

# A method for the asynchronous analysis of a voice source based on a two-Level autoregressive model of speech signal

Vladimir Vasilyevich Savchenko  · Lyudmila Vasilyevna Savchenko 

Received: 14 December 2023 / Accepted: 20 February 2024 / Published online: 15 July 2024

© Springer Science+Business Media, LLC, part of Springer Nature 2024

## Abstract

We consider the problem of analysis of the voice source of speech within the range of short-term observations. The problem of insufficient speed of the available methods for the analysis of voice source is described, regardless of the method of data preparation: either synchronous with the main tone of speech sounds or asynchronous. We propose a method for the analysis of voice sources based on the two-level autoregressive model of the speech signal. We describe a software realization of the developed method based on the Berg-Levinson high-speed procedure of numerical calculations. It is shown that this procedure is characterized by a relatively low level of computation costs and its application does not require synchronization of the sequence of observations with the main tone of speech signal. With the help of software implementation of the proposed method, we designed and performed full-scale experiment aimed at analyzing the vowel sounds in the speech of a reference speaker. The results of this experiment confirmed the elevated speed of the proposed method and enabled us to formulate the requirements to the duration of speech signal for the real-time voice analysis. Thus, the optimal duration of the speech signal should vary within the range 32–128 msec. The obtained results can be used for the development and investigation of digital speech communication systems, systems of voice control, biometrics, biomedicine and other speech systems in which specific voice features of speaker's speech are of primary importance.

**Keywords** Speech acoustics · Speech signal · Speech analysis · Voice analysis · Vocal tract · Fundamental tone · Fundamental tone frequency

## Introduction

A voice source is understood by experts as a signal of excitation of acoustic vibration in the vocal tract of the speaker in the production of voiced speech sounds, especially vowels [1–3]. As an object of acoustic analysis, voice sources serve as interesting objects for the researchers in various fields of activity: from the digital speech processing and synthesis to biomedical systems and technologies [3–6]. The general system problem of these and similar studies is connected with the time instability of the fine structure of speech signal under the influence of numerous random (uncontrollable) factors [7, 8]. The general problem explains and, at the same time, stimulates the development of various theories and experimental tools aimed the analysis of voice speech [9–11].

The aim of the present paper is to develop a rapid method that can be used for the real-time voice analysis.

Translated from *Izmeritel'naya Tekhnika*, Vol. 73, No. 2, pp. 55–62, February, 2024. Russian DOI: <https://doi.org/10.32446/0368-1025it.2024-2-63-72>



## Analysis of voice sources

The acoustic theory of speech formation [12] and its model of a voice source in the form of a quasiperiodic (or periodic for bounded time intervals) sequence of excitation pulses of the vocal tract of a speaker [13] are now extensively used in the field of information speech technologies. The parameters of the indicated sequence, namely, the repetition rate  $F_0$  and the shape of pulses, specify the fundamental tone and the fine structure of vocalized segments of the speech signal. Various values of the parameters correspond to different speech sounds and the indicated correspondence has a strictly individual (speaker-dependent) character. Therefore, the analysis of vocal source is both a nontrivial problem and an urgent task [14–16].

As a widespread procedure used to solve the posed problem, we can mention voice inverse filtering of the speech signal [17, 18]. The idea of filtering is to decompose the model of the observed signal into two independent components, namely, the voice source and the vocal tract [19]. In this case, the vocal tract is modeled by a linear recursive filter of relatively low order  $p_1 = 8 \dots 12$  [12, 20]. There are two known approaches to the realization of this decomposition [21, 22]. They differ by the procedure of acoustic measurements.

In the first approach, the sequence of observations (readings) of the speech signal is synchronized with its fundamental tone [12, 16]. In this case, the frequency of the fundamental tone  $F_0$  is regarded as *a priori* specified or preliminarily measured, and the shape of excitation pulses is computed within the period of synchronous observations with duration  $\tau = T_0$  of a single period  $T_0 = F_0^{-1}$  of the fundamental tone. However, under the conditions of *a priori* uncertainty, this task is practically unsolvable, at least for the real-time voice analysis of speech [22–24].

The second approach is based on simulation of speech signals in the frequency region [25]. In particular, the autoregressive model [13, 15] is used fairly extensively. This model is based on the description of stationary segments of speech signals of relatively large length  $\tau \gg T_0$  and, therefore, does not require synchronization with the fundamental tone. In this case, we encounter the problem of proper choice of the order  $p$  of the autoregressive model [20]. Under the conditions of *a priori* uncertainty in the fine structure of the speech signal, the order  $p \gg 1$  should be sufficiently large [26]. However, in the case of application of a high-order autoregressive model as a tool for the statistical data processing, we get a general system problem of small samples of the data of observations [27, 28]. In the analyzed case, this problem is strongly complicated by the conditions of finite duration  $\tau$  of the period of observations in which the speech signal can be regarded as stationary [8].

Thus, the problem of development of a rapid method for the asynchronous analysis of the voice source of speech aimed at the real-time application proves to be quite urgent. For this purpose, the authors of the present paper propose to apply a two-level autoregressive model of speech signals [29]. Its order  $(p_1, p_2)$  is determined by a pair of noticeably different values  $p_2 \gg p_1$ , which enables us to expect the possibility of combination, within a single method, of the advantages of both methods of acoustic measurements, namely, synchronous characterized by the potential accuracy of the results of analysis and asynchronous capable of decreasing the computational costs required for its realization.

