**Indonesian Journal of Electrical Engineering and Informatics (IJEEI)**
Vol. 10, No. 2, June 2022, pp. 385~397
ISSN: 2089-3272, DOI: 10.52549/ijeei.v10i2.3683


# Multimodal Based Audio-Visual Speech Recognition for Hard-of-Hearing: State of the Art Techniques and Challenges

**Shabina BHASKAR[1], Thasleema T M[2]**
[1,2]*Department of Computer Science, Central University of Kerala, India, 671320*

**Article Info**

***Article history:***

Received Jan 25, 2022
Revised Apr 4, 2022
Accepted May 11, 2022

***Keyword:***

Automatic speech recognition,
Sign language,
Audio-visual systems,
Speech recognition,
Neural networks.

**ABSTRACT**

Multimodal Integration (MI) is the study of merging the knowledge acquired by the nervous system using sensory modalities such as speech, vision, touch, and gesture. The applications of MI expand over the areas of Audio-Visual Speech Recognition (AVSR), Sign Language Recognition (SLR), Emotion Recognition (ER), Bio Metrics Applications (BMA), Affect Recognition (AR), Multimedia Retrieval (MR), etc. The fusion of modalities such as hand gestures- facial, lip-hand position, etc., are mainly used as sensory modalities to develop hard-of-hearing multimodal systems. This paper encapsulates an overview of multimodal systems available within the literature for hard-of-hearing studies. This paper also discusses some of the studies related to hard-of-hearing acoustic analysis. It is observed that very few algorithms have been developed for hard-of-hearing AVSR than normal hearing. Thus, the study of audio-visual-based speech recognition systems for hard-of-hearing is highly demanded by people trying to communicate with natively speaking languages. This paper also highlights the state-of-the-art techniques in AVSR and the challenges faced by the researchers in the development of AVSR systems.



***Corresponding Author:***

Shabina Bhaskar,
Department of Computer Science,
Central University of Kerala,
Kasaragod, India, 671320.
Email: shabinabhaskar@cukerala.ac.in

## 1. INTRODUCTION

Multimodal Integration (MI) offers more natural, accessible, and transparent ways of combining information from different sources. The MI system emerged approximately 30 years before. The MI experiments were initiated in Audio-Visual Speech Recognition (AVSR) to improve the robustness of speech recognition in a noisy acoustic environment. The AVSR has been an active research area for many years, and its transition from machine learning to deep learning improved the recognition rate of AVSR systems. Even though many applications have been developed for people with hearing loss to recognize the speech of normal hearing, only very few works have been reported on analyzing speech uttered by the hard-of-hearing person [1]. The development of language vocabulary in a person depends on the ability to hear. For a normal-hearing person, language vocabulary develops at their infant stage. The situation changes when a person experience hearing loss and this causes a delay in developing language and communication skills. This may cause a negative impact on language and vocabulary development for such person. According to the Hearing Loss Association of America, the degree of hearing loss can be classified as

- Slight hearing loss (16-25 dB)
- Mild hearing loss (26-40 dB)
- Moderate hearing loss (41-55 dB)
- Moderately severe hearing loss (56-70 dB)
- Severe hearing loss (71-90 dB)
- Profound hearing loss (90 above).

A person with hearing loss in the range of mild to severe (hard of hearing) can partially understand the speech, but the person with profound hearing loss cannot hear anything and is entirely deaf [2]. For the Deaf, communication is through sign language, but for those who are not trained in sign language, the semantical interpretation is difficult to attain. It makes a communication gap between normal hearing and Deaf people. So, there is an enormous scope for research in assistive system development, which helps to communicate with such people. Most of the studies associated with the Deaf are in the area of Sign Language Recognition (SLR) [3]. Hard-of-hearing people may not use sign language for communication, and they have speech impairments due to their inability to hear. With the help of assistive listening and learning devices, we can improve their communication skills. In such cases, speech recognition systems are crucial because they can be used to recognize the speech of both normal and hard-of-hearing individuals. But, hard-of-hearing people's speech recognition is challenging because their speech and speaking styles are entirely different from normal hearing people.

Many studies have been conducted on hard of hearing to test the literacy skill in children, check their speech perception level, improve the speech enhancement techniques for hearing aids, and test their hearing sensitivity [4-5-6-7]. Researchers have already conducted studies related to speech recognition using acoustic data for them [1-37-38-50-51]. Less attention is given to the development of methods for visual speech recognition. In the case of noise or missing data either in the audio or visual modality, the AVSR techniques can provide a reliable enhancement in the recognition rate of the system. Even though the AVSR study for normal hearing started three decades ago, the researchers were less focused on the area of hard-of-hearing AVSR systems. The recently published papers on AVSR review also look at the normal hearing speech recognition [8-9-74]. Our aim is to improve the existing communication systems for hard-of-hearing people by utilizing the techniques of AVSR to create systems that are accessible to them. Therefore, it is necessary to review available systems for the hard-of-hearing as well as AVSR systems. Our paper discusses various audio, visual, and multimodal systems available to hard-of-hearing people and all these systems aim to enhance or assist hard-of-hearing people in communicating.

In the multimodal studies, a great deal of attention is devoted to automatic sign language recognition (ASLR), cued speech recognition (CSR), etc. This paper provides an overview of MI systems developed for people with hearing difficulty. Also, it will explain the works related to acoustic-based speech recognition systems designed for hard-of-hearing. In state-of-the-art techniques, most AVSR methods are used for normal-hearing speech recognition, and only very few works are reported so far for hard-of-hearing. We, therefore, conducted a preliminary study on AVSR systems for normal-hearing people to identify the latest advancements in this field. The rest of this paper is organized as follows.

Section 2 discusses multimodal systems, and section 3 reviews different types of available multimodal systems for hard-of-hearing. Section 4 outlines the tools and techniques related to acoustic data analysis for hard-of-hearing. Section 5 elucidates audio-visual speech recognition systems and their integration techniques. Section 6 explains the challenges in audio-visual speech recognition systems, and finally, in section 7, the paper is concluded with an emphasis on future directions.

## 2. MULTIMODAL SYSTEMS

Modality refers to the manner or way of representing information through some medium. For example, visual modality can be represented using mediums such as images or videos. Multimodality means processing and combining data from at least two different modalities [10]. Single modality such as acoustic-based speech recognition is complicated for hard-of-hearing because of the enormous variation in their speech as compared with normal hearing. In this regard, MI techniques are crucial for the recognition of hard-of-hearing speech. In a multimodal, one modality can provide supplementary information rather than complementary information to the other modality. MI systems can overcome the limitations of single modality systems. MI began with Audio-Visual Speech Recognition (AVSR), and it is now extending to a variety of other fields such as Emotion Recognition (ER), Sign Language Recognition (SLR), Affect Recognition (AR), and so on. The recent studies in multimodal machine learning and the challenges in MI are discussed by Baltrusaitis et al. [10]. The different types of multimodal interactions, challenges, and issues in multimodal integration are briefly reviewed by the author Turk [11]. Table.1 tabulates the details of these challenges and their various associated applications.

Affect Recognition (AR) is a significant research area in multimodal signal processing. Sentiment Analysis (SA), ER, and Opinion Modeling (OM) are various parts of AR. Previously, sentiment extraction was based on text data in AR research, and ER was based on audio or visual data. After introducing MI, the combination of modalities such as text, audio, visual and physiological signals (EEG, ECG, etc.) brought a good result in Affect Recognition. A recent survey paper has reviewed different types of frameworks that are used in multimodal big data affective analytics [12]. The article discusses various big data models such as HACE, IBM's 4V model, and big data analytics frameworks such as Hadoop, Storm, and Spark. The paper contains a detailed survey combining different modalities such as text, audio, video, and physiological signals.

