International Scientific Journal of Engineering and Management
Volume: 02 Issue: 04 | April – 2023 DOI:
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
www.isjem.com

Silent Speech Database For Mobile Interaction

Rutveej Rajendra Jadhav[1]
Electronics & Telecommunication
JSPM'S Rajarshi Shahu College of Engineering
Pune,India
jadhavrutveej@gmail.com

Soham Sopan Haran[2]
Electronics & Telecommunication
JSPM'S Rajarshi Shahu College of Engineering
Pune,India
sohamharan@gmail.com

Aman Shahid Pathan[3]
Electronics & Telecommunication
JSPM'S Rajarshi Shahu College of Engineering
Pune,India
amanpathan2626@gmail.com

Mrs. Shilpa Sonawane
Professor Department of Electronics & Telecommunication JSPM's Rajarshi Shahu College Of Engineering,
Pune, India

Abstract:

We have developed a module called Silent Speech that enables users to control their smartphone with their lip movements. Our module analyzes the user's lip images and matches them with a predefined dataset to produce an output. Our system supports 26 commands for various tasks such as launching apps, adjusting settings and handling popups. We evaluated the performance of Silent Speech with user tests and measured its recognition accuracy. We also compared it with voice and gesture commands in terms of privacy and usability. We found that Lip-Interact can help users interact with their phone discreetly, communicate with others without disturbing them, and free their hands for other activities.

Keywords: Silent Speech, Data collection, Lip-reading

Introduction:

Silent Speech Mobile Interaction refers to a technology that enables users to interact with their mobile devices without speaking out loud or making any audible sounds. Instead, it uses silent speech recognition techniques to interpret the user's internal speech or subvocalizations. The idea behind silent speech interaction is to capture and analyze the electrical signals generated by the muscles involved in speech

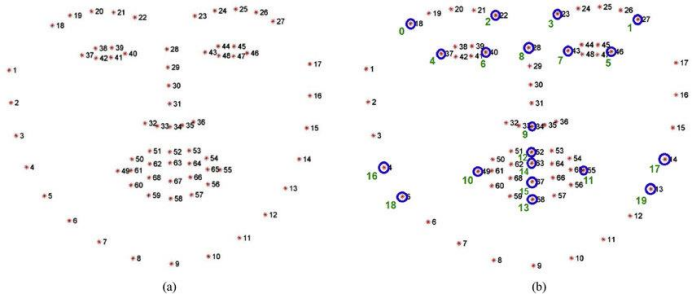

Figure 1.1

production, such as those in the throat, tongue, and face. These signals are then processed by specialized algorithms and converted into understandable commands or text input for the mobile device. Dlib is a C++ library that provides a set of tools and algorithms for machine learning, computer vision, and image processing tasks. One of its most popular features is its ability to perform face detection and facial landmark detection, which includes lip ***detection. Face*** detection in dlib is achieved using a Histogram of Oriented Gradients (HOG) based object detection algorithm. This algorithm searches for patterns in the image that correspond to the shape of a face. It then uses a sliding window approach to detect faces at different scales and orientations. Once a face is detected, dlib's facial landmark detection algorithm is used to identify key points on the face, such as the corners of the eyes, nose, and mouth. This algorithm is based on a set of 68 points that are pre-defined and trained on a large dataset of facial images. One of these 68 points corresponds to the center of the upper lip, and another corresponds to the center of the lower lip.Using these 68 points, dlib can also perform more specific tasks such as lip detection.

