AUDIO QUALITY MEASUREMENTS IN COMUNICATION SYSTEMS

Ivo Mateljan

Faculty of electrical engineering, University of Split
R. Boskovica bb, 21000 Split, Croatia (ivo.mateljan@fesb.hr)

Abstract: *ITU gives P-class recommendations for the measurement of audio quality in communication systems. The recommendations are based on system transfer function estimation, nonlinear distortion measurement and perceptual evaluation of audio signal degradation. This work shows and discusses implementation of these methods in real-time "point-to-point" testing of communication systems. A refinement of existing Fourier analyzer technique, a multitone distortion testing, and modulation quality evaluation in speech testing are presented. Modern digital communications introduce new types of signal degradation: coding distortions, time varying delays, time variance of system parameters (gain and noise floor) and even a time clipping in VoIP systems. Echo canceling systems impose restriction on selection of excitation signals in distortion testing, as well as on selection on minimal length of signal analysis window. After analysis of these factors, the work suggests use of excitation signals that are slightly different from those proposed in ITU_T recommendations. The work also presents new perceptual method for the evaluation of speech quality called MQE. It is based on speech modulation quality evaluation and can be used in real time.*

Key words: audio quality measurements, point-to-point audio testing in communications

1. INTRODUCTION

This paper will present various measurement methods for the estimation of audio quality in communication systems, with emphasize given to speech transmission systems.

The estimation of audio quality can be done with objective methods and with subjective testing of audio quality. Special methods, called perceptual methods, are objective methods, based on perceptual and cognitive hearing models that give quality rating expressed as equivalent subjective rating.

Following measurements were considered important:

- measurement of frequency response, impulse response and input/output delay ,
- nonlinear distortions measurements with sine and multitone signals,
- perceptual evaluation of speech quality.

From results of these measurements, other system characteristics can be estimated (SLR, RLR, T60, STI..).

For subjective and objective quality estimation ITU gives recommendations in ITU_T P-class documents. Following these recommendations a measurement system for "point-to-point” communication system testing has been made for the Croatian Telecommunication Agency. Emphasize is given to methods that can be realized in real-time measurement system.

In this work it was found that many of ITU proposed methods need refinement and validation. It is especially true for measurement of frequency response, where ITU gives recommendation for one method (I/O autospectrum with composite speech-noise signal excitation) but also allows other measurement methods (Fourier analyzer with random noise excitation, swept-sine measurement, MLS and crosscorrelation) [1]. After trying all these methods it was found that they give different results.

Differences in measurements were consequences of following communication system characteristics:

- automatic gain control,
- automatic noise reduction,
- echo suppressor,
- coding distortions,
- time variant system characteristics,
- speech activation.

After analysis of various methods we propose method for measuring frequency and impulse response with interrupted periodic noise. The same method is used for determination of system I/O delay.

Besides ITU_T recommendations, solutions from known measurement systems (i.e. Rhode&Schwarz, Neutrik, and Micronix) have been applied, notably the multitone signal for measurement of frequency response and total distortion [4]. Measurements with multitone signals gain more and more attention [5], but still there is no standardized test signal for measurement of total distortion. In this work we propose excitation signal for measuring the total distortion of coded speech systems.

Finally, a method for perceptual evaluation of speech quality called MQE – modulation quality evaluation [6] is presented. We offer it as replacement for ITU_T recommended method PESQ [7]. It has clear theoretical foundation and simple implementation algorithm.

2. POINT-TO-POINT TESTING

All measurements in this work refer to "point-to-point" testing of communication system. Figure 1 shows input/output connection in cases of testing classical phone system, ISDN or GSM system.

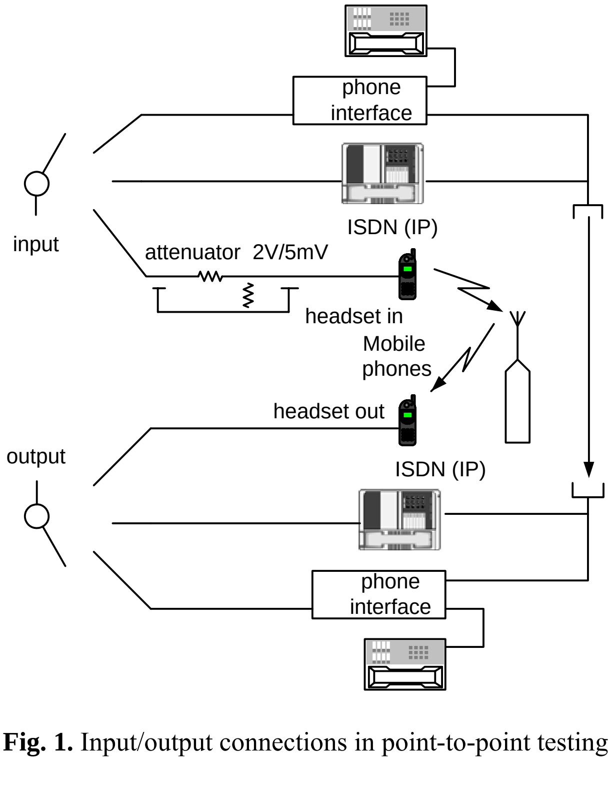

Fig. 1. Input/output connections in point-to-point testing

All measurements were done with PC soundcard and custom build software. Interface of soundcard with mobile phone is made using a phone headset input and attenuator 2V/5mV. The interface of soundcard to standard 600-Ω phone line is made with circuit shown on Figure 2.

