https://doi.org/10.31891/csit-2023-4-7
UDC 004.934

Olesia BARKOVSKA, Vladyslav KHOLIEV
Kharkiv National University of Radio Electronics

NEURAL NETWORK ARCHITECTURE FOR TEXT DECODING BASED ON SPEAKER'S LIP MOVEMENTS

The paper analyses the impact of using a speechless access interface (SSI), which provides the definition of the initial phase of the sound series associated with the beginning of speech based on the analysis of visemes, on the accuracy of voice command recognition in different sound environments. The analysis of the methods for recognizing the speech pattern of a speaker has shown that recent studies are based on the use of neural network architectures (CNN, LSTM) to analyze a predefined region of interest - the speaker's mouth.

In this paper, we tested a command recognition system using the SSI approach and conducted a series of experiments on modern solutions based on ALR interfaces. The main goal was to improve the accuracy of speech recognition in cases where it is not possible to use the speaker's non-noisy audio sequence, for example, at a great distance from the speaker or in a noisy environment. The obtained results showed that training the neural network on a GPU accelerator allowed to reduce the training time by 26.2 times using a high-resolution training sample with a size of the selected mouth area of 150 × 100 pixels. The results of the analysis of the selected speech recognition quality assessment metrics (word recognition rate (WRR), word error rate (WER), and character error rate (CER)) showed that the maximum word recognition rate of the speaker's speech is 96.71% and is achieved after 18 epochs of training. If we evaluate the character regonition rate of viseme recognition, the highest rate can be obtained after 13 epochs of training. Future research will focus on the use of depth cameras and stereo vision methods with increased frame rates to further improve the accuracy of voice command decoding in conditions of high background noise.

To further develop this work, we can apply noise reduction algorithms to the audio signal or solve the problem of detecting visemes in conditions of low brightness or a different angle of the face.

Keywords: NLP, automated lip reading, feature detection, audio processing, neural network, mode

Олеся БАРКОВСЬКА, Владислав ХОЛЄВ
Харківський національний університет радіоелектроніки

НЕЙРОМЕРЕЖЕВА АРХІТЕКТУРА ДЛЯ ДЕКОДУВАННЯ ТЕКСТУ ЗА РУХОМ ГУБ СПІКЕРА

У статті проаналізовано вплив використання інтерфейсу безмовного доступу (SSI), який забезпечує визначення початкової фази звукового ряду, що асоціюється з початком мовлення, на основі аналізу візерунків, на точність розпізнавання голосових команд у різних звукових середовищах. Аналіз методів розпізнавання мовного патерну диктора показав, що останні дослідження базуються на використанні нейромережевих архітектур (CNN, LSTM) для аналізу заздалегідь визначеної області інтересу - рота диктора.

У роботі протестовано систему розпізнавання команд з SSI-підходом та проведено ряд експериментів над сучасними рішеннями на основі ALR інтерфейсів. Головною метою було покращення точності розпізнавання мови у таких випадках, коли немає можливості використрвувати незашумлений аудіоряд спікера, наприклад на великій відстані від того, хто говорить, або у шумному оточенні. Отримані результати показали, що тренування нейронної мережі на графічному прискорювачі дозволило скоротити час навчання у 26,2 рази, використовуючи навчальну вибірку із високої роздільної здатності та розміром виділеної зони рота, що становить 150 × 100 пікселів. Результати аналізу обраних метрик оцінки якості розпізнавання мови (послівна точність розпізнавання (WRR), послівна помилка розпізнавання (WER) та посимвольна помилка розпізнавання (CER)) показав, що максимальна точність послівного розпізнавання промови спікера становить 96,71% та досягається після 18 епох навчання. Якщо оцінювати посимвольну точність розпізнавання візем, то найвищий показник можна отримати після 13 епохи навчання. Майбутні дослідження будуть зосереджені на використанні камер глибини та методів стереозору із збільшеною частотою кадрів задля подальшого збільшення точності декодування голосової команди в умовах великого фонового зашумлення.

Для подальшого розвитку цієї роботи можна застосувати алгоритми шумозаглушення до аудіосигналу або вирішити проблему виявлення виразів обличчя в умовах низької яскравості або іншого кута нахилу обличчя.

Ключові слова: NLP, автоматичне читання по губах, виявлення ознак, обробка звуку, нейронна мережа, режим

Introduction

Natural language processing is a general field of artificial intelligence and mathematical linguistics that studies the problems of computer analysis and synthesis of natural languages [1-2]. Solving these problems means creating a more convenient form of human-computer interaction, so there are SSI-based systems that use ultrasound or optical cameras and capture facial or neck movements as a signal regardless of the incoming audio.

