*data*

MDPI

*Data Descriptor*

# Visual Lip Reading Dataset in Turkish

**Ali Berkol [1,*], Talya Tümer-Sivri [1], Nergis Pervan-Akman [1], Melike Çolak [1] and Hamit Erdem [2]**

[1] Defense and Information Systems, BITES, Neighbourhood of Mustafa Kemal, Dumlupınar Avenue, METU Technopolis, Ankara 06530, Turkey
[2] Electrics and Electronics Department, Başkent University, Baglica Campus, Fatih Sultan District, Ankara 06790, Turkey
* Correspondence: ali.berkol@bites.com.tr

**Abstract:** The promised dataset was obtained from daily Turkish words and phrases pronounced by various people in videos posted on YouTube. The purpose of compiling the dataset was to provide a method for the detection of the spoken word by recognizing patterns or classifying lip movements with supervised, unsupervised, and semi-supervised learning, and machine learning algorithms. Most of the datasets related to lip reading consist of people recorded on camera with fixed backgrounds and the same conditions, but the dataset presented here consists of images compatible with machine learning models developed for real-life challenges. It contains a total of 2335 instances taken from TV series, movies, vlogs, and song clips on YouTube. The images in the dataset vary due to factors such as the way people say words, accents, speaking rate, gender, and age. Furthermore, the instances in the dataset consist of videos with different angles, shadows, resolution, and brightness that are not created manually. The most important feature of our lip reading dataset is that we contribute to the non-synthetic Turkish dataset pool, which does not have wide dataset varieties. Machine learning studies can be carried out in many areas, such as education, security, and social life with this dataset.

**Dataset:** https://doi.org/10.17632/4t8vs4dr4v.1.



**Keywords:** lip reading; visual speech recognition; Turkish dataset; face parts detection



**Citation:** Berkol, A.; Tümer-Sivri, T.; Pervan-Akman, N.; Çolak, M.; Erdem, H. Visual Lip Reading Dataset in Turkish. *Data* **2023**, *8*, 15. https://doi.org/10.3390/data8010015

Academic Editor: Muhammad Irfan

Received: 2 November 2022
Revised: 15 December 2022
Accepted: 15 December 2022
Published: 5 January 2023



## 1. Summary

Lip reading is the ability to perceive speech by observing and analyzing lip movements without auditory information. People who are called lip reading experts use this skill to solve judicial events. For instance, it is crucial to understand what a suspected person says from lip movements in camera recordings examined for security issues. In addition, due to the rapid development of deep learning (DL) techniques, researchers have shown great interest in this field. The dataset used in DL applications developed with image processing techniques is important for the real-life performance of the application. The applications developed with a fixed angle light and background data will not be sufficient for environment conditions with real-life variables. Our aim in this study is to provide a new Turkish dataset that will help develop a visual lip reading system that is not affected by real-life difficulties.

When languages are classified according to their structures, Turkish is a language in the family of agglutinative languages. For this reason, according to Turkish grammar rules, suffixes are an issue that significantly affects a sentence's meaning. Furthermore, in Turkish, if a word starting with a vowel comes after a word ending with a consonant letter, the edge effect that occurs when these two letters are connected and read is called liaison. For example, "top aldı" (bought a ball) and "topaldı" (was lame) mean different

things, with liaisons in the letters "p" and "a" despite having the same letter order. Some words' pronunciations in a sentence and some syllables at the level of words in a higher tone than others is called word stress. For example, while a person says the word "afiyet olsun" with a different word stress in a sentence, it may not fully coincide with real-life in cases where only the relevant word is recorded by saying it. In addition, almost every province in Turkey has its own dialect; therefore, the lip movements of the same word may vary in every province. As this article aims to solve a real-life problem, each word or phrase has features such as liaison, word stress, and different regional accents such as East Anatolia, Northeast Anatolia, and West Anatolia. Our dataset was created by hundreds of individuals and includes differences in liaison, word stress, and dialect, unlike the dataset in [1], which was only generated by speakers saying the relevant word or phrase.

