UDC 004.934:007.52

A. Yudov, K. Ostapchenko

TRANSCRIPTION MODULE FOR VOICE COMMANDS IN AUTOMATIC DEVICES ENVIRONMENT

Abstract: The purpose of the work is to analyze the efficiency of the voice command transcription module in a multilingual environment. The study uses voices from Google's database in various languages to improve the implementation of voice commands for automatic “turn on” / “turn off” equipment. Voice commands are performed in 9 random languages, depending on the availability of the Google Voice database using the recognition module. The influence of volume and distance on the performance of the voice recognition module is analyzed. The effectiveness and influence of the choice of command language from the distance between the microphone and the speaker in the range of approximately 5 cm, 10 cm and 15 cm, as well as the volume of voice commands in Google Voice at 30%, 50% and 100%.

Keywords: voice command robot, speech transcription, voice recognition.

Introduction

Transcription of audio stream into text is a solution for a lot of important processes, such as technical translation, robotic system of automatic devices, Internet of Things, etc. [1]. At present, transcribing human speech into text is a task that requires a lot of resources and qualified experts. Thanks to the involvement of experts to prepare and configure recognition methods, automatic speech transcription can be used live. In these methods, expert speakers teach appropriate speech recognition software by recognizing their voices. They then change the speech to untrained speakers, like the process used in simultaneous translation. Current research shows that the level of accuracy of software trained on qualified voices of experts makes voice change a real tool for automatic speech transcription.

The methods of automatic speech transcription are steadily improving, with the error rate falling by as much as 25% per year in various applications. However, fully automatic transcription of audio streams in all areas remains an elusive goal and requires expert intervention to correct errors.

Analysis of existing language transcription solutions

For today, there are several different approaches to solving the problem of speech recognition in the audio data stream [2].

The first approach is the automatic detection of a voice signal using mathematical Fourier analysis, in which the signal is divided into components of sine waves of different frequencies [3]. This method of analysis means the mathematical transformation of the representation of the signal from time to frequency, and the Fourier transform can be considered as rotation in the functional space of the signal from time to frequency. Then the created signal is compared with the already known representations of the signal of each letter, from which further words are formed and whole sentences are obtained for further processing.

The second approach is speaker recognition of the voice signal. This approach differs in that its implementation requires prepared resources of test voice data to identify the speaker by his pronunciation. You need to have enough memory to store and compare speaker pronunciation data. In this approach, the speaker's voice signal is first obtained, after which it is analyzed to establish the unique qualities of pronunciation (signal). Then there is a corresponding pronunciation and after finding this correspondence, the detected pronunciation is compared only with the corresponding speaker, which greatly reduces the time for transcription [4].

The last one approach in speech automation technology is currently considered as the most promising way of language transcription, compared with the solution of the first approach in the design of voice signal.

Problems with creating a transcription module

The standard language recognition process requires the speaker to read the prepared text aloud. The recorded sound is then used to teach the speaker model. A typical registration process involves reading a single text and usually takes 20-30 minutes. Reading more than 3-4 texts rarely improves decoding accuracy. Further improvement of the speaker's model can occur when the speaker himself corrects speech errors that have occurred during dictation.

This type of language recognition process has the following disadvantages. Many people do not want to spend hours learning a speech recognition module to achieve higher accuracy. An acoustic model based on the text read may not reflect certain characteristics of free speech, individual style of the speaker or the conditions of voice recording.

To overcome these problems, specialized speech recording tools are used. It is not the voice signal of the prepared text that is recorded, but the natural language of the speaker, for example, when talking on the phone or during speeches, meetings. This recorded stream is transcribed manually. Then the recorded voice signal and the transcribed language are compared, and a model of the speaker is created [5].

Thus, the transcription of voice commands based on the process of their recognition can be used in automatic systems for more effective control of the various devices in their composition. For example, voice command transcription can be used to support the automation of household appliances [6, 7], from automatic room cooling, smart door opening, lighting, and smart home management in general. Thus, it can increase safety, comfort, and energy savings in the home.

However, the most important problem in creating a transcription module and its correct operation is the quality of pronunciation of words by the speaker, pronunciation volume (distance to the microphone), and available "reference" data with pronunciation signals of each letter and syllable for different languages, including Ukrainian. There is also a need to save volume and to update and expand voice data sets.

Therefore, the basis for the study of the effectiveness of transcription of voice commands is the analysis of the parameters of the transcription module for automatic device control using commands such as "turn on / turn off".

