Precision automated phonetic analysis of speech signals for information technology of text-dependent authentication of a person by voice

Oleg Bisikalo[a], Olesia Boivan[b], Nina Khairova[c], Oksana Kovtun[b] and Viacheslav Kovtun[a]

[a] *Vinnytsia National Technical University, Khmelnitske Shose str., 95, Vinnytsia, 21000, Ukraine*
[b] *Vasyl' Stus Donetsk National University, 600-richchya str., 21, Vinnytsia, 21000, Ukraine*
[c] *National Technical University "Kharkiv Polytechnic Institute", 2, Kyrpychova str., Kharkiv, 61002, Ukraine*

Abstract

A model of the process of phonetic analysis of speech signals in the frequency and temporal spaces is highlighted in the article for the first time. The generalization of the spectral characteristics of the studied speech signals is formalized in the represented model as an optimization task of minimizing the functional of relative entropy in contrast to the existing models. The obtained mathematical apparatus made it possible to formulate metric for quantitative estimation of the quality of the phonetic analysis results and to propose an adaptive method of automated phonetic analysis with an integrated mechanism for counteracting the influence of Gaussian-type noise, found in the studied speech signal, on the final result. The adequacy and functionality of the proposed model and method have been proved empirically. The analysis of the experiments results also showed that it is possible to assess the suitability of the studied speech materials for the task of authenticating a person by voice or speech recognition, focusing on the value of the coefficient of variability, which is included in the metric proposed by the authors and determined for the studied database of phonograms with recordings of voiced syllables of speech. Also, the values of this coefficient determined for the studied phonemes can be used to estimate the degree of their vocalization.

Keywords

Automated phonetic analysis, clustering of language units, computational linguistics, information technology, authentication of a person by voice.

1. Introduction

Modern methods of computational linguistics [1-5] are created with a focus on the use of technologies that process speech material automatically without constant human control. However, it requires upgrading of the computer speech technologies to a fundamentally new level, which can be achieved only by complete automation of the process of phonetic analysis of speech signals. Phonemes form the basic level of language description and determine its information and communicative characteristics. It is confirmed, in particular, by the method of forming speech corpora, in which phonograms of speech signals are accompanied by their transcription, which is nothing more than a sequence of phonemes. However, like any physiological process, speech is characterized by considerable variability, so there are a great number of options for phonemes sounding. This circumstance explains the fact that no any theoretical and software complex has been

IntelITSIS'2021: 2nd International Workshop on Intelligent Information Technologies and Systems of Information Security, March 24–26, 2021, Khmelnytskyi, Ukraine
EMAIL: obisikalo@gmail.com (O. Bisikalo); olesiaboivan@gmail.com (O. Boivan); khairova@kpi.kharkov.ua (N. Khairova); o.kovtun@donnu.edu.ua (O. Kovtun); kovtun_v_v@vntu.edu.ua (V. Kovtun)
ORCID: 0000-0002-7607-1943 (O. Bisikalo); 0000-0002-3512-0315 (O. Boivan); 0000-0002-9826-0286 (N. Khairova); 0000-0002-9139-8987 (O. Kovtun); 0000-0002-7624-7072 (V. Kovtun)

created yet for the effective automatic phonetic analysis of coherent speech, although its necessity is extremely urgent.

So the object of the research presented in the article is the process of clustering centers-phonemes in the spectral representation of the speech signals which are studied.

2. State-of-the-Art

The mechanism of automated phonetic analysis is an obligatory structural unit of information technology focused on solving such fundamental and applied phonetic problems as, for example, sounds recognition, the study of suprasegmental linguistic characteristics, speech synthesis, ets. As it was noted the final solution of computer phonetic analysis has not been found yet, but its relevance encourages research teams to creative search.

There are a number of studies [6-9] based on the mathematical apparatus of digital signal processing. Inside of them, speech signals are analyzed directly, without taking into account their physiologically determined phonetic structure. The speech signal is interpreted as a non-stationary multi-frequency signal and processed in order to determine the transmission function that generates it. After its determination, the transition from signal analysis in temporal space to its study in frequency space is carried out, where phonemes are determined, as a rule, on the basis of summarizing the results of analysis of the energy of a signal. For the transition from the temporal to the frequency space, as a rule, variations of the Fourier transformation, based on linear perceptual coefficients or the wavelet-transformations are used. These signal processing methods are mentioned in order to increase information content and computational complexity. The advantage of such studies is their strict mathematical adequacy, but they completely ignore the physiological mechanism of the speech signal generation, therefore their automatic application for phonetic analysis demonstrates the average quality results.

The second direction of research assumes the presence of a priori information about the transmission function of the articulatory tract. In this direction, let us single out the methods [10, 11] for representing a speech signal as a vector of states of an a priori given dynamic system (articulatory tract) with the help of a recursive filter, which allows to smooth out emissions at formant frequencies in time. In this context, an improved version of the method of phonetic-format analysis of the structure of a speech signal using linear perceptual coefficients with additional smoothing using the Newton-Raphson algorithm is also proposed [12]. The studies [13, 14] describe the PRAAT algorithm, which automatically finds the smoothest formant trajectory for short segments of the speech signal. The method is based on a variational polynomial approximation of short-time fragments of speech signals with the subsequent selection of the smoothest of them using the appropriate criterion. The method is fully automatic. However, the effectiveness of all these methods is mainly determined by the reliability of the applied a priori information.

There are many well-known methods based on the acoustic-frequency model of the vocal tract, created as a result of studies of the acoustic-physiological direction [15-17]. However, this model was created as a tool for synthesizing speech signals, so its application for their analysis did not show outstanding results in terms of quality. The direction [18] of research, based on the analysis of speech signals with the interpolation of the appearance of phonemes based on the data of the energy peaks of the formants of the studied signal in the passband is known. To increase the effectiveness of these methods, corpora of regional dialects of North America has been created, which contains 134000 formants identified by human experts. However, the effectiveness of these methods is generally determined by the presence of this specialized corpora. A fundamentally new direction of research is the study of the representation of phonemes by neurolinguistic structures of the human brain [18, 19]. In this context, phonetic analysis can also be viewed as a new tool for studying psychological and physiological phenomena. The methods that implement this concept are based on machine learning algorithms. Their effectiveness is completely determined by the representativeness of the training information used, the generalization of which has just begun.

