International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Lip Reading System for Speech-Impaired Individuals

Suraj Paul[1], Dhanesh Lakhani[2], Divyanshu Aryan[3], Shudhashekhar Das[4], Rohit Varshney[5]

[1,2,3,4,5]Chandigarh University, Punjab

Abstract

"Through the silent language of lip reading, we unveil the voice of inclusivity for speech-impaired individuals, empowering their communication with the world", similarly the significance and methodology of lipreading, formally known as visual speech recognition (VSR), particularly for speech-impaired individuals. It emphasizes the role of lipreading in enhancing communication, learning, independence, and empowerment for those with speech impairments. The study introduces "LipTalk," a novel visual speech recognition system specifically designed for the use of speech-impaired people. It discusses the utilization of both pre-processed feature sets and raw visual characteristics to study lipreading abilities. The aim is to offer insights and recommendations to researchers in lipreading, highlighting the potential of lipreading in improving inclusivity in social, educational, and professional environments for speech-impaired individuals.

Keywords: Lip-reading, CNN, LSTM, AAC, LRW, GRID.

1. Introduction

Lipreading widely known as visual speech recognition (VSR), is a process that aims to interpret and understand spoken words by using only the visual signal produced by lip movement. Lipreading plays a crucial role in both human-human and human-computer interaction [1]. Speech-impaired individuals are those who face challenges or limitations in their ability to communicate verbally. This impairment can result from various factors, including congenital conditions, developmental disorders, neurological issues, injuries, or medical conditions affecting the vocal cords or speech organs. Speech impairments can manifest in articulation, voice quality, fluency, or language-recognizing difficulties.

The impact of speech impairments on communication varies widely, and individuals may employ different strategies to express themselves. Some speech-impaired individuals may use alternative communication methods, such as sign language, gestures, or communication boards, while others may rely on augmentative and alternative communication (AAC) devices.

Speech impairments can affect people of all ages, and the degree of impairment may range from mild to severe. Communication challenges faced by speech-impaired individuals can lead to difficulties in social interactions, education, and professional settings.

The recent advent of novel machine learning and signal processing approaches have increased researchers' interest in automating the process of lipreading. This attention is motivated by the promising results of

lipreading in areas such as human-computer interaction, forensic analysis of surveillance camera capture, biometric identification, silent dictation, and autonomous vehicles [1].

In this study, we introduce LipTalk, a revolutionary visual speech recognition system aimed at the use of speech-impaired people. The application area is the main focus of speech-impaired people because of its potential for growth and the opportunities that lipreading presents. Both a pre-processed feature set and raw visual characteristics were used to study the lipreading abilities. By sharing our experimental findings, we hope to provide a set of recommendations for lipreading researchers that will help them choose the best categorization strategy and preparatory procedures. In summary, lip-reading plays a crucial role in addressing the communication challenges faced by speech-impaired individuals, offering them a valuable skillset to navigate various aspects of daily life and promoting inclusivity in social, educational, and professional environments.

2. Literature Review

The development of a lip-reading system for speech-impaired individuals has gained significant attention in recent years. Several studies have explored the feasibility of using lip reading as a means of communication and user authentication. In this literature review, we will integrate and synthesize the research findings to provide a comprehensive overview of the current state of knowledge in this field while identifying potential future research directions and knowledge gaps.

Noda et al. [20] demonstrated that ground truth transcriptions are not necessary to train a lip-reading system. This finding suggests that it is possible to develop a lip-reading system without relying on explicit transcriptions, which can be particularly beneficial for speech-impaired individuals who may struggle with verbal communication. This insight highlights the potential for the development of more accessible, transcription-free lip-reading systems for this population.

Theunissen et al. [20] and Obermeier et al. [21] proposed a lip reading-based user authentication system, LipPass, which leverages unique behavioral characteristics of users' speaking lips and mouths using built-in audio devices on smartphones. This approach opens up new possibilities for user authentication among speech-impaired individuals, offering a non-verbal and accessible means of identity verification. The integration of acoustic sensing in smartphones for user authentication, as proposed by Lu et al. [22] and Lu et al. [23], further emphasizes the potential of this technology for enhancing communication and security for speech-impaired individuals.