Description of the structure of the transcription module

The voice command transcription module consists of 3 main units, namely the input unit (microphone with signal processing module), voice recognition and transcription module, command implementation processor (Fig. 1). The input unit functions as a sensor that is used to detect input sound. The sound detection sensor can be a microphone. The transcription module will work better when processing sound if you use the same sample of sound as during the recording process. The data output from the microphone will be redirected to the voice recognition and transcription module. When developing this structure, the Easy VR 3.0 voice recognition module is used as a transcription module [8].

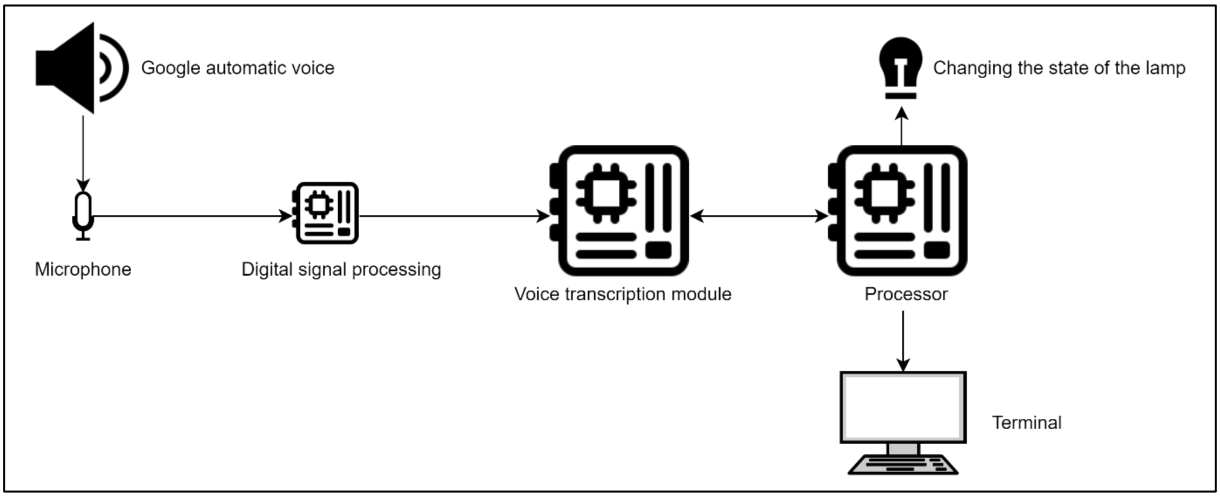

Google automatic voice
Changing the state of the lamp
Microphone
Digital signal processing
Voice transcription module
Processor
Terminal

Figure 1. Scheme of the transcription module

Voice recognition uses a variety of acoustic characteristic sounds of people. Acoustic models reflect both anatomy (eg, size and shape of the throat and mouth) and pronunciation models (eg, voice tone and speech style) [9]. Before recognizing a speaker's voice, this method requires some training when the system studies the speaker's voice, accent, and tone. This is done by recording a series of words or text commands by the user speaking through an external microphone.

A digital signal processing (DSP) module that can detect endpoints (word boundaries) to separate words and convert initial waveforms to a frequency domain representation, as well as to scale, filter, and compress data. The task of the DSP is to improve and preserve only those components of the spectral representation that are useful for the purposes of further transcription. This reduces the amount of information that the command generation algorithm should receive. This set of speech parameters for a single interval (10-30 milliseconds) is called a speech frame.

Voice recognition and transcription module is a module for creating commands based on voice recognition, for further use in the control of automatic devices. In the voice transcription module, voice commands are stored in a large set, such as a library. Voice recognition is limited to only 7 voice commands that can be imported simultaneously and efficiently. However, up to 80 voice commands can be increased during voice training of the voice recognition module.

The next block is the processing processor block. The main component is a microcontroller mounted on the Arduino Uno platform, which acts as a data processor from the input unit and the transcription module. The output of the sound sensor in the form of an analog signal will first be processed by the voice transcription module using a database of sound samples. The processor is also responsible for matching the output data to the sound database, after which the appropriate commands are sent to the light bulb and to the terminal.

Results of experimental analysis of efficiency of the transcription module

The efficiency of the transcription module and its correct operation depends on the quality of pronunciation of words by the speaker, the volume of pronunciation (distance to the microphone) and the pronunciation of voice commands in different languages such as Ukrainian, in particular. Therefore, the purpose of the experimental analysis is to check the recognition of voice commands by the module in the use of common languages and compare them with the Ukrainian language.

