Journal of COMPUTER AND FORENSIC SCIENCES

CFS 2022, Vol. 1, Issue 1, pp. 7–16
https://doi.org/10.5937/1-42696
Original research paper

Application of Convolutional Neural Networks to Spoken Words Evaluation Based on Lip Movements without Accompanying Sound Signal

Dušan Perić[1], Nemanja Maček,[2*] and Mitko Bogdanoski[3]

[1] DP SOFTWARE, Belgrade, Serbia; dusan.peric98@gmail.com

[2] Academy of Technical and Art Applied Studies, School of Electrical and Computer Engineering, Belgrade, Serbia & University Business Academy in Novi Sad, Serbia & SECIT Security Consulting, Serbia; macek.nemanja@gmail.com

[3] Military Academy General Mihailo Apostolski, Skopje, NR Macedonia; mitko.bogdanoski@ugd.edu.mk

* Corresponding author: macek.nemanja@gmail.com

Received: 2022-11-14 • Accepted: 2022-12-05 • Published: 2022-12-30

Abstract: This paper proposes an approach to evaluate spoken words based on lip movements without accompanying sound signals using convolutional neural networks. The main goal of this research is to prove the efficiency of neural networks in the field, where all data is received from an array of images. The modeling and the hypotheses are validated based on the results obtained for a specific case study. Our study reports on speech recognition from only a sequence of images provided, where all crucial data and features are extracted, processed, and used in a model to create artificial consciousness.

Keywords: machine learning; convolutional neural networks; lip reading.

1. INTRODUCTION

For centuries, people have been searching for a solution to the problems that everyday life brings them. Today, human civilization increasingly relies on artificial intelligence systems. It has become our present and has great potential to change the world, improve people's lives in various areas, and respond to numerous social needs in education, medicine, agriculture, economy. A couple of decades ago, artificial intelligence was reserved only for sci-fi movies, and today there are great technological achievements that we employ every day, such as smartphones with applications that help us in our daily activities, smart homes and home appliances, electronic autonomous cars with various integrated intelligent systems, as well as numerous applications on the Internet. As the application of artificial intelligence grew, the challenges it solved became more and more complex. One such challenge is the software's ability to read people's lips and thus bring this rare skill closer to a large number of users.

Citation: D. Perić, N. Maček, and M. Bogdanoski, "Application of Convolutional Neural Networks to Spoken Words Evaluation Based on Lip Movements without Accompanying Sound Signal," *Journal of Computer and Forensic Sciences*, 1(1), https://doi.org/10.5937/1-42696.

People with impaired hearing, as well as people who work in a very noisy environment, most often communicate visually, using sign language or lip-reading techniques. For both methods of communication, the key role is played by the sense of sight, which accepts information and the movements of hands or lips and converts them into a series of images at a one-time interval.

Speech reading is a complex psychophysiological process in which three moments are important: the visual perception of oral movements, the kinesthetic memorization of speech movements, and the psychological act of recognizing a word in order to better understand it [1, 2].

Lip reading can best be described as the visual segmentation of words into time sequences nested within a time period. The application of lip reading can also be seen as an inevitable skill in all spy movies, as well as in some renowned television programs such as BBC News, where, with the help of lip reading techniques, artificial intelligence recognizes the speech from the lips and prints the subtitle of the speaker.

In this paper, we will demonstrate the creation of software that will solve the problem of lip reading with great precision and make this technique available to more users.

2. RELATED WORK

Aggravating circumstances when reading lips are a very common problem in people, and these are: limited movements of the articulator (jaw and lips), fast or slow speech, poor mimicry, specific head movements, inadequate distance, and poor lighting [1, 2]. Many researchers are trying to create the perfect system for lip reading. A lot of related scientific papers can be found addressing this topic. Several approaches related to lip reading are briefly addressed in this paper.

