Advanced Multimodal Interfaces Design Using Speech Control

Oleg Korsun[1,2,*] and *Emma* Glukhova[1]

[1]State Research Institute of Aviation Systems, Moscow, Russia
[2]Moscow Aviation Institute (National Research University), Moscow, Russia

Abstract. The paper discusses methods of developing human-machine interfaces for air-craft cabins using voice control. The use of voice control is relevant for promising multimodal aviation interfaces. The features of methods for recognizing voice commands are considered, as well as system design of human-machine interfaces based on methods for analysing crew tasks.

1 Introduction

Modality in human-machine interaction is usually understood as a channel of control and human perception of information: visual, audio, tactile and other channels. Thus, multimodal is any interaction in which more than one modality is used simultaneously.

The first multimodal system is considered to be the "Put That There" system created in the USA in the 1980s [1]. Since this first demonstration of a multimodal system, which processed speech in parallel with touchpad indications, many other multimodal systems have been developed.

In [2, 3], multimodal control refers to a single control task using several different channels. For example, selecting an object on a map: a gesture is used to select an area on the map and a confirmation command is given by voice [1].

In a broader problem statement, multimodal interaction refers to human-machine interaction realized through several different channels. For example, controlling an aircraft using a human-machine interface with speech control, touchscreen display, 3D audio and gaze control technologies, face image processing for estimation pilot's emotional states [4]. The main advantages of such interfaces are the possibility of simultaneous use of different human resources in the case of parallel tasks (e.g., voice control of avionics and simultaneous manual piloting), which presumably reduces the total load on pilots and leads to a reduction in the number of crew errors and fatigue [5].

2 Multimodal interface development process

* Corresponding author: marmotto@rambler.ru

At the initial stages of human-machine interface development it is necessary to analysed already existing interfaces, the expected operating conditions of the technical system, as well as the possibilities and technological limitations of the interfaces implementation [6, 7].

Then functional requirements to the system are analysed and a list of its functions is formed. On the basis of the analysis, the operator's work scenarios are developed together with the experts. Then the interface is designed [6, 7].

2.1 Analyses of existing aviation interfaces that use voice control

Speech command systems for voice control are already being implemented in onboard information systems of aircraft. Intensive development of a speech interface is being carried out by Eurofighter GmbH in the European Union for the Eurofighter Typhoon aircraft, Lockheed Martin Corporation in the USA for the F-16 and F-35 fighters, as well as other companies.

Since 2005, the Eurofighter Typhoon has been operating a speaker-dependent Direct Voice Input (DVI) system [8] based on reference comparisons. The system has a vocabulary of over 100 commands - those not directly related to the flight process or weapons use. Direct Voice Input is used to control auxiliary onboard equipment: radar operation modes, instrument panel and graphic screens, navigation aids, setting frequencies for radio equipment tuning, radar identification system, and so on [9].

Also within the program of Advanced Fighter Technology Integration, a Voice-Controlled Interactive Device system created by Lear Siegler [9] was developed for F-16A and F-35 fighters. This system has a vocabulary of up to 256 words, allows recognition of commands in the form of phrases and shows a recognition probability of more than 90%. The main obstacle to improve the recognition quality is the noise level in the cockpit, which can reach 120 dB during manoeuvres. In addition, a few pilots retain the ability to speak at overloads greater than 5 g.

Similar voice control systems are used on the French Dassault Rafale fighters [11] and the French-British Aerospatiale Gazelle helicopter [12].

2.2 Specifics of the expected operating conditions

The peculiarity of using voice control in the cockpit interface is that it is necessary to ensure a high probability of correct recognition. This is achieved by rational choice of the recognition method, reduction of the dictionary size, and application of noise protection methods. To reduce the size of the dictionary it is advisable to apply a hierarchical clustering of the command system in such a way that at each level only a small number of commands are used for each task. This is achieved by applying the method of hierarchical analysis of crew tasks for systems control.

It is also expedient to use fixed sets of words and phrases rather than free speech recognition. Firstly, this makes recognition much easier. Secondly, fixed commands are well suited to the systems control task, since semantic interpretation of freely constructed phrases is a separate problem.