So, let’s try to take into account the strengths and weaknesses of the above mentioned methods by defining the subject of study as the methods of acoustic theory of speech formation and information theory, the results of which will be generalized using the methods of probability theory and mathematical statistics.

3. Materials and Methods

3.1 Statement of research

According to the provisions of the complex theory of phonation, the quantum of oral speech is the phoneme, the number of which is finite and different for all languages. It is the combination of phonemes that forms the semantic quantum of speech – the morpheme.

Neurolinguistic research suggests that despite the psycho-physiologically determined variability of phoneme pronunciation, in the human mind in the process of learning the appropriate language for each r-th phoneme formed a cluster of its speech pattern with center x_r^*, $x_r^* \in X_r = \{x_{rj}\}$, $j = \overline{1, J_r}$, $r = \overline{1, R}$, where x_{rj} is the j-th allophone of the r-th phoneme; J_r is the power of the studied set of allophones of the r-th phoneme; R is the total number of phonemes in the studied language. Accordingly, in the process of perception of a speech signal $X(t)$ by a person in discrete time, the first is represented by a set of the characteristic vectors $x(t)$ extracted from the corresponding segment of the original speech signal of duration Δt. The value of Δt is chosen so as to consider the fragment of the speech signal limited in this way to be quasi-stationary with a duration of $\tau \in [10, 20]$ ms greater than the average duration of the phonemic utterance in the studied language. Then the phonetic analysis of the speech signal $X(t)$, represented by the set $x = (x_1.x_2, \ldots, x_L)$ of characteristic vectors $x(t)$, $t = 1, 2, \ldots, L$, will be understood as the task of assigning each segment $x(t)$ to one of the classes from the set X_r: $x(t) \equiv x_v^* \in X_v$, where X_v is a subset of the set X_r determined as a result of classification, $v \le R$. This classification task can be solved using machine learning methods or methods of information theory and mathematical statistics, for example, based on the functional of relative entropy [20].

Let the density of probability distribution P_x of the multidimensional matrix x belong to some set of alternative multidimensional densities of distributions of probability distributions P_r defined on a finite set $\{X_r\}$ of phonemes of the studied language: $P_x \in \{P_r\}$. Therefore, the task of phonetic analysis of the speech signal represented by the set x can be reduced to finding such a probability density from a certain set of alternatives $\{P_r\}$, the difference of which from P_x is minimal according to the selected metric μ. If we consider the distribution law of the appearance of each phoneme normal with zero mathematical expectation and autocorrelation matrix K_r of dimension $n \times n$, $n \le L$: $P_r = Norm(0, K_r)$, then a necessary condition for solving the above task is to calculate the set of differences $\mu(P_x | P_r)$ between the empirical distribution P_x and each alternative from the set $\{P_r\}$. When studying speech signals in frequency space, a characteristic parameter for estimating the desired differences for the set of alternatives $\{P_r\}$ is the set of values of power spectral densities $\{G_r(f)\}$, $\forall f \in \{\overline{1, F}\}$ $G_r(f) > 0$, where F is the upper limit of the frequency range for the empirical speech signal. We take this into account by defining the required estimates of differences as a functional of relative entropy:

$$\mu(P_x | P_r) = \frac{1}{2F} \int_{-F}^{F} \left(\frac{G_x(f)}{G_r(f)} - \ln \frac{G_x(f)}{G_r(f)} - 1 \right) df , \qquad (1)$$

where $r = \overline{1, R}$, $G_x(f)$ is the estimate of the power spectral density of the empirical speech signal represented by the set x.

It is possible to identify a stationary stochastic process P_r in metric (1) by means of spectral analysis of data $\{x_{rj}, r = \overline{1, R}, j = \overline{1, J}\}$:

$$G_r(f)=\frac{1}{J}\sum_{j=1}^{J}G_{rj}(f),\ r=\overline{1,R}. \quad (2)$$

According to Wiener-Hitchkin theorem, the estimation of the power spectral density (2) is related to the autocorrelation matrix K_r of the empirical speech signal x by the discrete Fourier transform. However, such an estimate is possible only under the condition $n>>1$, while the estimation of the power spectral density $G_r(f)$ for a small amount of experimental data is of practical interest: $J<<\infty$. So, the aim of the study is the analytical formalization of the process of phonetic analysis of a speech signal in frequency and temporal spaces based on the functional of relative entropy, oriented towards the use of text-dependent voice authentication in information technology. The ***objectives*** of the research are: - creation of a model of the process of phonetic analysis of a speech signal in frequency and temporal spaces; - formalization of the metric for estimation of the quality of the results of the process of phonetic analysis based on the functional of relative entropy; - formalization of the adaptive method of automated phonetic analysis focused on achieving optimal results, assessed in the created metric; - empirical proof of the adequacy of the created mathematical model and analysis of the functionality of the created method.

