# Application of Parameters of Voice Signal Autoregressive Models to Solve Speaker Recognition Problems

**Evgeny Novikov, Vladimir Trubitsyn**

***Abstract*— An approach to the formation of the voice signal (VS) informative features of the Vietnamese language on the basis of stationary autoregressive model coefficients is described. An original algorithm of VS segmentation based on interval estimation of speech sample numerical characteristics was developed to form local stationarity areas of the voice signal. The peculiarity is the use of high order autoregressive coefficients, the set of which is determined on the basis of discriminant analysis.**

***Tóm tắt*—Bài báo mô tả một cách tiếp cận để tạo ra các đặc trưng thông tin tín hiệu thoại (VS-voice signal) của tiếng Việt trên cơ sở các hệ số của mô hình tự hồi quy dừng. Một thuật toán độc đáo để phân đoạn tín hiệu thoại dựa trên ước tính khoảng của các đặc trưng số mẫu tiếng nói đã được phát triển để tạo ra các vùng tĩnh cục bộ của tín hiệu thoại. Điểm đặc biệt là việc sử dụng các hệ số tự hồi quy bậc cao, tập hợp của chúng được xác định trên cơ sở phân tích biệt thức.**

***Keywords*— voice signal; voice signal segmentation; informative speech features; autoregression models; discriminant data analysis.**

***Từ khóa— tín hiệu thoại; phân đoạn tín hiệu thoại; các đặc tính thông tin thoại; mô hình tự hồi quy; phân tích dữ liệu khác biệt*.**

## I. INTRODUCTION

Despite quite a large number of studies in the field of speaker recognition, this problem has not yet been fully solved, since the required accuracy of identification or verification is not provided.

The fundamental stage of speaker recognition is VS parameterization, which means the formation of informative speech features demonstrating the individual characteristics of the speaker. Currently, the following parameterization methods of speaker recognition are most commonly used [1-4]:

- Formant methods;
- Methods of analyzing primary tone statistics;
- Methods based on linear prediction coefficients;
- Cepstral methods.

These methods are widely studied for the Slavic and Germanic languages and automatic speaker recognition systems are developed on their basis. However, the recognition accuracy of such systems does not allow their industrial implementation. The main reasons for this situation are the following:

- The absence of formalized criteria of selecting the length of the window for the original VS decomposition;
- Ambiguity of choosing the basic VS conversion functions;
- Instability of informative speech features relative to noise;
- Transformation of the original VS, leading to an increase in resource capacity and significant errors in calculating informative speech features;
- Significant variability of informative feature values for the same speaker.

At the same time, the existing systems have not been tested for the Vietnamese language, because they are not focused on the Vietnamese speech and do not take into account its features. Therefore, there is a contradiction between the capabilities of existing voice recognition methods and the need in ensuring the required values of voice biometrics accuracy.

When creating modern systems of the speaker's voice recognition, the advantage is

This manuscript is received July 19, 2019. It is commented on October 22, 2019 and is accepted on October 31, 2019 by the first reviewer. It is commented on December 17, 2019 and is accepted on December 27, 2019 by the second reviewer.

given to VS stochastic models, which are based on the assumption that the signal can be well described as a parametric random process and that its parameters can be estimated accurately enough [5, 6].

This approach, taking into account the peculiarities of Vietnamese speech (words are monosyllabic, speech is slower than Russian or English one, the temporal structure of vocalized sounds is quite stable) makes extensive use of the autoregressive method.

The VS autoregressive model is the most common form of speech path mathematical description for solving the problems of VS analysis and synthesis. It is explained by the adequacy of this model to the acoustic representation of the speech path in the form of pipe segments [5]. The method of estimating autoregressive model parameters allows applying them to solve the problems of VS recognition and synthesis. On the other hand, the coefficients of such a model can be interpreted as parameters of the speech path model generating VS. In this case, we can talk about the connection of the autoregressive method with the methods of identifying dynamic systems, which allow us to evaluate the structure and parameters of the identified models. Another prerequisite for the use of autoregressive models is the possibility of representing the VS in the form of a time series with time-varying probabilistic characteristics. The information above suggests that the possibility of describing VS by autoregressive models in the time domain is possible, as well as applying these models to solving speaker recognition problems.

## II. VOICE SIGNAL SEGMENTATION

The analysis of existing approaches to speaker recognition based on autoregressive models showed that the main reason for the lack of this method’s reliability is the presence of their significant parameters variations in different VS implementations. Among the influencing factors the key one is to determine the beginning and length of the speech segment, which will provide stable parameters of autoregressive models. At the same time, recent studies have shown that the best recognition of the speaker corresponds to the areas of speech associated with the characteristics of the primary tone (PT) [7-9]. Therefore, to form a VS mathematical model, it is initially necessary to carry out its segmentation in quasi-stationarity areas with a constant primary tone frequency (PTF).

The existing methods of VS segmentation do not allow obtaining robust estimates of autoregression coefficients, as they have significant errors in determining the position of boundaries within 4-6 milliseconds in comparison with manual marking. In this regard, an original algorithm for VS segmentation was developed, based on interval estimating the standard deviation of the speech sample values.

The algorithm is represented by the sequence of the following steps.

1. Selecting active speech fragments and dividing them into sections corresponding to vocalized (vowels) and non-vocalized sounds. To solve this problem, you can use one of the existing algorithms for determining the temporal boundaries of voice activity (Voice Activity Detection – VAD). In particular, the algorithm based on analyzing the characteristics of the VS autocorrelation function (ACF) [10-12] is applied.

2. Determining the preliminary size of the sliding window, within which the mean square deviation (MSD) of speech samples will be estimated.

