entropy

MDPI

Article

Entropy-Argumentative Concept of Computational Phonetic Analysis of Speech Taking into Account Dialect and Individuality of Phonation

Viacheslav Kovtun [1], Oksana Kovtun [2,*] and Andriy Semenov [3]

[1] Department of Computer Control Systems, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Khmelnitske Shose Str., 95, 21000 Vinnytsia, Ukraine; kovtun_v_v@vntu.edu.ua
[2] Department of the Theory and Practice of Translation, Faculty of Foreign Languages, Vasyl' Stus Donetsk National University, 600-Richchya Str., 21, 21000 Vinnytsia, Ukraine
[3] Department of Information Radioelectronic Technologies and Systems, Faculty of Information Electronic Systems, Vinnytsia National Technical University, Khmelnitske Shose Str., 95, 21000 Vinnytsia, Ukraine; semenov.a.o@vntu.edu.ua
* Correspondence: o.kovtun.work@gmail.com

Abstract: In this article, the concept (i.e., the mathematical model and methods) of computational phonetic analysis of speech with an analytical description of the phenomenon of phonetic fusion is proposed. In this concept, in contrast to the existing methods, the problem of multicriteria of the process of cognitive perception of speech by a person is strictly formally presented using the theoretical and analytical apparatus of information (entropy) theory, pattern recognition theory and acoustic theory of speech formation. The obtained concept allows for determining reliably the individual phonetic alphabet inherent in a person, taking into account their inherent dialect of speech and individual features of phonation, as well as detecting and correcting errors in the recognition of language units. The experiments prove the superiority of the proposed scientific result over such common Bayesian concepts of decision making using the Euclidean-type mismatch metric as a method of maximum likelihood and a method of an ideal observer. The analysis of the speech signal carried out in the metric based on the proposed concept allows, in particular, for establishing reliably the phonetic saturation of speech, which objectively characterizes the environment of speech signal propagation and its source.

Keywords: relative entropy; computational linguistics; computational phonetic analysis of speech; phonetic fusion; recognition of language units; individual phonetic alphabet

Citation: Kovtun, V.; Kovtun, O.; Semenov, A. Entropy-Argumentative Concept of Computational Phonetic Analysis of Speech Taking into Account Dialect and Individuality of Phonation. *Entropy* **2022**, *24*, 1006. https://doi.org/10.3390/e24071006

Academic Editor: Stanisław Drożdż

Received: 12 June 2022
Accepted: 15 July 2022
Published: 20 July 2022

1. Introduction

Computational phonetic analysis is a fundamental component of most information technologies for natural language recognition, cognitive speech analysis, automated speech transcription, and so on. The high reliability of phonetic analysis is a guarantee of a qualitative result of the functioning of all of these types of systems. The primary phonetic-morphological analysis of inflected languages and speech is especially relevant. The main source of errors in this process is a fusion [1–4]. This phenomenon characterizes the high variability of the individual sounding of phonemes, especially at the junction of morphemes. The phenomenon of fusion is objectively determined by the phonological evolution of natural language and cannot be ignored in the creation of precision technologies for computational phonetic analysis of speech.

The task of computational phonetic-morphological analysis of language or speech is objectively complicated, firstly, by the peculiarities of language itself as a process of physiologic-cognitive human activity, and, secondly, by the peculiarities of the profile information technologies involved.

We note the main integral factors of the first source of complications [5–9].

Homonymy of inflections. Inflexions can be homonymous if they belong to a single world-changing paradigm or characterize a single lexical and grammatical category, but belong to different world-changing paradigms, and they are sometimes found in the paradigms of different parts of language. This factor is a source of ambiguity in phonetic-morphological analysis. The negative impact of this factor can be reduced by using information technologies of linguistic context analysis and computational phonetic analysis.

Internal flexion. This type of inflexion is manifested when using the basic collection of language units, the representativeness of which depends on the content of word forms. If the collection is not used, it is necessary to formulate the rules of linguistic polymorphism inherent in the studied language.

Complex lexemes. Lexemes, the phonation (inscription) of which includes specific articulation techniques (special symbols), require the definition of declension for each component in the word form.

Analytical word forms. Analytical word forms are found in many languages and can cause significant complications in phonetic-morphological analysis because the components of the word form can be separated and even be located in different positions in the sentence.

Large lexical fund of language. Despite the rapid positive dynamics of computing power characteristics and the large memory capacity of modern computer technology, working with a basic collection of language units of the studied language (especially with a basic collection of word forms) in the implementation of phonetic-morphological analysis remains a task of high computing.

Variability of the lexical level of language. Updating the collections of language units for the appropriate type of phonetic-morphological analysis system does not keep up with the polymorphism of natural language (especially if we take into account dialects), which is manifested in the everyday phenomenon of new lexemes (specifically, terms) and word forms. Systems of computational phonetic-morphological analysis of a no collection type suffer less from the influence of this complicating factor.

We have mentioned only the most common factors of natural linguistic origin which negatively affect the effectiveness of computational phonetic-morphological analysis of speech. Depending on the information technology involved, this list is expanding.

We investigate the current state of the theoretical and analytical basis of current information technologies of computational phonetic-morphological analysis. Based on the results of information retrieval [10,11], we distinguish two relevant approaches—rationalistic and empirical. The first approach uses linguistic knowledge to analyze and synthesize language units. The second approach is based on the generalization of empirical data, for example, in the form of a statistical model of language (speech) [12,13]. However, in modern computational linguistics, technologies that integrate both of these approaches in a certain proportion are the most productive. According to the content of the used collection of language units, as a system-forming element for the implementation of computational phonetic-morphological analysis, technology-analyzers can be divided into [14–17]: (1) systems with a collection of phonemes and morphemes; (2) systems with a collection of lexemes and word forms; and (3) systems without basic collections.

The central element of the systems of the first type is a collection of relatively phonetically and linguistically stable language units (morphemes, phonemes, selected allophones) of the studied language. The corresponding technology analyzer decomposes the speech signal (text) into a certain sequence of indivisible portions, carrying out a recognition procedure for each of them. Such an elementary combinatorial model is most often used for the analysis of inflectional and agglutinative languages. The order of parts of lexemes is defined as the concatenation of the corresponding classes of morphemes in the collection. To determine the order of transition between classes of morphemes, the mathematical apparatus of finite state machines or Markov chains are usually used [18]. The number of classes of morphemes in the collection is determined by the result of the previous morphological classification of the studied language. In addition to declarative information on the composi-

tion of morphemes, the collection may also store procedural information. Such information determines the allowable range of variation of the patterns of morphemes and is most often designed as a system of production rules [10,11,14,19]. Empirical probabilistic-statistical methods are most often used for recognition in systems of this type [4,12,20].

Systems of the second type are focused on computational morphological analysis. Accordingly, the content of collections of etalons of language units in such systems is formed by morphemes and short lexemes. Systems of this type consider word forms as a sequence of such language units formed according to compositional and(or) production rules. When studying word forms, the system generates a lemma for it according to certain rules [21,22]. If such a lemma is present in the basic collection, then the word form is considered recognized. If we take into account the resource intensity, the effectiveness of such systems is determined mainly by the representativeness of the content of the basic collection. Collections of morphemes or lexemes are used in phonetic-morphological analysis to normalize the studied word forms. In the presence of a collection of morphemes, normalization is realized in the form of stemming. In the presence of a collection of lexemes, normalization is realized in the form of lemmas. We separately mention the subclass of systems of the second type, which uses a collection of word forms. The purpose of such systems is grammatical and morphological analysis, in which the collection presents a set of combinations of word forms, which is matched by a set of grammatical labels [11,19,23]. With a sufficiently rich collection, the source of analysis errors in these systems is only the homonymy of the complete word form.

The disadvantage of all systems of phonetic-morphological analysis of the first and second types is the use of large collections of language units. However, according to this criterion, systems focused on the use of phonetic and morphological collections look better if the efficiency of the recognition process is acceptable.

Systems of the third type perform phonetic-morphological analysis exclusively based on mathematical methods of machine learning (support vector machines, EM-method, genetic algorithms, Kohonen networks, etc.) [24–34]. Any methods capable of graphemic analysis [24], the result of which is the automatic or automated formation of phonetic-morphological collections, are acceptable. The advantage of the third type of system is the methodologically determined high heuristics and adaptability, which potentially allow for recognizing language units in speech material with a clear uncertainty. The disadvantage of such systems is the complexity and instability of learning these pseudo-intelligent methods, as well as the need for initial data and computing resources, the volume of which exceeds that required for systems of the first and second type, not in times, but orders.

Below we formulate the main provisions of our study.

The ***object*** of study is the fusion of the process of merged speech.

Considering the mentioned advantages and disadvantages of systems of phonetic-linguistic analysis, we formulate the ***purpose*** of the study as formalization in the paradigm of information theory of a statistically adequate analytically rigorous concept of phonetic analysis of speech, the variability of which will be taken into account.

The ***subject*** of research will be methods of probability theory and mathematical statistics, information theory, pattern recognition theory and acoustic theory of language formation.