In the literature, there are many datasets created with different methods to be used in lip reading studies. However, it has been observed that no Turkish lip reading data have been created in any of the studies except the study of Atila and Sabaz [1]. The authors created two new datasets consisting of words and sentences using image processing techniques. The dataset consisting of sentences includes classes such as "Which department did you get?" and "May I help you?", and the dataset consisting of words includes classes such as "Programmer" and "Video". The point that distinguishes this dataset from ours is that all of the words and sentences are created in the same environment and light conditions. In addition, unlike our data, which we obtained from different YouTube videos for each sample and which includes hundreds of different speaking profiles as a result, these data were obtained from a total of 24 individuals—18 women and 6 men. Atila and Sabaz conducted experiments with convolutional neural network (CNN)- and LSTM-based models on the datasets they obtained and showed that ResNet-18 and Bi-LSTM algorithms gave the best results in both datasets with 84.5% and 88.55% accuracy scores, respectively.

Matthews et al. [2] created their own aligned audio–visual AVLetter database of isolated letters. It consists of three repetitions of all letters of the alphabet by each of 10 speakers—five male (two with mustaches) and five females—for a total of 780 expressions. In the study, internal and external contour methods were used. As a third method, they presented a novel bottom-up approach where features are extracted directly from pixel intensity from nonlinear scale space analysis. In addition, they trained a Hidden Markov Model (HMM) and obtained a 44.6% accuracy score. In [3], two new datasets were introduced and publicly released, namely, LRS2-BBC [4], consisting of thousands of natural phrases from British television, and LRS3-TED [5], which contains hundreds of fragments from over 400 h of TED and TEDx videos [6]. Both datasets contain unrestricted natural language sentences and wild videos from different people, unlike synthetic datasets obtained with standard background, light, and angle conditions. Researchers have shown that combining visual speech recognition (VSR) and audial speech recognition methods, especially in the presence of noise in the voice, leads to a significant improvement in lip reading studies.

Yang et al. [7] presented a natively distributed large-scale benchmark called the LRW-1000, which included 1000 classes with 718,018 samples from more than 2000 speakers. Each class corresponds to syllables of a Mandarin word consisting of one or more Chinese characters. This dataset was created with real-world conditions in mind and showed great variation in this comparison in several aspects, including the number of samples in each class, video resolution, lighting conditions, and speaker characteristics such as pose, age, gender, and makeup. Egorov et al. [8] created a Russian lip reading dataset, LRWR, which included 235 classes and 135 speakers. A detailed description of the dataset aggregation pipeline and dataset statistics is presented in their paper. In this way, they contributed to the visual lip reading dataset studies, which are dominated by English language lip reading studies, by creating a large-scale Russian dataset.

Chung and Zisserman [9] aimed to recognize words spoken by a speaking face without phonetic information. They developed a pipeline for fully automated data collection from TV broadcasts and created a dataset containing examples of more than a million words

spoken by different people. A two-stream convolutional neural network was developed that learns a common insertion between voice and mouth movements from the unlabeled data obtained, and the training results with this dataset and model exceeded the state-of-the-art of the publicly available datasets, namely, Columbia [10] and OuluVS2 [11]. Anina et al. [11] presented OuluVS2, a newly aggregated multi-image audiovisual dataset for non-rigid mouth movement analysis. OuluVS2 consists of more than 50 speakers saying English phrases, numbers, three phrases, and three sentences and thousands of videos recorded simultaneously from five different angles ranging from front to profile views. In addition, a VSR system developed the HMM and tested the database. As a result of the recognition, a 60° angle had the best accuracy score of 46%, while this score was 42% at a 90° angle (front view). The Arabic Visual Speech Dataset (AVSD) [12] contains 1100 videos for 10 daily communication words, for example, hello, welcome, and sorry, collected from 22 speakers. The dataset was taken under realistic conditions inside various rooms with different indoor illuminations. As a result of VSR on AVSD with a support vector machine (SVM), the average word recognition rate of the algorithm was 70.09%. Sujatha and Krishnan [13] prepared a dataset with 10 participants in stable ambient conditions saying 35 different words; 4900 samples (7 participants pronounced 20 samples of each of 35 words) were collected for training, and 2100 samples (3 participants pronounced 20 samples of each one of 35 words) were used for testing. The videos of the participants were given as inputs to the face localization module for the detection of the facial region, and then the mouth region was determined.