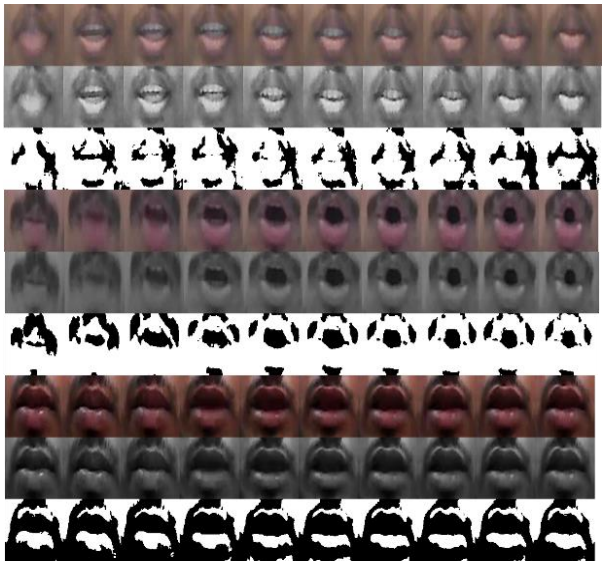

Figure 1.2

ISSN: 2583-6129
www.isjem.com

Lip detection in dlib involves identifying the position and shape of the lips in a given image or video frame. This can be useful in applications such as lip reading or emotion recognition.

Overall, dlib's face detection and facial landmark detection algorithms are highly accurate and widely used in industry and academia. The library is also open-source, which makes it accessible to developers and researchers around the world

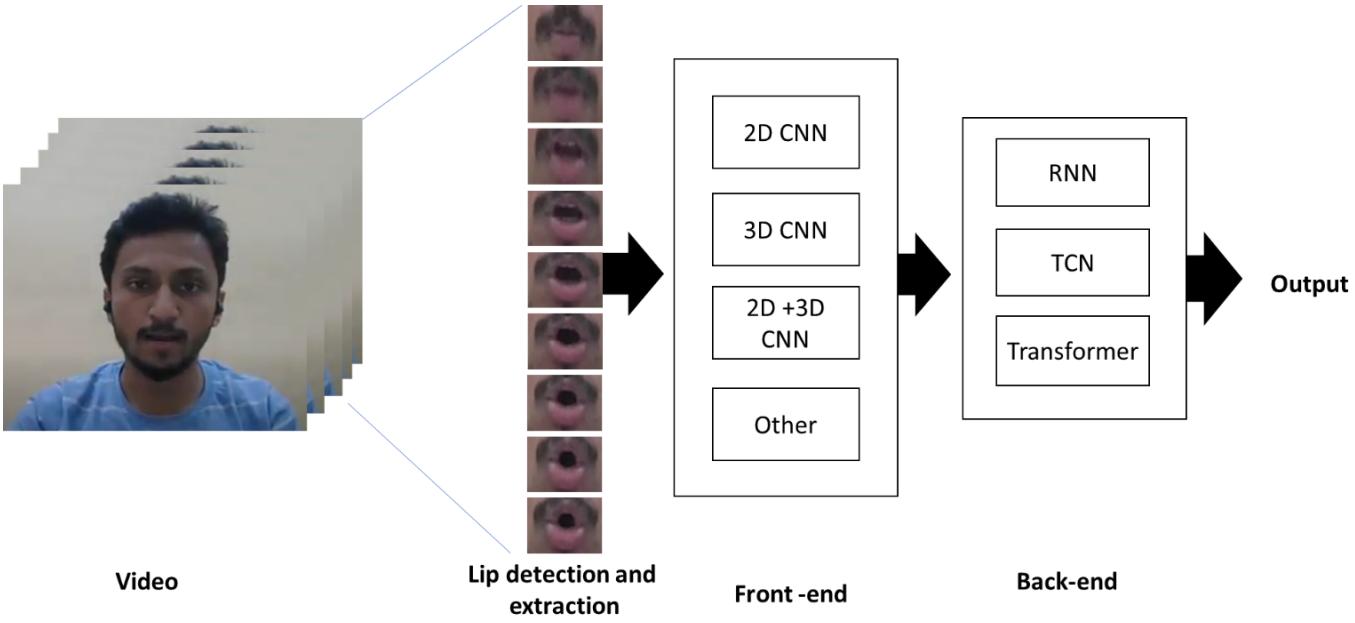

Figure 1.3

Model Architecture:

We use CNN and LSTM with several dense layer for lip reading. For lip reading, must realize the change of lip shape. So, we choose CNN and LSTM. CNN is realize lip shape and then each of the output of CNN become transform to sequence. Finally, LSTM realize pattern the sequence that contain the change of lip shape. Using this difference of sequences to classification. The above picture is simple architecture of our model. First, we transform the lip video to several frame images. We used the several frame images as input. Each fame image passes through the already trained CNN architecture. And then the output of CNN passes through Dense layer for transforming to LSTM layer input. The output of LSTM layer become next dense layer input. Finally, receive the output label by softmax activation function. Our train region is dense layer of CNN architecture to end part. This region is conceived in the above picture. We judged that the extraction feature of lip shape is almost same as the extraction feature of appearance. So, we decide using the transfer learning form the ImageNet trained model. And then only train the dense layer in the CNN. We choose the MobileNet for transfer learning model. MobileNet is compact model and user can resize within regular range. When we use VGG model, take about 3 hours for one epochs on our dataset and error occurred in LSTM layer. But, it take about 15 minute for one epochs. As a result, we choose MobileNet. MobileNet is a type of convolutional neural network architecture that is specifically designed for mobile and embedded devices with limited computational resources. It achieves this by using depth-wise separable convolutions, which separates the standard convolution operation into two separate layers: a depth-wise convolution and a point-wise convolution.

The depth-wise convolution applies a single filter to each input channel individually, while the point-wise convolution applies a 1x1 convolution to combine the results of the depth-wise convolution. This results in a significant reduction in the number of parameters and computations required compared to traditional convolutional layers, making it faster and more memory-efficient. MobileNet is often used as a pre-trained model for transfer learning in computer vision tasks, such as image classification and object detection, because of its compact size and efficiency. It can also be easily adapted to different input sizes, making it suitable for a wide range of applications, including lip reading.

Existing datasets:

Databases for word, phrase, and sentence recognition

GRID:

The GRID dataset, released in 2006, is a widely used dataset in the field of lip reading research. It is a corpus of videos containing spoken sentences, recorded in a

studio setting, by 34 different individuals. The sentences have a fixed structure and belong to 51 classes of words,

such as colors, numbers, and letters. The GRID dataset is primarily designed for testing the accuracy of lip reading models on clear and simple speech. The videos in the dataset have a resolution of 360x288 pixels and a frame rate of 25 frames per second. The videos are recorded in black and white and the speaker's face is the only visible object in the video. The dataset is accompanied by a set of text transcriptions of the

ISSN: 2583-6129
www.isjem.com

spoken sentences, which can be used to evaluate the accuracy of lip reading models.

The GRID dataset provides a valuable resource for researchers who are interested in developing and testing such algorithms. By using the GRID dataset, researchers can test their models on a large, standardized dataset and compare their results with other researchers working in the same field.

Overall, the GRID dataset is an important resource for the development and evaluation of lip reading models, and it has been widely used in research to improve the accuracy of these models.

MIRACL-VC1:

MIRACL-VC1 is a valuable dataset for researchers working on lip reading and speech recognition models. It was created in 2014 and includes spoken words and phrases by 15 different individuals using a Kinect sensor. The dataset comprises both color and depth images of the speakers' mouths, providing more comprehensive information for lip reading models.The dataset is divided into ten classes, each with ten different phrases or words, such as greetings, commands, and questions. This categorization helps researchers to develop models that can handle different types of speech accurately. The images in the dataset have a resolution of 640x480 pixels and a frame rate of 15 frames per second, which is sufficient to capture the subtle movements of the lips and mouth during speech.MIRACL-VC1 is unique from other lip reading datasets due to its inclusion of depth information. This additional data provides a more detailed representation of the speaker's mouth movements and helps to improve the accuracy of lip reading models. The dataset has been widely used by researchers to develop and evaluate various lip reading and speech recognition models.

