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, Building no: 280/G, Floor:4, Nu:411, METU Technopolis, Ankara, PC. 06530 TURKEY
[2] Electrics and Electronics Department, Başkent University, Baglica Campus, Fatih Sultan District, 18.km of Eskisehir road, Ankara, PC. 06790 TURKEY
* Correspondence: talya.tumer@bites.com.tr

Abstract: The promised dataset was obtained from the daily Turkish words and phrases pronounced by various people in the videos posted on YouTube. The purpose of collecting the dataset is to provide detection of the spoken word by recognizing patterns or classifying lip movements with supervised, unsupervised, semi-supervised learning and machine learning algorithms. Most of the datasets related with 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, accent, 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 the defense industry and social life, with this dataset.

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

Dataset License: CC BY 4.0

Keywords: Lip reading; Visual speech recognition; Turkish dataset; Face parts detection

1. Summary

Lip reading is the ability of speaking from observing and analyzing lip movements without auditory information. People who are called lip reading experts use this skill to solve judicial events. For example, it is important to understand what the suspected person says from the lip movements in the camera recordings examined for the security issues. In addition, due to the rapid development of Deep Learning (DL) techniques, the researchers show 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 on real-life variable environment conditions. 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 the 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." (She bought a ball.) and "Topaldı." (She was lame.) these sentences mean different things, with liaison in the letters "p" and "a" despite having the same letter order. Some words' pronunciation 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. Also, 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, which includes differences in liaison, word stress, and dialect, unlike [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, in the studies conducted, Turkish lip reading data was not created in any other study than [1]. Atila and Sabaz 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?", "May I help you?", and the dataset consisting of words includes classes such as "Programmer", "Video". The point that distinguishes the dataset from us 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 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.

Matthews et al. [2] they 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 present a novel bottom-up approach that features were extracted directly from pixel intensity from nonlinear scale space analysis. Also, they trained a hidden Markov model and got a 44.6% accuracy score. In [3], two new datasets have been introduced and publicly released. LRS2-BBC [4], consisting of thousands of natural phrases from British television and LRS3-TED [5], contains hundreds of fragments from over 400 hours 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 the 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.

Chung and Zisserman [7] 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 has been 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 exceed the state of the art on the publicly available datasets Columbia [8] and OuluVS2 [9]. Anina et al. [9] presents 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 and profile view. In addition, a VSR system was developed with the Hide Markov Model and tested on the database. As a result of the recognition, 60° angle has the best accuracy score with 46%, while this score is 42% at 90° angle (front view). Arabic Visual Speech Dataset (AVSD) [10] contains 1100 videos for 10 daily communication words for example hello, welcome, sorry collected from 22 speakers. The dataset was taken under realistic conditions inside various rooms in different indoor illuminations. As a result of VSR on AVSD with a Support Vector Machine, the average word recognition rate of the algorithm is 70.09%. Sujatha and Krishnan [11] prepared a dataset with 10 participants in stable ambient conditions saying 35 different words. 4900 samples (7 participants pronounced 20 samples of each one 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 input to the face localization module for the detection of the facial region, and then the mouth region was determined.

In [12], they created a dataset and they used it 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 employs VGG16 pre-trained Convolutional Neural Network (CNN) architecture for classification and recognition of data. The accuracy of the recommended model is 76% in VSR. Binyan Xu et al. [13], used Multi-Expansion Temporal Convolutional Networks (MD-TCN) to predict individual words in lip reading tasks. Their method includes a self-attention block after each convolution layer to further enhance the model's classification and scanning capabilities. On the LRW dataset [14], their technique achieves an accuracy of 85%, which has a 0.2% increase over other similarly structured networks [15]. Akman et al. [16], 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 had obtained as a result of the multiclass classification model for words and phrases is 59.80% for Dilated CNN (DCNN) and 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 [9][10][11][12][17][18] studies are obtained under controlled conditions. The contribution of our study to the literature, 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 have been meticulously examined, and if there is any factor that will block the lip movements of the person, like microphone, subtitles, or hands, they are not included in the sample dataset. The data 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. This dataset aids development in recognizing words or phrases being spoken by a talking face without the audio [19] and lip-motion recognition [14].

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, having balanced labels and distribution of each word’s frame, that we have taken into consideration.

Firstly, it is essential to have a balanced multi-class dataset. For example, working with a balanced dataset in terms of labels is less challenging. So, developers and researchers can easily focus on developing more optimal and diverse models. In this study, we pay extra attention to having an approximate amount of data for each label (see Figure 1). 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.