In [14], a dataset was created and used for achieving fruitful results for lip reading issues. This data contained interrelated audio and lip movement data in several videos of various contents reading the identical words, for example, book, come, and read. The proposed method employed VGG16 pre-trained CNN architecture for the classification and recognition of data. The accuracy of the recommended model was 76% in VSR. Xu et al. [15] used multi-expansion temporal convolutional networks (MD-TCN) to predict individual words in lip reading tasks. Their method included a self-attention block after each convolution layer to further enhance the model's classification and scanning capabilities. On the LRW dataset [16], their technique achieved an accuracy of 85%, which had a 0.2% increase over other similarly structured networks [17]. Akman et al. [18], using the dataset we proposed in this study, compared the performance of the DCNN model with the CNN model in their previous work. The test accuracy they obtained as a result of the multiclass classification model for words and phrases was 59.80% for dilated CNN (DCNN), and the CNN accuracy value they used in their previous study was 72%. It was stated that CNN outperformed DCNN in time and accuracy. The reason for the lower accuracy score was the use of a non-synthetic dataset with compelling features for the model. Many existing datasets produced for the lip reading problem studies [11–14,19,20] were obtained under controlled conditions. The contribution of our study to the literature is that this dataset is—to our best knowledge—the first non-synthetic Turkish lip reading dataset publicly available. This dataset was obtained by extracting from the natural speech flow of people. The videos used to form data were meticulously examined, and if there was any factor that blocked the lip movements of the person, such as a microphone, subtitles, or hands, they were not included in the sample dataset. The data consisted of faces only to describe lip movement. However, since it consisted of wide-framed images of people pronouncing various words, it can be used for different research problems after the necessary arrangements of the data. This dataset aids in the development of recognizing words or phrases being spoken by a talking face without audio [21] and without lip-motion recognition [16].

## 2. Data Description

This dataset consists of words and phrases commonly used in Turkish: "merhaba" (hello), "selam" (hi), "başla" (start), "bitir" (finish), "günaydın" (good morning), "teşekkür ederim" (thank you), "hoş geldiniz" (welcome), "görüşmek üzere" (see you), "özür dilerim"

(sorry), and "afiyet olsun" (enjoy your meal). There are two topics, namely, having balanced labels, and distribution of each word's frame, that we took into consideration.

Firstly, it was essential to have a balanced multi-class dataset. For example, working with a balanced dataset in terms of labels is less challenging. Thus, developers and researchers can easily focus on developing more optimal and diverse models. In this study, we paid extra attention to having an approximate amount of data for each label. Furthermore, Table 1 shows the number of each class in the dataset. Secondly, the normal distribution of these words' frame numbers in the dataset is crucial for a high-performance machine learning model. As the difference in the number of examples for each class instance is low, model results will give consistent results.

**Table 1.** Number of instances in the dataset.

| Words and Phrases | Number of Instances |
|---|---|
| başla (start) | 225 |
| bitir (finish) | 244 |
| merhaba (hello) | 268 |
| günaydın (good morning) | 232 |
| selam (hi) | 235 |
| hoş geldiniz (welcome) | 226 |
| özür dilerim (sorry) | 209 |
| görüşmek üzere (see you) | 224 |
| afiyet olsun (enjoy your meal) | 235 |
| teşekkür ederim (thank you) | 237 |

Secondly, in addition to the approximate percentage of each class, the number of frames for each word is another crucial point in machine learning models, especially for deep learning. In addition, it can be an effective parameter in recognizing the relevant word in real-time. In Figure 1, the distribution of each class according to frame number can be observed. From top to bottom, the first five labels identify phrases, for example, "teşekkür ederim" and "hoş geldiniz". The rest are words, for instance, "günaydın" and "selam". While the number of frames for a word varied between approximately 3 and 26, this number was between approximately 7 and 33 for phrases.