In this context, the ***objectives*** of the study are: to create a concept of the process of computational phonetic analysis of speech, taking into account dialects and the specifics of phonation introduced by the speaker; to formulate a criterion for the estimation of the phonetic saturation of speech based on the proposed model, taking into account the distorting effect of the channel of propagation of speech signals in the phonation process; and to prove the adequacy and functionality of the obtained theoretical results.

The ***main contribution*** of the research is the concept of computational phonetic analysis of speech. In the concept, in contrast to the existing methods, the task of addressing the multicriteria of the process of cognitive perception of speech by a person is strictly formally presented in the theoretical and analytical apparatus of information theory, pattern recognition theory and acoustic theory of speech formation. The obtained concept allows

for determining accurately the phonetic alphabet of a person, taking into account their inherent dialect of speech and individual features of phonation, as well as detecting and correcting errors in the recognition of language units and reliably assessing the phonetic saturation of speech.

The ***highlights*** of this research are:

- The entropy-argumentative concept (i.e., the mathematical model and methods) of computational phonetic analysis of speech, taking into account dialect and individuality of phonation;
- The entropy-argumentative concept of detection and correction of errors of computational phonetic analysis of speech.

2. Materials and Methods

2.1. Statement of Research

The functional purpose of typical modern information technology of computational analysis of speech patterns is realized by comparing the parameterized representation of the studied language unit and its corresponding etalon in a certain parametric space. The main source of uncertainty in the comparison process is the biological origin of the speech signal and its distortion during transmission and processing. However, the acoustic variability of phonation of language units (primarily, phonemes), due to the existence of dialects, is relatively stable. Based on this fact, we assume a simultaneous comparison of the studied pattern of the phonogram with the pronounced phoneme x with each element $x_{r,j}$ of the set of etalons $X_r = \{x_{r,j}\}$, where $j = \overline{1, J_r}$ is the index of the etalon that characterizes the corresponding dialect of the phoneme $r = \overline{1, R}$, where R is the capacity of the phonetic alphabet and J_r is the capacity of the set of recognized dialects for the phoneme r. Then, if the distance $\rho(x/x_{r,j})$, $r = \overline{1, J_r}$, between the studied pattern x and at least one of the elements $x_{r,j}$ of the cluster of the r-th phoneme does not exceed the specified threshold value

$$\frac{1}{J_r}\sum_{j=1}^{J_r}\rho\left(\frac{x}{x_{r,j}}\right) \le \rho_0, \tag{1}$$

then we can recognize the pattern x as the phoneme $r \in X_r$. Such a process of recognizing language units will be objective (in particular, insensitive to the dialects of phonation of language units), as the clusters $\{x_{r,j}\}$ for the phonetic alphabet X_r are representatively defined. Depending on the value of the threshold ρ_0, the result of the analysis of the studied pattern x according to Rule (1) will be: its recognition as one of the phonemes: $x = r$; its identification with several phonemes: $x = \{r_i\}$, $r_i \in X_r$, $i \le J_r$; or its recognition as marginal regarding the studied phonetic alphabet: $x \ne \forall r \in X_r$. To simplify the calculations, we convert Rule (1) into the form

$$\rho_r(x) = x_r^* = x_{r,v} : \frac{1}{J_r}\sum_{j=1}^{J_r}\rho\left(\frac{x_{r,j}}{x_{r,v}}\right) = \min_{i \le J_r}\frac{1}{J_r}\sum_{j=1}^{J_r}\rho\left(\frac{x_{r,j}}{x_{r,i}}\right) \triangleq \rho_r^* \le \rho_0, \tag{2}$$

where in the process of recognizing the pattern x within the cluster X_r one distance $\rho_r(x) \triangleq \rho(x/x_r^*)$ from it to the center of the cluster x_r^* is calculated, the coordinates of which determine the dialect-averaged phoneme etalon $r \in X_r$.

Based on Rule (2), we define the procedure of computational phonetic analysis of speech as a comparison of empirical (spoken by the person) $\{x_v^*\}$ and etalon $\{x_r^*\}$ sets of equal capacity, the pairwise elements of which generalize the corresponding phonemes of the studied language both on the speaker's side $v \in V$ and on the side of the etalon phonetic collection $r \in R$.

2.2. Entropy-Argumentative Concept of Computational Phonetic Analysis of Speech Taking into Account Dialect and Individuality of Phonation

Based on the provisions of information theory, we argue the solution rule (2) in the context of the relative entropy functional [35–37] (3):

$$\rho(x) \triangleq \int \ldots \int \ln \frac{dP(x)}{dP_r(x)} P(dx), \tag{3}$$

where $P(x)$ is the selective probability distribution of the studied (empirical) speech signal x relative to the etalon probability distribution $P_r(x)$, $r = \overline{1,R}$. Assume that the distribution law $P(x)$ is normal: $P(x) = N(K_X)$, where K_X is a sample matrix of autocorrelation of the speech signal x of dimension $n \times n$. Consider this in Expression (3): $\rho_r(x) = \frac{1}{2}\left(\mathrm{tr}\left(\frac{K_x}{K_r}\right) - \ln\left(\frac{K_x}{K_r}\right) - n\right)$, where $\mathrm{tr}(A)$ is the operation of finding a trace of the matrix A. If we assume that the studied speech signal is normalized to its entropy, then the last expression can be further simplified to the form

$$\rho_r(x) = \frac{1}{2}\left(\mathrm{tr}\left(\frac{K_x}{K_r}\right) - n\right).$$

We present Function (3) in frequency space as the optimal solving statistics [35]. For one sample of the studied speech signal, we obtain (4):

$$\rho_r(x) = \frac{1}{F}\left|\frac{1 - \sum\limits_{m=1}^{p} a_r(m) e^{-j\pi m \frac{f}{F}}}{1 - \sum\limits_{m=1}^{p} a_x(m) e^{-j\pi m \frac{f}{F}}}\right|^2, \tag{4}$$

where f is the discrete frequency value for the analyzed sample of the speech signal, F is the upper limit value of the speech signal frequency equal to half of its sampling frequency, and $\{a_r(m)\}$ and $\{a_x(m)\}$ are the vectors of linear autoregression coefficients of order p for etalon signal x_r^* and empirical signal x, respectively. The expression in the numerator of (4) is an amplitude-frequency characteristic of the bleaching filter tuned to highlight the features of the r-th phoneme x_r^*, $r = \overline{1,R}$.

Expressions (2) and (4) allow us to calculate quantitative characteristics, based on which it is possible to reasonably decide whether the studied pattern x belongs to the cluster x_r^* of the corresponding phoneme $r \in X_r$. It is possible to vary the errors of this recognition process by changing the value of the threshold ρ_0. Given the Gaussian approximation of the speech signal, the probability of error of the first kind α for the process of phoneme recognition taking into account the dialects of the studied language is proposed to be defined in terms of χ^2-criterion with M degrees of freedom:

$$\alpha \triangleq P\left\{\rho_r(x) \geq \rho_0|_{x \in X_r}\right\} = P\left\{\chi^2_M > M(1+\rho_0)\right\}, \tag{5}$$

where $P\{.\}$ is the probability of a random event, $M = const$.

In the general case, the value of the constant M is calculated by the expression $M \approx L - p$, where p is the order of the bleaching filter, and $L = 2F\tau$ is a parameter whose value depends on the number of stationary intervals τ allocated in the studied speech signal x. The value of error α determined by Expression (5) is inversely proportional to the value of the threshold ρ_0. For example, for a given value of $\alpha = 0.1$ at $\tau = 5$ ms, $F = 8$ kHz, $p = 20$, we obtain $L = 80$ and, accordingly, $M = 60$. Using the χ^2-distribution tables for the significance level $\beta = 1 - \alpha = 1 - 0.99 = 0.01$, we find the value of the quantile $\chi^2_{M;\beta} = \chi^2_{60;0.01} = 88.38$, using which we calculate the value of the threshold ρ_0: $\rho_0 = \chi^2_{M;\beta}/M - 1 = 0.473$.

The error of the second kind β in the context of the task of computational phonetic analysis of speech when taking into account dialects represents the probability of the confusion of phonemes r and v, $r, v \in X_r$, the centers of clusters x_r^* and x_v^* of which are close enough in the parametric space $\rho_{rv} \triangleq \rho_r(x)|_{x=x_v^*}$. Therefore, the value of error β is inversely proportional to the value of distance ρ_{rv}. Analysis of the results of a statistically representative number of experiments showed that the minimum value of ρ_{rv} the phonetic alphabets of the English language $\{x_r^*\}$ is in the range $[0.2; 0.3]$. Accordingly, in analogy with (5), we formalize the expression for calculating the error of the second kind β of the phoneme recognition process taking into account the dialects of the studied language:

$$\beta \triangleq P\Big\{ \rho_r(x) \geq \rho_0|_{x \in X_v} \Big\} = P\Big\{ \chi_M^2 < \frac{M(1+\rho_0)}{1+\rho_{rv}} \Big\}. \tag{6}$$

Summarizing the considerations embodied in Expressions (5) and (6), for practical use we choose the value of the threshold ρ_0 in the decision rule (2) based on the expression

$$p_0 = (1, \ldots, 2) \min_{r,v} \rho_{rv}. \tag{7}$$

The value of the threshold ρ_0, calculated by Expression (7), provides a balance between the values of errors of the first and second kind of the process of phoneme recognition from the phonetic alphabet X_r, taking into account the dialects of the studied language and the variability of the phonation process. However, the question of the influence of individual features of speakers' articulation on the result of phonetic analysis of speech requires more detailed analytical formalization.

