Voice command recognition for noisy environments by means of cross-correlation portraits

A.I. Armer[1], E.Yu. Galitskaya[2], N.A. Krasheninnikova[3]

[1] Ulyanovsk State Technical University, Severny Venets St., 32, Ulyanovsk, 432027, Russia
[2] Ulyanovsk Instrument Manufacturing Design Bureau, Krymov St., 10a, Ulyanovsk, 432071, Russia
[3] Ulyanovsk State University, Lev Tolstoy St., 42, Ulyanovsk, 432017, Russia

Abstract

Methods of voice command (VC) recognition in heavy noise environments are required for precise work of speech information systems on the factory floor and in transport. The paper considers a speaker-dependent way of VC recognition for VCs belonging to a limited vocabulary and being recognized in heavy noise environments. For this purpose, VCs are transformed into cross-correlation portraits (CCPs), i.e. special images. The VC under recognition is referred to a class with a minimal distance (metric) between CCP of this command and model CCPs of the class. The authors elaborated algorithms for VC transformation into CCPs, a method for defining VC boundaries, ways of model command optimization and metric choice. As a result, a rather precise VC recognition in heavy noise environment was obtained.

Keywords: voice command; intensive noise; recognition; cross-correlation portrait; metric; precise definition of boundaries; model command; optimization of VC library

1. Introduction

The growth of production and transport intensity leads to increase in operator burden. To reduce such workload, speech information systems are used. However, these systems often have to recognize VC precisely, especially for noisy environments. At present, a large number of speech recognition systems functioning in nearly noiseless environment have been developed. They include, for example, IBM Via Voice, its recognition accuracy is reported to be 97% and its recognition vocabulary includes up to 2,000 VCs; Dragon NaturallySpeaking or Dragon for PC, this software package accurately recognizes 70% of the vocabulary, which includes nearly 60,000 words; L&H Voice XPress, its accuracy is in the range of 90%-98% and its vocabulary size is nearly 1,000 words, etc. There are also user-friendly systems of continuous speech understanding and processing, such as VocalIQ, Siri, Google Now and Cortana. To compare VocalIQ with Siri, Google Now and Cortana the systems were given multiaspect requests in a natural language [1]. The correct recognition rate was more than 90% for VocalIQ, while Google Now, Siri and Cortana showed only 20% accuracy. Among home-grown technologies it is necessary to mention VoiceCom STC. It is reported to recognize 100-200 VCs in a speaker-dependent version and 30-50 VCs in a speaker independent one with accuracy 98%. However, these systems do not accurately work even in low loise environment. Recognition systems for VCs from a limited vocabulary under acoustic noise are currently being developed mainly for aviation and are used in voice control and flight control devices. Performance quality of such systems today is from 90 up tp 98% of accurate VC recognition, depending on the test conditions and vocabulary size. Almost all tested systems are speaker-dependent. According to the Air Force Research Laboratory - Wright-Patterson Air Force Base, flight tests of an ITT VRS-1290 speaker dependent, continuous speech recognition system and a Verbex VAT31 showed the following results: average word accuracy for VRS-1290 was 92-98%, if the vocabulary consisted of 50 commands; average word accuracy for VAT31 was up to 97% (no information on vocabulary size is available). In 1997, flight test results of the VC recognition system produced by National Research Council (Canada) were obtained. The system was integrated into Bell 412HP Avionics Management System and showed an average 95% accuracy for vocabulary consisting of 80 words, which were divided into 24 groups. According to the Smiths Industries Speech Recognition Module system built into the CAMU of the Eurofighter, the accuracy of VC recognition in a standard aircraft flight is at least 95% for a vocabulary consisting of 250 words, 25 of which can be simultaneously active. Currently, Thales Avionics develops a VC recognition system for Rafale fighters. The VC recognition accuracy is required to be above 95% for a vocabulary of 50-300 words. A 5-th generation jet fighter F-35 was equipped with DynaSpeak

