Artificial neural network technology for lips reading

Anna Pyataeva[1]* and *Anton* Dzyuba[1]

[1]Siberian Federal University, Institute of Space and Information Technology, 26, Kirensky st., Krasnoyarsk, 660074, Russian Federation

Abstract. The paper presents the use of neural networks for the task of automated speech reading by lips articulation. Speech recognition is performed in two stages. First, a face search is performed and the lips area is selected in a separate frame of the video sequence using Haar features. Then the sequence of frames goes to the input of deep learning convolutional and recurrent neural networks for speech viseme recognition. Experimental studies were carried out using independently obtained videos with Russian-speaking speakers.

1 Introduction

Visual speech recognition is a most important task when you communicating with hearing impaired people. According to the World Health Organization [1], as of 2020, more than 5% of all people in the world (360 million people) have serious hearing impairment problems. Hearing aids cannot fully solve the problem of these people because of hearing loss may be associated with neuropathology. There are diseases in which people lose the ability to make sounds. In such situations, speech recognition by lips articulation is one way to keep these people communication ability.

Speech recognition by articulation is based on the detection of visemes [2], i.e. mimic realizations of phonemes. There are 42 phonemes in Russian. 6 of them are vowels (а, и, о, у, ы, э) and 36 consonants (б, б', в, в', г, г', д, д', ж, з, з', э(й), к, к', л, л', м', м, н, н', п, п', р, р', с, с', т, т', ф, ф', х, х', ц, ч, ш, щ). Visemes and phonemes do not have a one-to-one correspondence. There are situations, when several phonemes correspond to one viseme and look the same on the speaking person face. The reason for this phenomenon is in fact that phonemes are reproduced inside the mouth or throat.

Traditionally, automatic lips reading (ALR) systems relied on the visual features extraction using hidden Markov models [4]. Currently, deep learning neural networks are used to solve the problems of automated recognition of various human actions (including lip reading) [3]. Speech recognition systems by articulation are used for dictating messages on smartphones [5], visual recognition of silent passwords [6], transcribing silent films [7], synthesizing the voice of people with speech impairments based on the movements of their lips [8], and developing systems for tracking the lips of the interlocutor for people with hearing impairments [9] and in other areas.

*Corresponding author: anna4u@list.ru

2 Algorithm for recognizing human speech by articulation

Lips speech recognition consists of two stages: face detection with lips highlighting and viseme recognition.

2.1 Face detection and lips area highlighting

To detect faces we used the OpenCV computer vision library [10] and the NumPy library [11]. The algorithm for face detection and lips area selection consists of the following steps.

1. Each frame of the video sequence is converted to grayscale since color does not play a significant role in detecting a face on the scene, but its presence negatively affects the execution speed of the algorithm.
2. Next step is detected a face in the examined frame using the Haar features [12] and then in the found area of the face algorithm is detected for lips.
3. The area with the detected lips is enlarged by 15 pixels on each side to prevent the lips from being cropped.
4. The lip areas obtained for each video sequence are stored in a NumPy data.

The resulting sequence of frames with images of lips is transmitted to a convolutional neural network for further processing.

2.2 Speech phoneme recognition by articulation

To recognize phonemes of speech by articulation, two neural networks are jointly used (Fig. 1): the MobileNet convolutional neural network (CNN) [13] and the recurrent neural network of the LTSM architecture [14]. The MobileNet architecture was chosen as the transfer learning model due to its light weight. The MobileNet convolutional neural network is used to normalize images. At the end of the CNN model is a fully connected layer, which output is used as input to the LSTM layer, so the ReLu function is chosen as the CNN activation function. The fully connected CNN layer to bring the data into the required dimension, has 1024, 128 neurons. After converting the videos using CNN, the video data looks like (586, 20, 128, 128, 3), where 586 is the total number of pronunciations for all tags, 20 is the number of frames for speaking, 128 and 128 are the width and height of the image, 3 is the number of color channels of the video image.