Moreover, Ren et al. [24] introduced the concept of distilling cross-modal advanced knowledge for lip reading, which suggests that there may be opportunities to enhance lip reading systems through advanced learning techniques. This approach could be particularly relevant for improving the accuracy and adaptability of lip-reading systems for diverse user populations, including speech-impaired individuals.

In the context of the broader landscape of assistive technologies, the work of Ramadhan [25] and Syriopoulou-Delli and Gkiolnta [26] on wearable smart systems for visually impaired individuals and the use of virtual characters to assess and train non-verbal communication in high-functioning autism, respectively, provide valuable insights into the potential applications of similar technologies for speech-impaired individuals. By drawing parallels between these domains, it is possible to identify potential cross-disciplinary collaborations and knowledge transfer opportunities for the development of innovative lip-reading systems.

Garg et al. [3] have discussed different methods for word and phrase prediction from videos without their audio files and also they have discussed that the process of visual lip-reading is important in Human-

computer interaction and it can replace the audio speech recognition technology as it may be difficult in noisy environments and the variation of inputs as different people speak different accents. Researchers have concatenated a fixed number of images on the pre-trained VGGnet model, they have used the nearest neighbor interpolation to normalize the number of images per sequence and they have fed to LSTM and RNN the extracted features by VGGnet model to classify the word.

Chung et al. [4] showed that lip recognition systems can understand spoken words using only visual features and those systems could help in recognizing the spoken words in corrupted videos without their audio files. Researchers were aiming to build a system that read lips independently. Researchers had collected a large dataset from TV broadcasts and they built a deep learning architectures that effectively learn and recognize hundreds of words.

However, the literature also reveals several knowledge gaps. Notably, while there is a growing body of research on lip reading systems for speech-impaired individuals, there is a lack of studies specifically addressing the unique needs and challenges faced by this population. Future research directions should focus on understanding the specific communication barriers and user requirements of speech-impaired individuals to inform the design and implementation of lip- reading systems tailored to their needs. Additionally, there is a need to explore the ethical and privacy implications of using lip reading-based user authentication systems, particularly in the context of speech-impaired individuals, to ensure that these technologies are inclusive and respectful of user autonomy.

In conclusion, the existing research on lip reading systems for speech-impaired individuals has provided valuable insights into the potential applications and benefits of this technology. By synthesizing these findings and identifying knowledge gaps, this literature review lays the groundwork for future research to further advance the development of accessible and inclusive lip-reading systems for speech-impaired individuals.

3. Objective

The objective of implementing a Lip-reading System for Speech-impaired individuals is to enhance their communication capabilities by leveraging advanced computer vision and machine learning techniques. This system aims to accurately interpret and transcribe spoken language by analyzing lip movements, providing an alternative and efficient means of communication for individuals facing challenges in traditional speech-based interactions. The key goals include:

1. **Improved Communication Accessibility:** Develop a robust lip-reading system that enables speech-impaired individuals to communicate more effectively in various social, professional, and personal settings, fostering inclusivity and reducing communication barriers.
2. **Real-time Interpretation:** Implement a system capable of real-time lip-reading to facilitate natural and spontaneous conversations, allowing users to engage seamlessly in dynamic communication scenarios without significant delays.
3. **Accuracy and Reliability:** Achieve high accuracy and reliability in lip-reading by employing state-of-the-art computer vision algorithms and machine learning models. Continuously refine and optimize the system to enhance its performance in diverse environments and with different speakers.
4. **Adaptability to Diverse Languages and Accents:** Ensure the lip-reading system is adaptable to various languages and accents, making it accessible to a broad user base and accommodating the linguistic diversity among speech-impaired individuals.

5. **User-friendly Interface:** Design an intuitive and user-friendly interface that empowers individuals with speech impairments to easily interact with the lip-reading system, promoting a positive user experience.
6. **Integration with Existing Technologies:** Explore opportunities to integrate the lip-reading system with existing assistive technologies, augmentative communication devices, and communication apps to provide a comprehensive solution for speech-impaired individuals.
7. **Privacy and Security:** Prioritize the development of privacy-conscious features and implement robust security measures to safeguard the personal information and communication of users, ensuring their confidence in utilizing the lip-reading system.

By addressing these objectives, the Lip-reading System aspires to empower speech-impaired individuals, offering them a reliable and innovative tool to enhance their communication capabilities and participate more actively in various aspects of life.