Since the task of speech segmentation in quasi-stationary sections is set, the size of the sliding window should correspond to the speaker's PTF. In general, during recognition process the speaker does not identify himself, so at the initial stage it is necessary to use a common assessment for all registered users' PTF. It is advisable to apply as such assessment the minimum PTF value for the intended speakers. For example, the PTF variation limits of Vietnamese men range from 110 to 300 Hz [13]. Then, as the initial size of the sliding window, the value of 400 samples at a given sampling rate of 44100 Hz should be used.

3. Point and interval estimation of VS samples’MSD.

For a particular speaker, the point and interval estimates of speech samples in the first selected section of the vowel sound are calculated using expressions of the following form [14]:

$$s(x) = \sqrt{\frac{\sum_{i=1}^{T}\left[x(t) - \bar{x}(t)\right]^2}{T-1}} \qquad (1)$$

$$P\left[\frac{s(x)\sqrt{T-1}}{\chi^2(\alpha_1, \vartheta)} < \sigma(x) < \frac{s(x)\sqrt{T-1}}{\chi^2(\alpha_2, \vartheta)}\right] = \gamma \qquad (2)$$

where $s(x)$ – is a point estimate of the speech samples' MSD, $x(t)$ – VS sample at time *t*, $\bar{x}(t)$ – is the estimate of samples' mathematical expectation, $T$ – is the sliding window size, $\vartheta = T - 1$, $\alpha_1 = (1-\gamma)/2$, $\alpha_2 = (1+\gamma)/2$, $\gamma$ – is a confidence probability.

4. One sample time window shift and calculation of the MSD point estimation of the obtained VS samples by expression (1).

5. Analyzing the calculated value of the VS samples' MSD estimation.

The calculated value of speech samples' MSD is analyzed to identify getting into the interval defined by formula (2). If the estimate value is within the interval, then the actions of step 3 are performed. Otherwise, using expression (2), a new interval estimate of MSD is determined.

6. The maximum duration segment is selected from the obtained preliminary VS segments and the period of PT by analyzing the VS ACF is estimated on its basis [5]:

$$B(\tau) = \sum_{t=x_{\text{in}}^{\text{pr}}}^{x_{\text{fin}}^{\text{pr}}} x(t) \cdot x(t+\tau), \qquad (3)$$

where $x_{\text{in}}^{\text{pr}}$ and $x_{\text{fin}}^{\text{pr}}$ – are the numbers of the initial and final speech samples of the preliminary quasi-stationary segments, respectively.

7. The formation of VS segments with a constant PTF.

The refined VS segments are formed using the procedure described in steps 2, 3 and 4. The value of the PT estimated period $T_{\text{pt}}$ is taken as the size of the sliding window. On the basis of the maximum duration segment, a section is formed, its size is a multiple of the PT period $[x_{\text{in}}^{\text{pr}}; x_{\text{fin}}^{\text{pr}}]$.

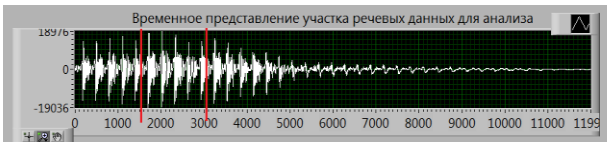


Fig.1. The results of forming the quasi-stationarity segments of the sound $\breve{a}$ in the word $\breve{a}n$

An example of forming quasi-stationarity segments is shown in Fig.1 (implemented in the LabVIEW modeling environment).

## III. FORMATION OF A VOICE SIGNAL MATHEMATICAL MODEL

The maximum error in determining the boundaries of the described algorithm's quasi-stationary segments in comparison with the "manual" marking was about 0.5 milliseconds. This result allows us to obtain robust values of autoregressive model coefficients within their standard errors. Thus, based on the properties of the VS selected areas (periodicity and quasi-stationarity), a hypothesis is put forward about the possibility of using autoregressive models AR(*p*) for VS modeling, which have the following general form:

$$x(t) = c + \sum_{i=1}^{p} a_i \cdot x(t-i) + \varepsilon(t), \ t = \overline{x_{\text{in}}^{\text{pr}}; x_{\text{fin}}^{\text{pr}}}, (4)$$

where $a_i$ – is the unknown autoregressive coefficient (AC), $\varepsilon(t)$ – is the white noise process, $c = \mu \cdot (1 - a_1 - \ldots - a_p)$, $\mu$ – is speech sample mathematical expectation.

The Box-Jenkinson methodology is used for VS mathematical description, according to which the modeling process consists of three stages – identification, evaluation of parameters and verification of the model adequacy.

Identification refers to the selection of the model structure (4) by the observed values of speech samples. To substantiate the possibility of applying autoregressive models and using the VS ergodicity property, the stationarity of the VS selected segments was estimated. For this purpose, visual analysis of speech segments' ACF and the extended Dickey-Fuller test were applied.

The ACF is formed by the values of the selected speech sample correlation coefficients [15]:

$$r(\tau) = \frac{\sum_{t=x_{\text{in}}^{\text{pr}}}^{t=x_{\text{fin}}^{\text{pr}}-\tau} [x(t) - \bar{x}(t)] \cdot [x(t+\tau) - \bar{x}(t+\tau)]}{\sqrt{\sum_{t=x_{\text{in}}^{\text{pr}}}^{t=x_{\text{fin}}^{\text{pr}}-\tau} [x(t) - \bar{x}(t)]^2 \sum_{t=x_{\text{in}}^{\text{pr}}}^{t=x_{\text{fin}}^{\text{pr}}-\tau} [x(t+\tau) - \bar{x}(t+\tau)]^2}}. \qquad (5)$$

In the case of a stationary series, the ACF converges to zero (Fig. 2), and for autoregression this process represents damping exponents [16].

