# Objective Assessment of the Quality of Transmission and Informativeness of a Speech Signal According to Statistical Parameters

Vladimir Taktakishvili, Alexey Ovchinnikov, Oleg Popov, Valentin Abramov
Moscow Technical University of Communications and Informatics (MTUCI)
Moscow, Russian Federation
vladimir@smartvend.ru, ovchinnikovmsk@gmail.com,
olegp45@yandex.ru, vabramov44@mail.ru

*Abstract*—**In this paper, the possibility of an objective assessment of the quality of the transmission of a speech signal (SS) is discussed based on the use of the complex statistical estimation method. The statistics of the energy parameter SS – relative average power (RAP) is investigated. Comparison of RAP for SS estimates and transmission quality on the MOS scale confirms the possibility of predicting transmission quality by the median RAP value of channel. A study of the informativeness of speech signals by their statistical parameters. The statistical characteristics of the speech signals of speakers, which are used as lecture notes by leading lecturers of the "Radio and Television" Faculty of MTUCI, are studied. A number of recommendations have been formulated for optimizing the selected parameters in order to improve the perception of information by the audience.**

## I. Introduction

In the process of development of algorithms for compact representation and processing of a speech signal (SS), the study of distortions in the transmission channel and an operational assessment of the SS sound quality is necessary. Sound assessment methods can be divided to two types:

- Subjective assessment methods. In this type there are two basic methods: *Pair comparison* method and *Scoring* method (also called as *method of point estimation*). Both those methods can't be used for operational assessment, as they are statistical in nature and require large amount of subjective tests with different people to be made. Also there is large deviation, typical for all subjective methods.
- Objective assessment methods. Those are divided into three different classes: Intrusive, Nonintrusive and Model-based ones.

Objective methods are divided into intrusive, requiring a standard (sound sample, that is used at transmission side), non-intrusive, passive, not requiring a standard and modeled. In the intrusive methods, a comparison of two speech signals, reference and distorted during transmission over a communication network, is carried out. Intrusive methods for assessing the quality of speech signals include *Perceptual evaluation of speech quality (PESQ)* [1] and *Perceptual objective listening quality assessment (POLQA)* [2]. Non-intrusive methods, based on a narrow-band telephony application, allow to evaluate speech quality without comparison with a reference signal. Simulated methods based on the *E-model* [1], developed for the development of networks and communication systems, are successfully used in monitoring the quality of voice transmission.

PESQ takes into account the following causes of speech quality degradation:

- coding distortion;
- transmission errors;
- packet loss;
- time delay fluctuations, including out-of-order packet reception;
- signal filtering (in analog networks).

It does not take into account the effect of changes in the signal level in the network on the communication quality, the presence of an echo signal, and the round-trip delay [1].

As shown in [3], POLQA has significant flaw, that annoying micropauses (10-300 microseconds) can be stretched by some delayed packet. Listener will feel that as bad connection, but POLQA shows that everything is OK.

The simulated method of analyzing the quality of voice communication is based on the E-model calculation algorithm [1], which takes into account more than twenty different parameters, including:

- signal transit time;
- delay variation (jitter);
- packet loss;
- packet loss peaks (Bursts) and others.

The results of the assessment have a good correlation with the subjective MOS-scores, but are less accurate than the P.563 algorithm [2].

POLQA is designed for IP-based networks, fixed and mobile communications and allows you to evaluate the quality of voice communications between GSM /UMTS /LTE and CDMA networks [2].

POLQA provides support for two modes; when operating in the first mode, POLQA issues a rating on a narrowband scale, and in the second, on an ultra-wideband scale. The POLQA algorithm uses an advanced psychoacoustic model that mimics the human perception of sound and transforms it into a neural internal representation. It takes into account the characteristics of human hearing - the critical bands and frequency masking. At the initial stages, speech distortions are determined and weighted at the level of a person's perception: noise, spectral distortions, and reverberation of sound. After the primary process of idealization, which aligns the reference and distorted signals according to the level and shape of the spectrum, psychoacoustic analysis takes place. The secondary idealization process suppresses noise, which guarantees the comparison of speech information in the block of internal signal representation. In conclusion, the algorithm determines how annoying the distortions are for user, and a quality score based on the MOS-scale is formed [2].

