Research Article GU J Sci 37(3): x-x (2024) DOI: 10.35378/gujs.1239207

Gazi University

Journal of Science

http://dergipark.gov.tr/gujs

# Lip Reading Using Various Deep Learning Models with Visual Turkish Data

Talya TUMER SIVRI[1], Ali BERKOL[1*], Hamit ERDEM[2]

[1]*BITES Defense and Information Systems, 06530, Ankara, Turkey*
[2]*Baskent University, Electrical Electronics Engineering Department, 06790, Ankara, Turkey*

## Highlights

- This paper focuses on the classification of daily words and phrases in Turkish.
- The dataset used in this work was initially created to solve and use in related problems.
- Different models developed and compared.
- Image augmentation techniques are used to improve the data.

**Article Info**

*Received: 19 Jan 2023*
*Accepted: 15 Nov 2023*

**Keywords**

*Visual lip reading*
*Turkish dataset*
*Deep learning*
*Image augmentation*
*HCI*

**Abstract**

In Human-Computer Interaction, lip reading is essential and still an open research problem. In the last decades, there have been many studies in the field of Automatic Lip-Reading (ALR) in different languages, which is important for societies where the essential applications developed. Similarly to other machine learning and artificial intelligence applications, Deep Learning (DL) based classification algorithms have been applied for ALR in order to improve the performance of ALR. In the field of ALR, few studies have been done on the Turkish language. In this study, we undertook a multifaceted approach to address the challenges inherent to Turkish lip reading research. To begin, we established a foundation by creating an original dataset meticulously curated for the purpose of this investigation. Recognizing the significance of data quality and diversity, we implemented three robust image data augmentation techniques: sigmoidal transform, horizontal flip, and inverse transform. These augmentation methods not only elevated the quality of our dataset but also introduced a rich spectrum of variations, thereby bolstering the dataset's utility. Building upon this augmented dataset, we delved into the application of cutting-edge DL models. Our choice of models encompassed Convolutional Neural Networks (CNN), known for their prowess in extracting intricate visual features, Long-Short Term Memory (LSTM), adept at capturing sequential dependencies, and Bidirectional Gated Recurrent Unit (BGRU), renowned for their effectiveness in handling complex temporal data. These advanced models were selected to leverage the potential of the visual Turkish lip reading dataset, ensuring that our research stands at the forefront of this rapidly evolving field. The dataset utilized in this study was gathered with the primary objective of augmenting the extant corpus of Turkish language datasets, thereby substantively enriching the landscape of Turkish language research while concurrently serving as a benchmark reference. The performance of the applied method has been compared regarding precision, recall, and F1 metrics. According to experiment results, BGRU and LSTM models gave the same results up to the fifth decimal, and BGRU had the fastest training time.



## 1. INTRODUCTION

Lip reading is a natural ability that allows humans to understand what people are saying without hearing the important part of speech in society. It is primarily performed by looking at the shape of the mouth and the shape of the sounds created by the person's lips, teeth, and tongue. Also, it can be expressed as a classification of patterns that naturally occur while speaking. It is possible to make them learn this ability on computers with the help of deep learning algorithms using the visual data features of natural speech patterns. Moreover, lip reading is an open problem for different languages and a challenging task because we need not only knowledge of the underlying language but also visual clues to predict spoken words.

*Corresponding author, e-mail: ali.berkol@bites.com.tr

Visual speech information is critical when voice data is noisy, difficult to acquire, or lacking context. People find it challenging to understand what someone is saying merely by watching their mouth motions [1]. For instance, adults who are deaf or hard of hearing only achieve an accuracy of approximately 17% for a limited sample of 30 monosyllabic words and approximately 21% for 30 complicated [2].

In addition to understanding or recognizing the words and phrases by the listener as a research question, lip reading can be applicable to many areas in the industry, such as information security [3, 4], speech recognition [5, 6, 7], and driver assistance [8]. Moreover, it gives people with hearing problems a new way to communicate with the outside world [9, 10]. Regular people who do not have hearing problems can also benefit from lip reading in settings where speaking aloud is improper, such as a meeting room [10]. Lip reading has recently been used as a novel biometric identification method for mobile devices [11, 12]. As a result, lip reading and its applications are inseparable from society.

Lip reading models that use multi-model data are widely used in the research field, e.g., [13] and [14]. Despite the advantages of working with multi-model data, there are significant drawbacks, such as separating noise from data captured from crowded environments and requiring higher data storage, which also affects the model training efforts. Furthermore, even while voice-image-based lip reading has shown its usefulness, only-image-based lip reading also demonstrated promising results [15, 16]. However, a challenging problem, distinguishing similar lip movements for different words or phrases, reveals itself when the dataset contains only-image data. Distinguishing sounds with similar lip movements is a challenging problem. Additionally, since the algorithm can handle one person's data, it can be challenging to decide who is talking and whom the algorithm will take into account when there is more than one person on the camera screen. However, it is still easier to preprocess image data.

In this work, the Turkish lip reading model is trained and tested on only-image based dataset to increase the classification success rate for various deep learning models, which are Convolutional Neural Networks (CNN), Long-Short Term Memory (LSTM), and Bidirectional Gated Recurrent Units (BGRU). In acknowledgment of the paramount significance attributed to data quality and diversity within our research, we followed a meticulous strategy. Three robust image data augmentation techniques namely, the sigmoidal transform, horizontal flip, and inverse transform, were systematically integrated into our methodology. Beyond the enhancement of data quality, these techniques played a crucial role in enriching the dataset's versatility and utility. This augmented dataset served as the bedrock for our subsequent aim. With an augmented dataset that now encapsulated a broader scope of real-world scenarios and intricacies, we embarked on an exhaustive exploration of DL models. Our aim was to harness the full potential of the latest DL advancements. This phase of our study was characterized by a resolute push towards the vanguard of research, where we actively sought to capitalize on the most recent DL innovations in order to extract profound insights and knowledge from our enriched dataset. The following sections cover the data preprocessing stages and the modeling experiments in detail.

## 2. RELATED WORK

Artificial Intelligence (AI) researchers have recently become interested in the lip reading problem. Each language has a different structure since lip reading is sensitive in terms of language and sound. Because of that reason, there are various works for some languages [17-19]. Additionally, many state-of-the-art studies are available in terms of data types and languages. Some important models and approaches are as follows.

