About the methods of forming a test signal in the instrumental evaluation of speaker clearance

V A Trushin, V E Khitsenko
Novosibirsk State Technical University, Novosibirsk, Russia

Abstract. This work presents the analysis of the possibility of creating a speech-like test signal for formant methods of assessing speech intelligibility based on the tables of syllables and words, random selection of speech elements from the database, the synthesis of sound files. Experimental research of the synthesized speech spectra, as well as its probability density, has been carried out. It is shown that the spectral and statistical characteristics of the synthesized speech-like test signals of the three voices of “speech choir” type coincide with the similar characteristics of real speech signals. Articulation tests of speech intelligibility by using synthesized speech-like interference with different signal-to-noise ratios showed the possibility of reducing the integral level of interference by 12 ÷ 15 dB in comparison with noise-like interference. There is discussed the direction of further research of improving the efficiency of generation of speech-like interference.
Keywords – speech intelligibility, speech-like signal, syllable and word bases, spectral probability density, distribution density.

1. Introduction

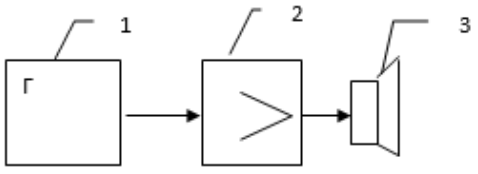
Various methods to assess the speech intelligibility are used and can be subdivided these methods into two large groups: subjective (expert) methods and objective (instrumental) methods.

The first ones are based on conducting articulation tests involving teams of announcers and respondents and using articulation tables of speech elements: syllables, words, phrases [1, 2].

Objective methods are based on measuring the objective parameters of a speech signal, while speakers and respondents are "replaced" by technical devices, which simulate real speech-forming paths and human hearing organs (for example, artificial voice, artificial ear) [2].

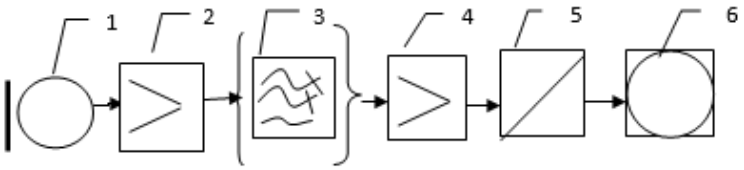
In Russia, formant intelligibility, based on the theory of the additive contribution of different frequency bands to the total speech intelligibility, is used to determine speech intelligibility. The differences between the particular methods are that different characteristics of speech and the forms of their interdependencies are chosen.

Below is shown a generalized block diagram of the implementation of formant methods of assessing speech intelligibility. The scheme includes a test signal generation device that simulates speech-forming paths and the measuring part that is the model of the human peripheral auditory system.



a) Test signal generation device

1 – Tone or white noise generator;
2 – Amplifier;
3 – Acoustic emitter.



b) Measuring part

1 – Primary converter;
2, 4 – Amplifiers;
3 – Set of bandpass filters;
5 – RMS detector;
6 – Display device with logarithmic calibration.

Figure 1. Generalized block diagram of implementation of formant method

Tonal signals corresponding to the middle of octave bands or white noise with a normal distribution of value probabilities are used in the accepted approach as a test signal (Figure 1a).
Such approximation is possible with dividing the frequency range into a large number of bands (for example, equal articulation), but in general, it is inadequate to the real speech signal.

2. Problem statement

In this regard a unified approach to the creation of test signal, formed on the basis of syllabic or verbal tables in accordance with State Standards [3, 4] is appropriate.
Articulation syllabic and verbal tables of State Standard 16600-72 and State Standard R 50840-95 are taken as a basis for creating a speech-like signal [3, 4]. For comparison, semantic texts are also used [5]. RNG Cripto-ServiceProvider method has been used as the algorithm of random selection of elements of speech from the corresponding base and Vocalizer as a synthesizer program. Audio files were recorded at a sampling rate of 44 KHz, 16 Bit Mono. For audio processing and obtaining spectra the program Adobe Audition 3.0 has been used. The signal was generated at the average speech level Ls=70 dB.

3. Theory and Experiment

As an example, we will consider test signals in the form of a speech choir based on syllables, words and semantic texts (two male and one female voices). The smoothed energy spectra of the signals are shown in Figure 2. It can be seen that the spectra are almost the same (the difference is 1-1.5 dB) and correspond to the average spectrum of the Russian language.