The tasks solved in the process of forming the evaluation using the POLQA algorithm are closest to the tasks solved at the TV&SB Department of MTUCI and are associated with an objective assessment of the quality of audio signal transmission in modern channels with the elimination of statistical and psychophysical redundancy and intensive audio processing of the signal. The disadvantages of POLQA, like of many other methods of objective evaluation of a sound broadcasting signal, are determined primarily by the method of spectral analysis that does not provide sufficient accuracy and resolution and is almost an order of magnitude different from the corresponding ability of human hearing. In addition, the POLQA algorithm uses a signal perception model that is based on the results of a narrow-band noise or harmonic signal perception study.

When perceiving a real sound signal, perception occurs according to completely different laws with other hearing thresholds, frequency masking curves [1], etc. Therefore, an integrated method for assessing sound quality based on changes in the statistical characteristics of a signal corresponding to the perception of SS by the listener has been developed at the TV&SB Department. In forming the assessment, it is proposed to use energetic, spectral, and cepstral parameters, as well as the shape parameters of the analytical envelope of the audio signal. The effectiveness of this approach has been confirmed practically.

Work was carried out on the study of the possibility of using the Department's workings in assessing the quality of speech signal transmission through channels with the elimination of psychophysical and statistical redundancy leading to non-preservation of the waveform.

The changes in the statistical characteristics of all the above signal parameters were investigated. The possibility of forming an integrated quality assessment based on the criteria for noticability of signal distortion, score and preference of sound was confirmed. In the process of analysis, statistical distributions of the parameter are formed at the input and output of the distorting channel, and their integral deviation serves as a measure of the change, in the particular case of distortion, of the signal.

## II. Relative Average Power as main parameter

The most obvious are the changes in the energy parameters of the SS, which are determined by its relative average power, i.e. the ratio of SS power to the power of a harmonic signal with an amplitude equal to the nominal for the channel (RAPk), or to the peak amplitude of SS for the duration of the measurement (RAPs).

The measurement duration should be selected as the time of formation of the sensation of sound in the auditory analyzer, which is approximately equal to 200 ms. RAPk determines the volume of the message, and volume variations in the form of the standard deviation of the RAPk determine the dynamic range of message, as well as the emotionality of this message. Figures 1 and 2 show the examples of integral deviations of the RAP distribution after a compact representation of the audio broadcasting signal and after LF filtering [1], calculated using the *Estim* program developed at the TV&SB Department [4].

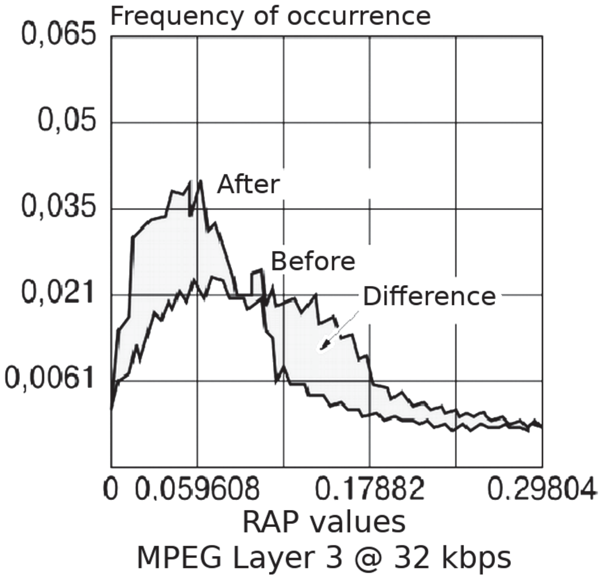

Fig. 1. Integral deviation of the distribution of the RAP after MPEG Layer 3 encoding, that corresponds with audible distortion

The algorithm includes the following operations on both initial and processed signals:

- Segmentation of the signal into intervals corresponding to the time of formation of the volume T (about 200 ms);
- Calculations of the RAPs for each of the intervals;
- Splitting the range of values of integral RAP (IRAP) into N intervals;
- Calculation of statistical frequencies of IRAP in each of N intervals;
- Calculation of estimates of statistical mean, standard deviation and median.