There are also requirements for autonomy, i.e., lack of communication with external networks, servers and cloud computing facilities, and consideration of limitations on the performance of onboard processors.

2.3 Speech command recognition methods: opportunities and limitations

Currently, three main groups of speech recognition methods [13] are known that can be used for voice control in aviation interfaces.

Historically, the first group is methods based on comparison with a pattern. For each word in the dictionary, a reference is compiled, such as a parametric word portrait obtained through frequency-time quantization [13], shown in Figure 1. The task of recognition is to select the word whose reference is closest in the sense of some metric to the input signal. The advantage of this method - simplicity, small amounts of training samples, good noise immunity when using special methods. The disadvantage is relatively small dictionary size. According to [13], for methods of this group, the probability of correct recognition of individual words was 95...96% in the speaker-independent variant in conditions both without noise and with noise of a significant level, comparable to the level of the useful signal after the application of additional algorithmic techniques. This result was achieved by using a second microphone remote from the pilot and registering noise in the cockpit. Unfortunately, such impressive noise-resistant characteristics were obtained only for isolated words and for a very small vocabulary of 4...5 words [13].

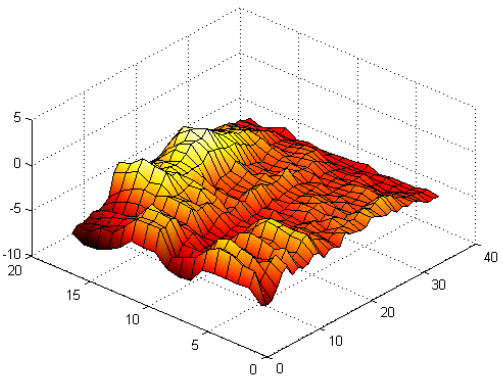

Fig. 1. Parametric portrait of the word "pilotage".

The second group is hidden Markov models. In them, the input speech is considered as a sequence of phonemes with certain transition probabilities. Recognition is performed by searching for the most probable sequence of phonemes for a given input signal. This approach shows good results for medium vocabulary sizes. It recognizes well not only single words but also phrases. It requires an average volume of training samples. The disadvantage is high sensitivity to acoustic noise. For the systems of this group considered in [12], the probability of correct recognition of commands-phrases of 3-4 words was 90...92% in the speaker-independent variant in noise-free conditions. In the presence of noise, the number of errors increased rapidly depending on the signal-to-noise level.

The third group of methods is based on artificial neural networks, first of all on convolutional deep learning neural networks. The essence of the methods consists in finding a solving function that determines from the input signal its belonging to a certain class. They provide a high probability of recognising both individual words and phrases. It is possible to achieve a high level of noise immunity when using a training sample containing noise of a given type. The disadvantage is that neural networks require very large training samples. An example of convolutional neural network architecture [13] is shown in Fig. 2.

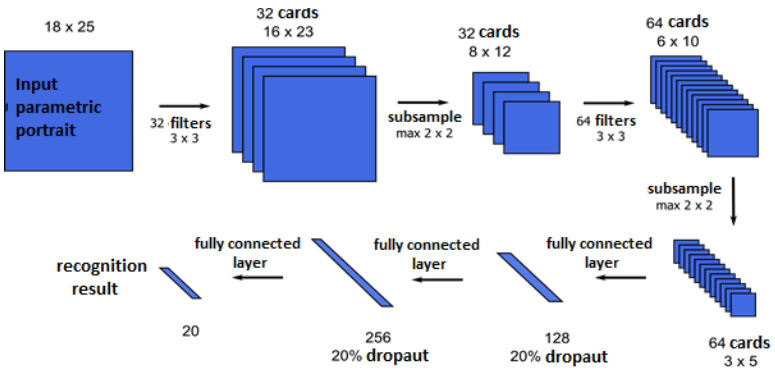

Fig. 2. Architecture of convolutional neural network for speech command recognition.

