Ways to increase the probability of correct recognition of noisy speech commands by their cross-correlation portraits

Ekaterina Galitskaya
Applied mathematics and computer science
Ulyanovsk State Technical University
Ulyanovsk, Russia
katrisa@yandex.ru

Viktor Krasheninnikov
Applied mathematics and computer science
Ulyanovsk State Technical University
Ulyanovsk, Russia
kvrulstu@mail.ru

***Abstract*—Currently, the field of application of voice information-control systems is being intensively expanded, for which recognition of speech commands (SC) is necessary. This recognition is very difficult in the presence of intense acoustic noise. We consider a method for recognizing noisy SCs by cross-correlation portraits (CCP), which is used for speaker-dependent recognition from a limited vocabulary of commands. In this method, SCs are converted into CCPs, which are special images. The probability of correct recognition directly depends on the choice of command standards. The standards should accurately reflect the entire class of commands, for which the library of standards is optimized. The standards are stored as CCPs. Recognized SC is converted into CCP and the closest portrait is found from the set of standard portraits. In this case, a sufficiently accurate matching of the standard and the recognized SC portraits is required. For this, two methods are proposed: phonemic alignment and variation of the boundaries of SCs, given that its boundaries can be estimated ahead or delayed. The experiments showed that the proposed modernization of the algorithm significantly increases the probability of correct recognition.**

Keywords—speech command, recognition, standard, cross-correlation portrait

I. INTRODUCTION

At present, the management of many technical systems is impossible without the participation of a human operator, despite the significant success of robotization. In this case, it is desirable to facilitate the operator work using a voice information management systems (VIMS), in which it is possible to obtain information about the state of the system and manage it by SC. However, SC recognition is required for this. To date, many speech recognition systems have been developed that are used to enter information into a computer, control robots, etc. [1-7]. However, most of these systems are inoperative in the presence of noise.

At the same time, there is a need for VIMS operating in conditions of very strong acoustic noise, for example, aviation, noisy production, etc. Installing VIMS in the cockpit can help reduce the workload of the pilot. Honeywell has tested the VIMS on its Embraer 170 aircraft (recognition accuracy of this VIMS is 90%) [8]. There are examples of VIMS using in military aircraft Eurofighter Typhoon. Lockheed Martin also developed the F-35 cab with speech recognition. Airbus Defense and Space considered adding a cockpit assistance system with voice recognition technology to its recently developed Sferion helicopter [9]. The TOUCH-FLIGHT 2 project is exploring the use of voice control as an alternative mode of interaction between pilots and cockpit avionics [10]. These programs are developed in English [11-17], which makes them impossible to use at Russian facilities, the pilots of which communicate in Russian. There are no open data on the using of Russian VIMS on aircraft. Thus, the problem of creating methods and algorithms for recognizing SC in the presence of strong interference remains relevant.

Various features of speech signals are used in recognition algorithms: spectral analysis, wavelets, hidden Markov chains, cepstral analysis, artificial neural networks, etc. Typically, SC recognition systems in severe interference conditions work with a limited dictionary. Some standards of the SCs from this dictionary are constructed, and the recognized SC refers to the closest of these standards. In this paper, we consider the method for recognizing highly noisy SCs by cross-correlation portraits (CCPs). In this method, SCs are converted into CCPs, which are special images that reflect the acoustic features of the SCs [18-20]. This makes it possible to apply image processing methods to recognition of the SC. There is an extensive literature on image processing, for example, [21-25].

Standard SCs are stored in the computer memory in the form of CCPs (the standard CCPs). Recognized SC is also converted into CCP and the nearest to standard CCPs is located. The probability of correct recognition essentially depends on the choice of standard SCs. Therefore, the library of standards is to be optimized. Sufficiently accurate matching of the portraits of the standard and the recognized SC is required to find nearest standard CCP. For this, two methods of refining alignment are proposed: phonemic alignment and variation of the boundaries of the SC, given that its boundaries can be estimated ahead or delayed. The experiments showed that the proposed modernization of the method significantly increases the probability of correct recognition.