Conventional approaches typically rely on handcrafted features, which are too complicated and time-consuming to train neural networks [18]. The images are converted into numerical features that can be fed into deep learning algorithms for classification. Both visual and sound data were used to train the model, a combination of a spatiotemporal convolution layer and SE-ResNet-18 network with a BGRU back-end, 1D convolutional layer, and fully connected layers performed on the Daily Mandarin Conversation Lip Reading dataset [19].

The tiny and intricate signal patterns created by mouth motion are well captured by the data collection approach developed in [20]. The authors also suggest a set of algorithms to extract signal profiles linked to

mouth motions and reduce interference factors like multi-path. Then, to improve the recognition accuracy at the word level, a carefully crafted set of features, including time-domain statistical and frequency-domain features, are retrieved from the signal. A transfer-learning-based strategy is utilized to improve the model's robustness in cross-environment situations and lower training costs when employed in a new environment. [21] suggests a network with channel-temporal attention and deformable 3D convolution, where channel-temporal attention takes advantage of the inherent correlation of features to force the network to focus on necessary keyframes and deformable 3D convolution adapts the sample position adaptively based on the lip architecture.

[22] proposes a complete Bayesian learning approach to account for the underlying uncertainty in LSTM-RNN and Transformer Language Models (LMs). LSTM-RNN or Transformer LMs are used to model the uncertainty surrounding their model parameters, choice of neural activations, and hidden output representations. In order to automatically choose the best network internal components for Bayesian learning utilizing neural architecture search, efficient inference methods were applied. Additionally, a minimum of one sample of a Monte Carlo parameter was used. These make it possible to reduce the computing expenses associated with Bayesian NNLM training and evaluation.

[15] is a valuable survey for contrasting different approaches concentrating on neural networks and feature extraction. The authors' key finding is that Attention-Transformers and Temporal Convolutional Networks benefit from Recurrent Neural Networks. They concentrate on both audiovisual and merely visual information. Additionally, they mentioned letter-based, word-based, and sentence-based approaches that applied to English, Chinese, German, and Arabic, among other languages. From a different perspective, data augmentation techniques such as "salt and paper", "gaussian", and "speckle" noise adding, and "median" filtering were used to increase the dataset size [23]. Moreover, they used AlexNet and GoogleNet pre-trained CNNs on the AvLetters dataset.

[24] stated that they use digits or letters and words or sentences as targets for the problem. They developed an end-to-end algorithm dominated by RNNs, and achieved approximately 40% advancement in the word prediction rates. The algorithm from [25] only uses visual signals and lacks language. Visemes in continuous speech are recognized using a uniquely developed transformer with a unique topology. The use of perplexity analysis to translate visemes into words. Authors 15% decreased word error rate and enhanced performance. The model uses spatiotemporal CNNs, RNNs, and the connectionist temporal classification (CTC) loss [26] and operates at the character level. The public sentence-level dataset GRID Corpus, published in [27], was used for experiments. Another model designed for improved speed and accuracy is LipType [16]. In this work, poor light conditions are taken into consideration. As a first step, a spatial-temporal feature extraction technique was applied, which includes a correction for facial landmarks using Kalman filtering, 3D-CNN, and 2D SE-ResNet. Following that, Bidirectional Gated Recurrent Neural Network with CTC was used.

## 3. METHODOLOGY

### 3.1. Dataset

Numerous works use multi-type data for different languages like German, English, Urdu, etc. There is a dataset that has been published recently in Turkish. This dataset [28] contains five daily words and phrases which have only-visual information. In our experiments, we used six classes of the visual lip reading dataset in Turkish, which are three word classes "selam" (hi), "merhaba" (hello), "günaydın" (good morning), and three phrase classes which are "hoş geldiniz" (welcome), "özür dilerim" (sorry), and "teşekkür ederim" (thank you). We use this dataset because, unlike many others, it contains images taken from real-world conditions that simulate the real lip reading challenges. There are no fixed light conditions, right angles, or places in the real world. Since the dataset we used was captured from the Youtube [29] videos, it has various simulations such as men/women, mustache/make-up, inside/outside, close/far from the camera, and different angles between camera and lip. It is important to mention that the dataset is a family member of the Ural-Altaic language family, which has a narrow scope of application and research.

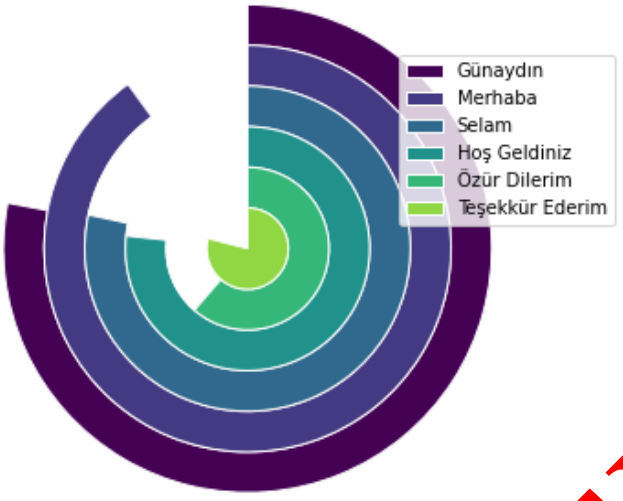


***Figure 1.*** *Data size for each class*

Dataset's every class has approximately equal data number (see Figure 1, i.e., each class's radial bar plot ending with a close radian and see Table 1 for an exact size.) It is essential to mention that the dataset instance size is not equal to the version taken from [28] since we applied some necessary preprocessing steps to solve the lip reading problem with DL algorithms. The steps are explained in the following sections in detail. Also, we have more data examples for some classes since we added some noisy examples from our local data storage. Data will be updated as a new version. Additionally, we observe that the data sequence length for each data sample depends on the length of the words and phrases. It can be concluded that the word or phrase length and frame number are highly correlated. Also, since the speakers are collected from a wide range of people, the dataset's classes are right-skewed, such as "merhaba" and "selam" which shows the speaker's speech speed differs. Besides the length of each class, we observed that data distributions according to frame lengths of each class have mostly normal distribution, and some classes have outliers, see Figures 2 and 3. More specifically, "merhaba" has a right-skewed distribution and outliers on the maximum side; see Figure 2. Since the dataset gathered different speaker resources, it is a natural result. Additionally, "selam" also has outliers, and "selam",and "günaydın" have a normal distribution. Similarly, we observe that phrases have normal distributions and they have outliers as well; see Figure 3.