Measurement configuration can be applied to testing VoIP system. In that case, the use of ISDN phone with dedicated audio input/output connection is recommended.

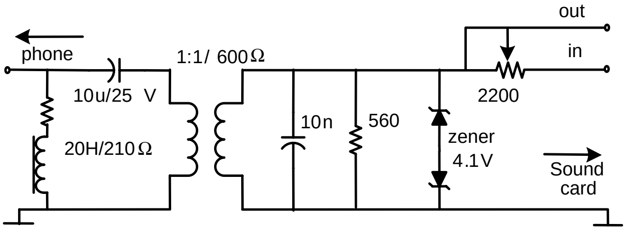

Fig. 2. Interface standard telephone line to PC soundcard

3. DISTORTION MEASUREMENTS

It is usual to express nonlinear system distortions as total harmonic distortion (THD), total harmonic distortion plus noise (THD+N) and intermodulation distortion (IMD). In first two cases excitation signal is sine; in third case excitation signal is sum of two sine signals. Distortion is expressed as percentage of square root of ratio of power of distortion power (+noise) to signal power.

For testing distortion in coded system it is common practice to use multitone signal [4], [5]. The multitone belongs to class of multisine signals, defined with following equation:

$$m(t) = \sum_{k=1}^{M} A_k \cos(2\pi f_k t + \varphi_k) \qquad (1)$$

It is composed of sum of M sine signals with amplitude A_k and predefined phases φ_k that are optimized to give the lowest crest factor. In a special case, when phase φ_k is random variable, resulting signal is periodic noise with normal amplitude distribution.

All multitone components were generated digitally with inverse DFT and sampling frequency f_s, so that each sine frequency f_k coincides with frequency of DFT bins that are Δf apart. That is way; in analyzing the response to multitone we do not need to analyze long signal sequence nor apply signal windowing to get high resolution of distortion components.

Following multitone signals were used in analysis of system response:

1. Wideband multitone - 1/3 octave spaced sine signals (from $20\Delta f$ to $fs/2$). crest factor 12 ± 1dB.
2. Speech multitone - linearly spaced sine signals from 100Hz to 500Hz, plus 1/3 octave spaced sine signals from 500Hz to 8kHz. Phases optimized for crest factor 10 ± 1dB.
3. ITU_T O.81 - 39 sine signals with frequencies spaced 100Hz (from 100 Hz to 3800Hz). Crest factor 10 ± 1dB.

For testing speech transmission channels we defined and used the "speech multitone" rather than ITU_T O.81 signal. The reason is simple: in system like GSM, that uses automatic echo suppression, multitone signal with linearly spaced components can push system to positive feedback and oscillations, as demonstrated on Fig. 3.

To quantify nonlinear distortions of multitone signal we used the total distortion measure TD+N (total distortion + noise), defined as percentage of square root of the ratio of power of distortion+noise to power of multitone signal.

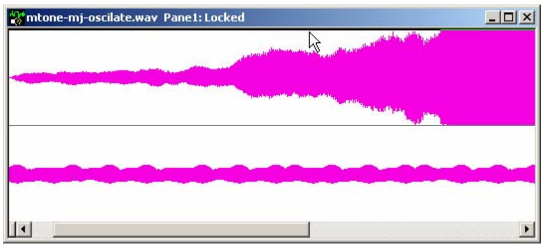

Fig. 3. Time record of oscillation build-up in a GSM system (upper trace), for the excitation with a multitone signal ITU_T O.81 (lower trace).

Figures 4 and 5 show distortion spectrum of sine signal in phone system PABX Ericsson MD-110 and GSM systems. Surprisingly, GSM has lower THD than MD-110.

Figures 6 and 7 show distortion spectrum of speech multitone signal in phone system MD-110 and in GSM system. Note that GSM system has total distortion TD+N twenty times larger than phone system MD-110.

Obviously, TD+N is a better distortion measure than THD, because subjectively speech quality in MD-110 system is much better than in GSM system.

The TD+N has not been accepted in any standard yet. The problem is that level of distortions depends on system bandwidth, so it is not easy to define TD+N measure that have equal meaning for wideband and narrowband system.

In a low-bit-rate coded system the TD+N can be higher than 100%, although the speech quality can be fair. Obviously, it will be problem to accept such a distortion measure. An alternative way to percentage measure of TD+N is a dB ratio.