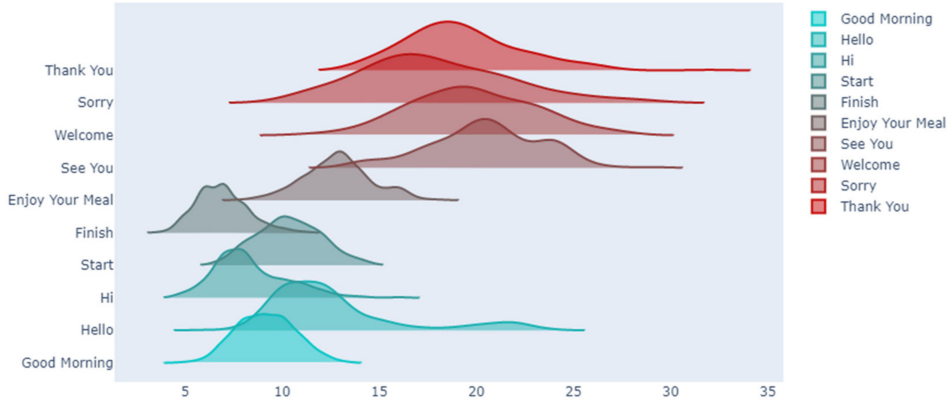


**Figure 1.** Frame number distribution for each word such as "hello" (merhaba), "hi" (selam), "start" (başla), "finish" (bitir), and "good morning" (günaydın) and phrases such as "thank you" (teşekkür ederim), "welcome" (hoş geldiniz), "see you" (görüşmek üzere), "sorry" (özür dilerim), and "enjoy your meal" (afiyet olsun).

In this study, the Pandas skew() method, which returns unbiased skew values, was used to examine the frame number distributions. The skewness coefficients for the words "günaydın", "merhaba", "selam", "başla", and "bitir" were, respectively, 0.06, 1.46, 0.86, 0.09, and 0.54 and for the phrases "afiyet olsun", "görüşmek üzere", "hoş geldiniz", "özür

dilerim", and "teşekkür ederim" were, respectively, 0.10, −0.16, 0.07, 0.48, and 0.72. The words "selam", "merhaba", and "teşekkür ederim" had high skewness coefficient values, so it can be said that the distributions were right skewed, while other words' and phrases' skewness coefficients were close to 0, and their distributions were normal. The frame number's mean of the word "merhaba" was 12.7, the median was 12, and the mode was 10, i.e., it had a normal distribution. The mean, mode, and median values of "günaydın" were, respectively, 9.1, 9, and 9, respectively, i.e., it had a normal distribution. In the case of "merhaba", some of the videos were recorded from children's songs in which speakers spoke relatively slowly, according to other recordings. Moreover, not having normal distributions for some classes showed that speakers were composed of more diverse samples. In addition, they included a variety of video types that we recorded, i.e., vlogs, TV series, or clips.

Finally, a correlation matrix was extracted by taking example sequences for all classes to show if there was a linear relationship between the classes. The steps of finding causal or non-causal relationships were as follows: Firstly, relatively clear examples, which are preferred for accurate results, were selected from the dataset for each class. After that, lips were cut from the original image, since capturing and analyzing the movements of the lips are essential, and provide the ability to work with fewer data. As a next step, we lowered the sequence of arrays to a one-dimensional summarized array by taking the median for each index of images. The Pearson correlation method was applied to finalized arrays for each class. The Pearson correlation coefficient can vary between −1 and 1. If the value is closer to 1, there is a positive relationship between the variables, i.e., they have a positive causal relationship. If the value is closer to −1, there is a negative causal relationship between the variables. If the value is closer to 0, both from the negative and positive sides, it can be concluded that there is a non-causal relationship between the variables. In other words, there is no linear relationship.

In Figure 2, we illustrated the Pearson correlation using a heatmap. As can be seen, some classes had high positive correlations. For example, "afiyet olsun" and "günaydın" were highly correlated and had correlation values of approximately 0.9. Similarly, "merhaba" and "başla" had a correlation value of approximately 0.6. However, we observed no strong relationship between classes in general. Additionally, there was no strong negative correlation such as we had in the positive examples. According to the method applied for linear relationships on specific examples, it is important to highlight that the dataset is useful for solving problems such as classification, since the patterns are different and solvable for various methods, including deep learning and machine learning algorithms.