The authors [3] presented the method of detecting lips and using the cropped images as a dataset for the training set for Convolutional Neural Networks. Also, they discussed different methods of evaluation that can be used. In [4], an interesting approach to lip detection by skin color and Kalman image filtering used to eliminate noises and unwanted information from the picture is presented. In Artificial Intelligent systems, effectiveness is a crucial aspect because of the more frequent use of those systems in more and more responsive jobs that were reserved for people only until now. Zhao et al. [5] presented how mutual information maximization facilitates effective lip reading. Real-time speech detection is a hardware-consuming task that can be done with the OpenCV library, which is widespread in movement tracking software. WenJuan et al. [6] presented how lip movement can be detected and used for lip reading. Rahmani and Almasganj [7] showed us a hybrid system for lip reading that uses a combination of image-based and model-based features. A lot of researchers thought that lip reading was only possible with the en face pictures of the face. Saitoh and Konishi [8] presented the solution for training the lip reading system from profile pictures. A detailed explanation of backpropagation, which neural networks are using to learn lip reading, is given by Rathee in [9]. One of the hot topics in the software industry today is mobile applications development. The work of Matsunaga and Matsui [10] depicts how lip reading can be implemented on mobile devices.

3. MATERIALS AND METHODS

Neural networks are well suited for learning from large amounts of raw data, such as images or sequences of images. An image can be represented as a series of pixels, and the task of the neural network is to create useful information from that series of pixels by extracting the features with the greatest weight of information, such as the shape or the edges of the object in the image.

Algorithms based on neural networks have several advantages compared to standard algorithms, but they are dependent on human expertise in the field for which the system is being developed. Of course, all algorithms that use neural networks, due to their nature of using a large amount of information for training the system, have one major drawback, which is the need for very powerful and expensive hardware.

Neurons are the main component of deep learning systems. In them, calculations are performed on the input data that the neuron receives from the previous neuron, and the output parameters are forwarded as a result. Those output parameters are further supplied to the next neuron, except in the case when the neuron is in the first hidden layer. In that case, the input data is raw data on which only pre-processing has been performed.

"In a biological neuron, an electrical signal is given as an output when it receives a more influential input. To capture that functionality in a mathematical model of neuron, we need to have a function that operates on the sum of the input multiplied by the appropriate weights and responds with the appropriate value based on the input. If an input with a higher impact is received, the output should be higher, and if an input with a lower impact is received, the output should be lower, which can be represented as a switch. An activation function is a function that takes a combination of inputs, applies a function over them, and passes an output value." In this way, it tries to imitate the activation/deactivation function. Thus, the activation function determines the state of the neuron by computing the activation function on the combined input [11]. The three activation functions that are far commonly used than the others are "Sigmoid", "Tanh" and "ReLU", which are depicted in Figure 1.

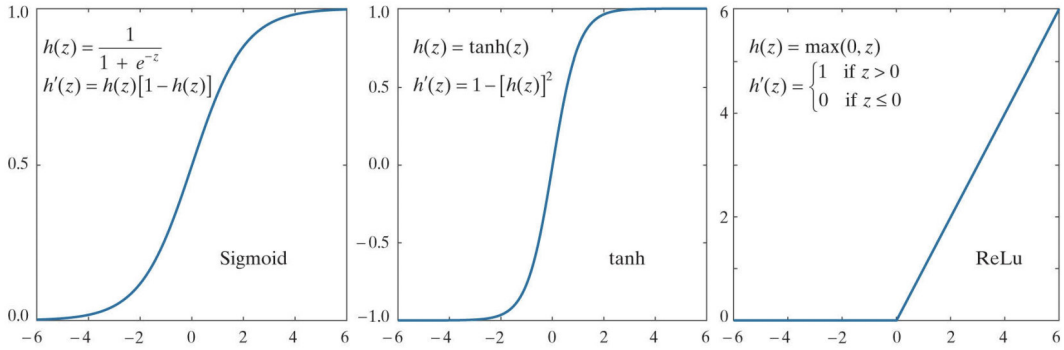

Figure 1. *Activation functions (Sigmoid,Tanh and ReLU). Adopted from [10].*

A neural network consists of one input and one output layer, as well as one or more hidden layers. Depending on the complexity of the task for which the neural network needs to create awareness, the hidden layers can vary.