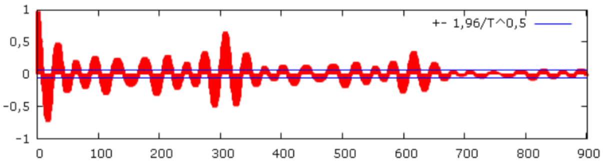
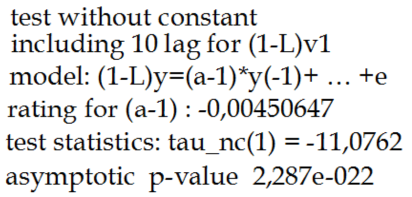


Fig. 2. The ACF example of the vowel sound segment $\breve{a}$ of the word $\breve{a}n$

The extended Dickey-Fuller test allows us to test the assumption that the studied VS segment corresponds to one of two types: deterministic or stochastic trend. For this purpose a statistical hypothesis about the presence of the model characteristic equation unit root (4) is put forward.

For verification, test statistics are used, the distribution of which is calculated by McKinnon. Since the specification of the VS model (4) is unknown, the Dickey-Fuller test was used for various types of models: without constant, with constant, with constant and trend.

Fig.3 shows an application example of the Dickey-Fuller test. It can be seen from the Fig that the studied area of the VS is stationary, since the significance level of the obtained observed value of the Dickey-Fuller statistics $tau\_nc(1) = -17.5542$ is equal to $4.168 \cdot 10^{-37}$, which is significantly less than the specified significance level $\alpha = 0.05$.



```
test without constant
including 10 lag for (1-L)v1
model: (1-L)y=(a-1)*y(-1)+ … +e
rating for (a-1) : -0,00450647
test statistics: tau_nc(1) = -11,0762
asymptotic  p-value  2,287e-022
```

Fig.3. An example of using the Dickey-Fuller test to analyze the segment stationarity of the sound $\breve{a}$ of the word $\breve{a}n$

An experimental analysis of the ACF of speech segments obtained for a sufficiently large number of Vietnamese words showed that they have the same structure corresponding to the ACF shown in Fig.2 Moreover, the use of the Dickey-Fuller test for these segments did not allow us to reject the hypothesis of the absence of a unit root. The results obtained show that the VS sections allocated on the basis of the developed segmentation algorithm are stationary time series and autoregressive models can be used to describe them. However, the use of arbitrary segments for parameterization of 20 ms does not allow us to make an assumption about their stationarity.

The final step in identifying the VS model is to evaluate its order. It is known that if the analyzed speech segment is a process of AR ($p$) type, then the values of the partial correlation coefficients up to order p are different from zero: $\rho_{\text{part}}(p) \neq 0$, and their values of order $m$ are equal to zero: $\rho_{\text{part}}(m) = 0, m > p$ [16]. This allows us to estimate the order of the autoregression process according to the schedule of a private autocorrelation function (PACF) (Fig. 4).

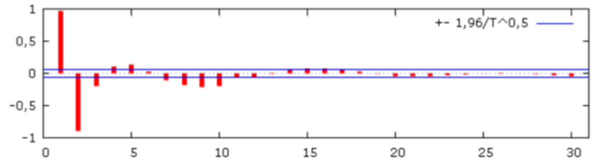


Fig.4. An example of a PACF segment of the sound $\breve{a}$ in the word $\breve{a}n$

PACF is formed by the values of the selective coefficients of partial correlation [15]:

$$r_{\text{part}}(\tau) = \frac{\begin{vmatrix} 1 & r(1) & \cdots & r(1) \\ r(1) & 1 & \cdots & r(2) \\ \vdots & \vdots & \ddots & \vdots \\ r(\tau-1) & r(\tau-2) & \cdots & r(\tau) \end{vmatrix}}{\begin{vmatrix} 1 & r(1) & \cdots & r(\tau-1) \\ r(1) & 1 & \cdots & r(\tau-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(\tau-1) & r(\tau-2) & \cdots & 1 \end{vmatrix}}, \qquad (6)$$

where $r(\tau)$ – are sample correlation coefficients.

From Fig.4 it is seen that in accordance with the behavior of the PACF for the selected segment of the word sound, it is necessary to use a 10th-order autoregressive model: AR (10).

On the whole, the results of the PACF study of the vowels' selected segments of various Vietnamese words showed that for their mathematical description it is sufficient to use the autoregressive models of the form (4), the order of which ranges from 10 to 14.

The estimation of the model parameters was carried out on the basis of the maximum likelihood method, in accordance with which, as an estimate of unknown AC $a_1,\ldots,a_p$, such values $\hat{a}_1,\ldots,\hat{a}_p$ are adopted that maximize the likelihood function of the form:

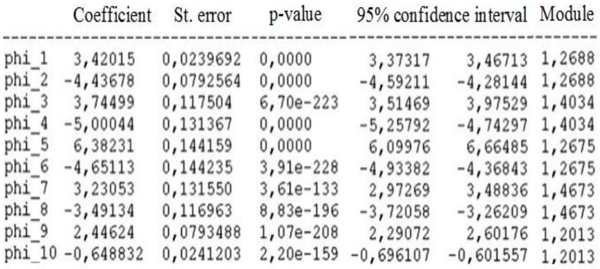
$$L(x_{\text{in}}^{\text{pr}},\ldots,x_{\text{fin}}^{\text{pr}};a_1,\ldots,a_p) = \\ = \prod_{t=x_{\text{in}}^{\text{pr}}}^{x_{\text{fin}}^{\text{pr}}+p} \frac{e^{\left(-\frac{1}{2\sigma^2}\cdot[x(t)-(\hat{a}_1\cdot x(t-1)+\ldots+\hat{a}_p\cdot x(t-p))]^2\right)}}{\sqrt{2\pi\sigma^2}} \to \max. \qquad (7)$$

Since the values of the coefficients of the VS model described by formula (4) are estimated using selective VS implementations, it is mandatory to check the statistical significance of the evaluation results. For this, the Student criterion is applied:

$$t = \hat{a}_i / \sigma_{\hat{a}_i} , \qquad (8)$$

where $\sigma_{\hat{a}_i}$ – is the standard error of the *i*-th AC equation (4).

