JOURNAL OF BUSINESS IN THE DIGITAL AGE

Vol. 5 Issue 2 2022

eISSN: 2651-4737 *dergipark.gov.tr/jobda*

Research Article

LIP READING USING CNN FOR TURKISH NUMBERS

Hadı POURMOUSA[1,] | Üstün ÖZEN[2]

1 Doktora Öğrencisi, Yönetim Bilişim Sistemleri, Atatürk Üniversitesi, Erzurum, Türkiye, pourmousahadi@gmail.com, ORCID: 0000-0001-6713-5872

2 Prof. Dr., Yönetim Bilişim Sistemleri Bölümü, İktisadi ve İdari Bilimler Fakültesi, Atatürk Üniversitesi, Erzurum, Türkiye, uozen@atauni.edu.tr, ORCID: 0000-0002-7595-4306

Article Info:
Received April 12, 2022
Revised August 03, 2022
Accepted: September 02, 2020

Keywords:
Lip reading,
Turkish language,
deep learning,
convolutional neural networks

Anahtar Kelimeler:
Dudak okuma,
Türkçe dili, d
erin öğrenme,
Evrişimli sinir ağları

DOI: 10.46238/jobda.1100903

ABSTRACT

Recently, lip reading has become one of the most important fields of study in the field of artificial intelligence. In this study, lip reading process was performed in Turkish language using convolutional neural networks (CNNs). For this purpose, people were asked to record the numbers video (61 video), and 9 video also collected from YouTube. The dataset was collected for 20 numbers. In this study, only the video was used and the sounds were completely removed. Due to the small dataset, it was tried to reproduce with different methods. The model was trained on the train dataset and 56.25% success was achieved on the test dataset.

ÖZET

Son zamanlarda, dudak okuma, yapay zeka alanında en önemli çalışma alanlarından biri haline gelmiştir. Bu çalışmada, evrişimli sinir ağları kullanılarak Türkçe dilinde dudak okuma işlemi gerçekleştirilmiştir. Bu kapsamda kişilerden sayıların videosunu (61 video) çekip göndermeleri istenmiş ve onun yanı sıra YouTube'dan 9 video toplanmıştır. Veri seti 20 sayı için toplanmıştır. Bu çalışmada sadece video kullanılmış ve sesler tamamen çıkarılmıştır. Veri setinin küçük olması nedeniyle farklı yöntemlerle çoğaltılmaya çalışılmıştır. Model eğitim veri setin üzerinde eğitilmiş ve test veri setinde %56,25 başarı sağlanmıştır.

[1] **Corresponding Author,**
E-mail: pourmousahadi@gmail.com (H. POURMOUSA)

1| INTRODUCTION

In recent years, human action recognition is one of the most important fields of study in computer vision. One of the most important areas of human action recognition is lip reading (Agrawal & Omprakash, 2016, s. 753). Automated lip reading, which has begun to be used in various applications is the process of reading or interpreting of lip movement and recognize speech without audio stream (Agrawal & Omprakash, 2016, s. 753; Garg and et al, 2016, s. 1; Li and et al, 2016, s. 1; Martinez and et al, 2020, s. 6319; Ozcan & Basturk, 2019, s. 195). Lip reading becomes even more important in noisy conditions where audio speech cannot be understood (Garg and et al, 2016, s. 1; Martinez and et al, 2020, s. 6319; Noda and et al, 2014, s. 1149). Therefore, it is of great importance to this approach that will help people who cannot hear each other in noisy environments or hearing-impaired people (Li and et al, 2016, s. 1; Yargıç & Doğan, 2013, s. 1).

Automatic lip-reading can be used in many different areas, such as, decoding multi-speaker simultaneous speech, transcribing and re-dubbing of silent films in the archive, security, biometric identification, and dictating instructions or messages to a phone in a noisy environment (Agrawal & Omprakash, 2016, s. 753; Chung et al, 2017, s. 3444; Faisal & Manzoor, 2018, s. 1). But there are many challenges to be met in the field of lip reading. The most important challenges are: (Faisal & Manzoor, 2018, s. 1)

- Lack of context
- Extraction of spatio-temporal features
- Some people are more expressive with their lips, while others are not (person who cannot speak visually)
- Generalization among speakers
- Guttural sounds (like K and G consonants)
- Humming sounds

In this study, a dataset consisting of Turkish numbers was classified using a designed convolutional neural network.