3.1. Convolutional Neural Networks

Convolutional neural networks, often represented by the abbreviation CNN, are a special type of multilayer neural networks designed to recognize visual patterns directly from images with minimal use of computer resources. There are several ways to connect neurons to create a convolutional neural network. One of the ways is the feed-forward network method [12]. The method is based on the fact that neurons from each layer feed the next layer with their output values. This process is repeated until we get the final resulting values.

Each convolutional neural network consists of the three most important layers for its functioning, namely the convolutional layer, the pooling layer, and the fully connected layer. The task of the convolutional layer is to extract features from the image that is used as input information, so we can consider it the most important layer because it does most of the calculations in the convolutional neural network. The activation function that performs exceptionally well for CNN is ReLU because of its features that perfectly fit the description of the input data for CNN, which are images. The pooling layer selects an area from the image and then selects one representative value using a maximum or average pooling technique. The architecture of a convolutional neural network can be represented as several convolutional layers connected in a cascade style. In each cascade, there is a convolutional layer with a ReLU activation function, then a pooling layer, thus simulating cascades several times. Finally comes the fully connected layer. The output from each convolutional layer is a set of objects called feature maps, generated by a single convolutional filter. Feature maps can be used to define new input data for the next layer. Thus, convolutional neural networks work by extracting features from the image that is the input information by dragging a filter over it. The result of multiplying the values of the filter matrix and the image area of the same size gives the output values for the next convolutional layer. In this way, in each subsequent step, we get an increasingly precise contour of the object we are looking for.

4. RESULTS

The testing platform consisted of an "upgraded" Dell Precision laptop based on an Intel Ii7 processor with 64 GB RAM, 3.5 terabytes in solid state drives, and an external NVidia 3080 RTX graphic card. We have used common testing platforms like Python and various libraries for visualization and deep learning (like Tensorflow).

Figure 2 depicts a sequence of images that form one word, viewed from left to right and spoken by one person. The optimal image size here is not of any interest. The functions embedded will, according to our algorithm, crop the surface around lips, hence the size is inconsistent. Yet, a bit of preprocessing to crop it to a proper fixed-size rectangular image may provide some help (i.e., an improvement in accuracy, etc.) in certain scenarios.

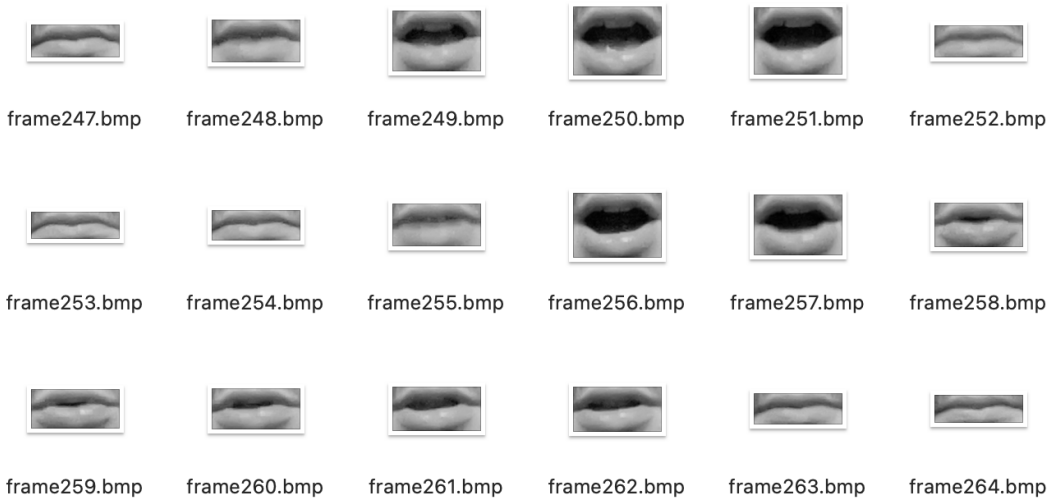

Figure 2. *Sequence of images used to represent one word. Images presented in Figure 2 and all others used in our experiments are generated from several databases [13–15].*

