Research of the Influence of Phonation Variability on The Result of the Process of Recognition of Language Units

Oleg Bisikalo[a], Olesia Boivan[b], Oksana Kovtun[b] , 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*

Abstract

The limited use of profile services in the corporate and government segments of cyberspace suggests that the task of recognizing the speech of more than one speaker in non-laboratory conditions is still relevant. The article presents the technology of improving the process of recognition of language units by integrating the model of the variability of their phonation in the decision rule. In the proposed technology, in contrast to existing ones, recognition occurs at the level of comparison of sound schemes of empirical and etalon language material in the common parametric space of acoustic, generative and language models. This allowed us to formalize the concepts of taking into account the influence of phonation variability in determining the etalon sound schemes of language units in the paradigm of pattern recognition theory and to formulate a UML activity diagram of the mechanism for calculating the parameters of these concepts. The classification results demonstrated in the test sample with high variability of speech material prove the functionality of the author's mechanisms to compensate for the influence of phonation variability at the level of the decision rule and increase the accuracy of recognition by 5-8% (from the original 52% to 57-60%, respectively). Experiments have shown that for all test samples, the decision-making rules formulated based on the author's concept, which took into account the optimal and suboptimal etalon sound schemes, respectively, exceeded the solving rule, which took into account the etalon sound schemes, but their frequency was ignored. It turned out that it is not advisable to use the author's mechanisms to compensate for the influence of phonation variability in the classification of speech material with a low or moderate degree of variability.

Keywords

Computational linguistics; language units recognition; phonation variability; decision rule; sound scheme

1. Introduction

At the present stage of the development of computer technology, automated recognition and synthesis of speech signals are probably one of the most relevant services of human-machine interfaces of control systems, in particular, in case of emergency. Indeed, the built-in control system of the signal recognition and synthesis subsystem will save the time needed to enter information, alert subscribers and make decisions, and thus prevent or at least reduce the damage caused by an emergency. Consider the current systems of speech signal recognition in more detail.

The profile for our study system of speech signal recognition [1-5] in general can be described as an asynchronously functioning conglomerate of acoustic models, vocabulary, language model and

IntelITSIS'2022: 3rd International Workshop on Intelligent Information Technologies and Systems of Information Security, March 23–25, 2022, Khmelnytskyi, Ukraine

EMAIL: obisikalo@gmail.com (O. Bisikalo); olesiaboivan@gmail.com (O. Boivan); 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); 0000-0002-9139-8987 (O. Kovtun); 0000-0002-7624-7072 (V. Kovtun)

classifier. If acoustic models estimate the probabilities of recognition of individual language units of a certain level in speech, then language models estimate the probability of the order of location of those language units in the signal. The dictionary must contain all possible variants of pronunciation of language units that will be recognized during the operation of the profile system. The classifier determines the best hypothesis in the recognition network. It is a software mechanism that operates on large amounts of data and has to decide in the shortest possible time on the sequence of segments of the phonogram of the input speech signal. The functionality of language unit recognition systems is determined mainly by the speed of the recognition process and its accuracy.

When determining the system of speech signal recognition, it is necessary to take into account several aspects, namely [3, 6-10]:

1. The size of the vocabulary. The larger the size of the vocabulary with which the speech signal recognition system operates, the higher the frequency of errors in the recognition of language units. For example, the frequency of errors in vocabulary recognition in one hundred thousand lexemes can reach 45%. The uniqueness of lexemes in the vocabulary should be taken into account. If the lexemes are phonetically similar, the recognition error will increase.

2. Speaker addiction. There are speaker-dependent and speaker-independent profile systems. The first type of system is intended to be operated by only one user (a person whose speech material was used to train the system), while systems of the second type are focused on the operation by an arbitrary user. At the current stage of development of speech signal recognition systems, the frequency of errors in speaker-independent systems is 5-8 times higher than a similar quality indicator for speaker-dependent systems.

3. The level of structural representation of the speech signal. Phrases, lexemes, two or three phonemes, diphones, allophones, etc. can act as structural units in speech signal recognition systems. Profile systems in which whole lexemes or phrases are analyzed are called templates. They are usually speaker-dependent, and their implementation is much less time-consuming than creating systems that recognize speech signals at the phonetic level (a sequence of phonemes, diphones, allophones).