The input array is a two-dimensional parametric portrait, like presented in Fig.1. Here, unlike perceptrons, there is no need to convert a two-dimensional parametric portrait into a one-dimensional array. Next comes the convolutional layer with 32 convolutional filters of size 3×3. Then follows a subsample layer with a kernel size of 2, that is, each section of each feature map of size 2×2 is converted to one element in a new feature map. After that, there is a convolutional layer with 64 filters, each of which is applied to all feature cards and at the output of this layer 64 feature cards are obtained. Next is another subsample layer with a kernel size of 2, followed by three fully connected layers with 128, 256 and 20 elements, respectively, where 20 is the number of recognizable words. Thus, the total number of parameters of this neural network is $(3 \times 3 \times 1 + 1) \times 32 + (3 \times 3 \times 32 + 1) \times 64 + (64 \times 3 \times 5 + 1) \times 128 + (128 + 1) \times 256 + (256 + 1) \times 20 = 179988$.

In convolutional layers, as well as in multi-layer perceptron, the ReLU activation function is used. The first two fully connected layers use the hyperbolic tangent as an activation function, which is given by the equation $f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Also, after both hidden, fully connected layers, regularization is applied to prevent overfitting of the network. For the last fully connected layer, the softmax function is used to normalize the obtained output probabilities. To train this type of neural network, cross entropy is used as a loss function. The learning process uses the Adam optimization for stochastic gradient descent.

The number and sizes of filters shown above, as well as the dimensions of the final fully connected layers are taken from empirical recommendations for image recognition problem solved by convolutional neural networks. A parametric portrait is a some kind of image(see Fig.1), so it was decided to use similar parameters.

For such a network and a vocabulary size of 20 isolated words and 12 phrases of 3...4 words each, it was possible to obtain probabilities of correct recognition of 99% without noise, and 95% in conditions with noise at a signal-to-noise ratio of 6dB [13].

2.4 Analysing aircraft functions and crew tasks

The functions analysis is performed in such a way that then to determine for each function the crew tasks that are performed for its realization. In [14], a formalization of system function analysis using Petri nets is proposed.

Various methods of task analysis [15-17] and formalization of operator performance [6] are used to analyze crew performance. Different methods differ in the principle of task definition and notation of the record.

For example, hierarchical task analysis (HTA) is used in [18-20]. IRIT-group uses its own notation and its own modelling software [21].

System approaches are also used in many applications, for example the failure analysis [22].

For designing the voice controlled cockpit interface it is convenient to use the method of hierarchical task analysis. The task definition is based on a similar aircraft standard operation procedures (SOP). Based on these documents, a task tree of the aircraft crew is developed. For the same activity, it is possible to build different task trees depending on the analysis problem statement. Task definition is largely a heuristic process.

When designing an interface, the main question of analysis is "What exactly is the operator doing?". The answer to this question should be followed by a list of display and control elements that are required by the operator.

In the notation of the Hierarchical Task Analysis (HTA) method [23], a task tree is an oriented graph whose vertices are crew tasks. The arcs of the graph connect tasks to their subtasks, resulting in a hierarchical tree structure. Tasks are numbered hierarchically. For example, Figure 3 shows the coordinate correction task, which includes subtasks consisting in correction from different sources.

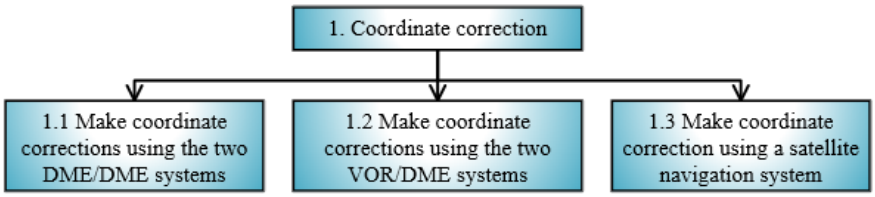

Fig. 3. Fragment of the task tree for coordinate correction.

After the task tree is compiled, it should be checked for completeness. For this purpose, expert interviews and, if necessary, methods of expert evaluation are used.

2.5 Interface development

When designing an interface for each task, it is necessary to select its modalities and specific interface elements. In prospective interfaces, some tasks can be performed in several different modalities, such as traditional haptic and voice. The choice of task modalities depends on the requirements of the tasks themselves and the characteristics of the modalities. Since voice control has lower performance reliability, it cannot be used for mission-critical tasks. Also, voice control cannot convey a smooth parameter change, such as in pen piloting. Therefore, voice control is not applicable for manual piloting tasks. Thus the whole task tree is analyzed and those tasks for which voice control is applicable and convenient are selected.