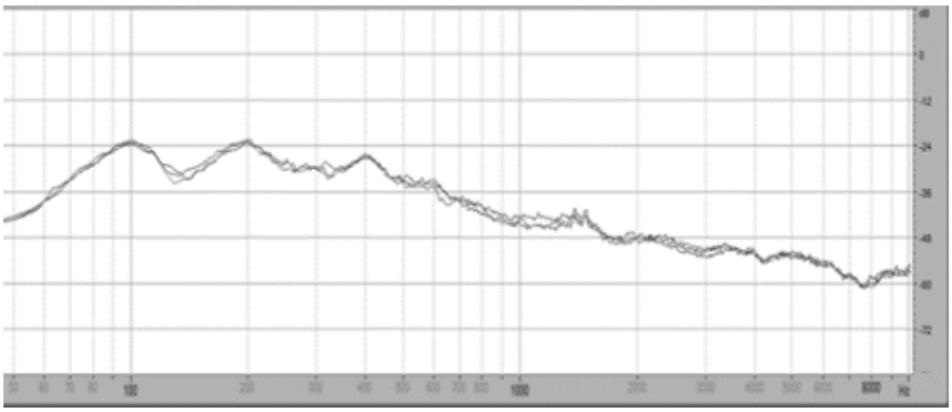


Figure 2. The energy spectra of the synthesized signals

An important statistical characteristic of speech is the probability density distribution of its values, which obeys the Laplace's law.
When the law of distribution of synthesized signals has been clarified, its significant instability in different segments, which is peculiar to speech signals, is revealed. In [6] it is recommended to maximize

the sample size to calculate entropy characteristics. The problem was overcome by random values mixing. Figure 3(a) shows a histogram with noticeable deviations from the Laplace distribution and asymmetry.

Indeed, the verification of the hypothesis of agreement with the Laplace distribution law by the $\chi 2$ criterion gives a negative result (Figure.3 (b)) in all attempts for different signal types and sample volumes. The reason is the asymmetry of the distributions, which can be the result of the imposition of external noise (the signal clearly possesses a composition of the distribution laws).

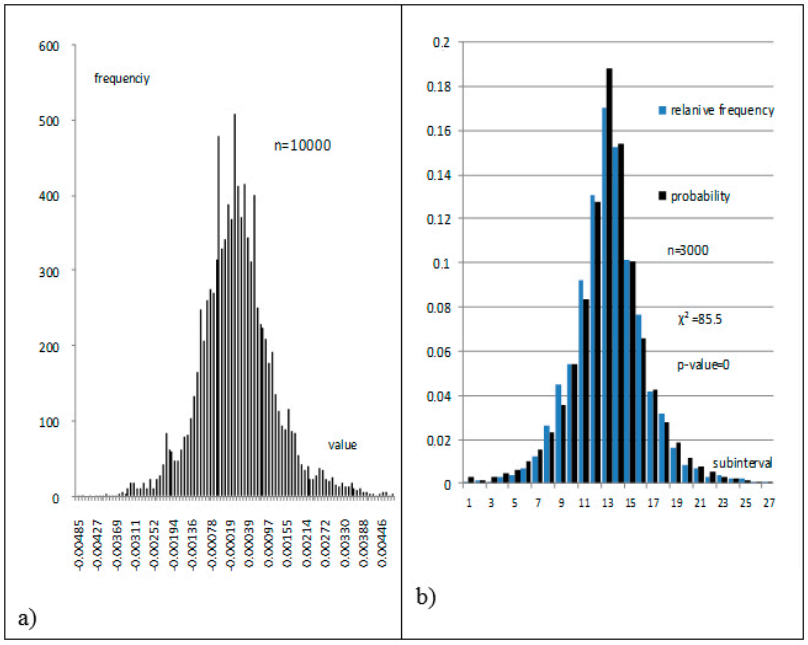


Figure 3. The frequency histogram constructed on 10000 values of a "word" signal (a) and result of check of a correspondence of a "syllable" signal with Laplace distribution (b)

However, the visual similarity with Laplace's law should be manifested in the proximity of the differential entropy estimates to the theoretical value for Laplace's law which equals to:

$$H(X) = \ln(e\sigma\sqrt{2}). \quad (1)$$

In [7,8], unbiased estimates of the so-called entropy coefficient of the distribution type H(X/σ) for the sample normalized by the standard deviation are proposed:

$$\hat{H}(X/\sigma) = -\sum_{i=1}^{k} \frac{n_i}{n} \ln \frac{n_i}{n} + \ln h, \quad (2)$$

where n is the sample size; n_i is the number of values in the i-th digit; k is the number of digits; h is the width of the digits. In this case, the number of digits k with asymmetry of the law is determined by the formula obtained by statistical tests [8].

Multiple estimation of the entropy coefficient of the synthesized signals has been carried out using samples with the volume of 3000 values. The results are summarized in table.1. The number of digits varied from 21 to 32 depending on the asymmetry and the interdecile sample range.

Table 1. Assessment of type distribution entropy coefficient

Type	$\hat{H}(X/\sigma)$	**95% confidence interval**	η
syllables	1.359	(1.357; 1.362)	0.942
words	1.373	(1.371; 1.379)	0.955
phrases	1.404	(1.403; 1.407)	0.985

The value of the entropy coefficient for the Laplace distribution according to (1) at σ =1 is 1.347. Thus, the asymmetry slightly enhances the entropy and increases during the transition from fragmented to continuous speech.

Comparing with the entropy coefficient of the normal distribution that equals to:

$$H_{normal}(X/\sigma) = \frac{1}{2}\ln(2\pi e) = 1.419 \quad (3)$$

we find the entropy coefficients of the signal quality relatively to the normal distribution:

$$\eta = \frac{\exp \hat{H}(X/\sigma)}{\exp H_{normal}(X/\sigma)}, \quad (4)$$

shown in Table 1.

