Automated Word-level Lip Reading using Convolutional Neural Networks with New Turkish Dataset

Nergis Pervan-Akman , Ali Berkol * , Hamit Erdem

Posted Date: 20 April 2023

doi: 10.20944/preprints202304.0645.v1

Keywords: Lip Reading; Multiclass Classification; Turkish Lip Reading Dataset; Deep Learning; Convolutional Neural Networks; Lip Detection

Article

Automated Word-Level Lip Reading Using Convolutional Neural Networks with New Turkish Dataset

Nergis Pervan-Akman [1], Ali Berkol [1,*] and Hamit Erdem [2]

[1] Defence and Information Systems, BITES, METU Technopolis, Ankara 06530, Türkiye
[2] Electrics and Electronics Department, Başkent University, Ankara 06790, Türkiye
* Correspondence: ali.berkol@bites.com.tr

Abstract: Automated lip reading is a research problem that has developed considerably in recent years. Lip reading is evaluated both visually and audibly in some cases. The lip reading model is a field of use for detecting specific words using images from security cameras, but it is not possible to use audio-visual databases in this situation. It is not possible to obtain the sound input of the pronounced word in all cases. We collected a new Turkish dataset with only the image in this study. The new dataset is produced using Youtube videos, which is an uncontrolled environment. For this reason, images have difficult parameters in terms of environmental factors such as light, angle, color, and personal characteristics of the face. Despite the different features on the human face such as mustache, beard, and make-up, the visual speech recognition problem was developed on 10 classes including single words and two-word phrases using Convolutional Neural Networks (CNN) without any intervention on the data. The proposed study using only-visual data obtained a model which is automated visual speech recognition with a deep learning approach. In addition, since this study uses only-visual data, the computational cost and resource usage is less than in multi-modal studies. It is also the first known study to address the lip reading problem with a deep learning algorithm using a new dataset belonging to the Ural-Altaic languages.

Keywords: Lip Reading; Multiclass Classification; Turkish Lip Reading Dataset; Deep Learning; Convolutional Neural Networks; Lip Detection

1. Introduction

Speech is the most commonly used method of communication between people. Although speaking is carried out audibly, the sight also has a great impact on understanding spoken expressions. McGurk and MacDonald observe people's reactions by showing them audio and visual data pointing to different texts [1]. Audio narration and vision are input data that support each other. Automated Visual speech recognition is a more challenging problem in terms of ensuring generalizable word variety and accuracy than voice speech recognition and audio-video speech recognition, so their accuracy performance is lower. One of the troublesome situations in visual speech recognition is homophones with similar expressions, that is, expressions with similar lip movements. In addition, the quality of the image, and the absence of the face and lips of the person in the image are also challenging factors.

Problems such as dictating messages to smartphones in noisy environments [2,3], using visual silent passwords [4–6], transcribing silent films [7,8], synthesizing sound based on lip movements for speech-impaired people [9–12], and analyzing lip appearances to help hearing-impaired people [13] are among the application areas of automated lip reading systems.

Audio datasets, visual datasets, and audio-visual datasets contain various attributes according to what tasks they can be used for, the number of speakers, word size, information, and the total duration of recordings. The first databases in this area had limited knowledge as they were created as alphabets and digits [14–16]. However, datasets created in this way are a complex approach to solving

 doi:10.20944/preprints202304.0645.v1

the continuous speech recognition problem. Since letter and number recognition has a limited number of classes, it has been used very often and is useful for quickly evaluating the results of different algorithms. however, the models created are not generalizable since they have a limited scope. This situation reveals the need for inputs that are close to real-world problems involving words, expressions, and sentences. Examples of databases are available in many languages, such as English [17–20], French [21], and Czech [22], that contain this type of word or phrase. Apart from the generalizability of the class content, multi-image datasets are also important for the generalizability of the image. In real images, we can never guarantee that the camera is looking directly at the human face from the front. In multi-view datasets, there are contents with attributes such as number of speakers, number of classes, language, shooting angle of the camera as digit [23], word [24], phrase [23], and sentence [25] inputs [26]. In this study, we also handled to make the estimation using a word-level dataset consisting of only discrete words.

In this study, our main contribution is to both reduce the computational cost by using only visual data and to develop a lip reading system in case the corresponding audio data is not available. Furthermore, there is no known deep learning model obtained from natural human speech images used in Turkish lip reading studies and created using such dataset. We propose a word-level CNN model trained with inputs containing different lip representations.

The rest of the paper is organized in the following order: Related work is presented in Section 2 to see and have information about similar studies. Then, the detailed analysis of dataset, detection of lips, mouth representation, CNN model and training parameters are explained in Section 3. Metrics such as the number of test data, performance scores, and confusion matrix used are shown in Section 4. Finally, the evaluation of this study and its contributions to the literature are summarized in Section 5.

2. Related Work

Visual speech recognition problem studies generally consist of two steps as feature extraction and classification. In current studies, feature extraction processes of lip images are generally created automatically with deep learning approaches that work well on images. The first step of feature extraction is to correctly identify the face and then the lips. In lip reading datasets derived from natural images, we cannot expect every image has the frontal face. Since some face images are right, left, down, or up, it is difficult to get the right lip image in such cases. It is effective to apply face frontalization on the data before the lip detection step. The Robust Face Frontalization (RFF) [27] approach estimates the position and 3D deformable shape of the input face and warps it onto a frontally viewed synthetic face.

Early lip reading methods relied on Support Vector Machines (SVMs) [28] or Hidden Markov Models (HMMs) [29] to classify manually acquired geometry of lips. Existing approaches now consist mostly of innovative machine learning and deep learning approaches. Atila and Sabaz [30] studied Turkish lip reading problem using deep learning models. They collected two new Turkish datasets one containing 111 words and the other 113 sentences. In this study, feature extraction using CNN and classification using Bidirectional Long Short-Term Memory (Bi-LSTM) are performed. With Bi-LSTM, 88.55% accuracy was achieved and the best results were achieved on sentence and word datasets. They declare that lip reading for sentence detection obtained better performance than word detection.

Yargic and Dogan [31] suggested another Turkish study. With the images taken from the MS Kinect device, a dataset was created in which Turkish color names are pronounced. The features were created by using the distance data between the lip points of this dataset, which was obtained in 3D. With this collected dataset, a 78.22% accuracy model with 15 classes was produced using the K-Nearest Neighbor (KNN) algorithm and this model was developed to support hearing-impaired children. In addition to this Turkish dataset, LRS (Lip Reading Sentences) [7], the largest known English audio-visual dataset, is built for continuous speech recognition. This database, containing more than 100,000 expressions and more than 1000 different people, was collected from BBC broadcasts.