4. The principle of allocation of language units in speech. In modern profile systems, several approaches are used to extract the language units from the phonogram of a speech signal. The most common approach is based on the Fourier transform, which translates the input signal from the amplitude-temporal space into the frequency-temporal space. For the analysis of the speech signal in the temporal area, the linear prediction method is most often used, which allows describing the analyzed signal as a model of autoregression. However, Fourier analysis has several shortcomings, which are manifested in the loss of important information about the short-term amplitude-frequency characteristics of the processed signals. Therefore, the use of, for example, wavelet transform, which allows for the analysis of the properties of the studied signal in both temporal and frequency spaces, is justified for the selection of language units.

5. Classification mechanism. After segmentation of the input speech signal, the sequence of the received fragments of the phonogram is parameterized and the software mechanism-classifier performs a probabilistic estimation of the affiliation of each of them to the reference elements from the vocabulary. The most widespread in modern systems of speech signal recognition have become various methods of machine learning, among which we note the hidden Markov models and artificial neural networks.

The field of application of speech recognition systems is constantly expanding – from software applications for converting speech information into text and ending with on-board hardware control devices. Depending on the area of application, the following classes of profile systems are distinguished [2, 7, 11-13]:

1. Software cores for hardware implementations of speech signal recognition systems. Depending on the purpose, systems of this class are divided into Text-to-Speech (TTS) and Automatic Speech Recognition (ASR). TTS cores are focused on converting text into a speech signal, and ASR cores are designed to represent the speech signal as text.

2. Libraries of utilities for the development of specialized software services for speech signal recognition, which are later integrated into human-machine interfaces.

3. Independent user programs designed for voice control and/or conversion of the speech signal into text.

4. Focused on critical use programs for speech signal recognition.

5. Devices for speech recognition, such as neural network microcontrollers VP-2025 from Primestar Technology Corporation.

Thus, the problem of creating a universal system for recognizing speech signals is relevant and far from being solved. Based on the analysis of existing analogues, we formulate the ***object*** of study as a speaker-dependent process of phonation of the speech signal. The ***subject*** of the study is the provisions of the theory of pattern recognition and the theory and mathematical statistics.

2. Models and methods

2.1. Research statement

An applied result of automated phonetic analysis of the phonogram of the speech signal is the sound scheme of the latter. However, the sound scheme characterizes a certain lexeme both semantically and acoustically. The variability of sound schemes, due to the speaker-dependence of speech, is a source of uncertainty for the task of recognition of language units in speech. We formalize this variability in the mathematical apparatus of pattern recognition theory [14, 15].

Let $X=\{x_t\}$, $t=\overline{1,T}$, be a parameterized pattern of the phonogram of the speech signal, and $W=\{w_i\}$, $i=\overline{1,N}$, be a phrase or a sequence of lexemes, which is presented in the vocabulary of the corresponding language.

The result of the recognition of the empirical pattern X is finding the most probable sequence of lexemes W^*, which can be analytically described by the expression

$$W^*=\arg\max P(W|X)=\arg\max_{W}\frac{P(W)}{P(X)}P(X|W), \tag{1}$$

where the relative probability $P(X|W)$ characterizes the plausibility of empirical data in the parametric space of the selected acoustic model of a corresponding sequence of lexemes; the probability $P(W)$ characterizes the etalon phonation of a corresponding sequence of lexemes generated by the acoustic model; probability $P(X)$ characterizes the representation of the empirical phonogram of the speech signal by an acoustic model and performs in expression (1) the function of normalization. In this context, we define the acoustic model of phonation of the lexeme w as a sound scheme t^w. The variability of phonation leads to the fact that the lexeme w will be characterized not by a single sound scheme t^w, but by their plural, generalized by the set T^w. Continuing this symbolic chain, the variability of the phonation of the entire vocabulary of lexemes is characterized by the set T^W, where the parameter t^W, $t^W\in T^W$, identifies a certain individual trajectory of the phonation in the set of sound schemes of the vocabulary W.

In current systems of automated speech signals recognition, when defining the criterion (1) substitution of concepts is carried out, which can be described by the expression

$$t^{W^*}=\arg\max_{t^W}\frac{P(t^W)}{P(X)}P(X|t^W), \tag{2}$$

That is, the actual result of recognition is not a sequence of lexemes, but a sequence of sound schemes defined in the selected a priori imperfect acoustic model. The desired result (1) based on (2) is formed as a result of the literal application of the operations of lexeme classification of the form:

$$t^{W^*}\to W^*. \tag{3}$$

If the phenomenon of the variability of phonation of language units can be neglected, then concepts (1) and (2) become identical:

$$W^*=\arg\max_{t^W}\frac{P(t^W|W)P(W)}{P(X)}P(X|t^W), \tag{4}$$

where $P\left(t^W \middle| W\right)P(W)=P\left(t^W\right)$. Based on this thesis, we state that the variability of phonation in the task of recognition of language units in speech is determined by the composite probability $P\left(T^W \middle| W\right)=\left\{P\left(t^W \middle| W\right), t^W \in T^W\right\}$.