II. RECOGNITION OF SPEECH COMMANDS BY THEIR AUTOCORRELATION AND CROSS-CORRELATION PORTRAITS

The use of autocorrelation portraits (ACPs) was proposed in [18,19] for SCs recognizing on the background of strong noise. Let $X = \{x_1, x_2 \dots x_N\}$ be SC, consisting of N values. The ACP of X is the two-dimensional array (image) R. We divide X into $M+1$ segments of the length $L = [N/(M+1)]$, where $[u]$ is the integer part of the number u. Each row of R is a sequence of sample correlation coefficients $r(t,k)$ of the segment $X_t = \{x_{(t-1)L+1}, \dots, x_{tL}\}$ and segments $X_{t,k} = \{x_{(t-1)L+1+k}, \dots, x_{tL+k}\}$ shifted by k samples relative to X:

$$r(t,k) = \frac{1}{L\sigma_t\sigma_{t+k}}\left(\sum_{j=0}^{L-1} x_{(t-1)L+j}x_{(t-1)L+j+k}\right) - \mu_t\mu_{t+k}, \quad (1)$$

where $t = 1,..., M$, $k = 1,..., K$, μ_t, μ_{t+k} are the sample means, and σ_t^2, σ_{t+k}^2 are sample variances of X_t and $X_{t,k}$. Thus, the ACP is an $M \times K$ array (image) of the sample autocorrelation coefficients of one SC

Fig. 1 shows the ACPs of once spoken SCs "Cab" (Cabina) and "Engine" (Dvigatel) and two pronunciations of "Air Conditioning" (Conditsioner) at different times. *Note, that in this paper all SC are spoken in Russian.* There are $M = 100$ rows and $K = 50$ columns (i.e. shifts) in these ACPs. The range of correlation coefficient $[-1;1]$ is converted into the range of brightness $[0;255]$ in Fig. 1. The image row reflects the change of the correlation between the values of the speech signal at shifts by $k = 1,..., K$ samples, that is, local correlations. The sequence of rows reflects the process of changing correlations with the time, for example, characterizes the sequence of phonemes.

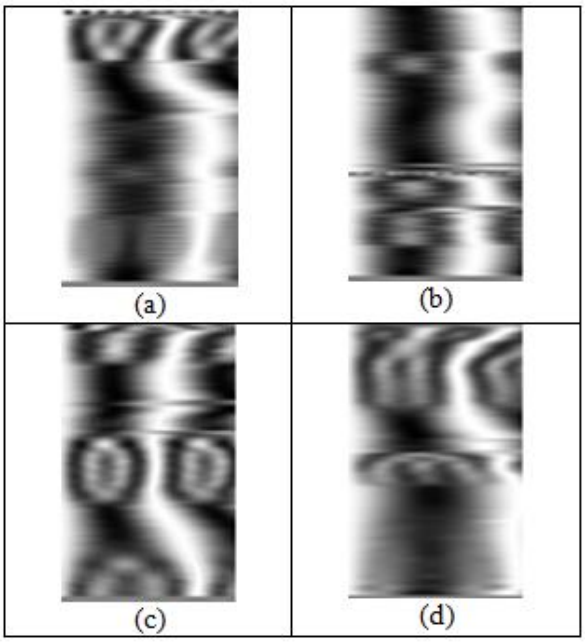

Fig. 1. Examples of autocorrelation portraits of speech commands: (a) “Cab”; (b) “Engine”; (c) “Air Conditioning 1”; (d) “Air Conditioning 2”.

