XVIII IMEKO WORLD CONGRESS
Metrology for a Sustainable Development
September, 17 – 22, 2006, Rio de Janeiro, Brazil

COMPUTATIONALLY EFFICIENT NON-INTRUSIVE ALGORITHM FOR SPEECH TRANSMISSION QUALITY MEASURMENT

Jan Holub [1], *Michael Street* [2], *Ondrej Tomiska* [3]

[1] Department of Measurement 13138, FEE CTU, Prague, Czech Republic, holubjan@fel.cvut.cz
[2] NATO C3A The Hague, The Netherlands, michael.street@nc3a.nato.int
[3] Department of Measurement 13138, FEE CTU, Prague, Czech Republic, tomiso1@fel.cvut.cz

Abstract: A new, wavelet-based, non-intrusive method for speech transmission quality measurements is described in the abstract. It models human perception of quality of transmitted speech signal. The deployed Discrete Wavelet Transform, in comparison with Fourier Transform, enables to reduce the computational power. In comparison with standardized methods (based on ITU-T P.563 algorithms), the described method saves about 90% of operations needed, achieving about 90% of the results of P.563. Thus, it is suitable for operational continuous assessment, or even for applications embedded in the mobile terminal.

Keywords: Speech transmission quality of service, wavelet transform

1. INTRODUCTION

The assessment of speech quality is mainly of interest for the evaluation of speech transmission systems which offer 100% or very near-to-100% speech intelligibility, because these systems cannot be distinguished by speech intelligibility measures [1], [2].

Speech transmission during any call in the telecommunication network is affected by many impairments like delay, echo, various kinds of noise, speech (de)coding distortions and artifacts, temporal and amplitude clipping etc. Each transmission impairment has a certain perceptual impact on the speech transmission quality. The overall quality can be evaluated and expressed in terms of a Mean Opinion Score (MOS) covering the range from 1 (bad) to 5 (excellent).

The following three groups of speech transmission quality measurement can be distinguished: Listening and conversational tests, intrusive objective measurements and non-intrusive objective measurements.

1.1. Listening and Conversational Tests

A trivial method of measuring quality would be to ask callers for their opinion after a call has been made. Due to obvious practical problems related to this approach, listening and conversational tests have been standardized instead 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 [3] 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.

1.2. 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. A typical example of an intrusive algorithm is PESQ [4],[5]. The correlation coefficient between the PESQ MOS estimate and the related MOS from 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.

1.3. Non-Intrusive Objective Measurements

Passive monitoring of on-going calls in the network is a basic principle of 3SQM – ITU-T P.563 [6]. The 3SQM (Single-Sided Speech Quality Measurement) combines three non-intrusive algorithms and achieves a correlation coefficient with listening tests of around 0.8.

The computational requirements of 3SQM are high – typically, for 20s speech sample the calculation on common PC (PIV, 3 GHz, 512 MB RAM), lasts another 10-15s.

2. PURPOSE

None of the above methods is suitable for operational measurements in network- or area-wide measurements when many (millions) of call records are to be processed and assessed. Principally, the most suitable candidate is the class of non-intrusive measurements (1.3) but the computational power required there is still too high. Our goal is to design a new algorithm, suitable both for streaming and sample processing that would save computational power without significant compromising the result accuracy.

3. METHODS

Based on our positive experience [7] with DWT [8,9] applied for intrusive measurements of speech transmission quality and on previous experiments with histogram of signal packet spectra [10], we have decided to combine both approaches to final non-intrusive, packet DWT-based algorithm.

The algorithm consists of the following steps:

1. Raw level alignment
2. Stream segmentation
3. DWT calculation
4. Scale power histograms generation
5. Parameter extraction
6. Perceptual synthesis of MOS estimate

3.1. Raw Level Alignment

The (PCM) data sample/stream is aligned to the level of -26dBoV. There are various solutions available; some of them require voice activity detection. To simplify the calculation, we do not detect voice activity, supposing speech activity factor above 50% and keeping the measurement window long enough (1000 ms).

3.2. Stream packetization, DWT calculation

The level-aligned stream is segmented to 16 ms packets with 50% overlap. DWT coefficients using discrete approximation (FIR-based) of Meyer wavelet are calculated at 6 scales (see Tab. 1). Meyer wavelet ensures orthogonal analysis.