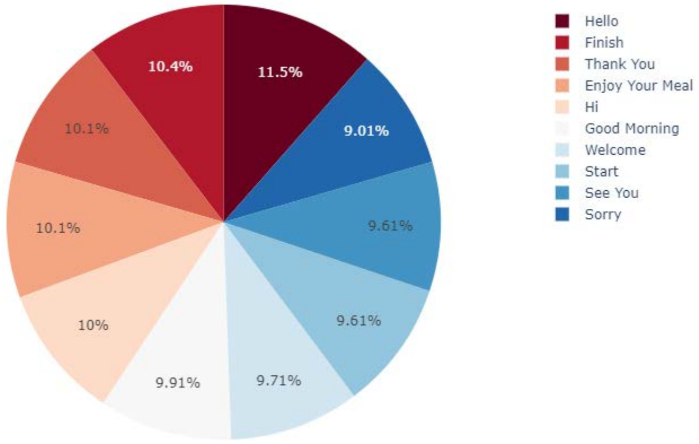


Figure 1. The size of each class in the dataset. These classes are, "hello" (merhaba), "hi" (selam), "start" (başla), "finish" (bitir), "good morning" (günaydın), "thank you" (teşekkür ederim), "welcome" (hoş geldiniz), "see you" (görüşmek üzere),"sorry" (özür dilerim) and "enjoy your meal" (afiyet olsun).

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. Also, it can be an effective parameter in recognizing the relevant word in real-time. In Figure 2, 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". Rest are words, for instance, "günaydın" and "selam". While the number of frames for a word varies between approximately 3-26, this number is between approximately 7-33 for phrases. Figure 2 shows that phrase classes have normally distributed numbers of frames. On the other hand, for "selam", and "merhaba", we have right-skewed distribution and normal distribution for others. In the case of "merhaba", some of the videos are recorded from children's songs in which speakers speak relatively slowly, according to other recordings. Moreover, not having normal distribution for some classes shows that speakers are composed of more diverse samples. Also, it presents a variety of video types that we record, i.e., vlogs, TV series, or clips.

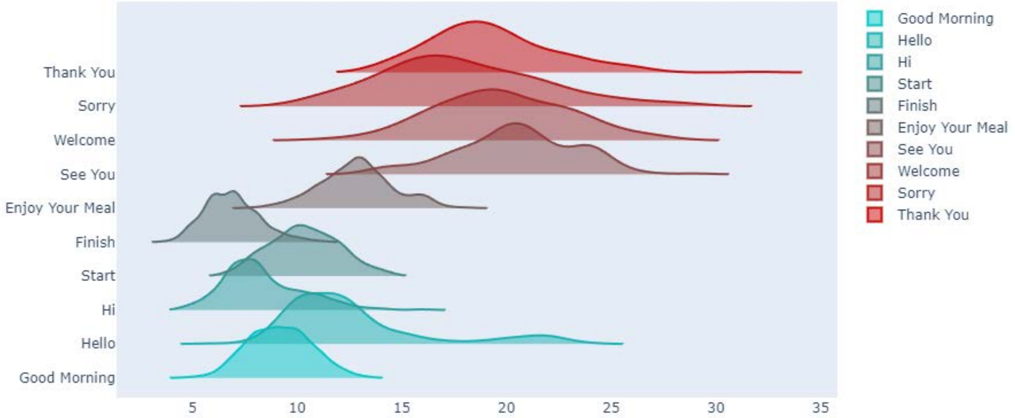


Figure 2. Frame number distribution for each word such as "hello" (merhaba), "hi" (selam), "start" (başla), "finish" (bitir), "good morning" (günaydın), and phrases such as "thank you" (teşekkür ederim), "welcome" (hoş geldiniz), "see you" (görüşmek üzere),"sorry" (özür dilerim) and "enjoy your meal" (afiyet olsun).