Fung and Mak [32] used CNN and Bi-LSTMs with max-out activation units for sentence-level classification. It showed an improvement of 3.1% from the previous known auto encoder BiLSTM

 doi:10.20944/preprints202304.0645.v1

study and achieved 87.6% accuracy in the Ouluvs2 corpus. This is the first low-resource lip reading approach that does not require a discrete feature extraction step or a pre-training phase.

Özcan and Basturk [33] used the CNN algorithm to classify AvLetters[14] which is alphabet level lip reading dataset. They also used pre-trained CNN model. As far as is known, there is no lip reading study using AlexNet using transfer learning.

Margam et al. [34] proposed a 3D-2D-CNN-BLSTM architecture configuration for decoding ASCII characters to predict spoken sentences from the GRID corpus, As first approach, the proposed network was trained on characters using Connectionist Temporal Classification (CTC) loss (ch-CTC) and in a conventional automatic speech recognition training pipeline, the BLSTM-HMM model was trained on bottleneck lip features. The same 3D-2D-CNN-BLSTM network was trained with CTC loss on word labels in the second technique (w-CTC). Using the unseen speaker test set, a better result was obtained than LipNet with a Word Error Rate (WER) difference of 24.5%.

Pertridis and Pantic [35] developed the visual speech recognition feature extraction and classification steps end-to-end, without any manual feature extraction steps. An LSTM model, models for both streams, and a Bidirectional LSTM model combines the two streams.

Martinez et al. [36] abolished the limitations of Bidirectional Gated Recurrent Unit (BGRU) layers and improved performance. Initially, BGRU layers were replaced with Temporal Convolutional Neural Networks (TCN) layers. Secondly, they reduced GPU runtime from 3 weeks to 1 week with a simpler training process. Finally, A variable length augmentation technique was applied to generalize the trained model.

3. Materials and Methods

3.1. Dataset

The dataset [37] utilized in the study is the Turkish lip reading dataset that we used in our previous studies and collected on Youtube. The dataset consists of 10 classes of daily Turkish 5 words and 5 phrases. These classes are “afiyet olsun” (enjoy your meal), “başla” (start), “bitir” (stop), “görüşmek üzere” (see you), “günaydın” (good morning), “hoş geldiniz” (welcome), “merhaba” (hello), “özür dilerim” (sorry), “selam” (hi), “teşekkür ederim” (thank you). Screenshots of the parts of the relevant word are recorded in various images such as TV series, movies, and blog videos published on Youtube. Then, the recorded screenshots are saved to frame-level files using Python. The videos containing the relevant word are 1 or 2 seconds and the number of images obtained for each sample is in the range of 40-60. Since the collected images are recorded during the pronunciation of the word in a sentence flow, in some cases, the end or beginning of a word can join with another word. For this reason, frames before the beginning of the word and the frames after its end were manually deleted by making a frame-by-frame elimination. After this elimination, a significant reduction in the image sequence was observed for each sample. While collecting the data, it was tried to balance as much as possible with an equal number of samples for each class (see Table 1).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2023 doi:10.20944/preprints202304.0645.v1

Table 1. Data Distribution of Classes.

Classes	Number of Samples
afiyet olsun	235
başla	235
bitir	244
görüşmek üzere	224
günaydın	232
hoş geldiniz	226
merhaba	268
özür dilerim	209
selam	235
teşekkür ederim	237

Since the dataset contains both single-word and 2-word classes, the pronunciation duration of the phrases varies. For example, since the phrases "teşekkür ederim" and "özür dilerim" are longer, their pronunciation durations and the number of frames they occupy in the dataset are more than the word "selam". While the frame count of the word "özür dilerim" exceeds 30, the frame count of the word "selam" does not exceed 15 (see Figure 1). It is critical to consider this distribution to make a balanced representation when classifying.

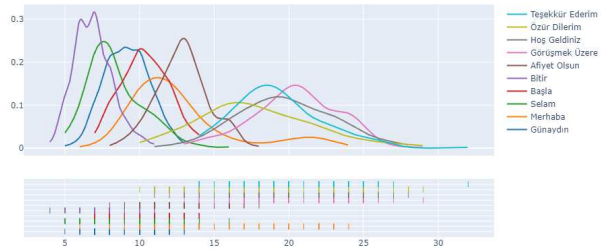

Figure 1. Data Frequencies.

Due to these data collected for the lip reading problem are obtained from the videos of the speakers who continue in their natural flow, the images are challenging in terms of diversity (see Figure 2). In some cases, speakers do not turn their face directly to the camera. Furthermore, there are situations such as light differences in the image, image quality, and the speaker being far away. In addition to these, there is also a problem that creates personal diversity such as objects such as microphones coming in front of the speaker in the images obtained, the speaker's mustache and lipstick.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2023 doi:10.20944/preprints202304.0645.v1

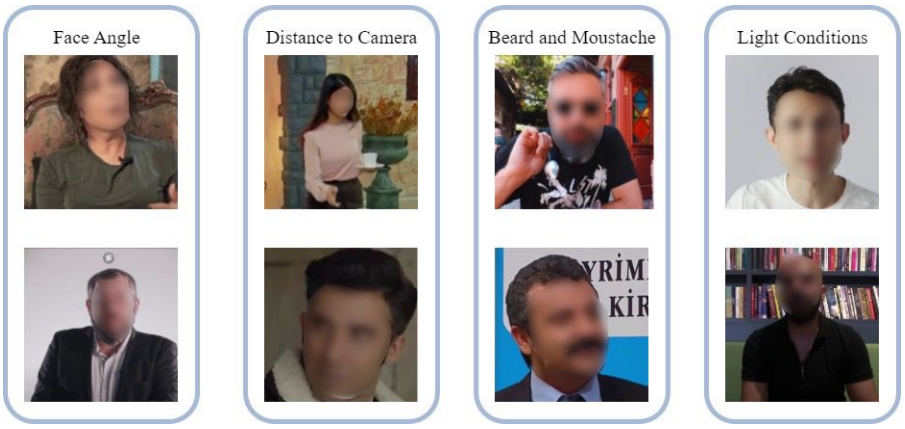

Figure 2. Data Challenges.

3.2. Detection of Lip

In the lip-reading problem, the RGB images are not important for the continuity of the studies. Images are converted to gray scale in order to reduce computational and time costs in face and lip detection studies and later during deep learning model training.