It turned out, that the ACPs are individual, resistant to noise and weakly sensitive to the pronunciation volume. The main advantage of ACPs for SCs recognizing is strong rows correlation, which makes it possible to use image processing methods for filtering, recognition, etc. The standards SCs are stored in the computer's memory as ACPs. Recognized SC is also converted into ACP. This ACP refers to the nearest of the standard ACPs according to some metric. The distance between two ACPs is defined as the sum of the distances between the corresponding rows. Any metric can be used, which allows to determine the distance between two rows as vectors: Euclidean, squared, angle between the vectors, etc. When constructing the ACP, the SC is divided into $M + 1$ segments. Each segment contains some part of SC. Due to variability of the pronunciation rate, the same phonemes of SC can have different row numbers in ACPs of standard and recognized SC. As a result, the distance between these portraits will be distorted. Therefore, the matching of portraits rows should be made. The dynamic programming algorithm was used for this matching according to the criterion of the minimum distance between the matching portraits

However, there is a significant drawback: an ACP reflects the features of one SC pronunciation. This is noticeable in two portraits of the SC "Air Conditioning 1" and "Air Conditioning 2" in Fig. 1, built from the pronunciations obtained at different times. During this time, the voice timbre of the speaker, his health status, etc. could change significantly. The standards seemed to be “aging”, so the ACPs of the standard SCs and the portraits of the same recognized SC could vary significantly, which reduced the quality of recognition. Therefore, the standards need to be updated from time to time.

More complete properties of SCs are presented in its CCPs, which are built using two pronunciations [20]. Let X and Y be two pronunciations of the same SC by one speaker at different times. They are divided into the same number of $M + 1$ segments with lengths L_X and L_Y, respectively. Each CCP row is a sequence of sample correlation coefficients $r(t,k)$ of the segment $X_t = \{x_{(t-1)L_X+1},..., x_{tL_X}\}$ of SC X with the segments $Y_{t,k} = \{y_{(t-1)L_Y+1+k},..., y_{tL_Y+k}\}$ of SC Y:

$$r(t,k) = \frac{\sum_{j=0}^{L_X-1} x_{(t-1)L_X+j}y_{(t-1)L_Y+j+k} - \mu_{X,t}\mu_{Y,t+k}}{L_X\sigma_{X,t}\sigma_{Y,t+k}} \quad (2)$$

where $t = 1,..., M$, $k = 1,..., K$, $\mu_{X,t}$, $\mu_{Y,t+k}$ are sample means, and $\sigma_{X,t}^2$, $\sigma_{Y,t+k}^2$ are the corresponding sample variances. Thus, CCP is the $M \times K$ array (image) of sample cross-correlation coefficients of two SCs X and Y. If $Y = X$, then CCP coincides with ACP. Fig. 2 shows the CCPs of SCs using two of their pronunciations with the number of split segments (i.e. rows) $M = 100$ and the number of shifts (i.e. columns) $K = 50$. It is noticeable that the CCPs of the various SCs are individual, which makes them a good basis for recognition. At the same time, they to a greater extent reflect the variability of pronunciation, as they are built from two pronunciations, which are advisable to take at different times. It is noticeable that the CCPs "Air Conditioning 1 + Air Conditioning 3" and "Air Conditioning 2 + Air Conditioning 3" in Fig. 2 are less different than the portraits ACPs " Air Conditioning 1" and "Air Conditioning 2" in Fig. 1.

The standards SCs are stored in the computer's memory as CCPs. Recognized SC is also converted into CCP in pair with some pre-read pronunciation, for example, from the standards. This CCP refers to the nearest of the standard CCPs according to some metric. The distance between two CCPs is defined as the sum of the distances between the corresponding rows, similar to the ACPs case.

III. METHODS TO INCREASE THE PROBABILITY OF CORRECT COMMANDS RECOGNITION

The described recognition method gives an almost absolute correct recognition in the noise absence. The

presence of strong noise significantly reduces it for a number of reasons. Let us consider some of the interfering factors and methods to reduce their influence. Some of these methods were applied to improve the recognition of SCs by their ACPs [18,19,26,27].