An example of assessing the significance of the AC is shown in Fig.5.



|  | Coefficient | St. error | p-value | 95% confidence interval |  | Module |
|---|---|---|---|---|---|---|
| phi_1 | 3,42015 | 0,0239692 | 0,0000 | 3,37317 | 3,46713 | 1,2688 |
| phi_2 | -4,43678 | 0,0792564 | 0,0000 | -4,59211 | -4,28144 | 1,2688 |
| phi_3 | 3,74499 | 0,117504 | 6,70e-223 | 3,51469 | 3,97529 | 1,4034 |
| phi_4 | -5,00044 | 0,131367 | 0,0000 | -5,25792 | -4,74297 | 1,4034 |
| phi_5 | 6,38231 | 0,144159 | 0,0000 | 6,09976 | 6,66485 | 1,2675 |
| phi_6 | -4,65113 | 0,144235 | 3,91e-228 | -4,93382 | -4,36843 | 1,2675 |
| phi_7 | 3,23053 | 0,131550 | 3,61e-133 | 2,97269 | 3,48836 | 1,4673 |
| phi_8 | -3,49134 | 0,116963 | 8,83e-196 | -3,72058 | -3,26209 | 1,4673 |
| phi_9 | 2,44624 | 0,0793488 | 1,07e-208 | 2,29072 | 2,60176 | 1,2013 |
| phi_10 | -0,648832 | 0,0241203 | 2,20e-159 | -0,696107 | -0,601557 | 1,2013 |

Fig.5. An example of assessing the model parameters significance of the sound segment $\breve{a}$ in the word $\breve{a}n$

The results of assessing the significance of a sufficiently large number of models made it possible to establish that all the coefficients are significant, except for the constant *c*. Then the VS model will have the following form:

$$\hat{x}(t) = \sum_{i=1}^{p} \hat{a}_i \cdot x(t-i), \; t = \overline{x_{\text{in}}^{\text{pr}}, x_{\text{fin}}^{\text{pr}}}, \; p = \overline{10{,}14} . \quad (9)$$

It was also found that models of VS segments generate stationary sequences of speech samples since the roots of the characteristic equation $1 - (\hat{a}_1 \cdot z + ... + \hat{a}_p \cdot z^p) = 0$, obtained from the equation (9) by introducing the shift operator $L^s x(t) = x(t-s)$, are modulo larger than 1: $z_i > 1$ (Fig.5).

The assessment of the model adequacy (9) is based on testing the hypothesis that errors $\varepsilon(t)$ are a process of Gaussian white noise.

In order to check the absence of model errors autocorrelation, the ACF analysis of error estimates and Lyung-Box statistics were applied:

$$\chi^2 = T \cdot (T+2) \cdot \sum_{\tau=1}^{M} r_\varepsilon^2(\tau) / (T-\tau) , (10)$$

where $r_\varepsilon(\tau) = \dfrac{\sum_{t=x_{\text{in}}^{\text{pr}}}^{x_{\text{fin}}^{\text{pr}}-\tau} \hat{\varepsilon}(\tau) \cdot \hat{\varepsilon}(t+\tau)}{\sum_{t=x_{\text{in}}^{\text{pr}}}^{x_{\text{fin}}^{\text{pr}}-\tau} [\hat{\varepsilon}(\tau)]^2}$, $M$ – maximum order of ACF errors $\hat{\varepsilon}(t)$, $T = \overline{x_{\text{in}}^{\text{pr}}, x_{\text{fin}}^{\text{pr}}}$ .

If all the values of AC errors $\hat{\varepsilon}(t)$ are within the $\pm 1{,}96 / \sqrt{T}$ range, then we can assume the absence of their autocorrelation (Fig.6).

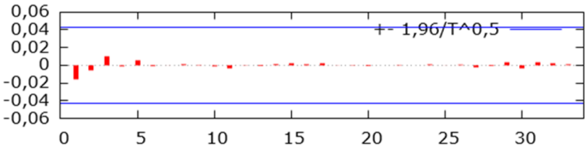


Fig.6. An example of ACF model errors of the sound segment $\breve{a}$ in the word $\breve{a}n$

Since the asymptotic distribution of the Lyung-Box statistics (10) was inferred under the assumption that the errors of model (9) represent Gaussian white noise, it is necessary to verify the hypothesis of a normal distribution of errors. For this, the Pearson criterion is applied:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i} , \qquad (11)$$

where $m$ – is the number of intervals of an errors series $\hat{\varepsilon}(t)$ of the VS model, $n_i$ and $n \cdot p_i$ – are the empirical and theoretical error rates in the *i*-th grouping interval.

As a result of evaluating the adequacy of various VS models, it was found that the errors are a process of white noise.

To assess the accuracy of the models, the coefficient of determination was calculated:

$$B = \frac{\sum_{t=x_{\text{in}}^{\text{pr}}}^{x_{\text{fin}}^{\text{pr}}} [\hat{x}(t) - \bar{x}(t)]^2}{\sum_{t=x_{\text{in}}^{\text{pr}}}^{x_{\text{fin}}^{\text{pr}}} [\hat{x}(t) - \bar{x}(t)]^2} . \qquad (12)$$

It has been experimentally established that the values of the determination coefficient for various models of words in the Vietnamese language vary from 0.94 to 0.99. This result indicates a high accuracy of autoregressive models, which allows them to be used for mathematical description of VS in Vietnamese speech.

## IV. FORMATION OF INFORMATIVE SPEECH FEATURES

The analysis obtained on the basis of the AC model (9) showed that their values are resistant to possible variations when the speaker pronounces the same phrase. However, a comparison of calculated AC values obtained by pronouncing identical phrases by different speakers made it possible to establish the presence of slight differences between the corresponding coefficients (Table 1).