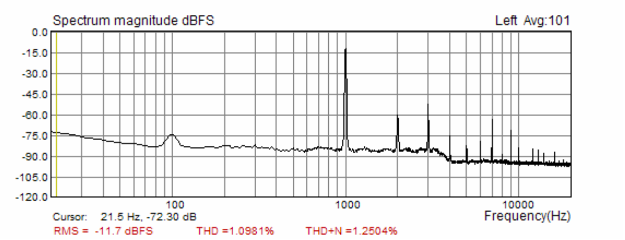

Fig. 4. The spectrum of sine signal passed through phone system PABX Ericsson MD-110. Distortions: THD = 1.090%, THD+N=1.2504%.

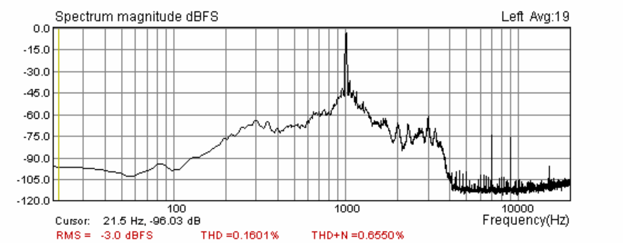

Fig. 5. Spectrum of sine signal of frequency 1000Hz passed through GSM system and two mobile phones Sony J70 and Motorola V300. Distortions: THD = 0.16%, THD+N=0.655%.

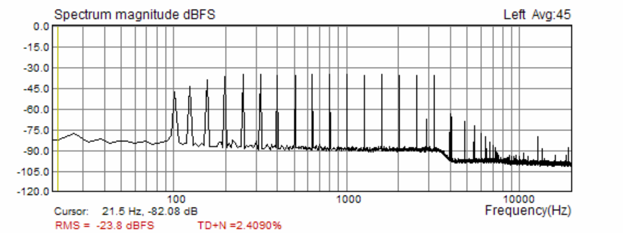

Fig. 6. The spectrum of "speech multitone" passed through phone system PABX Ericsson MD-110. Total distortions: TD+N = 2.409%.

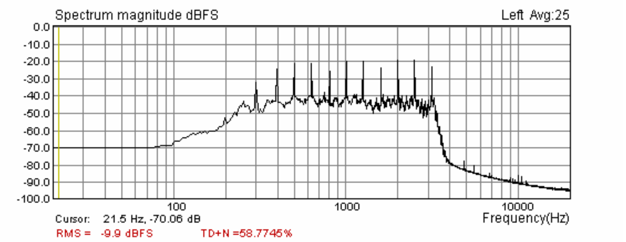

Fig. 7. The spectrum of "speech multitone" passed through GSM system and two mobile phones Sony J70 and Motorola V300. Total distortions: TD+N = 58.775%.

4. FREQUENCY RESPONSE MEASUREMENTS

Frequency response measurement can be done in various ways:

1. by swept-sine or stepped-sine response measurement,
2. by estimating the magnitude of frequency response from the multitone response.
3. by using Fourier analyzer and variety of excitation signal,
4. by using maximum-length sequence signal (MLS) and input/output crosscorrelation.

In first two cases, usually only the magnitude of frequency response is estimated.

The MLS signal is usually used for the estimation of impulse response from input/output crosscorrelation. It has been popular in acoustical measurement, but it cannot be applied in a speech activated systems.

Measurement with swept-sine or stepped-sine does not give reliable results in GSM and some other coded systems at higher frequencies [4]. Some manufacturers (Neutrik, Rhode&Schwarz) use multitone response for estimation of frequency response. But, as shown on Fig. 7, in coded systems multitone response is close to noise floor at the passband margins.

What seems to be the best choice is a Fourier analyzer method, but with necessary modification to account for speech activation and time-variant behavior. Fourier analyzer estimate frequency response and also gives the impulse response (as inverse Fourier transform of frequency response).

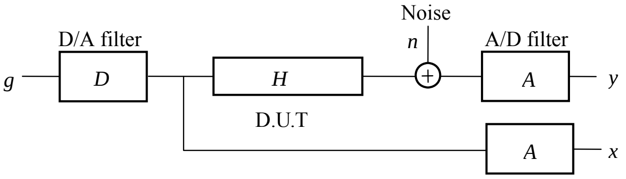

Fig. 8. Block diagram of measuring system in Fourier analyzer – *x* is input signal, *y* is output signal, *n* is noise, *h* is impulse response and *H* is frequency response

Fig. 8 shows typical measuring system. The computer generated signal *g*, after D/A filtering with transfer function *D*, is applied to the test system that has transfer function *H*. Note that *H* represent best linear fit of the possible nonlinear transfer function. The generator noise is neglected. The output from the test device, together with additive system noise *n*, is acquired by the computer as a discrete signal sequence *y*. The input to the test device is acquired by the computer as a discrete signal sequence *x*. The acquisition process implies the use of an antialiasing filter that has transfer function *A*.

4.1. Measurement with ITU CSS composite signal excitation

ITU_T recommendation P.501 defines composite signal – CSS - as shown on Fig. 9. It is periodic signal that consists of three parts. First part – speech activation signal – is speech like signal, defined in P.501, with duration 50ms. It activates and stabilizes automatic gain control. Second part is periodic noise signal of duration 200ms. Third part is a pause of duration at least 100ms.