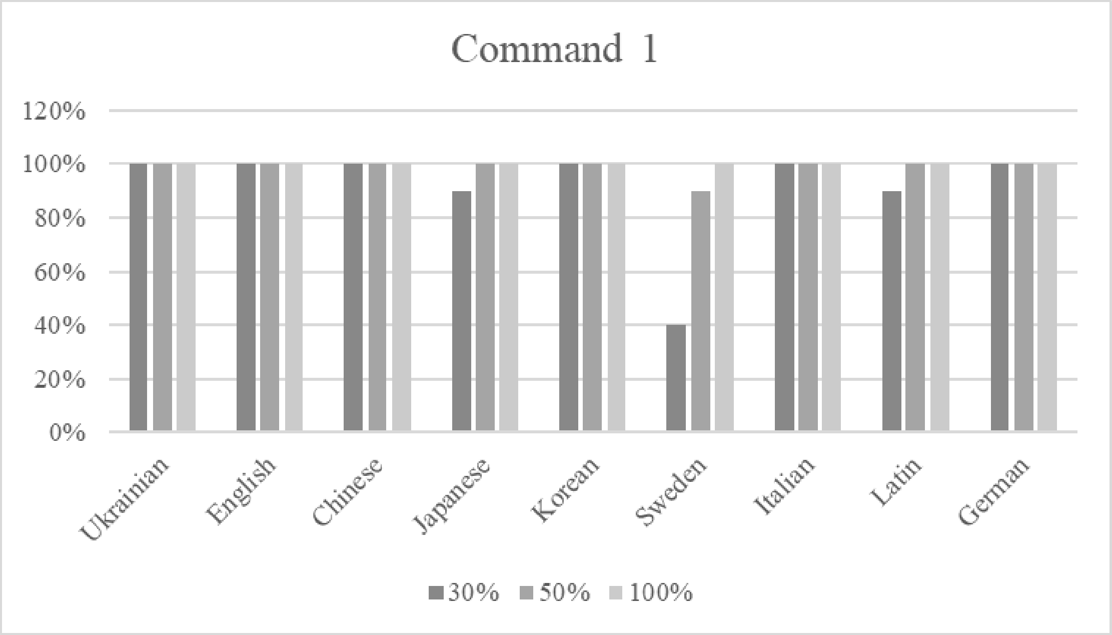
Verification of the effectiveness of transcription of voice commands is performed by changing the distance and volume of the signal from the Google voice database in 9 languages. Evaluation is carried out with the help of two voice commands in different languages. Voice commands consisted of "lights on", which means that the lights are turned on automatically, and "lights off" means that the lights are turned off automatically.

The study is conducted by sampling 10 times for each language and counting the number of successfully transcribed voice commands. The result of testing in this study is the percentage of successful commands in different languages. The percentage of successful voice commands recognized by the module at a distance of 5, 10, 15 cm between the microphone and the speaker is shown in Fig. 2-4.

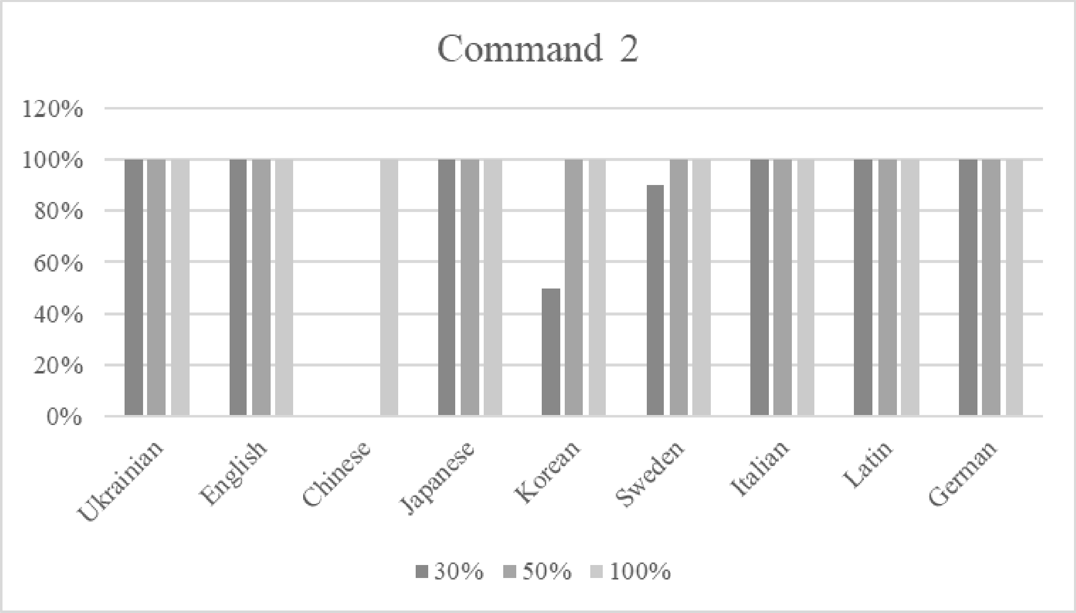
Based on the above results, it can be stated that the effectiveness of voice commands at distance of 5 cm at full volume (100%) reaches 100% for all languages tested for both voice command 1 and voice command 2. A 0% is also found in Chinese language, when the volume is 30% regulated by a voice command 2.