TABLE 1. AC MODEL VALUES OF THE SOUND SEGMENT *ă* OF THE WORD *ăn*

| **Model parameters** | **Speaker 1** | **Speaker 2** | **Speaker 3** |
|---|---|---|---|
| $\hat{a}_1$ | 3.54 | 2.84 | 3.59 |
| $\hat{a}_2$ | -4.34 | -2.28 | -4.39 |
| $\hat{a}_3$ | 2.37 | 1.34 | 2.10 |
| $\hat{a}_4$ | -2.83 | -4.35 | -2.51 |
| $\hat{a}_5$ | 4.68 | 4.84 | 5.32 |
| $\hat{a}_6$ | -1.98 | -0.80 | -3.16 |
| $\hat{a}_7$ | -1.02 | 1.44 | -0.29 |
| $\hat{a}_8$ | -0.36 | -3.60 | -1.78 |
| $\hat{a}_9$ | 0.53 | 0.39 | 3.26 |
| $\hat{a}_{10}$ | 1.93 | 1.34 | -0.45 |
| $\hat{a}_{11}$ | -2.00 | 0.50 | -0.28 |
| $\hat{a}_{12}$ | 0.16 | -0.39 | 0.44 |
| $\hat{a}_{13}$ | 0.43 | -0.65 | -0.08 |
| $\hat{a}_{14}$ | -0.13 | 0.35 | -0.36 |

The correlation analysis of the model parameters showed that there is a fairly strong correlation dependence between the AC sets (Table 2). In this case, the main contribution to this correlation is made by lower-order coefficients (Table 2), while there is a weak correlation between higher-order coefficients (Table 2). This situation is explained by the fact that lower-order coefficients characterize the pronounced sound, while higher-order coefficients contain information about the speaker's features [17].

Then we can conclude that the recognition of the speakers lies in the high-frequency region of the VS, which is characterized by higher-orders of the AC.

TABLE 2. A CORRELATION MATRIX OF THE AC MODEL OF THE SOUND SEGMENT *ă* IN THE WORD *ăn*

| $\hat{a}_1-\hat{a}_{14}$ | **Speaker 1** | **Speaker 2** | **Speaker 3** |
|---|---|---|---|
| **Speaker 1** | 1 | 0.78 | 0.86 |
| **Speaker 2** | 0.78 | 1 | 0.77 |
| **Speaker 3** | 0.86 | 0.77 | 1 |
| $\hat{a}_1-\hat{a}_6$ | Speaker 1 | Speaker 2 | Speaker 3 |
| **Speaker 1** | 1 | 0.94 | 0.98 |
| **Speaker 2** | 0.94 | 1 | 0.93 |
| **Speaker 3** | 0.98 | 0.93 | 0.98 |
| $\hat{a}_7-\hat{a}_{14}$ | Speaker 1 | Speaker 2 | Speaker 3 |
| **Speaker 1** | 1 | 0.23 | 0.19 |
| **Speaker 2** | 0.23 | 1 | 0.17 |
| **Speaker 3** | 0.19 | 0.17 | 1 |

Based on this, in order to increase the determinancy of speaker recognition, it was proposed to consider models of VS segments that include only given AC of higher-orders:

$$\begin{cases} \omega_j = \left\{x_{\text{in}}^{\text{pr}}; x_{\text{fin}}^{\text{pr}}\right\} \Big| T_{\text{от}} = const, L(\omega_j) = n\cdot T_{\text{pt}}, n = \overline{1,N}, \\ \hat{x}(t) = \sum_{i=1}^{p} \hat{a}_i \cdot x(t-i), t = \overline{x_{\text{in}}^{\text{pr}}; x_{\text{fin}}^{\text{pr}}}, p = \overline{10,14}, \\ \hat{a}_i(\omega_j), i = \overline{h,p} \Big| Y = \max, \end{cases} \qquad (13)$$

where $\omega_j$ – $j$-th VS segment, $L(\omega_j)$ — is the length of the $j$-th VS segment, $N$ – is the number of allocated periods of the PT in the VS segment, $h$ – is the highest initial order of autoregression, $Y$ – is the accuracy indicator of speaker recognition.

The determination of a specific AC set which will be informative speech features can be implemented by various methods. In our opinion, it seems appropriate to use the AC justification of such a method that allows us to simultaneously solve two problems – a priori estimation of the separation power of informative speech features and the construction of a classification rule for speakers. This property is fully possessed by the discriminant data analysis.

Among the possible AC sets of higher-orders, the best is the set that has minimal redundancy, maximum discriminant power, and classification accuracy (Table 3).

To select the minimum set of informative features, a tolerance indicator $1-r_{\hat{a}_i,\hat{a}_1\ldots\hat{a}_j\ldots\hat{a}_p}$, is used,where $r_{\hat{a}_i,\hat{a}_1\ldots\hat{a}_j\ldots\hat{a}_p} = \sqrt{1-\frac{|\mathbf{R}|}{|\mathbf{R}_{11}|}}, i=\overline{1,p}, i \neq j$ is the multiple correlation coefficient, $\mathbf{R}$ – is the AC correlation matrix.

Tolerance is a measure of the redundancy of a variable in a model, therefore, the greater its value, the less redundant the analyzed AC. The definition of the lower limit of tolerance was made on the basis of the properties of the multiple correlation coefficient and amounted to 0.3.

For example, Tables 3 and 4 show the results of evaluating the redundancy of the AC initial set $\hat{a}_7-\hat{a}_{12}$ and the set $\hat{a}_8-\hat{a}_{11}$, that was obtained after eliminating the excess coefficients $\hat{a}_7$ and $\hat{a}_{12}$. Table 4 shows that the second set is not redundant.