2| RELATED WORKS

There has been no lip-reading study so far with deep learning on Turkish language. However, some studies have been carried out in different languages.

Garg et al in their paper propose various methods to lip reading in English language video without any audio signal. In the study, lip region frames were extract from the MIRACL-VC1 dataset and arranged in an image in a sequential manner. So, all the frames of a word have been lined up in one image. In this study, 56% accuracy in only words, 33% accuracy in only phrase and 44.5% accuracy in both were obtained in the test dataset (Garg and et al, 2016, s. 1-8). Li et al in their paper propose novel lip-reading method in Japanese language. There are 216 words from 2620 male in their dataset. They obtained 71.76% accuracy in their model (Li and et al 2016, s. 1-5). Faisal and Manzoor in their paper propose a method to lip-reading in Urdu language. There are 1000 videos for words and 1000 phrases in their dataset. They obtained 62% accuracy for words and 72% accuracy for digits in LSTM based model (Faisal & Manzoor, 2018, s. 1-5). Petridis et al in their paper propose lip-reading method in English language. They used OuluVS2 and CUAVE datasets in their study. They obtained 84.5% accuracy for OuluVS2 dataset and 78.6% accuracy for CUAVE dataset in their model (Petridis and et al, 2017, s. 2592-2595). Yargıç and Doğan develop a model to recognize color names in Turkish language. They used KNN classifier with Manhattan and Euclidean distances for their model. They obtained 78.22% accuracy with all angles of lip and 72.44% accuracy with best four angles of lip (Yargıç & Doğan, 2013, s. 1-5). Chen et al in their paper propose lip-reading method in Chinese language for words. Their proposed model is composed of a 3D convolutional layer with DenseNet and residual Bi-LSTM. They obtained 53.11% accuracy with ResNet+Bi-LSTM model, 62.57% accuracy with SqueezeNet+Bi-LSTM, 57.43% accuracy with VGG+Bi-LSTM model and 53.11% accuracy with DenseNet+Bi-LSTM (Chen and et al, 2020, s. 981-988). Elrefaei et al in their paper propose lip-reading method in Arabic language for 10 daily communication words. Their dataset contains 1100 videos from 22 speakers that recorded using smartphones' cameras. They obtained 79% accuracy in all words and 84.5% accuracy in "Thank you" word (Elrefaei and et al, 2019, s. 400-408).

DATASET

The dataset used in this section has been defined and some of its properties and pre-processes are explained.

Dataset Properties

The dataset used in this study consisted of Turkish numbers. The numbers in the dataset are presented in table 1. There are not any datasets for Turkish number. Therefore, we created the dataset ourselves. People were asked to record video of these numbers by smartphones camera. 61 people recorded and sent videos. In addition, 9 videos were collected from channels that provide Turkish learning on YouTube.

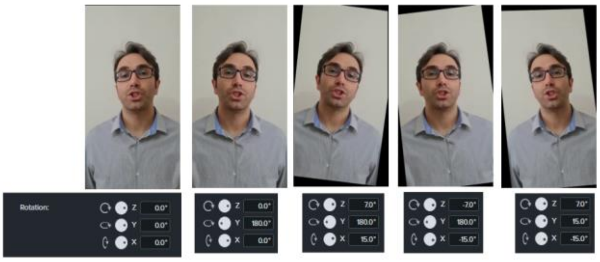
Table 1: Turkish Numbers Data Set

#	Number in Turkish	Number in English	#	Number in Turkish	Number in English
1	Sıfır (0)	Zero	11	On (10)	Ten
2	Bir (1)	One	12	Yirmi (20)	Twenty
3	İki (2)	Two	13	Otuz (30)	Thirty
4	Üç (3)	Three	14	Kırk (40)	Forty
5	Dört (4)	Four	15	Elli (50)	Fifty
6	Beş (5)	Five	16	Altmış (60)	Sixty
7	Altı (6)	Six	17	Yetmiş (70)	Seventy
8	Yedi (7)	Seven	18	Seksen (80)	Eighty
9	Sekiz (8)	Eight	19	Doksan (90)	Ninety
10	Dokuz (9)	Nine	20	Yüz (100)	One Hundred

Data Pre-processing

First of all, each number was recorded in separate videos with the Camtasia application, and because the dataset was small the videos were rotated at 3 different angles and 5 videos were obtained from each video (fig 1).

