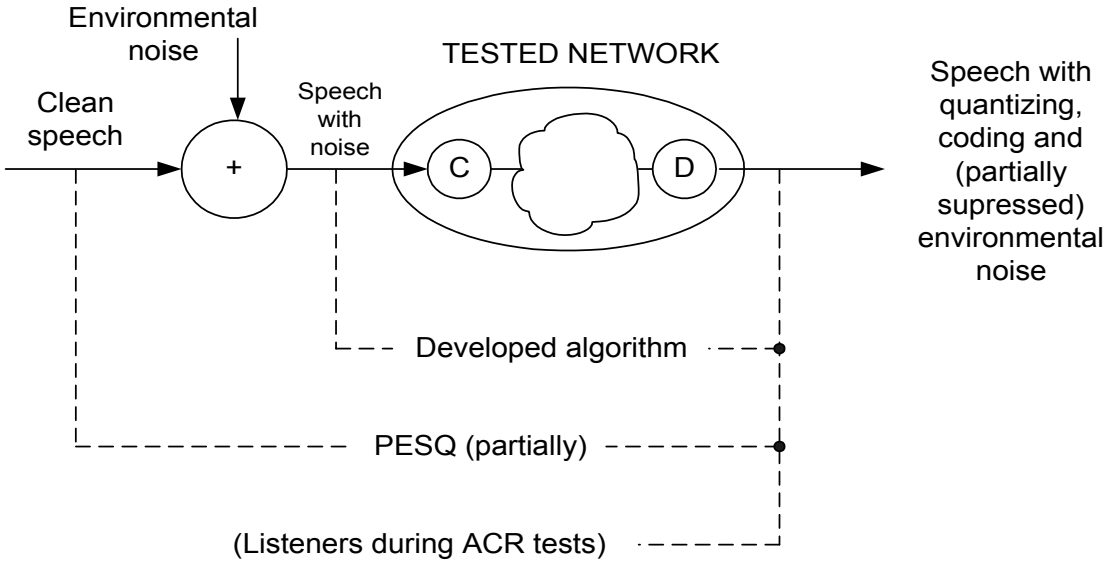
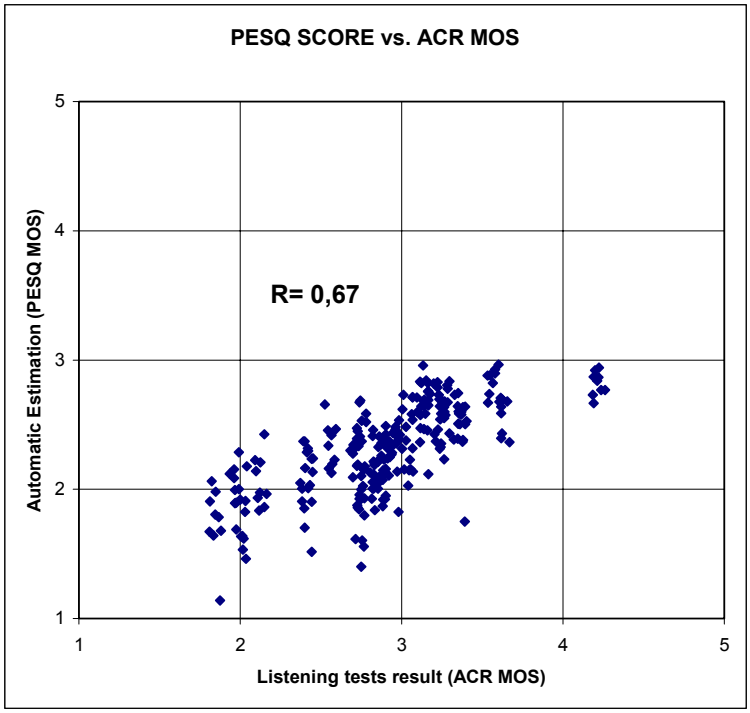


Fig. 4 Algorithm for speech transmission quality measurements in LBRN – applicability overview

4.2 Results

The correlation coefficient between ACR MOS and our estimates is 0.92 (with maximum absolute difference 0.60) on the training group of 30 samples and 0.89 (with maximum absolute difference 0.98) on the remaining 270 samples (for comparison: plain PESQ, applied to the same 300 samples, achieves correlation coefficient 0.67 and maximum absolute difference 1.67). The achieved results as described above are depicted in the Fig. 5. For comparison purposes, plain PESQ results for the same set of samples are also given.