Voice commands fit well into the concept of building a natural language user interface [3]. In addition, such technologies have already become widespread in various spheres of life (figure 1)

This list could be constantly updated, but such systems are most commonly used as voice assistants in the form of software applications on smartphones, PCs, or special devices similar to audio speakers.

This paper discusses in detail the applications where high noise levels or lack of sound signal are a problem, such as for people with disabilities, aviation, cars with insufficient noise insulation, and dialog restoration in silent movies [4-6].

Most often, speech recognition is defined as the conversion of an audio sequence of a human voice recording into text data. However, using not only audio information but also video can significantly improve the quality of recognition or even replace audio.

The main problem for a large number of SSI-based command recognition projects is that they use uncomfortable devices that need to be attached to the skin, which allows for experimental use only on patients, military, or astronauts, thus limiting mass adoption [7]. Many of these systems have a limited number of commands, such as a couple dozen in a NASA research project, and have variable accuracy for the same words from the same user due to the shifting of sensors on the body during use. Therefore, for convenience, it is better to use other devices (video camera or depth camera).

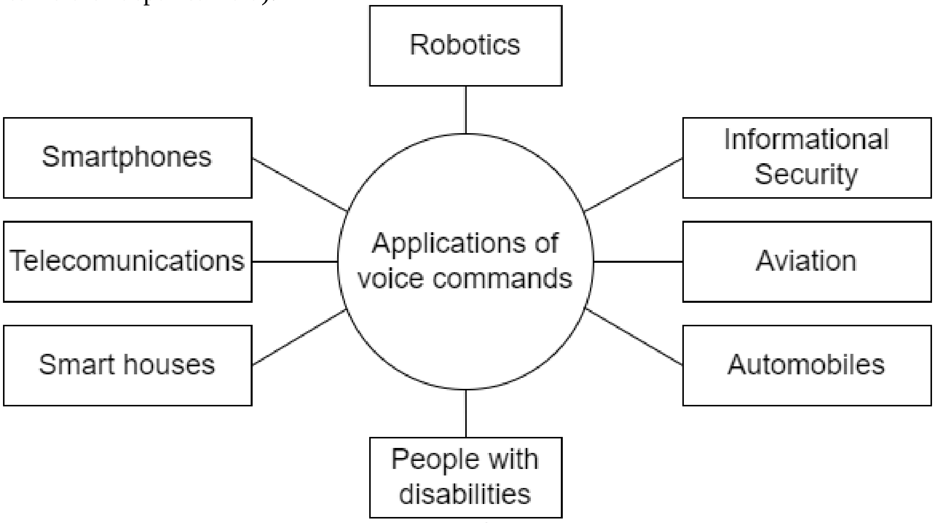


Fig.1. Applications of voice commands

The problem of high noise levels can be solved by using an algorithm for automated lip reading from a video stream, combining audio signal processing with viseme recognition in the frame [8-9]. On the other hand, such systems still have the problem of homophones, which include the same lip movements for different words, which can be solved by adding sound processing, but this method does not work well in noisy environments without additional filtering. Therefore, conducting research in this area and developing our own solution for recognizing voice commands, combining work with audio and video sequences depicting the speaker, is a relevant task.

Related works

Visual speech recognition or lip reading plays an important role in human communication, especially in noisy environments, and can be extremely useful for people with hearing impairments, so ALR technology was chosen for camera-based command recognition. To identify a word or sentence, the system must be trained using data collected for a particular language and vocabulary. However, ALR uses visemes instead of phonemes.

ALR (automated lip reading) is the process of decoding text from the speaker's mouth movement. Machine lip reading is complex because it requires extracting spatio-temporal characteristics from the video, namely the position and movement of the lips. It also complicates the process of recognizing the position of the tongue and teeth, as in many cases they are hidden behind a closed or covered mouth, so they are difficult to recognize without context.

Recent research in the field of ALR has shown a surge in end-to-end deep learning approaches for lipreading that focus on word-level prediction using a combination of convolutional and recurrent networks [10]. Therefore, in the further work we will follow this approach.

