ARCHIVES OF ACOUSTICS
13. 3 4. 221–234 (1988)

STATISTIC MODEL OF SPEECH-LIKE SIGNAL GENERATION([1])

PAWEŁ BIEŃKOWSKI, WOJCIECH MYŚLECKI

Institute of Telecommunication and Acoustics, Technical University of Wrocław
(50–317 Wrocław, ul. B. Prusa 53/55)

This paper presents a parametric model of a generator of a signal with probabilistic characteristics of a speech signal. Optimal parameters of the model have been selected for a case of a Polish speech signal and research results of the model in the form of a computer programme are given.

W pracy przedstawiono parametryczny model generatora sygnału o cechach probabilistycznych zbliżonych do wyróżnionych cech probabilistycznych sygnału mowy. Dobrano optymalne parametry modelu dla przypadku sygnału mowy polskiej oraz przedstawiono wyniki badań modelu zrealizowanego w postaci programu na EMC.

1. Introduction

The authors of this paper attempt to determine generation rules of signal with chosen statistic distributions of parameters close to distributions occuring in a speech signal. Such synthetic signals find application in studies concerned with mechanisms of speech signal perception and with speech signal transmission in acoustic and telecommunication channels. A model of generation of a signal with statistic characteristics significant from the point of view of speech transmission quality in telecommunication channels is presented on the basis of own research and research results given in papers [3, 9, 10].

This model includes only long-term probabilistic characteristics of the speech signal. Hence, the speech signal is a realization of a stationary stochastic process, which is an ergodic process, as well. According to the ergodicity definition probabilistic characteristics can be obtained from a singular realization in time $t \to \infty$. In practice we have a time-limited process realization in speech signal

([1]) Research performed within problem CPBP O2.03., 9.3.

analysis. Therefore, the notion of weak stationarity was used in this paper. The minimal time after which probabilistic characteristics of the signal do not change significantly was called the minimal time interval of signal stationarity or minimal time interval of signal stationarity with respect to a definite characteristic. Stated above assumptions concern research referred to in the paper as well as studies performed by the authors.

2. Theoretical foundations of the model

The shape of the probability density function of instantaneous amplitudes is one of the basical probabilistic characteristics of a speech signal. It distinguishes clearly the speech signal from other stochastic processes. The probability density function $p(x)$ of a random signal $x(t)$ determines the probability of the event that the signal's values in an arbitrary moment are contained in a definite interval. Probability density functions of instantaneous amplitudes of a speech signal, measured for various languages, show a great consistence of shape. Fig. 1 presents density distributions for 3 languages: English (according to DAVENPORT [2]), Russian (according to SZITOW and BIELKIN [5]), Polish (according to BRACHMAŃSKI and MAJEWSKI [1]).

Two parts can be distinguished in the curve which represents density distribution of instantaneous amplitudes of a speech signal. The first part characterizes the distribution for great amplitude values and can be approximated with the Laplace distribution; while the second one, representing low amplitudes, has the shape of a sharp vertex and is described by a normal, Laplace or delta distribution. The physical interpretation of the shape of the amplitude distribution curve is given by DAVENPORT [2]. The shape vertex of the distribution curve can result from the number of pauses in speech of which the character of the signal is determined by the reverberation and noises in the room, and channel's noises. A different interpretation was presented by RIMSKIJ-KORSAKOW [9]. He assumed that the achieved distribution differs greatly from the normal distribution, because speech, as well as music, is a complex random process. Apart from fluctuations of signal's phrases and amplitudes, related with the incoherence of separate sources of vibrations and transitions between individual sounds (what should give a normal distribution), also relatively slow variance changes occur. In the case of speech they result from a prosodic modulation of voice strength and modulation, stress and tempo, influence of the expiration process etc. Rimskij-Korsakow's hypothesis supplemented with WOLF's consideration [10] can be presented in analytic form as follows.

Let us assume that the amplitude distribution is normal but with dispersion values s varying in chosen time segments of the speech signal:

$$p_i(x;\, s_i) = \frac{1}{\sqrt{2\pi}\, s_i} \exp\left(-\frac{x^2}{2s_i^2}\right). \tag{1}$$