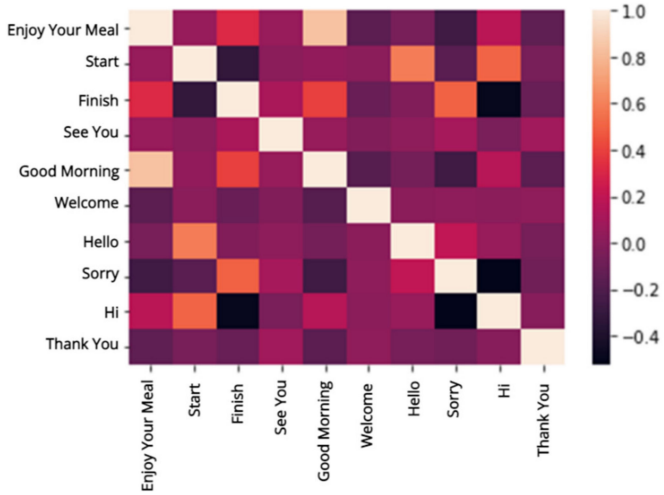


**Figure 2.** Distance matrix for each class such as "hello" (merhaba), "hi" (selam), "start" (başla), "finish" (bitir), "good morning" (günaydın), "thank you" (teşekkür ederim), "welcome" (hoş geldiniz), "see you" (görüşmek üzere), "sorry" (özür dilerim), and "enjoy your meal" (afiyet olsun) based on the image features.

## 3. Methods

Since the dataset was shaped from very raw to ready to use, we performed some important steps. These steps were video recording, frame extracting, and cropping noisy data from frames. Details are described in the following sections.

### *3.1. Dataset Collection*

The data collection process started with the detection of YouTube [22] videos with the specified words and screen recording. While collecting the data, we gave extra importance to creating samples using a wide variety in terms of male/female, adult/child/old, outdoor/indoor, light/dark, with/without a mustache, with/without makeup, and face position with a slight angle. The second when the word in the video was spoken for the first time was attempted to be selected so as not to include the lip movements of other words as much as possible. Then, according to the determined second, the frames were converted to an image with a simple code to be extracted, as explained in the next section. In many screen-recorded videos, where the lip image was not included and it was prevented from appearing on various objects, it was eliminated. For example, there were many situations such as hand movements blocking the face, the image of the lip that protruded from the field of view at one point in the word, or the default subtitle covering the lip.

### *3.2. Frame Extraction from Videos*

After 2335 instances were collected, they were split into frames with the help of the Python library, OpenCV. While splitting the videos into frames, a script was written that took the second where the word started and determined the fps (frame per second) of the video. Then the frames within 2 s after the determined second were recorded and saved as images. These images produced differed according to the fps value. In general, since the fps value of the videos we produced was 30, 60 frames were obtained for every 2 s block.

Knowing the directory structure of the dataset is helpful for understanding and reading the data. In the directory hierarchy, the first folder starts with a word or phrase tag, such as "başla", or "teşekkür ederim". A subdirectory of each word contains the instance names, and the instances are named by three-digit sequential numbering. The latest file of the dataset architecture contains the processed frames of the relevant video and these frames are named sequentially with two-digit numbering such as "01.jpg, 02.jpg, . . . , 28.jpg". Figure 3 shows the directory architecture of the dataset.