TABLE 3. AN EXAMPLE OF THE AC REDUNDANT SET CALCULATED BY THE SEGMENT MODEL OF THE SOUND *ă* IN THE WORD *ăn*

| N=500 | Results of the analysis of discriminant functions<br>Variables in the model: 6; group.: Speaker (10gr.)<br>Wilks' lambda: .00004 abt. F (54.2477)=288.44 p <0.0000 | | | | | |
|---|---|---|---|---|---|---|
| | **Wilks' lambda** | **Private lambda** | **F-expulsion (9.485)** | **P-level** | **Tolerance** | **1- tolerance (R-sq.)** |
| $\hat{a}_{11}$ | 0.000159 | 0.252119 | 159.8553 | 0.000000 | 0.992158 | 0.007842 |
| $\hat{a}_{9}$ | 0.000088 | 0.454148 | 64.7704 | 0.000000 | 0.995123 | 0.004877 |
| $\hat{a}_{10}$ | 0.000436 | 0.091728 | 533.5973 | 0.000000 | 0.986232 | 0.013768 |
| $\hat{a}_{7}$ | 0.000040 | 0.991295 | 0.4732 | 0.892698 | 0.000035 | 0.999965 |
| $\hat{a}_{12}$ | 0.000041 | 0.984736 | 0.8353 | 0.583771 | 0.000026 | 0.999974 |
| $\hat{a}_{8}$ | 0.000145 | 0.275083 | 142.0119 | 0.000000 | 0.987771 | 0.012230 |

TABLE 4. AN EXAMPLE OF THE AC NON-REDUNDANT SET
CALCULATED BY THE SEGMENT MODEL OF THE SOUND IN THE WORD

| N=500 | Results of the analysis of discriminant functions<br>Variables in the model: 4; group.: Speaker (10gr.)<br>Wilks' lambda: .00004 abt. F (36.1826)=701.20 p <0.0000 | | | | | |
|---|---|---|---|---|---|---|
| | **Wilks' lambda** | **Private lambda** | **F-expulsion (9,487)** | **P-level** | **Tolerance** | **1- tolerance (R-sq.)** |
| $\hat{a}_{8}$ | 0.000149 | 0.275758 | 142.1155 | 0.000000 | 0.992911 | 0.007089 |
| $\hat{a}_{9}$ | 0.000090 | 0.453074 | 65.3200 | 0.000000 | 0.995838 | 0.004162 |
| $\hat{a}_{10}$ | 0.000445 | 0.092043 | 533.7809 | 0.000000 | 0.990489 | 0.009511 |
| $\hat{a}_{11}$ | 0.000162 | 0.252499 | 160.1908 | 0.000000 | 0.994934 | 0.005066 |

Assessing the power of discrimination of non-redundant ACs is based on a step-by-step discriminant analysis with exceptions based on statistics of the following form [18]:

$$F = \frac{n-g-s}{g-1}\cdot\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}, \qquad (14)$$

where $\Lambda = |\mathbf{W}| / |\mathbf{T}|$, **W** and **T** – are the intra-group and inter-group correlation matrices of the AC, respectively, $s$ – is the number of AC, $n$ – is the number of VS realizations for all speakers.

Then, when conditions of the following form are satisfied:

$$F > F_{\text{expul}}(\alpha, \nu_1 = g-1, \nu_2 = n-g-s) \qquad (15)$$

discriminant power of the AC is significant. Otherwise, an insignificant AC must be excluded from the list of informative features.

## V. QUALITY ASSESSMENT OF SPEAKER RECOGNITION

Assessing the quality of speaker recognition on the basis of the generated informative ACs will be considered using the classification problem as an example.

The quality of the speakers classification can only be assessed as posteriori. For this purpose, at the training stage, informative ACs are calculated for each PW (password word) implementation for all the speakers registered in the system. After that, the average AC values for each *j*-th password word and *k*-th speaker $\bar{a}_{jk}$ .

At the classification stage, the AC values for the target speaker $\hat{a}_h, \hat{a}_{h+1}, ..., \hat{a}_p$ are calculated. The obtained values using the selected proximity measure must be compared with the average AC values calculated at the training stage and, based on the results, decide on assigning the target speaker to a specific class. To assess the degree of proximity of AC values calculated for the target speaker to their reference values obtained at the training stage, it is necessary to use some measure to find the distance between two points in the multidimensional space of AC values. In the general case, it is advisable to use the Mahalanobis distance, since, firstly, the standard deviations of the AC values may be unequal, and secondly, these values can be correlated. To calculate the proximity degree of the target speaker to a particular class, an expression of the following form is used:

$$D^2(\mathbf{A}|\mathbf{C}_k)=\sum_{i=h}^{p}\sum_{j=h}^{p}(w^{-1})_{ij}(\hat{a}_{ik}-\bar{\hat{a}}_{ik})(\hat{a}_{jk}-\bar{\hat{a}}_{jk}) ,(16)$$

where $D^2(\mathbf{A}|\mathbf{C}_k)$ – is the square of the distance from the AC values vector **A** of the target speaker to the center of the class of vectors characterizing the *k*-th speaker, $(w^{-1})_{ij}$ – is the element of the matrix inverse to the intragroup covariance AC matrix, $\hat{a}_{ik}$ – is the value of the *i*-th AC in the class *k*, $\bar{\hat{a}}_{ik}$ – is the average value of the *i*-th AC in class *k*.

In the particular case when the intragroup covariance matrix is single, the Mahalanobis distance is the Euclidean distance.

According to expression (16), the degree of proximity of the target speaker to each speaker registered in the system is estimated, and the target speaker is assigned to the class for which the value of the proximity measure is minimal. Then the indicator of the classification quality will be the proportion of correctly assigned AC vectors to the corresponding class of speakers. It is clear that the closer its value is to 1, the more accurate is the separation of speakers.