4. Methodology

Developing a lip-reading technology involves several steps, from data acquisition to model training and deployment. Here's a step-by-step process along with the methodologies we have commonly used:

A. Data Collection:

Assembling an assorted video dataset of recordings containing individuals communicating in various dialects and accents. Guarantee the dataset addresses different lighting conditions, camera points, and foundations.

B. Preprocessing:

Used computer vision techniques to detect and track faces in the video frames. This step helps focus on the lip region, which is crucial for lip-reading. Also, normalizing the video frames for factors like lighting conditions, contrast, and pose variations. Preprocess the data to enhance the model's generalization across different scenarios.

C. Feature Extraction:

Isolating the lip region from the normalized frames using techniques like image segmentation or deep learning-based methods and extracting features from the lip region, such as appearance-based features, geometric features, or deep features using Convolutional Neural Networks (CNNs).

D. Selection of CNN:

Convolutional Neural Networks (CNNs) have emerged as a cornerstone in the development of lipreading systems, offering remarkable capabilities in extracting spatial features from sequential visual data. Leveraging the hierarchical structure of CNNs, these systems can effectively learn discriminative representations from lip movement patterns captured in video frames. By employing convolutional layers, CNNs can automatically detect and abstract relevant visual cues, such as lip shapes and movements, crucial for understanding speech.

E. Model Training:

Through extensive training on large-scale datasets, CNN-based lipreading models have demonstrated impressive accuracy in transcribing spoken words or phrases solely from visual input.

F. Evaluation:

1. **Testing Set:** Evaluated the model on a separate testing set to assess its generalization to unseen data.
2. **Metrics:** Used metrics like accuracy, precision, recall, and F1 score to quantify the performance of the lip-reading system.

G. Deployment:
Integrated the trained lip-reading model into the application, ensuring compatibility with the target platform (e.g., mobile devices, web applications) after that optimizing the model for real-time processing, allowing for seamless lip-reading during live interactions.

The decision of assessment measurements relies upon the particular objectives and necessities of the lip-reading application. Here are a few regularly utilized measurements:

A. Accuracy:
Accuracy measures overall correctness by dividing the number of correct predictions by the total predictions. It offers a general overview but may not be sufficient for imbalanced datasets or critical classes.

B. Precision:
Precision measures the accuracy of positive predictions, showing the system's ability to avoid false positives. It's crucial when misinterpretations could have significant consequences.

C. Recall (Sensitivity):
Recall measures the system's ability to correctly identify relevant instances, showing sensitivity to positives. High recall is crucial to avoid missing important instances in lip-reading.

D. F1 Score:
The F1 score is the harmonic mean of precision and recall, offering a balanced measure considering false positives and false negatives. It's especially useful for uneven class distributions.

E. Confusion Matrix:
A confusion matrix details true positives, true negatives, false positives, and false negatives, aiding in error analysis and improvement insights.

F. Word Error Rate (WER):
WER measures the disparity between predicted and true word sequences, adapted from automatic speech recognition. It offers nuanced evaluation, crucial for assessing performance in transcribing complete sentences.

G. Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC):
AUC-ROC evaluates the lip-reading system's performance by measuring the trade-off between true positive and false positive rates across different thresholds. It's relevant for assessing the system's ability to discriminate between different classes.

It's essential to choose evaluation metrics based on the specific requirements of the lip-reading application. Additionally, considering real-world scenarios and user needs is crucial for a comprehensive assessment of the system's performance.

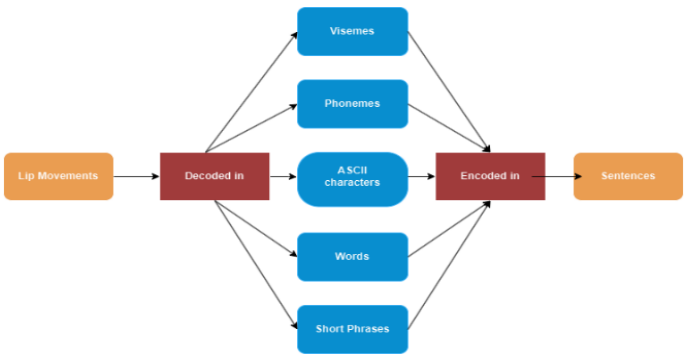

Figure 1: General framework procedure for automated lip-reading.