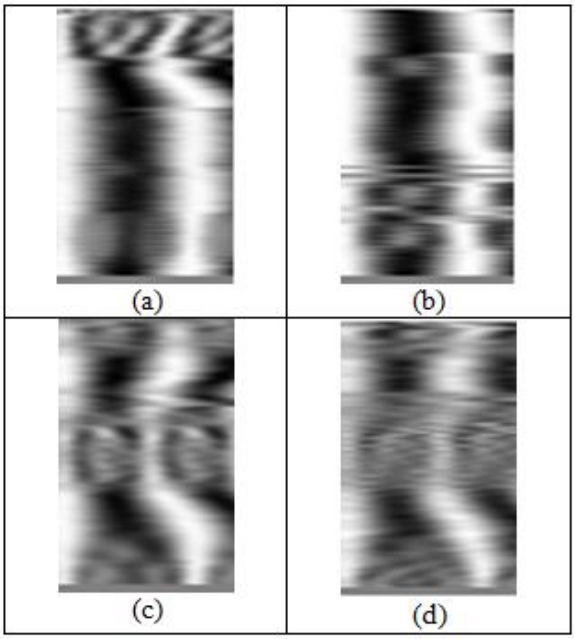

Fig. 2. Examples of cross-correlation portraits of speech commands: (a) “Cab 1 + Cab 2”; (b) “Engine 1 + Engine 2”; (c) “Air Conditioning 1 + Air Conditioning 2”; (d) “Air Conditioning 2 + Air Conditioning 3”.

The varying of the recognized SC boundaries. To compare the standards with the recognized command P, first of all, it is necessary to determine its beginning (start) and ending (end). At the same time, due to strong noise, errors are inevitable: advancing or delaying. It is especially difficult to find the ending of an SC, as it is usually pronounced quieter than the beginning. To mitigate the influence of these errors, trial additions and deletions of t samples of the signal at the estimated boundaries were applied. The value of the parameter t was chosen empirically, taking into account the fact that too large a value of it can change the command itself. In the process of recognition, the command $P_{(start\ ,end\ )}$ is converted into 9 commands: $P_{(start\ ,end\ )}$, $P_{(start\ -t,end\ )}$, $P_{(start\ +t,end\ )}$, $P_{(start\ ,end\ -t)}$, $P_{(start\ -t,end\ -t)}$, $P_{(start\ +t,end\ -t)}$, $P_{(start\ ,end\ +t)}$, $P_{(start\ -t,end\ +t)}$, $P_{(start\ +t,end\ +t)}$, where “start” and “end” are the estimated bounds of the command. For each of the 9 received variants of the command, its own CCP is built. The variant that has the smallest distance to the standard CCPs is taken as the true CCP of the recognized SC.

The CCPs width optimization. The width of the CCP (the number of columns in the portrait) K is chosen empirically. However, as the practice has shown, the optimal value of the parameter K depends on the length of the SC. Therefore, all dictionary commands were divided into groups of approximately the same length, and each group used its own value of this parameter.

The phonemes matching. When building CCPs, the SCs are divided into $M+1$ segments. Each segment contains some part of a phoneme. Due to the variability of the pronunciation rate, the segments of CCPs can begin with different phonemes, so the correlation coefficient can have a “false” value and the CCP will be distorted. To avoid this distortion, the dynamic phonemes matching algorithm was used. As a result, the beginning of the segment of one SC is shifted so that this segment is maximally correlated with the segment of the second SC.

Optimization of the standards library. The quality of recognition directly depends on how well the standard CCPs present features of pronouncing commands. In this regard, an additional problem arises of choosing the "best" standards. To do this, first, several standards of each command are built and directional was applied to achieve the best library of the standard CCPs [26]. To perform this operation, it is desirable to have a large number of recognized pronunciations SC, which requires a large time expenditure of speakers. In [25,28], the methods for obtaining realizations of quasiperiodic processes in the form of autoregressive models of cylindrical images are described. Phonemes of speech signals are also quasiperiodic processes, which made it possible to simulate many variants of pronouncing the SC even from one of its real pronunciations by a speaker.