However, this assessment of quality in this case is biased. There is a significant dependence of neighboring values, autocorrelation, which significantly reduces the differential entropy in real speech signals. Under these conditions, a multidimensional distribution law should be used to describe a random signal.

So for the sequence X = (X1, X2, ... Xn) of n values of a normal random signal with an autocorrelation matrix K, the exact value of the entropy coefficient of the type of distribution per one value is equal to:

$$H^{n}_{norm} = \ln[(2\pi e)^{n/2} \det K]/n. \quad (5)$$

It coincides with (3) in the absence of autocorrelation, when det K reaches a maximum value of one. For other distribution laws, there are no exact analytical expressions that take into account the correlation.

But the problem is not only in the presence of autocorrelation, but also in its instability. In other words, the correlation structure of the speech signal and det K change over time, and random mixing, turning the signal into white noise is unacceptable here as the speech-like disappears.

For an arbitrary joint law of distribution of a sequence, the following estimate of the differential entropy of a sequence X is proposed in [9] as:

$$H^{n}(X) = \sum_{i=1}^{n} \ln \sigma_{Xi} + \sum_{i=1}^{n} H(X_i/\sigma_{Xi}) + \frac{1}{2}\sum_{k=2}^{n} \ln(1 - R^2_{X_k/X_1X_2\ldots X_{k-1}}). \quad (6)$$

In relation to our problem, we can assume that the standard deviations σXi and the estimates of H (Xi / σXi) are constant, then:

$$H^{n}(X) = n(\ln \sigma_X + H(X/\sigma_X)) + \frac{1}{2}\sum_{k=2}^{n} \ln(1 - R^2_{X_k/X_1X_2\ldots X_{k-1}}), \quad (7)$$

where $R^2_{X_k/X_1X_2\ldots X_{k-1}}$ - the coefficients of determination of autoregressive models of communication of the k-th signal value with all previous ones. Essentially, these coefficients reflect the prediction accuracy of the X_k value by model

$$X_k \approx \alpha_1 X_{k-1} + \alpha_2 X_{k-2} + \ldots + \alpha_k X_1 \quad (8)$$

and represent the square of the correlation coefficient between the exact and predicted values of the signal.

The parameters of the model α_k are determined from the autocorrelation function of some fragments of a pre-processed speech signal [10]. Thus, these estimates of the entropy characteristics are unstable.

Nevertheless, the possibility of selecting and using the dynamics of their change, as an objective characteristic of a speech-like test signal, is not excluded.

In this regard, it is interesting to use the autoregressive speech signal models to identify a speaker in the article [11] and other publications by Bryukhomitsky Y.A., which can be used to obtain objective methods for assessing legibility.

While creating a speech-like test signal it is also important to take into account the peculiarities of the formation of the continuous speech flow and the possibility of modern technologies of reconstruction of speech signals [9].

4. Conclusion

In the approach proposed, either the tone signals corresponding to the middle of the octave bands, or white noise with a normal probability distribution of values, are used as a test signal (Figure 1a). This approximation is possible with dividing the frequency range into a large number of bands (for example, equal-articulation), but in general is inadequate concerning the real speech signal.

5. References

[1] Rashevskiy Y.I., Kargashin V.L, “Review of foreign methods of determining speech intelligibility”, *Special technique*. 2002 , No. 4, pp 37–46
[2] Pokrovskiy N.B., “Calculation and measurement of speech intelligibility”, *Svyazizdat*, 1962 392 p
[3] State Standard R50840-95. Speech transmission on communication paths. Methods of quality assessment, intelligibility, recognition. Moscow, *Gosstandart*, 1995,
[4] State Standard 16600-72. Speech transmission on the radio telephone line paths. Requirements for speech intelligibility and methods of articulation tests. Moscow, *Gosstandart*, 1972
[5] Batsula A.P.,Trushin V.A.,Ivanov A.V.,Reva I.L, “On the reliability assessment of speech information security from leak on technical channels”, *Reports of Tomsk state University of control systems and Radioelectronics*, 2010. No. 1 (21) pp 89–93
[6] Kropotov Y.A. “One-dimensional probability density model of speech signals”, *Control, communication and security systems*, 2015. No. 4, pp 158–170
[7] Kozlachkov S B, Dvoryankin S V, Bonch-Bruevich A M, “Principles of formation of test speech signals in assessing the effectiveness of noise cleaning technologies”, *Cybersecurity issues*, No. 3(27), pp 9
[8] Tyrsin A N, “Entropy modeling of multidimensional stochastic systems”, *Scientific book*, 2016, 156 p
[9] Tyrsin A N, Klyavin I.A. “Improving the accuracy of entropy estimation of random experimental data”, *Management systems and information technology*, 2010, No. 1(39), pp 87–90
[10] Rabiner L.R., Shafer R.V. “Digital processing of speech signals”, *Radio and communication*, 1981. 495 p
[11] Bryukhomitskiy Y A Immunological approach to identification by dynamic biometric parameters, *Izvestiya of Southern Federal University. Technical science*, 2017. No. 5 (190), pp 56–66