

A METHOD OF MEASURING THE INDEX OF ACOUSTIC VOICE QUALITY BASED ON AN INFORMATION-THEORETIC APPROACH

V. V. Savchenko

UDC 534.6:53.082.4

A new method is proposed for measurements of an index of acoustic voice quality using the Kulbaka–Leybler information metric. The primary benefit of this method consists in its high-level dynamic properties, as calculated by eliminating the problem of a small number of observational samples. A theoretical study of the method's efficiency was conducted, and its conclusions were confirmed experimentally. It has been established that, in order to provide a sufficiently precise assessment of the voice quality of the speaker, a speech signal of duration 2–3 minutes is required.

Keywords: *speech acoustics, acoustic measurements, small samples problem, information-theoretic approach.*

Speech, as a method of communication between people, has steadily attracted the attention of the scientific community for many years. Currently, automatic speech processing systems are broadly distributed around the world [1–3]. The efficiency of such systems depends on many factors, the first of which is the quality of speech that is manifest in all its main acoustic properties, such as recognizability, loudness, timbre, and intelligibility [4, 5]. The amplitude-phase spectrum of the minimal sound units, phonemes and their allophones, is the universal characteristic of the acoustic properties of speech [6]. Hence, specialists properly associate the evaluation of acoustic voice quality with measurements of the complex frequency spectrum $S(j, f)$ of speech signal $x(t)$ at current time t for limited time intervals of duration $T = 10\text{--}20$ msec per speech frame [7], where f is the cyclic frequency.

The first-ever technical device capable of distinguishing the vowel sounds of Russian oral speech from the distribution of energy of a speech signal on 14 parallel bands of the sound range of frequencies up to 4 kHz was created in 1946 in the USSR [8]. The spectrum-based orientation of acoustic research was continued by scientists in the USA, within the Visible Speech technology of dynamic spectrography, created by them one year later [9]. Hence, a considerable amount of material in the field of speech acoustics had been accumulated by the middle of the 20th century. In the 1960–70s, researchers already understood that automatic speech processing is a very challenging scientific and technical problem, the solution to which requires a great volume of specialized knowledge and a powerful technical base. At the same time, acute problems in this line of research were also designated [10]. In particular, essential distinctions were discovered between the linear orderliness of phoneme chains in a speech stream and the physical partitioning of a speech signal into frames in which these phonemes are implemented. The study of the correspondence between linguistic markers of phonemes and their physical correlates turned out to be even more complex. It is shown in [11] that the very same sound unit can have a multivariate (variable) articulation and acoustic embodiment depending on the functional status of the speaker and his specific voice features. In specifically that way, the effect of within-speaker acoustic variability of speech, and moreover variability of random and unpredictable nature, is revealed [12, 13]. Hence, a non-trivial approach based on the likelihood model of a speech signal is required in order to evaluate the acoustic parameters of voice quality. This article is devoted to its research and execution within a new method of acoustic measurements.

Unlike methods based on traditional spectral characteristics of speech sounds such as distortion of the frequency range of a speech signal, log-spectral distance, Itakura–Saito distance, and others, a frequency equivalent of the Kulbaka–Leybler measure lies at the base of the research that was conducted. This equivalent embodies a universal information-theoretic approach and is coordinated with standard auditory estimates of acoustic voice quality [4–6, 14–19]. This made it possible to formulate the developed method in terms of a likelihood model of speech sounds of the speaker, taking into account the statistical stability of the results of measurements from sample to sample, and also the consistency of these results [12, 18]. Here, the solution of the small observation sample problem has essential significance in the new method [6, 7].