In the context of the provisions of information theory, we consider the speaker as a source of discrete messages X, defined on the set of etalons of language units $\{x_r^*\}$. Such a source can be comprehensively characterized by the amount of information per language unit generated by it.

If we ignore the influence of individual features of the speaker's articulatory apparatus on the phonation process and assume that the speech message is transmitted in the absence of acoustic ambient noise, the required amount of information is defined as Shannon entropy for a discrete message source [35]:

$$H(X) \triangleq -\sum_{r=1}^{R} P(X = x_r^*) \log P(X = x_r^*) = -\sum_{r=1}^{R} p_r \log p_r. \tag{8}$$

If we mention the normalization $\sum\limits_{r=1}^{R} p_r = 1$, then, considering the equally probable appearance of language units $\forall r \leq R$: $p_r = 1/R$, we obtain a simplified form of Expression (8): $H(X) = \log R$. However, in real conditions, it is impossible to ignore articulatory conditioned variability of phonation. The speech signal at the output of the articulatory tract of the speaker X' may differ significantly from the etalon X: $X' \neq X$.

This axiom is true even for individual phonemes, not to mention more massive language units. Under such conditions, an adequate mathematical model of a discrete source of speech messages should be created based on phonemes defined by Expression (5), clearly clustered in the parametric space: $q_r \triangleq P(X' \neq x_r^*)$, $r = \overline{1, R}$, and taking into account the probability of an abstract, R+1-th, language unit, which includes cases of the unreliable recognition of a signal X': $q_{R+1} \triangleq P(X' \neq x_r^*, \forall r \leq R)$. We summarize these considerations for the decision rule (2):

$$\begin{gathered} q_r = \sum_{v=1}^{R} q_{rv} = \sum_{v=1}^{R} P(X' = x_r^*; X = x_v^*) = \sum_{v=1}^{R} P(X = x_v^*) P(X' = x_r^* | X = x_v^*) \\ = P(X = x_r^*) P(X' = x_r^* | X = x_r^*) = (1-\alpha) p_r, \\ q_{R+1} = \sum_{v=1}^{R} P(X' = x_v^*; X = x_v^*) = \sum_{v=1}^{R} P(X = x_v^*) P(X' \neq x_v^* | X = x_v^*) = \sum_{v=1}^{R} \alpha p_v = \alpha, \\ \sum_{r=1}^{R+1} q_r = (1-\alpha) \sum_{r=1}^{R} p_r + \alpha \equiv 1, \end{gathered} \tag{9}$$

where $P(X' = x_r^* | X = x_r^*) = 1 - \alpha$ is the conditional probability of recognizing the r-th phoneme, provided that the variability of its phonation introduced by the speaker is ignored.

Note that Expression (8) characterizes a discrete source of speech messages without taking into account the disturbing effect of the channel of their distribution on the final result of phonation. Consider this information using as a basic expression [35]:

$$I(X, X') \triangleq H(X) - H(X|X'), \tag{10}$$

where X is a specimen of the phonation of the etalon x_r^* of the phoneme $r \in X_r$, X' is a specimen of the phonation of this phoneme by the speaker (empirical specimen), and $H(X|X')$ is a posteriori entropy, which characterizes the scattering of useful information of a phonation process due to disturbing effects in its distribution channel. Taking into account Expression (9), we formulate the equivalent representation of Expression (10):

$$\begin{gathered} I(X, X') = H(X) + H(X') - H(XX') = H(X) - \sum_{r=1}^{R+1} q_r \log q_r + \sum_{v=1}^{R} \sum_{r=1}^{R+1} q_{rv} \log q_{rv} = H(X) \\ -(1-\alpha) \sum_{r=1}^{R} p_r \log(p_r(1-\alpha)) - \alpha \log \alpha + \sum_{r=1}^{R} q_{rr} \log q_{rr} + \alpha \sum_{v=1}^{R} p_v \log(p_v \alpha) = H(X) \\ +(1-\alpha) H(X) - (1-\alpha) \left((1-\alpha) - \alpha \log \alpha + (1-\alpha) \sum_{r=1}^{R} p_r \log(p_r(1-\alpha)) \right) - \alpha H(X) \\ +\alpha \log \alpha = (1-\alpha) H(X). \end{gathered} \tag{11}$$

Based on Expression (11), we can say that the a posteriori entropy of information scattering in the phonation of the speech message $H(X|X')$ is in direct proportion to the entropy of the discrete speech message source (8):

$$H(X|X') = \alpha H(X). \tag{12}$$

Based on Expression (12), we can say that with an equally probable distribution of phonemes in the phonetic alphabet of the speaker, the upper limit of scattering of useful information in the phonation process can be described by the expression

$$\sup H(X|X') = \alpha \log R. \tag{13}$$

The obtained result correlates with the known Fano inequality [38] for arbitrary solution rules:

$$H(X|X') \leq -\alpha \log \alpha - \beta \log \beta + \alpha \log(R-1). \tag{14}$$

The last statement can be proved empirically by comparing the calculated values of the right-hand sides of Expressions (13) and (14) for the experimental data for $0 \leq \alpha \leq 1$ and $1 < R < \infty$.

Thus, the decision rule (2), the decision statistic (4) and Expressions (7)–(9) together form the desired concept of the process of computational phonetic analysis of speech, taking into account dialects and the specifics of phonation introduced by the speaker. The central element of the concept is the matrix of information mismatch $\|\rho_{r,v}\|$ of dimensions $R \times R$. The data from the matrix $\|\rho_{r,v}\|$ are the basis for calculating the threshold ρ_0 using

Expression (7). With a known value of ρ_0 based on Expressions (2) and (5), the procedure of segmentation of the phonetic alphabet $X_r = \{x_{r,j}\}$ into a set of phonemes, which with probability $\beta = 1 - \alpha$ are reliably recognized despite the above-described disturbing factors, and another set of phonemes, which with probability α are not reliably recognized. A significant factor for such segmentation is the probability of error of the first kind, which is calculated by Expression (5). The probability of error of the second kind (6) in this procedure is taken into account indirectly as a limitation in determining the threshold ρ_0 by Expression (7). The use of Expressions (9) and (10) allows for clarifying the result of the segmentation procedure, taking into account the variability of the phonation of the studied language units caused by the individual features of the articulation of a particular speaker. Note that although the presented concept was formulated based on phonemes, the provisions underlying it are consistent and for the analysis of speech about the content of such language units as morphemes and lexemes. Based on the proposed concept (8)–(10), Rule (11) allows us to estimate the error of the first kind (5) and the personalized entropy of the phonetic dictionary (8) as a result of the analysis of empirical data, the sample size of which is $N = 2FT$. The statistically representative volume $N = 10^6$ in the study of the phonetic alphabet of $R = 10^2$ elements by Rule (11) as a result of analysis of phonograms of speech signals with a sampling frequency of 16 kHz is achieved with a censored duration.

2.3. Entropy-Argumentative Concept of Detection and Correction of Errors of Computational Phonetic Analysis of Speech

Let $X_r = \{x_{r,j}\}$, $r = \overline{1,R}$, $j = \overline{1,M}$ be a set of independent classified samples of type $x_{r,j} = \left[x_{r,j^{(1)}}, x_{r,j^{(2)}}, \ldots, x_{r,j^{(n)}}\right]^T$ with a capacity n of $R \geq 2$ Gaussian distributions $P_r = N(K_r)$ with zero mathematical expectation and unknown autocorrelation matrix $K_r = E_X\left(x_{r,j}x_{r,j}^T\right)$ of dimension $n \times n$, where j is the identifier of the cycle of observations of the r-th distribution, T is the transposition operation, E_X is the mathematical expectation of the sample of sets X. Denote by X_0 a sample of the form X_r with capacity M_0 for the studied signal with an unknown distribution $P(X) \subset \{P_r\}$. The task of recognizing the signal X_0 involves R-alternative testing of statistical hypotheses W_r regarding the distribution law of this signal:

$$W_r : P(X) = P_r,\ r = \overline{1,R}. \tag{15}$$

Let $R = 2$, i.e., two competing hypotheses, $W_1 : P(X) = P_1$ and $W_2 : P(X) = P_2$, are tested for a priori unknown autocorrelation matrices K_1 and K_2. The verification will be performed using the asymptotic minimax criterion of the likelihood ratio [35–37] based on data from a sample $X\{X_i\}$, $i = \overline{0,2}$. Under such conditions, the hypothesis W_1 will be considered true if the condition

$$W_1 : \lambda_1(X) \triangleq \frac{\underset{K_1}{\sup}\underset{K_2}{\sup}(p(X|W_1))}{\underset{K_1}{\sup}\underset{K_2}{\sup}(p(X|W_2))} \equiv \frac{\underset{K_1}{\sup}(p(X_0|W_1)p(X_1))\underset{K_2}{\sup}(p(X_2))}{\underset{K_1}{\sup}(p(X_0|W_2)p(X_1))\underset{K_2}{\sup}(p(X_1))} > 1, \tag{16}$$

is satisfied, where $p(X_0|W_r)$ is the plausibility function of the signal X_0 provided that hypothesis W_r is confirmed, and $p(X_r)$ is the plausibility function of the signal X_r.