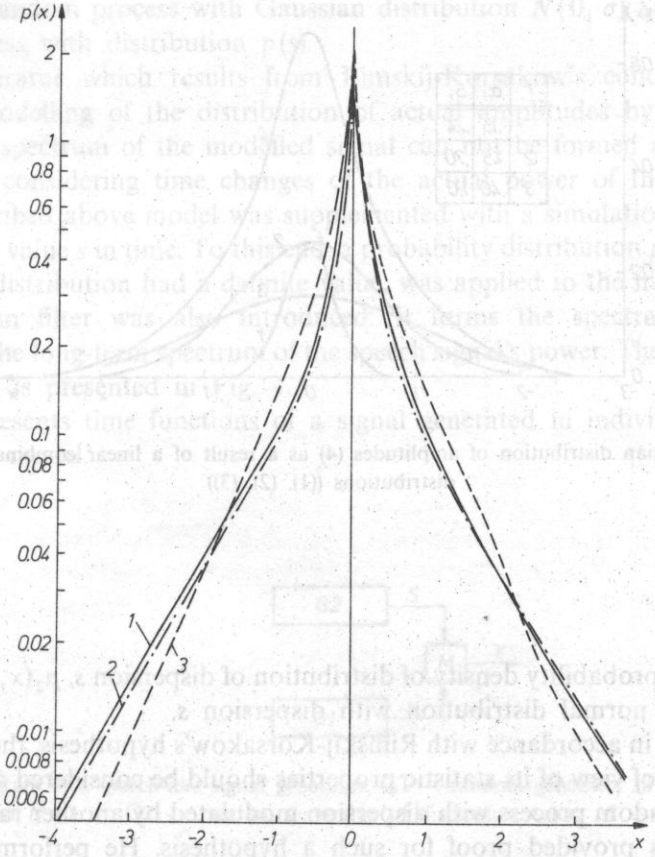


Fig. 1. Probability density distribution of actual values of a speed signal. According to: 1 — DAVENPORT, 2 — SZITOW and BIELKIN, 3 — BRACHMAŃSKI and MAJEWSKI

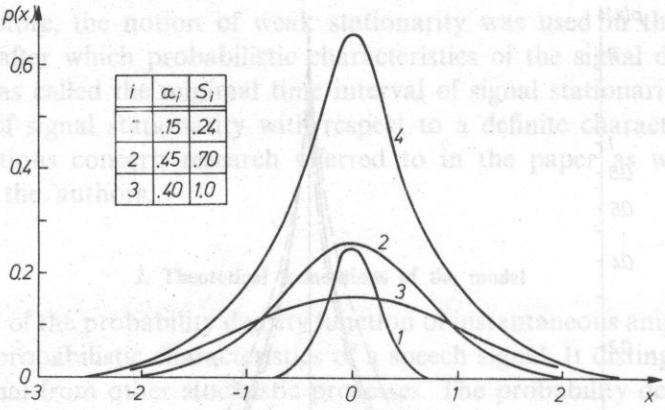
A weighted sum of these distributions

$$p(x) = \sum_{i=1}^{n} \alpha_i p_i(x; s_i); \sum \alpha_i = 1 \tag{2}$$

gives a distribution different from Gaussian (Fig. 2).

In the case of constant change of dispersion s the total density distribution is described by formula:

$$p(x) = \int_0^{\infty} p_1(s) p_2(x, s) ds, \tag{3}$$



i	α_i	S_i
1	.15	.24
2	.45	.70
3	.40	1.0

Fig. 2. Non-Gaussian distribution of amplitudes (4) as a result of a linear combination of Gaussian distributions ((1), (2), (3))

where $p_1(s)$ – probability density of distribution of dispersion s, $p_2(x, s)$ – probability density of normal distribution with dispersion s.

Therefore, in accordance with Rimskij-Korsakow's hypothesis, the speech signal from the point of view of its statistic properties should be considered as a realization of a normal random process with dispersion modulated by another random process. INBLIN [3] has provided proof for such a hypothesis. He performed theoretical analysis and experimental verification of two excluding one another hypotheses. In the first hypothesis he assumed the speech signal to be a simple exponential random process with time-dependent dispersion: while in the second hypothesis speech was to be a normal random process with varying in time dispersion. The author proved the second hypothesis to be true on the basis of an analysis of the excess coefficient of density distributions in following time intervals of the signal. He also found that probabilities of instantaneous values of the speech signal are described by a normal distribution with constant dispersion in time intervals of about 15–30 ms. These results confirm the Rimskij-Korsakow hypothesis. Inblin's analysis has probabilistic sense and thus it can not be accepted that the distribution of amplitudes is a normal distribution in every time interval of 15–30 ms.

A model of signal generation according to rules can be proposed on the basis of Rimskij-Korsakow's conception. Such a model is shown in Fig. 3.

The following relationship describes it:

$$Y = X \cdot S, \tag{4}$$

where X – random process with Gaussian distribution $N(0, \sigma)$ S – modulating random process with distribution $p(s)$.

The generator which results from Rimskij-Korsakow's conception enables parametric modelling of the distribution of actual amplitudes by changing $p(s)$. Whereas, the spectrum of the modelled signal can not be formed and there is no possibility of considering time changes of the actual power of the signal.

The described above model was supplemented with a simulation of changes of the dispersion value s in time. To this end, a probability distribution $p(T)$ of the time, in which the distribution had a definite value, was applied to the modulated signal S. A Gaussian filter was also introduced. It forms the spectral characteristic according to the long-term spectrum of the speech signal's power. The block diagram of this model is presented in Fig. 4.