VC recognition system developed by SRT International. The developers report the recognition accuracy to be 98%. A multipurpose 4-th generation Eurofighter is equipped with a voice control system developed by Logica. The vocabulary consists of 250 words, and the average VC accuracy is not less than 95%. The developers declare, that for the export version of the Rafale Block 05t, Thales Avionics has developed a speech control system with recognition accuracy not less than 95% for a 300 VC vocabulary, but no information on its implementation is available. Patent US 6529866 B1, 4 March 2003, The United States of America as Represented by the Secretary of the Navy, describes a method and system for transformation of an audio signal into speech. Audio signals are said to contain both VC units and noise, but test and implementation information is not available. Patent WO 1999040571 A1, 3 February 1999, Qualcomm Incorporated, describing a system and method for improving speech recognition accuracy in noisy environment also provides no test or implementation data. Among home-grown technologies the following ones should be noted. First of all, it is a VC recognition system tested on the Mikoyan MiG-29 (Fulcrum). Recognition accuracy is reported to be 56-81%, no information on the vocabulary is available. Patent RF 2267820 1, 25 April 2006, Ulyanovsk State Technical University. Recognition accuracy is reported to be 92%, vocabulary size is 23 VCs, and noise level is 3dB. No information on implementation is available. Patent RF 2271578 2, 10 March 2006, Speech Technology Center. The invention relates to speech analysis under adverse environmental conditions, e.g. in moving transport or high level noisy workplaces. No test information is available. Despite the available developments, there is no information on the actual application of VC recognition systems in avionics, since in real flights the systems developed showed substantially less efficiency than anticipated. Thus, developing VC recognition systems for noisy environments remains a challenging task. This paper examines a speaker dependent technique of VC recognition for a limited vocabulary. A method of VC transformation into portraits, i.e. images, is used.

2. Methods of VC recognition

The problems of speech recognition, in particular VC recognition, are widely discussed in modern literature. The first methods of automatic sound recognition were obtained in the first half of the 20-th century [2]. Among speech recognition techniques one can distinguish the following approaches: spectral methods [3, 4, 5, 6, 7], wavelet transform [3], statistical methods [5, 8, 9, 10, 11], and neural networks [12, 13].

This paper deals with VC recognition based on their transformation into portraits, i.e. flat images, and further implementation of image processing techniques [14, 15, 16, 17, 18].

3. Autocorrelation portraits

Let $S = s_0, s_1, s_2, s_3, ..., s_{N-1}$ be digital VC readouts. Then, a two-dimensional image $X(i,k) = \{x_{ik} : i = 1,2,3,...; k = 1..K\}$ will be its autocorrelation portrait (ACP). This image is obtained in the following way. Let us divide VC S into M segments and perform the following transformations

$$X(i,k) = \frac{Cov(S_n, S_{n+k})}{\sigma_n \sigma_{n+k}}, \tag{1}$$

where $Cov(S_n, S_{n+k})$ is a sample covariation of signal S intervals S_n, S_{n+k}, which are spaced $k\Delta t$ apart, $\sigma_n^2, \sigma_{n+k}^2$ are sample dispersions of segments S_n, S_{n+k} respectively. Thus, the $k-$th element of the $i-$th ACP line is equal to the correlation coefficient between the $i-$th segment S_i and the segment shifted left with respect to S_i on k readouts. Fig. 1 shows ACP examples.