Using the known computational algorithm [38] under the condition of independence of observations $X_r = \{x_{r,j}\}$, we write a system of equations of the form

$$\begin{cases} \ln(p(X_0|W_r)) = -\frac{M_0}{2}\left(\ln|K_r| + \mathrm{tr}\left(\frac{S_0}{K_r}\right) + n\ln(2\pi)\right), \\ \ln(p(X_r)) = -\frac{M_r}{2}\left(\ln|K_r| + \mathrm{tr}\left(\frac{S_r}{K_r}\right) + n\ln(2\pi)\right), \end{cases} \tag{17}$$

where $|K_r|$ is the determinant of the matrix K_r, and $S_r \triangleq \frac{1}{M_r}\sum\limits_{j=1}^{M_r} x_{r,j}x_{r,j}^T$ is the estimate of the maximum likelihood for the matrix K_r determined on the sample X_r, $r = \overline{0,2}$. We describe based on Rxpression (17) the fact that the upper limits $\ln(p(X_r))$ are reached at $K_r = S_r$:

$$\sup_{K_r}(p(X_r)) = -\frac{M}{2}(\ln|S_r| + nc), \tag{18}$$

where $r = \{1,2\}$, $c = \ln(2\pi) + 1$.

Similarly, we obtain the expression for determining the upper limits for $\ln(p(X_0|W_r)p(X_r))$:

$$\begin{gathered}\sup_{K_r}(\ln p(X_0|W_r)p(X_r)) \\ = -\frac{1}{2}\left((M_0+M)(\ln|S_{0r} + n\ln(2\pi)|) + M_0\mathrm{tr}\left(\frac{S_0}{S_{0r}}\right) + M\mathrm{tr}\left(\frac{S_0}{S_{0r}}\right)\right) \\ = -\frac{M_0+M}{2}(\ln|S_{0r}| + nc),\end{gathered} \tag{19}$$

where $r = \{1,2\}$, and $S_{0r} = \frac{M_0}{M_0+M}(S_0 + S_r)$ is the estimate of the maximum likelihood for the matrix K_r determined on the combined sample $X_{0r} + \{X_0, X_r\}$ with capacity $M_0 + M$.

Substitute Expressions (18) and (19) into Expression (16) and obtain the condition under which the hypothesis W_1 will be considered correct:

$$\begin{gathered}W_1(X) : \lambda_1(X) = \frac{1}{2}((M_0+M)\ln|S_{01}| - (M_0-M)\ln|S_{02}| - M\ln|S_1| + M\ln|S_2|) < 0 \\ \equiv M_0\gamma_{1,01} + M\gamma_{1,01} < M_0\gamma_{2,02} + M\gamma_{2,02},\end{gathered} \tag{20}$$

where $\gamma_{k,0r} = \frac{1}{2}\left(\mathrm{tr}\left(\frac{S_k}{S_{0r}}\right) - \ln|S_k| + \ln|S_{0r}| - n\right) \geq 0$ is the value of the relative entropy functional between two hypothetical probability distributions with autocorrelation matrices S_k and S_{0r}.

We scale rule (20) for the task of recognizing signals of the form in (15) with an arbitrary number of hypotheses $R \geq 2$:

$$W_v(X) : (M_0\gamma_{0,0r} + M\gamma_{r,0r})|_{r=v} = \min,\ r = \overline{1,R}. \tag{21}$$

Assuming the homogeneity of the pair of signals X_0 and X_r in the sample X_{0r} and considering that $\gamma_{0,0r} \leq \gamma_{0,r}$, $\gamma_{r,0r} \leq \gamma_{r,0}$ and $M = M_0$, we present Rule (21) in the form

$$W_v(X) : \lambda_v(X) \triangleq (M_0\gamma_{0,r} + M\gamma_{r,0})|_{r=v} \triangleq \gamma_{0,r} + \gamma_{r,0}|r = v = \min,\ r = \overline{1,R}. \tag{22}$$

where the solving statistics of the relative entropy functional

$$\gamma_{0,r} = \frac{1}{2}\left(\mathrm{tr}\left(\frac{S_0}{S_r}\right) - \ln|S_0| + \ln|S_r| - n\right), \tag{23}$$

$$\gamma_{r,0} = \frac{1}{2}\left(\mathrm{tr}\left(\frac{S_r}{S_0}\right) - \ln|S_r| + \ln|S_0| - n\right) \tag{24}$$

are determined on the R-set of pairs of sample distributions $N(S_0)$, $N(S_r)$, $r = \overline{1,R}$.

An alternative to Expressions (23) and (24) may be to take into account the principle of the minimum value of information non-directional mismatch $J(X_0, X_r) \triangleq \frac{1}{2}(\gamma_{0,r} + \gamma_{r,0})$ between stochastic signals X_0 and X_r, $r = \overline{1,R}$, in the rule (22):

$$\widetilde{W}_v(X) : \widetilde{\lambda}_v(X) \triangleq \gamma_{0,r}|_{r=v} = \min,\ r = \overline{1,R}, \tag{25}$$

where the decision statistics $\gamma_{0,r}$ are determined by Expression (23).

Expression (25) is a particular case of Criterion (22), provided that with an unlimited increase in the volume of training samples M, the second term in Expression (21) asymptotically reduces to zero: $\gamma_{r,0r} \to \gamma_{r,r} = 0 \forall r \leq R$. Thus, the transition from Rule (22) to (25) is

appropriate provided that there is a significant asymmetry in the values of the decision statistics (23), (24).

The probability $\alpha_{v\to r} \triangleq P(W_r(X)|W_r)$ of confusion of the v-th and r-th signals, $v \neq r \leq R$, from the user database of a priori data $\{X_r\}$ in the formalism of Rule (22) can be described by the expression

$$\alpha_{v\to r} = P\{\gamma_{0,v} + \gamma_{v,0} > \gamma_{0,r} + \gamma_{r,0}|W_v\} = P\{2\gamma_{v,v} > \gamma_{v,r} + \gamma_{r,v}\}. \quad (26)$$

If we take into account that the empirical signal before recognition is normalized to the value of its specific entropy, then the system of asymptotic equations $\forall r \leq R : \frac{1}{n}\ln|S_r| = \frac{1}{n}\ln|S_0| \underset{n\to\infty}{=} \ln\sigma_0^2 = const$ is satisfied. We take this fact into account by presenting the solving statistics $\gamma_{v,r}$ in the χ^2-distribution formalism with $K \leq M$ degrees of freedom: $\gamma_{v,r} = \frac{1}{2}n\left(\frac{\sigma_{r,v}^2\sigma_0^2\chi_{r,v}^2(K)}{M} - 1\right)$, where $\sigma_{r,v}^2 \triangleq \frac{\sigma_0^2}{n}\lim\limits_{n\to\infty}\left(M\mathrm{tr}\left(\frac{S_v}{S_r}\right)\right)$ is an auxiliary variable. Substitute the obtained expression for statistics $\gamma_{v,r}$ into Expression (26):

$$\alpha_{v\to r} = P\left\{\sigma_0^2\chi_{v,v}^2 > \frac{1}{2}\sigma_{r,v}^2\chi_{r,v}^2 + \frac{1}{2}\sigma_{v,r}^2\chi_{v,r}^2\right\} = P\left\{2\chi_{v,v}^2 > (1+\rho_{r,v})\chi_{r,v}^2 + (1+\rho_{v,r})\chi_{v,r}^2\right\}, \quad (27)$$

where $\rho_{r,v} \triangleq \frac{\sigma_{r,v}^2}{\sigma_0^2} - 1$ and $\rho_{v,r} \triangleq \frac{\sigma_{v,r}^2}{\sigma_0^2} - 1$ are the specific values of the information discrepancy for the studied pair of distributions $N(S_0)$ and $N(S_r)$ at $n \to \infty$, and $\sigma_{v,r}^2 \triangleq \frac{\sigma_0^2}{n}\lim\limits_{n\to\infty}\left(M\mathrm{tr}\left(\frac{S_r}{S_v}\right)\right)$ is an auxiliary variable of the same type as $\sigma_{r,v}^2$. If we assume the mutual noncorrelation of the three χ^2-distributions in Expression (27), then Expression (26) for calculating the probability of confusion $\alpha_{v\to r}$ can be represented as $\alpha_{v\to r} = P\left\{\frac{1}{2}((1+\rho_{r,v})F_{r,v}(1,K) + (1+\rho_{v,r})F_{v,r}(1,K)) < 1\right\}$, where $F_{r,v}(1,K) = \frac{\chi_{r,v}^2}{\chi_{v,v}^2}$ and $F_{v,r}(1,K) = \frac{\chi_{v,r}^2}{\chi_{v,v}^2}$ are statistics of the F-distribution with $(1,K)$ degrees of freedom. Accordingly, the upper limit of the probability of confusion $\alpha_{v\to r}$ can be estimated by the expression

$$\begin{aligned}\alpha_{v\to r} &\leq P\left\{\tfrac{1}{2}\max[(1+\rho_{v.r})F_{v,r}(1,K);(1+\rho_{r.v})F_{r,v}(1,K)]\right\}\\ &= P\left\{F(1,K) < \tfrac{2}{\max[(1+\rho_{v.r});(1+\rho_{r.v})]}\right\} =\\ P\left\{F(K,1) \geq \tfrac{1}{2}\max[(1+\rho_{v.r});(1+\rho_{r.v})]\right\} &= 1 - \Phi_{K,1}\{\max[(1+\rho_{v.r});(1+\rho_{r.v})]\},\end{aligned} \quad (28)$$

where $F(1,K) = \max[F_{r,v}(1,K);F_{v,r}(1,K)]$ and $F(K,1) = \frac{1}{F(1,K)}$ are statistics of the F-distribution with $(1,K)$ and $(K,1)$ degrees of freedom, respectively; $\Phi_{K,1}$ is the integral function of the F-distribution with $(K,1)$ degrees of freedom.