To check the quality of speaker recognition based on the selected ACs, we analyzed the classification matrix, which contains information on the number and percentage of correctly classified observations in each group. For example, Tables 5 and 6 show the results of applying the developed procedure for comparing a set of coefficients $\hat{a}_7-\hat{a}_{12}$, calculated for the sound $\breve{a}$ in the word $\breve{a}n$ and a set $\hat{a}_8-\hat{a}_{11}$ obtained after removing excess coefficients $\hat{a}_7$ and $\hat{a}_{12}$.

TABLE 5. SPEAKERS CLASSIFICATION RESULTS BY AC $\hat{a}_7-\hat{a}_{12}$ CALCULATED BY THE SEGMENT MODEL OF THE SOUND $\breve{a}$ IN THE WORD $\breve{a}n$

| Group | Classification matrix<br>Line: observable classes<br>Columns: predicated classes | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **accuracy rate** | **speaker 1** | **speaker 2** | **speaker 3** | **speaker 4** | **speaker 5** | **speaker 6** | **speaker 7** | **speaker 8** | **speaker 9** | **speaker 10** |
| **speaker 1** | 96.08 | 49 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **speaker 2** | 87.75 | 0 | 43 | 0 | 0 | 4 | 0 | 2 | 0 | 0 | 0 |
| **speaker 3** | 84.00 | 0 | 0 | 42 | 2 | 0 | 0 | 0 | 5 | 1 | 0 |
| **speaker 4** | 88.00 | 0 | 4 | 2 | 44 | 0 | 0 | 0 | 2 | 2 | 0 |
| **speaker 5** | 88.00 | 2 | 2 | 0 | 0 | 44 | 0 | 0 | 0 | 0 | 0 |
| **speaker 6** | 98.00 | 1 | 0 | 0 | 0 | 0 | 49 | 0 | 0 | 0 | 0 |
| **speaker 7** | 98.00 | 0 | 1 | 0 | 0 | 0 | 0 | 49 | 0 | 0 | 0 |
| **speaker 8** | 88.00 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 44 | 0 | 0 |
| **speaker 9** | 98.00 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 49 | 0 |
| **speaker 10** | 100.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 50 |
| **Total** | 92.60 | 52 | 49 | 50 | 47 | 48 | 50 | 51 | 51 | 52 | 50 |

TABLE 6. SPEAKERS CLASSIFICATION RESULTS BY AC $\hat{a}_8 - \hat{a}_{11}$ CALCULATED BY THE SEGMENT MODEL OF THE SOUND *ă* IN THE WORD *ăn*

| **Group** | Classification matrix<br>Line: observable classes<br>Columns: predicated classes | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **accuracy rate** | **speaker 1** | **speaker 2** | **speaker 3** | **speaker 4** | **speaker 5** | **speaker 6** | **speaker 7** | **speaker 8** | **speaker 9** | **speaker 10** |
| **speaker 1** | 98.04 | 50 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **speaker 2** | 97.96 | 0 | 48 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **speaker 3** | 100.00 | 0 | 0 | 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **speaker 4** | 100.00 | 0 | 0 | 0 | 50 | 0 | 0 | 0 | 0 | 0 | 0 |
| **speaker 5** | 100.00 | 0 | 0 | 0 | 0 | 50 | 0 | 0 | 0 | 0 | 0 |
| **speaker 6** | 100.00 | 0 | 0 | 0 | 0 | 0 | 50 | 0 | 0 | 0 | 0 |
| **speaker 7** | 100.00 | 0 | 1 | 0 | 0 | 0 | 0 | 49 | 0 | 0 | 0 |
| **speaker 8** | 100.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 50 | 0 | 0 |
| **speaker 9** | 100.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 50 | 0 |
| **speaker 10** | 100.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 50 |
| **Total** | 99.40 | 50 | 49 | 50 | 50 | 50 | 50 | 51 | 50 | 50 | 50 |

It can be seen from the tables that the quality of the speakers classification using AC $\hat{a}_8 - \hat{a}_{11}$ (Tab. 6) significantly exceeds the quality of the speakers classification based on the initial set $\hat{a}_7 - \hat{a}_{12}$ (Tab. 5).

In case of significant errors in the speakers classification, it is advisable to analyze the sets of autoregression coefficients in order to identify the observations that caused these deviations.

If there are incorrect classifications, they must be excluded from the training set of VS implementations. The procedure of excluding from training samples is that there is the AC set which should be excluded from the sample and the number of its belonging to this group is removed from the table of initial data, after which the process of assessing the quality of speakers classification is repeated. When a regular observation is deleted from the class, new incorrectly assigned coefficient vectors may appear, which were taken into account as correctly assigned before removal. The procedure of excluding observations must be continued until the classification accuracy indicator reaches its maximum value. This approach allows, when forming informative ACs, identifying and excluding VS implementations for each speaker which for various reasons (illness, emotional state, etc.) differ from the others in this class.

## VI. CONCLUSION

The article presents an approach to modeling VS to solve the speaker recognition problem. It is shown that as informative features characterizing speakers one can use the parameters of autoregressive time series models describing voice signals.

An algorithm is proposed for VS automatic segmentation into quasistationary sections based on interval estimation of speech samples standard deviation. At the same time, to solve the speaker recognition problem, segments are allocated to which the unchanged PTF and maximum energy correspond, since they contain the basic information about the features of the speaker's features.

It is demonstrated that the segments formed on the basis of the developed algorithm are stationary time series, and it allows using autoregressive models of various orders to describe them. In order to reduce the uncertainty in the formation of the decisive rule for speaker recognition, it is proposed to include only higher-order ACs in the model, since they are the ones that characterize VS high-frequency variations and contain basic information about the speaker's features. The possibility of using multivariate discriminant analysis to substantiate an AC set of coefficients is shown.