Let us note some ACP characteristics, which make them favorable for VC recognition. VC portraits are unique, i.e. ACPs of different VCs are unlike, whereas ACPs of the same VCs pronounced at different time intervals are the same. Autocorrelation transformation normalizes a signal, as a result ACPs are nearly insensitive to noisiness and slowly varying additives. If we consider additive white noise with dispersion σ_θ^2, then its ACPs and VC ACP readouts distorted by noise will differ by a constant factor. However, ACPs also have some negative characteristics, e.g. the dependence of element brightness on the differences in the tone of VC pronunciation, as well as geometric ACP distortions due to variations in speech rate. These distortions can be steadied by modifying ACP development, e.g. taking into account loudness extremum. VC recognition by their ACPs is conducted in the following way. ACPs of model VCs are stored in the memory. VC under recognition is transformed into ACP and it is referred to the class with a minimal distance between its model portrait and ACP of a recognized VC. This distance (metric) between two ACPs (i.e. images) is calculated as follows. At first, two images are aligned, i.e. for each line of one image a corresponding line of another image is found. The average distance (e.g. Euclidean) between the corresponding lines is considered to be the distance between the portraits. Such a correspondence for ACP of one and the same command means the proximity of VC fragments, so the distance is relatively small, since it only occurs from the difference in pronunciation and surrounding noise. If ACPs of different VCs are compared, then this distance is usually much more visible due to the larger difference in sounds. In the process of command alignment dynamic programming based on minimum distance criterion was applied. While testing the accuracy of VC recognition, commands were pronounced by the speaker in real time. The vocabulary used consisted of ten VC groups, and there were 4-23 aviation commands in each group. In total, the vocabulary included more than 100 VCs. Aircraft engine noise recorded in a flight mode was used as a background and reference noise, the signal-to-noise ratio was 5-0 dB. Four male speakers took part in the tests. Before the experiment each speaker recorded model VCs, each VC belonging to the given vocabulary was pronounced twice. During the experiment on VC recognition each speaker pronounced all the commands from the given vocabulary three times, all in all, more than 1,200 VCs were recorded during the experiment. Average command accuracy was more than 95%. However, further processing has shown that the probability of accurate VC recognition can be significantly reduced in the course of time. This problem is connected with model aging, i.e. speaker's voice pattern can change with time, and previously pronounced command models will not reflect the peculiarities of the speaker's voice at the very time of VC recognition. Therefore, it is required to update the commands from time to time (e.g. before the flight), which, of course, has certain inconveniences. One VC model does not reflect all the possible variants of its pronunciation, so the number of VCs was increased, i.e. the speaker pronounced each VC more than once at different periods of time. The totality of all these patterns somehow reflected pronunciation diversity. However, the increase in model number complicates and slows down the recognition algorithm, but it is permissible only to a certain extent. Therefore, the model number should be limited. Besides, these models should reflect the pronunciation diversity as much as possible. It turns out, that recognition accuracy depends greatly on the correctness of model choice, and recognition deviations can be more than 10%. Thus, among several pronunciations it is necessary to choose a certain number of VCs as model ones, so that the obtained model library contributed to the best VC recognition accuracy. This problem of model library optimization was examined in [19, 20, 21]. Technically it is impossible to conduct complete enumeration of all library patterns. That is why, a method of direct enumeration giving an almost optimal result has been developed. Sometimes it is possible to change the VCs themselves, using their synonyms. This problem was also considered and its solution was found while analyzing the synonym rings.

4. Cross-correlation portraits

Another way to decrease the impact of VC pronunciation variability is to use a different kind of portraits instead of ACPs. In the process of ACP development correlation coefficients between the segments of the same VC (autocorrelation) are found. When ACPs are used for recognition, the distances between the ACP of a recognized command and the ACP of a model command are found. If the distance between the ACP of the command under recognition and the ACP of its model is found, the ACPs of two different pronunciations of this command will be compared. These ACPs can significantly differ from each other (the distance will be large). Therefore, when comparing portraits it is desirable to minimize the difference in pronunciation. For this purpose, it is necessary for pronunciation variability to be somehow reflected in portraits. Let us consider a cross-correlation portrait (CCP), which consists of correlation coefficients between segments of two VCs (cross-correlation) [15, 16, 17, 22]. Let there be two VCs $S1$ and $S2$. Let us segment each command into M segments of the same length and determine the sample correlation coefficients x_{ik} between the $i-$th segment of VC $S1$ and a VC segment $S2$, beginning with the $k-$th readout of the VC $S2$ $i-$th segment. As a result, we get a two-dimensional array (image) $X = \{x_{ik}\}$, called a CCP of VCs $S1$ and $S2$. Let us consider CCP development in detail. As an example, let us consider the CCP development of two pronunciations of one avionics VC, the first pronunciation is $S1$ and the second pronunciation is $S2$. Let us divide each VC into equal segments, whereas $N1$ is the length of each interval for signal $S1$, and $N2$ is the length of each interval for signal $S2$. Let $N = min\{N1, N2\}$ be the minimal of these lengths. While specifying the number of intervals for each command M it should be taken into account that if the segment length is too small it will not include the whole phoneme; otherwise, if the segment length is big enough it will include several phonemes. Such segmentation will negatively affect the correlation coefficient between separate phonemes in different VCs while developing CCPs. Let's determine correlation coefficients of signal $S1$ i-th segment and signal $S2$ i-th segment, shifted $k = 0..K$ readouts right.