Fig. 5 presents time functions of a signal generated in individual functional blocks of the model.

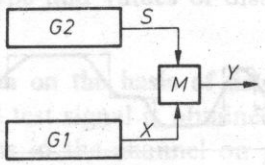


Fig. 3. Block diagram of a speech-like signal generator: G1 – random generator of noise with normal distribution N (0, σ); G2 – random generator of dispersion value, M – modulator

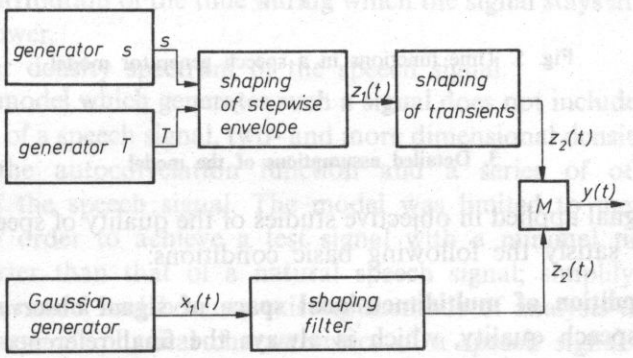


Fig. 4. Block diagram of a signal generator

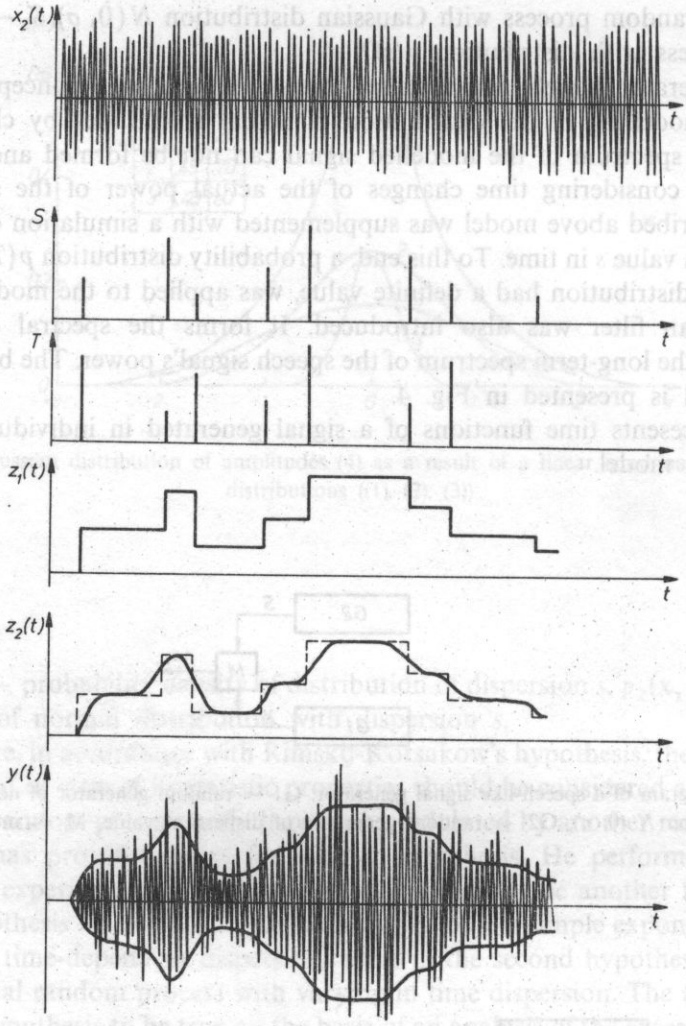


Fig. 5. Time functions in a speech generator model

3. Detailed assumptions of the model

The test signal applied in objective studies of the quality of speech transmission systems should satisfy the following basic conditions:

C1 — Condition of multidimensional space of signal observation. Subjective assessment of speech quality, which is always the final reference measure of an objective quality estimator, is a function defined in a multidimensional space of auditory perception where a numerous set of physical characteristics of the speech signal is analysed [7]. Hence, the test signal has to undergo multidimensional

analysis in order to determine an objective estimator of transmission quality. The following relationship describes the process of determining such an estimator

$$w \leftrightarrow \tilde{w} = D\{d^n[y(t),\ y^*(t)]\}, \tag{6}$$

where: w – reference measure of transmission quality subjective measurements; $\tilde{w}$ – objective estimator of quality measure, d^n – distance in a n-dimensional space of signal observation between the transmitted, $y(t)$, and received, $y^*(t)$, test signal, D operator of transformation of d^n into w.

C2 – Condition of the ability to respond to disturbances and distortions of the channel. Possibly to the same extent as a signal of natural speech.

C3 – Condition of convergence of statistic spectral and time characteristics with natural speech characteristics.

This condition has to be satisfied in order to ensure work conditions of the channel close to conditions existing during a normal transmission of a speech signal, because it strongly influences the type and values of disturbances and distortions occuring in the channel.