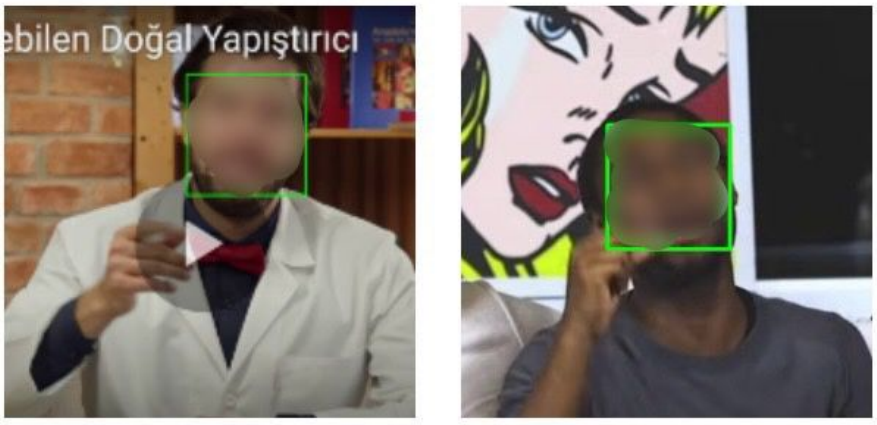

Figure 3. Face detection with HOG+SVM.

First, we cut the faces from the human images we collected using the dlib library, which is a ready-made library, since the faces on the images need to be handled. The get_frontal_face_detector() function we use does not receive face detection without taking any parameters. When this function is called, it returns the pre-trained HOG+Linear SVM face detector of the dlib. HOG+LINEAR SVM works fast and effectively. Due to the nature of the HOG, it adapts to rotation and viewing angle situations. This detector is built using a Histogram of Oriented Gradients (HOG) and a linear SVM. It is suitable method for real-time face detection due to its its rapid detection. As can be seen in Figure 3, even if the faces are angled or if there is an obstacle in front of the face, an accurate face detection can be made, including the lips.

In lip-cutting studies, using the OpenCV library, the contour of the relevant region is drawn by specifying a series of points to take the lip part. Since the 49-68 range corresponds to the lip region in the landmarks, the relevant range on the obtained face image is cut. Then, with the help of the boundingRect() function, a rectangular image of the determined region is taken. Figure 4 shows firstly, the original raw images, the faces detected in the second step, and the cut lip images at the end. Lip detection is also less than the number of raw images, as there is no corresponding face detection for each raw image. Although there are similar images in terms of angle and light in each image, it was observed that face detection could not be performed for each of them.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2023 doi:10.20944/preprints202304.0645.v1

Finally, the lip images obtained are recorded in 100X200 size to be used in the next steps.

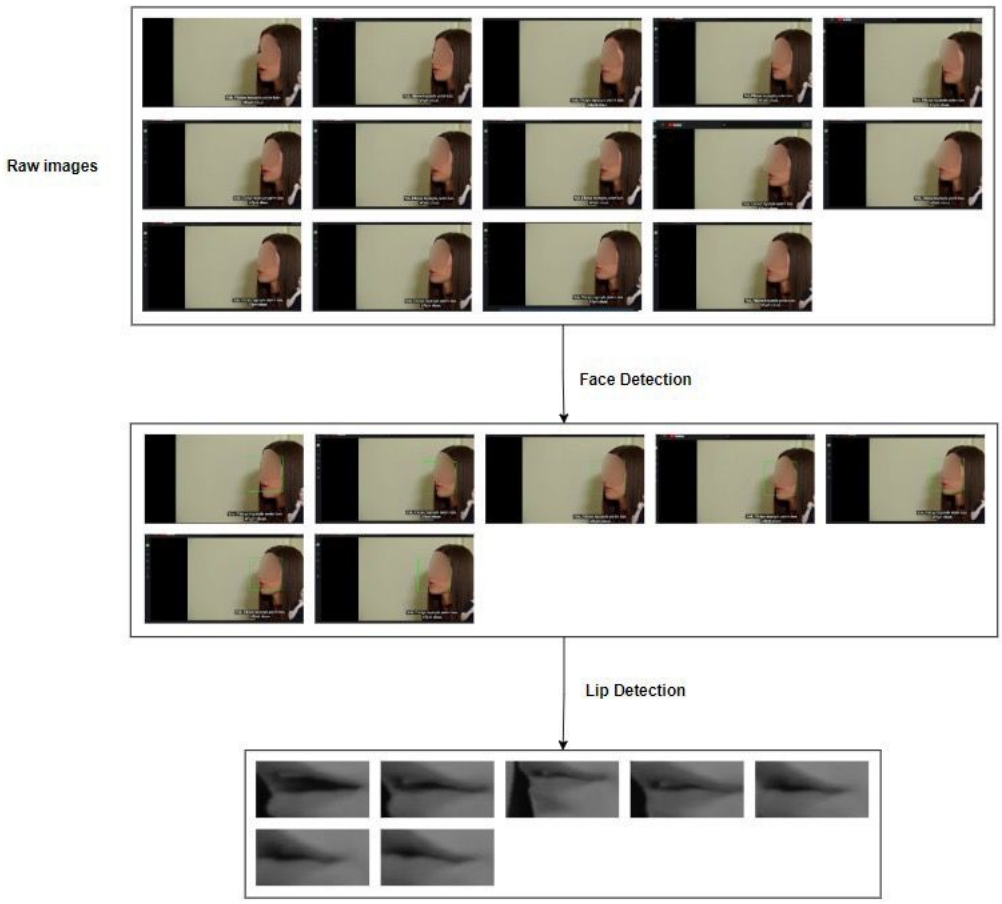

Figure 4. Lip Detection.

3.3. Lip Representation

In the first approach developed, each sequential image of a sample is used in the deep learning model so that the flow is preserved. As a second approach, 15 images are combined and used as a single smaller image. After the concatenation process, each 100x200 image is resized to 20x40 in order not to obtain a very large image. If the frame number of the relevant sample of the lip is less than 15, an image filled with 0 values on the gray scale is added. If it is more than 15 it is removed. When 15 frames are sequentially combined as 3 rows and 5 columns, a 60x200 image is obtained. In Figure 5 shows 15 sequence images produced in combination. In the case of separate lips, these images are used as a series of 15 images, providing a stream instead of a single image.