3.2 Model of phonetic analysis of speech signals in temporal and frequency spaces

A fundamental issue for the analytical formalization of the process of phonetic analysis of speech signals is to determine the centers of clusters $x_r=x_{rv}=v_r$, $v\in\overline{\{1,J\}}$, for empirical realizations of $\{x_{rj}\}$, $j\in\{1,J\}$ in the metric $\mu(x_{rj})$:

$$v_r=\arg\min_{i\in[1,\ldots,J]}\sum_{j=1}^{J}\mu_i(x_{rj}) \quad (3)$$

where characteristic

$$\mu_i(x_{rj})=\frac{1}{2F}\int_{-F}^{F}\left(\frac{G_{rj}(f)}{G_{ri}(f)}-\ln\frac{G_{rj}(f)}{G_{ri}(f)}-1\right)df,\ i,j\in\overline{\{1,J\}}, \quad (4)$$

is a functional of relative entropy between i-th and j-th parametric interpretations of allophones of r-th phoneme in frequency space. Semantic generalization of expressions (1) and (4) allows us to determine the frequency range where the center of the cluster of the r-th phoneme is as

$$G_r^{(J)}=G_{rv}(f),\ v\le J,\ r=\overline{1,R}. \quad (5)$$

However, with the identified autocorrelation matrix K_r it is possible to analytically formalize the analysis of the studied speech signals in temporal space analog of criterion (3), the classification decision in which is made on the basis of a set of values of statistics defined by expression

$$\rho_r(x)=\frac{1}{2n}\left(tr\left(\frac{\hat{K}}{K_r}\right)-\log\left|\frac{\hat{K}}{K_r}\right|-n\right),\ r=\overline{1,R}, \quad (6)$$

where $\hat{K}$ is a selective estimate of autocorrelation matrices of the studied empirical speech signal $x=x(t)$, $t=1,2,\ldots,L$; $tr(A)$ is the trace of the matrix A. The sample estimate $\hat{K}$ is determined based on the following considerations. Let the speech pattern X_r of the r-th phoneme be determined on the basis of the analysis of the set of its utterances x_{rj}, $j=\overline{1,J_r}$: $X_r=\{x_{rj}\}$. In this case, each utterance x_{rj} is formed by a sequence of L samples $\{x_{rj}(t)\}$ obtained with the periodicity $T=(2F)^{-1}$. Divide this sequence into frames of duration n samples, $n<<L$, grouping them into a set of data vectors $\{x_{rji}\}$ of dimension $n\times L-n$.

Then the sample estimate of the hypothetical normal distribution is defined as the arithmetic mean in the form

$$\hat{K}_{rj} = \frac{1}{L-n}\sum_{i=1}^{L-n} x_{rji} x_{rji}^{T},\ j=\overline{1,J_r}, \tag{7}$$

where T symbolizes the transposition operation. Substituting the value of the sample estimate (7) into expression (6) we obtain for the pattern X_r a matrix of statistics of dimension $J_r \times J_r$:

$$\rho_{rjk} = \frac{1}{2n}\left(tr\left(\frac{\hat{K}_{rj}}{K_{rk}}\right) - \log\left|\frac{\hat{K}_{rj}}{K_{rk}}\right| - n \right),\ k,j=\overline{1,J_r}. \tag{8}$$

We find the sum of the values of the columns of the matrix (8): $\sum_{j=l}^{J_r}\rho_{rjk} = \rho_{rk}$, $k=\overline{1,J_r}$, and analytically formalize oriented on the description of the studied speech signal in the temporal space the analogous to criterion (3):

$$v_r = x_r^{*} = x_{r\theta} = \arg\min_{k}\rho_{rk},\ r=\overline{1,R}. \tag{9}$$

Determined according to expression (7) for $j=\theta$, the sample autocorrelation matrix $\hat{K}_{r\theta}$ for the center of the cluster x_r^{*} will determine the optimal decisive statistics when substituting in (6). After analyzing expressions (4) and (6), we can conclude that the entropy of the estimate of the center of the cluster of phoneme will decrease with increasing value of J.

Therefore, with the center of the cluster for the r-th phoneme determined by expression (3) or (9), it is possible to determine the optimal estimate of the power spectral density $G_r^{(J)}$ for this phoneme on the basis of expression (5). If such actions are implemented for all R phonemes of the studied language, then we obtain a phonetic-acoustic database, the universality of which will increase with increasing number of phonemes pronounced during the formation of the model, i.e. the parameter J.

3.3 Formalization of metrics for qualitative evaluation of the results of phonetic analysis of the studied language in the paradigm of the proposed model

Based on the provisions of the acoustic theory of speech formation, we present a model of the j-th utterance of the r-th phoneme by the autoregression function of the form

$$x_{rj}(l) = \sum_{s=1}^{S} a_{rj}(s)\, x_{rj}(l-s) + \eta_{rj}(l),\ l=1,2,\ldots,\ j=\overline{1,J},\ r=\overline{1,R}, \tag{10}$$

which is uniquely determined by the set of coefficients $\left\{a_{rj}(s), s=\overline{1,S}\right\}$ by power $S \le n$ and a variance σ_{rj}^2 of the generating Gaussian process $\left\{\eta_{rj}(l), l=1,2,\ldots\right\}$. A property of such a representation that is relevant for us is that the estimation of the power spectral density of the studied signal obtained on the basis of the autoregression model with a finite set of coefficients $\left\{a_{rj}(s), s=\overline{1,S}\right\}$ will always satisfy the condition of regularity: $G_r(f) > 0$. However, given the functional of relative entropy used by us for phonetic analysis, the possibility of normalizing the speech signals described by the autoregression model of the form (10) to the value of their specific entropy $h\left(x_{rj}\right) = 0.5\ln\sigma_{rj}^2$ is especially relevant for us and allows to achieve the desired level by the variance $\sigma_{rj}^2 = \sigma_0^2 = const$ of the generating process η_{rj}. Accordingly, if we take into account the fact that the variance σ_0^2 does not change when a person utters not just phonemes, but words or even short phrases, then expression (4) can be simplified to the form

$$\mu_i\left(x_{rj}\right)=\frac{1}{2F}\int_{-F}^{F}\left(\frac{G_{rj}(f)}{G_{ri}(f)}-1\right)df=\mu_{rji},\ i,j=\overline{1,J}. \tag{11}$$

When analyzing the speech signal in the temporal space analog of expression (11) will be the target adaptation of expression (6), namely:

$$\rho_i\left(x_{rj}\right)=\frac{\sigma_i^2\left(x_{rj}\right)}{\sigma_0^2}-1, \tag{12}$$

where the variance $\sigma_i^2\left(x_{rj}\right)$ is determined by expression

$$\sigma_i^2\left(x_{rj}\right)=\frac{1}{L-S}\sum_{l=S+1}^{L}\left(y_{rj}^{(i)}(l)\right)^2, \tag{13}$$

in which the parameter $y_{rj}^{(i)}(l)$ characterizes the change of the studied speech signal $x_{rj}=\left\{x_{rj}(l)\right\}$, $l\le L$, after its passage i-th bleaching filter

$$y_{rj}^{(i)}(l)=x_{rj}(l)-\sum_{s=1}^{S}a_{ri}(s)x_{rj}(l-s). \tag{14}$$

Expressions (11) and (12) through criteria (3) and (9), respectively, allow us to identify the optimal standards x_r for all studied phonemes $r\in R$. An additional positive point is that when applying criterion (9) to calculate statistics (12) a target bleaching filter (14) is used which allows to effectively reduce the sensitivity of the result of phonetic analysis to the potential presence of Gaussian noise in the empirical speech signal.