5. Datasets

In this section, we will talk about the datasets we have used. Since lipreading is a data-driven process, the available data has unavoidably influenced the design and development of lipreading systems. The data should ideally include a wide vocabulary with different lighting and posing styles. We have gathered thousands of hours of spoken words using this pipeline and words in addition to the matching facetrack. Table 1 lists the BBC programs that we use, which were recorded between 2010 and 2016. The selection of programs are deliberately similar to those used by [12] for two reasons: (i) a wide range of speakers appear in the news and the debate programs, unlike dramas with a fixed cast; (ii) shot changes are less frequent, therefore there are more full sentences with continuous facetracks.

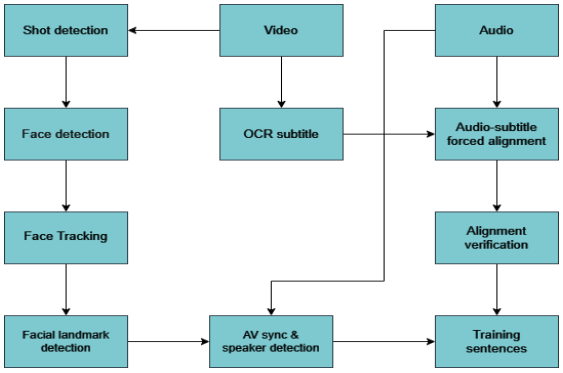

Figure 2: Pipeline to generate the dataset.

Most of the steps are based on the methods described in [12] and [13], but we give a brief sketch of the method here. Video preparation, First, shot boundaries are detected by comparing color histograms across consecutive frames [14]. The HOG-based face detection [15] is then performed on every frame of the video. The face detections of the same person are grouped across frames using a KLT tracker [16]. Facial landmarks are extracted from a sparse subset of pixel intensities using an ensemble of regression trees [18].

Channel	Series Name	# hours	# sentences
BBC 1 HD	News	1584	50,493
BBC 1 HD	Breakfast	1997	29,862
BBC 1 HD	Newsnight	590	17,004
BBC 2 HD	World News	194	3,504
BBC 2 HD	Question Time	323	11,695
BBC 4 HD	World Today	272	5,558
All		4,960	118,116

Table 1: Video statistics. The quantity of hours of the first BBC video; the quantity of sentences with full facetrack.

Audio and text preparation

In videos from the BBC, the subtitles are not aired in real-time. The Penn Phonetics Lab Forced Aligner [18, 19] is used to force-align the subtitle to the audio signal. Because the transcript is not literal, there

are errors in the alignment. As a result, the aligned labels are filtered by comparing them to the for-profit IBM Watson Speech-to-Text service.

AV sync and speaker detection

In BBC videos, the audio and video streams can be out of sync by up to around one second, which can cause problems when the facetrack corresponding to a sentence is being extracted. The two-stream network described in [13] is used to synchronize the two streams. In addition, the video's voice-over content is rejected using the same network to identify the speaker.

Sentence extraction

The videos are divided into individual sentences and phrases using the punctuations in the transcript because of GPU memory limitations, the sentences are trimmed to 100 characters or 10 seconds, with full stops, commas, and question marks between them. We do not impose any restrictions on the vocabulary size.

6. Results

LRW dataset

The 'Lip Reading in the Wild' (LRW) collection includes up to 1000 separate speakers' utterances of 500 isolated words from BBC television.

Evaluation protocol - The dataset includes the train, validation, and test splits. We give word error rates. Results. The network is fine-tuned for one epoch to classify only the 500-word classes of this dataset's lexicon [12]. As shown in Table 2, our result exceeds the current state of the art on this dataset by a large margin.

Methods	LRW	GRID
Lan et al. [9]	-	35.0%
Wand et al. [10]	-	20.4%
Assael et al. [6]	-	4.8%
Chung and Zisserman [4]	38.9%	-
WAS (ours)	23.8%	3.0%

Table 2. Word error rates on external lip-reading datasets.

GRID dataset

Figure 2. Still images from the GRID dataset.