Thus, the ***research aim***s to substantiate the probability $P\left(T^W \middle| W\right)$ in the context of the task of recognition of language units in speech. The ***objectives*** of the study are: - to mathematically define the recognition of language units in speech as a stochastic process of comparing the sound scheme of the empirical phonogram with the etalon sound scheme, determined taking into account the variability of phonation of language units from the acoustic-phonetic vocabulary; - to formulate the concept of applied use of the proposed model of the recognition process; - to conduct empirical research of the proposed approach to the recognition of language units in speech.

2.2. Mathematical formalization of the investigated process

Defined in general form in expression (4), the parametric space of the desired model of phonation variability of language units is a subspace of the general parametric space formed by three models – acoustic (performs direct-inverse representation of "phonogram"-"sound scheme"), generative (describes empirical phonogram). signal) and speech (describes the etalon phonogram, which probably corresponds to the empirical signal). The method of maximum a posteriori probability allows defining this subspace as in the first approximation as

$$W^* = \arg\max_{t^W} \frac{P\left(t^W \middle| W\right)P(W)P\left(X \middle| t^W\right)}{\sum\limits_{t^W \in T^W} P\left(t^W \middle| W\right)P(W)P\left(X \middle| t^W\right)}. \tag{5}$$

The parameters of acoustic, generative and speech models can be considered a priori independent because the first focuses on the parameterization of the speech signal, the second focuses on the reproduction of the speech signal, and the third focuses on determining the sound schemes of the vocabulary. The characteristic parameters of the language model $P(W)$ do not depend on the phonation of empirical speech signals. For their estimation, it is necessary to use the material of the profile language corpus. At the same time, the characteristic parameters of the generative model of phonation $P\left(T^W \middle| W\right)$ directly depend on the empirical speech signals. This fact determines the feasibility of studying generative and acoustic models together because the first provide empirical material, and the second determines the way of its compact presentation.

The applied use of criterion (5) is complicated by an objective problem – the potential lack of a representative corpus for the studied language. The authors are familiar with representative corpora for the English language, such as the TIMIT acoustic-phonetic Continuous Speech Corpus [16, 17]. For the Ukrainian language, the General Regionally Annotated Corpus of Ukrainian (GRAC) is a fundamentally comparable analogue.

The method of maximum likelihood [18, 19] allows estimating the configuration of the parametric space of the phonation variability model easier than the method of maximum a posteriori probability. We can limit ourselves to determining only the numerator of criterion (5). Suppose that there is such an acoustic-phonetic corpus X for the studied language that for an arbitrary phrase not only the sequence of sound schemes of lexemes $w_1 w_2 \ldots w_N$ is known, but also phonograms with their etalon phonation $t_1^w t_2^w \ldots t_N^w$. Analytical interpretation of the method of maximum likelihood for estimating the characteristic parameters of phonation variability will look like this:

$$p\left(t^w \middle| w\right) = \arg\max_{w,t^w} \prod_{w,t^w} p\left(t^w \middle| w\right) = \frac{\mathrm{COUNT}\left\{t^w\right\}}{\mathrm{COUNT}\left\{w\right\}}, \tag{6}$$

where $\mathrm{COUNT}\{\cdot\}$ is the function of counting the number of elements in the set-argument.

From expression (6) it can be concluded that the estimation of the probability of observing a model of a lexeme is directly related to the relative frequency of the presence of this lexeme in the etalon training material.

Let’s use the adapted form of the method of maximum likelihood (6) for "by-coordinate" estimation of the characteristic parameters of interdependent generative and acoustic models. "By-coordination" is implemented as follows:

1. Despite the variability of phonation but taking into account the a priori known order of lexemes in phrases, let's recognize empirical acoustic models based on etalon data from the relevant language corpus;
2. Determine the most probable sequences of lexemes (5) and by expression (6) to determine the most commonly used variants of their phonation;
3. Based on the information obtained in st. 2, update the default parameters of the acoustic model, focusing it on the most common version of the generation of speech signals inherent in the studied language.

This mechanism of adaptation of generative and acoustic models to the variability of phonation of speech signals is focused on the application by presenting in the form of UML activity diagrams (see Fig. 1).