The authors have also thoroughly reviewed the databases, fusion methods, and techniques related to the current research areas such as automated depression diagnosis, stress detection, lie detection, etc.

Table 1. Challenges associated with different Multimodal integration applications

| Challenges | | Algorithms | Modality | Applications |
|---|---|---|---|---|
| Representation: Joint | | Autoencoders LSTM, RNN, Multimodal DBN | Audio+ Visual | Object and scene recognition, pose estimation, AVSR, Affect recognition |
| Coordinated | | DNN, LSTM | Video+ Text | Visual object identification, Video description |
| Translation: | Example based | AAM+ Animation tool Integer linear programming | Text to visual, Visual to text | Visual speech synthesis Image captioning |
| | Generative | HMM+ Animation tool | Audio to video | Talking face generation |
| Alignment: | Explicit | Dynamic Time Wrapping | Video to text | Video retrieval |
| | Implicit | Markov Random Field | Visual to text | Natural science description |
| Fusion: | Model agnostic | SVM, Kernel Regression Decision Trees | Audio+ Visual+ Text | Multimedia event detection, Emotion recognition |
| | Model-based | Coupled HMM, LSTM, CNN | Audio+ Visual | Emotion Recognition, Affect Recognition, AVSR |
| Co-learning: | Parallel | Deep learning | Audio+ Visual | AVSR, Lip reading |
| | Non-Parallel | LSTM, RNN | Audio+ Visual | Action recognition |

## 3. MULTIMODAL SYSTEMS FOR HARD-OF-HEARING

Those with hearing loss communicate primarily through Sign Language (SL) or Cued Speech (CS). Moreover, the lip-reading capability helps them to understand normal-hearing persons' speech. Already some works have been introduced in MI to implement communication systems like Automatic Sign Language Recognition (ASLR), Cued Speech Recognition (CSR) system, etc. Multimodal interfaces are also being developed to improve their communication. In this section, we investigate multimodal interfaces and MI systems for helping the hard-of-hearing to communicate more effectively.

### 3.1. Multimodal Interfaces for Hard-of-Hearing

Different types of multimodal interfaces are available for the hard of hearing that help in either communication or interaction. The paper [13] presented a multimodal user interface that provides communication between disabled users and their interaction with the computer. It integrates both a sign language recognition-synthesis module for hearing impaired users and an AVSR-speech synthesis module for blind users. As an application scenario, the aforementioned technologies are integrated into a collaborative treasure hunting game which requires the interaction of the users at each level. The multimodal interfaces study for the blind and deaf person was conducted in [14]. The problems and issues associated with their communication and interaction are analyzed in their research. Also, the proposed prototype Tyflos Koufos is a combination of several technologies such as image understanding, natural language understanding, speech recognition, synthesis, etc. The authors Lee et.al, [15] introduced a lip-shape refinement technique for English vowels and which is implemented using a Support Vector Machine (SVM). The study was conducted on three different speakers; English native speakers, Korean normal hearing, and Korean hearing-impaired speakers. This paper proposed a new visual teaching method of English vowels for the hearing impaired. An assistive robotic system that makes use of a combination of modalities like audio, touch, and gesture for the effective communication of hearing-impaired persons was introduced by Edward et.al. The important feature of this system is it can interact with any type of hearing-impaired person without considering the degree of disability of users [16]. Different interfaces were developed by the researchers using the electroencephalography (EEG) signals for the effective communication reviewed by Bhuvaneshwari et.al. In which the communication systems developed by using machine learning and deep learning technologies are discussed very briefly [17]. A multimodal human-computer interaction system based on robot/virtual avatar tutors and a simple drumming-based interactive music tutoring game has been designed as a part of education and therapy. This training and feedback mechanism employed in the interactive games helped participants to enhance their performance and motivation [18].

### 3.2. Multimodal Integration for Hard of Hearing

This section discusses the multimodal systems by using audio-visual, hand-facial, and lip shape-hand position combinations. Communicating by manipulating lip shape and hand position is known as Cued Speech (CS). CS is an important communication mode for a visual-auditory communication system used by people with hearing loss to understand spoken language better. CS uses hand cues and natural mouth movements to specify phonemes [19]. The importance of CS in literacy and language development is intensely studied for children of age groups 6 to 14 years [20]. There are only a few resources available for Cued Speech Recognition (CSR) in any cued language. A remarkable work in French CSR was done using HMM-based vowel recognition [21], which gives an accuracy of 68%. HMM integration at the feature level provides a recognition rate of 85.1%, while Multistream HMM (MSHMM) based decision fusion provides a higher recognition rate (87.6%). They have extended their work to consonant, phoneme, and isolated word recognition tasks in addition to the vowel recognition system [22]. They have introduced three viseme classes for vowels. A total of eight handshapes are used in French CS along with five different hand positions. They used eight lip geometric features and four-hand position information for vowel recognition. The database consists of 262 sentences. During database collection, the participant wore a helmet to keep his head in the proper position. The lips were painted blue so that the shape of the lips could be tracked accurately. The eight geometric lip features and their first, second-order derivatives are used as lip features; thus, the dimension of the feature vector is 24. 12-dimensional feature vector produced for hand position by combining four basic parameters and their first, second-order derivatives. The combination of hand position and lip shape features is used as a concatenated feature vector for vowel classification. The HMM-based classification with the rest of the cases of consonant classification gives an accuracy of 78.9% for 32 Gaussian per state. The HMM-based phoneme recognition gives 74.4% accuracy. In the case of isolated word recognition, the result analysis is made for both normal hearing and hearing loss above 90 dB (deaf) people. The accuracy was 94.9% for normal hearing and 89% for hearing loss above 90 dB. Thomas Burger proposed an efficient method of CS recognition from video using a restricted number of keyframes to reduce computation time and cost [23]. It uses a 3D graphic system, including a text to animation conversion system. Another project is a system for interactive CS learning and practice for deaf children. Based on the text or speech inputted by the user, the system can synthesize face and hand animation for CS [24]. Resources are lacking in CS, which has limited the availability of research in this field. In 2017, a new method called Constrained Local Neural Field (CLNF) was proposed for CSR. Traditionally, lip coloring is used to extract correct lip features, but CLNF offers a novel feature extraction method for inner and outer lip features. The technique is the best suited for lipreading applications and attained more than 80% in CS recognition [25]. Deep learning was also utilized to analyze hand shape and lip features using CNN models for CS recognition from videos [26]. A Malay CSR system was introduced to help the deaf learn and practice the basics of cued speech consonants and vowels using hand gestures [27]. This paper is the most recent study in this field. It reviews the available systems for deaf people, including CS and SLR systems.

The most commonly used communication method among the deaf community is sign language. Most of the works in ASLR are related to manual sign language recognition, which uses hand and finger movement for sign representation. Non-manual sign language includes hand gestures and facial features like eyes, eyebrows, and mouth information. With 3D depth sensors, gloves sensors, and RGB camera techniques, it was possible to track correct hand and finger movements in ASLR. To improve the performance of hand segmentation in ASLR, the methods such as Gabor features, dynamic angular features, etc., are used. Apart from that, deep learning models such as Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) improved the dynamic hand gesture recognition performance. Also, MI techniques improved the performance of automatic SLR systems. Different modality combinations have already been tested in multimodal SLR, such as hand gestures and lips, hand gestures and body gestures, and hand gestures and facial expressions. A multimodal framework designed by [28] for SLR captured the features of finger and palm positions. This method reported overall accuracies of 97.85% for HMM and 94.55% for Bidirectional Long Short-Term Memory (BLSTM). For ASLR, two sensors such as Leap Motion and Intel RealSense were used to track the movements of hands. It is evaluated in the American sign language fingerspelling dataset, and the Intel RealSense showed better results and accuracy [29]. The work of Kulkarni et al. [30] aims to recognize hand gestures using the classical video analysis technique. This technique identifies the movement models based on skin and background pixels. Hand motion analysis with Gaussian Mixture Model (GMM) obtained a result of 93%. The key contributions of this paper are the automatic motion curve approximation, the automatic motion signature extraction (identification), and the automatic sign language space segmentation [31].