***Table 1.*** *Size of the each class in the dataset*

| Class | Number |
|---|---|
| günaydın | 234 |
| selam | 235 |
| merhaba | 270 |
| hoş geldiniz | 230 |
| özür dilerim | 184 |

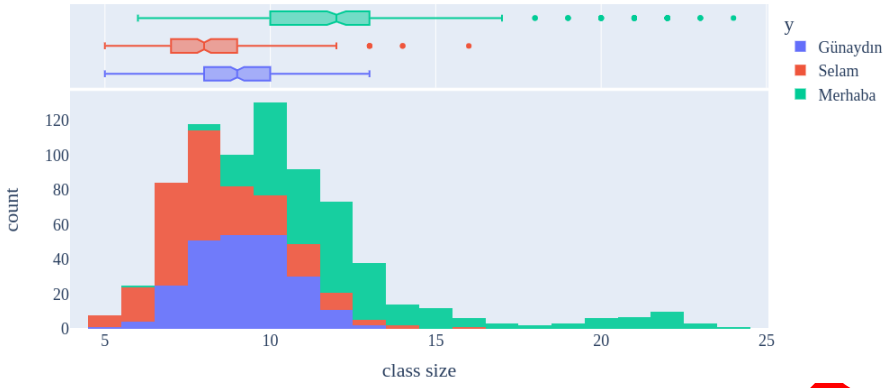


***Figure 2.*** *Data distributions according to frame length for each word class*

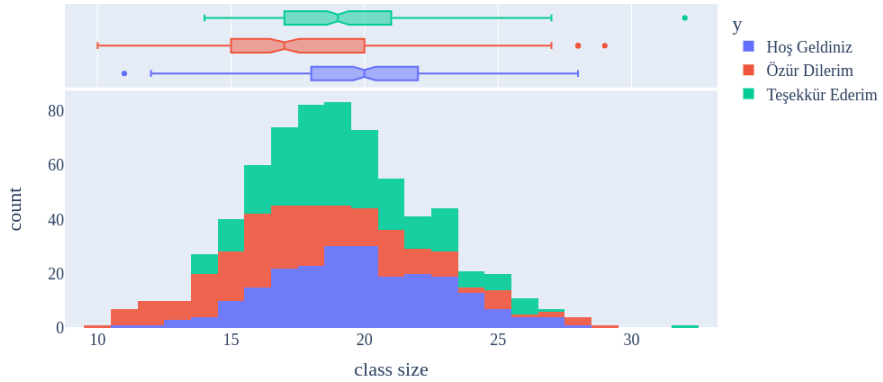


***Figure 3.*** *Data distributions according to frame length for each phrase class*

### 3.1.1. Data preprocessing

Since the dataset is in a raw form, some data preprocessing has been applied before trying to solve the classification problem. The first step is cutting the lip region with a face detection algorithm using dlib with 68 points facial landmark detector and OpenCV libraries because an important part of the lip reading problem is the lip part of the face. Just cutting the lips is beneficial for both the hyperparameter tuning of the models and the train time by minimizing the data, and it also allows us to capture the right focus area. Additionally, some data sequences can not be included in the cut mouth step since the mouth-cutting algorithm we used can not find the mouth part of the face correctly. Secondly, cut frames were converted from RGB to grayscale (see Figure 4). Finally, each image was set to a fixed size.

Additionally to preprocessing each image, each sample's number of frames has been fitted to a predetermined value. Some words are structured in a fixed size to ensure consistency because frame counts for some words vary depending on each speaker's speaking pace and the word's length. Based on the experiments, the value 15 produced the greatest outcomes for word lengths in this study. The samples are filled with empty frames if the number of frames is less than 15, and the ongoing frames are ignored if it is larger than 15.