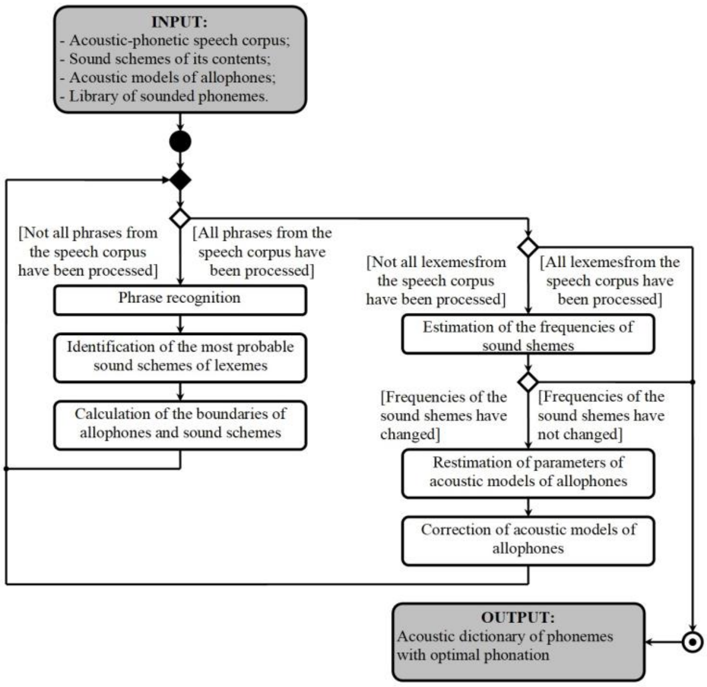

INPUT:
- Acoustic-phonetic speech corpus;
- Sound schemes of its contents;
- Acoustic models of allophones;
- Library of sounded phonemes.
[Not all phrases from the speech corpus have been processed]
[All phrases from the speech corpus have been processed]
Phrase recognition
Identification of the most probable sound schemes of lexemes
Calculation of the boundaries of allophones and sound schemes
[Not all lexemesfrom the speech corpus have been processed]
[All lexemesfrom the speech corpus have been processed]
Estimation of the frequencies of sound shemes
[Frequencies of the sound shemes have changed]
[Frequencies of the sound shemes have not changed]
Restimation of parameters of acoustic models of allophones
Correction of acoustic models of allophones
OUTPUT:
Acoustic dictionary of phonemes with optimal phonation

Figure 1: UML activity diagram for adaptation of generative and acoustic models to the phonation variability of speech signals

Naturally, speech is a dynamic object [20, 21], so the content of the language corpus must be periodically updated to reflect changes in generally accepted trends in the phonation of language units. If the language corpus focuses on the task of recognizing language units, then this update must be carried out based on expressions (2) & (3) according to the procedure defined in Fig. 1.

However, this approach, although strategically correct, is not computationally efficient. Let's try to get rid of this shortcoming. Convert expression (1) as follows:

$$P(W|X)=\frac{P(W|X)}{P(X)}=\frac{\sum_{t^W\in T^W}P(X,t^W)}{P(X)}=\frac{\sum_{t^W\in T^W}P(X|t^W)P(t^W)}{P(X)}. \qquad (7)$$

If we substitute expression (7) into criterion (4), then we define the most probable sequence of lexemes as

$$W^*=\arg\max_{W}\sum_{t^W\in T^W}P(t^W|X)P(t^W). \qquad (8)$$

Criterion (8) is directly focused on the task of speech signal recognition because it allows determining the most probable sequence of lexemes in the empirical phonogram of the speech signal, rather than the most probable sequence of sound schemes available in it (this is the criterion (2) & (3)). The concept of applied use of criterion (4) differed from the concept of applied use of criterion (2) & (3) in that the latter has to take into account the probability of realization of the sound scheme of the lexeme, i.e. the decision on the lexeme's plausibility is obtained as a weighted the sum of the plausibility of the implementation of all its sound schemes. In the implementation of the concept generalized by criterion (8), the sequence of actions presented in Fig. 1, will have to be supplemented by the operation of selecting the best sequence of lexemes.