$$x_{ik} = \frac{\frac{1}{N}\sum_{j=0}^{N-1} S1_{i\cdot N1+j} S2_{i\cdot N2+j+k} - \mu1_i\mu2_{i,k}}{\sigma1_i\sigma2_{i,k}}, \tag{2}$$

$$\mu1_i = \frac{1}{N}\sum_{j=0}^{N-1} S1_{i\cdot N1+j}, \tag{3}$$

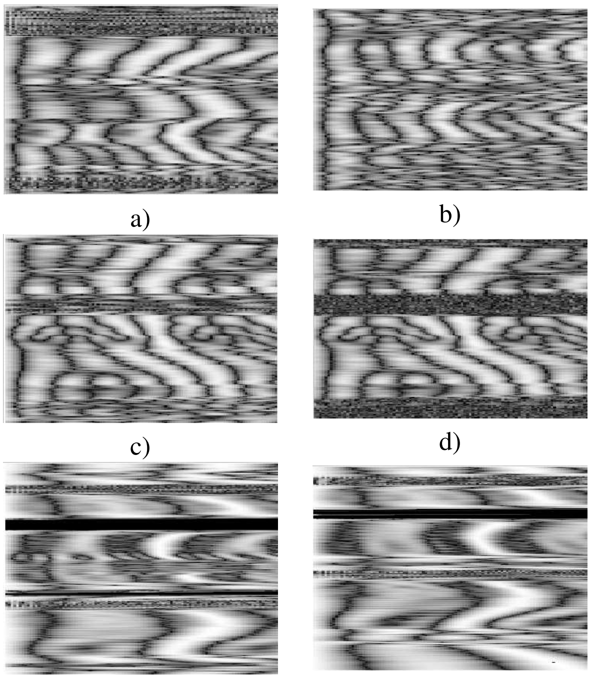
$$\mu2_{i,k} = \frac{1}{N}\sum_{j=0}^{N-1} S2_{i\cdot N2+j+k}, \tag{4}$$

$$\sigma1_i^2 = \frac{1}{N}\sum_{j=0}^{N-1} S1_{i\cdot N1+j}^2 - \mu1_i^2, \tag{5}$$

$$\sigma2_{i,k}^2 = \frac{1}{N}\sum_{j=0}^{N-1} S2_{i\cdot N2+j+k}^2 - \mu2_{i,k}^2. \tag{6}$$

While choosing parameter K, it is necessary to take into account the following fact: if its value increases, than value x_{ik} decreases. It is connected with correlation reduction of VC readouts along the line. This property proves the inadvisability of using large values K while developing CCPs (large K means that $K > N$).

Obviously, if CCPs of the same pronunciation ($S1 = S2 = S$) are developed, we get the ACP of a VC S. It is desirable to examine the CCP of two pronunciations of one and the same command. It depends on two pronunciations, so the pronunciation variability affects the portrait form. Fig. 2 shows CCPs of several VCs. For example, in the picture Manevr3 + Manevr4 'plus' means that this very CCP was obtained from the third and fourth pronunciations of the VC *"Manevr"*.



a)

b)

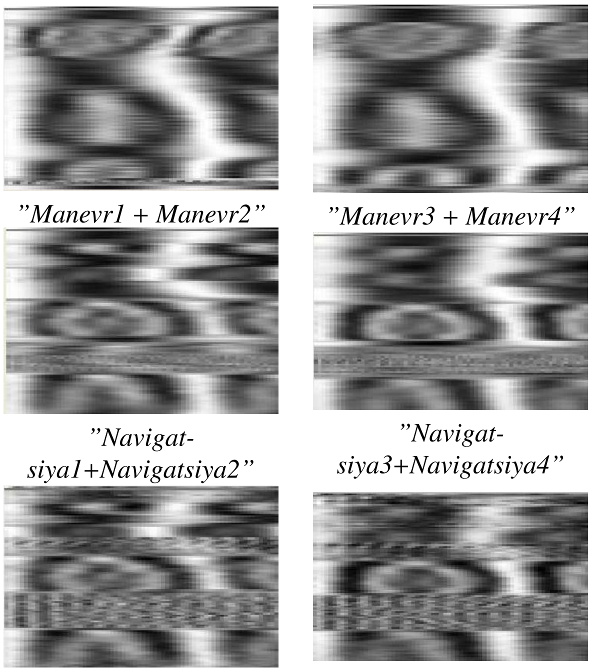
c)

d)

e)

f)

Figure 1: Autocorrelation portraits: a) VC *"Svet bol'she"*, b) VC *"Svet bol'she"* on the background of aircraft engine noise, c) VC *"Konditsioner"*, d) VC *"Konditsioner"* on the background of white noise with a mean zero and dispersion equal to five, e), f) VC *"Vysota absolyutnaya"* pronounced at different times.