The algorithm developed for speech transmission quality measurements in low bit-rate networks that is based on original voice activity detection approach and that can also deploy core algorithm based on Wavelet Transform as mentioned above is the only one that can be seriously and reliably used in the NATO environment where commonly available tools like PESQ generally fail.



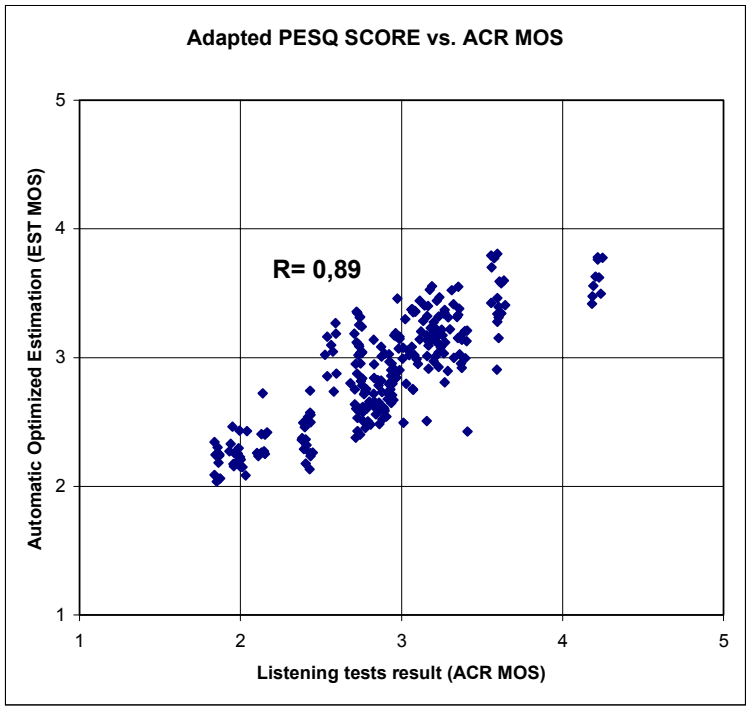


Fig. 5 Plain PESQ application to LBRN signals (top) and the modified noise- and noise-cancellation resistant version (bottom)

CONCLUSIONS

The presented set of designs shows achievements and also potential limits in selected areas of speech transmission quality measurements.
The design and verification of a new algorithmic approach based on Wavelet and the algorithm that has been developed for speech transmission quality measurements in low bit-rate networks based on original voice activity detection approach are the main achievements of the research team managed by J. Holub for the purpose of the project GACR 102/01/1355 Advanced Measurements in Mobile Networks. Further continuation of the research in cooperation with C3A NATO is expected on this topic.
The results of the presented study show that objective speech transmission quality measurements are balancing on the border between exact mathematical world of digital signal processing and the real world of human perception - that will be probably never fully described and understood. Therefore, it becomes rich and exciting research area with many simplifications and open points that are still waiting for investigation, identification and solution.

BIBLIOGRAPHY

[1] ITU-T, Methods for subjective determination of transmission quality. Series P: Telephone Transmission Quality, Recommendation P.800, ITU, Geneva, 1996.