LRW: This is a dataset that was presented in 2016 for lip reading research. It has words that are spoken by over 1000 different people in natural videos from the web. The words belong to 500 classes and are aligned with subtitles, which can help with training lip reading models. The images have a resolution of 256x256 pixels and a frame rate of 25 frames per second. This dataset is challenging for lip reading models because it involves

diverse speakers, backgrounds, lighting conditions, and head poses34.

Lip reading is the process of recognizing speech from visual information only3. Lip reading can have many applications, such as helping people with hearing impairments, enhancing speech recognition in noisy environments, or enabling silent communication24. However, lip reading is also very difficult because of

many factors, such as mumbling, accents, foreign languages, jargon, and homophenes. Homophenes are words that look similar when lip read, but have

different sounds. For example, “pat” and “bat” are homophenes because they have the same mouth shape but different sounds.

Vocabulary word collection:

The dataset we have consisting of speech recordings from three individuals, with each person speaking 26 different words. These words are related to mobile operations, which suggests that the dataset may be intended for use in developing speech recognition or natural language processing (NLP) systems for mobile devices.

Speech Data Collection:

Setup for acquisitionthe recording was done in a well-lit environment with minimal background noise or visual distortion. In room temperature and normal humidity, the video was recorded at a ratio of 4:3 with a frame rate of 30 frames per second. The speaker was instructed to take a seat 12-15 ***centimeters*** in front of the camera**.**

We recorded using the Open Camera application. Open Camera interface offers users a number of recording options, such as video resolution, frame rate, and file formats. Depending on the intended usage, these options can be tweaked to optimize the quality and size of the output video file. The program also offers capabilities for manipulating video files, such as trimming or clipping parts, altering brightness, contrast or saturation levels, applying filters and exporting the results.

ISSN: 2583-6129
www.isjem.com

Sr.No.	Words	Sr.No.	Words	Sr.No.	Words
1	Back	11	Open Whatsapp	21	Delete
2	Home	12	Open Browser	22	Check
3	Screenshot	13	Open Camera	23	Answer
4	WiFi	14	Open Gpay	24	Hangup
5	Mute	15	Open Camera	25	Yes
6	Flashlight	16	Open Music	26	No
7	Notification	17	Open Message		
8	Recent Apps	18	Open Mail		
9	Bluetooth	19	Open Photos		
10	Lock	20	Open Alarm		

Table:1

The dataset developed contains 26 words from 3 distinct speakers. To ensure that consumers receive the most accurate data possible, each word must be spoken ten times. Thus, the complete dataset has 546 different video samples, and each sample is recorded in a period of 1-2 second.

The dataset can be used for research in speech recognition, natural language processing, and other related fields. It can also be used as a benchmark

dataset for evaluating the performance of different algorithms and models. The videos can be pre-processed to extract features such as facial expressions, lip movements, and speech patterns, which can be used as inputs to machine learning models for developing speech recognition or NLP systems.

Conclusion:

This paper presents an optimal design and development of a system for silent speech mobile interaction. The paper also introduces a methodology for creating a dynamic vocabulary dataset for mobile application that can serve as a model for future work. Moreover, the paper demonstrates how the system can convert silent videos into words that can benefit researchers in related fields.This paper proposes a new method for silent speech mobile interaction that uses deep learning and computer vision to convert silent videos into words. The paper also describes how the dataset of silent speech videos was created and collected from different speakers and environments. The paper evaluates the accuracy and robustness of the system and shows its potential for various use cases such as speech therapy, voice authentication, and silent communication.

Acknowledgement:

This study is being completed at JSPM's Rajarshi Shahu College of Engineering, Tathawade, under the supervision of the Department of Electronics and Telecommunication. The author would like to thank all of the speakers for their time spent gathering the speech samples.