Task statement. The amplitude-phase spectrum $S(f) = |S(i, f)|$ of a speech signal $s(t)$ over an observation interval of finite duration $t < \infty$ is determined by the modulus of its spectral density in the function of frequency oscillations f , measured in Hz [7, 8]:

$$S(j, f) = \int_0^T s(t) \exp(-i2\pi ft) dt.$$

Different amplitude-phase spectra correspond to different speech sounds. The amplitude-phase spectrum of a speech signal on a finite J -set of realizations $s_j(t)$, $j \leq J$, of a specific (the same) sound unit is described by J different dependences $S_j(f)$ in which the effect of acoustic variability appears. Consequently, one may draw a conclusion on the quantitative measure of this sound unit: this is a certain distance

$$\rho_{j,k} = \rho[S_j(f), S_k(f)]$$

between single-phoneme realizations $s_j(t)$ and $s_k(t)$ of a speech signal in the frequency domain, averaged over the set of corresponding spectral characteristics $\{S_j(f)\}$, according to the expression

$$\rho(J) \stackrel{\Delta}{=} J^{-2} \sum_{j=1}^J \sum_{k=1}^J \rho_{j,k}, \quad (1)$$

where the symbol $\stackrel{\Delta}{=}$ designates equality by definition.

For example, in the case of classical Euclidean metrics, the mathematical description of the material form of a speech signal has the form

$$\rho_{j,k} \stackrel{\Delta}{=} F^{-1} \int_0^F [S_j(f) - S_k(f)]^2 df, \quad j, k \leq J,$$

where F is the the upper frequency of f in the speech signal spectrum.

However, this metric is insufficiently sensitive to distinctions in the spectral characteristics of speech sounds outside the regions of their primary (dominant) formant [14, 15]. This indicated shortcoming is excluded from the Itakura–Saito distance, which is broadly distributed in acoustics research and is coordinated with auditory estimations of speech quality [5]:

$$\rho_{j,k} \stackrel{\Delta}{=} F^{-1} \int_0^F \{S_j^2(f)S_k^{-2}(f) + \ln[S_k^2(f)S_j^{-2}(f)] - 1\} df. \quad (2)$$

It has been confirmed in [19] that the admissible degree of variability of speech sounds within a set of realizations $\{S_j(f)\}$ is defined in the metric of (2) by the threshold condition $\rho_{j,k} \leq 0.1$. Here, in recalculating to the symmetric form of COSH distance [15]

$$\tilde{\rho}_{j,k} \stackrel{\Delta}{=} 0.5F^{-1} \int_0^F [S_j^2(f)S_k^{-2}(f) + S_k^2(f)S_j^{-2}(f) - 2] df \quad (3)$$

we write, in accordance with (1),

$$\rho(J) = F^{-1} J^{-2} \sum_{j=1}^J \sum_{k=1}^J \int_0^F \{S_j^2(f)S_k^{-2}(f) + \ln[S_k^2(f)S_j^{-2}(f)] - 1\} df =$$

$$\begin{aligned}
&= F^{-1}J^{-2} \int_0^F \left\{ \sum_{j=1}^J \sum_{k=1}^J S_j^2(f) S_k^{-2}(f) + \sum_{j=1}^J \sum_{k=1}^J [\ln S_k^2(f) - \ln S_j^2(f)] - 1 \right\} df = \\
&= F^{-1}J^{-2} \int_0^F \left[\sum_{j=1}^J \sum_{k=1}^J S_j^2(f) S_k^{-2}(f) - 1 \right] df = J^{-2} \sum_{j=1}^J \sum_{k=1}^J \tilde{\rho}_{j,k}. \tag{4}
\end{aligned}$$

Currently, the COSH distance is one of the most widespread numerical measures of spectral variability in speech acoustics. However, its practical use is limited in this case by the unique nature of the results obtained (calculation on a specific speech stream, and the defined conditions of its production and the specific speaker). This is a direct consequence of the deterministic models (2) and (3) of the speech signal that are the foundation of the COSH distance (4). Hence, hereafter the author proposes in its stead a different index of speech sound quality, within which a universal information-theoretic approach is implemented.