Figure 1: 5 Videos of Number 60 with Rotation at Different Angles



Rotation: Z 0.0° Y 0.0° X 0.0°

Z 0.0° Y 180.0° X 0.0°

Z 7.0° Y 180.0° X 15.0°

Z -7.0° Y 180.0° X -15.0°

Z 7.0° Y 15.0° X -15.0°

First, video of each number opened by OpenCV and lip regions were detected and extracted by mediapipe library in every frame of video. Then each frame of lip regions resized to 64*64. (fig 2). Then all frames of each number were sequentially added in one image and only one image was obtained for each number (fig 3).

Figure 2: All Frames of Number 60 (64*64)

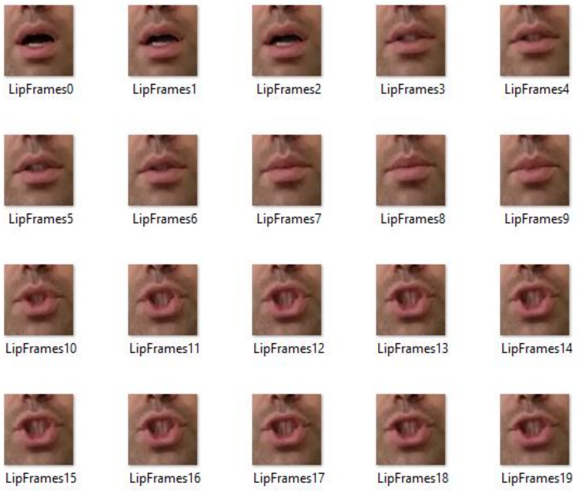
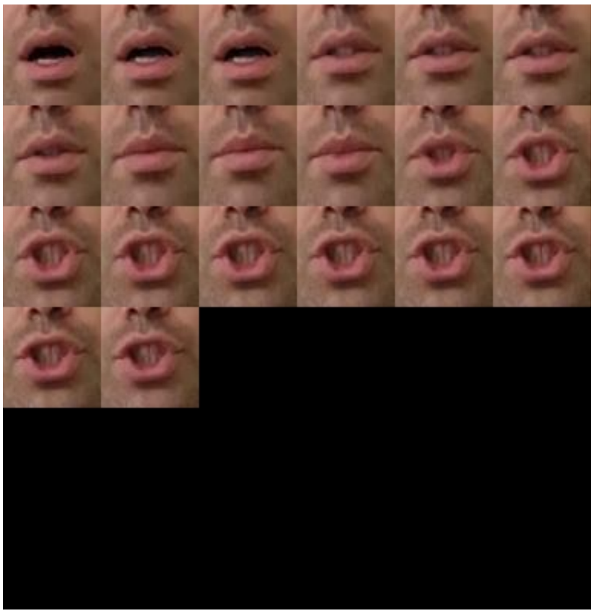


Figure 3: All Frames of Number 60 in One Image (384*384)

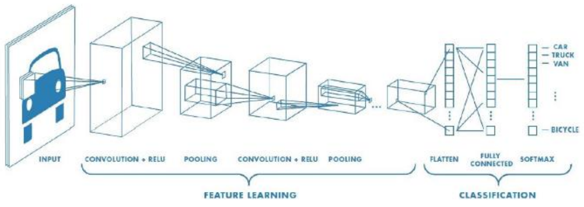


There are 300 instances in training dataset, 25 instances in validation dataset and 25 instances in test dataset.