## Statement of the problem

Let  $x(t)$  be a vocalized speech signal given by a sequence  $\{x(n)\}$  of  $x(n)$  readings at discrete times  $n = 0, 1, \dots, N - 1$  within the interval of observations  $t \leq \tau$  of finite length  $\tau = NT$ , where  $T$  is the period of time sampling of the signal. The Fourier spectrum of the sequence  $\{x(n)\}$  as a function of linear frequency  $f$  is given by the formula [30]:

$$S_x(jf) = T \sum_{n=0}^{N-1} x(n) \exp(-j2\pi n f T), \quad |f| \leq 0.5F, \tag{1}$$

where  $F = T^{-1}$  is a sampling rate and  $j$  is the imaginary unit.

In the linear model of vocal tract [8] specified by the complex transmission coefficient  $K(jf)$ , we get the following equality:

$$S_x(jf) = K(jf)S_z(jf), \quad |f| \leq 0.5F,$$

where  $S_z(jf)$  is a frequency spectrum of the sequence of excitation pulses  $z(n), n = 0, 1, \dots, N - 1$ , and

$$S_z(jf) = K^{-1}(jf)S_x(jf), \quad |f| \leq 0.5F. \tag{2}$$

As a result of the inverse Fourier transformation, we obtain

$$z(n) = \int_{-0.5F}^{0.5F} S_z(jf) \exp(j2\pi n f T) df, \quad n = 0, 1, \dots, N - 1, \tag{3}$$

In integral form, we can write

$$y(n) = y(n - 1) + z(n), \quad n = 0, 1, \dots, N - 1 \tag{4}$$

The set of expressions (2)–(4) determines the vocal source of speech by the method of inverse filtering [31]. In this case, Eq. 3 describes the sequence of excitation pulses of the vocal tract, while Eq. 4 specifies the volumetric velocity of the airflow passing through the glottis. The problem is thus reduced to the determination of the right-hand side of expression (2). In this case, it is necessary to explicitly determine the complex transmission coefficient of the vocal tract filter and substitute it in expression (3). Under the conditions of *a priori* uncertainty of the fine structure of speech signal, this is a nontrivial problem, and its solution requires the application of a universal probability-theory approach [13].

### Statistical model of the vocal tract

The main difficulty encountered in solving the posed problem is connected with the acoustic variability of speech signal [8]. Due to the influence of various random (uncontrollable) factors on the speaker in the process of speech production, the signal  $x(t)$  cannot be regarded as stationary or stable with respect to its parameters even for relatively small observation periods  $\tau = (3 \dots 5)T_0$ .

In the work [24] devoted to the analysis of voice timbre, the author justified the procedure of modeling the vocal tract by using a scheme of recursive filter with complex transmission factor

$$K(jf) = \left( 1 + \sum_{i=1}^{p_1} a_{p_1}(i) \exp(-j2\pi i f T) \right)^{-1}, \quad |f| \leq 0.5F \tag{5}$$

The order of this filter  $p_1$  is comparable with the double number of formants  $L_1$  in the spectrum of speech signal [20]. Thus, within the frequency band of a standard telephone channel 4kHz in width, for the vowel

speech sounds, we have  $L_1 = 4 \dots 6$  [32] and, hence,  $p_1 = 8 \dots 12$ . At the same time, the vector of filter coefficients (5) is determined by the  $p$ -vector of coefficients from the autoregressive equation

$$y(n) = - \sum_{i=1}^p a_p(i) y(n-i) + \eta(n), \quad n = 0, 1, \dots \quad (6)$$

of the same order  $p = p_1$ . Here,  $\{y(n)\}$  is a random (hypothetical) time series simulating the speech signal in discrete time  $n$ ;  $\{\eta(n)\}$  is the generating white noise with variance  $\sigma_\eta^2 = \text{const}$ . Assume that the preliminary autoregressive coefficients  $\{a_p(i)\}$  are adapted to the speech signal  $x(t)$  according to a vector of its readings  $\{x(n)\}$  of finite dimension  $N$ . In the theory of parametric estimation, there exists a specially developed mathematical procedure [33]. In particular, we can mention Berg's method<sup>1</sup> widely used in practice and based on the Levinson recursion [30]:

$$\begin{aligned} \forall q = \overline{1, p}: a_q(i) &= a_{q-1}(i) + c_q a_{q-1}(q-i), \quad i = 1, 2, \dots, q \\ c_q &= -2S_q^{-2} \sum_{n=q+1}^N \eta_{q-1}(n) v_{q-1}(n-1), \\ S_q^2 &= \sum_{n=q+1}^N \left[ \eta_{q-1}^2(n) + v_{q-1}^2(n-1) \right] \\ \eta_q(n) &= \eta_{q-1}(n) + c_q v_{q-1}(n-1), \\ v_q(n) &= v_{q-1}(n-1) + c_q \eta_{q-1}(n) \end{aligned} \quad (7)$$

in the case of its initialization by the system of equalities  $v_0(n) = \eta_0(n) = x(n-1)$  for all  $n \leq N$ . The final values of recursion (7) for  $q = p_1$  determine the adaptive autoregressive model (5) of the vocal tract in the frequency region, which should be substituted in expression (2). However, this is only the first step in solving the posed problem.