According to P.501: "The basic idea for using such a signal is to place the device under test in a well-defined, reproducible state for the period of measurement and to secure that the transfer functions of the device do not change appreciably during the actual measurement (quasi-stationarity)".

This signal assures that communication system is in active state during periodic noise excitation and that coding algorithm for that signal always has the same characteristics (quasi-stationarity). The measurement must take place during periodic noise excitation. Then, according to P.501, magnitude of frequency response can be estimated from input and output autospectrums;

$$|H(f)|=\frac{|Y(f)|}{|X(f)|} \quad \text{(magnitude estimator)} \qquad (2)$$

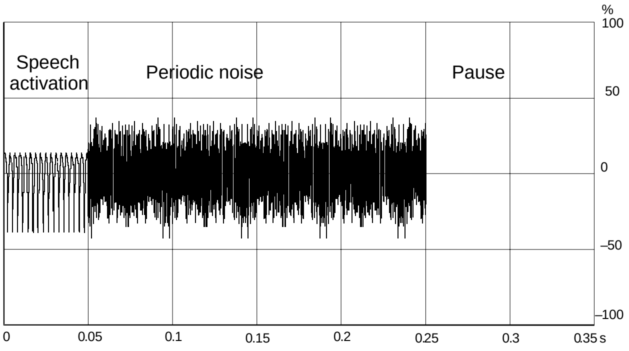

Fig. 9. ITU_T P.501 composite CSS signal

4.2. Fourier analyzer with continuous and periodic noise excitation

In a classical Fourier analyzer the excitation is a random noise and a frequency response is estimated by dividing the averaged cross-spectrum X^*Y with averaged auto-spectrum X^*X of N input and output discrete signal sequences x_i and y_i. We define the H_1 estimator as:

$$H_e(\omega)=\frac{\sum_{i=1}^{N} Y_i(f)X_i^*(f)}{\sum_{i=1}^{N} X_i(f)X_i^*(f)} \qquad (H_1\,estimator) \qquad (3)$$

where $H_e(f)$ denotes the estimated frequency response.

The H_1 estimator gives biased estimate of the real transfer function $H(f)$, which is dependant on noise, distortion and delay between input and output channel. When only noise contributes to bias, the effect of averaging can be expressed by the equation:

$$\begin{aligned} H_e(f) &\cong H(f)+\frac{\sqrt{n}\langle N_s(f)A(f)X^*(f)\rangle}{n\langle X^*(f)X(f)\rangle} \\ &\cong H(f)+\frac{1}{\sqrt{n}}\frac{\langle N_s(f)G^*(f)\rangle}{\langle G(f)G^*(f)\rangle}\frac{D^*(f)}{|D(f)|^2} \end{aligned} \qquad (4)$$

where brackets <> denote the averaged value. Note that signal term is summed coherently, while the stochastic part of the noise is power summed. The conclusion is that averaging lowers the noise level proportionally with a square root of number of averages, thus improving the measurement S/N by 10log(n).

We can have better insight in quality of measurements if we analyze the coherence function defined as:

$$\gamma^2 = \frac{Output\ power\ due\ to\ input}{Total\ output\ power} = \frac{|\langle S_{xy}(f)\rangle|^2}{\langle S_{xx}(f)\rangle \cdot \langle S_{yy}(f)\rangle} \tag{5}$$

The coherence function is a measure of the proportion of the power in output signal y that is due to linear operations on the input signal x. Maximum value of coherence is 1. When estimating transfer functions, the coherence function is a useful check on the quality of the data used.

Values of the coherence function less than one are possible if some of the following situations occur:

- Non correlated noise present
- Additional external signal source exist
- System has non linearity
- Additional inputs present in the system
- Error leakage not reduced with windowing.

To obtain a good accuracy of coherence function measurements, it is necessary to make frequency domain linear averaging.

From the definition of coherence function, the term $\gamma^2(f) \cdot S_{yy}(f)$ is the output power related to the input signal, and the term $\lfloor 1-\gamma^2(f) \rfloor \cdot S_{yy}(f)$ is the noise component of the output power, therefore, the signal to noise ratio of the system under test is so computable:

$$\frac{S}{N}(f) = \frac{\gamma^2(f)}{1-\gamma^2(f)} \tag{6}$$

Note that coherence value less then 0.5 means that noise (or distortion) is higher then measurement signal.

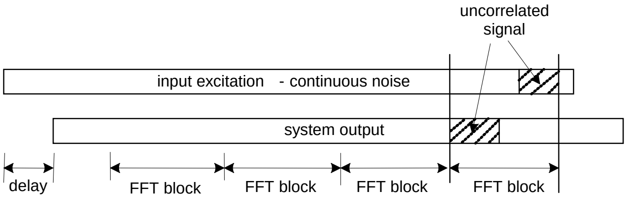

Fig. 10. Illustration of uncorrelated estimation in classical Fourier analyzer. FFT block denotes parts of input and output signal used in estimation of auto-spectrum and cross-spectrum.