C4 – Condition of signal generation on the basis of determined rules. When this condition is fulfilled then a standard test signal is obtained and the effect of certain types of disturbances and distortions in the channel on parameters of the output signal can be analysed mathematically.

The signal described in paragraph 2 satisfies condition C1 (i.e. it can be presented in a multidimensional space of observation), because it has several statistical characteristics of a speech signal which can be expressed by:

a) distribution of probability density of instantaneous values,
b) distribution of probability density of short-term average power of the speech signal,
c) probability distribution of the time during which the signal stays at the same level of average power,
d) average power density spectrum of the speech signal.

However, a model which generates such a signal does not include the density of zero – crossings of a speech signal, two- and more dimensional density distributions of amplitudes, the autocorrelation function and a series of other structural characteristics of the speech signal. The model was limited to mentioned above characteristics in order to achieve a test signal with a minimal time interval of stationarity, shorter than that of a natural speech signal; simplify the technical realization of the model and because existing methods of analysis do not make it possible to measure all physical characteristics of a speech signal.

In accordance with condition **C4** the presented signal is generated on the basis of determined rules. These rules can be accomodated to any language, while detailed parameters of the model have to be selected individually.

4. Measurements

Parameters which would ensure convergence of the characteristics of generated signal with the characteristics of the Polish speech signal are necessary to build the model. To this end we have to determine: distributions $p(s)$ and $p(T)$, frequency characteristic of the filter and to elaborate the method of determining function $z_2(t)$ on the basis of function $z_1(t)$ (see Fig. 4).

It was accepted that the distribution of probability density of the generated signal's amplitudes should be consistent with the density distribution of instantaneous amplitudes of the Polish speech signal given by BRACHMAŃSKI and MAJEWSKI [1]; while the average density spectrum of the power of the Polish speech signal should be approximated on the basis of results presented in papers [4, 6, 11]. No data on the distribution of probability density of short-term power $p(s)$ and the probability distribution $p(T)$ of the time the signal stays on the same level of average power can be found in literature. Therefore, suitable experimental studies were performed.

4.1. Research methods

A 5 –minute newspaper text read in turn by 20 speakers with a constant average level of sound intensity was the phonetic material. The recording was done in the recording room of the ITA at the Technical University of Wrocław. A UM57 Neumann condenser microphone and M601 SD ZRK tape recorder were used. Irregularities of the microphone's response characteristic did not exceed $\pm$ 2 dB in the band $30 \div 15000$ Hz; whereas irregularities of the frequency characteristic of the tape recorder's record-play channel did not exceed $\pm$ 2 dB in the band $80 \div 10000$ Hz. The microphone was placed near the speaker's mouth (in the near field).

The system shown in Fig. 6 was used for measurements of distributions of probability density $p(s)$ and $p(T)$.

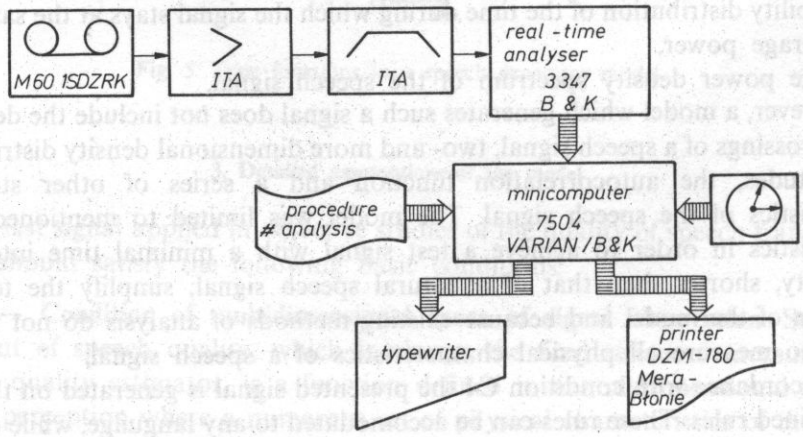


Fig. 6. System for measuring $p(s)$ and $p(T)$ characteristics

The analysed speech signal recorded on a tape recorder was subjected to band-pass filtration after being amplified what reduced the signal's spectrum to a band 200 Hz – 5000 Hz (the band deciding about the information content of the speech signal). The rms value of the signal was detected in the entire band in a 3347 Bruel-Kjaer synchronous analyser. So obtained rms values were sampled at 20 ms intervals and classified in 26 ranges; 2 dB each. At the same time events that following samples of the rms value were contained within the same 2 dB range were counted and classified.

The selection of the time-constant of the rms value detector should be discussed separately. In accordance with the Rimskij-Korsakow conception, on which the proposed model is based, the distribution of the probability density of the rms value of following speech segments had to be determined. INBLIN [3] has determined lengths of these segments as equal to 15–30 ms. The length of the segment was accepted as equal to 20 ms, so the applied time-constant of the rms value detector has to be equal to $\tau = 10$ ms.