## Statistical model of speech signal

The problem is connected with the fact that not only the complex transmission factor of the vocal tract but also the spectral density  $S_x(jf)$  of the speech signal  $x(t)$  on the right-hand side of Eq. 3 are not completely determined by expression (1) due to the insufficient volume  $N = \tau F$  of the sample of observations  $\{x(n)\}$ . Note that the duration of frames in the systems of digital processing and transmission of speech does not exceed  $\tau = 30\text{--}40$  msec (see GOST R 53556.3-2012<sup>2</sup>). Thus, by using a standard telephone communication line and a sampling frequency of speech signal  $F = 8$  kHz for substitution in expression (1), we get at most  $N = 240\text{--}320$  readings of observations. A frequency resolution  $\delta f = \tau^{-1} = 25\text{--}30$  Hz attained in this case is comparable with the lower limit of the fundamental tone frequency  $F_0 = 80\text{--}100$  Hz in male speech. However, this contradicts the requirements imposed on the accuracy of voice analysis in the frequency domain because the spectral density (2) has a linear form and consists of amplitude-modulated quasiharmonics with frequencies  $F_0, 2F_0, \dots, LF_0$ , where  $L = 0.5F/F_0 \gg 1$  [29]. Thus, for the sampling frequency  $F = 8$  kHz, there are  $L = 40$  quasiharmonics within the working frequency band with relative shifts (with respect to each other) by a frequency  $F_0 = 100$  Hz. In order to significantly decrease the value of  $\delta f$  in these conditions, it is necessary to additionally determine

<sup>1</sup> Researchers prefer Berg's method to other methods for the analysis of parametric time series due to its well-known advantages in the efficiency and, which is most important, in the stability of generated autoregressive models.

<sup>2</sup> GOST R 53556.3-2012. Part 3 (MPEG-4 AUDIO). Encoding of speech signals with the use of the CELP linear prediction.

(extrapolate [30]), within the framework of the posed problem (2)–(4), not only the vocal tract but also the speech signal itself outside the interval of its observations. For this purpose, a special mathematical apparatus of parametric methods of statistical analysis was theoretically developed in [33]. The methods of this kind are based on the statistical simulation of the time series  $\{x(n)\}$  with the help of a hypothetical (imaginary) random process  $\{y(n)\}$ . For this purpose, it is customary to use the linear autoregressive process (6) of order  $p_2 \gg 1$  [25, 26]. The power spectral density of this process is given by the expression

$$G(f) = \sum_{\eta} T \left| 1 + \sum_{i=1}^p a_p(i) \exp(-j2\pi ifT) \right|^{-2}, \quad |f| \leq 0.5F \tag{8}$$

where the coefficients  $\{a_p(i)\}$  are computed according to recursive relation (7) with  $p = p_2$ . The order of autoregression  $p_2 \geq 2L$  is determined with regard for the double (but less than a half of the sample volume  $N$ ) number of quasiharmonics  $L$  in the spectrum of speech signal [20]. Under the conditions of the previous example, we obtain  $80 \leq p_2 < 120$ . In the general case,  $p_2$  is much greater than the order of the vocal tract filter (5), namely,  $p_2 \gg p_1$ .

From expression (8), by the method of spectral factorization [30], we obtain an autoregressive model of speech signal in the frequency domain:

$$S_x(jf) = c_0 \left( 1 + \sum_{i=1}^{p_2} a_{p_2}(i) \exp(-j2\pi ifT) \right)^{-1}, \quad |f| \leq 0.5F, \tag{9}$$

where  $c_0 = \text{const}$  is an adjustable scaling factor.

The problem of resolving power  $\delta f \ll F_0$  in model (9) can be overcome due to the effect of superresolution in frequency [28, 34]. According to (2), by using (9), we can write

$$S_z(jf) = c_0 \frac{1 + \sum_{i=1}^{p_1} a_{p_1}(i) \exp(-j2\pi ifT)}{1 + \sum_{i=1}^{p_2} a_{p_2}(i) \exp(-j2\pi ifT)}; \quad |f| \leq 0.5F \tag{10}$$

Expression (10) defines the general system autoregressive moving-average (ARMA) model in the theory of statistical analysis of random time series [30]. In the analyzed case, this model describes the voice source in the frequency domain (2). Substituting (10) in expression (3), in the time domain, we obtain

$$\begin{aligned} z(n) &= c_0 \int_{-0.5F}^{0.5F} \frac{1 + \sum_{i=1}^{p_1} a_{p_1}(i) \exp(-j2\pi ifT)}{1 + \sum_{i=1}^{p_2} a_{p_2}(i) \exp(-j2\pi ifT)} \exp(j2\pi n f T) df = \\ &= c_0 \int_{-0.5F}^{0.5F} \frac{FT_N \{b_1\}}{FT_N \{b_2\}} \exp(j2\pi n f T) df = \\ &= c_0 IFT_n \left\{ \frac{FT_N \{b_1\}}{FT_N \{b_2\}} \right\} \triangleq z_N(n), \quad n = 0, 1, \dots, N - 1. \end{aligned} \tag{11}$$

In (11), we have used the following notation:

$$FT_N \{b_r\} \triangleq T \sum_{i=0}^N b_r(i) \exp(-j2\pi ifT) = T \left[ 1 + \sum_{i=1}^{p_r} a_{p_r}(i) \exp(-j2\pi ifT) \right]$$