The basic detection of visemas is based on the analysis of facial geometry. When the mouth is open, the distance between the corners of the mouth increases. Even though people have different mouth sizes, you can normalize this indicator by dividing it by the distance between the jaws and get a general ratio that can be used for different faces. Each image contains a large amount of raw information that is not used in speech recognition. Therefore, it is necessary to process each image and clearly identify the AOI - the area of the lips.

Among the existing software solutions that can be used to implement certain stages of ALR systems (for example, lip area extraction), we can highlight the Dlib library, MTCNN, Openface, LFW landmarks, etc. The latter provides fast processing but low accuracy. The Dlib library is preferred because it is open source, which is important for editing algorithms or changing parameters for recognition.

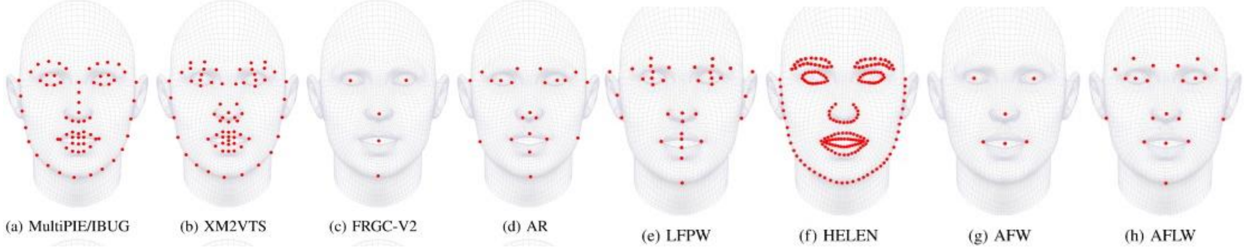


Fig.2. Examples of facial features markup

The dlib facial feature detector is developed using the classical histogram of oriented gradients (HOG) function in combination with a linear classifier, an image pyramid, and a sliding window detection scheme. The oriented gradient histogram is an image processing algorithm that performs feature extraction. Dlib contains information about the markup of dots in the face contour and applies them to the input frame, and in the output frame, it marks these dots if the image contains a mouth, eyes, or other facial features.

To recognize facial contours, including lips, we use shape_predictor_68_face_landmarks.dat, which is trained on the iBUG 300W image collection (figure 2). Other files can also be used, for example, based on HELEN, as it has a large number of dots that highlight the upper and lower lips, as well as the open mouth. Using these features, the algorithm obtains lip-centered images of 100 × 50 pixels, which will be sufficient for further processing by neural networks. The area with the detected lips is also enlarged by 10 or more pixels on each side so that the lips do not end up cropped.

To train a recognition system, a speech corpus is required, examples of which are shown in figure 3 [11-12]. GRID is a collection of tens of thousands of short videos in which 34 volunteers read nonsensical sentences in English with captions. Each file is three seconds long, and each sentence follows a pattern: command, color, preposition, letter, number, and adverb. Examples of such sentences: «place blue in M one soon», «set blue by A four please», and «place red at C zero again».

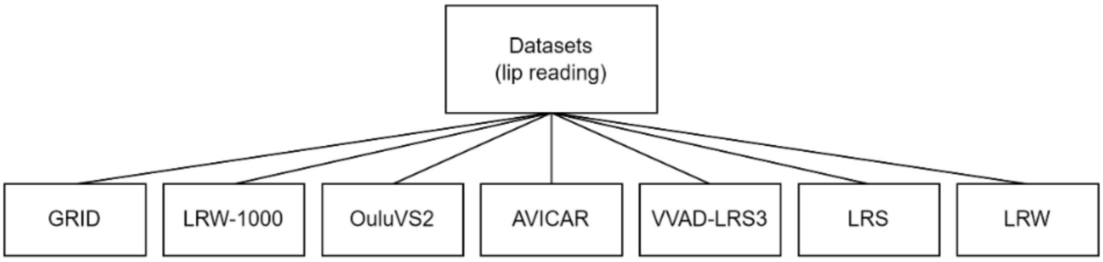


Fig.3. Examples of ALR corpora

The advantages of this corpus include a large number of videos in which the lip movements of different people are clearly visible in a bright room, which will be enough to build a recognition system, but for practical use it is better to use other corpora where the head position is not full-face, there are no template sentences and in dark lighting. To meet these conditions, we can use the OuluVS2 body, which was recorded by six cameras from five different views located between the frontal and profile views of 50 people, to analyze the non-smooth mouth movement, but the main problem will be the implementation of the ALR algorithm from the frames where the human face is in profile.