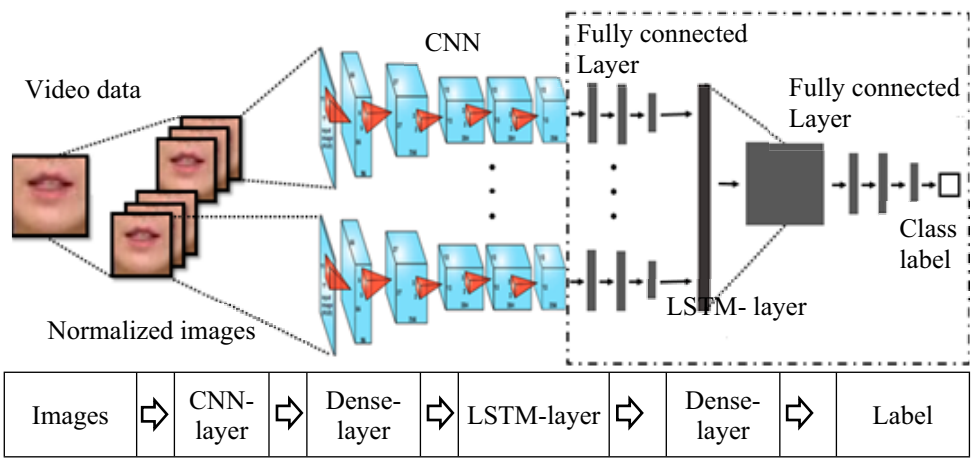

Fig. 1. Neural network architecture for speech recognition by articulation.

The LSTM-layer is used the TimeDistribute function supported by Keras. The TimeDistribute function takes ordered data and encodes its content. The CNN model is used as the object of coding. Combining CNN and LSTM models performed by using the Model function. Adam which has good performance was used as an optimizer. Since the model will perform the task of classification, the cross-entropy is chosen as the loss function.

The software implementation is performed with using the Keras deep learning library [15], which supports interaction with the Theano and Tensorflow packages. Keras functionality is very wide and has a convenient API, and its low speed is compensated by CUDA support. Keras is used to design, deploy and train a neural network, and to recognize speech visemes from a set of articulation frames.

3 Experimental research

Experimental studies used a dataset containing 768 different statements uttered by different speakers. The videos were obtained independently, as a dataset of Russian-speaking speakers was required. The utterances are labelled with the same labels as the training dataset. The training selection was 80%, the test selection was 20% of the total sample selection. The quality of the algorithm was evaluated using the "recognition accuracy" indicator. Test words were: «бегу» (in English means run), «пила» (in English means saw), «милый» in English means sweet), «усы» (in English means moustache), «вулкан» (in English means volcano), «банан» (in English means banana), «тонуть» (in English means drown). The best speech recognition accuracy by articulation was 93.7% for the word «банан» (banana) and the average accuracy was 68%, which suggests that the algorithm needs additional modification. Equations should be centred and should be numbered with the number on the right-hand side.

References

1. World Health Organization. https://www.who.int/
2. Speech Recognition: Watch what you say. https://www.econimist.com/ecience-and-technology/2015/01/23/watch-what-you-say
3. H. McGurk, J. MacDonald. Nature, 264, 7-18 (1976)
4. G. A. Fink. Berlin Heidelberg: Springer-Verlag. (2008)
5. A. Gabbay, T. Ephrat, S. Halperin, S. Peleg, *Seeing through noise: Speaker separation and enhancement using visually-derived speech,* in Proceedings of the International Workshop on Computer Vision for Audio-Visual Media, arXiv:1708.06767v1 (2017)
6. F. S. Lesani, F. F. Ghazvini, R. Dianat, *Mobile phone security using automatic lip reading* in Proceedings of the 9th International Conference on e-Commerce in Developing Countries: With focus on e-Business (ECDC), 16 April 2015, Isfahan, Iran (2015)
7. Y. M. Assael, B. Shillingford, S. Whiteson, N. De Freitas, *Lipnet: Sentence-level lipreading*, in Proceedings of the ICLR 2017 : 5th International Conference on Learning Representations, Toulon, France (2017)
8. A. Ephrat, T. Halperin, S. Peleg, *Improved speech reconstruction from silent video,* in Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW), 22-29 October, 2017, Venice, Italy (2017)
9. A. Britto Mattos, D. A. Mattos, *Multi-view mouth renderization for assisting lip-reading,* in Proceedings of the Conference the Internet of Accessible Things. (2018)
10. NumPy. https://numpy.org

11. OpenCV. https://opencv.org
12. P. Viola, M. J.Jones, *Rapid Object Detection using a Boosted Cascade of Simple Features,* in Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, Kauai, Hawaii (2001)
13. S. Hasim, W. Andrew, arXiv:1402.1128 (2014)
14. K. Yamaguchi, K. Sakamoto, T. Akabane, Y. Fujimoto *A Neural Network for Speaker-Independent Isolated Word Recognition,* in Proceedings of the First International Conference on Spoken Language Processing, 18-22 November, Kobe, Japan
15. Keras. https://keras.io