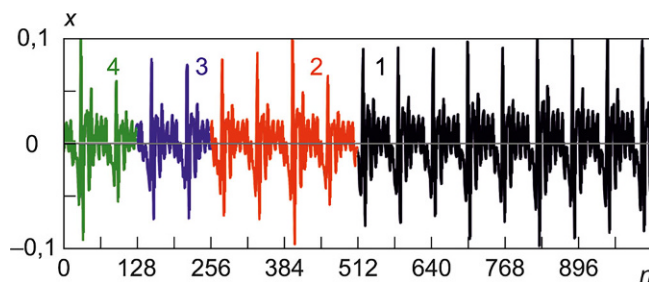
is the operator of Fourier transform,  $\mathbf{b}_r = \{b_r(i), i \leq J\} = [1, a_{p_r}(1), a_{p_r}(2), \dots, a_{p_r}(p_r), 0, 0, \dots, 0]$  is the vector of coefficients with dimension  $N+1$ , where  $r = 1, 2$ ;  $IFT_n\{\}$  is the operator of inverse Fourier transform of the spectral density  $S_{z,N}(jf) = K_N^{-1}(jf)S_{x,N}(jf)$ ,  $|f| \leq 0.5F$  of the excitation signal of vocal tract  $\{z_N(n)\}$ ; and  $K_N(jf) = FT_N^{-1}\{\mathbf{b}_1\}$  and  $S_{x,N}(jf) = c_0 FT_N^{-1}\{\mathbf{b}_2\}$  are the autoregressive models of the vocal tract (5) and speech signal (9), respectively, formed according to the results of recurrent processing (7) of the sequence of observations  $\{x(n)\}$  of finite volume  $N$ .

Expression (11), together with (7) and (10) specifies a method intended for the asynchronous analysis of a vocal speech source within the general formulation of the form (3). This method is based on the two-level autoregressive model aimed at the description of speech signals for two different levels of autocorrelation within the period of the fundamental tone (if the orders are equal  $p = p_1$ ) and in the interval of several consecutive periods (for  $p = p_2$ ). The problem of small samples in the proposed method is overcome due to the high rate of convergence of the Berg-Levinson recursion [31]. The problem of speed is solved by combining two computation procedures of different kinds within the framework of the common recurrence scheme (7). These procedures are aimed at the estimation, according to a sample  $\{x(n)\}$ , of the autoregressive coefficients  $\{a_{p_2}(i)\}$  and moving average  $\{a_{p_1}(i)\}$  as parameters of the ARMA model of the voice source (10). The efficiency of the proposed method was experimentally investigated by using the software specially developed by the authors<sup>3</sup>.

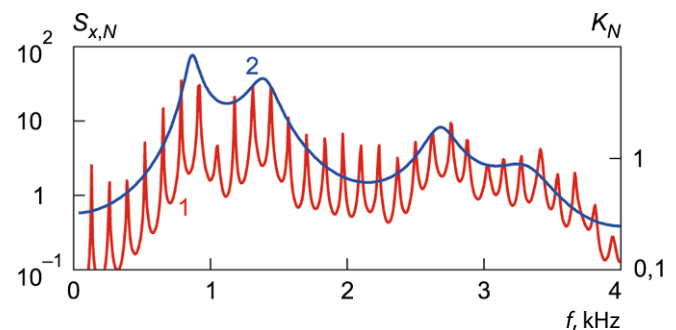
## Program and experimental results

As the object of the experimental investigations, we used the signals of six Russian vowel phonemes pronounced by a control speaker (one of the authors of the present paper): “a”, “i”, “o”, “u”, “y”, and “é”. A sufficiently large (3.5–4.0 sec) duration of these signals was chosen with an aim to be able to perform automatic partition of signals with a period of 16 msec into stationary segments of speaker’s oral speech of the same duration equal to  $\tau = 128$  msec. For the sampling frequency of speech signal  $F = 8$  kHz, the volume of experimental database for each vowel was not smaller than  $R = (3.5 - 0.128)/0.016 \approx 210$  frames of speech of the control speaker with dimensions  $N_0 = 8 \cdot 128 = 1024$ . For each frame, we formed four single-phoneme sound files  $x(t)$  of different duration  $\tau$ : 128, 64, 32 and 16 msec. In this case, the dimensions of  $N$  vectors of the same name  $\{x(n)\}$  were equal to  $N_1 = 1024$ ;  $N_2 = 512$ ;  $N_3 = 256$ , and  $N_4 = 128$  readings, respectively. All sound files of the frame of vowel speech sound “a” are depicted in Fig. 1. It is easy to see that any kind of synchronization of the data of observations with the fundamental tone of speech signal is excluded.

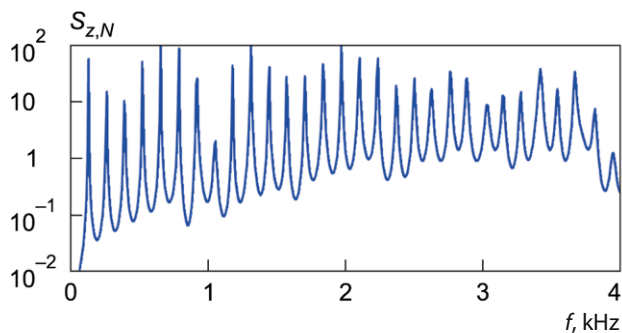
<sup>3</sup> Information system of phonetic analysis and speech training Phoneme Training: [site]. URL: <https://sites.google.com/site/frompldcreators/produkty-1/phonemetraining> (reference date: 18.02.2024).