In a system with large delay between input and output (see Fig. 10), i.e. when measuring response of communication systems with high delay, there will be low correlation between measured input and output signals. It is possible to delay acquisition of input channel, so this kind of error can be eliminated.

The biggest problem is in systems with voice activation and time varying coding algorithm, then, continuous noise excitation can not give reliable results. The problem can be eliminated by using the interrupted periodic noise excitation (Fig. 11), that always keeps the coded communication channel in active state. For the correct implementation of interrupted noise excitation following conditions must be met:

- Start of the acquisition must be after a preaveraging cycle that is necessary to activate system and reach the steady state response.
- After every acquired block, signal generation must be stopped, and new periodic noise sequence generated. Pause must have duration at least 100ms.
- The length of FFT block must be equal to length of the generated periodic noise sequence. This guaranties that generated and acquired signals are always correlated, so there will no bias due to the input/output delay.

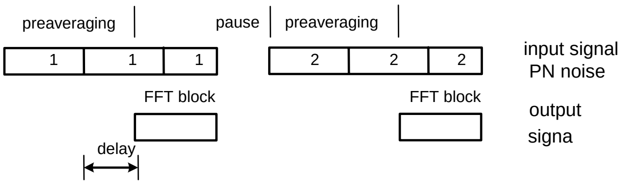

Fig. 11. Signal generation and acquisition in interrupted noise method

The excitation with interrupted periodic noise is the best choice for measurements of frequency response in communication systems that are voice activated and have time-variant signal processing (automatic gain control and noise reduction). Interrupted noise keep communication channel in "active" state, while measurements are taken in small interval of time to assure system stationarity.

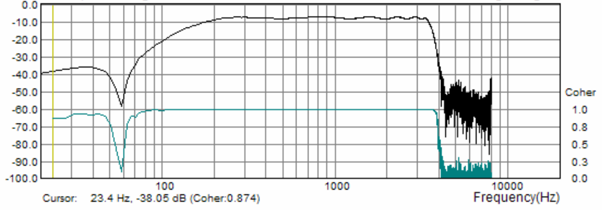

Fig. 12. Frequency response and coherence function of phone system MD-110.

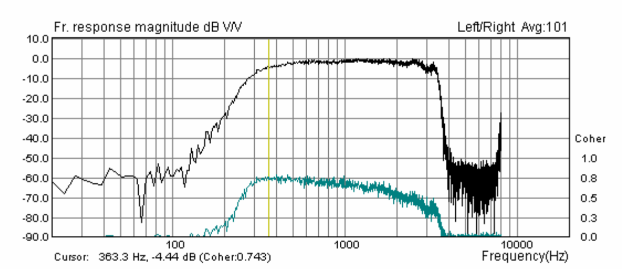

Fig. 13. Frequency response and coherence function of GSM system.

Fig. 11 shows measured frequency response of phone system MD-110. A high coherence value in system passband shows system with low distortion. In GSM system (Fig. 12) coherence function is low (0.4-0.8) because part of coding distortion is correlated with coded signal.

4.3. Fourier analyzer with swept-sine signal excitation

It was shown [3] that in acoustical measurements swept-sine excitation gives excellent results in time-varying environment. The principle of system excitation and measurement is shown on Fig. 14.

Swept sine, prefixed with speech activation signal P.501, is treated as nonperiodic signal. Time of measurement is much longer than excitation signal to account for system delay, echoes and reverberation.

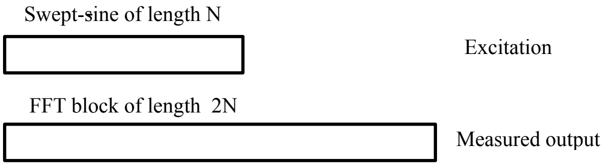

Fig. 14. Signal generation and acquisition of nonperiodic swept-sine

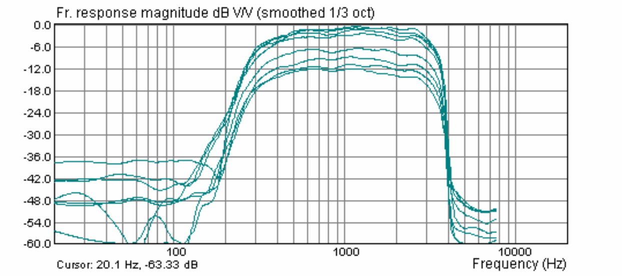

Fig. 15. Frequency response of GSM system measured with interrupted pink noise and relative input levels: 0dB, -3dB, -10dB, -12dB, -15dB, -17dB, -19dB,-20dB.