In some cases, hand gestures are not enough to understand sign language. For example, the hand gesture for "who" and "what" looks similar. In such a case, facial expression gives some additional information about the word. This multimodal combination leads to a more accurate system than a single modality. In [32], the authors introduced a decision fusion method for multimodal integration. Each modality is classified using separate classifiers, and the recognition rate is improved by applying the Independent Bayesian Classification

Combination (IBCC) algorithm. Here they used a Kinet camera to capture facial data and leap motion sensors to capture hand gestures. They have developed a database that contains 51 different dynamic signs corresponding to Indian Sign Language (ISL). Thirty-one signs represent double hand gestures, and the remaining 20 signs represent single hand gestures. The HMM-based feature fusion method in the case of single-hand gestures gives 95.79% accuracy, whereas IBCC-based decision fusion gives 96.05% accuracy. In the case of the double-hand experiment, the result was 92.71% for the feature fusion method and 94.27% for the decision fusion method. In multimodal studies, hand position and facial features are the most commonly used modalities in SLR, and for CSR, both lip and facial features are employed. Based on image and video processing, Caplier et al. presented a summary of available SLR and CSR systems for the communication of hard-of-hearing [33]. This paper also outlined the sign language and cued speech video synthesis methods for communication from normal hearing people to hard-of-hearing people. Face and hand features fusion using HMM and SVM implemented for sign language recognition and the average accuracy was above 50% for all experiments. Several features have been evaluated for classification in this work, and the overall best performing set has been determined using the sequential feature selection technique [34]. The work on Arabic sign language recognition translates isolated Arabic word signs into text, and it is evaluated based on the Euclidian distance. The system attains a recognition rate of 97% in signer independent mode [35]. A four-stream CNN model proposed by Ravi et al. for multimodal ASLR achieved the best result in RGB video data, and the results compared state-of-arts models for comparison [36].

Nowadays computer-based speech therapy systems or Virtual Speech Therapists (VST) getting popular because of their acceptability, effectiveness, and portability. VST can be developed for people with speech disorders due to hearing impairment, dysarthria, etc. But it is very difficult to develop such systems for people with speech impairment due to autism, down syndrome, etc. Even though we cannot replace speech-language pathologists (SLP), the VST can assist such people in speech training to some extent. The work introduced by Jiang et al. includes a lipread training system for hearing-impaired to correct their pronunciation. The multimodal diviseme instance selection algorithm is used to find the optimal representative diviseme instance for the viseme pairs in the input speech [37]. A three-dimensional talking head articulation training system developed for hearing impaired children proved to be very effective, and it has been tested among Mandarin children [38]. As a method of delivering speech therapy, virtual taking heads are the most popular technology invented. This animated guide describes how to use articulators and how to correct pronunciations. Work in this area is very limited and a detailed review of the VST systems like speech training systems, talking heads, etc. is conducted for speech impairment people by Chen et al. [39]. They suggested the methods like Word Recognition Rate (WRR), Phoneme Recognition Rate (PRR), and Word Error Rate (WER) to measure the improvements in speech recognition. Also, the problem related to speech production can be measured using sound pressure level, Percentage of Consonants Correct (PCC), Correctness of Vowel Production, Phonological Assessment Battery (PhAB), etc.

Mainly three types of multimodal studies such as CSR, ASLR, and VST are discussed in this section. Table.2 gives the details of multimodal studies for the hearing-impaired. The MI studies related to the hearing-impaired are more focused on ASLR rather than speech recognition. Even though some Automatic Lip Reading (ALR) systems are reported by [42], the studies for hearing-impaired AVSR systems are rarely available. But, the acoustic analysis of the hearing impaired is also carried out by many researchers so in the next section we will be discussing the techniques related to that.

Table 2. Multimodal Integration for Hard-of-Hearing

| Authors | Features | Modalities | Integration |
|---|---|---|---|
| [21,22,23,24] | Lip and Hand features | Visual | HMM-based early integration |
| [28,29,30,31] | Hand motion features | Visual | HMM-based feature fusion |
| [32,33,34] | Hand and facial features | Visual | HMM, IBCC decision fusion |
| [35,36] | Hand, lip, and facial features | visual | HMM, Deep learning model |
| [37,38,39] | Audio and facial feature | Audio+ visual | HMM, Deep learning model |
| [40] | Audio and hand features | Audio+ visual | Deep learning integration |
| [41,42] | Hand features | Visual | Deep learning integration |

## 4. ACOUSTIC ANALYSIS FOR HEAD-OF-HEARING SPEECH RECOGNITION

### 4.1. Speech Training Systems for Head-of-Hearing

In the previous section, we have discussed audio-visual-based multimodal systems for hearing-impaired people. Other than that, some experiments in speech have been carried out for hard-of-hearing persons. Among these, the language training systems helped the hearing-impaired to increase their learning performance. Computer Integrated Speech Training Aid (CISTA) is a commercially available system in Japan that provides speech training to deaf people. In the early days of developing speech training for deaf people,

this was one of the effective systems [43]. The Radar graphic Display System developed for speech learning is best suited for hearing-impaired primary school students to learn Chinese pronunciation [44]. The proposed system was based on speech parameters such as LPC, MFCC, etc., and Back Propagation Neural Network (BPNN). Also, the Self Organizing Map was utilized for the creation of a radar map. The system included both a microphone and a computer graphic display screen. Similar experiments are carried out in other languages, such as VOIS -a CNN based speech therapy app available in the Myanmar language [45].

A visual teaching method for the pronunciation of English vowels was introduced by Han et al. [46]. In their work, first, phonetic features were extracted and using the SVM technique, the sound is recognized. After that, the lip refinement method is introduced, which demonstrates the lip movement corresponds to each vowel. An interactive training system developed for hearing-impaired to improve their language skill in the Mandarin language provides a virtual reality environment for learning. Also, the pronunciation accuracy is evaluated by the system, and it also gives feedback to the speakers. By using the technique of Automatic Speech Recognition (ASR), the automatic articulation error detection system was proposed in the Punjabi language [47]. MFCC features and SVM classifier are used to recognize the word spoken by the hearing-impaired. The overall accuracy of predicting the word was 92.67%. By using visual features from the face and applying HMM model, a speech recognition system was developed in the Chinese language for hearing-impaired [48]. In this work, six Chinese vowels were used as experimental data, and the system gained an accuracy of 91.47%.

### 4.2. Speech Recognition system for Hard-of-Hearing

The acoustic signal is the main source of information to recognize the speech of a person. The speech recognition systems identify the uttered words or phrases by analyzing the acoustic signal produced during speech. Although the speech of a hard-of-hearing person does not contain sufficient audio information, the acoustic analysis plays a pivotal role in hard-of-hearing speech recognition. Many speech recognition systems are available within the literature for them to understand or recognize the speech produced by normal hearing people. But the study of the techniques or applications to recognize the speech of hard-of-hearing is still far behind. The main reason behind the lagging of study in this area is the large variation in energy, duration, pronunciation, and spectral components of hearing-impaired speech as compared to normal hearing people. Also, the unavailability of hard-of-hearing speech database made this study more complicated. In this section, we will discuss the acoustic analysis techniques that are applied in the case of hard-of-hearing speech recognition.