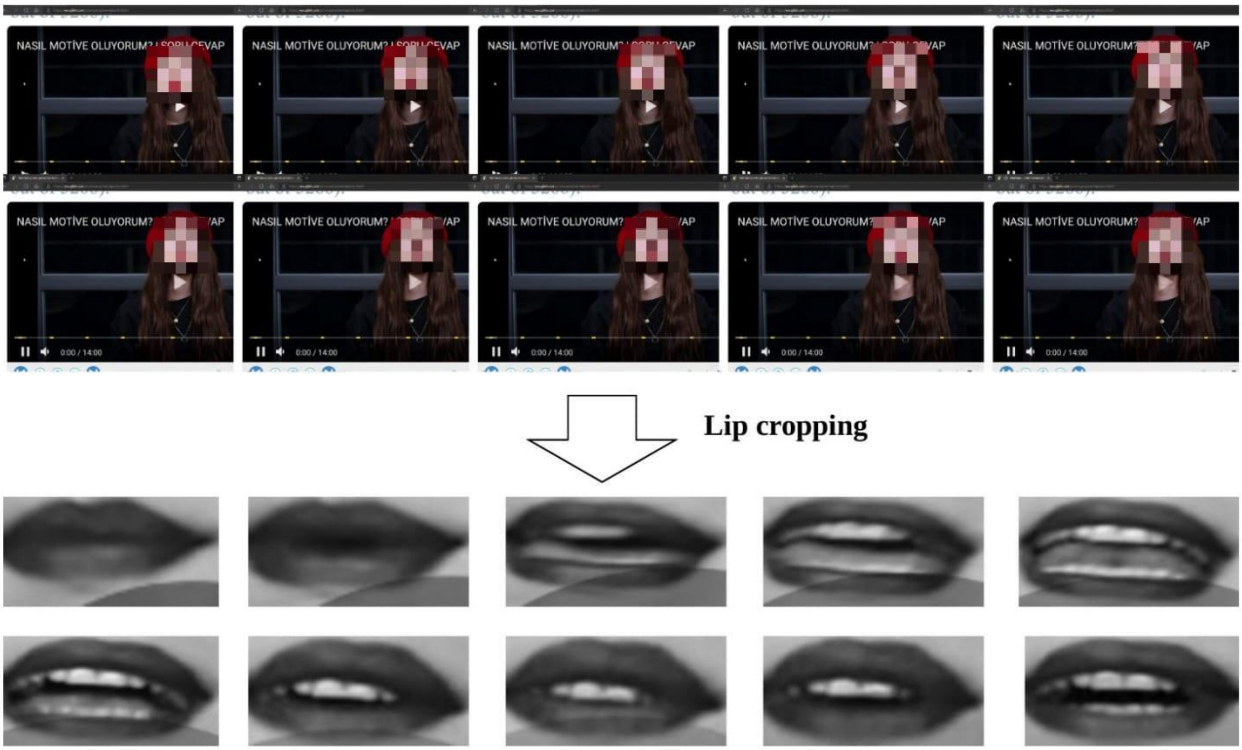


***Figure 4.*** *Lip cropping example*

### 3.1.2. Data augmentation

Data augmentation techniques are used when the dataset size is not enough to train deep learning algorithms or when the data quality or variety is not enough. With the help of augmentation techniques, classification results can be enhanced. In this work, we applied three different augmentation techniques to the dataset. It is important to note that augmentation techniques were implemented for the whole dataset since the visual lip reading problem concerns the sequence data where data are all images. The first augmentation technique is a horizontal flip (see Figure 5, the second row). A horizontal flip is a mirror reflection by the y-axis. The second augmentation technique is inverting the image by subtracting pixel values from 255 (see Figure 5, the third row). The last augmentation technique is sigmoid contrast (see Figure 5, the last row). The technique is applied with the sigmoid function in Equation (1), where the gain is (5, 10) and the cutoff is (0.4, 0.6). After applying the augmentation techniques, the dataset size expanded from 1390 to 5560 sets of example**s.**

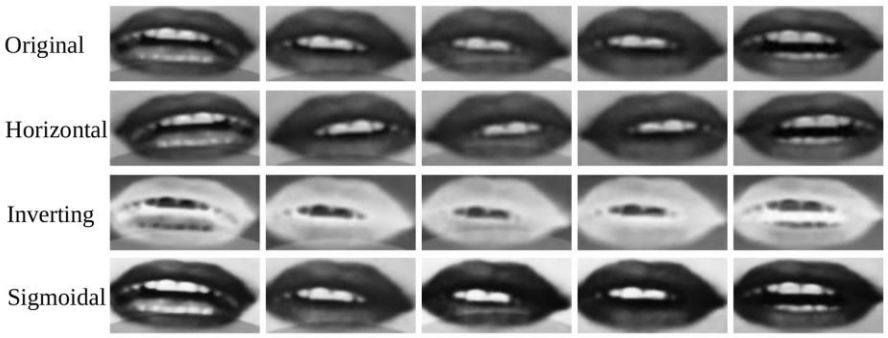


***Figure 5.*** *Data augmentation techniques applied on visual lip reading in Turkish dataset*

$$f(v) = 255 \times \frac{1}{(1 + exp^{(gain \times cutoff - \frac{v}{255})})} \quad (1)$$

### 3.2. Deep Learning Models

In this work, we tried three models to compare training time and classification scores. We worked on CNN, LSTM, and BGRU. Each model has its own advantages and disadvantages regarding training time, accuracy, adaptation to sequence data, etc.

**Convolutional Neural Networks.** Convolutional Neural Networks (CNNs) have become an increasingly popular deep learning technique for image and sequence data. The success of CNNs on these tasks is partly due to their efficient implementation, which allows them to achieve performance comparable to that of more complicated learning algorithms. CNNs are composed of a series of convolutional layers followed by pooling layers and fully connected layers for prediction. A convolutional layer is a set of filters applied to the input later. Each filter is applied to the input in turn and represents a different feature of the input data. These filters are called kernels. Pooling layers combine information across multiple pixels into a single output value. The final fully connected layer performs the final classification step.

**Long-Short Term Memory.** Long Short-Term Memory (LSTM) is a powerful deep learning algorithm that has proven efficient on sequence data. It produces results based on time steps, which makes it a good choice for tasks such as recognition and classification. Unlike CNNs, which are trained as a collection of static layers, the LSTM uses recurrent connections to improve accuracy. Recurrent connections allow the network to remember information over a period of time and adjust its behavior accordingly. There are three main components in an LSTM cell: input gate, forget gate and output gate. The input gate detects when a new input is available. When it detects a new input, it resets the state vector for the next iteration and updates the output. The forget gate determines whether or not the previous state vector should be forgotten. The output gate determines the state vector's state in the next iteration. When the forget gate decides that the state vector should not be forgotten, the output gate determines the new state vector based on the current input. Applications of LSTM include visual speech recognition and image classification.

**Bidirectional Gated Recurrent Unit.** One of the most important features of neural networks is their ability to learn and remember patterns. Gated Recurrent Units (GRUs) are a type of recurrent neural network that is especially adept at learning long data sequences, including sequential images over time. They are similar to LSTMs, but more straightforward and more flexible. Moreover, GRU provides a powerful method for modeling complex temporal relationships by providing gating control over memory. GRU model takes in input and produces output based on the current state of its memory and the input it receives at the current time step. In other words, GRU is modeled by a matrix multiplication operation whose input and output are the activation function values of each GRU cell. Like an LSTM, a GRU has a cell that can remember the input it receives at a past-time step. However, a GRU has only one forget gate and one update gate, unlike an LSTM. Each of these gates controls where a unit forgets its previous or updates its value based on the new input it receives at the current time step. Additionally, there are two types of GRU which are bidirectional and undirectional. A bidirectional gate connects two units in adjacent layers to allow information to flow both forward and backward in the network. GRUs are used in problems where the state space is larger than the problem's input. So it can be applicable to visual lip reading problems.

### 3.3. Model Architectures

The performance of deep neural networks improves by training them, which is a process that requires making several choices related to network architecture and hyperparameters. These choices are difficult to make precisely because they significantly impact the network's performance. In other words, it is crucial that researchers are able to optimize them. Our experiments use three different models, which require three different hyperparameter optimization steps. As a result of the hyperparameter tuning, the best model results are described below. The hyperparameter values are explained in detail in Table 2.