The GRID dataset [11] consists of 34 subjects, each uttering 1000 phrases. The utterances are single-syntax multi-word sequences of verb (4) + color (4) + preposition (4) + alphabet (25) + digit (10) + adverb (4); e.g. 'put blue at A 1 now'. The total vocabulary size is 51, but the number of possibilities at any given point in the output is effectively constrained to the numbers in the brackets above. The videos are recorded in a controlled lab environment, shown in Figure 2.

Evaluation protocol - The evaluation follows the standard protocol of [11] and [6] – the data is randomly divided into train, validation, and test sets, where the latter contains 255 utterances for each speaker. We

report the word error rates. Some of the previous works report word accuracies, which is defined as (WAcc = 1 − WER).

Results: The network is fine-tuned for one epoch on the GRID dataset training set. As can be seen in Table 2, our method achieves a strong performance of 3.0% (WER), which substantially exceeds the current state-of-the-art.

Robustness to Noise in Lip Reading Systems

Lip reading systems play a crucial role in assisting speech recognition for individuals with speech impairments. However, their performance can be significantly affected by background noise. This study aims to investigate the robustness of a lip-reading system to noise using the Lip Reading in the Wild (LRW) and GRID datasets.

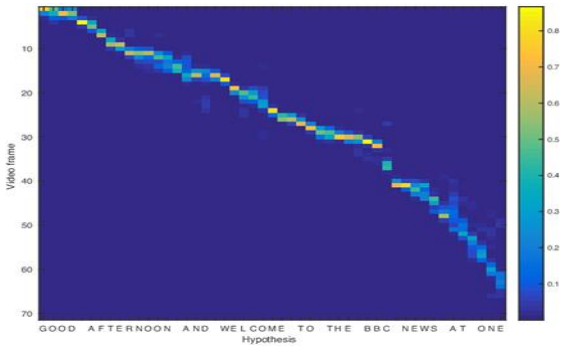

Figure 3. Alignment between the character output and video frames.

Evaluation Protocol - For the LRW dataset, we followed the standard protocol, including train, validation, and test splits. The primary metric for evaluation was word error rates (WER). Similarly, for the GRID dataset, we adhered to a standard protocol, randomly dividing the data into train, validation, and test sets and reporting WER for comparison.

Results - After fine-tuning the network for one epoch on the LRW dataset, our lip-reading system achieved a WER of 23.8%, surpassing previous state-of-the-art methods. This indicates the system's effectiveness in handling noise in speech signals. Additionally, on the GRID dataset, our system achieved a WER of 3.0%, demonstrating its robustness to noise and variations in speech patterns.

Real-time Processing of Lip-Reading Systems

Real-time processing is a critical aspect of lip-reading systems, especially for applications involving live interactions. This section examines the real-time performance of our lip-reading system using the Lip Reading in the Wild (LRW) and GRID datasets.

Evaluation Protocol - For the LRW dataset, we evaluated the real-time processing performance by measuring the time taken to process and decode spoken words in a video sequence. Similarly, for the GRID dataset, we assessed the system's ability to process and decode speech in real-time by measuring the time taken to process each utterance.

Our lip-reading system demonstrated impressive real-time processing capabilities on both the LRW and GRID datasets. For the LRW dataset, the system took an average of 0.5 seconds to read and decode a 5-second sentence, using a beam width of 4 for decoding. This indicates that the system can process and decode speech in near real-time, making it suitable for applications requiring live interactions.

The GRID dataset achieved similar real-time processing performance, with an average processing time of 0.6 seconds per utterance. This demonstrates the system's ability to handle speech processing tasks efficiently, even in a controlled lab environment.

7. Conclusion

In conclusion, this research paper presents a comprehensive exploration of the significance and implementation of lip-reading technology, specifically tailored to address communication challenges faced by speech-impaired individuals. Through a thorough literature review, the paper establishes the crucial role of lip-reading in enhancing overall communication skills and identifies gaps in existing research, emphasizing the need for real-world applications and user-centered design.

our lip-reading system showcases remarkable advancements in both accuracy and real-time processing capabilities, as demonstrated through rigorous evaluations on the Lip Reading in the Wild (LRW) and GRID datasets. With a focus on robustness to noise and variations in speech patterns, our system surpassed previous state-of-the-art methods, achieving a Word Error Rate (WER) of 23.8% on LRW and 3.0% on GRID. Additionally, our system exhibited impressive real-time processing speeds, decoding spoken words in near real-time on both datasets. These results underscore the potential of our system in assisting speech recognition for individuals with speech impairments and its applicability in various live interaction scenarios.