"Manevr1 + Manevr2"

"Manevr3 + Manevr4"

"Navigatsiya1+Navigatsiya2"

"Navigatsiya3+Navigatsiya4"

"Noised Navigatsiya1+Navigatsiya2"

"Noised Navigatsiya3+Navigatsiya4"

Figure 2: Cross-correlation portraits of VCs.

Note, that CCP characteristics are similar to those of ACP. But CCPs are less pronunciation dependent, as they combine two different pronunciations. VC recognition by means of CCPs is carried out in the same way as recognition by means of ACPs. For each VC, a model CCP made of two pronunciations of this VC is developed. These model CCPs are stored in the memory. For the VC under recognition CCPs are developed with one of pronunciations of each command group, then the distance between this CCP and the model CCP is found. The recognized VC is related to the group with the smallest distance.

5. Optimization of voice command recognition by means of their CCPs

The CCPs used have a number of characteristics, which come from both the properties of the speech signals themselves and the structure of their CCP development. Let us consider some techniques increasing the recognition accuracy by means of CCPs.

5.1. Noisy models

If a VC under recognition is too noisy, it increases the distance from its CCP to its model portrait formed by means of noiseless pronunciations. Therefore, 'noisy models' were used in the experiment, i.e. artificial noise was added to the model commands. It came from the microphone placed far from the operator. As a result, the distorted models and the command under recognition contained approximately the same noise, which significantly increased the recognition accuracy.

5.2. Precise definition of boundaries

While developing portraits, it is desirable for the VC time boundaries to be defined as precisely as possible. Then a more accurate portrait alignment can be attained. Among several known techniques of useful signal detection, the one, which shows the most accurate recognition results on the background of noise, was chosen. Besides, after definition of VC boundaries by means of this technique some boundary adjustments were made, which resulted in recognition accuracy.

5.3. Pause removal

In some VCs, e.g. those consisting of two words, there are micro-pauses between speech units. These pauses can differ in duration, but they do not contain any information. So, a special method for their removal was developed.

5.4. Optimization of portrait width

Portrait width, i.e. the line length, can be chosen arbitrary. So, it is desirable to choose the optimal length, which contributes to the best recognition accuracy. It turned out, the line length in the portrait of a VC under consideration should be equal to $K = D/(5M)$, where D is the length of the recognized VC, M is the number of lines in a portrait. The line lengths of model CCPs are a bit longer, but they are no less than K.

5.5. Choice of metric

VC recognition by means of their CCP is based on detection of the portraits, which are as similar as possible. Hence, there appears a problem to define the distance between two CCPs, i.e. metric defined on CCP. This distance is considered to be equal to the average distance between the corresponding CCP lines. Moreover, any metric defined on the lines, i.e. on finite sequences or vectors, can be used. Twelve known metrics (namely, Euclidean, Hilbert, Zhuravlev, etc.) and their variants were tested. For the purpose of the problem under consideration, five metrics showed the best results: Zhuravlev method (for $\varepsilon = 10$, $\varepsilon = 20$ and $\varepsilon = 30$), the Ruzicka distance and the Bray-Curtis distance. Besides, analyzing the recognition results obtained while using these metrics it was found out that certain recognition errors corresponded to certain metrics. Therefore, it is possible to improve recognition accuracy by using, for example, two metrics. If the recognition results coincide, then the command is considered to be recognized; if the recognition results differ, the command should be considered unrecognized. In such a case, the speaker should pronounce the command once again.

5.6. Optimization of a model library

As in the case of VC recognition by means of CCPs, the words included in the model library significantly affect the recognition accuracy. Therefore, it is required to optimize the model portrait library while recognizing VCs by means of CCPs.

5.7. Fourier analysis

Each CCP line is a sequence of correlation function sample values. Because of speech signal quasi-periodicity, the correlation function turns out to be similar to periodic. This quality was used to improve the portrait quality by removing insignificant harmonics from each CCP line spectrum. This operation was performed by means of FFT. The isolation of the most fundamental harmonics for each CCP line reduced the influence of speech signal pronunciation variability.