LRW-1000 used to be called CAS-VSR-W1k and is often used for recognition, but it is difficult to use for this paper because the authors of this paper do not speak Chinese, although LRW-1000 includes 18018 video samples from about 2000 people and is a good corpus for practical use because it contains different lighting, head positions and non-laboratory sentences.

The best speech corpus to compare with others in different sound conditions is AVICAR, as it records the faces of 86 people in different positions inside the car and from 7 microphones, and has 5 noise levels depending on the speed of the car. But the problem is the difficulty of obtaining all but a limited number of people's data, as links to the files were unavailable at the time of writing (last updated in 2004) or confirmation from the University of Illinois researchers is required.

Also common is the LRS corpora group, which has different versions (LRS, LRS2, LRS3-TED, MV-LRS, etc.) and was created by BBC television, with each sentence being 100 characters long. Due to the large number of camera positions, file size, and variety of content, the LRS2 dataset is more complex (75.2% of non-frontal face frames) than the LRS or MV-LRS dataset and is recommended for projects with the best recognition algorithms.

In addition to these data, there is a lack of a standardised set of similar videos in Ukrainian, which could be used to compare the results for different systems without creating our own corpus.

Aims and task of the work

The aim of this paper is to study the effect of using a speechless access interface (SSI), which provides the definition of the initial phase of the sound series associated with the beginning of speech, based on the analysis of visemes, on the accuracy of voice command recognition in different sound environments.

In order to achieve this goal, the following tasks must be solved:

- analysis of methods for recognising the speech pattern of a person speaking;
- creation of a model for recognising voice commands, improved by analysing the speech pattern of a person speaking;
- implementation of a voice command recognition model based on sound series analysis;
- implementation of a voice command recognition model based on a combination of sound series analysis and speaker's lip image;
- analysis of the obtained results.

To further develop this work, we can apply noise reduction algorithms to the audio signal or solve the problem of detecting visemes in conditions of low brightness or a different angle of the face.

Results and Discussion

To create a speech recognition system, a neural network architecture is used that maps sequences of visemes in video fragments of variable length into text sequences.

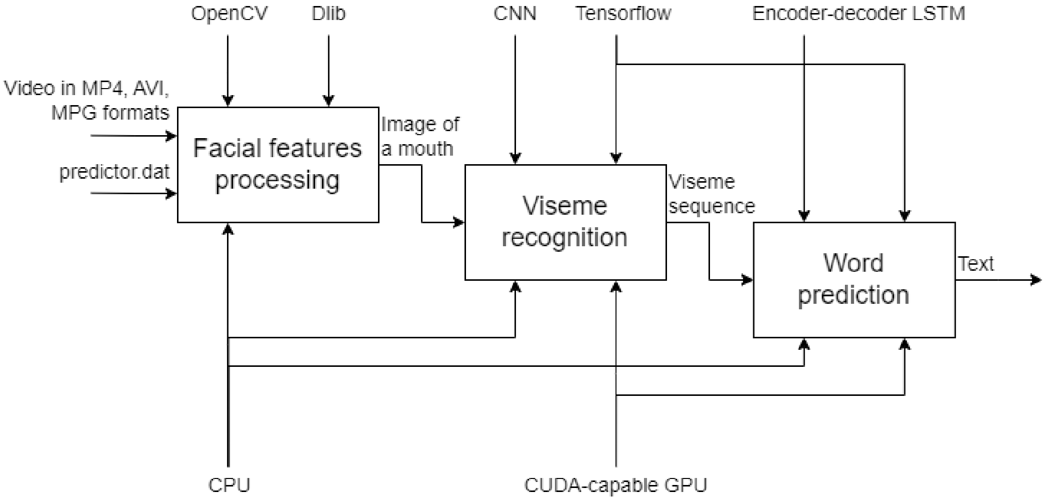


Fig.4. IDEF0 notation of the proposed system for viseme recognition

In the proposed system (figure 4), the video is decomposed into a sequence of frames containing lip images. In the next step, these frames are used as input to a convolutional neural network that has been trained on similar data. The data from the CNN is then passed through fully connected layers to form the input vector of the LSTM. The output of one layer becomes the input of the next recurrent layer. The last step is to decode the probability distribution vector of potential visemes in the LSTM, and as a result, a sequence of characters is formed, which are combined into words.