3.| METHOD

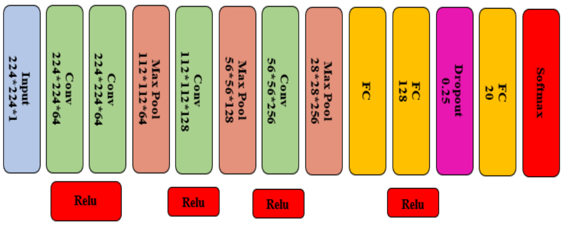
In this section, the method used to solve the lip-reading problem using Convolutional neural networks, is explained. CNN is a feed forward artificial neural network proposed by LeCun et al. (1989) for processing data consisting of matrices (multidimensional arrays) such as images and videos (Gu et al., 2018). In the following years, it has been used for different natural language processing tasks such as sentence modeling (Kalchbrenner, Grefenstette, & Blunsom, 2014) and sentence classification (Güven, 2019). Then, convolutional neural networks showed great success in computer vision subjects like object detection, classification and recognition. The initial function of a deep convolutional neural network is feature extraction and classification or regression. Convolutional neural network consists of two main functions, convolution and pooling layers to extract features; where multiple convolution and pooling layers are used to extract high-level features, and the output of each layer used as the input for the next layer. The extracted high-level features are then classified using fully connected layers (Ryczko et al., 2018; Karim, 2018). A convolutional neural network consists of some important parameters and functions such as convolution layer, pooling and fully connected layer, as seen in Figure 4. In the Convolution layer, the image is separated into subregions of a certain size and scanned by multiple filters. There are various activation functions and choosing among them depends on the model and this choice significantly affects the performance of the CNNs (Gu et al., 2018). Rectified Linear Unit (Relu) is the most commonly used activation function as it only passes positive values (Mishkin, Sergievskiy, & Matas, 2017). The pooling layer is placed between the convolution layers in the convolutional neural network structure. Pooling is a function to reduce the number of parameters and computations in the neural network and then reduce the spatial size of the data to control overfitting (Tran, Iosifidis & Gabbouj, 2018; Wu & Zhao, 2018). In deep learning architectures, pooling is used, after convolution, to combine all similar features in one space (Ramadan, 2019). Generally, maximum pooling and average pooling are used as two types of pooling methods (Karim, 2018). The fully connected layer is like a traditional neural network, but is used in deep learning after subtracting the decreasing important features and feature count (input data) in the final layer. In the fully connected layer, every node in the different layers is interconnected (Guo et al., 2018; Karim, 2018).

Figure 4: Convolutional Neural Network architecture (Ozcan & Basturk, 2019)



Convolutional neural networks were used because they presented great success in image processing and this study is image-based. Since the images are only the lip area, there are not many features, therefore they are converted to grayscale. For this reason, the photos will be given to the system in one dimension instead of 3 dimensions, so the system will learn more easily and faster. The model used in this paper developed by the user (fig 5).

Figure 5: Designed CNN Model



4| RESULT

The model presented in figure 4 was trained on our train dataset. Adam optimizer was used to train the model and the training was run 50 epochs. Training and validation accuracy and Training and validation loss shows in figure 6. The training accuracy was 95% and the validation accuracy was 91%. The reason for the high validation accuracy is that they are selected from the data rotated at different angles. However, data that are not found in the training and validation dataset were used in the test set. Therefore, the test accuracy was obtained as %56.25 (Fig 7). The reason for the low-test accuracy is that the dataset is very small and most of the data is obtained by rotating the same data in different angles. In deep learning, the large training set plays a very important role for the system to be successful. Confusion matrix figures for designed CNN model are shown as in Figure 8.

Figure 5: Training and Validation Accuracy and Training and Validation Loss

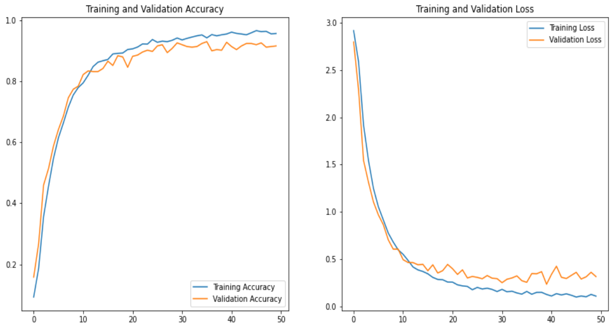
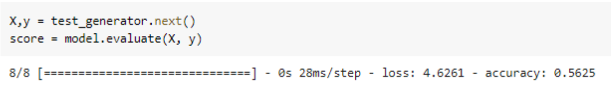


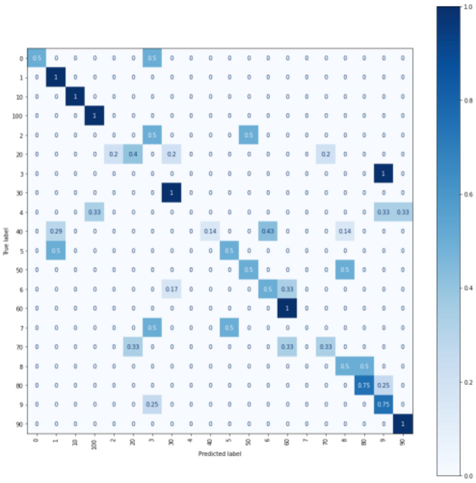
Figure 6: Test Accuracy



```
X,y = test_generator.next()
score = model.evaluate(X, y)

8/8 [==============================] - 0s 28ms/step - loss: 4.6261 - accuracy: 0.5625
```