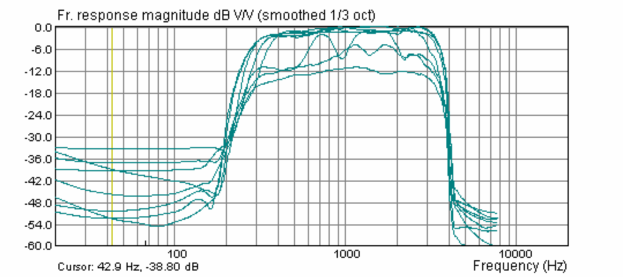

Fig. 16. Frequency response of GSM system measured with swept-sine and relative input levels: 0dB, -3dB, -10dB, -12dB, -15dB, -17dB, -19dB,-20dB.

Figures 15 and 16 illustrate effect of automatic gain control in GSM system on measurement results. Measurement with interrupted noise gives the same response pattern for different input levels, while measurement with swept-sine shows distorted frequency response. Obviously, swept-sine excitation gives bad results in communication system with automatic gain and noise reduction control.

5. IMPULSE RESPONSE AND DELAY

Methods for direct measurement of impulse response (MLS and direct impulse excitation) are not suitable for measurement of communication system impulse response.

Impulse response and input/output delay can be estimated more reliable from inverse Fourier transform of frequency response. Equation (4) shows that there will be high level of noise at frequency near fs/2 that is way; it is necessary to apply antialiasing filter to impulse response.

6. PERCEPTUAL EVALUATION OF SPEECH QUALITY

Methods for perceptual evaluation of audio quality are relatively new type of measurements in which original and degraded speech signals are compared using perceptual and cognitive models of hearing. The result is the quality rating on an equivalent subjective scale [6].

ITU proposed various perceptual methods: PSQM, MNB and PESQ [7], [8], and we have proposed the method called Modulation Quality Evaluation – MQE [9]. We believe that MQE method has clearer theoretical concept the PESQ method. It is based on paradigm that the most important speech characteristic is the modulation. Perceptual sensitivity to change of modulation is dependant on internal noise. This effect is modeled by using *just noticeable differences* [10].

Fig. 17 shows main components of a system for perceptual evaluation of speech quality. The system analyzes and compares the original and degraded speech signals in overlapped time frames of length 40ms. When testing GSM and VOIP speech transmission, proper delay estimation and frame synchronization is applied.

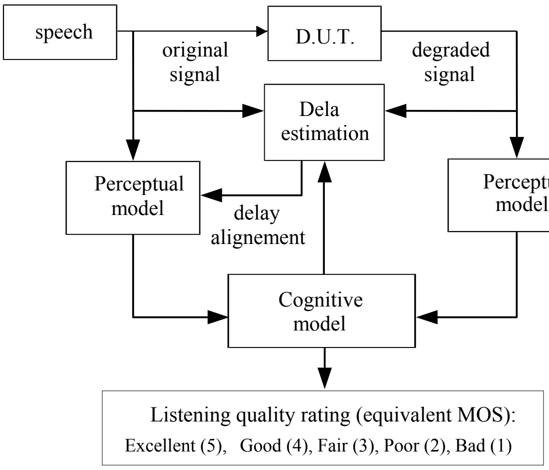

Fig. 17. The system for perceptual evaluation of speech quality

Perceptual modeling means that signal is transformed from the physical domain (sound intensity *I*) to the excitation of basilar membrane (excitation intensity *E*) and finally to the compressed neural excitation domain (loudness *N*) [10]. A cognitive modeling gives proper weight and logistic transformation to differences of original and degraded speech in order to get the listening quality rating on scale 1 to 4.5, that is as close as possible to mean opinion score (MOS) obtained from subjective listening tests [6].

Next, a description of perceptual distance measure and cognitive model of MQE method are given.

6.1 Perceptual distance measure

Perceptual distance measure, or frame distortion, is defined for *k*-th speech frame in frequency warped bark-loudness domain as *p*-norm measure:

$$D_p(k) = C\sum_{i=1}^{N} W(i)\cdot abs(S_{or,k}(i) - S_{\deg,k}(i)) \tag{7}$$

Where:

- i is index of bark band, i=1,2,..,N
- k is index of speech frame
- $W(i)$ is weighting factor for band i
- $S_{or}(i)$ is perceptual value of original speech in band i
- $S_{deg}(i)$ is perceptual value of degraded speech in band i
- C is arbitrary constant.

The total distance measure (or total distortion) is expressed as average value of all $D_p(k)$.

In a PESQ method and other ITU_T approved methods, the perceptual values are specific loudness of original speech and specific loudness of degraded speech that are normalized to the same total loudness. Main problem in PESQ is the assumption that degradation of speech quality is proportional to difference of loudness regardless of loudness value. This assumption is not correct, as it excludes the influence of internal noise.

MQE method analyzes internal hearing noise using concept of *just noticeable difference* (*JND*). A fundamental postulate of psychophysics is that all decision variables are random variables, drawn from some probability density function. From the signal detection theory premise the JND of loudness is $\Delta N_{\mathrm{JND}} = d'\sigma_N$ [11], where σ_N is standard deviation of *N*, and *d'* is discrimination constant.