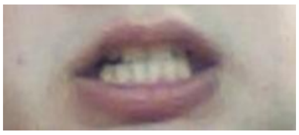
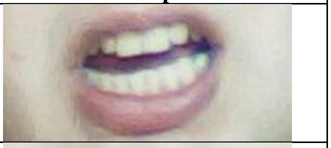
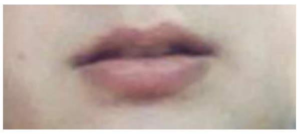
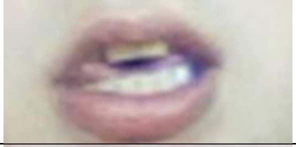
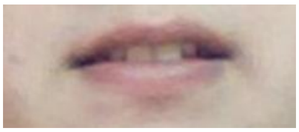
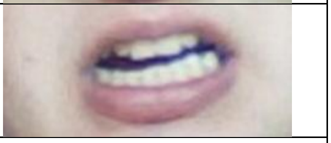
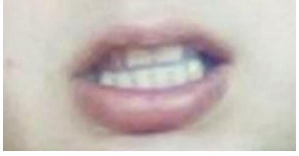
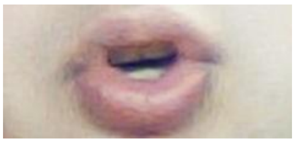
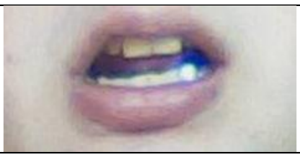
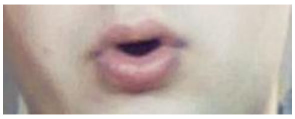
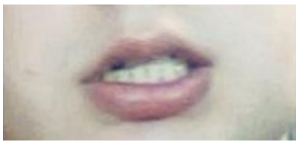
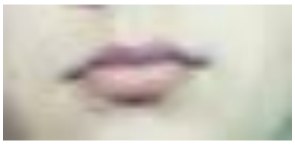
To solve the task, it is necessary to determine which words or phrases are pronounced from a fixed set of known phrases. The system's components use a sequence of images as input, and the output is words. Table 1 shows 11 visemes and a silence state that correspond to the phonemes of the English language and can be programmed into a dictionary, since phoneme groups do not differ in visual features.

The accuracy of viseme recognition will be low in cases where the distinguishing feature is the position of the tongue, such as $V_{D,T,S}$ and $V_{G,K,N}$, which requires an environment with good lighting and recognition in the dark will not be possible. For such conditions, it is possible to use a depth camera, since light does not affect the data and there are projects that use this device, such as Microsoft Kinect .

Before starting the testing, we downloaded a collection of videos with audio and text in a separate subtitle file. The training set consists of the first 30 archives of the GRID corpus without the 22nd archive (no video due to technical reasons), and the test set contains the last 3 archives, whose speakers were not considered in the training set. GRID is a collection of tens of thousands of short videos in which 34 volunteers read nonsensical sentences in English with captions. Each file is three seconds long, and each sentence follows a pattern: a command, a colour, a preposition, a letter, a number, an adverb. Examples of such sentences: «place blue in M one soon», «set blue by A four please» and «place red at C zero again». A pseudo-random number generator was also used to select a file from the test sample. The result of the system's efficiency testing in detecting the mouth area in the video and cropping it into a sequence of 150 × 100 pixels with the viseme detection is shown in table 2.

Table 1

English visemes and phonemes with examples

Consonants			Vowels		
Viseme	**Phoneme**	**Example**	**Viseme**	**Phoneme**	**Example**
$V_{J,C,H}$	/dʒ/ /tʃ/ /ʃ/ /ʒ/		V_A	/ɑː/ /aʊ/ /aɪ/ /ʌ/	
$V_{P,M,B}$	/p/ /b/ /m/		V_E	/e/ /eɪ/ /æ/	
$V_{F,V}$	/f/ /v/		V_I	/iː/ /ɪ/	
$V_{D,T,S}$	/d/ /t/ /s/ /z/ /θ/ /ð/		V_O	/ɔː/ /ɔɪ/ /əʊ/	
$V_{R,W}$	/r/ /w/		V_U	/ʊ/ /uː/	
$V_{G,K,N}$	/g/ /k/ /n/ /ŋ/ /l/ /y/ /h/		Silent		