At the last stage, the statistical characteristics of IRAP are compared for the processed and initial signals. Also, too form an assessment of the gradation changes in the dynamics, in

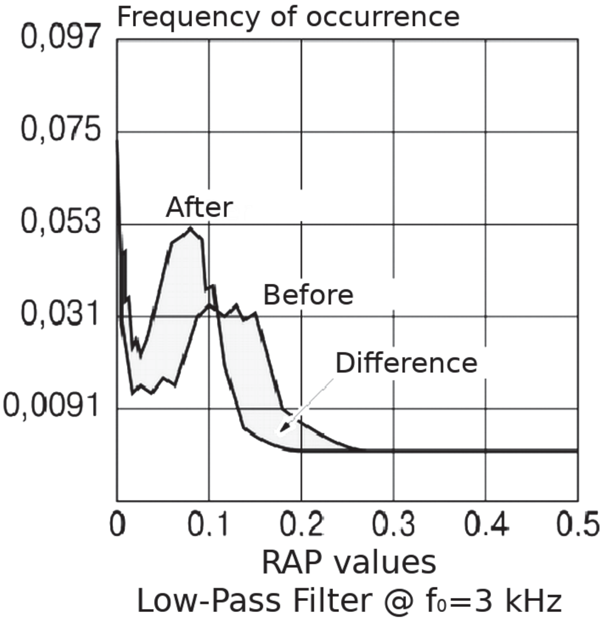

Fig. 2. Integral deviation of the distribution of the RAP after Low-Pass Filtering, showing audible distortion

addition to the IRAP, you can use the first difference of the IRAP function over time.

The integral deviation of the statistical characteristics of IRAP of the original and distorted signals given in Figures 3 and 4, allows us to predict a subjective assessment of the quality of transmission from the listener.

Formation of estimates of the parameters of the form, spectrum and cepstrum, are made similarly. Figure 5 shows the results of comparing the RAP estimates for SS transmitted at different rates and the transmission quality on the MOS scale, confirming the possibility of predicting the transmission quality by the median RAP value.

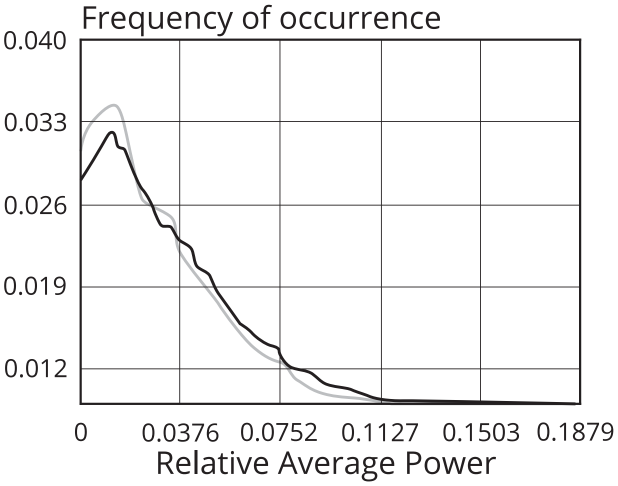

Fig. 3. Distribution of RAPk of the original SS (black) And transmitted over a channel with a transmission speed of 8 kbit/s (gray)