Accordingly, if each lexeme w from the language corpus X corresponds to the probability

$$P(w)=\sum_{t^w \in T^w} P\left(t^w \middle| X\right), \tag{9}$$

then the acoustic component of the language corpus can be represented by tree-like architecture, where the "tree"-the phrase is formed by "branches"-lexemes, each of which is characterized by "leaves"- variations of its phonation, characterized according to expression (9). To ensure the computational efficiency of the target operation of such a tree in the calculation of expression (9) should ignore the unlikely variants of phonation of lexemes. This can be achieved by replacing in expression (9) the weighted sum of the plausibility of the phonation models of lexemes by the corresponding value of the maximum plausibility:

$$W^* = \arg\max_{W,t^W} P\left(t^W\right)P\left(t^W \middle| X\right). \qquad `$$

It is the term $P\left(t^W \middle| X\right)$ in expression (10) that takes into account the unlikely variants of phonation of language units.

3. Results

Applied use of the decision-making models presented in the previous section will be based on the material of the acceptable language corpus GRAC. It will be recalled that the basic element that determines the possibility of using the proposed mathematical apparatus is the availability of data on the frequency of presence of certain language units in the acoustic vocabulary of the language corpus. We choose numerals as the focus set of lexemes for the study. The precondition for such a choice is the limited number and clearly defined structure of such lexemes, their prevalence in the language material of any style. However, this choice has its drawbacks. In particular, the pronunciation duration of most numeral lexemes is short, in the phrases of these language units is characterized by a high semantic load, so their pronunciation is treated with extreme care, which reduces the variability of phonation. The training sample included language material from 250 speakers (over $4 \cdot 10^4$ sentences). The focus group included 28 unique lexemes and their combinations: voiced numbers from “one” to “hundred”.

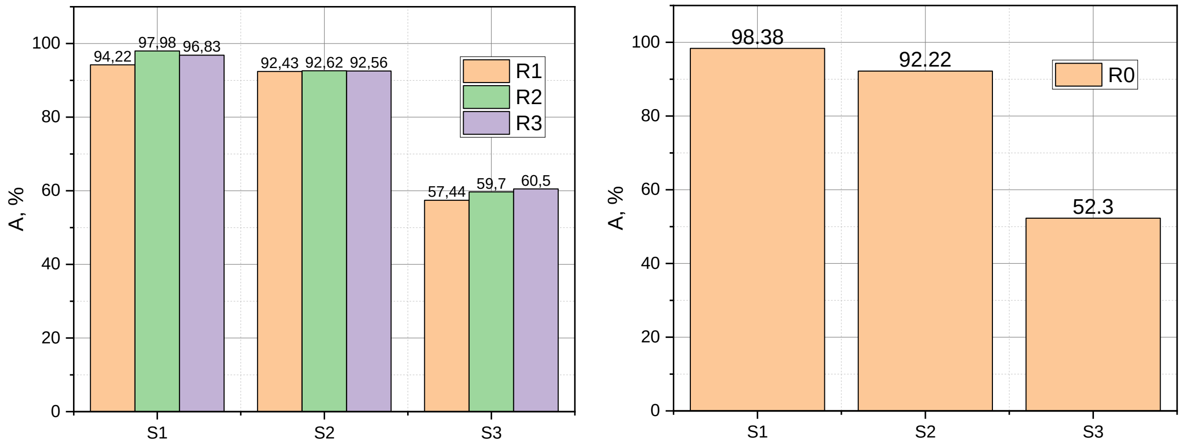
Considering that the results proposed in section 2 are aimed at improving the classification process in the task of recognition of language units in speech, as well as to ensure the reproducibility of experimental results, direct lost implementation of theoretical results was conducted in *Simon* (https://simon.kde.org/). It is an open-source speech recognition software. The software environment provides the ability to customize the classification process. It is possible to connect acoustic and generative models from such well-known specialized projects as KDE, CMU SPHINX, Julius, HTK [22, 23]. There is an interface for connecting language corpora based on dialects (sound schemes).

We experimented to recognize the mentioned alphabet of lexemes on a series of test samples, the content of which does not intersect with the material of the training sample. Since the variability of phonation affects the probability of both errors of the first and second kind, to assess the results of recognition chosen the basic for the classification task characteristic – the accuracy A.

During the experiment, the classifier of the recognition system consistently functioned in four modes: $\{R_0, R_1, R_2, R_3\}$, where: - in R_0 the phenomenon of phonation variability in the decision rule is not specifically taken into account (criterion (1)); - in R_1 the phenomenon of phonation variability in the decision rule is taken into account (criterion (4)); - in R_2 the phenomenon of phonation variability in the decision rule is taken into account (criterion (8)), - in R_3 the phenomenon of phonation variability in the decision rule is taken into account (criterion (10)). The phenomenon of phonation variability for the content of an arbitrary test sample was determined by the value of the parameter $V=\sum_{i=1}^{N} t_i / N$, where t_i is the number of phonation variants (sound schemes) for the i-th lexeme in the vocabulary, N is the number of the lexeme in the vocabulary. For the experiment, $N=28$ is the focus group of unique lexemes-numerals. For a random test sample we have: $V=\left\{V_0 \forall R_0; V_{\bar{0}} \forall R_1, R_2, R_3\right\}$.