From Expression (28), it follows that there are essentially unequal distributions of statistics $\chi_{v,v}^2$ and a pair of statistics $\chi_{r,v}^2$, $\chi_{v,r}^2$ provided that $r \neq v$. Thus, Expression (28) theoretically proves the correctness of Expressions (23) and (24) concerning the asymmetry of the value of information discrepancy, which is taken into account in the decision rule (22). This means that when the condition $\exists v, r \leq R : \rho_{v,r} >> \rho_{r,v}$ is satisfied, it is more appropriate to apply the decision rule (22) rather than (25) to make decisions about the recognition of language units in the speech signal parameterized in the paradigm of concept (8)–(10). This thesis will be tested in the experimental part of this article.

Assume that when recognizing the signal under study using the decision rule (25), the verdict was erroneously in favor of the hypothesis $W_\mu(X)$, not the hypothesis $W_v(X)$. Suppose also that when recognizing the same signal using decision rule (22), the verdict was made in favor of the hypothesis $W_v(X)$. The stated assumptions assume that according to Expressions (25) and (26), inequalities $\gamma_{v,v} \geq \gamma_{v,\mu}$ and $2\gamma_{v,v} \geq \gamma_{v,\mu} + \gamma_{\mu,v}$ were fulfilled simultaneously, which is possible only if the condition $\gamma_{\mu,v} >> \gamma_{v,\mu}$ is satisfied. Thus, an

analytical indication of the erroneousness of the decision made under Rule (25) concerning the analyzed sample X_0 may be inequality of the form $\overline{W}_\mu(X) : \gamma_{\mu,0} >> \gamma_{0,\mu}$ or

$$\overline{W}_\mu(X) : \frac{1+\widetilde{\gamma}_{\mu,0}}{1+\widetilde{\gamma}_{0,\mu}} \geq c_0, \tag{29}$$

where $\widetilde{\gamma}_{0,\mu} = \frac{2\gamma_{0,\mu}}{n}$, $\widetilde{\gamma}_{0,\mu} = \frac{2\gamma_{0,\mu}}{n}$ are the specific values of the solving statistics (23), (24), respectively; c_0 is the threshold value (minimum value of the asymmetry coefficient of the values (23) and (24) in Rule (22)), set depending on the maximum permissible error

$$\beta \triangleq P\left\{\frac{1+\widetilde{\gamma}_{\mu,0}}{1+\widetilde{\gamma}_{0,\mu}} \geq c_0 \middle| W_\mu\right\} \leq \beta_0.$$

Repeating the considerations that accompanied the transition from Expressions (26) to (28), we rewrite the defined expression to determine the probability β in terms of the F-distribution:

$$\begin{gathered}\pi_{v\to\mu} \triangleq P\left\{\frac{1+\widetilde{\gamma}_{\mu,0}}{1+\widetilde{\gamma}_{0,\mu}} \geq c_0 | W_v\right\} = P\left\{\frac{1+\widetilde{\gamma}_{v,\mu}}{1+\widetilde{\gamma}_{\mu,v}} \geq \frac{1}{c_0}\right\} \\ = P\left\{\frac{1+\widetilde{\gamma}_{\mu,\mu}}{1+\widetilde{\gamma}_{0,0}} \geq c_0\right\} = P\left\{\frac{\chi^2_{\mu,\mu}(K)}{\chi^2_{0,0}(K)} \geq c_0\right\} = 1 - \Phi_{K,K}(c_0) \leq \beta_0.\end{gathered} \tag{30}$$

Analyzing Expression (30), we obtain an equation $\min c_0 = f_{K,K}(1-\beta_0)$, where $f_{K,K}(1-\beta_0)$ is the quantile of the F-distribution with (K,K) degrees of freedom and the level of significance $1-\beta_0$. For example, for $K = 100$ and $\beta_0 = 0.01$ from the tables for F-distribution, we have: $c_0 \geq f_{100,100}(0,99) = 1.59$.

Thus, Rule (29) allows us to estimate the probability of the event of marginal recognition of the correct result of the phoneme recognition procedure, employing the decision rule (25). The stochastic estimate of such an event is characterized by the expression

$$\pi_{v\to\mu} = P\left\{\frac{\chi^2_{v,\mu}(1)}{\chi^2_{\mu,v}(1)} \geq \frac{1+\rho_{v,\mu}}{c_0(1+\rho_{\mu,v})}\right\} = 1 - \Phi_{1,1}\left(\frac{1+\rho_{v,\mu}}{c_0(1+\rho_{\mu,v})}\right) \tag{31}$$

and is determined by the result of comparing the opposing elements $\rho_{r,v}$ and $\rho_{v,r}$ in the matrix $\|\rho_{r,v}\|$.

3. Results

We use Rule (11) based on the concept (8)–(10) to estimate the phonetic saturation of speech of persons in a team of 30 people. The personnel composition of this team was formed in a balanced way. It took into account such criteria as age (three age groups: 20–29, 30–39, 40–49 years), gender (male, female), higher education (university), native language (Ukrainian), and level of English language proficiency according to CEFR-B2. Each person listened to a phonogram of an 1800-character English-language journalistic text pronounced by a Google Translate service once. Subsequently, each person recounted the heard text for recording on a personalized digital phonogram lasting 3 min. The phonation of the retelling took place at the same tempo and timbre and with a clear fixation on language units. The phonograms were recorded using an AKG P420 microphone without an amplifier connected to a Creative Audigy Rx sound card integrated into a personal computer with a sampling frequency of 16 kHz. Each phonogram was saved in a .wav format file. For further analysis, the phonograms were split into segments of duration $\tau = 5$ ms ($L = 80$ samples). Based on the analysis of the corresponding phonograms of retellings, individual phonetic alphabets $\{X_r\}$ were formed for each person, for which the centers of clusters of phonemes $\{x^*_r\}$ were determined by Expression (2). Two variants of the individual phonetic alphabet were formed for each person with hard and soft conditions of formation. These conditions were caused by the level of a mismatch $\Delta\rho = \{0,5; 1,0\}$ for phonemes of the same name and their minimum duration $\Delta L = \{8L; 4L\}$, $\tau = \{40; 20\}$.

The values of the autoregression coefficients $\{a_r(m)\}$, $\{a_v(m)\}$ required for the calculation of the information mismatch matrix $\|\rho_{r,v}\|$ were determined using the Berg–Levinson recurrent procedure with an unambiguously determined order of models $p = 20$.

Figure 1 visualizes fragments of the resulting matrices for person №1, calculated with the selected hard (Figure 1a) and soft (Figure 1b) sets of formation conditions. The capacities of the phonetic alphabets were $R^1_{hard} = 32$ and $R^1_{soft} = 87$ language units, respectively. The minimum value of information discrepancy between phonemes was $\Delta\rho_{rv}^{R^1_{hard}} = 0.324$.

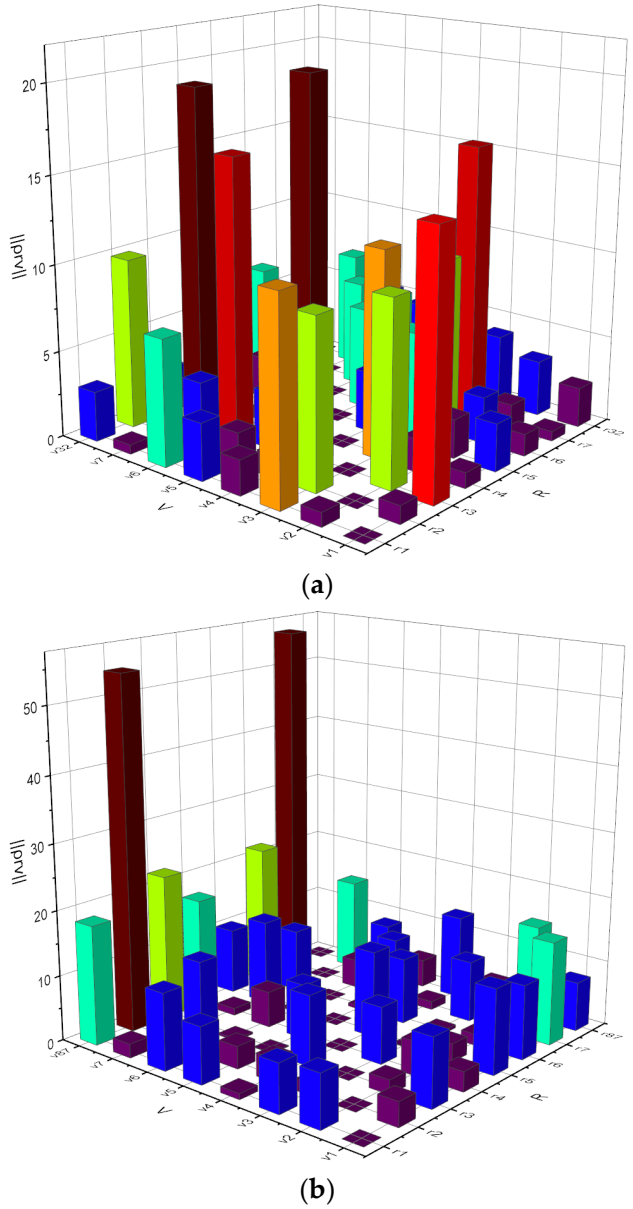

Figure 1. Visualization of fragments of information mismatch matrices $\|\rho_{r,v}\|$ for person №1, calculated with the selected hard (**a**) and soft (**b**) sets of formation conditions.