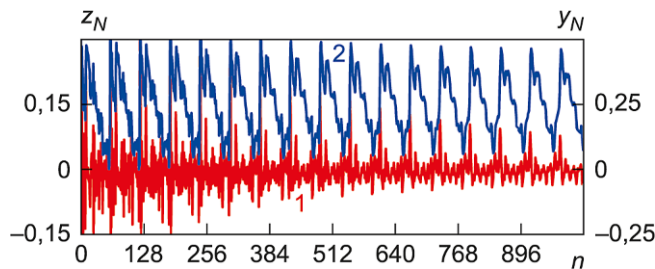
**Fig. 1** Signal of the Russian vowel phoneme “a” for the observation intervals equal to  $N=1024$ , 512, 256, and 128, regions 1–4, respectively



**Fig. 2** Amplitude spectrum of the speech signal (1) and the amplitude-frequency response of the vocal tract filter (2) according to the results of processing of the data of observations with a volume  $N = 1024$



**Fig. 3** Amplitude spectrum of the model of voice source of the Russian vowel phoneme “a” for the sample size  $N = 1024$



**Fig. 4** Model of voice source of the Russian vowel phoneme “a” based on the results of processing of the data of observations with a volume  $N = 1024$  in two versions of the description: excitation pulses (1) and pulses of the volumetric velocity of air flow (2) at the entrance of the vocal tract

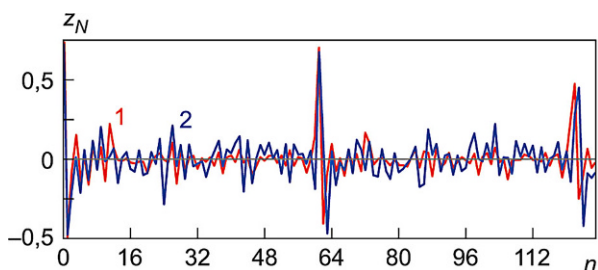
The software implementation of the voice source (11) with parameters  $p_1 = 10$  and  $p_2 = 90$  was experimentally investigated. In this case, the operators of direct and inverse Fourier transforms are realized on the basis of rapid algorithms of Fourier transformations with dimension  $M = 2^{10}$  and the frequency selectivity  $\Delta f = FM^{-1} = 7.8125$  Hz. The purpose and principle of action of both operators are illustrated in Figs. 2, 3 and 4.

In Fig. 2, we present the amplitude spectrum  $S_{x,N}(f) = |S_{x,N}(jf)|$  of the Russian vowel phoneme “a” signal and the amplitude-frequency characteristic  $K_N(f) = |K_N(jf)|$  of the vocal tract filter (5) for  $c_0 = \sqrt{10}$  constructed according to the results of processing the data of observations  $\{x(n)\}$  with the following volume:  $N = 1024$ . In Fig. 3, we display the corresponding amplitude spectrum  $S_{z,N}(f) = |S_{z,N}(jf)|$  of the model of voice source (11). The envelope of the amplitude spectrum characterizes the shape of the excitation pulses  $z_N(n)$ , whereas the repetition period of its quasiharmonics characterizes the frequency of the fundamental tone of the signal  $x(t)$ . In the analyzed case, it is approximately equal to  $F_0 \approx 132$  Hz. This fact is confirmed by the results of the profile work [29].

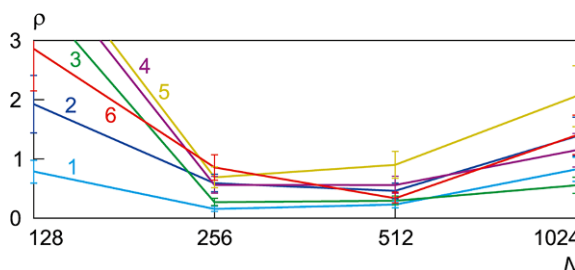
The same source (for the volume of observations  $N = 1024$ ) in the time domain is presented in Fig. 4 by two impulsive sequences: excitation of the vocal tract (11) and volumetric velocity of the air flow:

$$y_N(n) = y_N(n - 1) + z_N(n), \quad n = 0, 1, \dots, N - 1$$

The shape of excitation pulses on the enlarged scale is shown in Fig. 5 and compared with an impulsive sequence  $\{z_N(n)\}$  obtained for  $N = 256$ . It follows from Fig. 5 that both the shape and repetition frequency



**Fig. 5** Model of the voice source in the interval of the first two periods of the fundamental tone of speech signal according to the results of processing of the data of observations with sample sizes  $N=1024$  and 256; marks 1 and 2 respectively



**Fig. 6** Dependence of the parameter of accuracy of the ARMA model of voice source (11) on the duration of Russian vowel phonemes: “a” (1), “i” (2), “o” (3), “u” (4), “y” (5), and “e” (6)

$F_0 \approx 131.5$  Hz of the pulses of voice source (11) are stable with respect to the duration of the speech signal  $x(t)$  within a broad range  $\tau = 32 \dots 128$  msec. This conclusion is used as a foundation of the second (final) stage of the experimental investigation of the efficiency of the proposed method of voice analysis.