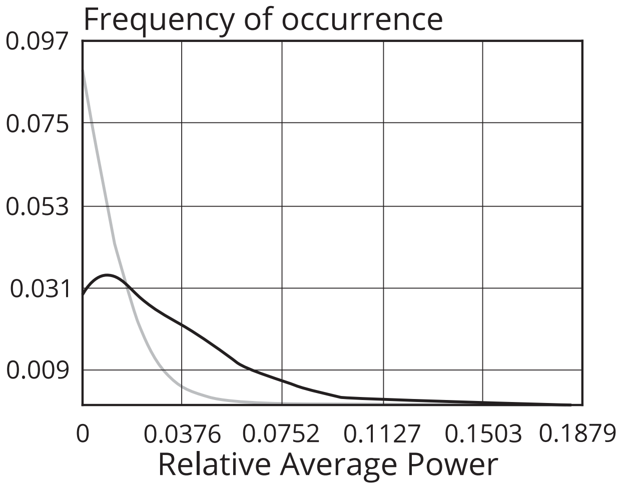

Fig. 4. Distribution of RAPk of the original SS (black) And transmitted over a channel with a transmission speed of 4.8 kbit/s (gray)

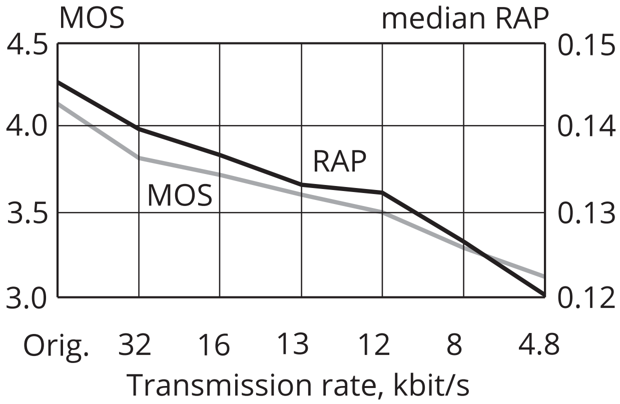

Fig. 5. Comparison of the transmission quality estimates on the MOS scale and the median RAPk value at various transmission rates

## III. Psychophysical perception

The integral information content of the speech signal (SS) is determined by the combination of the semantic and emotional components of this signal. Semantic informativeness, as a first approximation, is determined by the intelligibility of the message. Work on an objective (technical) assessment of emotional informativity began in the early 1930's at Moscow State University. There were studied specially recorded phonograms by Chaliapin, who sung with different emotional attitudes: fear, anger, joy, etc. The research was continued by Abraham Mole in 1950's [5], I.M. Kogan in the 1980's, O.B. Popov and S.G. Richter in the 2000's [6].

In the course of the work carried out at the TV&SB Department of MTUCI, a number of parameters of the audio signal were revealed, the statistical characteristics of which make it possible to predict the assessment of signal quality by the listener. In this paper, using these parameters, we analyzed the speech signals of a number of leading lecturers at the faculty of R&T of MTUCI, as well as professional readers and preachers. That made it possible to formulate recommendations for improving the communication of information to the listener.

Recall that RAPk determines the volume of the message, and the RAP standard deviation – dynamic range and emotionality (see Figure 6).

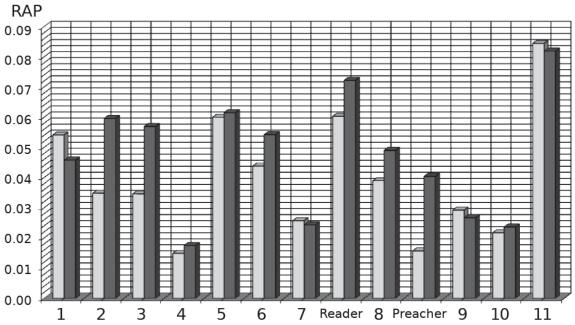

Fig. 6. RAPk (left column) and RAP standard deviation (right column) for SS of different lecturers from MTUCI (shown by numbers), professional reader and preacher

As can be seen from the histograms in Figure 6, the range of loudness among the lecturers is large enough, but for a professional reader and priest standard deviation of RAP – the dynamic range – is wider than that of most lecturers. RAP depends on the nature of the signal and is significantly different for signals with excitatory and inhibitory physiological effects, the distributions of which are shown in figure 7 [6].