Finally, a correlation matrix is extracted by taking example sequences for all classes to show if there is a linear relationship between the classes. The steps of finding causal or non-causal relationships are as follows. Firstly, relatively clear examples, which are preferred for accurate results, are selected from the dataset for each class. After that, lips are cut from the original image since capturing and analyzing the movements of the lips are essential, and it provides working with fewer data. As a next step, we lower the sequence of arrays to a one-dimensional summarized array by taking the median for each index of images. The Pearson correlation method is 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 3, we illustrated the Pearson correlation using a heatmap. As it can be seen, some classes have a high positive correlation. For example, "afiyet olsun" and "günaydın" are highly correlated and have correlation values of approximately 0.9. Similarly, "merhaba" and "başla" have a correlation value of approximately 0.6. However, we observe no strong relationship between classes in general. Additionally, there is no strong negative correlation like we have in 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 like classification since the patterns are different and solvable for various methods, including deep learning and machine learning algorithms.

3. Methods

Since the dataset is shaped from very raw to ready to use, we performed some important steps. These steps are video recording, frame extracting, and cropping noisy data from frames. Details are described in the following sections and Figure 4 shows an example of the steps.

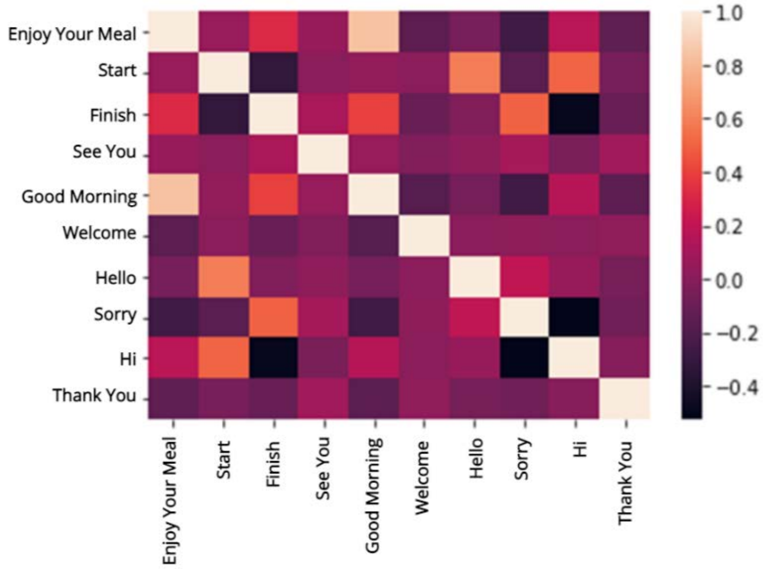


Figure 3. Distance matrix for each class such as "hello" (merhaba), "hi" (selam), "start" (başla), "finish" (bitir), "good morning" (günaydın), "thank you" (teşekkür ederim), "welcome" (hoş geldiniz), "see you" (görüşmek üzere),"sorry" (özür dilerim) and "enjoy your meal" (afiyet olsun) based on the image features.

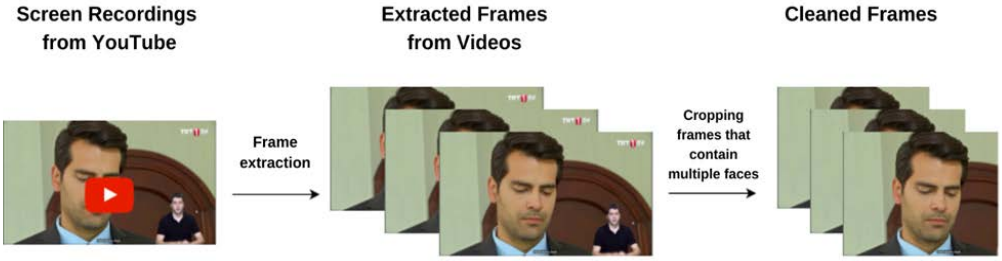

Screen Recordings from YouTube
Extracted Frames from Videos
Cleaned Frames
Frame extraction
Cropping frames that contain multiple faces

Figure 4. An example of data preprocess steps.

3.1. *Dataset collection*

The data collection process started with the detection of YouTube [20] 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 is spoken for the first time has been tried 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 is not included, and it is prevented from appearing on various objects, it has been eliminated. For example, there are many situations such as hand movements blocking the face, the image of the lip that protrudes 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 popular python library OpenCV. While splitting the videos into frames, a function was written that takes the second where the word starts and determines the fps (frame per second) of the video. Then the frames within 2 seconds after the determined second are recorded and saved as images. These images produced differ according to the fps value. In general, since the fps value of the videos we produce is 30, 60 frames are obtained for every 2-second block. The frame extraction part seen in Figure 4 corresponds to the process of extracting and filling the frames made in this step.