Indicator of acoustic speech quality. Expression (2) formally repeats the definition in the frequency domain of the informational mismatch (divergence), in accordance with Kulbaka–Leybler, of the Gaussian signals $s_j(t)$ and $s_k(t)$ [16]. Moreover, the essential position of the theory in relation to the speech signal model being used is concealed behind the formal side of the problem. If a deterministic model that has been specified by energy spectra or signal power spectra $S_j^2(f)$ and $S_k^2(f)$ was used within expression (2), then with reference to the Kulbaka–Leybler divergence one should mention the statistical estimates $G_j(f)$ and $G_k(f)$ of the power spectrum density (PSD) $G(f)$ of a hypothetical (modeled) random process over its interval of stationarity of duration T [17, 20].

At first sight, an inadmissible substitution of concepts has occurred: the initial deterministic model of a speech signal is unjustifiably replaced by a likelihood model. However, taking account of the asymptotic definition of the PSD:

$$G(f) = \mathbf{M}[T^{-1}S_j^2(f)]$$

as $T \rightarrow \infty$ ($\mathbf{M}[\cdot]$ is the symbol for population mean) through the normalized mean statistical value of the power spectrum of a signal, the replacement taken above is theoretically strictly conditioned by the equality of general form [20]:

$$G_j(f) = T^{-1}S_j^2(f) \quad \forall j \leq J.$$

Then by analogy with (2), one may write

$$\rho_{j,k} = F^{-1} \int_0^F \{G_j(f)/G_k(f) + \ln[G_k(f)/G_j(f)] - 1\} df. \tag{5}$$

In symmetric form, the expression has the form

$$\tilde{\rho}_{j,k} = 0.5F^{-1} \int_0^F [G_j(f)/G_k(f) + G_k(f)/G_j(f) - 2] df. \tag{6}$$

Substituting (6) into (1) within the calculation framework of (3), we derive the dependence

$$\begin{aligned}
\rho(J) &= J^{-2} \sum_{j=1}^J \sum_{k=1}^J \tilde{\rho}_{j,k} = F^{-1}J^{-2} \int_0^F \left[\sum_{j=1}^J \sum_{k=1}^J G_j(f)/G_k(f) - 1 \right] df = \\
&= F^{-1} \int_0^F \left[\overline{G}(f, J) \times \overline{G^{-1}}(f, J) - 1 \right] df, \tag{7}
\end{aligned}$$

where

$$\overline{G}(f, J) \stackrel{\Delta}{=} J^{-1} \sum_{j=1}^J G_j(f), \quad \overline{G^{-1}}(f, J) \stackrel{\Delta}{=} J^{-1} \sum_{j=1}^J G_j^{-1}(f)$$

is the mean arithmetic and mean inverse (harmonic) value over the set $\{G_j(f)\}$ of PSD estimates.

Expression (7), taking into account equalities (5) and (6), characterizes the probabilistic Gaussian model of the minimum sound units [6]. This model will be coordinated with results of field studies of the sounds of Russian speech, based on the systemic principle of the minimax of entropy [12, 21].

The task consists in strictly defining the criterion for evaluating the voice quality of the speaker in the form of objective and statistical characteristic using acoustic parameters (7). With a sufficiently large volume J of the sample, this characteristic is extended with a high degree of confidence to a given speaker's speech flow, obtained at different times and in various conditions. The speed of convergence of index (7) to its asymptotic limit value $\rho^* = \rho(\infty)$ thus is of crucial importance.

It is known from spectral analysis theory that for any finite $T < \infty$, both spectral estimates entering the right side of expression (7) do not possess the property of consistency [20]. However, they do converge asymptotically (as $J \rightarrow \infty$) to the optimum on dispersion of variations, since statistical averaging is used in them. Moreover, they converge relatively quickly under the $1/J$ power law. Therefore, it is possible to confirm that in the first approximation under the power law the statistical estimate (7) as a whole also converges to its limit value ρ^* . Here, the metric relationship is $\rho^* \geq 0$, with equality to zero only in the degenerate case, in the absence of any variability within the set of implementations of the PSD of the same index $\{G_j(t)\}$, i.e., when the equality $\overline{G}(f, J) = 1/\overline{G^{-1}}(f, J)$ is fulfilled.