Modify the analytical form of criterion (3) taking into account expression (11), resulting in

$$\mu_{rv}=\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{2F}\int_{-F}^{F}\left(\frac{G_{rj}(f)}{G_{rv}(f)}-1\right)df\right)=\frac{1}{2F}\int_{-F}^{F}\left(\frac{G^{RE}(f)}{G_{rv}(f)}-1\right)df=\mu_{rv}^{RE}. \tag{15}$$

The value of the functional of relative entropy (15) substituted in expression (5) allows us to identify the optimal estimate of the power spectral density of the r-th phoneme, which is potentially more reliable than the estimate of maximum likelihood calculated by expression (2). This thesis can be rationally substantiated by the fact that the reliability of the estimate (2) depends only on the representativeness of the empirical data, whereas when calculating the estimate (5) based on the functional (15), firstly, according to expression (14), the empirical data get rid of potentially present in the studied speech signal Gaussian noise, secondly, the empirical data are further generalized by autoregressive models of the form (10), the reliability of which can be increased by increasing the order of the models s in the range from 1 to J inclusive.

In fact, the estimate (2) allows us to determine the circumference of a center of the cluster for the power spectral density of the r-th formant, while the estimate (5) taking into account expression (15) allows changing the value of the order of s to find a center of the cluster of the r-th phoneme as a result of the solution optimization task. Let's generalize the just stated concept by modifying expression (15) taking into account expressions (12)-(14). We obtain

$$\mu_{rv}^{RE}=\frac{1}{J}\sum_{j=1}^{J}\frac{\sigma_v^2\left(x_{rj}\right)}{\sigma_0^2}-1=\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\sigma_0^2(L-S)}\sum_{l=S+1}^{L}\left(y_{rj}^{(v)}(l)\right)^2-1\right)=\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{M}\chi_{rjM}^2\left(1+\mu_{rjN}\right)-1\right), \tag{16}$$

where $\chi_{rjM}^2(l)$ is the χ^2-distribution of the stochastic quantity l with $M=L-S$ degrees of freedom. The greatest influence on the value of μ_{rv}^{RE} calculated by expression (16) is caused by the variability of the characteristics of allophones of the r-th phoneme, which is generalized by the coefficient of variability $\overline{\mu}_{rji}=1+\mu_{rji}$. The value of this coefficient depends on the individual speech characteristics of the persons whose speech signals are being studied and can vary widely. To some extent, the asymptotic properties of the χ^2-distribution will smooth out these fluctuations:

$$\mu_{rv}^{RE}=\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{M}\chi_{rjM}^2\left(1+\mu_{rjv}\right)-1\right)\Bigg|_{M>>1}\approx\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{M}\left(1+\mu_{rjv}\right)Norm_j\left(M,2M\right)-1\right)= \tag{17}$$

$$= \frac{1}{J}\sum_{j=1}^{J}\left(Norm_j\left(\left(1+\mu_{rjv}\right), \frac{2}{M}\left(1+\mu_{rjv}\right)^2\right)-1\right) = \frac{1}{J}\sum_{j=1}^{J} Norm_j\left(\mu_{rjv}, \frac{2}{M}\left(1+\mu_{rjv}\right)^2\right) =$$

$$= Norm\left\{\frac{1}{J}\sum_{j=1}^{J}\mu_{rjv}, \frac{1}{MJ^2}\sum_{j=1}^{J}\left(1+\mu_{rjv}\right)^2\right\}.$$

Let us denote the standard deviation of the coefficient of variability $\bar{\mu}_r$ of the characteristics of allophones of the cluster X_r of the corresponding phoneme as

$$SD\left[\mu_r\right] = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(1+\mu_{rjv}\right)^2}\,. \tag{18}$$

Expression (18) is in fact a Gaussian model for stochastic estimation of the coefficient of variation μ_{rv}^{RE}. The mathematical expectation of the characteristic of variability of μ_r of values (11) within the cluster of the r-th phoneme is determined by the expression

$$M\left[\mu_r\right] = \frac{1}{J}\sum_{j=1}^{J}\mu_{rjN}\,. \tag{19}$$

The standard deviation of the characteristic of variability of values (11) within the cluster of the r-th phoneme is determined by the expression

$$\sigma_r = \sqrt{\frac{2\sigma_r^2}{MJ}} = \sqrt{\frac{2M\left(\mu_r\right)}{MJ}}\,. \tag{20}$$

The statistical meaning of the parameter σ_r is as follows – the larger the value of σ_r, the lower the density of the cluster of the r-th phoneme. Using expressions (18), (19) and (20), we define the confidence interval Δ_r of the spectral estimation of the r-th phoneme as

$$\Delta_r = 2z_p\sigma_r = \sqrt{\frac{8}{MJ}}z_p SD\left[\mu_r\right], \tag{21}$$

where z_p is the coefficient of proportionality, p is the confidence probability. For example, for the Gaussian distribution with $p = 0,95$ the tabular value of the coefficient z_p is equal to $1,96$.