Knowing the directory structure of the dataset is helpful in 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. In the lowest file of the dataset, there are the processed frames of the related video and these frames are named sequentially with two-digit numbering. Figure 5 shows directory architecture of the dataset.

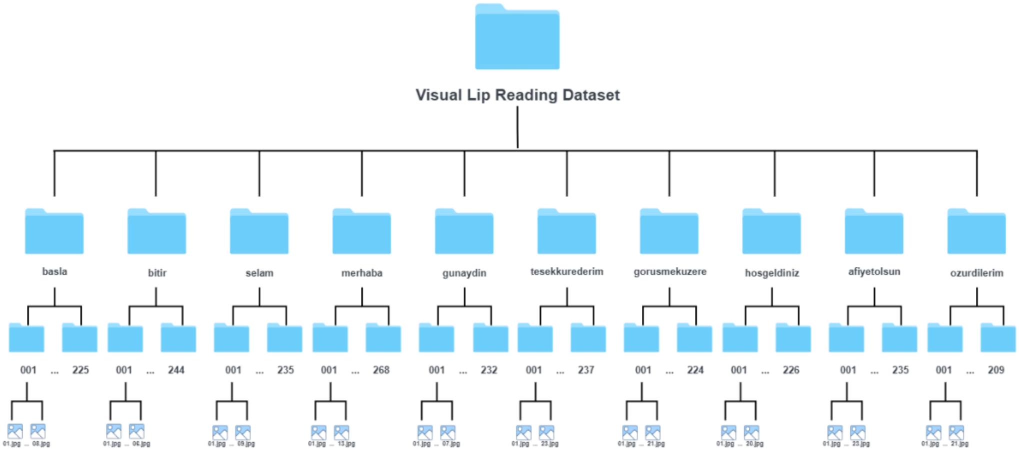
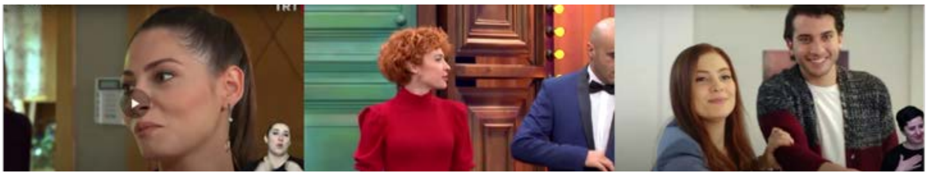


Figure 5. The directory architecture of the dataset. “Merhaba” (hello), “selam” (hi), “başla” (start), “bitir” (finish), “günaydın” (good morning), “teşekkür ederim” (thank you), “hoş geldiniz” (welcome), “görüşmek üzere” (see you), “özür dilerim” (sorry) and “afiyet olsun” (enjoy your meal) are words and phrases which appearing in the first step. A subdirectory are samples of words and phrases contained in 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 cut out, except for the human face, which should be considered. 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 are clearly visible, as shown in Figure 6 (a), with no other face in the frame, no object preventing lip movement, and no profile view are included in the dataset. In contrast, the frames in Figure 6 (b) are not included in the dataset due to the reasons mentioned above. Preserving the background and obtaining real-world instances without removing noise are emphasized while cropping the related part from noisy images. The cropping frame's part seen in Figure 4 corresponds to cutting out the frames in a single face. The final version of the dataset is shown in Figure 7.

(**a**)



(**b**)

Figure 6. Appropriate and inappropriate frames in the dataset. (**a**) Frame samples where lip movements are not blocked. (**b**) Frame samples where lip movements are blocked for different reasons such as having a profile view or having a button in front of the lip.

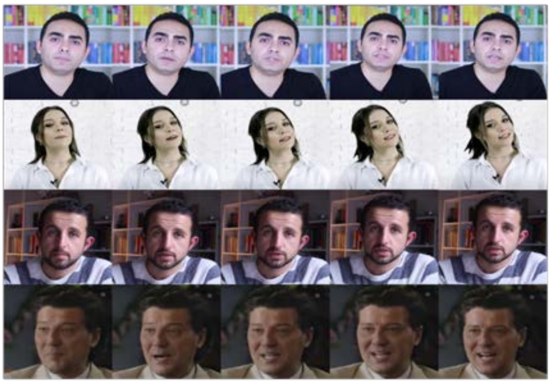


Figure 7. The content of the dataset.

