Visual Speech Recognition Using Lip Movement

Soham Akhade
Computer Engineering
Anantrao pawar college of engineering and research
Pune-411009
akhadesoham@gmail.com

Prof.J.C.Musale
Computer Engineering
Anantrao pawar college of engineering and research
Pune-411009
jitendra.musale@abmspcoerpune.org

Prof.S.J.Nawale
Computer Engineering
Anantrao pawar college of engineering and reseaarch
Pune-411009
sambhaji.nawale@abmspcoerpune.org

Omkar Jadhav
Computer Engineering
Anantrao pawar college of engineering and research
pune-411009
jadhav.om10903@gmail.com

Atharva Bhadale
Computer Engineering
Anantrao pawar college of engineering and research
Pune-411009
atharvabhadale1678@gmail.com

Prathmesh Gaikwad
Computer Engineering
Anantrao pawar college of engineering and research
Pune-411009
prathameshgaikwad902@gmail.com

ABSTRACT

Visual Speech Recognition(VSR) is a fleetly evolving field with different operations in mortal- computer commerce, availability, and security. This paper presents an innovative approach to VSR, fastening on the birth and analysis of lip movements for speech recognition. Traditional speech recognition systems calculate primarily on aural information, making them vulnerable to noisy surroundings and audio disturbances. In discrepancy, our proposed system leverages the visual modality by employing the rich information decoded in lip movements during speech product The study begins by collecting a comprehensive data-set of visual and audio recordings of speech in colorful languages and surrounds. latterly, a deep literacy armature is designed to process the visual data, emphasizing lip movements, and the corresponding audio data. The proposed model integrates conventional neural networks(CNN s) and intermittent neural networks(RNNs) to prize and fuse information from both modalities. This emulsion process enhances the robustness of the system by mollifying the limitations of traditional audio-only speech recognition We estimate the performance of the visualbased speech recognition system on a range of standard datasets and real- world scripts. The results demonstrate the efficacity of our approach, pressing its capacity to ameliorate recognition delicacy, particularly in noisy surroundings or situations where audio data is deficient or unapproachable In conclusion, our exploration contributes to the advancement of Visual Speech Recognition by introducing a new approach that emphasizes lip movement analysis. By using both audio and visual modalities, the proposed system provides a more robust and protean result for speech recognition, with the implicit to enhance operations in mortal- computer commerce, availability, and security.

Keywords— Visual Speech Recognition, Lip Movement Analysis, Multimodal Speech Recognition, Deep Learning, Convolutional Neural Networks (CNN)

I. INTRODUCTION

In our decreasingly connected world, the development of effective and protean speech recognition systems has surfaced as a vital frontier in mortal- computer commerce. These systems hold the implicit to transfigure the way we interact with technology, making it more intuitive and accessible for a broad range of operations. While audio-grounded speech recognition systems have made remarkable progress, they aren't without their limitations, particularly in noisy surroundings or for individualities with speech diseases. Visual speech recognition, a incipient yet promising field, offers a compelling result to these challenges by employing the intricate movements of the lips as a supplementary modality for speech understanding. The mortal capacity for speech perception is a remarkable feat of cognitive processing. When we engage in discussion, we don't solely calculate on audile information; our smarts painlessly integrate visual cues, similar as the movements of the speaker's lips, to enhance appreciation. Visual speech cues, though frequently taken for granted, give a wealth of information that can prop in disambiguating phonemes, words, and indeed entire rulings. Feting the eventuality of this multimodal approach, experimenters have turned their attention toward the development of visual speech recognition systems that harness the power of lip movement This innovative approach isn't only applicable for perfecting the robustness of speech recognition in noisy surroundings but also has significant counteraccusations for different operations, including mortal- computer commerce, assistive technology for those with speech impairments, and surveillance systems. By decrypting the visual cues handed by the lips, these systems can operate effectively in conditions where audio-only systems may falter. likewise, visual speech recognition has the implicit to break down walls for individualities who face challenges in using traditional audio- grounded speech recognition systems, therefore fostering inclusivity and availability in technology This paper explores the instigative sphere of" Visual Speech Recognition Using Lip Movement." It delves into the complications of this multimodal approach, encompassing data collection, preprocessing, deep literacy models, and the emulsion of audio and visual information. We bandy the challenges and complications essential in this field and highlight its implicit to transfigure the geography of speech recognition technology. As we trip through the realms of lip movement analysis and machine literacy, we will discover the pledge, openings, and challenges of employing visual cues to advance the borders of speech recognition

II. LITERATURE REVIEW

Kinfe Tadesse [1] developed a sub-word based isolated Amharic word recognition systems using HTK (Hidden Markov Model Toolkit). In this experiment, phones, triphones, and CV-syllables were used as the sub-word units and selected 20 phones out of 37 and 104 CV syllables for developing the system. The speech data of those selected recorded from 15 speakers for training and 5 speakers for testing. Average recognition accuracies of 83.07% and 78% were obtained for speaker dependent phone-based and triphone-based systems respectively.