The confidence interval (21) determines the reliability of the estimate (5) and, accordingly, criterion (1) for the spectral representation of the phonemes of the studied language. It is obvious that the variability of the spectral characteristics of the phonemes of the studied language will decrease with: - increasing the homogeneity of the studied speech material; - reducing the level of presence in the studied speech material of non-Gaussian noise; - increasing the number of persons-donors of speech material.

Note that the influence of the first factor analytically takes into account expression (21), the influence of the second factor analytically takes into account expression (14), while taking into account the outflow of the third factor on the variability of spectral characteristics of phonemes of the studied language is worth additional analytical explanations.

We introduce a modifier for the confidence interval described by expression (21), which will take into account the number of donors of speech material, which was studied by generalizing the spectral characteristics of the corresponding phonemes: $\Delta_r^{(I)}$, where $I = 1,2,\ldots$ is the index of donor of speech material. Accordingly, expression (21) was obtained for $\Delta_r^{(1)}$. For cases when $I = 2,3,\ldots$, based on expression (21) we obtain:

$$\delta_r\left(I\right) = \frac{\Delta_r^{(I)}}{\Delta_r^{(1)}} = \frac{\mu_r^{(I)}}{\mu_r^{(1)}} \geq 1\,. \tag{22}$$

Naturally, as I increases, the value of the coefficient δ_r will increase nonlinearly, eventually reaching saturation at $I = I^*$. The value of I^* will depend on the factors just mentioned, but the analytical formalization of this dependence requires additional theoretical research.

Therefore, the theoretical material presented in section 3.2 has found its application in the analytical formalization of the metric $\left\{\mu_r, \sigma_r, \Delta_r, \delta_r\right\}$, focused on the qualitative evaluation of the

results of phonetic analysis. The use of autoregressive analysis in deriving the functional of relative entropy μ_r allows to determine the spectral characteristics of the center of the cluster of the r-th phoneme as a result of solving the optimization task by changing the order s of regression models created to describe the studied speech signals. We can quantify the quality of the phonetic analysis performed in this way by calculating the value of the coefficient of variability δ_r.

4. Results

A group of 20 students from the Department of the Theory and Practice of Translation, Faculty of Foreign Languages, Vasyl' Stus Donetsk National University (Ukraine, Vinnytsia) was formed to conduct experiments. At the initial stage of the experiments, it was assumed that each student, using a microphone connected to a computer, would record phonograms with sequentially, repeatedly (ten times), at the same tempo, pronounced long English phonemes [i:], [a:], [u:], [ɔ:], [z:] (one phoneme - one phonogram). An AKG P420 microphone without an amplifier connected to a Creative Audigy Rx sound card integrated into a computer was used for the experiments. Sound recording processes were supported by Sound Forge Pro for Windows.

Phonograms were recorded with a sampling rate of 8000 Hz ($F = 4000$ Hz), quantization of 16 bits, mono and stored in .wav format. Subsequently, the phonograms were programmatically processed in order to form clusters for the corresponding phonograms $\{X_r\}$, $r = 1, 2, \ldots, R = 5$, and to determine the centers of the clusters based on model (3)-(11). Preliminary phonetic analysis of phonograms was performed using the Praat program developed by the Institute of Phonetic Sciences from the University of Amsterdam. To use the mathematical apparatus proposed in the article, phonograms were processed in frames with a duration of $L = 160$ (≈ 20 ms).

For spectral analysis of phonograms with speech signals using the autoregression model, the Berg's algorithm was used [21], known for its high resolution in the analysis of short-term signals and guaranteed stability of the calculated forming filter. To enable the comparison of power spectral densities according to criterion (1), the source speech material was presented in the Mel-space of the bank of the corresponding filters with triangular averaging functions.

As a result, the frequency characteristic parameters of the studied speech signals were obtained in the form of weighted sums of power spectral densities in uniform intervals lasting 55 mels (a total of 31 counts for overlapping the frequency range $[200, 3400]$ Hz).

As a result, 20 personalized phonetic databases $\{X_r\}$ of the same volume $R = 5$ were formed and also more than 1000 integrated phonetic databases $\{X_r\}_I$ were formed as a result of joint processing of phonograms of two, three, etc. students. For the primary set of phonetic data $\{\{X_r\}, \{X_r\}_I\}$, $r = \overline{1, R = 5}$, $I = \overline{2, 10}$, obtained as a result of the described actions, the values of the coefficient of variability were calculated by expression (15) and by expression (22) with a change in the order of the autoregression models used.

Empirical dependences of the value of the coefficient of reliability of the results of phonetic analysis δ_r on the number of students I, whose speech material was used to form the corresponding primary integral phonetic bases $\{X_r\}_I$, for phonemes [i:] and [a:] are presented in Figure 1.

Potentially, the experiment was oriented to arrange the phonemes of the studied language according to the level of their informativity for the task of authentication of the person by voice.

The second experiment aimed to recognize the isolated syllables of English words formed by as many studied phonemes as possible. The authors formed a working dictionary of 200 selected English words. Each of the students read the words from the working dictionary for recording in the phonogram. The word order for all recording procedures was the same. Each speaker-student repeated the reading-recording procedure five times.

Prerequisites for the reading process were: - clear diction; - stable pronunciation rate; - division of words into syllables with a clear fixation of pauses between the latter. Subsequently, the content of individual bases of phonograms was processed by the procedure of reducing variability (21).

Accordingly, individual databases of original and processed phonograms were formed for each student.

An integrated database of original phonograms was also formed by averaging the spectral characteristics of the content of individual databases of original phonograms. Subsequently, the content of the integrated database of original phonograms was processed by the method generalized (22), resulting in an integrated database of processed phonograms.