4. Conclusion

When previous studies were examined, only one dataset on Turkish lip reading was found. What distinguishes this study from the [1] dataset, 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 is 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, whether he uses liaison or word stress. For this reason, it is aimed to create a dataset suitable for real-life conditions by collecting samples from as many people as possible in the study. The dataset we have 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., N.P.A; framing, T.T.S, N.P.A, M.Ç.; data curation, M.Ç.; software, N.P.; visualization, M.Ç.; writing - original draft, M.Ç.; writing - review & edit, T.T.S, N.P.A.; project administration, A.B.; funding acquisition, A.B.; supervision, T.T.S, N.P.A, A.B., H.E.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://doi.org/10.17632/4t8vs4dr4v.1

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. Engineering Science and Technology, an International Journal, 101206. 2022 [Google Scholar] [CrossRef]
2. Matthews, I.; Cootes, T.; Bangham, J.; Cox, S.; Harvey, R. Extraction of visual features for lipreading. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002, 24, 2. 198–213. [Google Scholar] [CrossRef]
3. Afouras, T.; Chung, J. S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep Audio-visual Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019, 1,1. [Google Scholar] [CrossRef]
4. Lip Reading Sentences 2 (LRS2) dataset. Available online: https://www.robots.ox.ac.uk/%7Evgg/data/lip_reading/lrs2.html (accessed on 23/ 9/ 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/ 9/ 2022)
6. Afouras, T.; Chung, J. S.; Zisserman, A; LRS3-TED: a large-scale dataset for visual speech recognition. 2018, arXiv preprint arXiv:1809.00496. [Google Scholar] [CrossRef]
7. Chung, J. S.; Zisserman, A., Learning to lip read words by watching videos. Computer Vision and Image Understanding. 2018 173, 76–85. [Google Scholar] [CrossRef]
8. Chakravarty, P.; Tuytelaars, T., Cross-modal supervision for learning active speaker detection in video. In European Conference on Computer Vision. 2016, 285-301, Springer, Cham. [Google Scholar] [CrossRef]
9. Anina, I.; Zhou, Z.; Zhao, G.; Pietikainen, M.,OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis. 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). 2015 [Google Scholar] [CrossRef]
10. Elrefaei, L. A.; Alhassan, T. Q.; Omar, S. S., An Arabic Visual Dataset for Visual Speech Recognition. Procedia Computer Science. 2019, 163, 400–409. [Google Scholar] [CrossRef]
11. Sujatha, P.; Krishnan, M. R, Lip feature extraction for visual speech recognition using Hidden Markov Model. 2012 International Conference on Computing, Communication and Applications. 2012. [Google Scholar] [CrossRef]
12. R, S.; Patilkulkarni, S., Visual speech recognition for small scale dataset using VGG16 convolution neural network. Multimedia Tools and Applications. 2021, 80, 19, 28941–28952. [Google Scholar] [CrossRef]
13. Xu, B.; Wu, H., Lip Reading Using Multi-Dilation Temporal Convolutional Network. CONF-SPML Signal Processing and Machine Learning. 2022, 3150, 50-59 [Google Scholar] [CrossRef]
14. Chung, J. S.; Zisserman, A, Lip Reading in the Wild. Computer Vision – ACCV. 2016, 87–103. [Google Scholar] [CrossRef]
15. Feng, D.; Yang, S.; Shan, S.; Chen, X., Learn an effective lip reading model without pains. 2020. arXiv preprint arXiv:2011.07557 [Google Scholar] [CrossRef]
16. Akman, N. P.; Sivri, T. T.; Berkol, A.; Erdem, H., Lip Reading Multiclass Classification by Using Dilated CNN with Turkish Dataset. 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET). 2022. [Google Scholar] [CrossRef]
17. Cooke; Martin; Barker; Jon; Cunningham; Stuart; Shao Xu., The Grid Audio-Visual Speech Corpus (1.0), Data set. Zenodo. 2006 [CrossRef]
18. Rekik, A.; Ben-Hamadou, A.; Mahdi, W., A New Visual Speech Recognition Approach for RGB-D Cameras. Lecture Notes in Computer Science, 2014, 21–28. [Google Scholar] [CrossRef]
19. Desai D.; Agrawal P., Parikh P.; Soni P. K., Visual Speech Recognition. International Journal of Engineering Research & Technology (IJERT), 2020, 9, 4. 601-605. [Google Scholar] [CrossRef]
20. YouTube. Available online: https://www.youtube.com/ (accessed on 17/10/2022)