The study of acoustic characteristics analysis of hearing-impaired started a few years after the introduction of normal hearing speech recognition systems. The methods introduced for normal hearing speech recognition are also experimented with for hearing-impaired speech recognition. In the method proposed by Gudi et.al, with the aid of adaptive signal processing, the speech of disabled children is tuned into curve-fitted normal speech through a feedback mechanism [49]. Acoustical characteristics of speeches of deaf people are analyzed [50] to increase the speech recognition rate. Among Indian languages, Tamil language [1-49-50] contributed groundbreaking work for acoustic analysis of hard-of-hearing. In one of the main works, the speech samples were collected from ten deaf children of the age group 10-14 years, and the language material includes ten isolated digits, ten connected words, and 45 continuous speech sentences in Tamil [51]. The data was collected by a tactile method which is used by the speech therapist to train the hearing-impaired. The system was evaluated for normal hearing and hearing-impaired speakers. The articulation errors of hearing-impaired while uttering a sentence were analyzed using the tool SPHINX. Two methods were used for implementation, one is the most widely used technique Mel Frequency Cepstral Coefficient (MFCC) with Hidden Markov Model (HMM) classifier, and the other is Mel Frequency Perceptual Linear Predictive Cepstrum (MF-PLPC) features with K-means clustering. The expected and observed frequencies were analyzed using chi-square distribution. Another work carried out by [52] has introduced a Voice-Input Voice-Output Communication Aid (VIVOCA). This assistive system was developed for speech-impairment people to help them with communication. VIVOCA includes mainly three modules, and the first one is the speech recognition module, the second one is the message building module from the recognized speech, and the third module is the speech synthesis module.

The decision level feature fusion method implemented for the Tamil database reported enhanced speech recognition results [1]. For the acoustic analysis, they applied the methods like the Gamma Tone filters (GTF) and Mel Frequency Perceptual Linear Predictive Cepstral (MF-PLPC). They used four different features such as Equivalent Rectangular Bandwidth (ERB) spaced GTF, BARK spaced GTF, MEL spaced GTF and MF-PLPC. Feature reduction was performed by using the Vector Quantization (VQ) technique, and the Fuzzy C Means (FCM) algorithm was used for classification. Also, the Multivariate Hidden Markov Model technique gives a good result in hearing-impaired speech recognition. This work has been extended by [52], in which they modified the feature extraction part. They introduced Modified Group Delay Features (MGDF), and Discrete Cosine Stockwell Transform Cepstrum (DCSTC) as modified features, and this system provides a

better recognition result.

So far, we have discussed the research works corresponding to acoustic data analysis developed for hearing-impaired speech recognition. The problems faced in single modality speech recognition can be handled using the multimodal concept in speech recognition. The AVSR studies for normal hearing are a hot research area in multimodal speech recognition. Initiating the research on hearing-impaired AVSR, it is important to identify the methods and techniques available for normal hearing AVSR. So, in the next section, we will be discussing the techniques that are applied in normal-hearing AVSR studies, and this can be extended for Hearing impaired AVSR research.

## 5. AUDIO-VISUAL SPEECH RECOGNITION

Visual cues such as lips and facial information can be utilized to perceive speech in extremely noisy conditions. By combining both audio and visual cues, we can perceive speech in a better way, and this motivated the research in multimodal speech processing, especially in AVSR. So, improvements in speech recognition can be attained by combining both audio and visual information using sophisticated MI techniques. With multimodal speech processing, the most important aspect is to make use of the benefit of information from both modalities effectively while ignoring the drawbacks of each. A wide study has been carried out in AVSR, and several works have been introduced for audio-visual feature extraction, audio-visual classification, and integration techniques. In this session, we will discuss the techniques that are used in AVSR systems.

The first and most important step in AVSR research is database creation. There are plenty of AVSR data corpora available at present, but several of them have limitations in terms of recording quality and the number of participants. Initial research was conducted on letter and digit recognition. Eventually, researchers focused on predicting words, phrases, and sentences from isolated and continuous speeches. In the past, databases were recorded at low resolution, but today they provide high-resolution videos, making Visual Speech Recognition (VSR) significantly better. The use of real-time acoustic noises for the development of AVSR systems has also been considered by some researchers [59, 71]. Most of the available databases are developed for isolated word or digit recognition rather than continuous speech recognition. The databases such as TULIPS, CUAVE, AVCAR, XM2VTS, and OULU VS are most commonly used for isolated word or digit recognition. For continuous speech recognition, the most commonly used databases include AV-TIMIT, LRS, and LRS-TED. The list of currently available AV databases can be found in Table.3 and the overview can be obtained from [74].

Table 3. Audio-visual Databases

| Database | Year | Type | No. of speakers | Description |
|---|---|---|---|---|
| TULIPS [53] | 1995 | Isolated | 12 | First four numerals |
| M2VTS [54] | 1997 | Continous | 37 | Isolated numerals |
| AVLETTERS [55] | 1998 | Isolated | 10 | English alphabets |
| XM2VTS [56] | 1999 | Continous | 295 | 3 Sentences |
| CUAVE [57] | 2002 | Isolated | 30 | Isolated or connected numerals |
| BANCA [58] | 2003 | Isolated | 52 | Numbers, name, date of birth, address |
| AVICAR [59] | 2004 | Isolated and Continuous | 84 | Letters, phone numbers, TIMIT sentences |
| GRID [60] | 2005 | Isolated | 34 | Commands |
| CMU-AVPFV [61] | 2007 | Isolated | 10 | Words |
| AVLETTER2[62] | 2008 | Isolated | 5 | Alphabets |
| MV-LRS [63] | 2009 | Isolated | 42 | Sentences |
| OULU [64] | 2009 | Isolated | 20 | Ten daily used phrases |
| VIDTIMIT [65] | 2010 | Isolated | 34 | Sentences |
| AusTalk [66] | 2014 | Isolated | 50 | Digits, sentences |
| TCD-TIMIT [67] | 2015 | Isolated | 20 | TIMIT sentences |
| MODALITY [68] | 2015 | Isolated | 35 | Commands |
| LRS [69] | 2017 | Continuous | 1000+ | Videos from BBC television |
| LRS-TED [70] | 2018 | Continuous | 1000+ | Lecture videos from TED |
| GRID-Lombard [71] | 2018 | Isolated | 54 | Phrases |
| RGB-D [72] | 2019 | Isolated | 53 | Twenty daily used phrases |
| VG digits [73] | 2020 | Continuous | 6 | Digits |

In AVSR design there are mainly three modules like Audio Speech Recognition (ASR), Visual Speech Recognition (VSR), and Audio-Visual Fusion (AVF). In AVSR, feature extraction techniques play a crucial role in enhancing the performance of the system. Most of cases, MFCC is the most commonly used feature extraction procedure for ASR. Visual feature extraction can be mainly classified as modal-based and image-based feature extraction. In model-based feature extraction, the Active Appearance Model (AAM) gained more attention [76,80] and in image-based feature extraction, pixel-based features are extracted by Discrete Cosine

Transform (DCT), Discrete Wavelet Transform, Principal Component Analysis (PCA), etc [75,78,79]. In, audio-visual fusion preferred models are the Hidden Markov Model (HMM) and the Multi-Stream Hidden Markov Model (MSHMM) [75,76,77,78,79]. The introduction of deep learning models enhanced the performance of ASR, VSR, and AVSR systems. CNN is now taking the place of traditional MFCC, DCT, and AAM in aspects of audio and visual feature extractions. Also, HMM and MSHMM are currently being substituted by Long Short-Time Memory networks (LSTM) or Bidirectional Long-Short-Term Memory networks (Bi-LSTM) in aspects of time sequence modeling [83,84,85]. As the audio information is not completely available for the hearing- impaired, the use of a relevant acoustic feature extraction method is very significant for AVSR. For the implementation of the visual module, a lot of visual feature extraction methods are proposed, and the methods like DCT, AAM, and DWT features give good recognition results. By reviewing the integration techniques, the MSHMM method is the most common late fusion method, and it provides a better fusion of audio-visual modality in AVSR. A graph depicting the result analysis of HMM and MSHMM models proposed by different authors is shown in figure 1. A comparison of the different technologies used for AVSR systems proposed by different authors is shown in Table.4. The result comparison was made based on Word Recognition Rate (WRR) in percentage. This section provides a brief overview of existing AVSR techniques and databases for normal hearing. The following section explores the challenges associated with AVSR systems for both normal hearing and hard-of-hearing.