An acceptable result (at a significance level on the order of several percent) will generally be available for $J \geq 100$. And these are very low requirements from the point of view of practice of acoustic measurement [5–7]. The conclusions that have been drawn are confirmed by the successive results of physical testing.

Program and results of research. The author's Voice Self-Analysis software system was involved in the conducted experiment [22]. Based on algorithms (2)–(6) and the Berg–Levinson computational procedure [20] for $T = 10$ msec and the digitization frequency of the speech signal 8 kHz (coordinated with a 4 kHz pass band on a standard telephone communications channel), the speech quality index (7) estimated from acoustic markers is implemented in a relative (percentage) expression in this system:

$$\delta(J) = 100 / [1 + \rho(J)]. \quad (8)$$

This index is determined automatically: sequentially for segments of a speech signal of identical duration 20 sec. Here, the vowel speech sounds: A, O, U, I, Y, and E were used exclusively, as the most informative among all other sound units of Russian speech in the acoustic-articulation sense [18]. Human speech is oscillations of complex form that depend on the words being pronounced, the timbre of the voice, intonation, sex, and age of the speaker. Hence, the control group of speakers consisted of three people: men 60 and 31, and a woman 30 years old. Each of them read in on-line mode into a microphone the same artistic text from Chapter 1 of "The Captain's Daughter" by A. S. Pushkin. The results obtained for each speaker were displayed on the PC monitor in the form of a dependence graph (8) dynamically, sequentially for each segment of the speech signal in order. In order to monitor the information index (7) and record the moment at which the acoustic voice quality of the speaker reaches a high value, the corresponding index was placed in the software system. In the experiment being examined, recording such events required no more than two to three minutes.

The voice quality minimum, as estimated from acoustic markers, generally occurred at the beginning of the period of the speaker's reading the control text, when the reader had not yet managed to concentrate. On the contrary, the voice quality maximum was recorded most often by the program closer to the end of reading, i.e., during moments of maximum concentration of attention of the speaker. At the very end of this period, index (8) decreased, albeit insignificantly, in the process of natural exhaustion of the speaker in the course of continuous reading. In this, the high precision of the conducted measurements was ensured by a considerable volume of data test sample.

At the specified digitization frequency of a speech signal, $1.6 \cdot 10^5$ data samples were obtained within each of its segments, or on the order of $L = (8-9.6) \cdot 10^4$ (about 50–60%) with reference to vowel sounds. In recasting to one vowel (of