Tab 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

3.3. Scale power histograms generation

Scale rms values are stored to the histogram array. Also differences in rms values from the previous packet are stored to different histogram.. In total, it gives 6 (scales) x 2 (rms and rms delta) values for each packet. For up-to 10 s speech sample processing, the histogram is filled with data from the entire sample. In case of long sample processing or in case of streaming speech signal analysis, each 1-5 s the histogram values are moved to other arrays for further processing and the original two histograms are reset.

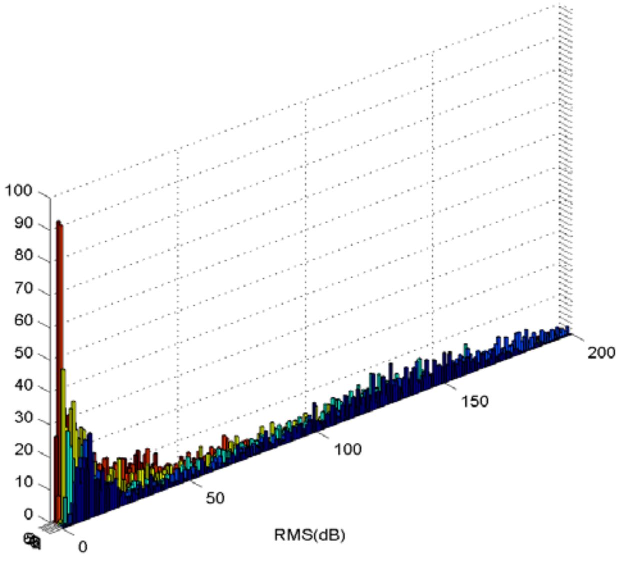

Fig. 1 Scale power histogram for clean male speech, 8kSa/s, 8s speech sample length, MOS=4.3

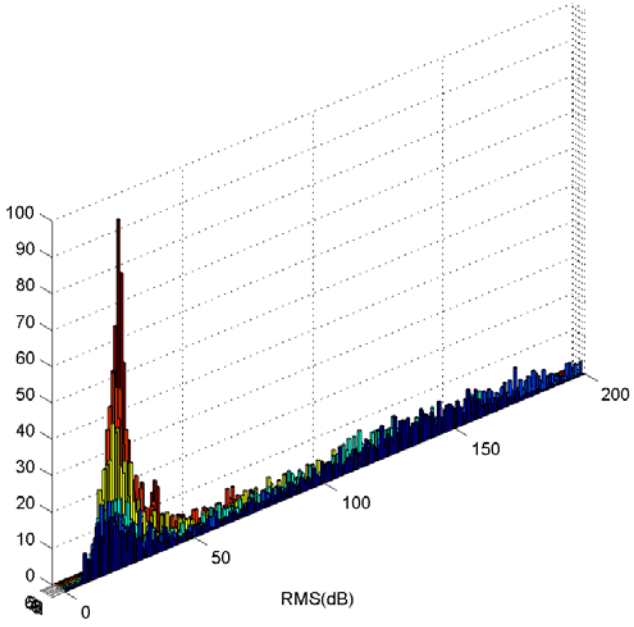

Fig. 2 Scale power histogram for noisy male speech, 8kSa/s, 8s speech sample length, SNR=8 dB, MOS=2.8

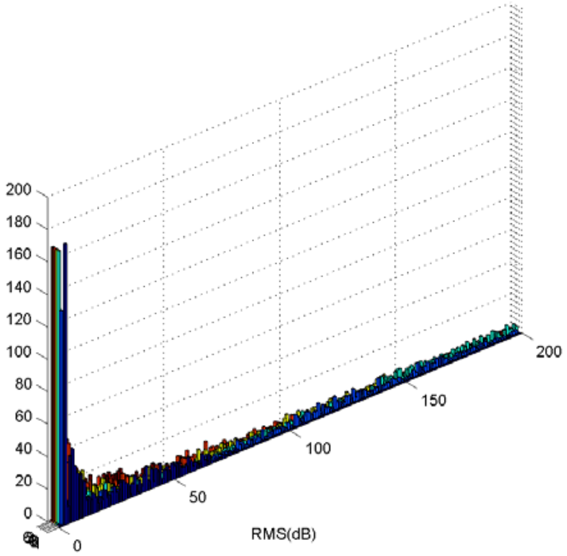

Fig. 3 Scale power histogram for encrypted GSM male speech, 8kSa/s, 8s speech sample length, MOS=2.3