ISSN: 2583-6129
www.isjem.com

References:

[1] Michael Wand, Jan Koutník, and Jürgen Schmidhuber, "Lipreading with long short-term memory" | Issue : Jan 2016, 1601.08188

[2] K. J. Ray Liu, "Speech and Signal Processing" (ICASSP),IEEE International Conference on. IEEE, 2016, 6115–6119.

[3] Joon Son Chung, "Deep Lip Reading: a comparison of models and an online application (arxiv.org)", Jun 2018,1806.06053

[4] Google. 2018. Google Assistant. (2018).https://assistant.google.com/

[5] Caroline Appert and Shumin Zhai. "Using Strokes As Command Shortcuts: Cognitive Benefits and Toolkit Support". In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09). ACM, New York, NY, USA, 2009, 2289–2298.

[6] Apple. 2018. iOS-Siri-Apple. (2018). https://www.apple.com/ios/siri/

[7] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas, "LipNet: end-to-end sentence-level lipreading", (2016),

[8] Patrick Baudisch and Gerry Chu. "Back-of-device Interaction Allows Creating Very Small Touch Devices". In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09). ACM, New York, NY, USA, 2009,1923–1932.

[9] Jonathan S Brumberg, Alfonso Nieto-Castanon, Philip R Kennedy, and Frank H Guenther. "Brain–computer interfaces for speech communication".Speech communication 52, 4 2010, 367–379.

[10] Alex Butler, Shahram Izadi, and Steve Hodges. 2008. SideSight: Multi-"Touch" Interaction Around Small Devices. In Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology (UIST '08). ACM, New York, NY, USA, 201–204.

[11] Xiang 'Anthony' Chen and Yang Li. "Bootstrapping User-Defined Body Tapping Recognition with Offline-Learned Probabilistic Representation" In Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16). ACM, New York, NY, USA, 2016, 359–364.

[12] Xiang 'Anthony' Chen, Julia Schwarz, Chris Harrison, Jennifer Mankoff, and Scott E. Hudson. "Air+Touch: Interweaving Touch & In-air Gestures". In Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14). ACM, New York, NY, USA, 2014,519–525.

[13] Cooke, J. Barker, S. Cunningham, and X. Shao, ''An audio-visual corpus for speech perception and automatic speech recognition,'' J. Acoust., Soc. Amer., vol. 120, no. 5, p. 2421, 2006.

[14]. Antonakos, A. Roussos, and S. Zafeiriou, ''A survey on mouth mod eling and analysis for sign language recognition,'' in Proc. 11th IEEEInt. Conf. Workshops Automat. Face Gesture Recognit. (FG), vol. 1,May 2015, pp. 1–7.

[15] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic,''End-to-End audiovisual speech recognition,'' in Proc. IEEE Int. Conf.Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 6548–6552.

[16] D. Kumar Margam, R. Aralikatti, T. Sharma, A. Thanda, P. A K, S. Roy,and S. M Venkatesan, ''LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models,'' 2019, arXiv:1906.12170. [Online]. Available:http://arxiv.org/abs/1906.12170

[17] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, ''CUAVE: A new audio-visual database for multimodal human-computer interface research,'' in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (CASSP), May 2002, p. II-2017.

[18] S. NadeemHashmi, H. Gupta, D. Mittal, K. Kumar, A. Nanda, and S. Gupta, ''A lip reading model using CNN with batch normaliza tion,'' in Proc. 11th Int. Conf. Contemp. Comput. (IC), Aug. 2018, pp. 1–6.

[19] K. Kumar, T. Chen, and R. M. Stern, ''Profile view lip reading,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), vol. 4, Apr. 2007, pp. IV-429–IV-432

[20] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, ''LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild,'' in Proc.14th IEEE Int.Conf. Autom. Face Gesture Recognit. (FG), May 2019, pp.18.