***Table 2.*** *Hyperparameters used in models. CCE: Categorical Cross Entropy*

| Hyperparameter Name | CNN | LSTM | BGRU |
|---|---|---|---|
| learning rate | 0.0002 | 0.0002 | 0.0001 |
| optimizer | ADAM | ADAM | ADAM |
| loss | CCE | CCE | CCE |
| hidden layer dropout | 0.4 | 0.5 | 0.25 |
| hidden layer neurons | 72 | 64 | 64 |
| hidden layer size | 2 | 2 | 1 |
| feature extraction layer | 2 | 2 | 1 |
| filter (CNN) /unit (LSTM, BGRU) s | 96 | 32 | 72 |
| feature extraction dropout prob. | - | 0.5 | 0.2 |
| activation function | ReLU | ReLU | ReLU |
| pooling size | 2 | - | - |
| patience | 4 | 5 | 3 |



The first model is the CNN model (see Figure 6). As it can be seen from Figure 6, two Conv3D layers with 96 filters and maxpooling3D layers are used as feature extraction layers. In the Conv3D layers, filters are applied with the size of (3, 3, 3), and strides are 1. Maxpooling3D layers applied with pooling size (2, 2, 2) and stride is 2. After Conv3D and maxpooling3D layers, flatten layer is applied. After that, two dense layers with 72 neurons were followed by a dropout layer with a probability of 0.4. Lastly, an output layer with 6 neurons is applied. ReLU activation function is used in all layers except the output layer. In the output layer, the softmax activation function is used since we perform a classification problem with 6 classes. The other hyperparameters are as follows: the learning rate is 0.0002, the optimizer is ADAM, and the loss function is categorical cross-entropy. Training is performed with early stopping monitoring validation accuracy, and patience is 4.

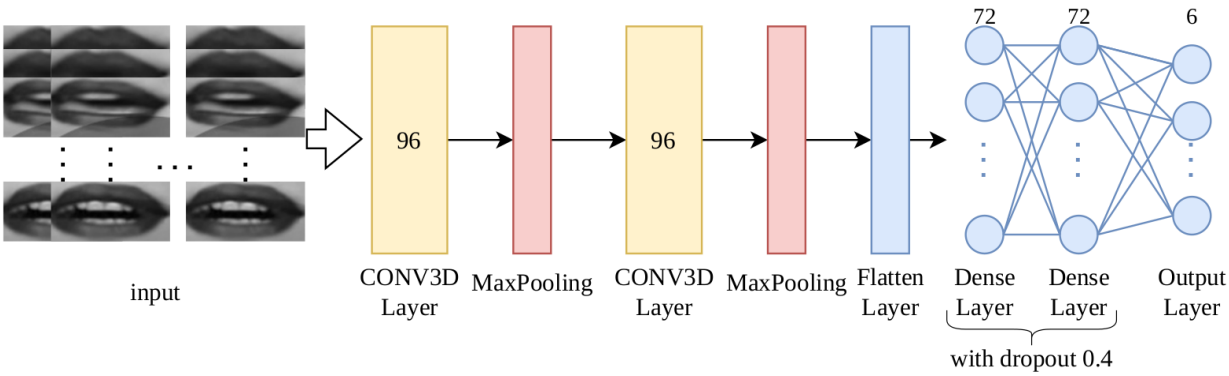


***Figure 6.*** *CNN model architecture*

The second model is the LSTM model (see Figure 7). The LSTM model is performed with two LSTM layers with 32 neurons and 0.5 dropout probability. Following that flatten layer is applied. The next layers are two dense layers with 64 neurons and 0.5 dropout probability. As an output layer, a dense layer with 6 neurons was applied. Except for the output layer, which uses the softmax function, the ReLU function is used in the fully connected layers. Other hyperparameters are as follows: the learning rate is 0.0002, the

optimizer is ADAM, and the loss function is categorical cross-entropy. Training is performed with early stopping monitoring validation accuracy, and patience is 5.

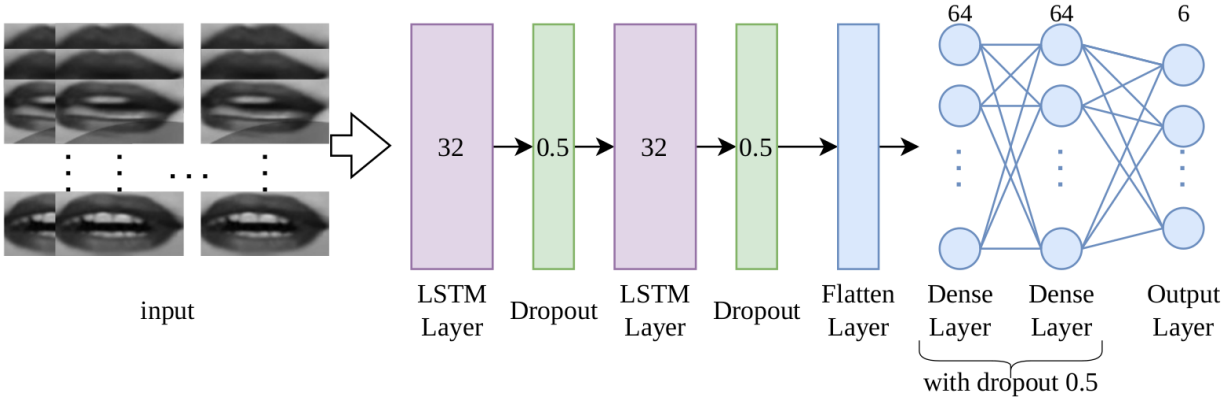


***Figure 7.*** *LSTM model architecture*

The last model is the BGRU model (see Figure 8). This model contains much fewer layers than the others. It uses a bidirectional GRU layer with 72 units and 0.2 dropout probability. Then, the flatten layer and dense layer with 64 neurons and 0.25 dropout probability. The last layer is again a dense layer with 6 neurons. As applied to the other models, the ReLU function is used in the hidden layer, and the softmax function is used in the output layer. The other hyperparameters are as follows: the learning rate is 0.0001, the optimizer is ADAM, and the loss is categorical cross-entropy. Similarly, the BGRU model is trained with early stopping monitoring validation accuracy, and patience is 3.