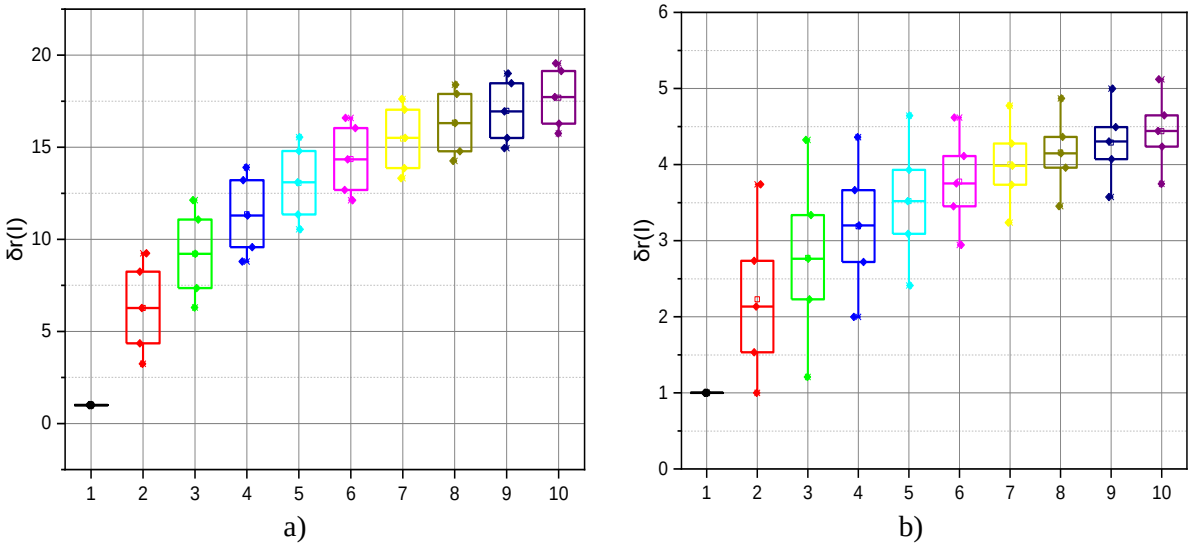

Figure 1: Dependence of the value of the coefficient of reliability of the results of phonetic analysis δ_r on the number of students for: a) phoneme [i:]; b) phonemes [a:]

Next, an iterative process was performed to add to the content of all databases of phonograms the Gaussian noise of such power as to obtain variants of all databases of phonograms with a signal-to-noise ratio of $5,10,15,\ldots,30$ dB, respectively. As a result, seven sets of personalized and integrated databases of original and processed phonograms were obtained. Speech recognition in the created sets of phonogram databases was carried out using the most popular currently professional APIs: Cloud Speech from Google and Microsoft Speech from Microsoft. The results of the experiments in the metric $\varepsilon(SNR)$ are shown in Figure 2.

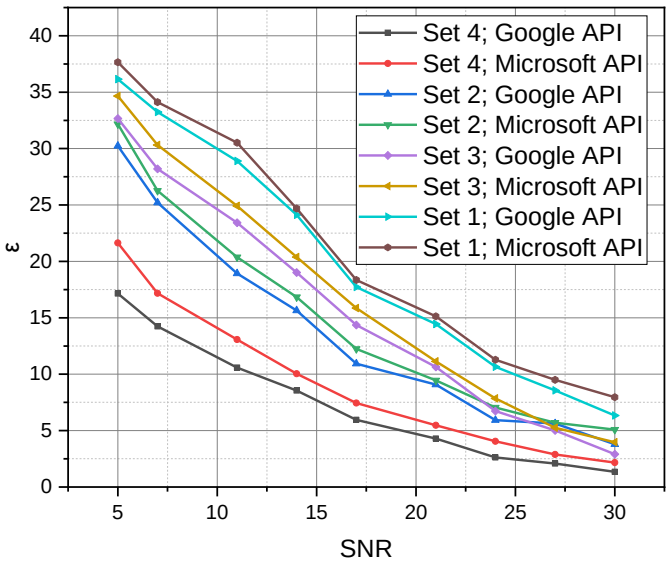

Figure 2: Dependence of the relative error of recognition of isolated syllables on the signal-to-noise ratio in the corresponding set of databases of phonograms

The relative error of recognition of isolated syllables for all sets of databases of phonograms (Set 1: personalized databases of processed phonograms; Set 2: personalized databases of processed phonograms; Set 3: database of integrated original phonograms Set 4: database of integrated processed phonograms) was calculated as the ratio of the absolute value of the difference between the number of correctly and incorrectly recognized syllables to the total number of syllables in the corresponding sets of databases of phonograms.

5. Discussion

From the empirical results of estimating the dependence of the value of the reliability coefficient of the results of phonetic analysis δ_r on the number of speaking students I shown in Figure 1 it is seen that the hypothesis formulated in the theoretical part of the article about the existence of phonetic data saturation threshold has been empirically confirmed. It is seen that the value of the characteristic $\delta_r(I) \to \sup \delta_r = \delta_r^*$ is limited from above by the value of δ_r^*, the value of which is essentially phonemic-dependent. Based on this fact, we can conclude that by changing the volume and method of forming an integrated database of phonograms of phonemes can be profiled for their intended use or in the task of authentication of the person by voice (low value of δ_r) or in the task of semantic analysis of text (high value of δ_r).

The received estimation will be not only qualitative, but also quantitative, which is especially relevant for information technology of text-dependent authentication of the person by voice, focused on the application in the structure of information system for critical use with authentication of the person-user by voice [22-26].

Note the values of the coefficients $\delta_r^{[i:]}$ and $\delta_r^{[a:]}$ shown in Figures 1a and 1b, respectively, for the same values of I. It is seen that the values of $\delta_r^{[i:]}$ are many times larger than the values of $\delta_r^{[a:]}$ and this tendency only increases with increasing I. Given that the value of the coefficient δ_r for the target phoneme characterizes the degree of density of its cluster from the volume and source of speech material, it can be argued that phonemes with a relatively high value of the coefficient δ_r (for the set $\left\{\delta_r^{[i:]}, \delta_r^{[a:]}\right\}$ is $\delta_r^{[i:]}$) carry more information about individual of speaker's voice. This should be taken into account when creating a representative dictionary of passphrases for information technology of text-dependent authentication of the person by voice.