Figure 5. Concatenated Mouths.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2023 doi:10.20944/preprints202304.0645.v1

3.4. CNN Model

In recent years, the use of CNN models has increased considerably, thanks to technological developments such as the increase in the computing power of machines, the expansion of hardware resources, and the use of GPUs. It is very useful to take advantage of this deep learning model, especially if you are working on images with high computational costs. The biggest difference from vanilla artificial neural networks is that it reduces the number of parameters by taking certain regions of the image. It is used in many problems where the input data is an image, such as image classification, object detection, and image segmentation in recent studies, as well as in studies where the input is texts. One of the biggest reasons why complex problems in this area are easily solvable is that studies can be carried out with CNN without any feature extraction on the image, without the need for an expert's knowledge, and without detecting attributes such as location and shape for any object on the image.

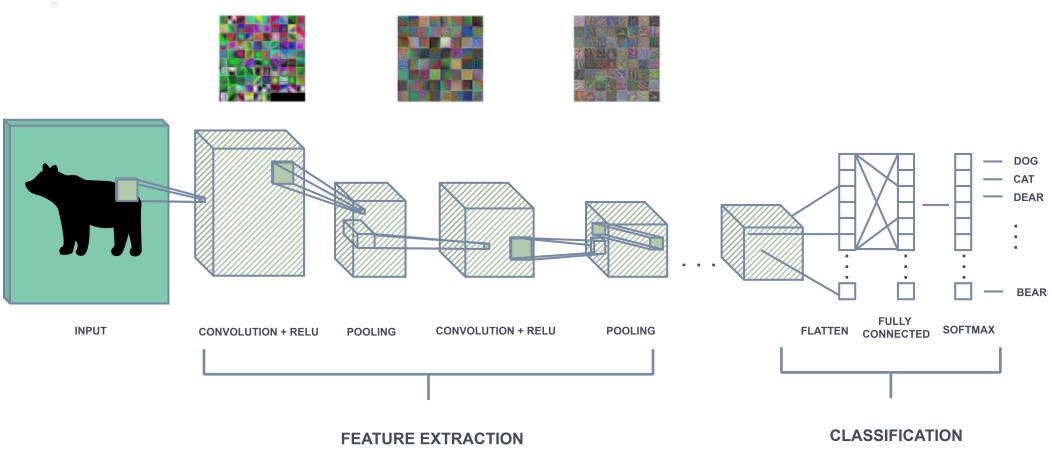

Figure 6. Traditional CNN Model.

Convolutional neural networks leverage four essential principles to exploit the features of natural signals: local connections, shared weights, pooling, and the usage of numerous layers [38]. The most important layer that distinguishes CNN from other neural network approaches is the convolution layer. It basically consists of input, convolution, pooling, and fully connected layers. A convolution layer's units are structured in feature maps, within which each unit is connected to local patches in the preceding layer's feature maps by a collection of weights. This locally weighted sum is then run via a non-linearity, such as a Rectified Linear Unit (ReLU). A feature map's units all use the same set of weights. Distinct the weights are used by different feature maps in a layer. The convolution layer detects local conjunctions of features from the preceding layer and the pooling layer merges semantically comparable features into one.

Classifier CNN models roughly consist of feature extraction and classification layers (see in Figure 6). The feature extraction layers produce meaningful outputs of the input image and are associated with other images, thanks to the convolution and pooling layers. In the classification layer, the number of classes and probabilities based on their meaningful images are now produced for the classification problem. Based on this, in a multi-class model, an estimation is made with the softmax function in the last layer.

3.5. Proposed Model

The proposed CNN architecture is two, for the concatenated lip images as a result of hyperparameter tuning, and for the lip images trained using discrete.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2023 doi:10.20944/preprints202304.0645.v1

3.5.1. CNN Model with Discrete Mouths Input

As we mentioned in the Section 3.3, discrete lip images are given to the input layer as a sequence. The architecture includes two convolution and two max-pooling layers. Convolution layers use ReLU as an activation function, the filter sizes are 128, and the stride used in filters is 1 with no padding. Max-pooling layers pool sizes are 3x3x3 with the stride of 2. Flatten layer follows these four layers and architecture continues with fully connected layers with dropout. The input vector consists of 15 images with a fixed size of 50x50. Random 128 filters are applied to these images in the convolution layer without padding and with a stride of 1 step. After the convolution process, an output of 13x48x48x128 is produced. Since there is a 3x3 pool size in the output of the max pooling layer following the convolution, it outputs as 6x24x24x128. After applying the conv3d, max-pooling, and flatten layers, respectively, a 15488 dimensional vector is obtained. Two fully connected layers with ReLU activation function and 0.5 ratio dropout layers used to avoid overfitting, especially in CNN models are implemented. Finally, since a multi-class classification problem is studied, the architecture is finalized with a fully connected layer that produces 10-dimensional vector output with the softmax activation function. In the output, probabilities are produced for 10 classes in the form of "afiyet olsun", "başla", "bitir", "görüşmek üzere", "günaydın", "hoş geldiniz", "merhaba", "özür dilerim", "selam", and "teşekkür ederim".

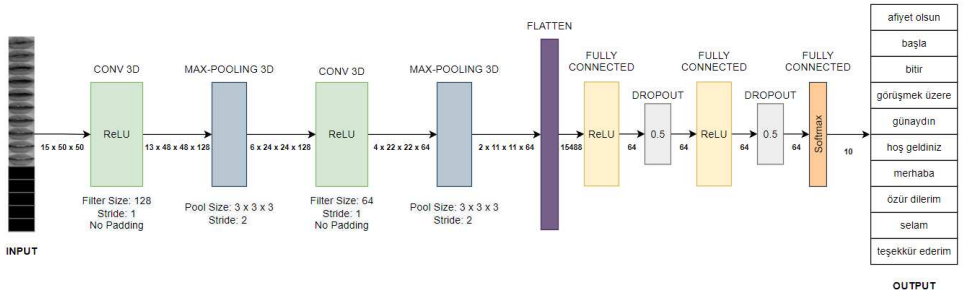

Figure 7. CNN Model using Discrete Represented Mouths.