6. Results

The experiments showed that using CCPs with the described above modifications significantly reduced the effect of VC pronunciation variability and model aging. The recognition accuracy was nearly the same as in the ACP recognition on newly-pronounced models. Thus, to evaluate the efficiency of the suggested method, an experiment was conducted. The recognition was tested on two groups of VCs consisting of 10 commands each. The first group included single-word commands, the second group of VCs included both single-word and two-word commands. Each VC was pronounced 100 times by a woman-speaker. The experimental results are represented in Table 1. The maximum VC recognition accuracy was 95.6%.

Table 1: VC recognition accuracy by means of CCPs.

Commands	Signal/noise ratio (dB)				
	5	4	3	2	1
Group 1	95.6	90.1	87.4	82.9	61.2
Group 2	94.6	92.1	87.4	83.5	67.2

7. Conclusion

The present work suggests and examines a speaker-dependent method for recognizing voice commands from a limited vocabulary in conditions of intense acoustic noise, e.g. on the background of an aircraft engine. This method implies transformation of digitized commands into certain images and further application of image processing methods. The method underwent various modifications in order to increase the recognition accuracy. Tests on a large number of voice commands showed rather high efficiency of the suggested method.

References

[1] Businessinsider. How apples vocaliq ai works [Electronic resource]. ”— 2017. ”— URL: http://uk.businessinsider.com/how-apples-vocaliq-ai-works-2016-5.

[2] Rabiner, L. Tsifrovaya obrabotka rechevykh signalov [Digital processing of speech signals]: translated from English. Edited by M.V. Nazarov and Yu.N. Prokhorov / L.R. Rabiner, R.V. Shafer. ”—Moscow, Russia. : Nauka, 1981. ”—P. 495. (in Russian)

[3] Boykov, F. Primenenie veyvlet-analiza signala v sisteme raspoznavaniya rechi [Wavelet analysis in speech recognition] / F.G. Boykov, Starozhilova T.K. // Trudy mezhdunarodnoy konferentsii Dialog 2003 [Proceedings of the international conference Dialogue 2003]. ”—Zvenigorod, Russia. ”—2003. ”—Pp. 12–19. (in Russian)

[4] Gudonavichyus, R. Raspoznavanie rechevykh signalov po ikh strukturnym svoystvam [Speech signal recognition by means of their structural characteristics] / R.V. Gudonavichyus, P.P. Kemeshis, A.B. Chitavichyus. ”—Leningrad, USSR. : Energiya, 1977. ”—P. 64. (in Russian)

[5] Myasnikova, E. Ob”ektivnoe raspoznavanie zvukov rechi [Objective recognition of speech sounds] / E.N. Myasnikova. ”— Leningrad, USSR. : Energiya, 1967. ”—P. 148. (in Russian)

[6] Pikone, D. Metody modelirovaniya signala v raspoznavanii rechi [Signal modeling methods in speech recognition] / D. Pikone. ”—Kemerovo, Russia, 2000. ”—P. 79. (in Russian)

[7] Potapova, R. Rech’: kommunikatsiya, informatsiya, kibernetika [Speech: communication, information, cybernetics] / R.K. Potapova. ”—Moscow, Russia.: Radio i svyaz’, 1997. ”—P. 568. (in Russian)

[8] Sorokin, V. Skrytye markovskie modeli v raspoznavanii rechi [Hidden markov models in speech recognition] / V.N. Sorokin, V.A. Sukhanov // Rechevaya informatika [Speech informatics]. Collected papers edited by V.V. Zyablov. ”— Moscow, Russia. ”—1989. ”— Pp. 104–118. (in Russian)

[9] Peinado, A. Discriminative codebook design using multiple vector quantization in hmm-based speech recognizers / A. Peinado, J. Segura, A. Rubio [et al.] // IEEE Trans. Speech and Audio Processing. ”— 1996. ”—Vol. IV, No. 2. ”—Pp. 89–94.

[10] Jelinek, F. Statistical Methods for Speech Recognition / F Jelinek. ”—Cambridge. : MIT Press, 1998.

[11] Shahshahani, B. A markov random field approach to bayesian speaker adaptation / B. Shahshahani // IEEE Trans. Speech and Audio Processing. ”— 1997. ”— Vol. V, No. 2. ”— Pp. 183–191.