Respectively, according to the decision rule (2), taking into account Expression (7), the value of the threshold $\rho_0 = 0.324$ is determined. Using the tables of the χ^2-distribution for the number of degrees of freedom $M = 60$, the probability of error of the first kind $\alpha = 0.047$ is determined. Then, according to Expression (13), the upper limit of the scattering of useful information of the phonation process for person №1 is equal to $\sup H(X|X') = \alpha \log R = 0.235$, and the upper limit of phonetic saturation of speech for person №1, according to Expression (11), is equal to $\sup I(X|X') = (1-\alpha) \log R = 4.765$.

Similar calculations were made for the rest of the persons in the team. For clarity of presentation, these results were averaged for each of the three age groups and visualized in Figure 2. In addition, for comparison, for persons from the first age group, the mismatch matrices were calculated with the selected soft set of formation conditions, and we performed all other computational operations described above. These results, referred to as «1^{AG}_{soft}», are also shown in Figure 2.

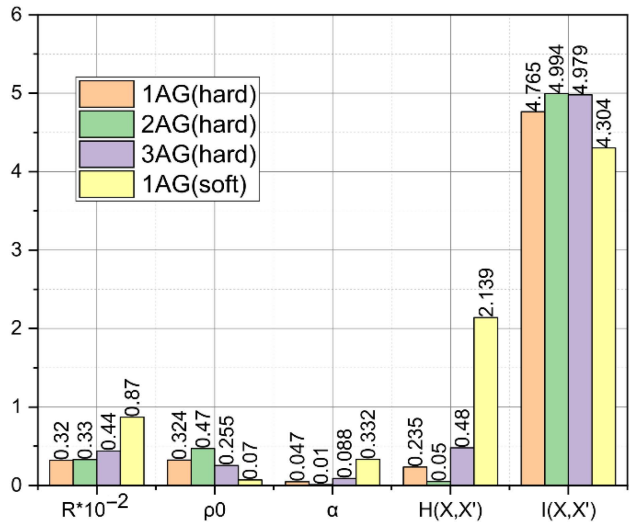

Figure 2. Estimation of phonetic saturation of personified speech (the format of the numbers is determined by the computing software used).

We investigate empirically the functionality of decision-making concepts generalized by solving Rules (22) and (25) in the task of the computational phonetic analysis of speech (statistical classification without a teacher in the concept (8)–(10) paradigm). The empirical material for the research was two phonograms with a recording of the same content of language material spoken by person №1. Phonograms were represented by samples X_0, X_r of equal capacity $M = 120$. First, the information mismatch matrix $\|\rho_{r,v}\|$ was calculated for four vowel phonemes of the person №1. The content of the matrix is visually presented in Figure 3. Allophones $[\mathrm{u}:]_1$ and $[\mathrm{u}:]_2$ represent person-specific dialects of pronunciation of the phoneme $[\mathrm{u}:]$.

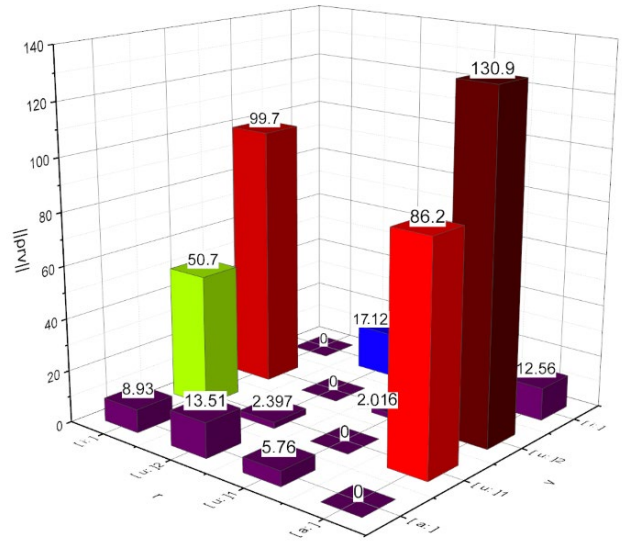

Figure 3. Visualization of a matrix $\|\rho_{r,v}\|$ calculated for instances of phonation of four phonemes by person №1 (the format of the numbers is determined by the computing software used).

Further use of the data presented in Figure 3 will be demonstrated by the example. Consider the data from the matrix $\|\rho_{r,v}\|$ for a pair of phonemes $([\text{a :}], [\text{u :}]_1)$. These data characterize the situation when the phoneme [a :] is recognized as a phoneme $[\text{u :}]_1$. Figure 3 shows that $\rho([\text{a :}], [\text{u :}]_1) = 5,76$. The number of degrees of freedom for the F-distribution in expression (28) is assumed to be equal to $K = M - p = 100$. If the decision on the result of phoneme recognition is made according to the solution rule (25), then $(K, 1) = (100, 1)$, and we have $\widetilde{\alpha}_{v\to r} = 1 - \Phi_{K,1}\{1 + \rho_{r,v}\} = 1 - \Phi_{100,1}\{5.76\} \approx 0.3$. If the decision on the result of phoneme recognition is made according to the decision rule (22), then we have $\max[(1 + \rho_{vr}); (1 + \rho_{rv})] = \max[87.2; 6.76] = 6.76$. According to Expression (28), we have $\alpha_{v\to r} \leq 1 - \Phi_{100,1}(5.76) \approx 0.12$. Thus, the probability of confusion when deciding on the result of phonetic analysis on the example of phonemes [a :] and $[\text{u :}]_1$ using the solution rule (22) in comparison with Rule (25) is almost three times less. Calculations similar to the above were performed for all pairs of phonemes of different names in Figure 3. For all implementations, Rule (22) allowed us to obtain a lower estimate of the probability of confusion compared to Rule (25).

Let us complete this stage of research by calculating by means of Expression (31) the probability of the event of marginal recognition of the correct result of the phoneme recognition procedure using the decision rule (25): $\pi_{r\to v} = 1 - \Phi_{1,1}\left\{\frac{1+86.2}{1.59(1+5.76)}\right\} = 1 - \Phi_{1,1}(8, 11) \approx 0.21$. It can be stated that the greater the asymmetry between the opposing elements of the matrix $\|\rho_{rv}\|$, the greater the value of probability $\pi_{r\to v}$.

We generalize the experimental section by verifying the models proposed in Section 2 in the paradigm of practical planning theory. We form certain sets of input influences (speech synals): $X^k = \left\{x_1^k, x_2^k, \ldots, x_n^k\right\}$ and $X^{\bar{k}} = \left\{x_1^{\bar{k}}, x_2^{\bar{k}}, \ldots, x_m^{\bar{k}}\right\}$. The system's response to input effects from the set X^k is predicted in the consept. Input influences from the set $X^{\bar{k}}$ are structurally identical to the generalized set X^k but differ in values that may exceed the limits set up in the system's design stage (extraneous noises, significant problems with diction, etc.) The system's reaction to the input influence from the set $X^{\bar{k}}$ can be incorrect speech unit recognition. The numbers of elements in the sets X^k and $X^{\bar{k}}$ are $n = 3000$ and $m = 7000$, respectively. Experiments were performed with fixation on the reaction of the system to the input influences from the sets X^k and $X^{\bar{k}}$ (in the matrix form $B_e^k = \left(B_{ij}^k\right)$, $i = \overline{1,n}$, and $B_e^{\bar{k}} = \left(B_{ij}^{\bar{k}}\right)$, $i = \overline{1,m}$, respectively). We calculate for the ith input influence the variance of the implementation of the situation of the incorrect speech unit recognition: $s_i^2 = M^{-1}\sum\limits_{j=1}^{M}\left(B_{ij} - B'_{ij}\right)^2$, where B_{ij} is the state defined in the model; B'_{ij} is the actual state. We calculate the average value of the variance for all input influences: $s^2 = N^{-1}\sum\limits_{i=1}^{N} s_i^2$. Evaluation of the substantial deviations s_i^2 from s^2 Fisher's criterion showed that all deviations do not exceed the tabular values, which confirms the adequacy of the proposed mathematical apparatus.