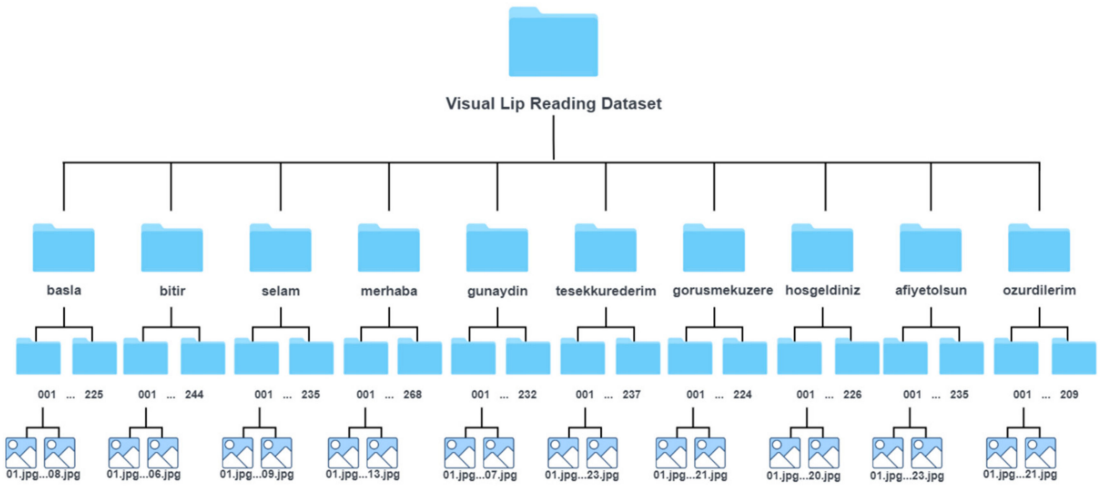


**Figure 3.** The directory architecture of the dataset; "merhaba" (hello), "selam" (hi), "başla" (start), "bitir" (finish), "günaydın" (good morning), "teşekkür ederim" (thank you), "hoş geldiniz" (welcome), "görüşmek üzere" (see you), "özür dilerim" (sorry), and "afiyet olsun" (enjoy your meal) are words and phrases that appeared in the first step. A subdirectory has samples of words and phrases contained within it. The last step of the architecture shows the frames of the related word.

### *3.3. Frame Cropping*

Having more than one human face in the same frame will cause complexity in identifying the person who is speaking and whose lips should be read. Therefore, the entire dataset was scanned, and the images with multiple faces were cropped, except for the human face that was the subject of consideration. This step was performed manually using image cropping applications. Attention was taken to ensure that the face of the person speaking the relevant word was entirely within the field of view while the clipping process was performed. Frames in which the lip movements of the speaker were clearly visible, with no other face in the frame, no object preventing lip movement, and no profile view, were included in the dataset. Preserving the background and obtaining real-world instances without removing noise were emphasized while cropping the related part from noisy images.

## 4. Conclusions

When previous studies were examined, only one dataset on Turkish lip reading was found. What distinguishes this study from the dataset [1], where all words and sentences were created under the same ambient and light conditions, is that it is a non-synthetic lip reading dataset that has not been created before. The methods of obtaining the data showed completely different approaches in the two studies. While the data are obtained by 24 speakers who say only certain words and phrases in [1], in our dataset for each sample, the seconds in which the relevant word occurs from the sentence of different people's YouTube videos are obtained. In addition, the way the words are pronounced in the Turkish language is affected by many factors such as the speaker's accent, and whether he/she uses liaison or word stress. For this reason, the aim is to create a dataset suitable for real-life conditions by collecting samples from as many people as possible in the study. The dataset we created contributes to visual lip reading studies and allows the developed studies to produce more realistic results due to the complex environmental conditions in real life. By developing lip reading studies with this dataset, researchers can help solve forensic cases, facilitate the lives of hearing-impaired people, and offer an innovative approach to language education. This dataset consists of faces only to describe lip movement. However, since it consists of wide framed images of people pronouncing various words, it can be used for different research problems after necessary arrangements on the data.

**Author Contributions:** Video recording, T.T.-S. and N.P.-A.; framing, T.T.-S., N.P.-A. and M.Ç.; data curation, M.Ç.; software, N.P.-A.; visualization, M.Ç.; writing—original draft, M.Ç.; writing—review and editing, T.T.-S. and N.P.-A.; project administration, A.B.; funding acquisition, A.B.; supervision, T.T.-S., N.P.-A., A.B. and H.E. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data available at https://doi.org/10.17632/4t8vs4dr4v.1 (accessed on 6 July 2022).