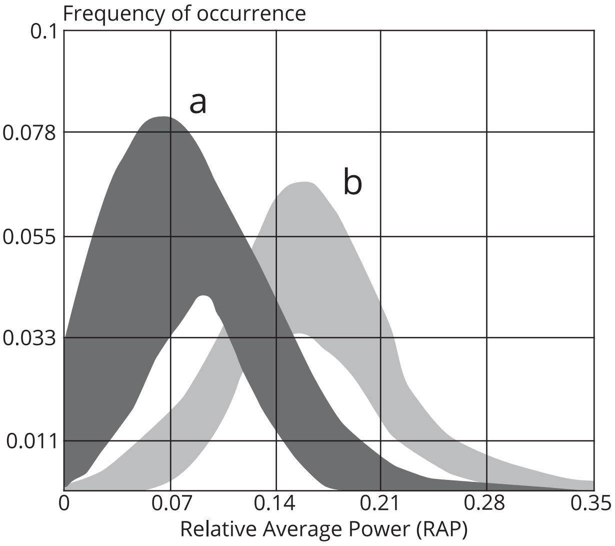

Fig. 7. Distributions of instantaneous values of RAPs for speech exerting a inhibiting (a) and a stimulating (b) effects

In figure 7, the integral distributions of instantaneous values of RAPs with and inhibiting (dark region, "a") and stimulating (light region, "b") effects are given [2]. According to the histograms of assessments of the RAPs shown in Figure 8, most of the lecturers have an inhibiting effect on the listener, unlike the professional reader and the priest.

The distribution of the relative amount of emotional (aesthetic) and semantic information per unit band is shown in figure 9. [1]

The SS spectrum largely determines the transmission of both semantic and emotional information.

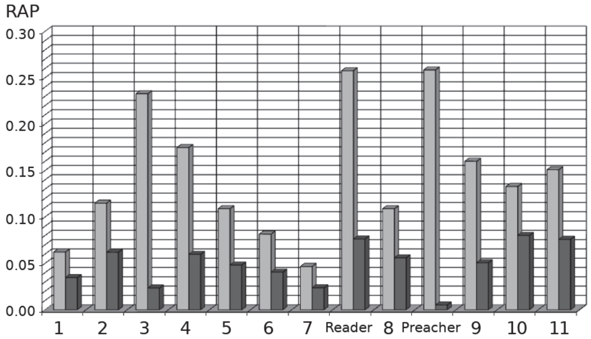

Fig. 8. A histogram of assessments of RAPs and RAPs standard deviations for SS of different lecturers from MTUCI (shown by numbers), professional reader and preacher

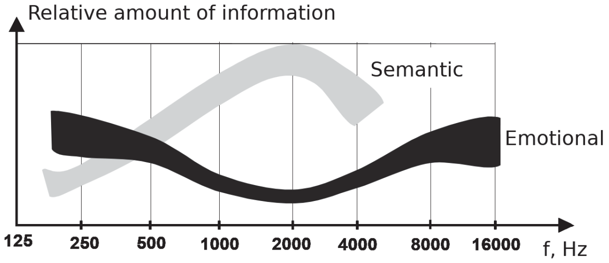

Fig. 9. Distribution of the relative amount of audio information per unit of band

## IV. Rhythmic components

Significant influence on the preference of the SS has a harmonious of the signal. As a rule, harmony is evaluated using cepstral analysis. A *cepstrum* is a repeated FFT transformation of a transformed (usually logarithmic) envelope of the amplitude spectrum of a signal. To form an estimate, the peak factor of cepstrum is usually used - the ratio of the peak value of the cepstral coefficient to the average. The calculations of the cepstrum peak factor for the MTUCI lecturers showed that most of them emit a rather harmonious signal that does not reach the capabilities of a professional reader, but is superior to a priest.