4. Discussion

The task of computational phonetic analysis of speech in the general case is reduced to a cyclically repeated procedure for estimating the deviation of the current segment of the studied speech signal from the etalons defined within a finite list of language units. The duration of the segments, by the sequence of which the output speech signal is presented, is selected based on the average duration of the studied language units; for example, for phonemes it is $\tau \in \left\{\overline{5,10}\right\}$ ms. In the paradigm of the Bayesian theory of pattern recognition, such a task is solved by testing stochastic hypotheses about the homogeneity of the distribution law of the speech signal. If the empirical distribution law can be reliably estimated by Gaussian approximation, then the above-mentioned task has an optimal solution. If the procedure of comparing the empirical segment with the etalon is trivial, then the question of determining the etalon for the language unit is a cornerstone.

There is no generally accepted definition of the etalon of a language unit in the context of computational phonetic analysis of speech. A typical approach is to determine the desired etalon based on one of the variations of the method of expert assessments. However, this approach examines not so much the phonation of the language unit as the environment of distribution of signal and format of its presentation. In this context, the derivation of the task of computational phonetic analysis in the subject area of information theory allows us to consider the definition of the etalon in the absolute metric of the criterion of relative entropy, rather than in the relative metric, as implemented in analogues.

From the empirical results shown in Figure 2, we can draw conclusions about the representativeness of the proposed metric $\{H(X|X'); I(X|X')\}$ for estimating the personified phonetic saturation of speech. It turned out that the highest phonetic saturation (11) is characterized by the speech of persons from the second age group. Of particular note are the data characterizing the phonetic saturation of speech of persons from the first age group, whose phonetic alphabets were determined by choosing a hard and soft set of formation conditions-1^{AG}_{hard} and 1^{AG}_{soft}, respectively. It is seen that the phonetic saturation of speech of persons from the first age group $jI^{1AG}_{hard} = 4.765$, estimated based on phonetic alphabets $R^{1AG}_{hard} = 32$, determined by choosing a hard set of formation conditions, was higher than the same indicator for the same group of persons $I^{1AG}_{soft} = 4.304$, estimated based on phonetic alphabets $R^{1AG}_{soft} = 87$, determined by choosing a soft set of formation conditions. This is without assuming that the capacity of the phonetic alphabet of the second variant $R^{1AG}_{soft} = 87$ exceeds the capacity of the phonetic alphabet of the first variant $R^{1AG}_{hard} = 32$ more than twice. This fact allows us to outline a promising direction for the investigation of the function $I(X, X') = f(R, \Delta\rho, \Delta L)$, the extremum of which can potentially indicate the elements of the personalized phonetic alphabet, in which the individuality and informativeness of speech are most pronounced.

Based on the relative entropy function, Section 2.3 theoretically substantiates two error-detectable approaches to decision-making $W_v(X)$ in the task of computational phonetic analysis of speech (15) based on decision rules (22) or (25). The results of the computational experiment presented in Figure 3 convincingly prove the functionality of both of these approaches. Of particular importance is Expression (31) to estimate the reliability of a decision made based on Rule (25). Indeed, if the solution $W_\mu(X)$ is found to be erroneous according to Expression (29), then this fact, according to the provisions of the theory of experimental planning, will oblige the researcher to repeat the experiment according to Scheme (15) with all already rejected distribution alternatives, because the decision on their marginality is compromised. The result of such a re-experiment

$$\widetilde{\widetilde{W}}_v(X) : \widetilde{\widetilde{\lambda}}_v(X) \triangleq \gamma_{0,r}|_{r=v\neq\mu} = \min, \tag{32}$$

determined on a reduced sample of alternatives with capacity $R - 1$, together with the solution rules (25), (29), defines the entropy-based concept of detecting and correcting errors in the computational phonetic analysis of speech. The potential inherent in the proposed concept and the demonstrated results prove its superiority over such Bayesian concepts of decision-making using Euclidean-type mismatch metrics as the method of maximum likelihood and the method of the ideal observer.

Finally, it should be noted that the mathematical apparatus proposed in this article is proved to be adequate because it is based on the verified mathematical apparatus of information theory. This fact, as well as the rigor and reversibility of the analytical transformations carried out in the formalization of the corresponding metric, substantiate the adequacy of the mathematical apparatus presented in the article.

5. Conclusions

The study of a cornerstone object for modern linguistics, the process of speech and textual interpersonal communication, considering the size of the infosphere of the twenty-

first century, is impossible without a thorough and purposeful involvement of information technology from other fields of knowledge, including computer science. Created as a result of relatively young science, computational linguistics aims to automatically analyze natural languages in all spectra of their implementation. From the long list of current tasks actively studied in the paradigm of computational linguistics, we mention the automation of compilation and linguistic processing of language corpora, the automated classification and abstracting of documents, the creation of accurate linguistic models of natural languages, ad the extraction of factual information from informal linguistic data. An effective, strictly formalized technology of computational phonetic analysis of linguistic information, especially speech information, is potentially the driving force behind the improvement of the results of solving these research tasks. This thesis is fully consistent with the content of the article, which proves the relevance of the presented scientific and applied results.

The proposed concept in this article (i.e., the mathematical model and methods) of computational phonetic analysis of speech defines the **scientific novelty** of the research. In the concept, in contrast with the existing methods, the task of addressing the multicriteria of the process of cognitive perception of speech by a person is strictly formally presented in the theoretical and analytical apparatus of information theory, pattern recognition theory and acoustic theory of speech formation. The obtained concept allows for determining accurately the phonetic alphabet of a person, taking into account their inherent dialect of speech and individual features of phonation, as well as detecting and correcting errors in the recognition of language units and reliably assessing the phonetic saturation of speech.

The proposed concept is represented by the decision rule (2), the decision statistics (4) and Expressions (7)–(9). The central element of the concept is the matrix of information mismatch $\|\rho_{r,v}\|$ of language units of the personalized phonetic alphabet of the speaker. The matrix $\|\rho_{r,v}\|$ is the basis for calculating the threshold ρ_0 for the implementation of computational phonetic analysis by Expression (7). With a known value of ρ_0, based on Expressions (2) and (5), the procedure of segmentation of the studied phonetic alphabet of a speaker into a set of phonemes, which with probability $\beta = 1 - \alpha$ are reliably recognized despite disturbing factors, and another set of phonemes, which with probability α are not reliably recognized. The use of Expressions (9) and (10) allows for clarifying the result of the segmentation procedure, taking into account the variability of the phonation of the studied language units, introduced by the individual features of the articulation of a particular speaker.

The study of the results of computational phonetic analysis based on the function of relative entropy allowed for substantiating theoretically two detectable errors of the process of recognition of language units (15) based on solving Rules (22) and (25). Note the possibility, formalized by Expression (31), to estimate the reliability of the decision made based on Rule (25). If the solution is found to be compromised according to Expression (29), then with the help of a computational procedure with Scheme (15), it is possible to find erroneously recognized unreliable results of phonetic analysis and rehabilitate them. Thus, the **practical significance** of the proposed concept of computational phonetic analysis of speech lies in the fact that with its help, it is possible not only to single out phonetic units in speech signals, taking into account the individual features of speech formation, but also to detect and correct errors in the results of such an analysis.

The potential inherent in the proposed concept and the experimental results presented after Figure 3 prove its superiority over such Bayesian decision-making concepts using Euclidean-type mismatch metrics as the maximum likelihood method and the ideal observer method. The analysis of the studied speech signal carried out in the metric $\{H(X|X'); I(X|X')\}$ based on the proposed concept allows for establishing reliably the phonetic saturation of speech, which objectively characterizes the environment of speech signal propagation and its source.

Further research is planned to analyze the function $I(X, X') = f(R, \Delta\rho, \Delta L)$, the extremum of which can potentially indicate the elements of the personalized phonetic alphabet, in which the individuality and informativeness of the speech of the person are

most apparent. The authors hope that the results of such an investigation will increase the practical value of the proposed system of models for the precision phonetic analysis of speech [39].

Author Contributions: Conceptualization, V.K. and O.K.; methodology, V.K. and O.K.; software, V.K. and O.K.; validation, V.K. and O.K.; formal analysis, V.K. and O.K.; investigation, V.K. and O.K.; resources, V.K. and A.S.; data curation, V.K. and O.K.; writing—original draft preparation, V.K. and O.K.; writing—review and editing, V.K. and O.K.; visualization, V.K. and O.K.; supervision, V.K. and O.K.; project administration, V.K. and O.K.; funding acquisition, A.S. and V.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Most data are contained within the article. All the data available on request due to restrictions, e.g., privacy or ethical.