Table 4. Audio-visual speech recognition techniques

| Authors | Feature Extraction | Integration | WRR in Percentage |
|---|---|---|---|
| [69] | MFCC+DCT | HMM | 72.4 |
| [68] | MFCC+AAM | HMM | 90.01 |
| [75] | AUTOENCODER+CNN | MSHMM | 75 |
| | | HMM | 85.04 |
| [76] | Gabor features | MSHMM | 98.89 |
| [77] | MFCC+DWT | HMM | 82 |
| [78] | MFCC+AAM | HMM | 80.8 |
| [79] | GFCC+OFA | HMM | 93.76 |
| | | MSHMM | 96.86 |
| [80] | MFCC+DBNF | DNN | 96.7 |
| [81] | MFCC+ AAM+ Geometric features | HMM | 80 |
| [82] | MFCC+3D lip features | MSHMM | 88.03 |
| [83] | CNN | LSTM | 95.5 |
| [84] | MFCC, ASM | RNN | 93.41 |
| [85] | Encoders | RNN | 96.7 |

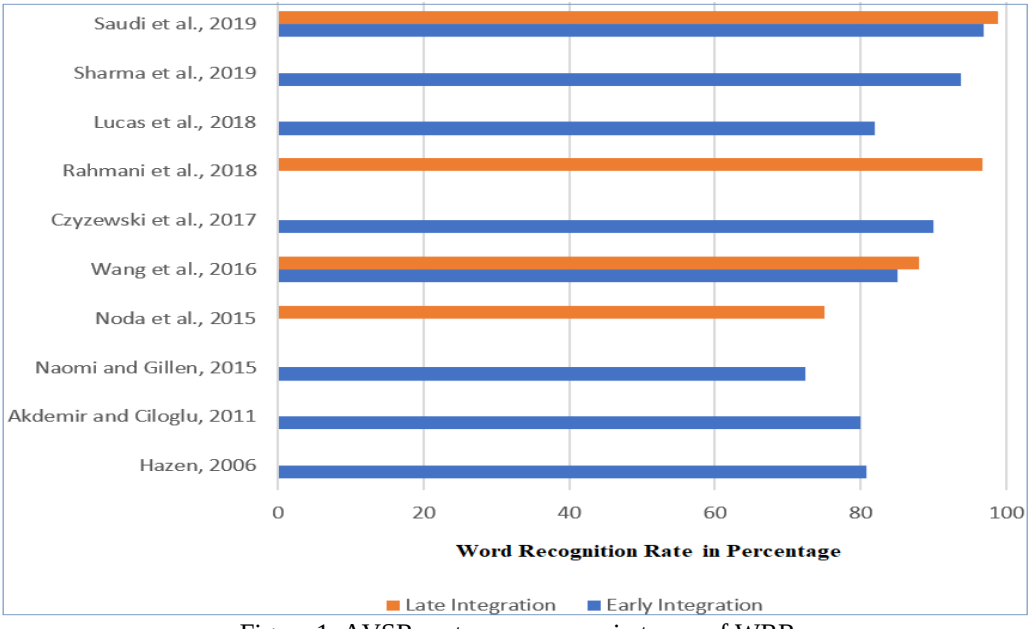


Figure 1. AVSR systems accuracy in terms of WRR

## 6. CHALLENGES IN AUDIO-VISUAL SPEECH RECOGNITION SYSTEMS

ASLR act as one of the most important multimodal system available within the literature. A lot of work was carried out for American sign languages, but in the case of ISL, it is in an infancy state. Dictionaries are created for ISL, which includes alphabets, grammar, family, fruits, etc., but all are developed based on American English. There are no standardized dictionaries not yet developed in the Indian language alphabet or grammar category. Alphabet signs are available for the Devanagari languages like Marathi, and Urdu, and also for the Dravidian language Tamil.

Malayalam, a classical language in India, is a Dravidian language spoken in the state of Kerala. According to Statistical profile 2016, the proposition of disabled persons to the total population of India is between 2.26 and 2.50 percent. 6.1% of children in Kerala were detected to have a hearing impairment, in 0.8% of children were detected to have severe hearing loss. So, standard sign language dictionary studies are more important for the Malayalam language. Also, normal hearing people communicate using their native language, and the accent varies from region to region in Kerala. By considering these, speech recognition becomes difficult for both hearing-impaired and normal hearing. In this case, it is a very difficult problem to be solved considering the state of Kerala.

As mentioned above, in the case of Indian languages like Malayalam, the main problem for SLR is there is no standardized format for sign language. Also, the system becomes more complex in the case of dynamic hand gesture recognition using both hands. In the case of CSR, the cued language is available for only a few languages. The CS is introduced for Indian languages like Hindi, Marathi, and Punjabi but no powerful repositories are available for these languages. The complexity associated with the VST is very high because the building blocks deal with ASR, speech synthesizer, etc. In such cases arises the importance of AVSR for the hard-of-hearing communication system. New research trends and improvements in hard-of-hearing speech recognition can be gained by introducing the AVSR technologies in such a situation. Also, in AVSR, the lip area is selected as a Region of Interest (ROI) for visual feature extraction. But in the case of a hard-of-hearing AVSR study, we have to focus on facial feature extraction techniques because they are more expressive while speaking. Considering all these factors, the main challenges faced by AVSR systems are discussed below.

- The major obstacle in AVSR research is the scarcity of suitable databases. For normal hearing, only a few databases are publicly available. As compared with the speech database, the AVSR database contains a smaller number of speakers and limited vocabulary. Most of the works concentrated on isolated word recognition rather than continuous speech recognition in AVSR. Also, no database is publicly available for Malayalam AVSR. For database creation, there is no described standard to follow. In earlier work single camera was used for recording, but nowadays, multiple cameras are experimented with AVSR to improve multi-pose lipreading and artificial noise is included in the database. So, Camera position, video quality, and noise level are considered before the database creation. Moreover, ethical issues should be considered while designing a database.
- In the case of hard-of-hearing AVSR studies, the main barrier is also the unavailability of a relevant hearing-impaired audio-visual database. Moreover, hard-of-hearing speeches are highly distorted, and there is an enormous variation in their speech regarding pronunciation style, energy duration, spectral components, etc. Children with mild hearing loss cannot understand either high-frequency or low-frequency components of speech. This makes them trouble in constructing vocabularies with proper grammatical structure. In conducting studies in AVSR, it is important to identify the hearing level of a person because the system characteristics are entirely different for mild to profound hearing loss. For that, we need the help of an audiologist to perform hearing tests like Auto Acoustic Emission (AAE). In addition to that, for hard-of-hearing database development, we need the help of an intermediator, or we have to use tactile methods like a speech therapist while recording.
- The next challenge corresponds to feature extraction. Even though MFCC is the most accepted technique in acoustic feature extraction, the feature extraction techniques can vary from normal hearing to hearing loss person. Hard-of-hearing people are more expressive in their way of communication so it is very difficult to capture the lip movements compared to normal-hearing people. An alternative to this, the feature extraction from facial parts like mouth, lips, eyes, and nose leads to getting more relevant information for hearing-impaired AVSR study.
- Finding the best technique for audio-visual integration that effectively utilizes the information from both modalities is also a key problem in AVSR.