Also, based on the above results, it is seen that the distance is inversely proportional to the success of the commands. The greater is the distance between the microphone and the speaker from Google's voice base, the less successful the voice commands will be.

The results of the success rate of voice commands using the transcription module with a distance between the microphone and the Google voice speaker of 15 cm, are given in Fig. 3, it can be argued that the success rate of voice commands at a distance of 15 cm at full volume (100%) reaches 100% for all commands of the tested languages.

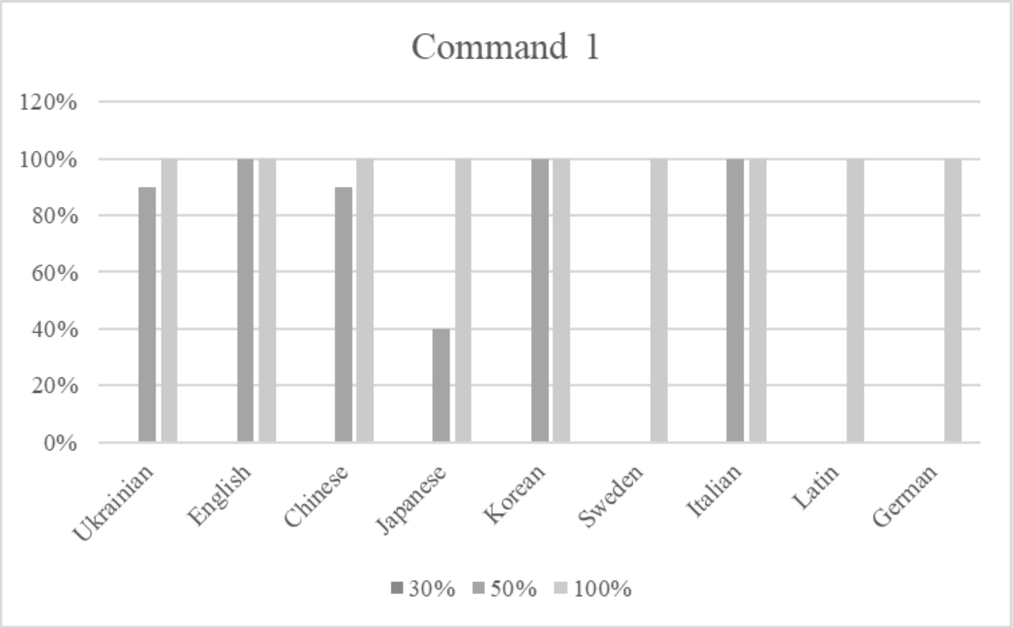


a)