Finally, we pay attention to the value of the confidence intervals for the values of the coefficients $\delta_r^{[i:]}$ and $\delta_r^{[a:]}$ shown in Figures 1a and 1b, respectively. Recall that the values of confidence intervals calculated by expressions (21), (22) depend on the order of the autoregression model (10) used to describe the studied speech signal and the degree of compensation of the influence of Gaussian noise present in the studied speech signal (14).

Accordingly, the low variability of the confidence interval for the phoneme [i:] indicates the high density of its cluster despite the different origin of the studied speech material and the potentially non-Gaussian form of the distribution function of the corresponding signal. In the context of the acoustic theory of speech formation, this is, in fact, a quantitative estimate of the degree of vocalization of this phoneme. At the same time, the high variability of the confidence interval for the phoneme [a:] (with increasing I from 2 to 10, the width of the confidence interval decreased by almost 20 times) can potentially indicate a significantly lower vocalization of this phoneme. Accordingly, the proposed mathematical apparatus provides a potential opportunity to organize the set of phonemes of the studied language by quantifying the degree of their vocalization.

Let's analyze the experimental results shown in Figure 2. It becomes obvious that the generalization of phonetic information, whether simple averaging of spectral characteristics in certain frequency ranges or analytically substantiated in Chapter 3 of the article generalization based on the value of the coefficient of reliability of phonetic analysis δ_r, has a positive effect on solving the task of automated recognition of syllables by the most modern specialized information systems. However,

it is seen that the effect of noise leads to a rapid nonlinear increase in the relative error of recognition of isolated syllables ε.

Depending on the studied set of databases of phonogram, the value of the relative error ε at the limit value of 5 dB of the studied range of the signal-to-noise ratio increased by 2.5-7 times. These results allow us to recommend the application of the generalized expression (14) approach to the filtering of Gaussian noise in the speech signal to empirical signals, the estimation of the level of the signal/noise ratio is in the range $[40,20]$ dB. It is also obvious that it is necessary to continue the search for more efficient methods of filtering or compensating for noise for processing empirical speech signals with a low signal-to-noise ratio – $SNR \in [15,5]$ dB.

Finally, the lowest values of the relative error of recognition of isolated syllables ε were obtained when working with the content of the integrated database of processed phonograms, which was obtained using the procedure proposed by the authors, generalized by expression (22). This result is an empirical proof of the adequacy of the mathematical apparatus presented in the article.

6. Conclusions

Without exaggeration, phonetic analysis is a "cornerstone", which is the basis of modern human-machine information technologies, focused on the target interpretation of speech signals. In particular, automated phonetic analysis is the basis of approaches to solving such tasks as authentication of the person by voice, speech recognition, determination of the speaker's emotional state, semantic interpretation of the text, etc. However, the quality of the results demonstrated by modern automated systems of phonetic analysis is inversely proportional to the amount of educational information available to them. Thus, the task of improving the quality of phonetic analysis in conditions of limited educational information is relevant.

A model of the process of phonetic analysis of speech signals in the frequency and temporal spaces is highlighted in the article for the first time. The generalization of the spectral characteristics of the studied speech signals is formalized in the represented model as an optimization task of minimizing the functional of relative entropy in contrast to the existing models.

The obtained mathematical apparatus made it possible to formulate metric for quantitative estimation of the quality of the phonetic analysis results and to propose an adaptive method of automated phonetic analysis with an integrated mechanism for counteracting the influence of Gaussian-type noise, found in the studied speech signal, on the final result.

The adequacy and functionality of the proposed model and method have been proved empirically. The analysis of the experiments results also showed that it is possible to assess the suitability of the studied speech materials for the task of authenticating a person by voice or speech recognition, focusing on the value of the coefficient of variability, which is included in the metric proposed by the authors and determined for the studied database of phonograms with recordings of voiced syllables of speech. Also, the values of this coefficient determined for the studied phonemes can be used to estimate the degree of their vocalization.

Further research is planned to be focused on the oriented to the task text-dependent authentication of the person by voice and on the phonetic analysis of the most common Germanic, Indo-European, Romance and Slavic languages with the help of the created theoretical and applied complex.

7. Acknowledgments

The authors note that the research results presented in the article were obtained while working on the research topic "Synchronic and Diachronic Studies of Language Units on Different Levels" of the Department of the Theory and Practice of Translation, Vasyl' Stus Donetsk National University (Ukraine, Vinnytsia).

The authors are grateful to the staff of this department for facilitating the research. The authors also thank the staff of the Department of Automation and Intelligent Information Technologies and the Department of Computer Control Systems of Vinnytsia National Technical University (Ukraine, Vinnytsia) for consulting in theoretical and applied aspects of the study.

8. References

[1] S.L. Kryvyi, N.P. Darchuk, A.I. Provotar, Ontological similar systems for analysis of texts of natural language, in Problems in programming, Iss. 2-3, 2018, pp. 132-139, doi: 10.15407/pp2018.02.132.

[2] O. Orobinska, J.-H. Chauchat, N. Sharonova, Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security), in 1st International conference Computational linguistics andintelligent systems (COLINS), Kharkiv, Ukraine, 2017, pp, 95-99.

[3] Y. Burov, V. Lytvyn, V. Vysotska and I. Shakleina, The Basic Ontology Development Process Automation Based on Text Resources Analysis, in 15th International Conference on Computer Sciences and Information Technologies (CSIT), Zbarazh, Ukraine, 2020, pp. 280-284, doi: 10.1109/CSIT49958.2020.9321910.

[4] A. Adala, N. Tabbane and S. Tabbane, A novel semantic approach for Web service discovery using computational linguistics techniques, in Fourth International Conference on Communications and Networking, ComNet-2014, Hammamet, 2014, pp. 1-6, doi: 10.1109/ComNet.2014.6840909.

[5] Y. Wang and R. C. Berwick, On formal models for cognitive linguistics, in 11th International Conference on Cognitive Informatics and Cognitive Computing, Kyoto, 2012, pp. 7-17, doi: 10.1109/ICCI-CC.2012.6311169.