As a parameter of efficiency, we use the objective measure of stability of the ARMA model of the voice source (10) regarded as a function of the sample size  $N$  of the data of observations  $\{x(n)\}$ :

$$\rho(N) \triangleq \sqrt{F^{-1} \int_{-0.5F}^{0.5F} S_{z,N}^2(f) \overline{S}_z^{-2}(f) df} \times \sqrt{F^{-1} \int_{-0.5F}^{0.5F} \overline{S}_z^2(f) S_{z,N}^{-2}(f) df} - 1 \geq 0 \quad (12)$$

Here,  $\overline{S}_z(f) \triangleq 0.25 \sum_{i=1}^4 S_{z,N_i}(f)$  is the mean value of the amplitude spectrum  $S_{z,N}(f)$  of the speech signal on the set of four versions  $S_{z,N_i}(f), i = \overline{1,4}$ , considered in the experiment. In [35], the authors showed the invariance of measure (12) to the scale of the excitation signal  $\{z_N(n)\}$ . The lower the value of  $\rho(N)$ , the higher the stability of the considered model in the dynamics [36]. Moreover, the stability of the ARMA model (10) guarantees the validity of the proposed method of voice analysis [7].

The obtained results are presented in Fig. 6 in the form of a family of plots of the function  $\rho(N)$  for six Russian vowel phonemes pronounced by the control speaker. The vertical segments at the control points of these plots specify the boundaries of the confidence interval of parameter (12) according to the results of multiple ( $R$ -fold) measurements. In this case, the relative length of the confidence interval  $\varepsilon = 1.65/\sqrt{R}$  [29] for a confidence level equal to 0.9 does not go beyond  $165 \cdot 210^{-\frac{1}{2}} = 11.38\%$ . The plots presented in Fig. 6 differ from each other only in details. However, they are similar in the main: in all versions, the optimal choice of the size of sample  $\{x(n)\}$  lies within the range  $N = 256-512$ . This volume corresponds to the length of the interval of observations  $\tau = 32-64$  msec. This is, in fact, the requirement of the proposed method to the duration of the speech signal  $x(t)$  in the problem of voice analysis (2)–(4). Moreover, the lower boundary  $\tau = 32$  msec of the acceptable signal duration is directly related to the period  $T_0$  of its fundamental tone [29]. In the experiments, for different vowels, it varied within the range 7–8 msec. At the same time, the upper boundary  $\tau = 64$  msec of acceptable duration reflects a natural requirement of the proposed method (7), (10), and (11) to the stability of fine structure of the speech signal in the interval of observations.

## Discussion of the obtained results

We now consider the speed of the developed method for the analysis of voice sources determined by the two factors: the duration  $\tau$  of the speech frame characterizing the period of updating the results of the voice analysis of speech in formulation (11) and the computational complexity of the proposed method caused by the total amount of calculations  $W_\tau = W_7 + W_{10}$  performed according to relations (7) and (11). For relations (7), we have about  $W_7 = 3Np_2 = 3\tau Fp_2$  elementary operations of multiplication and division of real numbers [30]. The cost of simulation of the vocal tract by the autoregressive model (5) of order  $p_1 < p_2$  is not taken into account in this case because, in the recurrence computational scheme, they are included in the computation cost of modeling of the speech signal (9) [36]. In the case of relation (11), the volume of computations includes the threefold cost of performing the  $M$ -point rapid Fourier transform for  $M \geq N$ , which correspond to  $W_{10} = 3M \log_2 M$  elementary operations. In total, we get  $W_\tau = 3(Np_2 + M \log_2 M)$  elementary operations within the interval of observations of length  $\tau$  or  $W = W_\tau/\tau = 3F(p_2 + N^{-1}M \log_2 M)$  operations per second. Thus, under the conditions of the performed experiment, for  $p_2 = 90$ ;  $F = 8$  kHz;  $\tau = 2^5$  msec;  $N = 2^8$ ; and  $M = 2^{10}$ , we get  $W = 3 \cdot 8000(90 + 2^{-8} \cdot 2^{10} \cdot 10) = 3.12 \cdot 10^6 c^{-1}$ , which gives, as a result of recalculation to the clock frequency of the computing device, 3.12 MHz. This result, with a significant margin (by an order of magnitude or more) corresponds to the efficiency of modern speech systems operating under the conditions of soft (with delays for the duration of a single frame) real-time mode [37].



## Conclusions

The proposed method for the analysis of voice sources of speech makes it possible to model the excitation signal (3) of the vocal tract of a speaker in real time. Its sufficiently high speed is explained by the use of a high-speed recurrence procedure (7) used to adjust the parameters of the ARMA model (10) for a sequence of excitation pulses (11) according to a speech signal  $x(t)$  of finite duration  $\tau$ . The proposed method does not require synchronization of the sequence of observations  $\{x(n)\}$  with the fundamental tone of the speech signal and is characterized by relatively small calculation costs required for the technical implementation. The performed full-scale experiment confirmed the high speed of the proposed method and, at the same time, allowed us to formulate the requirements to the duration of speech signals.

The obtained results are intended for applications in the development and investigation of modern systems of digital speech communication, voice control, biometrics, biomedicine, and other speech systems [7] in which the specific voice features of speaker's speech are of primary importance.

**Author Contribution** Both authors contributed to the concept and design of the present study. Material preparation, data collection, and analyses were performed by both authors (Savchenko V. V. and Savchenko L. V.). The first draft of the manuscript was written by Savchenko V. V. Both authors commented previous versions of the manuscript. Both authors have read and approved the final manuscript.