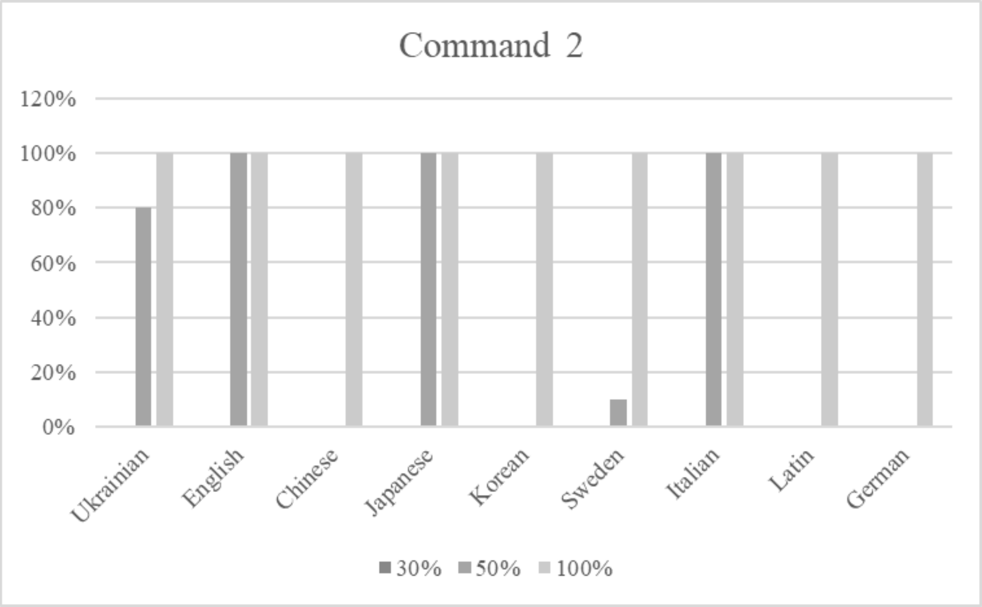


b)

Figure 2. Dependence between successful command recognition rate and volume at a distance of 5 cm

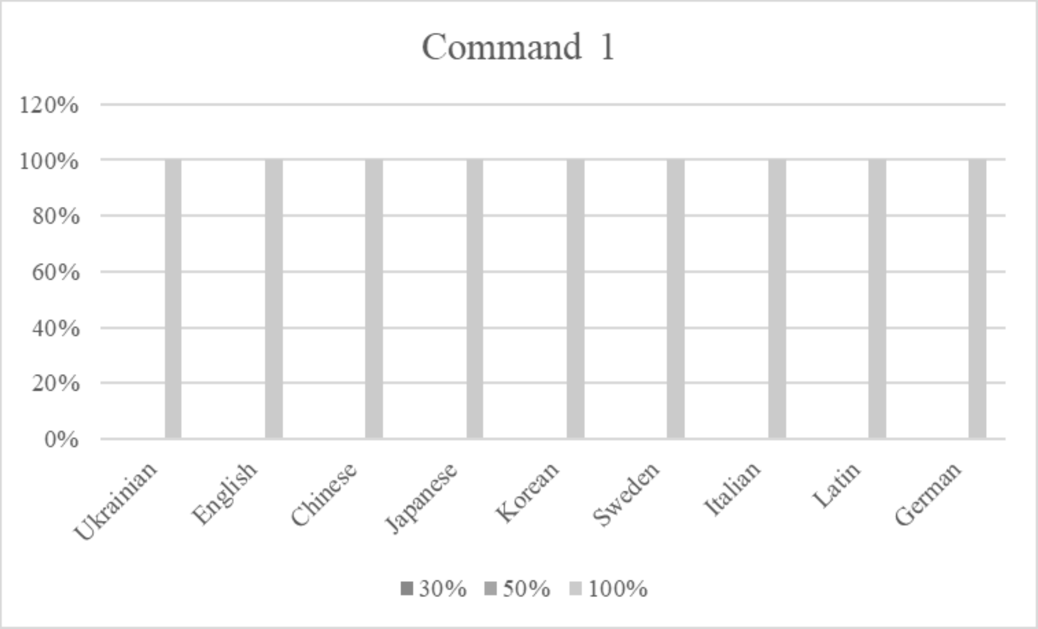


a)