The rhythmic parameters of the signal have a great influence on the SS perception process. Everyone knows the emotional impact of poetic speech or the influence of the rhythms of music on the listener. They directly affect the rhythms of the brain, imposing a certain state on the listener. In fig. 9 shows the estimates of the three most powerful rhythmic SS frequencies for the MTUCI lecturers with the boundaries of the zones determining the state of brain activity.

As can be seen from the histograms, most lecturers use SS rhythms, leading the listener to a state of non-critical perception (meditation), when information is remembered as ultimate truth without attempting to analyze it. Usually that's good, but this "feature" of human brain is frequently used in wrong way – usages vary from advertisments to crowd manipulation and propaganda. a professional reader and a priest seek to bring the audience to the same state of mind. The ability to bring listeners into such a state is taught by

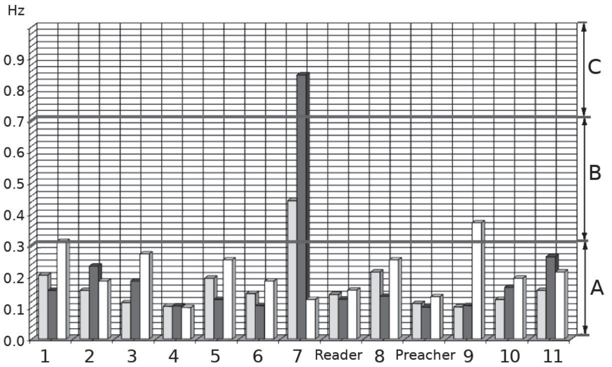

Fig. 10. Three main frequencies in capstrum of different lecturers from MTUCI (shown by numbers), professional reader and preacher. A – Zone of Uncritical Perception, B – Zone of Activation of the Brain, C – Zone of Active Perception

the creators of financial pyramids, and sometimes by public figures.

Interesting results were obtained when analyzing the number of sound objects (phonemes) synthesized by speakers per unit of time. Naturally, the teachers' desire to give students the maximum amount of information – number of phonemes per second. At the same time, the listeners' perceptions of SS are limited, which is taken into account by professional readers and priests (Fig. 11).

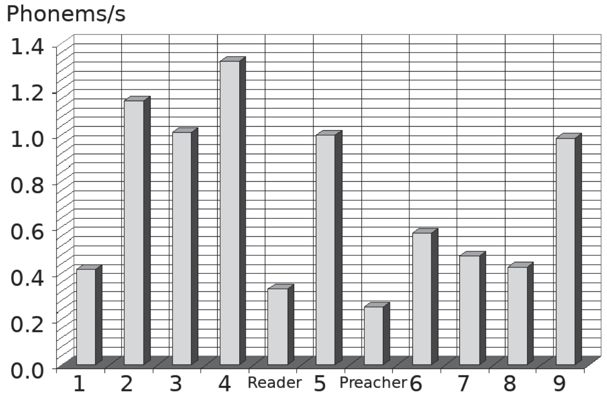

Fig. 11. A histogram of the average number of phonemes per second for different lecturers from MTUCI (shown by numbers), professional reader and preacher.

The statistical parameters of the SS used in the present work are formed using the program ”Estim” [4], developed at the TV&SB Department of MTUCI, to predict the assessment of the sound quality of radio programs by the listener. In forming the parameter estimates, a number of original algorithms are used to increase accuracy of received data [7], [8], [9]. This study showed that the results of the analysis were also useful for the evaluation of the public speaking (lecture) message. Thus, the evaluation of the statistical parameters of the speech signal of the speaker or lecturer allows an objective assessment of the effectiveness of his communication to the listener. According to the results of the study, the following recommendations can be formulated for speakers:

- the energy parameters of the relative average power (RAP) of speech signal in channel (RAPk), short-based relative average power (RAPs) and RAPk standard deviation can be improved by targeted training of speakers (lecturers), which allows maintaining a nearly constant level of relative average power for prolonged time and increasing the speech dynamic range, defining the message emotionality;
- rhythmic characteristics of speech can be improved by developing rhytmic sense of speech, training the speaker to speak under the metronome;
- in some applications it's possible to use audio processing systems, that can adjust and equalize dynamical range for better perception of speech by listeners;
- the change in the number of phonemes pronounced by the speaker or lecturer per unit of time and determining the volume of semantic information, is possible only by a targeted change in the form of presentation of the material, for example by usage of multimedia devices during lecture, which can extremely decrease time loses for drawing and other dispensable actions;
- The parameters of the SS, such as RAPk, RAPk standard deviation, RAPs, spectral, cepstral estimations and rhythmic characteristics can be purposefully changed using deep high-speed audio processing with the automatic Hilbert envelope regulator, such as ARGO, developed at the "Television and Sound Broadcasting" Department of Moscow Technical University of Communications and Informatics department [5], [6], [10].

## V. Conclusions

1) Speech Signal (SS) has many aesthetic parameters, that are still not described and not measured, while those parameters, being optimized, can dramatically improve speech content quality of lecturers and other speakers. Scientists try to extract from speech signal as much information as possible, but there is a lot of undechipered dependencies.
2) Existing methods for assessing the quality of speech signal transmission allow to reliably detect: gross distortion during coding; transmission errors; packet loss; delay time and fluctuations during packet transmission; signal filtering in analog networks, but do not allow to form an estimate of the quality of transmission on the MOS scale.
3) The methods associated with modeling the process of perception of a speech signal by a listener form a very approximate assessment since they use spectral analysis for all parameters almost an order of magnitude less accurate than in an aural human analyzer, and the perception model used is adequate only for perception of a single-frequency harmonic signal or narrowband noise.
4) The possibility of forming an integral assessment of the quality of transmission of a speech signal through channels with the elimination of redundancy based on changes in the statistical parameters of the signal,

in particular, the relative average power with the transition to the MOS scale, was confirmed.

5) Analysis of the statistical parameters of the SS allows to predict the effectiveness of communicating to the listener a speech message including both semantic and emotional components and recommend ways to improve their transmission.

## REFERENCES

[1] S.G. Richter, *Coding and transmission of speech in digital mobile radio systems. Textbook for universities*, Moscow: Hotline–Telecom, 2011, 304 pages.

[2] O.B. Popov, S.G. Richter, A.N. Terekhov and T.V. Chernysheva, *Methods of quality assessment in TV and radio channels*, Moscow: Hotline–Telecom, 2016, 232 pages.

[3] A.N. Terekhov, "The flaw of the intrusive method of evaluation of the quality of speech transfer and method of its elimination", Moscow: *T-Comm*, #5, 2016, p.98-102.

[4] V.A. Abramov, G.M. Ozhdikhin, O.B. Popov, K.V. Chernikov and A.V. Malov, "Parameters analysis of sound signals ESTIM". Certificate of registration of software No. 2013616645, 15.07.2013.

[5] A. Mol, *Theory of Information and Aesthetic Perception*, Moscow: Mir, 1966, 211 pages.

[6] O.B. Popov and S.G. Richter, *Digital processing and measurement of signals in the paths of audio broadcasting*, Moscow: Insvyazizdat, 2010, 320 pages

[7] V.A. Abramov, O.B. Popov and S.G. Richter, "A method for measuring the instantaneous and average values of the absolute and relative power of acoustic signals and a device for its implementation". RF patent No. 2458340 BI n10. 10.04.2012.

[8] V.A. Abramov and O.B. Popov, "The method of measuring the spectrum of information acoustic signals of broadcasting and device for its implementation". RF patent No. RU2573248 C2, BI n2. 20.01.2016.

[9] S.A. Litvin, O.B. Popov and T.V. Chernysheva *Audio processing of audio broadcast signals. Textbook for universities*, Moscow: Hotline–Telecom, 2016, 232 pages.

[10] V.A. Abramov, O.B. Popov and S.G. Richter, "A method for measuring the instantaneous and average values of the absolute and relative power of acoustic signals and a device for its implementation". RF patent No. 2458340 BI n10. 10.04.2012.