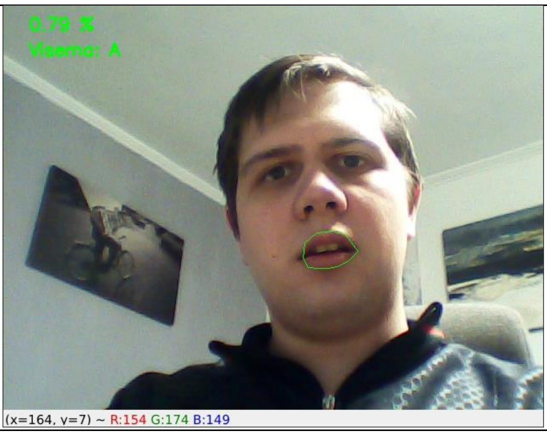
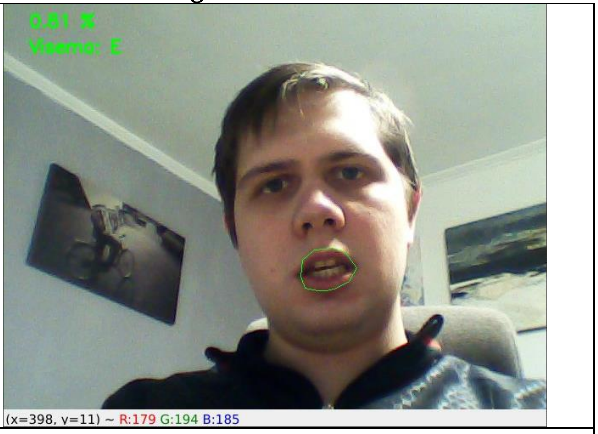
The next step is to train the neural networks on a high-resolution training set, which involves sequentially processing several hundred videos from each directory, selecting a 150 × 100 pixel mouth area. To accelerate the neural network training process and reduce the processing time by several dozen times, an NVIDIA GeForce GT 960m graphics card with Compute Capability of 5.0 was used. Training the neural network on the CPU took 64 minutes and 37 seconds, and on the above GPU – 2 minutes and 28 seconds.

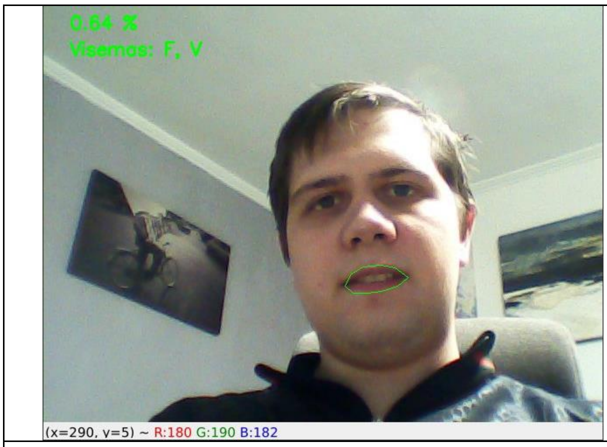
Neural networks were not trained on low-resolution video with a 75 × 50 selected mouth area, as [36] reported results showing a significant decrease in recognition accuracy when using the above settings.

The result is a text file with the corresponding recognised voice command for each video.

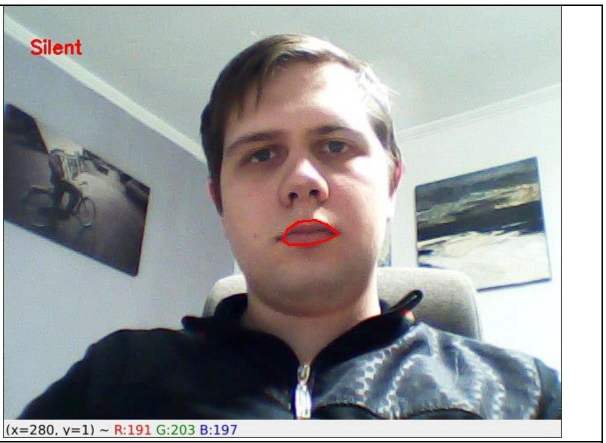
Table 2

Results of mouth area detection and viseme recognition

	
Viseme V_A	**Viseme** V_E



Viseme $V_{F,V}$



Silence

The system took 20 epochs to train, which was chosen empirically. The results are presented in table 3. We calculated the word recognition accuracy, word and character errors, and mean. The total number of video sequences for training in epoch 1 is 492 and corresponds to the total number of all videos with a speaker.