3.5.2. CNN Model with Concatenated Mouth Input

In this approach where lips are combined, 15 images are concatenated to form a single image input, unlike the case of discrete mouths as input. Therefore, it is quite convenient in terms of computational cost. Experiments were conducted using a shallower series of convolution layers compared to the previous CNN model, since a single image represents a sequence of images, reducing data complexity. It is sent to the convolution layer using a 50x50 image as input. Experiments were conducted using a shallower series of convolution layers compared to the previous CNN model, since a single image represents a sequence of images, reducing data complexity. It is sent to the convolution layer using a 50x50 image as input. Then the Flatten layer's input is 24x24x16 since the pool size is 2x2. Unlike the architecture in the approach where the lips are given separately, there is 1 fully connected layer and dropout after the Flatten layer, which has 9216 dimensional vector output. Finally, an output vector with 10 classes is produced.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2023 doi:10.20944/preprints202304.0645.v1

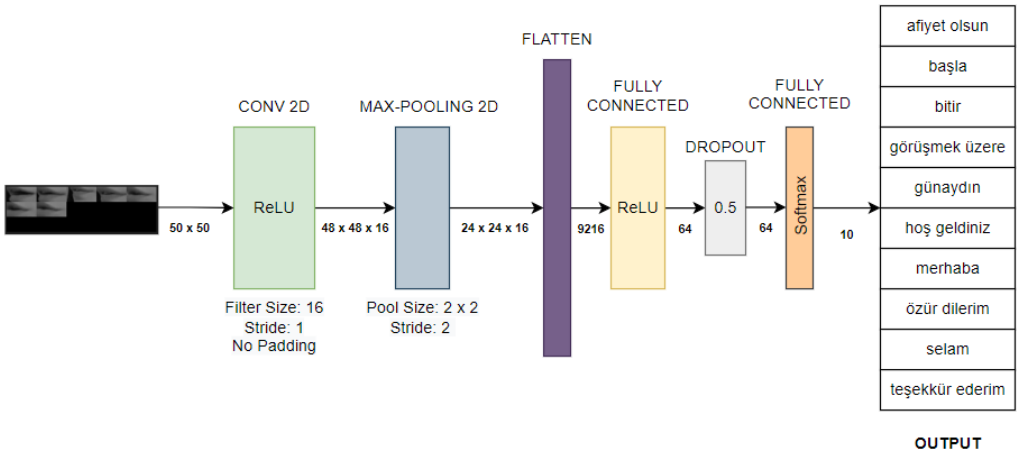

Figure 8. CNN Model using Concatenated Represented Mouths.

3.6. Training

In the training process, experiments were carried out on different hyperparameters for studies on two different approaches to training the lips separately and combining them. Mlflow, a Python library developed to manage the machine learning lifecycle, was used to evaluate the results of the experiments and make hyperparameter tuning. In Table 2 it is seen that the hyperparameter results for both approaches.

Different hyperparameters have been applied for the cases where the lips are joined and separate. Since the model capacity and complexity of the two approaches are different, parameters such as learning rate, batch size, and number of epochs varied.

Table 2. CNN Model Training Parameters.

Parameters	Discrete Mouth	Concatenated Mouth
Number of train samples	1606	1606
Number of validation samples	345	345
Number of test samples	344	344
Learning rate	0.0002	0.002
Batch size	32	16
Word length	15	15
Input dimension	50	50
Loss function	Categorical cross entropy	Categorical cross entropy
Optimizer	Adam	Adam
Total Trainable Parameters	1,220,938	590,698

4. Results and Discussion

It is difficult to make an accurate assessment in studies where language is involved, such as lip reading, because there are different pronunciations and variations in the language. It is possible to make an evaluation, especially when there are many studies and data in the English language. However, there is no comparable word-level dataset in terms of our studies in Turkish.

In our studies, we basically aimed to develop a CNN architecture for the Turkish lip reading problem. All experiments based on CNN architecture run on GPU. NVIDIA 1650 Ti graphics card with 4GB memory. The improvements were made using the Python Keras library. In addition to these, Plotly and Seaborn libraries were used for visualization, and OpenCV libraries were used for image processing studies.

 doi:10.20944/preprints202304.0645.v1

Experiments were performed with the number of samples in Table 3 in both of the representation steps, where the lips are joined and the lips are separated.

Table 3. Number of Test Samples of Each Class.

Classes	Number of Samples
afiyet olsun	29
başla	35
bitir	46
görüşmek üzere	23
günaydın	32
hoş geldiniz	34
merhaba	43
özür dilerim	31
selam	37
teşekkür ederim	35

4.1. Training Results with Discrete Lips

Looking at the training results, the accuracy and loss value changes for which the epoch number is determined using early stopping are shown in the Figure 9. The training process, which was stopped after the improvement in Loss value did not improve in 3 epochs, ended in 68 epochs. If the training continues further, there is no need to make further calculations as the model will be overfitting.

When the results of the predicted classes in the test data are examined, it is seen that the incorrectly determined classes are generally collected in the "afiyet olsun" class, see Figure 10. Especially for instances of classes whose actual class is "başla", "günaydın", and "özür dilerim", the wrong predictions concentrated on "afiyet olsun". Mistakes made in the "afiyet olsun" class were generally made for 6 examples in the "teşekkür ederim" phrase. Contrary to these, there is no example of an incorrectly guessed "afiyet olsun" in the "hoş geldiniz" phrase.

"hoş geldiniz", "merhaba", "selam", and when looked at the "başla", "bitir", "özür dilerim" classes that follow them, it is seen that the precision scores are high, see Figure 11. Thus, we can interpret that the majority of positive predictions for these classes are correct. In general, we see that the "afiyet olsun" class error rate is high based on the confusion matrix. There may not be a clear lip movement in the vocalization of these phrases in the dataset, or it may be interpreted as one of the more challenging expressions compared to Turkish grammar rules. Since f1-score is the harmonic mean of precision and recall metrics, it is generally seen as f1-score high when precision and recall are high at the same time, or low when f1-score is low at the same time such as "hoş geldiniz" and "günaydın". Although there is no class imbalance in terms of the number of samples in this dataset, the prediction performances vary according to the classes, as there are situations that create diversity for each class, such as the differences in speakers, viewing angles, and light differences, just like real-life scenes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2023 doi:10.20944/preprints202304.0645.v1

Figure 9. Training and Validation Accuracy and Loss per Epoch with Discrete Lips.