Figure 7: Confusion Matrix Figures for Designed CNN Model



4| CONCLUSION

In this study, it was tried to design a visual lip reading system for Turkish Language. In this framework, a model based on CNN is developed and trained with the training dataset. Although the training set was small, 56.25% success was achieved. When the results were examined, the system guessed the number 3 completely wrong and showed the number 9 as the answer. The reason for this is that the lip movements of these two numbers are similar to each other. Likewise, 25% of the number 9 is estimated as the number 3. Also, the number 4 was guessed 100% incorrectly and showed the numbers 100, 9 and 90 as the answer. The reason for this is that the lip movements of 4 number is similar to this numbers. Also, data augmentation did not improve the result and the vgg16 model was used as transfer learning but achieved 5% success.

For future works, Turkish words and Turkish word groups dataset can be collected and the system can be trained and the results can be compared with numbers dataset. Also, other pre-trained models can be run on the number datasets.

REFERENCES

Agrawal, S., & Omprakash, V. R. (2016, July). Lip reading techniques: A survey. In 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) (pp. 753-757). IEEE.

Chen, X., Du, J., & Zhang, H. (2020). Lipreading with DenseNet and resBi-LSTM. Signal, Image and Video Processing, 14(5), 981-989.

Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017, July). Lip reading sentences in the wild. In 2017 IEEE conference on computer vision and pattern recognition (CVPR) (pp. 3444-3453). IEEE.

Elrefaei, L. A., Alhassan, T. Q., & Omar, S. S. (2019). An Arabic visual dataset for visual speech recognition. Procedia Computer Science, 163, 400-409.

Faisal, M., & Manzoor, S. (2018). Deep learning for lip reading using audio-visual information for urdu language. arXiv preprint arXiv:1802.05521.

Garg, A., Noyola, J., & Bagadia, S. (2016). Lip reading using CNN and LSTM. Technical report, Stanford University, CS231 n project report.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Chen, T. (2018). Recent advances in convolutional neural networks. Pattern Recognition, 77, 354-377.

Guo, Y., He, Y., Song, H., He, W., & Yuan, K. (2018). Correlational examples for convolutional neural networks to detect small impurities. Neurocomputing, 295, 127-141.

Güven, F. (2019). Using Text Representation And Deep Learning Methods For Turkish Text Classification. (Master's Thesis), Adana: Çukurova University.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188.

Karim, A. M. (2018). A New Framework By Using Deep Learning Techniques For Data Processing. (PhD dissertation), Ankara: Ankara Yıldırım Beyazıt University.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

Li, Y., Takashima, Y., Takiguchi, T., & Ariki, Y. (2016, June). Lip reading using a dynamic feature of lip images and convolutional neural networks. In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (pp. 1-6). IEEE.

Martinez, B., Ma, P., Petridis, S., & Pantic, M. (2020, May). Lipreading using temporal convolutional networks. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6319-6323). IEEE.

Mishkin, D., Sergievskiy, N., & Matas, J. (2017). Systematic evaluation of convolution neural network advances on the imagenet. Computer vision and image understanding, 161, 11-19.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2014). Lipreading using convolutional neural network. In fifteenth annual conference of the international speech communication association, 1149-1153.

Ozcan, T., & Basturk, A. (2019). Lip reading using convolutional neural networks with and without pre-trained models. Balkan Journal of Electrical and Computer Engineering, 7(2), 195-201.

Petridis, S., Li, Z., & Pantic, M. (2017, March). End-to-end visual speech recognition with LSTMs. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 2592-2596). IEEE.

Ramadan, M. (2019). Video Analysis By Deep Learning (Doctoral dissertation), University of Pittsburgh.

Ryczko, K., Mills, K., Luchak, I., Homenick, C., & Tamblyn, I. (2018). Convolutional neural networks for atomistic systems. Computational Materials Science, 149, 134-142.

Tran, D. T., Iosifidis, A., & Gabbouj, M. (2018). Improving efficiency in convolutional neural networks with multilinear filters. Neural Networks, 105, 328-339.

Wu, H., & Zhao, J. (2018). Deep convolutional neural network model based chemical process fault diagnosis. Computers & chemical engineering, 115, 185-197.

Yargıç, A., & Doğan, M. (2013, June). A lip reading application on MS Kinect camera. In 2013 IEEE INISTA (pp. 1-5). IEEE.