[6] S. Hara and H. Nishizaki, Acoustic modeling with a shared phoneme set for multilingual speech recognition without code-switching, in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, 2017, pp. 1617-1620, doi: 10.1109/APSIPA.2017.8282284.

[7] C. Zhao, H. Wang, S. Hyon, J. Wei and J. Dang, Efficient feature extraction of speaker identification using phoneme mean F-ratio for Chinese, in 8th International Symposium on Chinese Spoken Language Processing, Kowloon, 2012, pp. 345-348, doi: 10.1109/ISCSLP.2012.6423485.

[8] J. M. McQueen and M. A. Pitt, Transitional probability and phoneme monitoring, in Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP ‘96, Philadelphia, PA, USA, 1996, pp. 2502-2505 vol.4, doi: 10.1109/ICSLP.1996.607321.

[9] S. Chen, B. Song, L. Fan, X. Du and M. Guizani, Multi-Modal Data Semantic Localization With Relationship Dependencies for Efficient Signal Processing in EH CRNs, in IEEE Transactions on Cognitive Communications and Networking, vol. 5, no. 2, 2019, pp. 347-357. doi: 10.1109/TCCN.2019.2893360.

[10] K. S. Sai Vineeth, V. Phaneendhra and S. Prince, Identification of Vowel Phonemes for Speech Correction Using PRAAT Scripting and SPPAS, in International Conference on Communication and Signal Processing (ICCSP), Chennai, 2018, pp. 0850-0853. doi: 10.1109/ICCSP.2018.8524273.

[11] B. Wang and C.-C. J. Kuo, SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models, in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, 2020, pp. 2146-2157. doi: 10.1109/TASLP.2020.3008390.

[12] S. Chakrasali, U. Bilembagi and K. Indira, Formants and LPC Analysis of Kannada Vowel Speech Signals, in 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 2018, pp. 945-948, doi: 10.1109/RTEICT42901.2018.9012641.

[13] M. Pleva, J. Juhár and A. S. Thiessen, Automatic Acoustic Speech segmentation in Praat using cloud based ASR, in 25th International Conference Radioelektronika (RADIOELEKTRONIKA), Pardubice, 2015, pp. 172-175, doi: 10.1109/RADIOELEK.2015.7129000.

[14] M. A. Kutlugün and Y. Şirin, Turkish meaningful text generation with class based n-gram model, in 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2018, pp. 1-4. doi: 10.1109/SIU.2018.8404801.

[15] B. Bharathi, S. Kavitha and S. Sugapriya, Bilingual Speech Recognition System for Isolated Words Using Deep Neural Network, in International Conference on Computer, Communication,

and Signal Processing (ICCCSP), Chennai, India, 2018, pp. 1-4, doi: 10.1109/ICCCSP.2018.8452832.
[16] L. Chen, H. Yang and H. Wang, Research on Dungan speech synthesis based on Deep Neural Network, in 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei City, Taiwan, 2018, pp. 46-50. doi: 10.1109/ISCSLP.2018.8706713.
[17] M. Fry, Modeling the Acquisition of Intonation: A First Step, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, 2018, pp. 5124-5128. doi: 10.1109/ICASSP.2018.8462541.
[18] Z. Zhou, X. Song, R. Botros and L. Zhao, A Neural Network Based Ranking Framework to Improve ASR with NLU Related Knowledge Deployed, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6450-6454, doi: 10.1109/ICASSP.2019.8682727.
[19] Y. N. Seitkulov, S. N. Boranbayev, B. B. Yergaliyeva, S. K. Atanov, H. V. Davydau and A. V. Patapovich, The base of speech structural units of Kazakh language for the synthesis of speech-like signals, in IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), Almaty, Kazakhstan, 2018, pp. 1-4. doi: 10.1109/ICAICT.2018.8747120.
[20] W. Zhu, J. Dai, J. Li, J. Wang and F. Hou, Analysis of α Wave in Normal and Epileptic EEG Signals Based on Symbol-Relative Entropy, in 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Beijing, China, 2018, pp. 1-9, doi: 10.1109/CISP-BMEI.2018.8633179.
[21] A. P. Berg and W. B. Mikhael, An efficient structure and algorithm for the mixed transform representation of signals, in Conference Record of The Twenty-Ninth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 1995, pp. 1056-1060 vol. 2, doi: 10.1109/ACSSC.1995.540861.
[22] M. M. Bykov, V. V. Kovtun, I. D. Ivasyuk, A. Kotyra and A. Mussabekova, The automated speaker recognition system of critical use, in International Society for Optical Engineering, Vol. 10808, 2018, 108082V, doi: 10.1117/12.2501688.
[23] M. M. Bykov, V. V. Kovtun, A. Raimy, K. Gromaszek and S. Smailova, Neural network modelling by rank configurations, in The automated speaker recognition system of critical use, in International Society for Optical Engineering, Vol. 10808, 2018, 1080821, doi: 10.1117/12.2501521.
[24] O. V. Bisikalo, V. V. Kovtun, M. S. Yukhimchuk and I. F. Voytyuk, Analysis of the automated speaker recognition system of critical use operation results in Radio Electronics, Computer Science, Control, Zaporizhzhia, Ukraine, No, 4, 2018, pp . 71-84, doi: 10.15588/1607-3274-2018-4-7.
[25] O. V. Bisikalo, V. V. Kovtun and M. S. Yukhimchuk, Modeling the security policy of the information system for critical use, in Radio Electronics, Computer Science, Control, Zaporizhzhia, Ukraine, No. 1, 2019, pp. 132-149, doi: 10.15588/1607-3274-2019-1-13.
[26] V. V. Kovtun, M. S. Yukhimchuk, P. Kisała, A. Abisheva and S. Rakhmetullina, Integration of hidden markov models in the automated speaker recognition system for critical use, in Przeglad Elektrotechniczny, Wydawnictwo SIGMA, Poland, 2019, No. 1, 2019, pp. 178-182, doi: 10.15199/48.2019.04.32.