The noise adding to the standards. The standards are usually built in advance by pronunciations in the absence of noise. The recognized SC contains significant noise, therefore, its CCP inevitably differs from the standard CCPs. Therefore, the distances between the CCPs are distorted and the quality of recognition is reduced. To correct the distances, the noise addition to the standard SCs was applied before their conversion into CCPs. In this case, the noise for the standards came from an additional microphone far from the operator’s mouth while pronouncing the recognized SC, which ensured the similarity of the noise characteristics in the compared CCPs. The disadvantage of this method is the calculation of all noisy standards for each incoming recognized SC.

IV. The Results of the Experiments

The following experiment was conducted to assess the significance of the considered methods of the correct recognition probability increasing. There was a dictionary consisting of 41 SCs on aviation topics. The dictionary was divided into 4 groups, containing 10, 5, 8, and 19 SCs, respectively. Each SC was pronounced 30 times (in total 1230 SCs participated in recognition). The SCs were additively noisy with the noise of an aircraft engine with a signal-to-noise ratio of 4. When constructing the CCPs, the first two pronunciations were chosen as standard ones. As a result of recognition (without applying the methods described above) 158 SCs were not recognized. Using the methods described above, 67 of the unrecognized SC were recognized. At the same time, the SCs recognized correctly in the first case were also recognized by the improved method. As a result, the probability of correct recognition increased from 87% to 93% (significance was tested by Student's criterion with a significance level of 0.05).

V. Conclusions

The paper proposes the use of the conversion of the SCs into CCPs for commands recognition on the background of strong noise. The CCP of two SCs is two-dimensional images, rows of which consist of cross-correlation coefficients between these SCs. The use of two

pronunciations in the CCP allows you to take into account the variability of pronunciations. The standard CCP is constructed for each SC. Recognition is carried out by comparing the CCP of the recognized SC with the standard CCPs. The performed experiments showed that the use of several modifications of this method significantly increases the probability of correct recognition.

ACKNOWLEDGMENT

The reported study was funded by the RFBR, project number 20-01-00613.

REFERENCES

[1] A. Zhdanov, "Speech input as an alternative to keyboard input," 2020 [Online]. URL: https://compress.ru/article.aspx?id=11907.

[2] M.V. Mikhaylyuk, "Ergonomic voice control interface for anthropomorphic robot," 2020 [Online]. URL: https://cyberleninka.ru/article/n/ergonomichnyy-golosovoy-interfeys-upravleniya-antropomorfnym-robotom/viewer.

[3] A. Gerasimov, "Smart home from Apple, Google and Yandex - voice control," 2020 [Online]. URL: https://voiceapp.ru/articles/smarthome.

[4] SpeechKit - Yandex speech technology, 2020 [Online]. URL: https://yandex.ru/company/technologies/speech_technologies/.

[5] D. Geer, "5 impacts of speech recognition system in various fields," 2020 [Online]. URL: https://thenextweb.com/contributors/2017/09/05/5-impacts-speech-recognition-system-various-fields/.

[6] S. Rustamov, E. Gasimov, R. Hasanov, S. Jahangirli, E. Mustafayev and D. Usikov, "Speech recognition in flight simulator," 2020 [Online]. URL: https://www.researchgate.net/publication/329485063_Speech_recognition_in_flight_simulator.

[7] 8 Innovative Ways to Use Speech Recognition for Business, 2020 [Online]. URL: https://www.transcribeme.com/blog/8-innovative-ways-to-use-speech-recognition-for-business.

[8] L. Savvides, "Hey Siri, take off! Get ready for more-advanced planes," 2020 [Online]. URL: https://www.cnet.com/news/honeywell-tests-gear-for-even-more-high-tech-planes/.

[9] Woodrow Bellamy III. Rockwell Collins Rapidly Advancing Cockpit Voice Recognition Technology, 2020 [Online]. URL: https://www.aviationtoday.com/2014/11/13/rockwell-collins-rapidly-advancing-cockpit-voice-recognition-technology/.