Solomon Berhanu [2].The author developed isolated Consonant-Vowel syllable Amharic recognition system which recognizes a subset of isolated consonantvowel (CV) syllable using HTK (Hidden-Markov Modeling Toolkit). The author selected 41 CV syllables of Amharic language out of 234 and the speech data of those selected CV syllables were recorded from 4 males and 4 females with the age range of 20 to 33 years. The average recognition accuracies were 87.68% and 72.75% for speaker dependent and independent systems, respectively.

Asratu Aemiro [3] developed two types of Amharic speech recognition (ASR) systems, namely canonical and enhanced speech recognizers. The canonical ASR system is developed based on the canonical pronunciation model which consists of canonical pronunciation dictionary and decision tree. The canonical pronunciation dictionary is prepared by incorporating only a single pronunciation for each distinct word in the vocabularies. The canonical decision tree is constructed by only considering the place of articulations of phonemes as it was commonly used by the previous Amharic ASR researchers.

Petajan [4] A geometric features-based approach includes the first work on VSR done by Petajan in 1984, who designed a lip reading system to aid his speech recognition system. His method was based on using geometric features such as the mouth's height, width, area and perimeter.

Werda et al [5] where they proposed an Automatic Lip Feature Extraction prototype (ALiFE), including lip localization, lip tracking, visual feature extraction and speech unit recognition. Their experiments yielded 72.73% accuracy of French vowels, uttered by multiple speakers (female and male) under natural conditions.

Hazen et al [6] developed a speaker-independent audio-visual speech recognition (AVSR) system using a segment-based modelling strategy. This AVSR system includes information collected from visual measurements of the speaker's lip region using a novel audio-visual integration mechanism, which they call a segment-constrained Hidden Markov Model (HMM).

Gurban & Thiran [7] developed a hybrid SVM-HMM system for audio-visual speech recognition, the lips being manually detected. The pixels of down-sampled images of size 20 x 15 are coupled to get the pixel-to-pixel difference between consecutive frames.

Saenko et al [8] proposed a feature-based model for pronunciation variation to visual speech recognition; the model uses dynamic Bayesian network DBN to represent the feature stream.

Sagheer et al [9] introduced an appearance-based lip reading system, employing a novel approach for extracting and classifying visual features termed as "Hyper Column Model" (HCM).

Yau et al [10] described a voiceless speech recognition system that employs dynamic visual features to represent the facial movements. The system segments the facial movement from the image sequences using motion history image MHI (a spatio-temporal template). The system uses discrete stationary wavelet transform (SWT) and Zernike moments to extract rotation invariant features from MHI.

III. METHODOLOGY

Creating a visual speech recognition device involves a complex exploration methodology that encompasses colorful stages, from data collection to model development and evaluation. Then's a general exploration methodology that you can follow.

1. Problem description and thing Setting easily define the objects of your exploration and the specific problem you end to address with the visual speech recognition device. Identify the target operation and stoner conditions.

2. Collection Gather a substantial and different dataset of audiovisual speech recordings. This dataset should include colorful languages, accentuations, and speaking conditions to insure robustness.

3. Data Preprocessing Perform data preprocessing to clean and format the collected data. This may include coinciding audio and visual data, removing noise, and annotating phonetic information.

4. point birth Excerpt applicable features from the visual data, similar as lip shape, position, and stir circles. Consider using computer vision ways to prisoner these features effectively.

5. Deep Learning Model Selection Select applicable deep literacy infrastructures for the task. Common choices include convolutional neural networks(CNNs) for visual data and intermittent neural networks(RNNs) for successional data. Explore variants like Convolutional intermittent Neural Networks(CRNNs) for combined audiovisual processing.

6. Model Training Train the chosen deep literacy models using the preprocessed data. apply strategies like data addition to increase the model's robustness. insure that lip movement and audio data are accompanied during training

7. Cross-Modal Fusion Develop ways for integrating visual and audio data effectively. probe how to freight and combine these modalities for optimal recognition performance.

8. Evaluation Assess the performance of your visual speech recognition device using applicable evaluation criteria , similar as word error rate(WER) or phoneme error rate(PER). estimate it on standard datasets and real- world scripts.

9. Speaker Independence Testing Conduct trials to estimate how well your device performs across a range of speakers. Assess its conception capability.

10. Noise and Environmental Testing estimate the device's performance in noisy surroundings, low- light conditions, or scripts with background distractions. Explore styles to enhance noise robustness.

11. Real- Time Processing Optimize your device for realtime processing, as this may be pivotal for practical operations.

12. Ethical Considerations Address ethical and sequestration considerations associated with visual data. apply data protection and anonymization measures where necessary.