All tasks for which voice control is applicable are entered into a table. Then, for each of them tactile control and voice control commands are defined. The voice commands are developed taking into account the limitations of the recognition method. An example for the tasks shown in Fig. 3 is presented in Table 1.

Table 1. Tactile and voice interface commands.

Task	Traditional button control	Voice command
1.1 Make coordinate corrections using the two DME/DME systems	Press button “CORRECTION DME/DME”	DISTANCE
1.2 Make coordinate corrections using the two VOR/DME systems	Press button “CORRECTION VOR/DME”	OMNI-DIRECTION
1.3 Make coordinate correction using a satellite navigation system	Press button “CORRECTION SNS”	SATELLITE

The principle of formation of voice commands - they should be short, significantly different in sound, and reflect the physical meaning of the task, which facilitates recognition. For example, the DISTANCE command follows from the meaning of DME - Distance measurement equipment. The OMNI-DIRECTION command indicates the key term of the VOR system - omnidirectional. The SATELLITE command is derived from SNS - Satellite Navigation System.

3 Implementation of the interface on the modelling complex and its evaluation

Human-machine interface is a system, emergent properties of which consist in convenience and reliability of operator's work. Therefore, along with verification of the interface for compliance with relevant normative documents, it is necessary to carry out many experimental studies.

A significant part of research is carried out on modelling complexes [24]. In this case, it is necessary to investigate the performance of individual tasks, as well as a comprehensive assessment of the crew's performance throughout the flight under various external conditions.

There are various approaches to the evaluation itself. In [25] the application of Markov processes is proposed. In [26] - observation of operators' work by independent experts. In [27, 28] expert methods of operator performance evaluation are widely discussed.

4 Conclusion

1. The paper considers the main approaches to speech command recognition as part of multimodal cockpit interfaces. Their main advantages and limitations are presented, as well as the characteristics obtained from the authors' research results.
2. Descriptions of the main methods used for system analysis and structuring of interfaces are presented.
3. The feasibility of applying system analysis methods for interface design in the development of multimodal interfaces using speech control is shown.

References

1. R. A. Bolt, Put-That-There: Voice and Gesture at the Graphics Interface, Comput. Graph., **14**(3), 262–270 (1980)
2. S. L. Oviatt, Ten Myths of Multimodal Interaction, Commun. ACM., **42**, 74–81 (1999)
3. J. Barbé, A. Clay, A. Aissani, R. Mollard, WHY and HOW to study multimodal interaction in cockpit design, Ergo'IA '16, July 06 - 08, 2016, Bidart, France.
4. O.N. Korsun, V.N. Yurko, M.H. Om et al., Estimation of the interrelation between the pilot state and the quality index of piloting, AS **5**, 465–471 (2022). https://doi.org/10.1007/s42401-022-00135-z
5. C. D. Wickens, Multiple resources and performance prediction, THEOR. ISSUES IN ERGON. SCI., **3**(2), 159-177 (2002)
6. B. Adelstein, A. Hobbs, J. O'Hara et al., Design, development, testing, and evaluation: human factors engineering. NASA/TM-2006-214535. NASA, 2006.