The total measuring error of the rms value of a random signal depends on errors of the instrument and statistical inaccuracy due to a finite time of averaging. The total error of the 3347 analyser does not exceed $\pm$ 0.6 dB. The estimation of the statistical error is expressed by the relationship

$$\varepsilon = \frac{1}{2\sqrt{BT}}, \tag{6}$$

where B – width of analysing filter, T – effective time of averaging.
A 68% confidence level was determined from relationship (6) when the condition $BT > 15$ satisfied. For used in measurements values: $B = 5$ kHz and $T = 20$ ms, the reading error is negligible with respect to the observation resolution of the p (s) distribution accepted at 2 dB.

4.2. *Results*

Measurements of the p (s) distribution were carried out for speech signal intervals of 5, 3, 1 min. Statistic analysis has proved that a 3 minute interval of a speech signal is sufficient to determine the $p(s)$ distribution. $p(s)$ distributions for following voices were normalized in relation to the average value and then were averaged. Fig. 7 presents the distribution of probability density of short-term average power p (s), averaged for 20 men's voices.

The direct application of the achieved p (s) distribution in the model is possible, but that would require a very complicated random-number generator. Therefore, it had to be approximated by a different distribution with known analytical form which would be more simple to generate. The application of the analytical p (s) distribution also makes it possible to perform a mathematical analysis of the effect

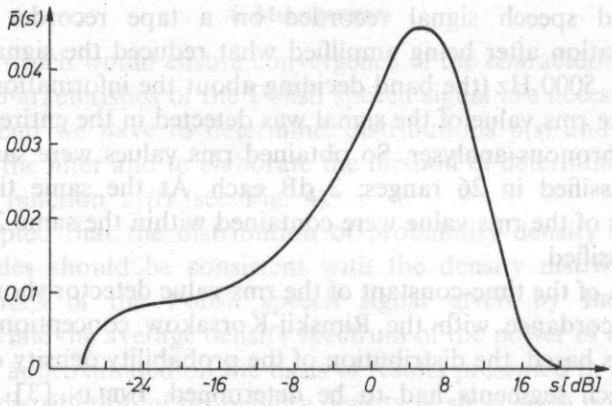


Fig. 7. Distribution of the average probability density of short-term average power of a Polish speech signal

this distribution has on the distribution of the probability density of amplitudes of the output signal. Two types of approximating distributions were considered: Rayleigh's distribution and gamma distribution. The probability density of the output signal's amplitudes for the Rayleigh's distribution in the form

$$p(s) = \frac{s}{\sigma_0^2} \exp\left(-\frac{s^2}{2\sigma_0}\right) \tag{7}$$

achieved from relationship (4) – is expressed by Laplace's distribution

$$p(x) = \frac{1}{2\sigma_0} \exp\left(-\frac{|x|}{\sigma_0}\right). \tag{8}$$

This distribution is frequently used to approximate amplitude densities of a speech signal. The parameter σ_0 in relationship (8) was determined from the approximation of the curve of amplitude density of Polish speech [1] by the Laplace's distribution. The correlation coefficient was used as the similarity measure of these curves. Because the segment corresponding to greater values of amplitudes is the most significant part of the curve of speech amplitude's density, the correlation coefficient was maximized in the interval $0.3 < |x| < 4$. The point 0.3 was determined as shown in Fig. 8. The maximal value of the correlation coefficient ($R_{\max} = 0.986$) was achieved for $\sigma_0 = 0.74$; thereby determining the value of the σ_0 parameter for Rayleigh's distribution.

The possibility of applying the gamma distribution in the following form

$$p(x) = \frac{a^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-ax}, \quad a > 0, x > 0, \alpha > 0, \tag{9}$$

where $\Gamma(\alpha)$ – Euler's gamma function, α, a – parameters of the distribution, was considered.

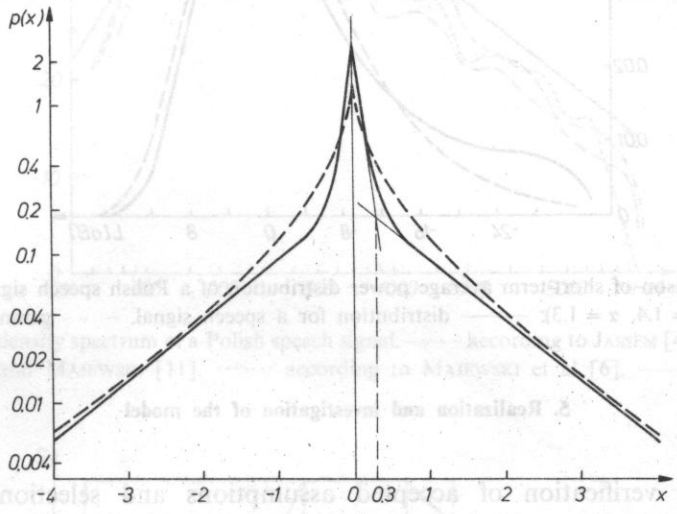


Fig. 8. A comparison of amplitude density distribution of Polish speech with a distribution resulting from the application of the gamma distribution to the signal model; ——— distribution for a speech signal, – – – – theoretical distribution