The final step in data preprocessing is data normalization, which we can do by nesting the range of pixel values between 0 and 1. This type of optimization is called "colormap normalization" and is used when pixel values need to be represented by values between 0 and 1.

```
[[192. 187. 186. ... 170. 180. 186.]
 [187. 187. 184. ... 165. 172. 181.]
 [185. 186. 182. ... 164. 169. 178.]
 ...
 [173. 170. 169. ... 162. 161. 159.]
 [170. 170. 169. ... 162. 160. 159.]
 [169. 169. 169. ... 161. 160. 159.]]

[[  0.   0.   0. ...   0.   0.   0.]
 [  0.   0.   0. ...   0.   0.   0.]
 [  0.   0.   0. ...   0.   0.   0.]
 ...
 [  0.   0.   0. ...   0.   0.   0.]
 [  0.   0.   0. ...   0.   0.   0.]
 [  0.   0.   0. ...   0.   0.   0.]]
```

Figure 3. *Data before normalization (data screenshot).*

```
[[0.81818182 0.79292929 0.78787879 ... 0.70707071 0.75757576
  0.78787879]
 [0.79292929 0.79292929 0.77777778 ... 0.68181818 0.71717172
  0.76262626]
 [0.78282828 0.78787879 0.76767677 ... 0.67676768 0.7020202
  0.74747475]
 ...
 [0.72222222 0.70707071 0.7020202  ... 0.66666667 0.66161616
  0.65151515]
 [0.70707071 0.70707071 0.7020202  ... 0.66666667 0.65656566
  0.65151515]
 [0.7020202  0.7020202  0.7020202  ... 0.66161616 0.65656566
  0.65151515]]

[[0.         0.         0.         ... 0.         0.
  0.        ]
 [0.         0.         0.         ... 0.         0.
  0.        ]
 [0.         0.         0.         ... 0.         0.
  0.        ]
 ...
 [0.         0.         0.         ... 0.         0.
  0.        ]
 [0.         0.         0.         ... 0.         0.
  0.        ]
 [0.         0.         0.         ... 0.         0.
  0.        ]]
```

Figure 4. *Data after normalization (data screenshot).*

After we have loaded, processed, and normalized our data, it is time to define a learning model that will create awareness from that data. For the purposes of this algorithm, a sequential model composed of layers that can be viewed as groups of neurons was used. The sequential model allows adding different types of layers on top of each other, creating a single stack.

The optimizer is the most important part of our model. Its functionality is to calculate and change the weights of the neurons so that the losses are as small as possible. The back-propagation process we mentioned earlier is actually an optimization algorithm. Two of the most well-known and widely used deep learning optimizers are Stochastic Gradient Descent (SGD) and Adam (Adaptive Moment Estimation).

For outcomes with categories, the prediction will be the name of the class, i.e., the category. In this algorithm, the results are the words that the system should recognize based on the movement of the lips, which indicates that the regressive standard will not be used but a categorical one.

In this research, categorical cross-entropy is used to calculate the losses because there are more than two possible prediction outcomes.

The results of model training through epochs are depicted in Figure 5.

```
Epoch 54/450
4/4 [==============================] - 1s 310ms/step - loss: 0.9456 - accuracy: 0.5833 - val_loss: 0.9692 - val_accuracy: 0.5000
Epoch 55/450
4/4 [==============================] - 1s 306ms/step - loss: 0.8478 - accuracy: 0.6333 - val_loss: 0.9602 - val_accuracy: 0.5000
Epoch 56/450
4/4 [==============================] - 1s 305ms/step - loss: 0.8521 - accuracy: 0.6333 - val_loss: 0.9561 - val_accuracy: 0.5667
Epoch 57/450
4/4 [==============================] - 1s 306ms/step - loss: 0.8979 - accuracy: 0.6667 - val_loss: 0.9713 - val_accuracy: 0.5667
Epoch 58/450
4/4 [==============================] - 2s 737ms/step - loss: 0.8018 - accuracy: 0.7333 - val_loss: 0.9234 - val_accuracy: 0.6333
```

Figure 5. *Model training through epochs (results screenshot).*

We can present the performance of our model visually through two graphs. Figure 6 depicts the accuracy of the model across epochs, while Figure 7 depicts the losses.