## 7. CONCLUSION

In this paper, we reviewed the methods for hard-of-hearing person's acoustic analysis, multimodal interfaces, and multimodal integration systems, as well as techniques used in normal hearing AVSR studies. In the available acoustic analysis studies, the system evaluation is carried out for small vocabulary and, it gives an average recognition rate of 80%. Also, the experiments with MSHMM in the AVSR system provide a

minimum recognition rate of 75%. The paper studied the multimodal integration techniques, and the multimodal systems developed to support the hard-of-hearing in applications like ASLR, CSR, VST, etc. These multimodal experiments achieved an average improvement of 8% over the single-modal recognition system. Also, we have addressed the significant challenges faced by AVSR system studies. In these state-of-art techniques, AVSR systems can make a breakthrough in hard-of-hearing multimodal studies.

Future multimodal studies should focus on the development of an audio-visual speech database for the head-of-hearing. We also need to consider the development of techniques which is more relevant to the audio-visual feature extraction. The 3d feature extraction mechanism, which extracts features directly from the video, can improve the AVSR speech recognition results. The DNN has primarily driven advances in the video and audio domain, but the required amount of data for training such a system is much higher than the traditional algorithms. So, the large vocabulary database development is an important criterion that should be considered for future improvements in AVSR. Giving preference to facial feature extraction techniques can have the advantage of getting semantic information, which can be more useful in the case of hard-of-hearing speech recognition. Also, the inclusion of the prosodic acoustic features will provide additional content to semantic information. Furthermore, to model real-world problems, we must consider the actual and complex situations during database creation. In the end, the AVSR system helps to bridge the communication gap between hard-of-hearing and normal hearing people by providing native-language communication.

## ACKNOWLEDGEMENT

The authors are grateful to Miss. Ambili. K.R, Faculty Department of English, EKNM Govt. College for helping to revise the manuscript.

## REFERENCES

[1] Arunachalam. R, “A strategic approach to recognize the speech of the children with hearing impairment: different sets of features and models”, Multimedia Tools and Applications 78, pp.0787–20808, 2018.
[2] Carey. W. B, Crocker. A. C, Elias. E. R, Feldman. M, Coleman. W. L, “Developmental-Behavioral Pediatrics E-Book”, Elsevier Health Sciences, 2009.
[3] Papastratis, Ilias, Christos Chatzikonstantinou, Dimitrios Konstantinidis, Kosmas Dimitropoulos, Petros Daras. "Artificial Intelligence Technologies for Sign Language." Sensors 21, no. 17, 2021.
[4] Goehring. T, Bolner. F, Monaghan. J.J, Van Dijk. B, Zarowski. A Bleeck. S, “Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users”, Hearing research 344, pp.183–194, 2017.
[5] Lai. Y.H, Zheng. W.Z, “Multi-objective learning-based speech enhancement method to increase speech quality and intelligibility for hearing aid device users”, Biomedical Signal Processing and Control 48, pp.35–45, 2019.
[6] Kang, Yuyong, Nengheng Zheng, and Qinglin Meng, "Deep Learning-Based Speech Enhancement with a Loss Trading Off the Speech Distortion and the Noise Residue for Cochlear Implants", Frontiers in Medicine 8, 2021.
[7] Li, Lieber Po-Hung, Ji-Yan Han, Wei-Zhong Zheng, Ren-Jie Huang, and Ying-Hui Lai, "Improved Environment-Aware–Based Noise Reduction System for Cochlear Implant Users Based on a Knowledge Transfer Approach: Development and Usability Study", Journal of medical Internet research 23, no. 10, 2021.
[8] Xia, Linlin, Gang Chen, Xun Xu, Jiashuo Cui, and Yiping Gao, "Audiovisual speech recognition: A review and forecast", International Journal of Advanced Robotic Systems 17, no. 6, 2020: 1729881420976082.
[9] Seong, Thum Wei, and M. Z. Ibrahim. "A review of audio-visual speech recognition." Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 10, no. 1-4, pp 35-40, 2018.
[10] Baltrusaitis. T, Ahuja. C, Morency. L.P, “Multimodal machine learning: A survey and taxonomy”, IEEE Transactions on pattern analysis and machine intelligence 41, pp.423–443, 2018.
[11] Turk. M, “Multimodal interaction: A review”, Pattern Recognition Letters 36, pp.189–195, 2014. doi:10.1016/j. patrec2013 .07.003.
[12] Shoumy. N. J, Ang. L.M, Seng. K.P, Rahaman. D.M, Zia. T, “Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physio-logical signals”, Journal of Network and Computer Applications 149, pp.102447, 2019. doi: 10.1016/j.jnca.2019.102447.
[13] Argyropoulos. S, Moustakas. K, Karpov. A.A, Aran. O, Tzovaras. D, Tsakiris. T, Varni. G, Kwon. B, “Multimodal user interface for the communication of the disabled”, Journal on Multimodal User Interfaces 2(2), pp.105–116, 2008. doi:10.1007/s12193-008-0012-2.
[14] Bourbaki. N, Esposito. A, Kabraki. D, “Multimodal interfaces for interaction-communication between hearing and visually impaired individuals: problems and issues”, In:19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), vol. 2, pp. 522–530, 2007. doi:10.1109/ICTAI.2007.178.
[15] Lee. K.M, Han. K.I, Park. H. J, “Speech recognition and lip shape feature extraction for English vowel pronunciation of the hearing-impaired based on SVM technique”, Journal of rehabilitation welfare engineering & assistive technology 11(3), pp.247–252, 2017. doi:10.1109/BIGCOMP.2016.7425931.
[16] John ES, Rigo SJ, Barbosa J, “Assistive robotics: Adaptive multimodal interaction improving people with communication disorders”, IFAC-PapersOnLine, 49(30), pp.175-80, 2016.