Acknowledgments: The authors would like to thank the anonymous reviewers who helped better present the research results.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Almutiri, T.; Nadeem, F. Markov Models Applications in Natural Language Processing: A Survey. *Int. J. Inf. Technol. Comput. Sci.* **2022**, *2*, 1–16. [CrossRef]
2. Bhanja, C.C.; Laskar, M.A.; Laskar, R.H. Modelling multi-level prosody and spectral features using deep neural network for an automatic tonal and non-tonal pre-classification-based Indian language identification system. *Lang. Resour. Eval.* **2021**, *55*, 689–730. [CrossRef]
3. Umasankar, C.D.; Ram, M.S.S. Speech Enhancement through Implementation of Adaptive Noise Canceller Using FHEDS Adaptive Algorithm. *Int. J. Image Graph. Signal Process.* **2022**, *3*, 11–22. [CrossRef]
4. Firooz, S.G.; Reza, S.; Shekofteh, Y. Spoken language recognition using a new conditional cascade method to combine acoustic and phonetic results. *Int. J. Speech Technol.* **2018**, *21*, 649–657. [CrossRef]
5. Sunitha, P.; Prasad, K.S. Speech Enhancement based on Wavelet Thresholding the Multitaper Spectrum Combined with Noise Estimation Algorithm. *Int. J. Image Graph. Signal Process.* **2019**, *11*, 44–55. [CrossRef]
6. Pujar, R.S. Wiener Filter Based Noise Reduction Algorithm with Perceptual Post Filtering for Hearing Aids. *Int. J. Image Graph. Signal Process.* **2019**, *11*, 69–81. [CrossRef]
7. Bender, E.M.; Drellishak, S.; Fokkens, A.; Poulson, L.; Saleem, S. Grammar Customization. *Res. Lang. Comput.* **2010**, *8*, 23–72. [CrossRef]
8. Al-Bakeri, A.A. ASR for Tajweed Rules: Integrated with SelfLearning Environments. *Int. J. Inf. Eng. Electron. Bus.* **2017**, *9*, 1–9. [CrossRef]
9. Moran, S.; Grossman, E.; Verkerk, A. Investigating diachronic trends in phonological inventories using BDPROTO. *Lang Resour. Eval.* **2020**, *55*, 79–103. [CrossRef]
10. Peleshko, D.; Rak, T.; Izonin, I. Image Superresolution via Divergence Matrix and Automatic Detection of Crossover. *Int. J. Intell. Syst. Appl.* **2016**, *8*, 1–8. [CrossRef]
11. Chittaragi, N.B.; Koolagudi, S.G. Automatic dialect identification system for Kannada language using single and ensemble SVM algorithms. *Lang. Resour. Eval.* **2020**, *54*, 553–585. [CrossRef]
12. Izonin, I.; Trostianchyn, A.; Duriagina, Z.; Tkachenko, R.; Tepla, T.; Lotoshynska, N. The Combined Use of the Wiener Polynomial and SVM for Material Classification Task in Medical Implants Production. *Int. J. Intell. Syst. Appl.* **2018**, *10*, 40–47. [CrossRef]
13. Kurimo, M.; Enarvi, S.; Tilk, O.; Varjokallio, M.; Mansikkaniemi, A.; Alumäe, T. Modeling under-resourced languages for speech recognition. *Lang. Resour. Eval.* **2017**, *51*, 961–987. [CrossRef]
14. Masmoudi, A.; Bougares, F.; Ellouze, M.; Estève, Y.; Belguith, L. Automatic speech recognition system for Tunisian dialect. *Lang. Resour. Eval.* **2018**, *52*, 249–267. [CrossRef]
15. Elvira-García, W.; Roseano, P.; Fernández-Planas, A.M.; Martínez-Celdrán, E. A tool for automatic transcription of intonation: Eti_ToBI a ToBI transcriber for Spanish and Catalan. *Lang. Resour. Eval.* **2016**, *50*, 767–792. [CrossRef]
16. Hu, Z.; Mashtalir, S.V.; Tyshchenko, O.K.; Stolbovyi, M.I. Clustering Matrix Sequences Based on the Iterative Dynamic Time Deformation Procedure. *Int. J. Intell. Syst. Appl.* **2018**, *10*, 66–73. [CrossRef]
17. Aissiou, M. A genetic model for acoustic and phonetic decoding of standard arabic vowels in continuous speech. *Int. J. Intell. Syst. Appl.* **2020**, *23*, 425–434. [CrossRef]

18. Hu, Z.; Tereykovski, I.A.; Tereykovska, L.O.; Pogorelov, V.V. Determination of Structural Parameters of Multilayer Perceptron Designed to Estimate Parameters of Technical Systems. *Int. J. Intell. Syst. Appl.* **2017**, *9*, 57–62. [CrossRef]
19. Chittaragi, N.B.; Koolagudi, S.G. Acoustic-phonetic feature based Kannada dialect identification from vowel sounds. *Int. J. Speech Technol.* **2019**, *22*, 1099–1113. [CrossRef]
20. Kleynhans, N.T.; Barnard, E. Efficient data selection for ASR. *Lang. Resour. Eval.* **2015**, *49*, 327–353. [CrossRef]
21. Hu, Z.; Ivashchenko, M.; Lyushenko, L.; Klyushnyk, D. Artificial Neural Network Training Criterion Formulation Using Error Continuous Domain. *Int. J. Mod. Educ. Comput. Sci.* **2021**, *13*, 13–22. [CrossRef]
22. Vinola, F.A.F.; Padma, G. A probabilistic stochastic model for analysis on the epileptic syndrome using speech synthesis and state space representation. *Int. J. Speech Technol.* **2020**, *23*, 355–360. [CrossRef]
23. Mehrabani, M.; Hansen, J.H.L. Automatic analysis of dialect/language sets. *Int. J. Speech Technol.* **2015**, *18*, 277–286. [CrossRef]
24. Rello, L.; Baeza-Yates, R.; Llisterri, J. A resource of errors written in Spanish by people with dyslexia and its linguistic, phonetic and visual analysis. *Lang. Resour. Eval.* **2016**, *51*, 379–408. [CrossRef]
25. Chaki, J. Pattern analysis based acoustic signal processing: A survey of the state-of-art. *Int. J. Speech Technol.* **2020**, *24*, 913–955. [CrossRef]
26. Bhangale, K.B.; Mohanaprasad, K. A review on speech processing using machine learning paradigm. *Int. J. Speech Technol.* **2021**, *24*, 367–388. [CrossRef]
27. Verma, P.; Das, P.K. i-Vectors in speech processing applications: A survey. *Int. J. Speech Technol.* **2015**, *18*, 529–546. [CrossRef]
28. Drugman, T.; Dutoit, T. The Deterministic Plus Stochastic Model of the Residual Signal and Its Applications. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *20*, 968–981. [CrossRef]
29. Chen, X.; Bao, C. Phoneme-Unit-Specific Time-Delay Neural Network for Speaker Verification. *IEEE ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1243–1255. [CrossRef]
30. Hu, Z.; Tereikovskyi, I.; Chernyshev, D.; Tereikovska, L.; Tereikovskyi, O.; Wang, D. Procedure for Processing Biometric Parameters Based on Wavelet Transformations. *Int. J. Mod. Educ. Comput. Sci.* **2021**, *13*, 11–22. [CrossRef]
31. Omer, A.I.; Zampieri, M.; Oakes, M.M. Phonetic differences for dialect clustering. In Proceedings of the 9th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 3–5 April 2018; pp. 145–150. [CrossRef]
32. Viacheslav, K.; Kovtun, O. System of methods of automated cognitive linguistic analysis of speech signals with noise. *Multimedia Tools Appl.* **2022**, 1–20. [CrossRef]
33. Bisikalo, O.; Boivan, O.; Kovtun, O.; Kovtun, V. Research of the Influence of Phonation Variability on The Result of the Process of Recognition of Language Units. *CEUR Workshop Proc.* **2022**, *3156*, 82–93.
34. Kannadaguli, P.; Bhat, V. A comparison of Bayesian multivariate modeling and hidden Markov modeling (HMM) based approaches for automatic phoneme recognition in kannada. *Recent Emerg. Trends Comput. Comput. Sci.* **2015**, 1–5. [CrossRef]
35. Laleye, F.A.A.; Ezin, E.C.; Motamed, C. Automatic Text-Independent Syllable Segmentation Using Singularity Exponents And Rényi Entropy. *J. Signal Process. Syst.* **2016**, *88*, 439–451. [CrossRef]
36. Kang, J.; Zhang, W.-Q.; Liu, W.-W.; Liu, J.; Johnson, M.T. Lattice Based Transcription Loss for End-to-End Speech Recognition. *J. Signal Process. Syst.* **2017**, *90*, 1013–1023. [CrossRef]
37. Qian, Y.; Ubale, R.; Lange, P.; Evanini, K.; Ramanarayanan, V.; Soong, F.K. Spoken Language Understanding of Human-Machine Conversations for Language Learning Applications. *J. Signal Process. Syst.* **2019**, *92*, 805–817. [CrossRef]
38. Cui, Y.; Sirén, J.; Koski, T.; Corander, J. Simultaneous Predictive Gaussian Classifiers. *J. Classif.* **2016**, *33*, 73–102. [CrossRef]
39. Bisikalo, O.; Boivan, O.; Khairova, N.; Kovtun, O.; Kovtun, V. Precision Automated Phonetic Analysis of Speech Signals for Information Technology of Text-dependent Authentication of a Person by Voice. *CEUR Workshop Proc.* **2021**, *2853*, 276–288.