Three test samples $S=\left\{S_1, S_2, S_3\right\}$ were formed for the experiment. The test sample S_1 included 800 phrases sounded by one speaker-man (400) and one speaker-woman (400). The test sample S_2 included the material of the test sample S_1, supplemented by 800 phrases sounded by members of a gender-symmetrical team of 10 speakers. The test sample S_3 included the material of the test sample S_1, supplemented by 800 phrases sounded by members of 50 speakers team balanced on gender, age group (1: 16-20 years; 2: 25-40 years; 3: 45-60 p.) and dialect. The sounded language material of the sample S_1 was included in the test samples S_2 and S_3 for normalization. In all test samples from the set S, each phrase included from 2 to 20 lexemes, at least one of which was a numerator from the focus group. The variability of the test samples $S=\left\{S_1, S_2, S_3\right\}$ was characterized by such values of the parameter V as $V_{\bar{0}}=\left\{1{,}3: S_1, R \in\left\{R_1 \div R_3\right\}; 1{,}9: S_2, R \in\left\{R_1 \div R_3\right\}; 3{,}1: S_3, R \in\left\{R_1 \div R_3;\right\}\right\}$ and $V_0=\left\{1{,}0: S \in\left\{S_1 \div S_3\right\}, R=R_0\right\}$ (phonation variability is not taken into account in a decision rule R_0).

The results of the experiments $A=f\left(S_1 \div S_3; R_1 \div R_3; V_{\bar{0}}\right)$ and $A=f\left(S_1 \div S_3; R_0; V_0\right)$ are presented in the diagrams in Fig. 2a and 2b, respectively.



a) b)

Figure 2: The results of the experiment: a) $A=f\left(S_1 \div S_3; R_1 \div R_3; V_{\bar{0}}\right)$ b) $A=f\left(S_1 \div S_3; R_0; V_0\right)$

4. Discussion

Before proceeding to the direct analysis of the experimental results, let us recall that with the growth of the test sample index, the degree of phonation variability in the sounded language material also increased. Note that the total duration of a sounded language material in test samples from the set S does not exceed 15% of the total duration of a sounded language material in the training sample, which is sufficient for effective use of not only simple classifiers (k-Means, Support Vector Machines) [22] but complex classifiers (Bayesian Classification, Monte Carlo Classification, Neural Network Classification, etc. [24]).

The generalization of the classification results was carried out by one of the $R=\{R_0 \div R_3\}$ decision rules proposed in section 2, where: - in the classical decision rule R_0 the etalon sound schemes were not taken into account; - the decision rule R_1 took into account the etalon sound schemes, but their frequency (expression (6)) was ignored; - the decision rule R_2 took into account the optimal etalon sound scheme, which was determined according to expression (7); - the decision rule R_3 took into account the suboptimal etalon sound scheme, which was determined according to expression (9). As the expected accuracy of the classification increases, these decision rules can be arranged as follows: R_0, R_1, R_3, R_2. As the computational complexity of the classification process increases, these decision rules can be arranged as follows: R_0, R_1, R_3, R_2. Let's analyze whether the results of the experiments confirmed these expectations.

First, pay attention to the results presented in Fig. 3. The adequacy of these results is convincingly proved by the realities of the modern cybersphere, in which speech recognition systems are confidently used in personalized software environments (operating systems of smartphones, laptops, personal computers) (one speaker), but not for, for example, automated stenography, concerts, etc., (many speakers, disturbing factors). As the amount of sounded language material from different speakers increases (sequential transition from the test sample S_1 to S_3), the recognition accuracy decreases from a high 98% to an unacceptable 52%. The R_0 decision-making rule was used in this study, i.e. the possibility of adapting the classifier to the specifics of phonation was solely due to its cognitive properties at the training stage. Moreover, the variability of phonation was perceived as an additional source of disturbances (noise). The demonstrated results convincingly prove the relevance of the study of the influence of phonation variability on the result of the process of recognition of sounded language units.