[17] Bhuvaneshwari M, Kanaga EG, Anitha J, Raimond K, George ST, "A comprehensive review on deep learning techniques for a BCI-based communication system. Demystifying Big Data, Machine Learning, and Deep Learning for Healthcare Analytics", pp.131-57, 2021.
[18] Ince G, Yorganci R, Ozkul A, Duman TB, Köse H., "An audiovisual interface-based drumming system for multimodal human-robot interaction", Journal on Multimodal User Interfaces 15(4), pp.413-28, 2021.
[19] LaSasso. C.J, Crain. K.L, Leybaert. J, "Cued Speech and Cued Language Development for Deaf and Hard of Hearing Children", 1st edn. Plural Publishing, Inc, LosAngeles, 2010.
[20] Colin. S, Leybaert. J, Ecalle. J, Magnan. A, "The development of word recognition, sentence comprehension, word spelling, and vocabulary in children with deafness: A longitudinal study". Research in developmental disabilities 34, pp.1781–1793, 2013.
[21] Heracleous. P, Aboutabit. N, Beautemps. D, "Lip shape and hand position fusion for automatic vowel recognition in cued speech for French, IEEE Signal Processing Letters 16, pp.339–342, 2009. doi:10.1109/LSP.2009.2016011.
[22] Heracleous. P, Aboutabit. N, Beautemps. D, "Cued speech automatic recognition in normal-hearing and deaf subjects", Speech Communication 52, pp.504–512, 2010.
[23] Burger T, Caplier A, Perret P. Cued speech gesture recognition: a first prototype based on early reduction. EURASIP Journal on Image and Video Processing, pp.1-9, 2007.
[24] Arsov I, Jovanova B, Preda M, Preteux F, "On-line animation system for learning and practice Cued Speech", In International Conference on ICT Innovations, pp. 315-325, Springer, Berlin, Heidelberg, 2009.
[25] Liu L, Feng G, Beautemps D, "Inner lips feature extraction based on CLNF with hybrid dynamic template for Cued Speech", EURASIP Journal on Image and Video Processing (1), pp.1-5, 2017.
[26] Papadimitriou K, Parelli M, Sapountzaki G, Pavlakos G, Maragos P, Potamianos G, "Multimodal Fusion and Sequence Learning for Cued Speech Recognition from Videos", In International Conference on Human-Computer Interaction 2021 Jul 24 (pp. 277-290). Springer, Cham.
[27] Twahir MG, Yusof ZM, Ahmad I., "Malay Cued Speech Recognition Using Image Analysis: A Review. Advanced Materials and Engineering Technologies", pp. 319-25, 2022.
[28] Kumar. P, Gauba. H, Roy. P.P, Dogra. D.P, "A multimodal framework for sensor-based sign language recognition", Neurocomputing 259, pp.21–38, 2017. doi: 10.1016/j.neucom.2016. 08.132.
[29] Quesada Quirós L, López Herrera G, Guerrero Blanco LA. Automatic recognition of the American sign language fingerspelling alphabet to assist people living with speech or hearing impairments, 2017.
[30] Kulkarni S, Manoj H, David S, Madumbu V, Kumar YS, "Robust hand gesture recognition system using motion templates", In2011 11th International Conference on ITS Telecommunications, pp. 431-435, 2011.
[31] Boulares M, Jemni M, "Automatic hand motion analysis for the sign language space management", Pattern Analysis and Applications, pp.311-41, 2017
[32] Kumar, P., Roy, P.P., Dogra, D.P.: "Independent Bayesian classifier combination-based sign language recognition using facial expression. Information Sciences 428, pp.30–48, 2018. doi:10.1016 /j.ins.2017 .10.046.
[33] Caplier. Alice, Sébastien Stillittano, Oya Aran, Lale Akarun, Gérard Bailly, Denis Beautemps, Nouredine Aboutabit, Thomas Burger, "Image and video for hearing impaired people", EURASIP Journal on Image and Video Processing, pp.1-14, 2008.doi: 10.1155/2007/45641.
[34] Fagiani M, Principi E, Squartini S, Piazza F, "Signer independent isolated Italian sign recognition based on hidden Markov models", Pattern Analysis and Applications, pp.385-402, 2015.
[35] Ibrahim NB, Selim MM, Zayed HH, "An automatic Arabic sign language recognition system (ArSLRS)", Journal of King Saud University-Computer and Information Sciences, pp.470-7, 2018.
[36] Ravi S, Suman M, Kishore PV, Kumar K, Kumar A, "Multimodal spatiotemporal co-trained CNNs with single modal testing on RGB–D based sign language gesture recognition", Journal of Computer Languages, pp.88-102,2019.
[37] Jiang D, Ravyse I, Sahli H, Verhelst W, "Speech driven realistic mouth animation based on multi-modal unit selection", Journal on Multimodal User Interfaces, pp.157-69, 2008.
[38] Liu X, Yan N, Wang L, Wu X, Ng ML, "An interactive speech training system with virtual reality articulation for Mandarin-speaking hearing-impaired children", In2013 IEEE International Conference on Information and Automation (ICIA), pp. 191-196, 2013.
[39] Chen. Y.P.P, Johnson. C, Lalbakhsh. P, Caelli. T, Deng. G, Tay. D, Erickson. S, Broadbridge. P, El Refaie. A Doube. W, Morris. M.E, "Systematic review of virtual speech therapists for speech disorders", Computer Speech and Language 37, pp.98–128, 2016. doi:10.1016 / j. csl.2015.08.005.
[40] Tapu. R, Mocanu. B, Zaharia. T, "Deep-hear: A multi-modal subtitle positioning system dedicated to deaf and hearing-impaired people". IEEE Access 7, pp.88150–88162, 2019. doi:10.1109/ACCESS. 2019 .2925806.
[41] Cardenas. E.E, Chavez. G.C, "Multimodal hand gestures recognition combining temporal and pose information based on CNN descriptors and histogram of cumulative magnitudes", Journal of Visual Communication and Image Representation19, pp.21–38. 2020. doi: 10.1016/j.jvcir.2020.102772.
[42] Puviarasan. N, Palanivel. S, "Lip reading of hearing-impaired persons using HMM". Expert Systems with Applications 38, pp.4477–4481, 2011. doi: 10.1016/j.eswa.2010.09.119.
[43] Yamada Y, Javkin H, Youdelman K, "Assistive speech technology for persons with speech impairments", Speech communication, 30(2-3), pp179-87, 2000
[44] Yang, Hui-Jen, Yun-Long Lay, Chern-Sheng Lin, Pei-Yuan Hong, "The radar-graphic speech learning system for hearing impaired." Expert Systems with Applications 36, vol.no. 3, pp.4804-4809. doi: 10.1016/j.eswa.2008.05.053.
[45] A. Thida, N. N. Han, S. T. Oo, S. Li, C. Ding, "VOIS: The First Speech Therapy App Specifically Designed for Myanmar Hearing-Impaired Children," 2020 23rd Conference of the Oriental COCOSDA International Committee