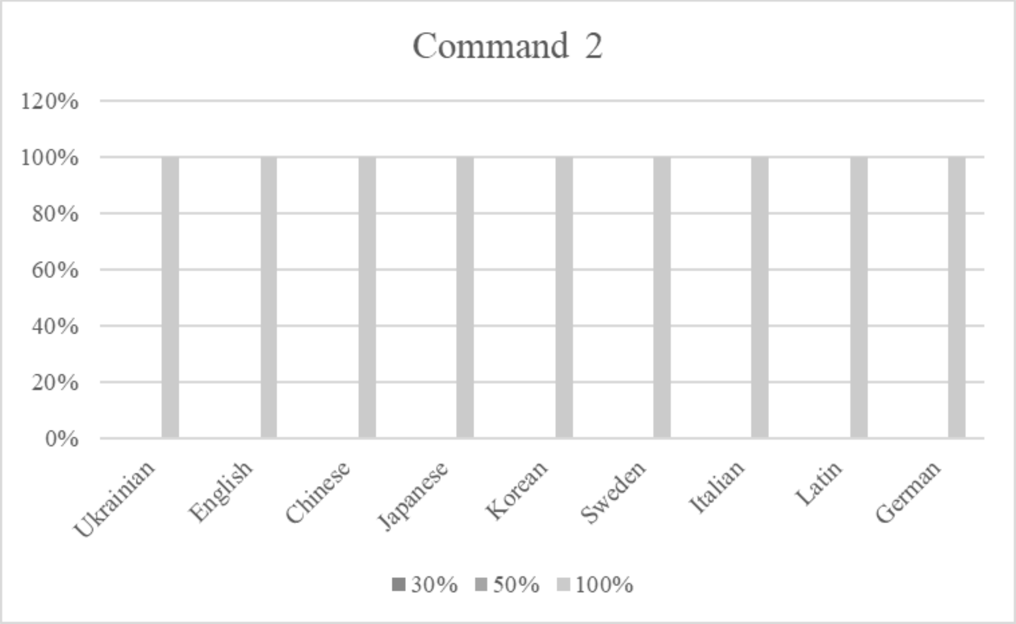


b)

Figure 3. Dependence between successful command recognition rate and volume at a distance of 10 cm



a)



b)

Figure 4. Dependence between successful command recognition rate and volume at a distance of 15 cm

Conclusions

Transcription like “turn on” / “turn off” commands based on the process of recognizing their voice signals with the ability to present in different languages is a promising and quite effective direction when used in systems of automatic devices controlled by voice commands.

An experimental study of the effectiveness of the transcription module and the implementation of voice commands in 9 random languages from the Google Voice voice database. The effects of volume and distance on module performance were studied. The results showed that the volume of the signal from the Google voice database on the mobile phone is directly proportional to the success rate of voice commands. While the test results from the distance of the microphone are inversely proportional.

In general, it can be argued that the voice transcription module can work well at a distance of 5 cm even at a voice volume of 30% for the vast majority of languages. Except for Chinese, because the loud pronunciation of the word sounds weak in it. Thus, the clarity of the pronunciation of voice commands affects the success of recognition and speed of their execution in the control systems of automatic devices.

REFERENCES

1. Sen, S. Design of an intelligent voice-controlled home automation system / S. Sen, S. Chakrabarty, R. Toshniwal, A. Braumik // International Journal of Computer Applications. 2015. Vol. 121, No 15. Pp. 39-42. DOI: https://doi.org/10.5120/21619-4904

2. Owens, F. Signal Processing of Speech / F.J. Owens. New York, US: McGraw-Hill Inc, 1993. URL: https://link.springer.com/book/10.1007/978-1-349-22599-6

3. Kumar, S. An Approach for Automatic Voice Signal Detection (AVSD) using Matlab / S. Kumar, A. Shastri, R.K. Singh // International Journal of Computer Theory and Engineering. 2011. Vol. 3, No 2. Pp. 240-247. DOI: https://doi.org/10.7763/IJCTE.2011.V3.311

4. Campbell, J.P. Speaker Recognition. In: Jain, A.K., Bolle, R., Pankanti, S. (eds) Biometrics. Springer, Boston, MA., 1996. DOI: https://doi.org/10.1007/0-306-47044-6_8

5. Kanevsky, D. Speech Transcription Services / D. Kanevsky, S. Basson, S. Chen, A. Faisman, A. Zlatsin // International Conference on Speech and Computer SPECOM-2006. St. Petersburg, 25-29 June 2006. Pp. 37-43. URL https://www.researchgate.net/publication/228738432_Speech_Transcription_Services

6. Kamdar, H. A review on home automation using voice recognition / H. Kamdar, R. Karkera, A. Khanna, P. Kulkarni, S. Agrawal // International Research Journal of Engineering and Technology. 2017. Vol. 4, No 10. Pp. 1795-1799. URL: https://www.irjet.net/archives/V4/i10/IRJET-V4I10329.pdf

7. Karudaiyar, G. IOT Based Voice Controlled Smart Home Automation / G. Karudaiyar, S. Bhummireddi, C. Deepak // International Journal of Engineering Applied Sciences and Technology. 2017. Vol. 2, No 5. Pp. 44-45. URL: http://www.ijeast.com/papers/44-45,Tesma205,IJEAST.pdf

8. EasyVR 3 Plus Manual. RoboTechsrl. URL: https://fortebit.tech/docs/manuals/easyvr-3/ (accessed 14.01.2020)

9. King, R. Speech and Voice Recognition / R. King // Biometrics Research Group. 2014. URL: https://www.biometricupdate.com/wp-content/uploads/2014/05/Voice-Biometrics.pdf