Unfortunately, from the shown in Fig. 2 results, it can be seen that the implementation of the theoretical approaches proposed in section 2, embodied in the solution rules R_1, R_2, R_3, did not overcome the tendency to decrease the accuracy of recognition with increasing variability of phonation in the test language material. Moreover, the results demonstrated in test samples with low and moderate phonation variability (S_1 and S_2) showed that the use of authorial mechanisms $R_1 \div R_3$ to compensate for the effect of phonation variability at the level of the decision rule led to a slight decrease in accuracy of sounded focus group lexemes recognition (in comparison with the results presented in Fig. 3). A potential reason for this may be the redundancy of the factor space of the acoustic model, which leads to "blurring" the boundaries of clusters of language units. At the same time, the classification results demonstrated in the test sample with high phonation variability (S_3) prove the functionality of the author's mechanisms to compensate for the influence of phonation variability at the level of the decision rule and increase recognition accuracy by 5-8% (from the original 52% to 57-60 %, respectively).

Note also that for all test samples from the set S, the decision rules R_2 and R_3, which took into account the optimal and suboptimal etalon sound schemes, respectively, exceeded the decision rule R_1, which took into account the etalon sound schemes, but their frequency was ignored. The comparison of the solution rules R_2 and R_3 shows in favour of the latter, because, with close recognition accuracy, the amount of computational resources spent on classification according to rule R_3 is 20-30% less than the same as for rule R_2.

5. Conclusions

Experience with the use of speech signal recognition services in modern personal mobile and desktop operating systems shows that this task is currently being solved with acceptable accuracy. At the same time, the limited use of such services in the corporate and government segments of cyberspace unequivocally prove that the task of recognizing the speech signals of more than one speaker in non-laboratory conditions is still relevant.

The article presents the technology of improving the process of recognition of language units by integrating the model of the variability of their phonation in the decision rule. In the proposed technology, in contrast to existing ones, recognition occurs at the level of comparison of sound schemes of empirical and etalon language material in the common parametric space of acoustic, generative and language models. This allowed us to formalize the concepts of taking into account the influence of phonation variability in determining the etalon sound schemes of language units in the paradigm of pattern recognition theory and to formulate a UML activity diagram of the mechanism for calculating the parameters of these concepts.

The classification results demonstrated in the test sample with high variability of speech material prove the functionality of the author's mechanisms to compensate for the influence of phonation variability at the level of the decision rule and increase the accuracy of recognition by 5-8% (from the original 52% to 57-60%, respectively). Experiments have shown that for all test samples, the decision-making rules formulated based on the author's concept, which took into account the optimal and suboptimal etalon sound schemes, respectively, exceeded the solving rule, which took into account the etalon sound schemes, but their frequency was ignored. It turned out that it is not advisable to use the author's mechanisms to compensate for the influence of phonation variability in the classification of speech material with a low or moderate degree of variability.

Further research is planned to focus on finding methods for optimizing the factor space of acoustic, generative and speech models with an active mechanism to compensate for phonation variability.

6. Acknowledgements

The authors would like to thank the Armed Forces of Ukraine for providing security to perform this work. This work has become possible only because of the resilience and courage of the Ukrainian Army.

7. References

[1] J. H. L. Hansen and H. Bořil, On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks, Speech Communication 101 (2018) 94–108. doi: 10.1016/j.specom.2018.05.004.

[2] O. Scharenborg and M. van Os, Why listening in background noise is harder in a non-native language than in a native language: A review, Speech Communication 108 (2019) 53–64. doi: 10.1016/j.specom.2019.03.001.

[3] B. D. Sarma, S. R. M. Prasanna, and P. Sarmah, Consonant-vowel unit recognition using dominant aperiodic and transition region detection, Speech Communication 92 (2017) 77–89. doi: 10.1016/j.specom.2017.06.003.

[4] W. Liu, et al., Improved Phonotactic Language Recognition Using Collaborated Language Model, 5th International Conference on Cloud Computing and Intelligence Systems (CCIS) (2018) 747–751. doi: 10.1109/CCIS.2018.8691262.

[5] D. Liu, X. Wan, J. Xu and P. Zhang, Multilingual Speech Recognition Training and Adaptation with Language-Specific Gate Units, 11th International Symposium on Chinese Spoken Language Processing (ISCSLP) (2018) 86–90. doi: 10.1109/ISCSLP.2018.8706584.

[6] R. Asnawi and M. Said, Testing of other languages usage in addition to the default languages for the easy voice recognition module, International Conference on Electronics Technology (ICET) (2018) 321–324. doi: 10.1109/ELTECH.2018.8401476.

[7] S. A. E. El-Din and M. A. A. El-Ghany, Sign Language Interpreter System: An alternative system for machine learning, 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES) (2020) 332–337. doi: 10.1109/NILES50944.2020.9257958.