Now we define signal to noise ratio in loudness domain as:

$$SNR_N(N) = \frac{N}{\sigma_N(N)} = d'\frac{N}{\Delta N_{JND}} \tag{8}$$

If we suppose that between two successive overlapped speech frames loudness change is equal to Δ*N*, then perceptually significant value of this change can be expressed as increment of the signal to noise ratio (1).

$$S = \frac{\Delta N}{\Delta N_{JND}} \tag{9}$$

It is usual to define loudness *relative* JND function *J*(*N*):

$$J(N) = \frac{\Delta N_{JND}}{N} \tag{10}$$

then:

$$S = \frac{\Delta N}{N}\frac{1}{J(N)} \tag{11}$$

First factor represents the loudness modulation; second factor is reciprocal of relative JND (proportional to SNR_N). Experimental results [11] show that *J*(*N*) is constant for SPL above 60dB. That is way; the dominant value used in perceptual distance measure is a loudness modulation.

To apply this perceptual model it is necessary to express the perceptual value *S* as a function of the excitation *E* and relative JND function $J(E) = \Delta E_{JND}/E$ that is usually called Weber fraction.

Allen and Neely [11] show that relative JND functions in intensity and loudness domain are related as:

$$J(N) = \upsilon\, J(E) \tag{12}$$

where υ is a function that is equal to the slope of log-loudness vs. log-intensity curve;

$$\upsilon = \frac{d(\log N)}{d(\log E)} = \frac{dN}{N}\frac{E}{dE} \tag{13}$$

They took an approximation that Δ*N*/Δ*E* = d*N*/d*E*, as υ is a slow varying function of *logE*. Then, substitution of (12) and (13) in (11) gives:

$$S = \frac{\Delta E}{E}\frac{1}{J(E)} \tag{14}$$

This equation shows that relevant perceptual value can be estimated in the excitation domain. We get the excitation E by summing power spectrum (intensity) in each critical band and applying outer-inner ear filter:

$$\begin{aligned} EarFilter(f) &= -0.6\cdot 3.64(\frac{f}{1000})^{-0.8} \\ &+ 6.5\exp(-0.6(\frac{f}{1000}-3.2)^2) - 0.001(\frac{f}{1000})^{3.6}(dB) \end{aligned} \tag{15}$$

MQE method uses equation (14), as perceptual value in distance measure (6), in the following form:

$$S_k = \frac{\Delta E_k}{E_k}\frac{1}{J_n(E_k)} \tag{16}$$

where: k is a frame index, ΔE_k is excitation difference between k and k-1 speech frame, E_k is average value of excitation in k and k-1 frame. $J_n(E_k)$ is normalized relative JND function ($J_n = J / J_{min}$).

To get the normalized relative JND function, experimental data of Riesz [11] are used to set the following function:

$$J_n(E) = 1 + (\frac{20000E_{th}}{E})^{1/3} \tag{17}$$

where E_{th} is the excitation at threshold of hearing corrected with outer-middle ear filter. The choice of Riecz data for tone-like signals is approved in experimental work where it is shown that most of the speech frames have high tonality factor [14].

6.2 Cognitive model

The cognitive model of MQE determines weighting factors $W(i)$ and constant C of distance measure (7), the total distortion over all frames and transform of total distortion to equivalent MOS score.

The constant C is chosen as C=0.4125. Weighting factors are chosen to account for higher sensitivity to modulation change in frequency range from 1kHz to 2kHz [10]. For center bark frequencies: 350, 450, 570, 700, 840, 1000, 1170, 1370, 1600, 1850, 2150, 2500 and 2900 Hz, weighting factors are: 0.7, 0.8, 0.9, 1.0, 1.06, 1.125, 1.125, 1.125, 1.125, 1.125, 1.125, 1.06 and 1.0.

Two types of frames are analyzed: active frames and silent frames. Active frames have energy level higher than 20dB below maximum energy, and silent frames have energy level higher than 30dB below the level of active frames. Distortions of silent frames are scaled with factor 0.2.

Total distortion is the sum of average distortion in active frames and scaled average distortion of silent frames. The quality score, called MQEscore, is defined as:

$$MQE_{scor}e = 4.5 - \text{total\_distortion} \tag{18}$$

The same way is defined the PESQ score, to give the maximum quality score equal to 4.5.

Finally, the equivalent MOS score is calculated using the logistic function that is similar to PESQ logistic function [8]:

$$\text{MQE}_{\text{MOS}} = 1 + \frac{4}{1+\exp(-A\cdot MQE_{score}+B)} \tag{19}$$

where constants A and B are chosen as: A=1.35, B=4.1291. The choice of this logistic function is quite arbitrary, just to get the equivalent MOS as close as possible to PESQ method.

Fig. 18 shows equivalent MOS for noise modulated speech degradation of male voice. Pearson correlation coefficients results obtained with PESQ and MQE are high: r(PESQ,MQE)=0,9929. The similar results are for female voice. The speech modulation with noise is generated by ITU MNRU method - Modulated noise reference unit [13].