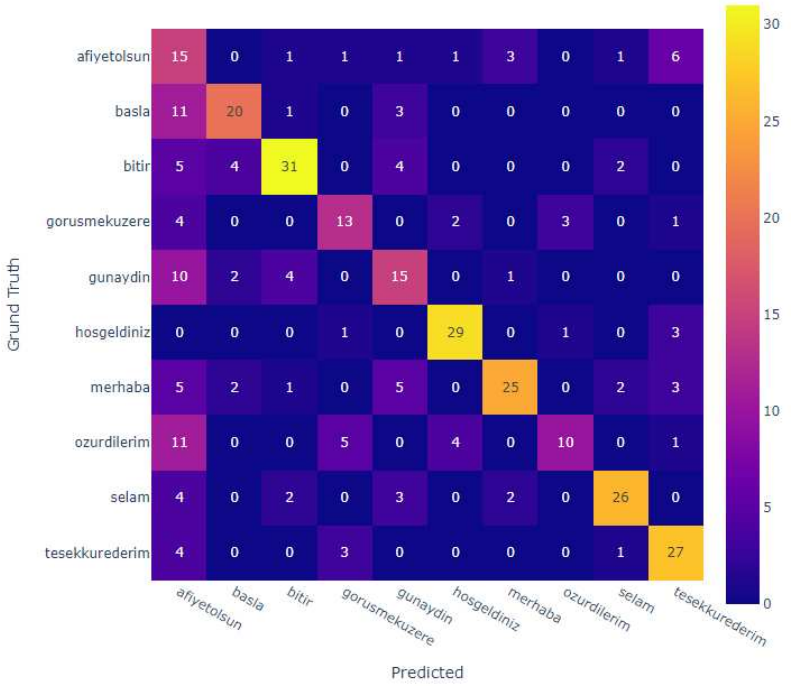

Figure 10. Confusion Matrix of Model Trained with Discrete Lips.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2023 doi:10.20944/preprints202304.0645.v1

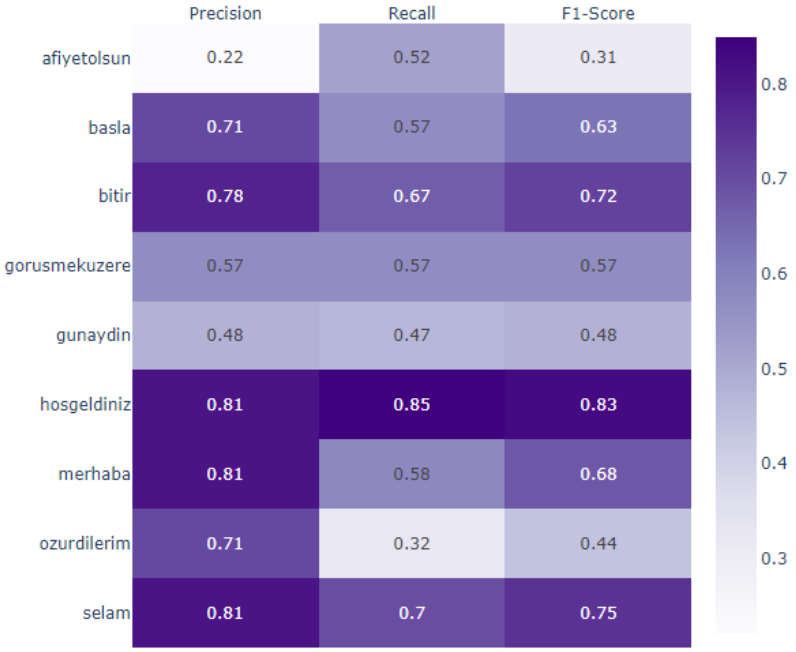

Figure 11. Classification Report of Model Trained with Discrete Lips.

4.2. Training Results with Concatenated Lips

In the CNN model training performed using joined lips, the process stopped with early stopping ended at 42 epochs (see Figure 12). In the last epochs, validation accuracy starts to decrease, while training accuracy increases. Therefore, if the training continues further, it will be inevitable to achieve a low test accuracy. Similarly, as in the dataset with split lip images, wrong predictions for many classes such as "bitir", "günaydın", "merhaba" and "özür dilerim" in the results of combined lip images were collected in the "afiyet olsun" class, see Figure 13. Apart from that, we can see that the estimations are generally high in the "başla", "görüşmek üzere", "hoş geldiniz", "selam" and "teşekkür ederim" classes and do not predominantly confused with other classes. As seen in the confusion matrix, it is observed in the classification report graph (Figure 14) that the precision, recall and f1-score values of the "afiyet olsun" class are low. To interpret the accuracy percentages of other classes, more balanced results are seen compared to training using split lips.

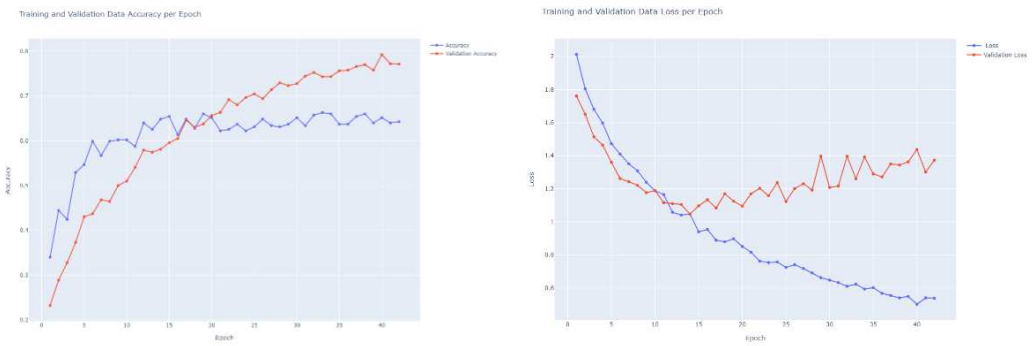

Figure 12. Training and Validation Accuracy and Loss per Epoch with Concatenated Lips.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2023 doi:10.20944/preprints202304.0645.v1