3.4. Parameter extraction

Using the previously described two histograms, the following signal parameters are extracted:

For basic histogram of rms values of DWT scales:

A. Position of maxima in each scale

B. Number of hits in the 1st and 2nd histogram bin for each scale

C. Mean value of bin over aech scale

D. Variance of bin hits over each scale

For the delta histogram:

E. Variance of bin hits over each scale in delta histogram

3.5. Perceptual synthesis of final MOS estimate

Due to the average speech frequency occupation, the speech itself contributes mostly to hits in scales B3, B4, B5 that means the frequency range 250…2000 Hz.

The position of maxima in all 6 scales defines the noise profile that is, after perceptual weighting (see Tab. 2), used to calculate a basic quality score (that reflects mostly the psophometrically weighted SNR).

Tab. 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

Other parameters (namely C, D, E and comparisons with their typical values for clean speech) are used to further precise the objective quality estimation that models human perception of transmission quality achieved.

Parameter B serves as detector of presence of zero-signal portions in the PCM stream, thus indicating non-recovered jitter or temporal clipping caused by badly performing VAD (Voice Activity Detector).

4. RESULTS

The algorithm has been tested on 877 speech samples fulfilling the P.80 requirements. Those samples were obtained partly on real transmissions in GSM networks, partly by artificial distortions (noise, amplitude and temporal clipping, echo, harmonic and non-harmonic distortion). Also samples from end-to-end encrypted GSM transmissions and samples acquired in low bit rate network environment (MELPe at 1.2 and 2.4 kb/s) have been used. Most of them has been calibrated by means of listening tests (app. 80%), otherwise PESQ-LQ (P.862.1) evaluation have been used as a reference.

The results are summarized in Tab. 3.

Tab 3: Non-intrusive, wavelet based speech quality estimation – correlation coefficient for different speech databases

Data-base	Content Description	Number of Speech Samples	Corellation Coefficient
CT	Clean, noisy, GSM, jitter, clipping	665	**0,81**
GSM-E	Encrypted GSM	60	**0,77**
ME12	MELPe 1200 bit/s	76	**0,78**
ME24	MELPe 2400 bit/s	76	**0,80**

5. DISCUSSION

The commercially available non-intrusive algorithm P.563 (3SQM) achieves correlation coefficient between 0.80-0.85, thus being slightly better than the developed, wavelet-based one. However, the calculation of MOS estimate using our algorithm requires slightly less than 10% of processing power required by 3SQM. The new algorithm also enables speech stream processing, providing MOS estimate update each 1-5 s.

6. CONCLUSION

The developed algorithm seems to be suitable for operational quality measurement where massive amount of call records must be evaluated. The algorithm models human perception of quality of transmitted speech signals. Due to low computational power required by the algorithm, also applications embedded directly in mobile terminals (where the computational power as well as battery life and available memory are limited) are possible. However, due to low number of speech samples in testing databases, further verification is necessary to prove algorithm robustness and general applicability.

ACKNOWLEDGMENTS

This work has been supported by the Czech ministry of Education: MSM 6840770014 ”Research in the Area of the Prospective Information and Navigation Technologies”.

REFERENCES

[1] Preminger, J.E. and D.J. Van Tasell: Quantifying the Relation between Speech Quality and Speech Intelligibility. J. Speech Hear. Res., 38, 1995

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

[3] ITU-T Rec. P. 800 “Methods for subjective determination of transmission quality”, International Telecommunication Union, Geneva, 1996

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

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

[6] ITU-T Rec. P. 563, “Single-ended method for objective speech quality assessment in narrow-band telephony applications”, International Telecommunication Union, Geneva, 2004

[7] Dresler, T., Holub, J., Šmíd, R.: Voice Transmission Quality Measurement based on Wavelet Transform, XVII IMEKO World Congress, June 2003, Dubrovnik, Croatia, p. 233

[8] Teolis, A. *Computational Signal Processing with Wavelets*. Birkhäuser, Boston, 1998.

[9] J. Lewalle. Tutorial on continuous wavelet analysis of experimental data. Technical report, Syracuse University, April 1995.

[10] Holub, J., New Methods for Speech Transmission Quality Measurements, Habilitation Thesis, FEE CTU, Prague, 2004