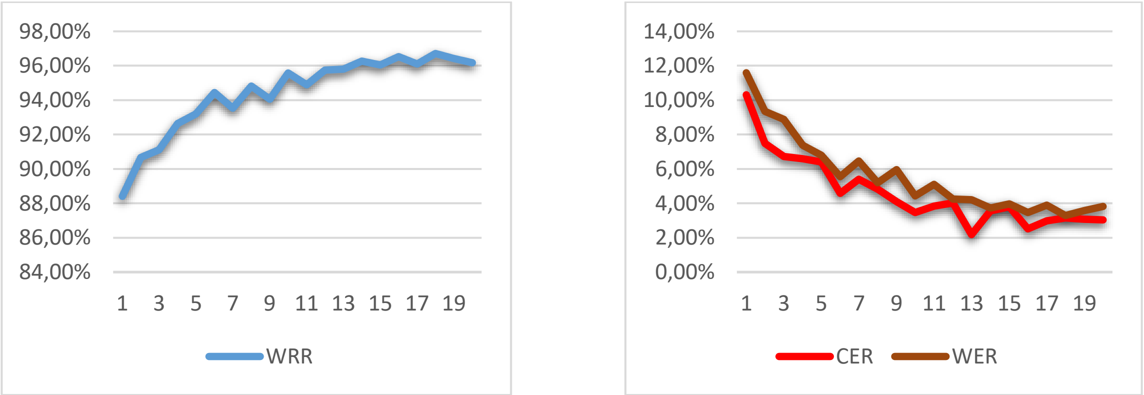
We use standard metrics for evaluating the quality of speech recognition, such as word recognition rate (WRR), word recognition error (WER) and character recognition error (CER).

Table 3

CER, WER and WRR values for 20 epochs

Epochs	CER	WER	WRR
1	10,31%	11,59%	88,41%
2	7,48%	9,34%	90,66%
3	6,73%	8,88%	91,12%
4	6,59%	7,37%	92,63%
5	6,41%	6,8%	93,2%
6	4,58%	5,55%	94,45%
7	5,4%	6,46%	93,54%
8	4,82%	5,19%	94,81%
9	4,1%	5,95%	94,05%
10	3,47%	4,43%	95,57%
11	3,84%	5,11%	94,89%
12	4,03%	4,25%	95,75%
13	2,18%	4,2%	95,8%
14	3,56%	3,73%	96,27%
15	3,8%	3,96%	96,04%
16	2,51%	3,47%	96,53%
17	2,98%	3,9%	96,1%
18	3,15%	3,29%	96,71%
19	3,07%	3,58%	96,42%
20	3,04%	3,82%	96,18%
Mean	4,6%	5,54%	94,46%

Increasing the number of epochs beyond 18 negatively affects the accuracy of the system and leads to network overtraining, which is evident in the last two results of table 3.



a)

b)

Fig.5. Performance dependency on 20 epochs: a) word recognition rate; b) word and character recognition error

The low accuracy with a small number of training epochs is explained by the fact that when using a recurrent neural network, input data from each frame is fed sequentially, so the system will not learn to correctly identify the word from the first frames, so further training is required to reduce the frequency of misinterpretations of the word.

We also checked the number of errors in characters and words, and found a significant impact of incorrect word recognition on the system's accuracy (figure 5). This may be due to the problem of visem-based word recognition in the module, which uses a recurrent neural network and requires changes to improve accuracy, such as replacing LSTM with bidirectional GRUs or other neural networks.

Analysis of the results showed that the largest number of errors were in homophones, for example, the letter v was replaced by f or b by p.

The average word recognition accuracy is 94.46%, which is a good indicator for modern recognition systems based on silent speech interfaces and is comparable to similar projects or software tools that recognise words based on audio only.

Conclusions

The paper analyses the impact of using a speechless access interface (SSI), which provides the definition of the initial phase of the sound series associated with the beginning of speech based on the analysis of visemes, on the accuracy of voice command recognition in different sound environments.

The analysis of the methods for recognising the speech pattern of a speaker has shown that recent studies are based on the use of neural network architectures (CNN, LSTM) to analyse a predefined region of interest - the speaker's mouth. In this work, we used the GRID corpus to perform experimental studies, since this collection of videos was taken with the best lighting conditions and short laboratory phrases, which makes it easy to test the proposed approach. In the future, at the stage of implementation of the research results, it is planned to train the proposed system with a neural network architecture on lower quality videos, for example, from the LRS collection.