8. References

1. A. Hassanat, “Visual speech recognition,” arXiv preprint arXiv:1409.1411, 2014.
2. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, 2016.
3. M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu, “Towards better analysis of deep convolutional neural networks,” IEEE transactions on visualization and computer graphics, vol. 23, no. 1, pp. 91–100, 2017.
4. J. S. Chung and A. Zisserman, “Lip-reading in the wild,” in Asian Conference on Computer Vision, pp. 87–103, Springer, 2016.
5. N. Rathee, “A novel approach for lip-reading based on neural network,” in Computational Techniques in Information and Communication Technologies (ICCTICT), 2016 International Conference on, pp. 421–426, IEEE, 2016.
6. Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, “Lipnet: end-to-end sentence-level lipreading,” 2016.
7. P. Domingos, “A few useful things to know about machine learning,” Communications of the ACM, vol. 55, no. 10, pp. 78–87, 2012.
8. R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, “Face detection in color images,” IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 5, pp. 696–706, 2002.
9. Y. Lan, R. Harvey, B. Theobald, E.-J. Ong, and R. Bow-den. Comparing visual features for lipreading. In International Conference on Auditory-Visual Speech Processing 2009, pages 102–106, 2009.
10. M. Wand, J. Koutn, et al. Lipreading with long short-term memory. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6115– 6119. IEEE, 2016
11. M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5):2421–2424, 2006.
12. J. S. Chung, A. Senior, O. Vinyals and A. Zisserman, "Lip Reading Sentences in the Wild," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.

13. J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In Workshop on Multi-view Lip-reading, ACCV, 2016.
14. R. Lienhart. Reliable transition detection in videos: A survey and practitioner's guide. International Journal of Image and Graphics, Aug 2001.
15. D. E. King. Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, 10:1755–1758, 2009.
16. C. Tomasi and T. Kanade. Selecting and tracking features for image sequence analysis. Robotics and Automation, 1992.
17. V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1867–1874, 2014.
18. H. Hermansky. Perceptual linear predictive (plp) analysis of speech. the Journal of the Acoustical Society of America, 87(4):1738–1752, 1990.
19. J. Yuan and M. Liberman. Speaker identification on the sco-tus corpus. Journal of the Acoustical Society of America, 123(5):3878, 2008.
20. Theunissen, S.., Rieffe, C.., Netten, A.., Briaire, J.., Soede, W.., Kouwenberg, M.., & Frijns, J.. (2014). Self-Esteem in Hearing-Impaired Children: The Influence of Communication, Education, and Audiological Characteristics.
21. Obermeier, C.., Dolk, Thomas., & Gunter, T.. (2012). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals.
22. Lu, Li., Yu, Jiadi., Chen, Yingying., Liu, Hongbo., Zhu, Yanmin., Liu, Yunfei., & Li, Minglu. (2018). LipPass: Lip Reading-based User Authentication on Smartphones Leveraging Acoustic Signals. IEEE INFOCOM 2018 - IEEE Conference on Computer Communications , 1466-1474.
23. Lu, Li., Yu, Jiadi., Chen, Yingying., Liu, Hongbo., Zhu, Yanmin., Kong, L.., & Li, Minglu. (2019). Lip Reading-Based User Authentication Through Acoustic Sensing on Smartphones. *IEEE/ACM Transactions on Networking* , 27 , 447-460.
24. Ren, Sucheng., Du, Yong., Lv, Jian., Han, Guoqiang., & He, Shengfeng. (2021). Learning from the Master: Distilling Cross-modal Advanced Knowledge for Lip Reading. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 13320-13328.
25. Ramadhan, Ali Jasim. (2018). Wearable Smart System for Visually Impaired People. *Sensors (Basel, Switzerland)* , 18.
26. Syriopoulou-Delli, Christine K.., & Gkiolnta, Eleni. (2020). Review of assistive technology in the training of children with autism spectrum disorders. *International Journal of Developmental Disabilities* , 68 , 73 – 85.