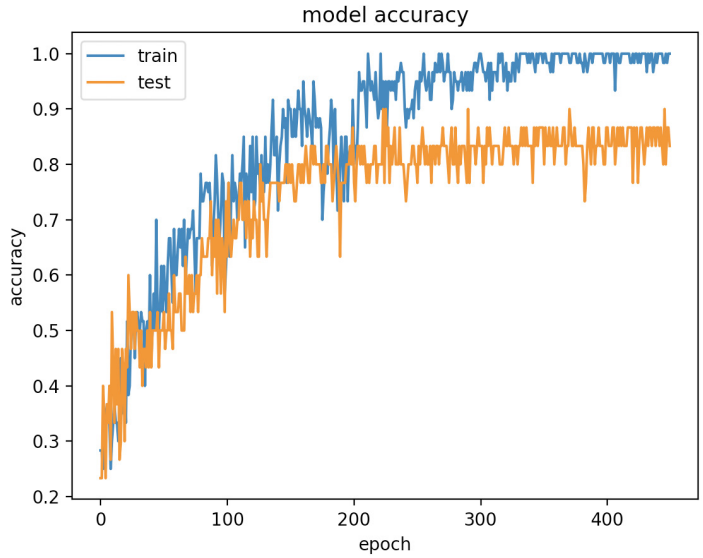

Figure 6. *Model accuracy.*

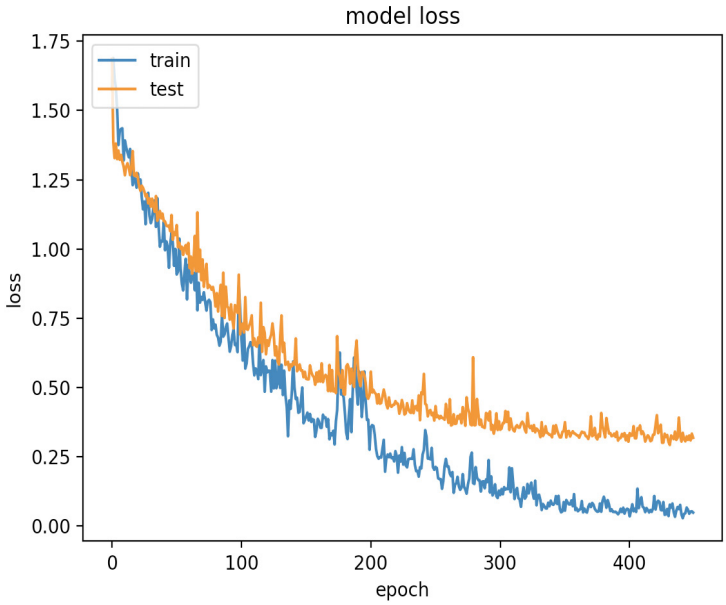

Figure 7. *Model losses.*

The effectiveness of our model is finally presented in the following screenshot:

```
Val:
2022-05-09 22:49:07.871541: W tensorflow/core/platform/profile_utils/cpu_utils.cc:128] F
2022-05-09 22:49:07.972552: I tensorflow/core/grappler/optimizers/custom_graph_optimizer
vice_type GPU is enabled.
1/1 [==============================] - 1s 874ms/step - loss: 0.3297 - accuracy: 0.8667
Test:
4/4 [==============================] - 1s 178ms/step - loss: 0.2320 - accuracy: 1.0000
Accuracy:1.0
```

Figure 8. *Model evaluation (results screenshot).*

5. DISCUSSION

Since there is almost no field of human activity where artificial intelligence is not used, modern man is required to master new knowledge and technologies. The purpose of the emergence of artificial intelligence is to facilitate the daily activities of people, which most often require mental fatigue, and to increase the performance of work in jobs that require the processing of a large amount of information. A human is capable of doing that kind of work for a few hours, but a machine, on the other hand, can do it non-stop for days or years.

The software we presented in this paper simulates human performance, which means that it performs the task at the same speed as a human. The main advantage is that the person has to take a break after some time and divide the work into several days, and the software will work non-stop until the task is completed. In this way, the skill of lip reading can be used much more effectively in long videos, as well as in cases where the video does not have a sound recording or the ambient noise is too great to record the sound with sufficient quality. The downside of the software is the same as with all complex neural networks, and that is the large training time of the model. As the number of categories that the model needs to predict increases, the training time also multiplies.