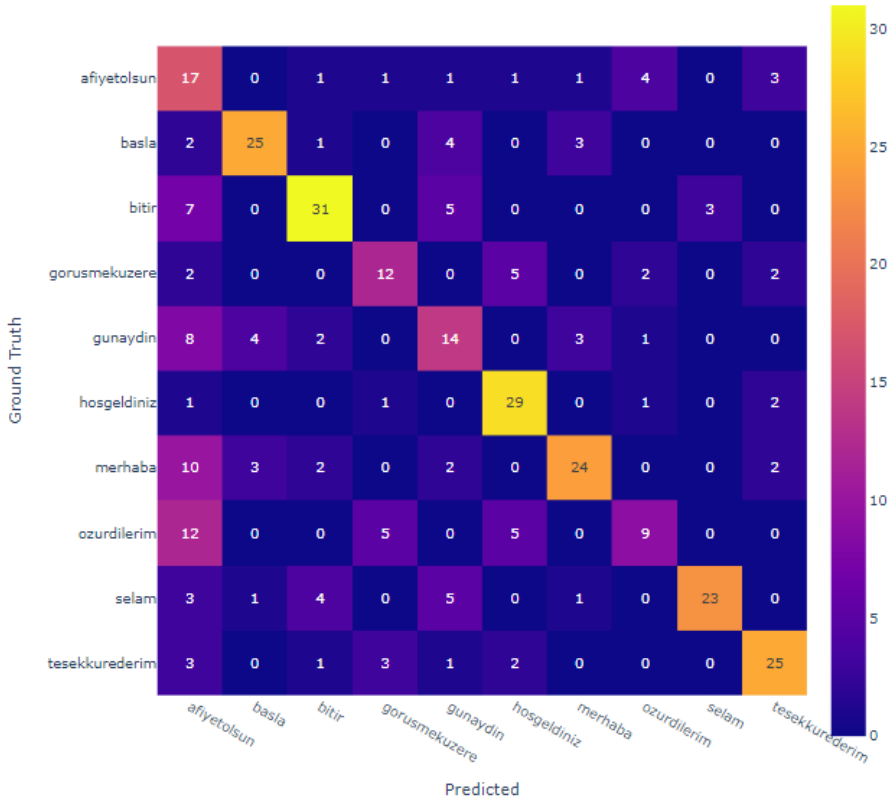

Figure 13. Confusion Matrix of Model Trained with Concatenated Lips.

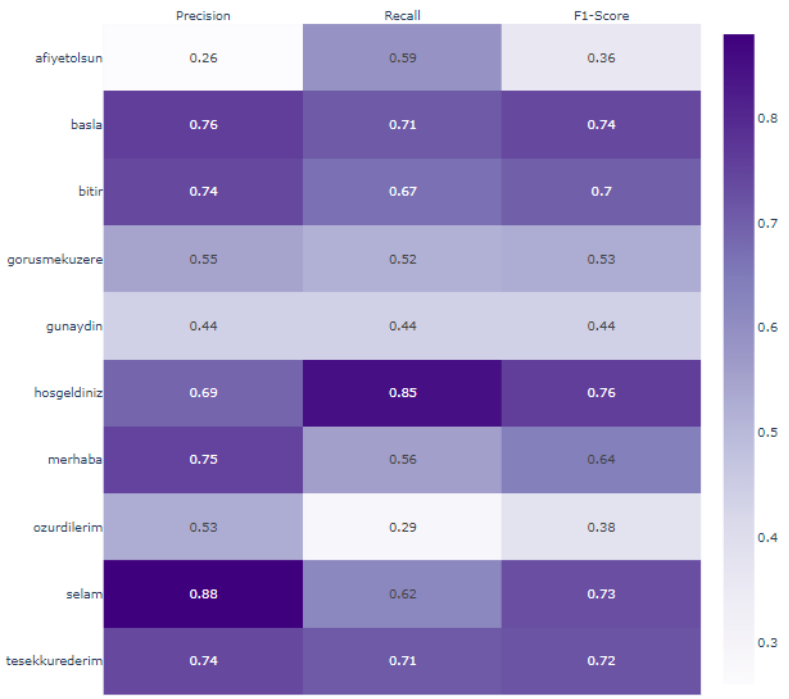

Figure 14. Classification Report of Model Trained with Concatenated Lips.

 doi:10.20944/preprints202304.0645.v1

4.3. Discussion

When the total experimental results are compared, the accuracy is 60.6% for concatenated lips and 61.7% for discrete lips. Again, the times are 18 seconds and 8 minutes, respectively. Since the training time of the discrete lips is long, it can be considered more burdensome in terms of computational cost, but it can be preferred in terms of performance because of its higher accuracy. In terms of image representation, 15 images of 50x50 size are used in one of the inputs, while 1 image of 60x200 size is used in the other. In a situation where simultaneous estimation is required, the use of representation using joined lips would be more appropriate, but for problems where accurate detection is important, the use of the CNN model using split lips is appropriate.

Compared to similar studies, the data contents used in terms of the dataset are quite challenging. In this novel dataset, faces are not viewed from the front, some images are very dark while others are quite bright, and at times it is not possible to accurately detect the face because the background is too mixed.

Table 4. Accuracy and Training Time of Two CNN Models.

	Accuracy	**Training Time**
Concatenated Lips	60.6%	18 seconds
Discrete Lips	61.7%	8 minutes

5. Conclusion

In this study, a CNN model is proposed for a new Turkish dataset. It also compares accuracy and computational cost with two different input representations. In the first of these, sequence lip images form the input of the model separately, while in the other, the lips are combined to form a single image. In terms of performance, split lips look better, but combined lips perform better in terms of time cost. In addition, the Turkish dataset collected from natural Youtube images is also challenging as it is closer to real-world images compared to other studies. The images collected in the studies in the literature were obtained with a fixed background and a fixed human pose by establishing a controlled environment. There is a known dataset that can be evaluated for Turkish, although it has more data, it was also collected in a controlled environment. Automatic lip-reading over natural videos is also of great importance in terms of automatic captioning for hearing-impaired people. In this study, a CNN model is proposed by performing lip reading from natural video images. Since the natural language and lip reading studies in Ural-Altaic languages are shallow, we have contributed with a unique study.

In future studies, it is planned to collect a larger dataset and classify with higher accuracy lip detection.

Author Contributions: Data curation, methodology, visualization, N.P.A; supervision, H.E.; project administration, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data available at https://data.mendeley.com/datasets/4t8vs4dr4v/1.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Networks
RFF	Robust Face Frontalization
SVMs	Support Vector Machines
HMMs	Hidden Markov Models
Bi-LSTM	Bidirectional Long Short-Term Memory
KNN	K-Nearest Neighbor
LRS	Lip Reading Sentences
CTC	Connectionist Temporal Classification
WER	Word Error Rate
BGRU	Bidirectional Gated Recurrent Unit
TCN	Temporal Convolutional Neural Networks
HOG	Histogram of Oriented Gradients
ReLU	Rectified Linear Unit