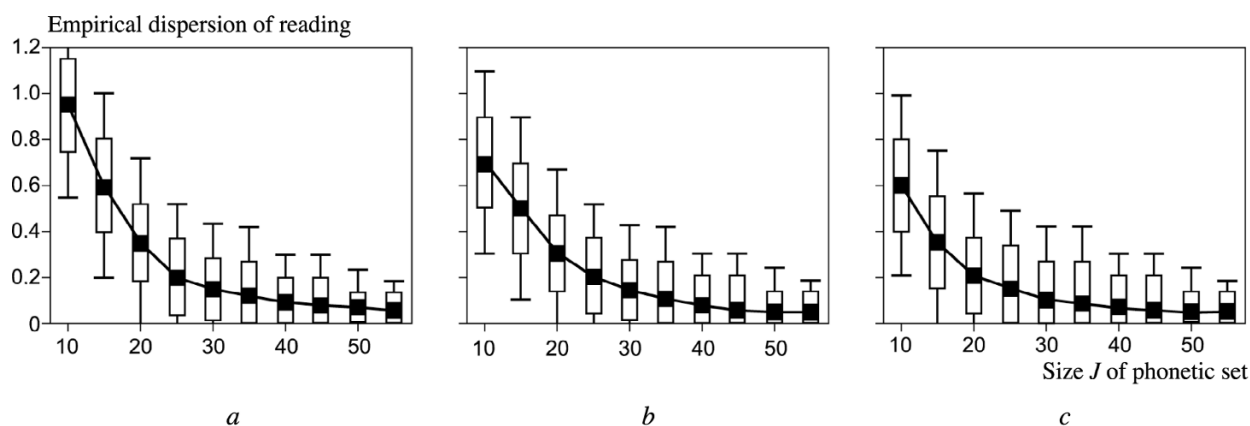


Fig. 1. Dynamics of the convergence on the dispersion of the estimate of the acoustic quality of the sound of the vowel O for the first (a), second (b), and third (c) readers.

the six), on the average this gave each Russian vowel not fewer than $J = L / 6 \cdot 80 = 200$ of its samples (allophones), which provides evidence of the sufficiently high precision of measurements of an index (8).

High speed or minimal requirements for the amount of the control sample of a speech signal is the chief advantage of the developed method among its modern analogs. This is particularly shown in Fig. 1, which presents three charts of the dispersion of the index of acoustic voice quality dynamically for all speakers and with reference to the vowel sound O. According to the presented results, the convergence of index (7) to its lower bound ρ^* is reached with a relatively small number $J = 40\text{--}50$ of samples of the speech sound. Moreover, for all other vowels spoken by different speakers, the dynamics of convergence by the character had practically no change. Thus, the pressing issue of small observation samples in the acoustic measurement task has been resolved successfully. Here, the acoustic features of different speakers are shown mainly in the value that is maximally achievable for each of them: the smaller this value, the less the variation in pronunciation of the specific speaker.

The experimental estimates that were obtained confirm as a whole the results of the theoretical study conducted previously. Along with the guaranteed reliability in the asymptote, the developed method possesses high dynamic properties and can be recommended for application in tasks with small samples, the amount of which is commensurate with the response time (or memory) of a signal processing system.

Discussion of the obtained results. In the majority of work so far on studying the quality of oral speech, the heuristic approach based on auditory or expert estimates still dominates. And although their authors formulate their results in terms of statistical estimates and conclusions, in the absence of a strict criterion of effectiveness they have an essentially deterministic, individual nature and are consequently not beyond the subjective characteristics of specific texts and the evaluation of specific speakers. This article presents a criterion of a different type. Its theoretical justification is based on the principle of the minimum of information mismatch by Kulbaka–Leybler. The main distinctive feature of the criterion is the presence of high dynamic properties in the calculations for resolving the issue of small samples. It was shown in practice, using an example of speech analytics, that when using the information index (7) it is necessary to have a speech signal duration of several minutes so that the evaluation of the acoustic voice quality of this speaker be sufficiently precise and informative.

Conclusion. Strict substantiation of a new method of measuring voice quality using the Kulbaka–Leybler information metric follows from the results of the test that has been conducted. Requirements have been defined for the volume of representative speech material. The proposed method, unlike the majority of known methods, extends to different speech streams, various times, and any conditions of speech production. The results obtained in this article are important for both the theory and the practice of acoustic measurement. The fields of application of the studied method includes the theory and technology of digital speech processing, as well as modern speech technologies in daily life, production, and transport, where voice quality is a key factor from the point of view of overall performance of system as a whole.