Training of the neural network on a GPU accelerator allowed to reduce the training time by 26.2 times using a high-resolution training sample with the size of the selected mouth area of 150 × 100 pixels. The results of the analysis of the selected speech recognition quality metrics (word recognition rate (WRR), word recognition error (WER), and character recognition error (CER)) showed that the maximum word recognition rate accuracy of the speaker's speech is 96.71% and is achieved after 18 epochs of training. If we evaluate the character recognition rate of word recognition, the highest rate can be obtained after 13 epochs of training.

To further develop this work, we can apply noise reduction algorithms to the audio signal or solve the problem of detecting visemes in conditions of low brightness or a different angle of the face.

References

1. Havrashenko, A., & Barkovska, O. (2023). Analysis of text augmentation algorithms in artificial language machine translation systems. *Advanced Information Systems*, 7(1), 47-53. https://doi.org/10.20998/2522-9052.2023.1.08

2. Barkovska, O., Kholiev, V., Havrashenko, A., Mohylevskyi, D., & Kovalenko, A. (2023). A Conceptual Text Classification Model Based on Two-Factor Selection of Significant Words. *In COLINS* (2) (pp. 244-255).

3. Mykhailichenko, I., Ivashchenko, H., Barkovska, O., & Liashenko, O. (2022, October). Application of Deep Neural Network for Real-Time Voice Command Recognition. *In 2022 IEEE 3rd KhPI Week on Advanced Technology (KhPIWeek)* (pp. 1-4). IEEE. https://doi.org/10.1109/KhPIWeek57572.2022.9916473

4. Hanifa, R. M., Isa, K., & Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, 90, 107005. https://doi.org/10.1016/j.compeleceng.2021.107005

5. Prabakaran, D., & Shyamala, R. (2019, February). A review on performance of voice feature extraction techniques. *In 2019 3rd International Conference on Computing and Communications Technologies (ICCCT) (pp. 221-231).* IEEE. https://doi.org/10.1109/ICCCT2.2019.8824988

6. Shraddha, C., Chayadevi, M. L., & Anusuya, M. A. (2019, July). Noise cancellation and noise reduction techniques*: A review. In 2019 1st International Conference on Advances in Information Technology* (ICAIT) (pp. 159-166). IEEE. https://doi.org/10.1109/ICAIT47043.2019.8987262

7. Gonzalez-Lopez, J. A., Gomez-Alanis, A., Doñas, J. M. M., Pérez-Córdoba, J. L., & Gomez, A. M. (2020). Silent speech interfaces for speech restoration: A review. IEEE access, 8, 177995-178021. https://doi.org/10.1109/ACCESS.2020.3026579

8. Fenghour, S., Chen, D., Guo, K., Li, B., & Xiao, P. (2021). Deep learning-based automated lip-reading: A survey. IEEE Access, 9, 121184-121205. https://doi.org/10.1109/ACCESS.2021.3107946

9. Sooraj, V., Hardhik, M., Murthy, N. S., Sandesh, C., & Shashidhar, R. (2020). Lip-reading techniques: a review. *Int. J. Sci. Technol.* Res, 9, 4378-4383.

10. Hao, M., Mamut, M., Yadikar, N., Aysa, A., & Ubul, K. (2020). A survey of research on lipreading technology. *IEEE Access*, 8, 204518-204544. https://doi.org/10.1109/ACCESS.2020.3036865

11. Fenghour, S., Chen, D., Guo, K., & Xiao, P. (2020). Lip reading sentences using deep learning with only visual cues. *IEEE Access*, 8, 215516-215530. https://doi.org/10.1109/ACCESS.2020.3040906

12. Fernandez-Lopez, A., & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, 78, 53-72. https://doi.org/10.1016/j.imavis.2018.07.002

Olesia Barkovska **Олеся Барковська**	Ph.D., Associate Professor of Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine e-mail: olesia.barkovska@nure.ua; https://orcid.org/0000-0001-7496-4353 Scopus Author ID: 24482907700, https://scholar.google.ru/citations?hl=ru&user=Uj96xp4AAAAJ	доцент к.т.н., доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна
Vladyslav Kholiev **Владислав Холєв**	Department of Electronic Computers**,** Kharkiv National University of Radio Electronics, Kharkiv, Ukraine e-mail: vladyslav.kholiev@nure.ua https://orcid.org/0000-0002-9148-1561 https://scholar.google.com/citations?user=Y1mLOvsAAAAJ	асистент кафедри електрониих обчислювальних машин**,** Харківський національний університет радіоелектроніки, Харків, Україна