References

1. McGurk, H.; MacDonald, J. Hearing lips and seeing voices. *Nature* **1976**, *264*, 746–748.
2. Gabbay, A.; Ephrat, A.; Halperin, T.; Peleg, S. Seeing through noise: Speaker separation and enhancement using visually-derived speech. *arXiv* **2017**, arXiv:1708.06767.
3. Stewart, D.; Seymour, R.; Pass, A.; Ming, J. Robust audio-visual speech recognition under noisy audio-video conditions. *IEEE Transactions on Cybernetics* **2013**, *44*, 175–184.
4. Lesani, F.S.; Ghazvini, F.F.; Dianat, R. Mobile phone security using automatic lip reading. In Proceedings of the 2015 9th International Conference on e-Commerce in Developing Countries: With focus on e-Business (ECDC), 2015; pp. 1–5.
5. Mathulaprangsan, S.; Wang, C.Y.; Kusum, A.Z.; Tai, T.C.; Wang, J.C. A survey of visual lip reading and lip-password verification. In Proceedings of the 2015 International Conference on Orange Technologies (ICOT), 2015; pp. 22–25.
6. Sengupta, S.; Bhattacharya, A.; Desai, P.; Gupta, A. Automated lip reading technique for password authentication. *International Journal of Applied Information Systems (IJAIS)* **2012**, *4*, 18–24.
7. Son Chung, J.; Senior, A.; Vinyals, O.; Zisserman, A. Lip reading sentences in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017; pp. 6447–6456.
8. Assael, Y.M.; Shillingford, B.; Whiteson, S.; De Freitas, N. Lipnet: Sentence-level lipreading. *arXiv* **2016**, arXiv:1611.01599.
9. Ephrat, A.; Halperin, T.; Peleg, S. Improved speech reconstruction from silent video. In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017; pp. 455–462.
10. Jaumard-Hakoun, A.; Xu, K.; Leboullenger, C.; Roussel-Ragot, P.; Denby, B. An articulatory-based singing voice synthesis using tongue and lips imaging. *ISCA Interspeech* **2016**, *2016*, 1467–1471.
11. Bocquelet, F.; Hueber, T.; Girin, L.; Savariaux, C.; Yvert, B. Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLoS Computational Biology* **2016**, *12*, e1005119.
12. Gabbay, A.; Shamir, A.; Peleg, S. Visual speech enhancement. *arXiv* **2017**, arXiv:1711.08789
13. Mattos, A.B.; Oliveira, D.A.B. Multi-view mouth renderization for assisting lip-reading. In Proceedings of the 15th International Web for All Conference, 2018; pp. 1–10.
14. Matthews, I.; Cootes, T.F.; Bangham, J.A.; Cox, S.; Harvey, R. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2002**, *24*, 198–213.
15. Cox, S.J.; Harvey, R.W.; Lan, Y.; Newman, J.L.; Theobald, B.J. The challenge of multispeaker lip-reading. In Proceedings of the AVSP, 2008; pp. 179–184.
16. Lee, B.; Hasegawa-Johnson, M.; Goudeseune, C.; Kamdar, S.; Borys, S.; Liu, M.; Huang, T. AVICAR: Audio-visual speech corpus in a car environment. In Proceedings of the Eighth International Conference on Spoken Language Processing, 2004.
17. Wong, Y.W.; Ch'ng, S.I.; Seng, K.P.; Ang, L.M.; Chin, S.W.; Chew, W.J.; Lim, K.H. A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities. *Pattern Recognition Letters* **2011**, *32*, 1503–1510.

18. McCool, C.; Marcel, S.; Hadid, A.; Pietikäinen, M.; Matejka, P.; Cernockỳ, J.; Poh, N.; Kittler, J.; Larcher, A.; Levy, C.; others. Bi-modal person recognition on a mobile phone: using mobile phone data. In Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops, 2012; pp. 635–640.
19. Rekik, A.; Ben-Hamadou, A.; Mahdi, W. A new visual speech recognition approach for RGB-D cameras. In *International Conference Image Analysis and Recognition*; Springer: 2014; pp. 21–28.
20. Estival, D.; Cassidy, S.; Cox, F.; Burnham, D.; et al. *AusTalk: An audio-visual corpus of Australian English*; 2014.
21. Petrovska-Delacrétaz, D.; Lelandais, S.; Colineau, J.; Chen, L.; Dorizzi, B.; Ardabilian, M.; Krichen, E.; Mellakh, M.A.; Chaari, A.; Guerfi, S.; others. The iv 2 multimodal biometric database (including iris, 2d, 3d, stereoscopic, and talking face data), and the iv 2-2007 evaluation campaign. In Proceedings of the 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems, 2008; pp. 1–7.
22. Trojanová, J.; Hrúz, M.; Campr, P.; Železnỳ, M. Design and recording of czech audio-visual database with impaired conditions for continuous speech recognition. 2008.
23. Petridis, S.; Shen, J.; Cetin, D.; Pantic, M. Visual-only recognition of normal, whispered and silent speech. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018; pp. 6219–6223.
24. Kumar, K.; Chen, T.; Stern, R.M. Profile view lip reading. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, 2007; Volume 4, pp. IV–429.
25. Afouras, T.; Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2018**.
26. Fernandez-Lopez, A.; Sukno, F.M. Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing* **2018**, *78*, 53–72.
27. Kang, Z.; Horaud, R.; Sadeghi, M. Robust face frontalization for visual speech recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021; pp. 2485–2495.
28. Zhao, G.; Barnard, M.; Pietikainen, M. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia* **2009**, *11*, 1254–1265.
29. Gurban, M.; Thiran, J.P. Information theoretic feature extraction for audio-visual speech recognition. *IEEE Transactions on Signal Processing* **2009**, *57*, 4765–4776.
30. Atila, Ü.; Sabaz, F. Turkish lip-reading using Bi-LSTM and deep learning models. *Engineering Science and Technology, An International Journal* **2022**, 101206.
31. Yargıç, A.; Doğan, M. A lip reading application on MS Kinect camera. In Proceedings of the 2013 IEEE INISTA, 2013; pp. 1–5.
32. Fung, I.; Mak, B. End-to-end low-resource lip-reading with maxout CNN and LSTM. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018; pp. 2511–2515.
33. Ozcan, T.; Basturk, A. Lip reading using convolutional neural networks with and without pre-trained models. *Balkan Journal of Electrical and Computer Engineering* **2019**, *7*, 195–201.
34. Margam, D.K.; Aralikatti, R.; Sharma, T.; Thanda, A.; Roy, S.; Venkatesan, S.M.; others. LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models. *arXiv* **2019**, arXiv:1906.12170.
35. Petridis, S.; Li, Z.; Pantic, M. End-to-end visual speech recognition with LSTMs. In Proceedings of the 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2017; pp. 2592–2596.
36. Martinez, B.; Ma, P.; Petridis, S.; Pantic, M. Lipreading using temporal convolutional networks. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020; pp. 6319–6323.
37. Berkol, A.; Tümer-Sivri, T.; Pervan-Akman, N.; Çolak, M.; Erdem, H. Visual Lip Reading Dataset in Turkish. *Data* **2023**, *8*, 15.
38. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.