REFERENCES

1. A. Yu. Tychkov, A. K. Alimuradov, and P. P. Churakov, "Adaptive signal processing method for speech organ diagnostics," *Izmer. Tekhn.*, No. 5, 26–29 (2016).
2. A. Yu. Tychkov, A. K. Alimuradov, M. V. Frantsuzov, and P. P. Churakov, "Software implementation of an algorithm for recognition of speech signals in the Labview graphics programming environment," *Metrologiya*, No. 3, 1–25 (2015).
3. V. V. Savchenko, "Vocoder testing using the criterion of minimal required redundancy of a speech signal," *Telekommunikatsii*, No. 1, 17–25 (2017).
4. A. K. Alimuradov and F. Sh. Murtazov, "Methods to improve the efficiency of recognition of speech signals in voice control systems," *Izmer. Tekhn.*, No. 10, 20–24 (2015).
5. GOST R 50840-95, *Speech Transmission on Communications Media. Methods of Evaluating Quality, Legibility, and Recognition*.
6. V. V. Savchenko, "The principle of the information-divergence minimum in the problem of spectral analysis of the random time series under the condition of small observation samples," *Radiophys. Quan. Electr.*, **58**, No. 5, 373–379 (2015).
7. B. A. Atayants and V. S. Parshin, "Measurement of the frequency of a harmonic signal received on the background of additive white noise," *Izmer. Tekhn.*, No. 6, 42–45 (2004).
8. L. L. Myasnikov, "Speech sounds and their objective recognition," *Vestn. LGU*, No. 3, 9–17 (1946).
9. R. Potter, G. Kopp, and H. Green, *Visible Speech*, Van Nortrand, New York (1947).
10. K. H. Davies, R. Biddulph, and S. Balashek, "Automatic speech recognition of spoken digits," *J. Acoust. Soc. Am.*, **24** (6), 637–642 (1952).
11. V. S. Fajn, "The possibility of correlating the acoustic code of speech with its acoustic signal," *Akust. Zh.*, **9**, No. 3, 430–433 (1973).
12. V. V. Savchenko and D. Yu. Akat'ev, "An adaptive cluster model of the minimal sound units in speech analysis and voice recognition tasks," *Nauka Obraz. MGTU Baumana*, No. 2, <http://dx.doi.org/10.7463/0213.0527867/>, acc. 07.01.2017.
13. Yu. V. Andreev and M. V. Koroteev, "The chaotic nature of speech sounds," *Izvest. VUZ. Prikl. Nelin. Dinam.*, **12**, No. 6, 44–59 (2004).
14. E. V. Shatalov, S. V. Popov, and N. A. Dubovitskii, "Method for evaluating the clarity of speech received through the channel of intercommunication devices of self-contained respirators," *Izmer. Tekhn.*, No. 6, 61–63 (2004).
15. H. Ozer, I. Avcibas, B. Sankur, and N. Memon, "Steganalysis of audio based on audio quality metrics," *SPIE Electronic Imaging Conf. on Security and Watermarking of Multimedia Contents*, Santa Clara (2003), Vol. 5, pp. 55–66.
16. V. V. Savchenko and D. A. Ponomarev, "Automatic segmentation of stochastic time series using a whitening filter," *Optoelectron. Instrum. Data Proc.*, **45**, No. 1, 37–42 (2009).
17. S. Kullback, *Information Theory and Statistics*, Dover Publ., New York (1997).
18. V. V. Savchenko and A. V. Savchenko, "Information-theoretic analysis of efficiency of the phonetic encoding-decoding method in automatic speech recognition," *J. Comm. Tech. Electron.*, **61**, No. 4, 430–436 (2016).
19. G. Chen, S. N. Koh, and I. Y. Soon, "Enhanced Itakura measure incorporating masking properties of human auditory system," *Signal Proc.*, **83**, 1445–1456 (2003).
20. S. L. Marple, *Digital Spectral Analysis*, Prentice Hall, Englewood Cliffs NJ (1987).
21. V. V. Savchenko, "The minimax principle of entropy in problems of statistical solutions with limited observations," *Radiotekh. Elektr.*, **35**, No. 9, 1892–1899 (1990).
22. VOICE SELF-ANALYSIS Software, <https://sites.google.com/site/frompldcreators/produkty-1/voice-self-analysis/>, acc. 07.01.2017.