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

## References

1. Atila, Ü.; Sabaz, F. Turkish lip-reading using Bi-LSTM and deep learning models. *Eng. Sci. Technol. Int. J.* **2022**, *35*, 101206. [CrossRef]
2. Matthews, I.; Cootes, T.; Bangham, J.; Cox, S.; Harvey, R. Extraction of visual features for lipreading. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 198–213. [CrossRef]
3. Afouras, T.; Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *1*, 1. [CrossRef] [PubMed]

4. Lip Reading Sentences 2 (LRS2) Dataset. Available online: https://www.robots.ox.ac.uk/%7Evgg/data/lip_reading/lrs2.html (accessed on 23 September 2022).
5. Lip Reading Sentences 3 (LRS3) Dataset. Available online: https://www.robots.ox.ac.uk/%7Evgg/data/lip_reading/lrs3.html (accessed on 23 September 2022).
6. Afouras, T.; Chung, J.S.; Zisserman, A. LRS3-TED: A large-scale dataset for visual speech recognition. *arXiv* **2018**, arXiv:1809.00496.
7. Yang, S.; Zhang, Y.; Feng, D.; Yang, M.; Wang, C.; Xiao, J.; Long, K.; Shan, S.; Chen, X. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019; pp. 1–8.
8. Egorov, E.; Kostyumov, V.; Konyk, M.; Kolesnikov, S. LRWR: Large-scale benchmark for lip reading in Russian language. *arXiv* **2021**, arXiv:2109.06692.
9. Chung, J.S.; Zisserman, A. Learning to lip read words by watching videos. *Comput. Vis. Image Underst.* **2018**, *173*, 76–85. [CrossRef]
10. Chakravarty, P.; Tuytelaars, T. Cross-modal supervision for learning active speaker detection in video. In *Lecture Notes in Computer Science, Proceedings of the 14th European Conference on Computer Vision 2016 (ECCV), Amsterdam, The Natherlands, 8–16 October 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016; Volume 9909, pp. 285–301.
11. Anina, I.; Zhou, Z.; Zhao, G.; Pietikainen, M. OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis. In Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; pp. 1–5.
12. Elrefaei, L.A.; Alhassan, T.Q.; Omar, S.S. An Arabic visual dataset for visual speech recognition. *Procedia Comput. Sci.* **2019**, *163*, 400–409. [CrossRef]
13. Sujatha, P.; Krishnan, M.R. Lip feature extraction for visual speech recognition using hidden Markov model. In Proceedings of the 2012 International Conference on Computing, Communication and Applications, Dindigul, India, 22–24 February 2012; pp. 1–5.
14. Shashidhar, R.; Patilkulkarni, S. Visual speech recognition for small scale dataset using VGG16 convolution neural network. *Multimed. Tools Appl.* **2021**, *80*, 28941–28952.
15. Xu, B.; Wu, H. Lip reading using multi-dilation temporal convolutional network. *CONF-SPML Signal Process. Mach. Learn.* **2022**, *3150*, 50–59.
16. Chung, J.S.; Zisserman, A. Lip reading in the wild. In *Lecture Notes in Computer Science, Proceedings of the 13th Asian Conference on Computer Vision 2016 (ACCV), Taipei, Taiwan, 20–24 November 2016*; Lai, S.H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Springer: Cham, Switzerland, 2017; Volume 10112, pp. 87–103.
17. Feng, D.; Yang, S.; Shan, S.; Chen, X. Learn an effective lip reading model without pains. *arXiv* **2020**, arXiv:2011.07557.
18. Akman, N.P.; Sivri, T.T.; Berkol, A.; Erdem, H. Lip reading multiclass classification by using dilated CNN with Turkish dataset. In Proceedings of the 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), Prague, Czech Republic, 20–22 July 2022; pp. 1–6.
19. Cooke, M.; Barker, J.; Cunningham, S.; Xu, S. *The Grid Audio-Visual Speech Corpus (1.0)*; Zenodo: Geneva, Switzerland, 2006.
20. Rekik, A.; Ben-Hamadou, A.; Mahdi, W. A new visual speech recognition approach for RGB-D cameras. In *Lecture Notes in Computer Science, Proceedings of the 11th International Conference on Image Analysis and Recognition (ICIAR 2014), Vilamoura, Portugal, 22–24 October 2014*; Campilho, A., Kamel, M., Eds.; Springer: Cham, Switzerland, 2014; Volume 8815, pp. 21–28.
21. Desai, D.; Agrawal, P.; Parikh, P.; Soni, P.K. Visual Speech Recognition. *Int. J. Eng. Res. Technol.* **2020**, *9*, 601–605.
22. YouTube. Available online: https://www.youtube.com/ (accessed on 17 October 2022).