The analytical form of the probability distribution of the output signal was not found, because an integral expression was impossible to solve. It resulted from numerical integration that the maximum value of the correlation coefficient at parameters of gamma distribution equal to $a = 1.4$ and $\alpha = 1.3$ was found between the curve of density distribution for Polish speech and the curve of probability density of the product of the gamma and normal distribution $N(0.1)$. The approximation with gamma distribution was proved to be better, because the correlation coefficient R_{max} was equal to 0.996 (for $0.3 < |x| < 4$). This approximation is shown in Figs. 8 and 9.

From the above considerations it was accepted that the generation of the short-term average power distribution will be done on the basis of the gamma distribution.

The distribution $p(T)$ of the time a signal stayed on the same level of average power was measured at the same time. Achieved distributions were averaged on a set of 10 speakers and approximated by an unilateral normal distribution. The best approximation was obtained at the value of 22.5 of the σ_0 parameter of this distribution.

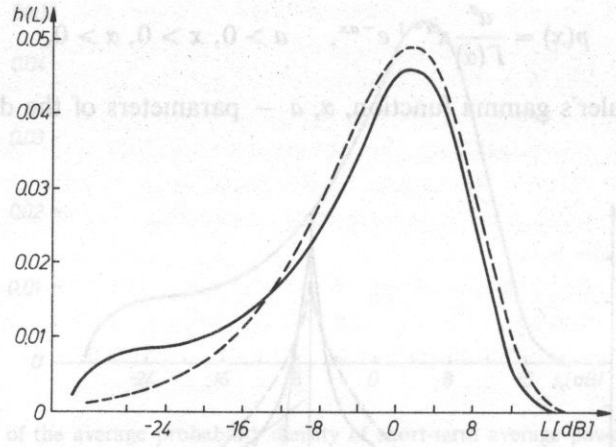


Fig. 9. A comparison of short-term average power distribution of a Polish speech signal with gamma distribution ($a = 1.4$, $\alpha = 1.3$); ——— distribution for a speech signal, – – – gamma distribution

5. Realization and investigation of the model

The final verification of accepted assumptions and selection of model's parameters can be done only on the basis of experimental studies. To this end a numeric parametric model of a signal generator was built in the form of a computer program. Individual blocks of this generator (see Fig. 4) were realized as follows:

"Generator T" and "Gaussian generator" were realized on the basis of the tribution was applied in accordance with conclusions from paragraph 4. Presented in paper [12] algorithm of generation was used.

"Generator T and "Gaussian generator" were realized on the basis of the FPMCRW generator found in the Fortran library of the Odra 1300 system [14].

"Spectrum forming filter" – the power density spectrum of a Polish speech signal was approximated by straight line segments with a slope of 12 and 6 dB/octave, as it is shown in Fig. 10. Connected in a cascade, a low- and a high-pass Butterworth's filter, respectively, were used in the program. "Shaping of a stepwise envelope "and" "shaping of transients" due to a lack of adequate experimental research and the simplicity of the technical realization, the $z_2(t)$ function was formed with a linear approximation between successive values of the $z_1(t)$ function; this is shown in Fig. 11.

The frequency of signal sampling of $f_p = 8$ kHz was accepted in calculations.

Before the experimental work on the optimization of the model was done, the minimum time interval of stationarity of the generated signal was estimated. Hence, the problem was to find how many samples have to be generated by generators $p(T)$ and $p(s)$ in order to stabilize analysed distributions. The minimum time interval of

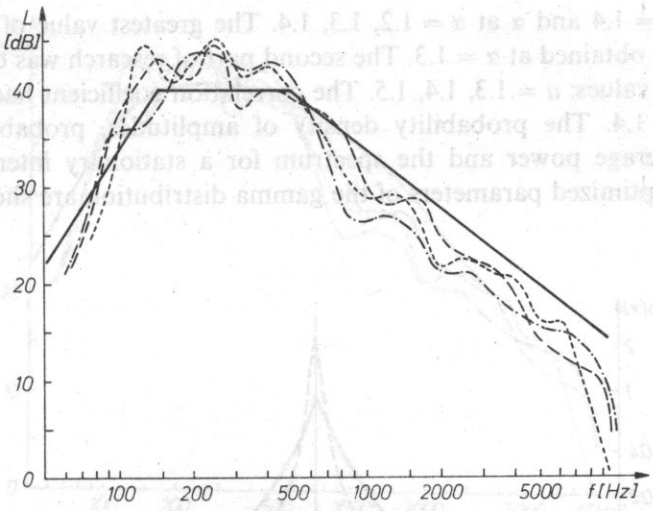


Fig. 10. Power density spectrum of a Polish speech signal. – – – according to JASSEM [4], –·–·– according to ZALEWSKI and MAJEWSKI [11], ····· according to MAJEWSKI et al [6], ——— approximation