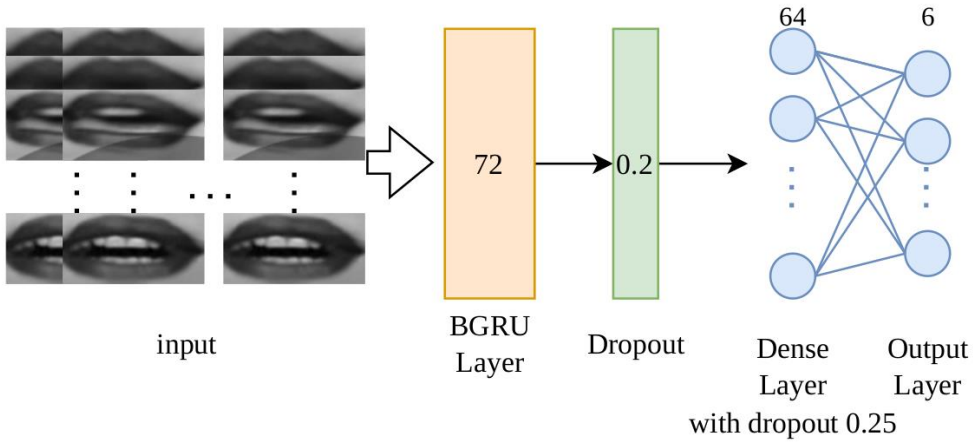


***Figure 8.*** *BGRU model architecture*

## 4. EXPERIMENTS

The model architectures were explained in detail in the previous section. Experiments are run on an NVIDIA Tesla T4 graphics card. The dataset is divided into three parts: train, validation, and test sets with percentages of 70%, 15%, and 15%, respectively. The training set contains 3892 sets of examples, while validation and test sets contain 834 sets of examples. The training epochs are different since each model is trained with early stopping to prevent the model from overfitting. The CNN model's epoch size is 62, LSTM's epoch size is 58, and BGRU's epoch is 29. The accuracy results and training times for each model are in Table 3. The accuracy scores are very close to each other, unlike the training time. LSTM and BGRU models' accuracy scores are the same as the sixth decimal, 0.7781. CNN, which is 0.7649, performed the worst among the three models. In this case, training time helps decide the models' performance. The BGRU

model is the fastest, approximately at 216 seconds, and the CNN model is the slowest, approximately at 863 seconds.

***Table 3.*** *Model accuracy and their training time results*

| Model | Accuracy | Training time (secs) |
|---|---|---|
| CNN | 76.49% | 862.84 |
| LSTM | 77.81% | 389.30 |
| BGRU | 77.81% | 215.59 |

Additionally, we evaluated each model by confusion matrix (see Figures 9,10,11). Since the accuracy scores are almost the same, we observed that the confusion matrices of the LSTM and BGRU models differ. Phrases and words performed well among themselves for the three models. Moreover, we evaluated the precision, recall, and f1 scores for each class trained with the three models (see Table 4). As it can be seen from Table 4, there is no strict way to draw a conclusion about which model is more accurate. For example, for classes "hoş geldiniz" and "selam" CNN's precision scores are higher than others, or for classes "teşekkür ederim" and "merhaba" LSTM's precision scores are higher than the other two. However, we can observe that for some metrics and models, there is a considerably high difference between results.

***Table 4.*** *Precision, recall, and f1 scores of models*

| Words | Size | Model | Precision | Recall | F1 score |
|---|---|---|---|---|---|
| hoş geldiniz | 153 | CNN | 0.6702 | 0.8366 | 0.7442 |
| | | LSTM | 0.6089 | 0.8954 | 0.7249 |
| | | BGRU | 0.6079 | 0.9020 | 0.7263 |
| özür dilerim | 105 | CNN | 0.6600 | 0.6286 | 0.6439 |
| | | LSTM | 0.8594 | 0.5238 | 0.6509 |
| | | BGRU | 0.8514 | 0.6000 | 0.7039 |
| teşekkür ederim | 139 | CNN | 0.8519 | 0.8273 | 0.8394 |
| | | LSTM | 0.8264 | 0.8561 | 0.8410 |
| | | BGRU | 0.8561 | 0.8129 | 0.8339 |
| merhaba | 167 | CNN | 0.8696 | 0.7186 | 0.7869 |
| | | LSTM | 0.8639 | 0.7605 | 0.8089 |
| | | BGRU | 0.8872 | 0.7066 | 0.7867 |
| selam | 141 | CNN | 0.8718 | 0.7234 | 0.790 |
| | | LSTM | 0.8382 | 0.8085 | 0.8231 |
| | | BGRU | 0.8014 | 0.8298 | 0.8153 |
| günaydın | 129 | CNN | 0.6993 | 0.8295 | 0.7589 |
| | | LSTM | 0.8220 | 0.7519 | 0.7854 |
| | | BGRU | 0.8197 | 0.7752 | 0.7968 |



For instance, "özür dilerim" class's precision score is much lower for the CNN model. On the other hand, "günaydın" class's recall score is much higher for the CNN model. For f1 score values, there are no such significant differences. To be more specific, the highest precision score, 0.88%, was obtained for "merhaba" with the BGRU model; similarly, the highest recall score, 90%, was obtained with the BGRU model on "hoş geldiniz", and the highest f1 score, 0.85%, was obtained with the LSTM model on "teşekkür ederim". If we consider the classes separately, we can conclude them as follows. Firstly, the phrases are evaluated. In the "hoş geldiniz" class, although the CNN model is the best in precision and f1 score, the recall score of the BGRU model is the highest among them. In the "özür dilerim" class, the LSTM model's precision is the best among all the models and metrics. CNN model is good at recall, and the BGRU model is good at the f1 score. The scores in the "teşekkür ederim" class are close. BGRU is the best in precision, and LSTM is the best for recall and f1 scores. Lastly, words are evaluated. In "merhaba" class, similar results with

"teşekkür ederim" occur. BGRU is the best in terms of precision, and LSTM is the best for recall and f1 scores. In the "selam" class, the precision score is the best with CNN, the recall score is the best with BGRU, and the f1 score is the best with LSTM. In the "günaydın" class, the precision score is the best with LSTM, the recall score is the best with CNN, and the f1 score is the best with BGRU.