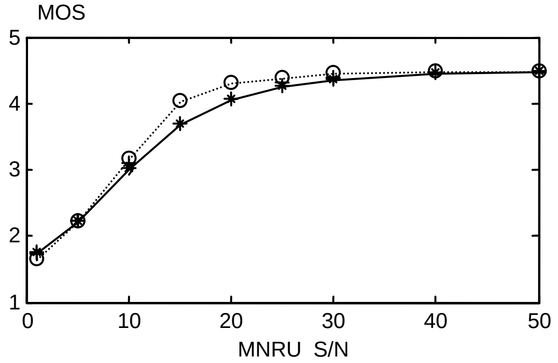

Fig. 18. Equivalent MOS for male speech degraded with noise modulation as a function of S/N, for PESQ (o), MQE (*).

We apply MQE method to various coded speech signals. Then, subjective tests have shown that MQE method is better than PESQ [9].

MQE method is not suitable for analysis of distortions that can be present during speech silence, as it is a modulation based method. That is way, it cannot be treated as general method for perceptual evaluation of speech quality, rather, as name suggests, it is a method for modulation quality evaluation.

7. CONCLUSION

This paper gives survey of fundamental measurement methods for testing audio quality in communication systems. Primary interests were systems for speech transmission and point-to-point testing of such systems.
Following measurements were considered important:

- measurement of frequency response, impulse response and input/output delay ,
- nonlinear distortions measurements with sine and multitone signals,
- perceptual evaluation of speech quality.

All other system parameters can be estimated from these measurements.

These measurements have lot in common with acoustical measurements, as we deal with systems that are not time-invariant and have large delays. But, there is one big difference: acoustical system is independent of excitation signal, while communication system characteristics depend on excitation signal (through automatic gain and noise reduction control). This leads to some fundamental differences from acoustical measurement. I.e. it is not recommended to use swept-sine excitation of Fourier analyzer in frequency response measurements of communication system, although it is a signal of choice in acoustical measurements.

This works shows that there is no ideal system for measuring frequency and impulse response in communication systems, but preferences are given to Fourier analyzer with interrupted periodic noise excitation. It allows use of concept of coherence function to monitor measurement S/N ratio. It also satisfied ITU requirement for excitation signal, which has to keep communication channel in active and quasi-stationary state.

Classical methods of measurement of nonlinear distortion with a THD and IMD are useless in coded systems. A much better way is estimation of total distortions of a multitone signal. We defined a "speech multitone" signal which has equally spaced tonal components on bark scale. A total distortion measure still needs to be standardized.

Finally, a method for perceptual evaluation of speech quality called modulation quality evaluation (MQE) is presented. It is based on a simple paradigm that quality of degraded speech signals can be predicted from changes of speech loudness modulation in critical bands.

Theoretical analysis has shown that MQE perceptual distance measure can be estimated in the excitation domain. There is no need to estimate the loudness. This results with simple implementation of fast MQE algorithm that can be used in real-time.

MQE can be used as a replacement for ITU recommended method PESQ in standard and GSM phone systems.

REFERENCES

[1] ITU-T Recommendation P.501: Test signals for use in telephonometry, ITU, 1996.

[2] ITU_T supplement 21: The Principles of a composite source signal as an example of a measurement signal to determine the transfer characteristics of terminal equipment, ITU, 1993.

[3] I. Mateljan, K. Ugrinović: The Comparison of Room Impulse Response Measuring Systems, Proceedings of AAAA Congress 2003, Portoroz 2003.

[4] Rohde & Schwarz Application Notes: Acoustic Measurements on GSM Mobile Phones with Audio Analyzer UPL and Digital Radiocommunication Tester CMD, Application Note1GA39_0D, 2004.

[5] Tan, Moore, Zacharov: The Effect of Nonlinear Distortion on Percived Quality of Music and Speech Signals, JAES, vol. 5, November, 2003.

[6] ITU_T P.800: Methods for subjective determination of transmission quality - MOS, ITU, 1996.

[7] ITU_T P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, ITU, 2001.

[8] ITU_T P.862.1: Mapping function for transforming P.862 raw result scores to MOS_LQO, ITU, 2003.

[9] I. Mateljan: The Modulation Approach in Perceptual Evaluation of Speech Quality, Proceedings of the Softcom 2004, Split-Dubrovnik-Venice (1994)

[10] E. Zwicker, H. Fastl: Psycho-acoustics, Facts and Models, Springer Verlag, Berlin, 1999.

[11] Jont B. Allen, Stephen T. Neely: Modeling the relation between the intensity just-noticeable differences and loudness for pure tones and wideband noise, JASA, vol. 102, December 1997.

[12] A. Farina: Simultaneous measurement of impulse response and distortion with a swept sine technique, 108 AES Convention, Paris, 2000.

[13] ITU_T P.810: “Modulated noise reference unit”, ITU, 1996.

[14] E . Terhardt, G. Stoll and M. Seewann: "Algorithm for extraction of pitch and pitch salience from complex tonal signals", J. Acoust. Soc. Am., vol. 71(3), March 1982.