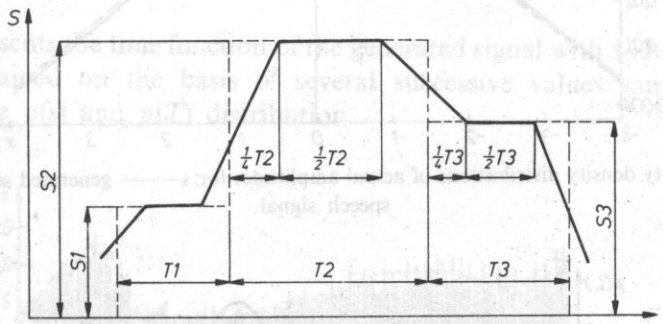


Fig. 11. Shape of transients; T1, T2, T3 – successive values from time generator ($p(T)$ distribution); S1, S2, S3 – successive values from level generator (p (s) distribution)

stationarity of the generated signal was determined at about 30 seconds on the basis of the properties' analysis of applied generators of $p(s)$ and $p(T)$ distributions.

Because calculations of a representative signal interval are very time-consuming, a single program of dynamic, sequential optimization [8] was applied. The correlation coefficient between the probability density curve of instantaneous amplitudes of the generated signal and of natural speech in the range of values of amplitudes $0.3 < |x| < 4$ was accepted as the objective function. The number of investigated parameters was reduced to two, which have the greatest effect on the objective function: parameter α and parameter a of the gamma distribution. In the first part of investigations three 30 s stationary intervals of a signal were generated. The earlier set central value of the parameter of the gamma distribution was

accepted at $a = 1.4$ and α at $\alpha = 1.2, 1.3, 1.4$. The greatest value of the correlation coefficient was obtained at $\alpha = 1.3$. The second part of research was done for $\alpha = 1.3$ and following values: $a = 1.3, 1.4, 1.5$. The correlation coefficient had the maximum value for $a = 1.4$. The probability density of amplitudes, probability density of short-term average power and the spectrum for a stationary interval of a signal generated at optimized parameters of the gamma distribution are shown in Figs. 12, 13, 14.

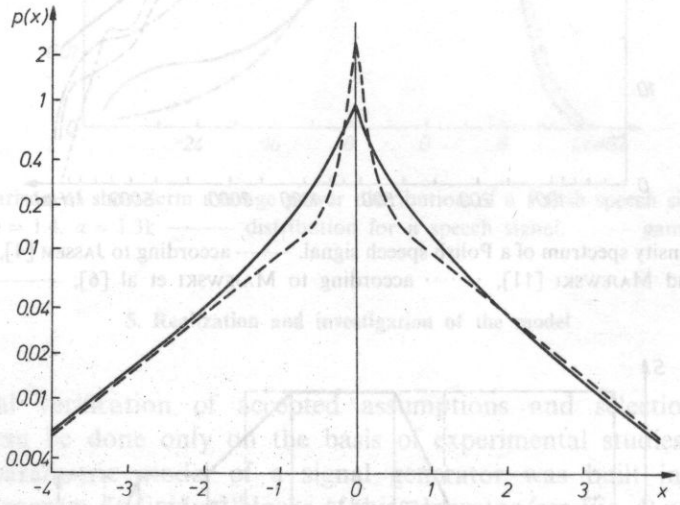


Fig. 12. Probability density distributions of actual amplitudes for: ——— generated signal, ----- Polish speech signal

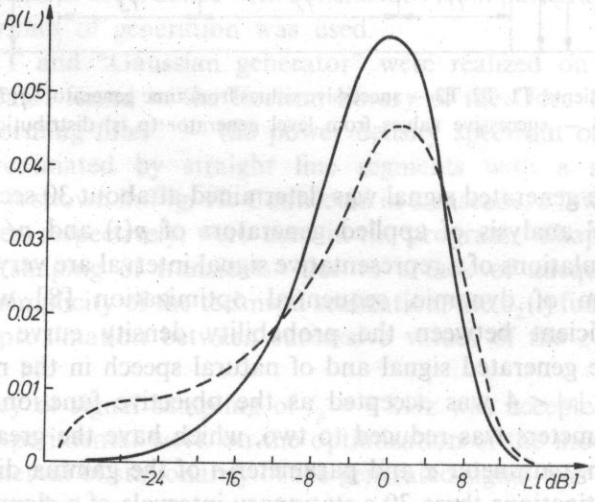


Fig. 13. Probability density distributions of short-term average power of a generated signal ——— and Polish speech signal -----