[2] ITU-T, Subjective performance assessment of telephone-band and wide-band digital coders. Series P: Telephone Transmission Quality, Recommendation P.830, ITU, Geneva, 1996

[3] ITU-T, Rec. P. 861 "Objective quality measurement of telephone-band (300- 3400 Hz) speech codecs", Series P: Telephone Transmission Quality, , ITU, Geneva, 1996

[4] ITU-T, Study Group 12, Correlation between the PSQM and the subjective Results of ITU-T 8kbit/s 1993 Speech Codec Test. Technical Report, ITU, Geneva. Question 13/12 SQEG, 1994

[5] ITU-T Rec. P.862 "Perceptual Evaluation of Speech Quality", International Telecommunication Union, Geneva, 2001

[6] ITU-T Rec. P.561, "In-service, Non-intrusive Measurement Device – Voice Service Measurements", International Telecommunication Union, Geneva, Switzerland, 1996

[7] ITU-T Rec. P.562, "Analysis and Interpretation of INMD Voice-service Measurements", International Telecommunication Union, Geneva, 2000

[8] ITU-T Study group 12, "ANIQUE: Lucent Technologies' candidate algorithm for ITU-T single-ended speech quality assessment model", Delayed contribution D.181, Geneva, 2003

[9] ITU-T Study group 12, "Performance of the P.SEAM Collaboration's Models on Known Databases", Delayed contribution D.172, Geneva, 2003

[10] Pennock, S.: Accuracy of the Perceptual Evaluation of Speech Quality (PESQ) Algorithm, MESAQIN 2002, Praha, CTU

[11] Rix, A., Beerends, J.G., Hollier, M.P., Hekstra, A. P.: Perceptual Evaluation of Speech Quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Salt Lake City, May 2001

[12] Street, M.D., Future NATO Narrow Band Voice Coder Selection: STANAG 4591, NC3A Technical Note, 2001

[13] Street, M.D., The NATO Post-2000 Narrow Band Voice Coder, Test and Selection of STANAG 4591, NC3A Technical Presentation, 2002

[14] Street, M.D., Future NATO Narrow Band Voice Coder (Stanag 4591) Selection Process (Phase Two), NC3A Technical Note 912, 2002

[15] Dresler, T., Holub, J., Šmíd, R.: Wavelet Transform in Voice Transmission Quality Measurements In: First ISCA Tutorial & Research Workshop on Auditory Quality of Systems. Academy Mont-Cenis, Germany, April 2003, p. 35-38.

[16] Holub, J., Šmíd, R. (ed.) Measurement of Speech and Audio Quality in Networks. Praha : Czech Technical University in Prague, 2002. 100 p. ISBN 80-01-02515-2.

[17] Holub, J., Očenášek, J., Šmíd, R.: A Novel Intrusive Voice Transmission Quality Test System for Mobile Networks, IEEE 9th International Workshop on Systems, Signals and Image Processing (IWSSIP), Recent Trends in Multimedia Information Processing. London: World Scientific Publishing Company, November 2002, Manchester UK, pp. 158-162

[18] Holub, J., Street, M., Šmíd, R. : Intrusive Speech Transmission Quality Measurements for Low Bit-Rate Coded Audio Signals, Paper 5954, AES115 Convention, New York, October 2003

Ing. Jan Holub, Ph.D.

Date and place of birth: 14 June 1973, Prague, Czechoslovakia

Education:	1996 graduated at Czech Technical University, Faculty of Electrical Engineering 1999 dissertation defence at CTU FEE
Employment:	1999 Assistant Professor – Department of Measurements, CTU FEE
Research Interests:	AD and DA conversion, digital signal processing, psychoacoustics, measurements in telecommunications
Scientific activities:	organizing/programme committee memberships: Dithering in Measurement 1998, Advanced A/D and D/A Conversion Techniques and their Applications ADDA EWADC 2002, Eurosensors 2002, Measurement of Speech and Audio Quality in Networks MESAQIN 2002, 2003, 2004
Grant Support:	GAČR 102/01/1355 Advanced Measurement in Mobile Networks GAČR 102/01/D087 New Methods for Analog to Digital Converters Testing by means of Stochastic Signals

Teaching experience:

Lectures:	Selected Chapters on Instrumentation
Seminars:	Electrical Measurements (Czech and English courses) Digital Signal Processing (Czech and English courses) Digital Measuring Modules Labs in Electronic Instruments Design Advanced Methods of Signal Digitalization and Processing
Diploma supervisor:	6 sucessfully graduated students
PhD supervisor:	currently 3 PhD students – supervisor specialist
Publications:	19 journal articles and 76 conference contributions (68% international)