[12] Fedyaev, O. Neyrosetevoy interpretator rechevykh komand dlya upravleniya programmnymi sistemami [Neural network interpreter of voice commands for program system processing] / O.I. Fedyaev, S.A. Gladunov // Proceedings of the 7th All-Russian conference ”Neural computers and their usage”, eduted by A.I. Galushkin. ”— Moscow, Russia. ”— 2001. ”— Pp. 298–301. (in Russian)

[13] Lippmann, R. Neural classifiers useful for speech recognition / R. Lippmann, B. Gold // in. Proc. IEEE First Int. Conf. Neural Net. ”— Vol. IV. ”— 1987. ”— Pp. 417–422.

[14] Krasheninnikov, V. Raspoznavanie rechevykh komand na fone intensivnykh shumov s pomoshch’yu avtokorrelyatsionnykh portretov [Speech command recognition on the background of noise using autocorrelation portraits] / V.R. Krasheninnikov, A.I. Armer, N.A. Krasheninnikova, A.V. Khvostov // Naukoemkie tekhnologii. ”— 2007. ”— 9. ”— Pp. 65–76. (in Russian)

[15] Krasheninnikov, V. Cross-correlation portraits of voice signals in the problem of recognizing voice commands according to patterns / V.R. Krasheninnikov, A.I. Armer, V.V. Kuznetsov, E.Yu Lebedeva // Pattern Recognition and Image Analysis. ”— 2011. ”— Vol. 21, No. 2. ”— Pp. 192–194. (in Russian)

[16] Krasheninnikov, V. Variatsiya granits rechevykh komand dlya uluchsheniya raspoznavaniya rechevykh komand po ikh krosskorrelyatsionnym portretam [Voice command variability for voice command recognition accuracy by means of their cross-correlation portraits] / V.R. Krasheninnikov, E.Yu. Lebedeva, V.K. Kapyrin // Izvestiya Samarskogo nauchnogo tsentra RAN. ”— 2013. ”— Vol. 4(4). ”— Pp. 928–930. (in Russian)

[17] Krasheninnikov, V. Povyshenie veroyatnosti pravil’nogo raspoznavaniya signalov po ikh krosskorrelyatsionnym portretam [Improvement of signal recognition accuracy by means of their cross-correlation portraits] / V.R. Krasheninnikov, N.A. Krasheninnikova, E.Yu. Galitskaya // Radiotekhnika. ”— 2014. ”— Vol. 7. ”— Pp. 107–110. (in Russian)

[18] Vasil’ev, K. Statisticheskiy analiz izobrazheniy [Statistical image analysis] / K.K. Vasil’ev, V.R. Krasheninnikov. ”— Ulyanovsk, Russia. : UlSTU, 2014. ”— P. 216. (in Russian)

[19] Armer, A. Ispol’zovanie ontologii dlya formirovaniya naborov etalonov rechevykh komand v zadache raspoznavaniya rechevykh komand na fone shumov [Using ontologies to generate a set of voice commands in the problem of speech recognition of voice commands in background noise] / A.I. Armer, V.S. Moshkin // Radiotechnika. ”— 2016. ”— Vol. 9. ”— Pp. 72–77. (in Russian)

[20] Armer, A. Podkhod k formirovaniyu naborov etalonov rechevykh komand s ispol’zovaniem ontologii [Formation of voice command model groups with ontology] / A.I. Armer, V.S. Moshkin // Ontologiya proektirovaniya. ”— 2016. ”— Vol. 6. ”— Pp. 270–277. (in Russian)

[21] Krasheninnikov, V. Optimization of dictionary and model library for recognition of speech commands / V.R. Krasheninnikov, N.A. Krasheninnikova, V.V. Kuznetsov, E.Yu. Lebedeva // Pattern Recognition and Image Analysis. ”— 2011. ”— Vol. 21, No. 3. ”— Pp. 505–507.

[22] Krasheninnikov, V. Optimization of dictionary and model library for recognition of speech commands based on cross-correlation portraits / V.R. Krasheninnikov, N.A. Krasheninnikova, V.V. Kuznetsov, E.Yu. Lebedeva // Pattern Recognition and Image Analysis. ”— 2013. ”— Vol. 23, No. 1. ”— Pp. 80–86.