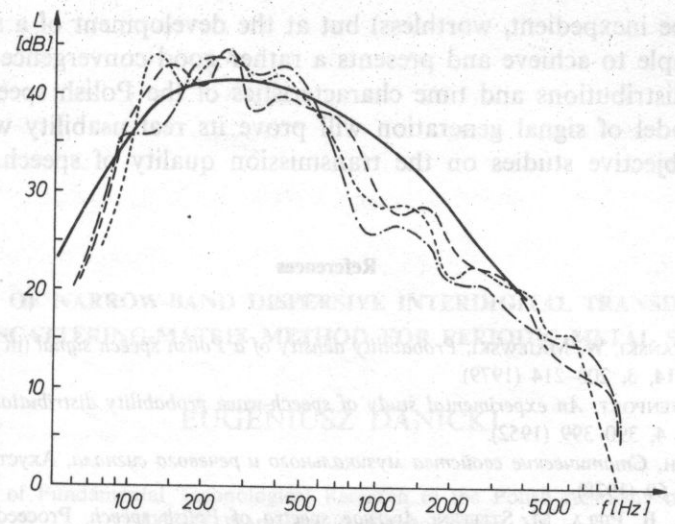


Fig. 14. Power density spectrum of a generated signal ——— and Polish speech signal – – – – according to JASSEM; –·–·– according to ZALEWSKI and MAJEWSKI; ····· according to MAJEWSKI et al.

Fig. 15 presents the time function of the generated signal with 140 mm duration which was obtained on the basis of several successive values sampled by the generator of the $p(s)$ and $p(T)$ distribution.

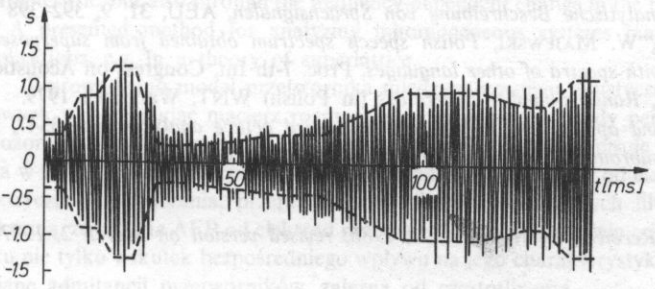


Fig. 15. A time function fragment of generated signal

6. Conclusion

A comparison between probability distribution and characteristics of the generated signal (paragraph 5) and corresponding distributions and characteristics of Polish speech demonstrates a lack of their full compatibility. Yet, our studies were not aimed at the development of a signal with characteristics of the speech signal

(this would be inexpedient, worthless) but at the development of a signal which is relatively simple to achieve and presents a rather good convergence with accepted probability distributions and time characteristics of the Polish speech signal. The proposed model of signal generation will prove its real usability when it will be applied in objective studies on the transmission quality of speech.

References

[1] S. Brachmański, W. Majewski, *Probability density of a Polish speech signal* (in Polish) Archiwum Akustyki, **14**, 3, 203–214 (1979).

[2] W. B. Davenport, *An experimental study of speech-wave probability distributions*, J. Acoust. Soc. Amer., **24**, 4, 390–399 (1952).

[3] Я. Инблин, *Статические свойства музикального и речевого сигнала*, Акустический Журнал, **25**, 5, 693–69 (1978).

[4] W. Jassem, B. Piela, M. Steffen, *Average spectra of Polish speech*, Proceedings of Vibration Problems, **2**, 59–71 1959.

[5] П. Г. Капотов, Ю. П. Максимов, В. В. Марков. *Нелинейные искажения речевого сигнала*, Электросвязь, **8**, 29–34 (1976).

[6] W. Majewski, H. B. Rothman, H. Hollien, *Acoustic comparisons of American English and Polish*, Journal of Phonetics, **5**, 247–251 (1977).

[7] B. J. Mc Dermott, *Multidimensional analyses of circuit quality judgements*, J. Acoust. Soc. Amer., **45**, 774–781 (1969).

[8] Z. Polański, *Contemporary methods of experimental studies* (in Polish) WP, Warszawa 1978.

[9] А. В. Римский-Корсаков, *Статические свойства радиовещательного сигнала*, Акустический Журнал, **6**, 3, 360–369 (1960).

[10] D. Wolf, *Analytische Beschreibung von Sprachsignalen*, AEU, **31**, 9, 392–398 (1977).

[11] J. Zalewski, W. Majewski, *Polish speech spectrum obtained from superposed samples and its comparison with spectra of other languages*, Proc. 7-th Int. Congress on Acoustics, Budapest 1971.

[12] R. Zieliński, *Random-number generators* (in Polish) WNT, Warszawa 1979.

[13] *Instruction and applications of the real time third octave analyzer type* 3347, Bruel-Kjaer, 1971.

[14] *Fortran – subroutine library* (in Polish) Zeszyty Elwro, nr 13044 2.

Received on November 2, 1986; revised version on March 2, 1987.