**Conflict of interest** V. Vasilyevich Savchenko and L. Vasilyevna Savchenko declare that they have no competing interests.

## References

1. Li, Y., Tao, J., Erickson, D., Liu, B., Akagi, M.: Noise-robust glottal source and vocal tract analysis based on ARX-LF model, In IEEE/ACM Transactions on Audio, Speech, Language Processing, vol. 29., pp. 3375–3383 (2021). <https://doi.org/10.1109/TASLP.2021.3120585>
2. Narendra, N.P., Airaksinen, M., Story, B., Alku, P.: Estimation of the glottal source from coded telephone speech using deep neural networks. *Speech Commun.* **106**, 95–104 (2019). <https://doi.org/10.1016/j.specom.2018.12.002>
3. Drugman, T., Alku, P., Alwan, A., Yegnanarayana, B.: Glottal source processing: from analysis to applications. *Comput. Speech Lang.* **28**(5), 1117–1138 (2014). <https://doi.org/10.1016/j.csl.2014.03.003>
4. Sadok, S., Leglaive, S., Girin, L., Alameda-Pineda, X., Séguier, R.: Learning and controlling the source-filter representation of speech with a variational autoencoder. *Speech Comm* **148**, 53–65 (2023). <https://doi.org/10.1016/j.specom.2023.02.005>
5. Mittapalle, K.R., Pohjalainen, H., Helkkula, P.: Glottal flow characteristics in vowels produced by speakers with heart failure. *Speech Commun.* **137**, 35–43 (2022). <https://doi.org/10.1016/j.specom.2021.12.001>
6. Rudzicz, F.: *Clear speech: technologies that enable the expression and reception of language*. Springer, Cham (2022). <https://doi.org/10.1007/978-3-031-01599-1>
7. Ternström, S.: Special issue on current trends and future directions in voice acoustics measurement. *Appl. Sci.* **13**(6), 3514 (2023). <https://doi.org/10.3390/app13063514>
8. Savchenko, V.V.: Acoustic variability of voice signal as factor of information security for automatic speech recognition systems with tuning to user voice. *Radioelectron. Commun. Syst.* **63**(10), 532–542 (2020). <https://doi.org/10.3103/S0735272720100039>
9. Serry, M.A., Alzamendi, G.A., Zañartu, M., Peterson, S.D.: An Euler-Bernoulli-type beam model of the vocal folds for describing curved and incomplete glottal closure patterns. *J Mech Behav Biomed Mater* **147**, 106130 (2023). <https://doi.org/10.1016/j.jmbbm.2023.106130>
10. Sundberg, J.: Objective characterization of phonation type using amplitude of flow glottogram pulse and of voice source fundamental. *J. Voice* **36**(1), 4–14 (2022). <https://doi.org/10.1016/j.jvoice.2020.03.018>
11. Yao, X., Bai, W., Ren, Y.N., Liu, X., Hui, Z.: Exploration of glottal characteristics and the vocal folds behavior for the speech under emotion. *Neurocomputing* **410**, 328–341 (2020). <https://doi.org/10.1016/j.neucom.2020.06.010>
12. Rabiner, L.R., Shafer, R.W.: *Theory and applications of digital speech processing*. Pearson, Boston (2011)
13. Gibson, J.: Mutual information, the linear prediction model, and CELP voice codecs. *Information* **10**(5), 179–189 (2019). <https://doi.org/10.3390/info10050179>
14. Südholt, D., Cámara, M., Zh, X., Reiss, J.D.: Vocal tract area estimation by gradient descent. In: Proc. of the 26th Internat. Conf. On digital audio effects (DAFx23). Denmark, Copenhagen (2023) <https://doi.org/10.48550/arXiv.2307.04702>