for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), pp.151-154,2020. doi:2020.10.1109/OCOCOSDA503 38 .2020.9295024.
[46] Han, Kyung-Im, Hye-Jung Park, and Kun-Min Lee, "Speech recognition and lip shape feature extraction for English vowel pronunciation of the hearing-impaired based on SVM technique." In 2016 International Conference on Big Data and Smart Computing (BigComp*)*, pp. 293-296. 2016.
[47] Singh Shailendra, Anshul Thakur, Dharam Vir, "Automatic articulation error detection tool for the Punjabi language with aid for hearing impaired people." International Journal of Speech Technology 18, no. 2, pp.143-156. 2015. doi:10.1007/s10772-014-9256-2.
[48] Wang, Xu, Zhiyan Han, Jian Wang, and Mingtao Guo, "Speech recognition system based on visual feature for the hearing impaired." In 2008 Fourth International Conference on Natural Computation, vol. 2, pp. 543-546. IEEE, 2008. 10.1109/ICNC.2008.550
[49] Gudi AB, Shreedhar HK, Nagaraj HC., "Signal processing techniques to estimate the speech disability in children", International Journal of Engineering and Technology, 169, 2010.
[50] Revathi. A, Sasikaladevi.N, "Hearing impaired speech recognition: Stockwell features and models", Int J Speech Technology 22, pp.979–991.2019. doi: 10.1007/s10772-019-09644-3.
[51] Jeyalakshmi. C, Revathi. A, "Efficient speech recognition system for hearing impaired children in classical Tamil language", Int. J. Biomedical Engineering and Technology 26, pp.84–99, 2017. doi: 10.1504/IJBET.2018.089261.
[52] Hawley. M.S, Cunningham. S.P., Green. P.D, Enderby. P, Palmer. R, Sehgal. S, O'Neill. P, "A voice-input voice-output communication aid for people with severe speech impairment", IEEE Transactions on neural systems and rehabilitation engineering 21, pp. 23–31, 2012. doi: 0.1109/TNSRE.2012.2209678.
[53] Movellan. J.R, "Visual speech recognition with stochastic networks". Advances in neural information processing systems, pp.851–858, 1995.
[54] Pigeon. S, Vandendorpe. L, "The M2VTS multimodal face database (Release 1.00)", In: Bigün J., Chollet G., Borgefors G. (eds) Audio- and Video-based Biometric Person Authentication, AVBPA 1997, Lecture Notes in Computer Science, vol 1206. Springer, Berlin, Heidelberg, 1997. doi:10.1007/BFb0016021.
[55] Matthews. I, Cootes. T.F, Bangham. J.A, Cox. S, Harvey. R, "Extraction of visual features for lipreading", IEEE Transactions on Pattern Analysis and Machine Intelligence 24(2), pp.198–213, 2002. doi:10.1109/34.982900.
[56] Messer. K, Matas. J, Kittler. J, Luettin. J, Maitre. G, "Xm2vtsdb: The extended m2vts database", In: Second International Conference on Audio and Video-based Biometric Person Authentication, vol. 964, pp. 965–966, 1999.
[57] Patterson. E.K, Gurbuz. S, Tufekci. Z, Gowdy. J.N, "Cuave: A new audio-visual database for multimodal human-computer interface research", In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 2017, 2002. doi: 10.1109/ICASSP.2002.5745028
[58] Bailly Bailli´ere. E, Bengio. S, Bimbot. F, Hamouz. M, Kittler. J, Mariethoz. J, Matas. J, Messer. K, Popovici. V, Poree. F, "The banca database and evaluation protocol", In: International Conference on Audio-and Video-based Biometric Person Authentication, pp. 625–638. 2003.
[59] Lee. B, Hasegawa-Johnson. M, Goudeseune. C, Kamdar.S, Borys. S, Liu. M, Huang. T, "Avicar: Audio-visual speech corpus in a car environment". In: Eighth International Conference on Spoken Language Processing, 2004.
[60] Antonioletti. M, Atkinson. M, Baxter. R, Borley. A, Chue Hong. N.P, Collins. B, Hardman. N, Hume. A.C, Knox. A, Jackson. M, "The design and implementation of grid database services in ogsa-dai". Concurrency and Computation: Practice and Experience 17(2-4), pp.357–376, 2005.
[61] Kumar, Kshitiz, Tsuhan Chen, and Richard M. Stern. "Profile view lip reading." 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07. Vol. 4. IEEE, 2007.
[62] Cox, Stephen J., Richard W. Harvey, Yuxuan Lan, Jacob L. Newman, and Barry-John Theobald, "The challenge of multispeaker lip-reading", In AVSP, pp. 179-184, 2008.
[63] Chung, Joon Son, and A. P. Zisserman.,"Lip reading in profile", 2017
[64] Zhao. G, Barnard. M, Pietikainen. M, "Lipreading with local spatiotemporal descriptors", IEEE Transactions on Multimedia1 1(7), pp.1254–1265, 2009. doi:10.1109/TMM.2009.2030637.
[65] C. Sanderson and B.C. Lovell,"Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference", Lecture Notes in Computer Science (LNCS), Vol. 5558, pp. 199-208, 2009.
[66] D. Estival, S. Cassidy, F. Cox, and D. Burnham, ''AusTalk: An audiovisual corpus of Australian English,'' in Proc. Int. Conf. Lang. Resour. Eval., 2014, pp. 1–13.
[67] Naomi. H, Gillen. E, "TCD-TIMIT-An audio-visual corpus of continuous speech", IEEE Transactions on Multimedia17, pp.603–615, 2015. doi:10.1109/TMM.2015.2407694.
[68] Czyzewski. A, Kostek. P, Band Bratoszewski, Kotus. J, Szykulski. M, "An audio-visual corpus for multimodal automatic speech recognition", Journal of Intelligent Information Systems 49, pp.167–192, 2017. doi: 10.1007/s10844-016-0438-z.
[69] Chung. J.S, Senior. A, Vinyals. O, Zisserman. A, "Lipreading sentences in the wild", In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3444–3453, 2017. doi: 10.1109/CVPR.2017.367.
[70] Afouras. T, Chung. J.S, Zisserman. A, "Lrs3-ted: a large-scale dataset for visual speech recognition", arXiv preprintarXiv:1809.00496, 2018.
[71] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, ''A corpus of audio-visual lombard speech with frontal and profile views,'' J. Acoust. Soc. Amer., vol. 143, no. 6, pp. EL523–EL529, Jun. 2018.
[72] Ahmed. N, "Rgb-d dynamic facial dataset capture for visual speech recognition". In: 2019 International Conference on Image and Video Processing, and Artificial Intelligence, vol.11321, pp.1132108, 2019.

[73] Y. Lu and H. Li, ‘‘Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory,’’ Appl. Sci., vol. 9, no. 8, p. 1599, Apr. 2019.
[74] Shabina. B, Thasleema. T.M, Rajesh. R, “A Survey on Different Visual Speech Recognition Techniques”, Lecture Notes in Networks and Systems, Springer, Singapore, 1st edn., pp.304–316, 2019.
[75] Noda. K, Yamaguchi. Y, Nakadai. K, Okuno. H.G, Ogata. T, “Audio-visual speech recognition using deep learning”, Applied Intelligence 42, pp.722–737, 2015. 10.1007/s10489-014-0629-7.
[76] Saudi. A.S, Khalil. M.I, Abbas. H.M, “Improved features and dynamic stream weight adaption for robust audio-visual speech recognition framework”, Digital Signal Processing 89, pp.17–29, 2019. doi: 10.1016/j.dsp.2019.02.016.
[77] Lucas. D, Gonzalo. D.S, Juan. C.G, “Robust front-end for audio, visual and audio–visual speech classification”, International Journal of Speech Technology 21, pp.293–307, 2018. doi:10.1007/s10772-018-9504-y.
[78] Hazen. T.J, “Visual model structures and synchrony constraints for audio-visual speech recognition”, IEEE Transactions on Audio, Speech, and Language Processing 14, pp. 1082–1089,2006. doi:10.1109/TSA.2005.857572.
[79] Sharma. U, Maheshwar. S, Mishra. A.N, Kaushik. R, “Visual speech recognition using optical flow and hidden Markov model”, Wireless Personal Communications 106, pp.2129–2147, 2019. DOI: 10.1007/s11277-018-5930-z.
[80] Rahmani. M.H, Almasganj. F, Seyyedsalehi. S.A, “Improved features and dynamic stream weight adaption for robust audio-visual speech recognition framework”, Digital Signal Processing 82, pp.54–63, 2018. DOI: 10.1016/j.dsp.2019.02.016.
[81] Akdemir. E, Ciloglu. T, “Bimodal automatic speech segmentation based on audio and visual information fusions”, Speech Communications 53, pp.889–902, 2011.DOI: 10.1016/j.specom.2011. 03.001.
[82] Wang. J, Zhang. J, Honda. K, Wei. J, Dang. J, “Audio-visual speech recognition integrating 3d lip information obtained from the Kinect”. Multimedia Systems 22, pp.315–323. 2016. doi:10.1007/s00530-015-0499-9.
[83] Makino, Takaki, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan., "Recurrent neural network transducer for audio-visual speech recognition", In 2019 IEEE automatic speech recognition and understanding workshop (ASRU), pp. 905-912. IEEE, 2019.
[84] Kumar, L. Ashok, D. Karthika Renuka, S. Lovelyn Rose, and I. Made Wartana, "Deep Learning based Assistive Technology on Audio Visual Speech Recognition for Hearing Impaired", International Journal of Cognitive Computing in Engineering, 2022.
[85] Sterpu, George, and Naomi Harte, "Taris: An online speech recognition framework with sequence-to-sequence neural networks for both audio-only and audio-visual speech", Computer Speech & Language, 2022: 101349.