[10] J. Gauci, "Aircraft control through the use voice commands," 2020 [Online]. URL: https://www.um.edu.mt/newspoint/news/features/2019/07/aircraftcontrolthroughtheuseofvoicecommands.

[11] R. Crist, "Talk to your house with these voice-activated smart-home systems," 2020 [Online]. URL: https://www.cnet.com/news/talk-to-your-house-with-these-voice-activated-smart-home-systems/.

[12] Talking to Your With Telligence Voice Control, 2020 [Online]. URL: https://www.transcribeme.com/blog/8-innovative-ways-to-use-speech-recognition-for-business.

[13] Speech Recognition Interfaces Improve Flight Safety, 2020 [Online]. URL: https://spinoff.nasa.gov/Spinoff2012/t_4.html.

[14] Pilot Speech Recognition, 2020 [Online]. URL: http://www.voiceflight.com/.

[15] N. McKeegan, "Speech recognition technology allows voice control of aircraft systems," 2020 [Online]. URL: https://newatlas.com/speech-recognition-technology-allows-voice-control-of-aircraft-systems/7484/.

[16] M. Peck, "Fly by Voice," 2020 [Online]. URL: https://aerospaceamerica.aiaa.org/departments/fly-by-voice/.

[17] Speech recognition technology for air traffic controllers, 2020 [Online]. URL: https://www.internationalairportreview.com/news/75900/voice-recognition-air-traffic/.

[18] V.R. Krasheninnikov, A.I. Armer, N.A. Krasheninnikova and A.V. Hvostov, "Recognition of noisy speech command by autocorrelation portraits," Naukoemkie tekhnologii, vol. 9, pp. 65-74, 2007.

[19] V.R. Krasheninnikov, A.I. Armer, V.V. Kuznetsov and E.Yu. Lebedeva, "Cross-Correlation Portraits of Voice Signals in the Problem of Recognizing Voice Commands According to Patterns," Pattern Recognition and Image Analysis, vol. 21, no.2, pp. 185-187, 2011.

[20] V.A. Soifer, S.B. Popov, V.V. Mysnikov and V.V. Sergeev, "Computer image processing. Part I: Basic concepts and theory," VDM Verlag, Dr.. Muller, 2009.

[21] R.C. Gonzalez and R.E. Woods, "Digital image processing," Pearson, Prentice-Hall, New York, 2017.

[22] R.G. Magdeev and A.G. Tashlinskii, "Efficiency of object identification for binary images," Computer Optics, vol. 43, no. 2, pp. 277-281, 2019. DOI: 10.18287/2412-6179-2019-43-2-277-281.

[23] V.V. Myasnikov, "Description of images using a configuration equivalence relation," Computer Optics, vol. 42, no. 6, pp. 998-1007, 2018. DOI: 10.18287/2412-6179-2018-42-6-998-1007.

[24] V.R. Krasheninnikov and K.K. Vasil'ev, "Multidimensional Image Models and Processing," Computer Vision in Control Systems-3. Intelligent Systems Reference Library 135, Springer International Publishing, pp. 11-64, 2018.

[25] V.R. Krasheninnikov, N.A. Krasheninnikova, V.V. Kuznetsov and E.Yu. Lebedeva, "Optimization of dictionary and model library for recognition of speech commands," Pattern Recognition and Image Analysis, vol. 21, no. 3, pp. 505-507, 2011.

[26] V.R. Krasheninnikov, A.V. Khvostov and A.I. Armer, "Preparation of Templates in Speech Command Recognition by Single- and Double-Channel Scheme in Background Noise," Pattern Recognition and Image Analysis, vol. 18, no. 4, pp. 580-583, 2008.

[27] V.R. Krasheninnikov, A.I. Armer, N.A. Krasheninnikova, V.R. Derevyankin, V.I.Kozhevnikov and N.N. Makarov, "Autoregressive Models of Speech Signal Variability in the Speech Commands Statistical Distinction," Internetional Conference on Computational Science and it's Applications, Springer-Verlag: Berlin Heidelberg, pp. 974- 982, 2006.