7. A.N. Anokhin, N.A. Nazarenko, Interface design, Biotekhnosfera, **2**, 21– 27 (2010)
8. Eurofighter. Eurofighter Typhoon - About Us / Eurofighter. 2005. URL: https://www.eurofighter.com/about-us.
9. Eurofighter. Eurofighter Typhoon - Direct Voice Input Description / Eurofighter. 2016. URL: https: / / www. eurofighter . com / news - and events/2016/08/the-human-factor.
10. L. Martin, F-16 AFTI (Advanced Fighter Technology Integration) 2014. URL: http://www.f-16.net/f-16_versions_article13.html.
11. J. Bennet, *G-Force: Flying the World's Greatest Aircraft: First hand accounts from the pilots who flew them in action* (Chartwell Books, 2016)
12. Aerospace. QinetiQ Speech Recognition Technology Allows Voice Control of Aircraft Systems / Aerospace, D. News, 2007. URL: http:// www . asdnews.com/news-12659/qinetiq_speech_recognition_technology_allows_ voice_control_of_aircraft_systems.htm.
13. G.G. Sebryakov, O.N. Korsun, G.A. Lavrova, A.O. Lavrov, A.V. Poliev, V.N. Yurko, Modern Audio Technologies in the Information and Control Field of the Cockpit (Publishing House of Zhukovsky Academy, Moscow, 2021)
14. A. Chernyaev, E. Alontseva, A. Anokhin, Application of Petri nets for formalization of NPP I&C functional design, Proc. of the Int. Symp. on Future I&C for Nuclear Power Plants: ISOFIC 2017 (Gyeongju, Korea, November 26-30, 2017)
15. B. Kirwan, L. K. Ainsworth, *A Guide to Task Analysis* (London: Taylor & Francis, 1992)
16. E. Hollnagel, *Handbook of Cognitive Task Design* (Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey, 2003)
17. N.A. Stanton, Hierarchical task analysis: development, application, and extensions, Appl. Ergon., **1**(37), 55-79 (2006)
18. O.N. Korsun, A.A. Pirozhkov, E.D. Glukhova, N.V. Skryabikov, System Methodologies for the Design of Human–Machine Interfaces for Advanced Aircraft. Recent Developments in High-Speed Transport, Aerosp. Sci. Technol., 23-32 (2023) https://doi.org/10.1007/978-981-19-9010-6_3.
19. S. Mamessier, K. Feigh, HTA-Based Tracking of Pilot Actions in the Cockpit, Int. J. Hum. Factors Ergon., 93-103 (2016) DOI: 10.1007/978-3-319-41694-6_10.
20. O. N. Korsun, E.D. Glukhova, V. D. Lyakhov, N. V. Skryabikov Methodology for obtaining time intervals of crew data entry tasks performance in modern avionic systems. International Journal of Open Information Technologies ISSN: 2307-8162 vol. 11, no.4, 2023
21. Creissac Campos J. Fayollas C., D. Harrison M., Martinie C., Masci P., Palanque P. Supporting the Analysis of Safety Critical User Interfaces: An Exploration of Three Formal Tools. ACM Transactions on Computer-Human Interaction. – Vol. 6. – No. 4. – pp 341–369. – URL: https://dl.acm.org/doi/10.1145/331490.331493
22. Wu, Y., Xiao, G. & Wang, M. State-based safety analysis method for dynamic evaluation of failure effect. AS 4, 49–65 (2021). https://doi.org/10.1007/s42401-020-00073-8
23. Annett, K. Duncan. Task analysis and training design. Occup. Psychol., 1967, 41. P. 211–221.
24. Kozyrev, A. D. Implementation of the voice control function in the information and control field of the airplane cockpit / A. D. Kozyrev, I. I. Greshnikov //

Neurocomputers: development, application. - 2022. - T. 24, № 1. - C. 16-24. - DOI 10.18127/j19998554-202201-02. - EDN RFRBRZ

25. Kuravsky L.S., Greshnikov I.I. Optimizing the Mutual Arrangement of Pilot Indicators on an Aircraft Dashboard and Analysis of this Procedure from the Viewpoint of Quantum Representations // Journal of Applied Engineering Science, 2021. N 4. Pp. 1-10. DOI: 10.5937/jaes0-31855.
26. Vicente K.J., Burns C.M., Mumaw R.J., Roth E.M. How do operators monitor a nuclear power plant? Proceedings of the 1996 ANS International Topical Meeting on Nuclear Plant Instrumentation, Control, and Human-Machine Interface Technologies (NPIC&HMIT`96) (Pennsylvania State University, USA, May 6—9, 1996). — La Grande Park: ANS Inc., 1996. — Vol.2. — P.1127—1134.
27. Anokhin A. N. Expert evaluation techniques. Obninsk: IATE MEPHI, 1996. – 148 p.
28. Greshnikov I. I., Salnikov T. D., Ivanov A. S. Expert assessment of the cockpit crew information and control field. –, Volume 1958, XI International Scientific & Technical Conference on Robotic and Intelligent Aircraft Systems Improving Challenges (RIASIC 2020) 10-11 December 2020, Moscow, Russia, p. 204-222, DOI 10.1088/1742-6596/1958/1/012018.