***Figure 9.*** *CNN model confusion matrix*

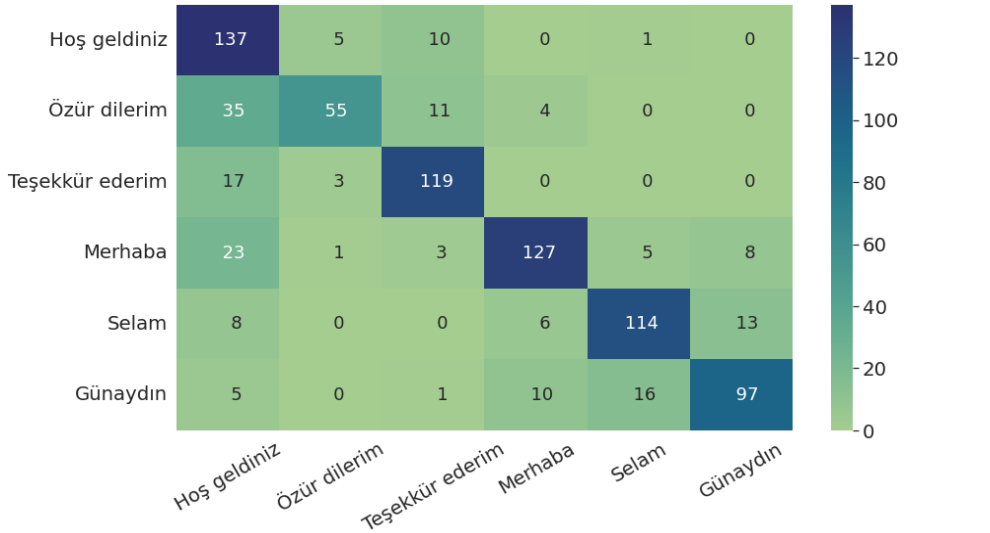


***Figure 10.*** *LSTM model confusion matrix*

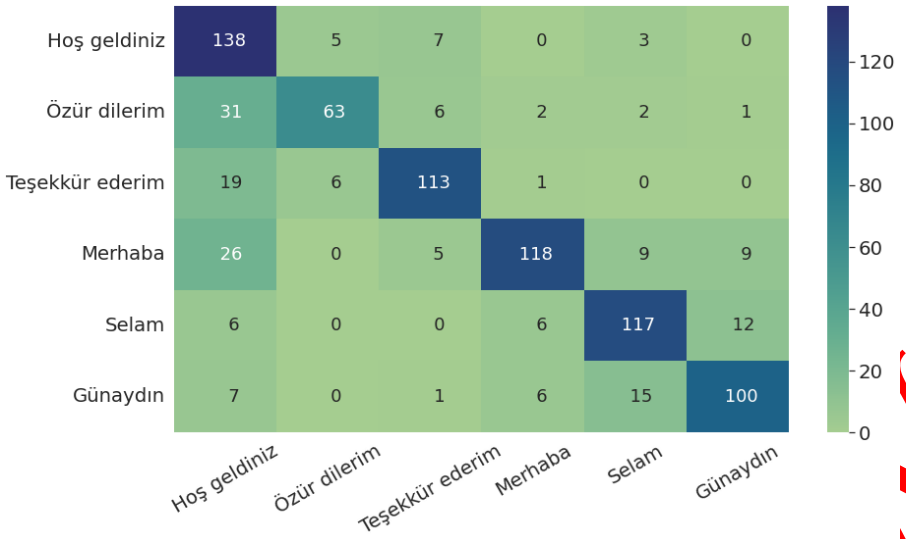


***Figure 11.*** *BGRU model confusion matrix*

## 5. RESULTS AND DISCUSSION

Lip reading is the process by which the brain learns to identify and interpret auditory information that has been perceived through the lips. While various techniques can facilitate the process of lip reading, it is possible to learn the patterns of computer systems. Deep learning has proven its efficiency in many fields, including image recognition and speech recognition. So, it is no surprise that it is also used for lip reading. Using deep learning for lip reading has numerous benefits. For example, it is able to resolve speech in noisy environments and does not require an expert lip reader to interpret speech. The biggest challenge here is that our real-world lifecycle is not always perfect for high-accuracy results. In this work, we showed that our model successfully works on the dataset which simulates real-world conditions.

In our research, we placed a strong emphasis on enhancing the richness and diversity of our dataset, strategically employing sigmoid contrast techniques. This deliberate approach was undertaken with the intention of elevating the data quality and expanding the range of variations present in our dataset. In this way, we aimed to create a resource that not only surpassed the traditional benchmarks but also facilitated more robust and nuanced research in the field of Turkish lip reading. Moreover, our investigation unveiled a crucial facet of our study: the applicability of diverse solution approaches within the exclusive realm of visual datasets. We showcased the versatility of our dataset by demonstrating its compatibility with both sequential and feature extraction techniques, thus illuminating the myriad possibilities it offers to researchers. In our rigorous experimentation, the results indicated a notable trend. Recurrent-based models, specifically the Long-Short Term Memory (LSTM) and Bidirectional Gated Recurrent Unit (BGRU) models, demonstrated their exceptional efficiency when compared to the convolutional-based feature extraction technique, Convolutional Neural Networks (CNN). Not only did they exhibit superior accuracy in classification tasks, but they also exhibited a noteworthy reduction in training time. Consequently, our findings underscored the BGRU model as the most efficient choice, both in terms of training time and overall classification performance. This finding has significant implications for the field of Turkish lip reading, as it highlights a promising avenue for further research and practical applications.

Furthermore, we used data augmentation techniques; however, there are many more techniques for this dataset. In the future, we can work on different augmentation techniques and observe their effect on the model results. Additionally, we can work on transfer learning to enhance the performance of the models. To sum up, as the lip reading problem becomes more adaptable to daily life and society, lip reading will be more critical for technological developments.

## CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

## REFERENCES

[1] Fisher, C. G., "Confusions among visually perceived consonants", Journal of Speech, Language, and Hearing Research, 11(4): 796–804, (1968).