Readers may also consider the following articles for further information on the topic [16–19].

FUNDING:

This research received no external funding. Data used to generate lip images in Figure 2 is generated from publicly available datasets licensed under CC BY-NC 4.0 license.

INSTITUTIONAL REVIEW BOARD STATEMENT:

Not applicable.

INFORMED CONSENT STATEMENT:

Not applicable.

CONFLICTS OF INTEREST:

The authors declare no conflict of interest.

REFERENCES

[1] S. Pažin, L. Isaković, S. Slavnić, and M. Srzić, "Specifičnost čitanja govora sa usana kod gluvih i nagluvih učenika različitog uzrasta," In Specificity of hearing impairment – new trends, 2020, pp. 219–233.

[2] A. Živanović, I. Sokolovac, M. Marković, S. Suzić, and V. Delić, "Prepoznavanje reči u govornoj audiometriji," In Specificity of hearing impairment – new trends, 2020, pp. 97–111.

[3] Y. Xiao, L. Teng, A. Zhu, X. Liu, and P. Tian, "Lip Reading in Cantonese," *IEEE Access*, Vol. 10, 2022, pp. 95020–95029. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9878070

[4] T. Thein and K. M. San, "Lip localization technique towards an automatic lip reading approach for Myanmar consonants recognition," In 2018 International Conference on Information and Computer Technologies (ICICT), 2018, pp. 123–127.

[5] X. Zhao, S. Yang, S. Shan, and X. Chen, "Mutual information maximization for effective lip reading," In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 420–427.

[6] Y. WenJuan, L. YaLing, and D. MingHui, "A real-time lip localization and tacking for lip reading," In 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 2010, Vol. 6, pp. V6–363.

[7] M. H. Rahmani and F. Almasganj, "Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features," In 2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), 2017, pp. 195–199.

[8] T. Saitoh and R. Konishi, "Profile lip reading for vowel and word recognition," In 2010 20th International conference on pattern recognition, 2010, pp. 1356–1359.

[9] N. Rathee, "Investigating back propagation neural network for lip reading," In 2016 International Conference on Computing, Communication and Automation (ICCCA), 2016, pp. 373–376.

[10] Y. Matsunaga and K. Matsui, "Mobile Device-based Speech Enhancement System Using Lip-reading," In 2018 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), 2018, pp. 1–4.

[11] J. Moolayil, *Learn Keras for Deep Neural Networks: A Dust-Track Approach to Modern Deep Learning with Python*, 1st edition. Apress, 2018.

[12] M. Sewak, R. Karim, and P. Pujari, *Practical Convolutional Neural Networks: Implement advanced deep learning models using Python*. Packt Publishing, 2018.

[13] J. S. Chung and A. Zisserman, "Lip Reading in the Wild," In Asian Conference on Computer Vision, 2016. Available: https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16/chung16.pdf

[14] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," In IEEE Conference on Computer Vision and Pattern Recognition, 2017. Available:

https://openaccess.thecvf.com/content_cvpr_2017/papers/Chung_Lip_Reading_Sentences_CVPR_2017_paper.pdf

[15] IPSJ SIG-SLP Noisy Speech Recognition Evaluation WG (2011): Audio-Visual Speech Recognition Evaluation Environment (CENSREC-1-AV). Speech Resources Consortium, National Institute of Informatics. (dataset). https://doi.org/10.32130/src.CENSREC-1-AV

[16] S. Ren, Y. Du, J. Lv, G. Han, and S. He, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13325–13333.

[17] F. Eguchi, K. Matsui, Y. Nakatoh, Y. O. Kato, A. Rivas, and J. M. Corchado, "Development of Mobile Device-Based Speech Enhancement System Using Lip-Reading," In International Symposium on Distributed Computing and Artificial Intelligence, 2021, pp. 210–220.

[18] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with densely connected temporal convolutional networks," In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2857–2866.

[19] K. R. Prajwal, T. Afouras, and A. Zisserman, "Sub-word level lip reading with visual attention," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5162–5172.