15. Li, Y., Sakakibara, K.I., Akagi, M.: Simultaneous estimation of glottal source waveforms and vocal tract shapes from speech signals based on ARX-LF model. *J. Signal Process. Syst.* **92**, 831–838 (2020). <https://doi.org/10.1007/s11265-019-01510-4>
16. Drugman, T., Bozkurt, B., Dutoit, T.: A comparative study of glottal source estimation techniques. *Comput. Speech Lang.* **26**, 20–34 (2019)
17. Freixes, M., Luis, J.O., Socoró, J.C., Francesc, A.P.: Evaluation of glottal inverse filtering techniques on OPENGLOT synthetic male and female vowels. *Appl. Sci.* **13**(15), 8775 (2023). <https://doi.org/10.3390/app13158775>
18. Zhang, Z., Lin, J.: Evaluation of glottal inverse filtering in the presence of source-filter interaction. *J. Acoust. Soc. Am.* **152**(4), A284–A284 (2022). <https://doi.org/10.1121/10.0016281>
19. Perrotin, O., McLoughlin, I.: A spectral glottal flow model for source-filter separation of speech. In: 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP 2019), pp. 7160–7164. Brighton, UK (2019) <https://doi.org/10.1109/ICASSP.2019.8682625>
20. Savchenko, V.V.: Method for reduction of speech signal autoregression model for speech transmission systems on low speed communication channels. *Radioelectron. Commun. Syst.* **64**(11), 592–603 (2021). <https://doi.org/10.3103/S0735272721110030>
21. Walker, J., Murphy, P.A.: Review of glottal waveform analysis. In: Progress in nonlinear speech processing, Lecture notes in computer science, vol. 4391. Springer, Berlin, Heidelberg (2007) [https://doi.org/10.1007/978-3-540-71505-4\\_1](https://doi.org/10.1007/978-3-540-71505-4_1)
22. Palaparthi, A., Titze, I.R.: Analysis of glottal inverse filtering in the presence of source-filter interaction. *Speech. Commun.* **123**, 98–108 (2020). <https://doi.org/10.1016/j.specom.2020.07.003>
23. Gupta, S., Fahad, M.S., Deepak, A.: Pitch-synchronous single frequency filtering spectrogram for speech emotion recognition. *Multimed Tools Appl.* **79**, 23347–23365 (2020). <https://doi.org/10.1007/s11042-020-09068-1>
24. Savchenko, V.V.: Measure of difference between speech signals according to the voice timbre. *Izmerit. Tekh.* **10**, 63–69 (2023). <https://doi.org/10.32446/0368-1025it.2023-10-63-69>
25. Nossier, S.A., Wall, J., Moniri, M., Glackin, C., Cannings, N.: A comparative study of time and frequency domain approaches to deep learning based speech enhancement. In: 2020 Internat. Joint conf. on neural networks (IJCNN), pp. 1–8. Glasgow, UK (2020) <https://doi.org/10.1109/IJCNN48605.2020.9206928>
26. Freixes, M., Arnela, M., Socoró, J.C., Alías, F., Guasch, O.: Glottal source contribution to higher order modes in the finite element synthesis of vowels. *Appl. Sci.* **9**(21), 4535 (2019). <https://doi.org/10.3390/app9214535>
27. Candan, Ç.: Making linear prediction perform like maximum likelihood in gaussian autoregressive model parameter estimation. *Signal Process.* **166**, 107256 (2020). <https://doi.org/10.1016/j.sigpro.2019.107256>
28. Cui, S., Li, E., Kang, X.: Autoregressive model based smoothing forensics of very short speech clips. In: IEEE Internat. Conf. on multimedia and expo (ICME), pp. 1–6. London, UK (2020) <https://doi.org/10.1109/ICME46284.2020.9102765>
29. Savchenko, A.V., Savchenko, V.V.: Adaptive method for measuring a fundamental tone frequency using a two-level autoregressive model of speech signals. *Izmerit. Tekh.* **6**, 60–66 (2022). *English translation: Meas. Tech.*, **65**(6), 453–460 (2022). <https://doi.org/10.1007/s11018-022-02104-6> <https://doi.org/10.32446/0368-1025it.2022-6-60-66>
30. Marple, S.L.: Digital spectral analysis with applications, 2nd edn. Dover Publications, Mineola, New York (2019)
31. Savchenko, V.V., Savchenko, A.V.: Method for measuring distortions in speech signals during transmission over a communication channel to a biometric identification system. *Izmerit. Tekh.* **11**, 65–72 (2020). *English translation: Meas. Tech.*, **63**(11), 917–925 (2021). <https://doi.org/10.1007/s11018-021-01864-x> <https://doi.org/10.32446/0368-1025it.2020-11-65-72>
32. Kathiresan, T., Maurer, D., Suter, H., Dellwo, V.: Formant pattern and spectral shape ambiguity in vowel synthesis: The role of fundamental frequency and formant amplitude. *J. Acoust. Soc. Am.* **143**(3), 1919–1920 (2018). <https://doi.org/10.1121/1.5036258>
33. Corey, R.M., Kozat, S.S., Singer, A.C.: Parametric estimation. In: Diniz, P.S.R. (ed.) *Signal processing and machine learning theory*, pp. 689–716. Academic Press, (2024) <https://doi.org/10.1016/B978-0-32-391772-8.00017-X>
34. Savchenko, V.V.: Method for comparison testing of parametric power spectrum estimates: spectral analysis via time series synthesis. *Izmerit. Tekh.* **6**, 56–62 (2023). *English translation: Meas. Tech.*, **66**(6), 430–438 (2023). <https://doi.org/10.1007/s11018-023-02244-3> <https://doi.org/10.32446/0368-1025it.2023-6-56-62>
35. Savchenko, A.V., Savchenko, V.V.: Scale-invariant modification of COSH distance for measuring speech signal distortions in real-time mode. *Radioelectron. Comm. Syst.* **64**(6), 300–306 (2021). <https://doi.org/10.3103/S0735272721060030>
36. Savchenko, V.V.: Improving the method for measuring the accuracy indicator of a speech signal autoregression model. *Izmerit. Tekh.* **10**, 58–63 (2022). *English translation: Meas. Tech.*, **65**(10), 769–775 (2023). <https://doi.org/10.1007/s11018-023-02150-8> <https://doi.org/10.32446/0368-1025it.2022-10-58-63>
37. Kumar, S., Singh, S.K., Bhattacharya, S.: Performance evaluation of a ACF-AMDF based pitch detection scheme in real time. *Int J Speech Technol* **18**, 521–527 (2015). <https://doi.org/10.1007/s10772-015-9296-2>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

Vladimir Vasilyevich Savchenko<sup>1</sup> · Lyudmila Vasilyevna Savchenko<sup>1</sup>

✉ Vladimir Vasilyevich Savchenko  
vvsavchenko@yandex.ru

<sup>1</sup> “Higher School of Economics”, National Research University, Nizhny Novgorod, Russian Federation