[2] Easton, R. D., and Basala, M., "Perceptual dominance during lipreading", Perception and Psychophysics, 32(6): 562–570, (1982).

[3] Lesani, F. S., Ghazvini, F. F., and Dianat, R., "Mobile phone security using automatic lip reading", 9th International Conference on e-Commerce in Developing Countries: With focus on e-Business, Isfahan, Iran, 2015, 1-5, (2015).

[4] Mathulaprangsan, S., Wang, C. Y., Frisky, A. Z. K., Tai, T. C., and Wang, J. C., "A survey of visual lip reading and lip-password verification", International Conference on Orange Technologies (ICOT), Hong Kong, China, 22-25, (2015).

[5] Bahdanau, D., Chorowski J., Serdyuk D., Brakel P., and Bengio Y., "End-to-end attention-based large vocabulary speech recognition", 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 4945-4949, (2016).

[6] Huang, J. T., Li, J., and Gong, Y., "An analysis of convolutional neural networks for speech recognition", IEEE International Conference on Acoustics, Speech and Signal Processing, South Brisbane, QLD, Australia, 4989–4993, (2015).

[7] Miao, Y., Gowayyed, M., Metze, and F., "EESEN: End-to-end speech recognition using deep RNN models and WFSTbased decoding", IEEE Workshop on Automatic Speech Recognition and Understanding, 167–174, (2016).

[8] Hyunmin, C., Kang, C. M., Kim, B., Kim, J., Chung, C. C., and Choi, W., "Autonomous Braking System via Deep Reinforcement Learning", ArXiv, abs/1702.02302, (2017).

[9] Soltani, F., Eskandari, F., and Golestan, S., "Developing a Gesture-Based Game for Deaf/Mute People Using Microsoft Kinect", 2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems, Palermo, Italy, 491-495, (2012).

[10] Tan, J., Nguyen, C. T., and Wang, X., "SilentTalk: Lip reading through ultrasonic sensing on mobile phones", IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, Atlanta, GA, USA, 1-9, (2017).

[11] Lu, L., Yu, J., Chen, Y., Liu, H., Zhu, Y., Kong, L., and Li, M., "Lip reading-based user authentication through acoustic sensing on smartphones", IEEE/ACM Transactions on Networking, 27(1): 447–460, (2019).

[12] Tan, J., Wang, X., Nguyen, C., and Shi, Y., "Silentkey: A new authentication framework through ultrasonic-based lip reading", Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2(1): 1–18, (2018).

[13] Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A., "Lip reading sentences in the wild", 2017 IEEE Conference on Computer Vision and Pattern Recognition, 6447-6456, (2017). DOI: https://doi.org/10.1109/cvpr.2017.367

[14] Iwano, K., Yoshinaga, T., Tamura, S., and Furui, S., "Audio-Visual Speech Recognition Using Lip Information Extracted from Side-Face Images", Hindawi Publishing Corporation EURASIP Journal on Audio, Speech, and Music Processing, 2007: 0-9, (2007).

[15] Fenghour, S., Chen, D., Guo, K., Li, B., and Xiao, P., "Deep learning-based automated lip-reading: A survey", IEEE Access, 9: 121184–121205, (2021).

[16] Pandey, L., and Arif, A. S., "LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model", In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, Article 1: 1–19, (2021).

[17] Chitu, A., and Rothkrantz, L., "Visual Speech Recognition Automatic System for Lip Reading of Dutch", Journal on Information Technologies and Control, 7(3): 2-9, Simolini-94, Sofia, Bulgaria, (2009).

[18] Faisal, M., and Manzoor, S., "Deep Learning for Lip Reading using Audio-Visual Information for Urdu Language", ArXiv, abs/1802.05521, (2018).

[19] Haq, M. A., Ruan, S. J., Cai, W. J., and Li, L. P. H., "Using Lip Reading Recognition to Predict Daily Mandarin Conversation", in IEEE Access, 10, 53481-53489, (2022).

[20] Zhang, S., Ma. Z., Lu. K., Liu. X., Liu. J., Guo. S., Zomaya. A. Y., Zhang. J., and Wang. J., "HearMe: Accurate and Real-time Lip Reading based on Commercial RFID Devices", in IEEE Transactions on Mobile Computing, early access, 1-14, (2022).

[21] Peng, C., Li, J., Chai, J., Zhao, Z., Zhang, H., and Tian, W., "Lip Reading Using Deformable 3D Convolution and Channel-Temporal Attention", 13532, In: Pimenidis, E., Angelov, P., Jayne, C., Papaleonidas, A., Aydin, M. (eds) Artificial Neural Networks and Machine Learning. Lecture Notes in Computer Science, Springer, Cham, 707-718, (2022).

[22] Xue, B., Hu, S., Xu, J., Geng, M., Liu, X., and Meng, H., "Bayesian Neural Network Language Modeling for Speech Recognition", in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30: 2900-2917, (2022).

[23] Ozcan, T., and Basturk, A., "Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models", Balkan Journal of Electrical and Computer Engineering, 7(2), (2019).

[24] Fernandez-Lopez, A., and Sukno, F. M., "Survey on automatic lip-reading in the era of Deep learning. Image and Vision Computing", Image and Vision Computing, 78: 53–72, (2018).

[25] Fenghour, S., Chen, D., Guo, K., and Xiao, P., "Lip reading sentences using deep learning with only visual cues", IEEE Access, 8: 215516–215530, (2020).

[26] Graves, A., Fernandez, S., Gomez, F., and Schmidhuber, J., "Connectionist temporal classification: labeling unsegmented sequence data with recurrent neural networks", In Proceedings of the 23rd international conference on Machine learning, Association for Computing Machinery, New York, NY, USA, 369–376, (2006).

[27] Cooke, M., Barker, J., Cunningham, S., and Shao, X., "An audio-visual corpus for speech perception and automatic speech recognition", The Journal of the Acoustical Society of America, 120(5): 2421–2424, (2006).

[28] Berkol, A., Tümer-Sivri, T., Pervan-Akman, N., Çolak, M., and Erdem, H., "Visual Lip Reading Dataset in Turkish", Data, 8(1): 15, (2023).

[29] https://www.youtube.com. Access date: 08.11.2022