[8] N. Sae Jong and P. Phukpattaranont, P. A speech recognition system based on electromyography for the rehabilitation of dysarthric patients: A Thai syllable study. Biocybernetics and Biomedical Engineering 39 1 (2018) 234–245). Doi: 10.1016/j.bbe.2018.11.010.

[9] W. Xu, et al., Fully automated detection of formal thought disorder with Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS), Journal of Biomedical Informatics 126 (2022) 103998). doi: 10.1016/j.jbi.2022.103998.

[10] V. Montani, V. Chanoine, J. Grainger and J. C. Ziegler, Frequency-tagged visual evoked responses track syllable effects in visual word recognition. Cortex 121 (2019) 60–77. doi:10.1016/j.cortex.2019.08.014.

[11] V.V. Kovtun, et al., Precision automated phonetic analysis of speech signals for information technology of text-dependent authentication of a person by voice, 2nd International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS 2021) 2853 (2021) 376–388. urn:nbn:de:0074-2853-7.

[12] K. Obelovska, O. Panova, and V. Karovič Jr., Performance Analysis of Wireless Local Area Network for a High-/Low-Priority Traffic Ratio at Different Numbers of Access Categories, 13 4 Symmetry (2021) 693. doi: 10.3390/sym13040693.

[13] O. Tymchenko, O. O. Tymchenko, B. Havrysh, O. Khamula, O. Sosnovska, and S. Vasiuta, Efficient Calculation Methods of Subtraction Signals Convolution, 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM) (2019). doi: 10.1109/cadsm.2019.8779250.

[14] I. Dronyuk, O. Fedevych, and B. Demyda, Signals and Images Protection Based on Ateb-Transforms in Infocommunication Systems, International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (2018). doi: 10.1109/infocommst.2018.8632043.

[15] O. Bisikalo, V. Kovtun, O. Boivan, and O. Kovtun, Method of Automated Transcribing of Speech Signals for Information Technology of Text-Dependent Authentication of a Person by Voice, 11th International Conference on Advanced Computer Information Technologies (ACIT) (2021). doi: 10.1109/acit52158.2021.9548627

[16] Dedry, M., Maryn, Y., Szmalec, A., Lith-Bijl, J. van, Dricot, L., & Desuter, G. (2022). Neural Correlates of Healthy Sustained Vowel Phonation Tasks: A Systematic Review and Meta-Analysis of Neuroimaging Studies. In Journal of Voice. Elsevier BV. https://doi.org/10.1016/j.jvoice.2022.02.008.

[17] Y. Zhao and L. Zhu, Speaker-Dependent Isolated-Word Speech Recognition System Based on Vector Quantization, 2017 International Conference on Computer Network, Electronic and Automation (ICCNEA) (2017). doi: 10.1109/ICCNEA.2017.103.

[18] H. H. O. Nasereddin and A. A. R. Omari, Classification techniques for automatic speech recognition (ASR) algorithms used with real time speech translation," 2017 Computing Conference, (2017). doi: 10.1109/SAI.2017.8252104.

[19] P. Vanajakshi and M. Mathivanan, A detailed survey on large vocabulary continuous speech recognition techniques, 2017 International Conference on Computer Communication and Informatics (ICCCI) (2017). doi: 10.1109/ICCCI.2017.8117755.

[20] L. Li, D. Wang, Y. Chen, Y. Shi, Z. Tang and T. F. Zheng, Deep Factorization for Speech Signal, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2018). doi: 10.1109/ICASSP.2018.8462169.

[21] M. Raczynski, Speech processing algorithm for isolated words recognition, 2018 International Interdisciplinary PhD Workshop (IIPhDW), (2018). doi: 10.1109/IIPHDW.2018.8388238.

[22] T. Dinushika, L. Kavmini, P. Abeyawardhana, U. Thayasivam and S. Jayasena, Speech Command Classification System for Sinhala Language based on Automatic Speech Recognition, 2019 International Conference on Asian Language Processing (IALP), (2019). doi: 10.1109/IALP48816.2019.9037648.

[23] R. Fu, J. Tao, Z. Wen and Y. Zheng, Phoneme Dependent Speaker Embedding and Model Factorization for Multi-speaker Speech Synthesis and Adaptation, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2019). doi: 10.1109/ICASSP.2019.8682535.

[24] B. Huang, Phonetic Feature Extraction and Recognition Model in Japanese Pronunciation Practice, 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), (2021). doi: 10.1109/ICOEI51242.2021.9452933.