The results of assessing the quality of speakers classification allow us to conclude that it is possible to use the described approach to create speaker recognition automatic systems for the Vietnamese language.

## REFERENCES

[1]. Sorokin V. N. “Voice recognition: analytical review” V. N. Sorokin, V. V. Vyugin, A. A. Tananykin, Information processes,.Vol. 12, No. 1. pp. 1–13, 2012.

[2]. Pervushin E. A. “Review of the main methods of speaker recognition” / E. A. Pervushin // Mathematical structures and modeling,, No. 3 (24), pp. 41–54, 2011.

[3]. I. Rohmanenko. “Algorithms and software for verifying an announcer using an arbitrary phrase: thesis ... cand. tech. sciences”. [Electronic resource]. URL: https://postgraduate.tusur.ru /system/file_copies/ files / 000/000/262 / original / dissertation.pdf Tomsk, , 111 pp. 2017.

[4]. Ahmad K. S. A “unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network” // Advances in Pattern Recognition (ICAPR), Eighth International Conference on..,pp.16. 2015

[5]. Markel, J. D. Linear Prediction of Speech: [trans. from English.] / J. D. Markel, A. H. Gray; under the editorship of Yu.N. Prokhorov and V. S. Zvezdin. – Moscow: Communication, 308 p,1980.

[6]. Lysak A. B. Identification and authentication of a person: a review of the basic biometric methods of user authentication of computer systems / A. B. Lysak // Mathematical structures and modeling.. No. 2 (26). – pp. 124–134,2012.

[7]. Meshcheryakov R. V. Algorithms for evaluating automatic segmentation of a speech signal / R. V. Meshcheryakov, A. A. Konev // Informatics and Control Systems.– No. 1 (31). – pp. 195–206. 2012.

[8]. Ding J., Yen C. T. Enhancing GMM speaker identification by incorporating SVM speaker verification for intelligent web-based speech applications // Multimedia Tools and Applications.. – Vol. 74. – No. 14. – pp. 5131-5140, 2015.

[9]. Trubitsyn VG Models and algorithms in speech signal analysis systems: dis. ... cand. tech. sciences. – Belgorod, 2013 .– 134 pp. [Electronic resource]. URL: http://dissercat.com/content/modeli-i-algoritmy-v-sistemakh-analiza-rechevykh-signalov.

[10]. Ganapathiraju A., Hamaker J., Picone J., Doddington G.R. and Ordowski M. Syllable-Based Large Vocabulary Continuous Speech Recognition. IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 4, pp. 358–366, 2001.

[11]. Tomchuk K. K. Segmentation of speech signals for tasks of automatic speech processing: dis. cand. tech. sciences. – St. Petersburg.– 197 pp. [Electronic resource]. URL: http://fs.guap.ru/dissov/tomchuk_kk/full.pdf, 2017.

[12]. Sorokin V. N. Segmentation and recognition of vowels / V. N. Sorokin, A. I. Tsyplikhin // Information Processes. Vol . 4. – No. 2. – pp. 202–220, 2004.

[13]. Nguyen An Tuan Automatic analysis, recognition and synthesis of tonal speech (based on the material of the Vietnamese language): dissertation ... Doctors of technical sciences. – Moscow– 456 pp. [Electronic resource]. URL: https: // dissercat.com/content/avtomaticheskii-analiz-raspoznavanie-i-sintez-tonalnoi-rechi-na-materiale-vetnamskogo-yazyka, 1984.

[14]. Gmurman V. Ye. Probability theory and mathematical statistics: textbook. manual for universities / V. E. Gmurman. – 12th ed., Revised. – M.: Yurayt, 2010 .– 478 p.

[15]. Boxing J., Jenkins G. Time Series Analysis / Per. from English; Ed. V.F. Pisarenko. M .: Mir, 1974.– 406 pp.

[16]. Kantorovich, G. G. Analysis of time series // Moscow, 2003. – 129 pp. [Electronic resource]. URL: http: // biznesbooks.com/components/com_jshopping/ files / demo_products / kantorovich-g-g-analiz-vremennykh-ryadov.pdf.

[17]. Novikov E.I. Parameterization of a speech signal based on autoregressive models / E.I. Novikov, Do Kao Khan, // XI All-Russian Interdepartmental Scientific Conference "Actual problems of the development of security systems, special communications and information for the needs of public authorities of the Russian Federation Federations”: materials and reports (Oryol, February 5-6, 2019). At 10 hours / under the general editorship of P. L. Malyshev. – Eagle: Academy of the Federal Security Service of Russia, –pp. 127–130, 2019..

[18]. Kim J.-O. Factor, discriminant and cluster analysis: Per. from English / J.-O. Kim, C.W. Muller, W.R. Kleck and others; Ed. I.S. Enyukova. – M.: Finance and Statistics, – 215 pp, 1989.

## ABOUT THE AUTHORS

**PhD. Evgeny Novikov**

Workplace: The Academy of Federal Guard Service of the Russian Federation.

Email: nei05@rambler.ru

The education process: Received his Ph.D. degree at the Research Institute of Radio-Electronic Systems of the Russian Federation in Sep 2010.

Research today: modeling of random processes, statistical data processing and analysis, decision-making.

The education process: received his Ph.D. degree in Engineering Sciences in Academy of Federal Guard Service of the Russian Federation in Dec 2013.

Research today: information security, unauthorized access protection, mathematical cryptography, theoretical problems of computer science.

**PhD. Vladimir Trubitsyn**

Workplace: The Academy of Federal Guard Service of the Russian Federation.

Email: gremlin.kop@mail.ru

The education process: received his Ph.D. degree at Belgorod technical University of the Russian Federation in Dec 2014.

Research today: modeling of random processes, information and coding theory, voice signal processing and analysis.