13. stoner Feedback and replication Gather stoner feedback and iteratively upgrade your device grounded on stoner input and realworld operation scripts.

14. Comparison to Baseline Models Compare the performance of your visual speech recognition device to birth models, including audio-only and audiovisual models, to demonstrate its efficacity.

15. Attestation and Reporting Document your exploration, methodologies, and findings in detail. Publish your results in exploration papers and reports.

16. Deployment and Integration If applicable, work on planting and integrating your visual speech recognition device into the target operations, similar as mortal-computer interfaces, assistive technology, or surveillance systems.

IV. Existing System

The being systems in Visual Speech Recognition using lip movement generally calculate on traditional audio-grounded speech recognition technologies. These systems generally use audio information without the significant objectification of visual cues. While they have been successful in a variety of operations, they face limitations in scripts with high noise situations, audio disturbances, or when audio information is inadequate. The traditional speech recognition systems are primarily base on the following

1. principle Acoustic Information Traditional systems generally concentrate on aural features uprooted from the audio signal, similar as phonemes, mel- frequence cepstral portions(MFCCs), and hidden Markov models(HMMs). These systems work well in controlled surroundings with clean audio data but tend to perform inadequately in noisy or adverse conditions.
2. Lack of Visual Data These systems don't effectively use the precious visual information conveyed by lip movements during speech. Lip movements are frequently not considered in the point birth or recognition processes, limiting the system's capability to perform in scripts where lip- reading could be salutary.
3. Limited Robustness The reliance on audio alone limits the robustness and rigidity of traditional speech recognition systems, making them less effective in operations where audio data may be corrupted or unapproachable.
4. Language and Context Dependency Traditional systems may struggle with feting languages or cants that they have not been specifically trained on, limiting their versatility.
5. Data Collection The system relies on expansive datasets of videotape clips featuring people speaking a wide range of words and rulings. These vids prisoner colorful speakers in different lighting conditions, with different accentuations and speech patterns. This different dataset is essential for training the deep literacy models effectively.
6. Preprocessing The input vids suffer preprocessing, where facial milestones and lip movements are tracked and uprooted. colorful computer vision ways are used to descry and detect the lips within each frame of the videotape.
7. Deep Learning Models Deep neural networks, similar as convolutional neural networks(CNN s) and intermittent neural networks(RNNs), are employed to reuse the lip movement data. These models are trained to fete patterns in the lip movements and prognosticate corresponding phonemes, words, or rulings.
8. Multi-Modal Fusion To enhance delicacy, this system may combine visual lip reading with traditional audio-grounded speech recognition. By integrating the information from both modalities, it becomes more robust in colorful realworld scripts.
9. Real- Time Processing The system is optimized for realtime processing, making it suitable for operations where immediate responses are needed. This is particularly important in surrounds like availability for the hail disabled or mortal- computer commerce.
10. Affair The honored speech is converted into textual or audile affair, making it accessible to druggies. This affair can be used for colorful operations, including furnishing real- time mottoes , controlling bias through speech commands, or abetting communication for individualities with hearing impairments.
11. Feedback and enhancement Visual speech recognition systems frequently incorporate feedback mechanisms to upgrade their delicacy over time. This feedback may involve stoner corrections, model retraining, or nonstop system updates to acclimatize to different speakers and conditions.
12. Integration Visual speech recognition can be integrated into a wide range of operations and diligence. It has operations in availability tools for the hail bloodied, mortal- computer commerce, surveillance, gaming, education, and more.

V. Problem In Existing System

In the being system of Visual Speech Recognition(VSR) using lip movement, several significant challenges and limitations persist, hindering the system's overall effectiveness. One of the most burning issues is the lack of a comprehensive and different dataset for training and testing VSR models. The problem can be described as follows inadequate Dataset Diversity and Size The being VSR systems heavily calculate on training datasets that are frequently limited in terms of diversity and size. This limitation poses several critical problems.

1. conception Issues numerous VSR models are trained on fairly small datasets that contain a limited range of speakers, languages, accentuations, and speech surrounds. As a result, these systems struggle to generalize effectively to a broader population, leading to reduced recognition delicacy when faced with speakers, accentuations, or languages not adequately represented in the training data.
2. LimitedCross-Cultural connection The lack of different datasets in terms of artistic backgrounds and languages

impedes the development of VSR systems that can be applied encyclopedically. Being systems tend to perform well for certain languages or societies but may perform inadequately for others, limiting theircross-cultural connection.

3. shy Representation of Real- World Variability VSR models trained on small datasets frequently fail to capture the full range of real- world variability in lip movements during speech. Factors similar as varying lighting conditions, facial expressions, and speaker movements are not adequately addressed, performing in reduced robustness in real- world scripts.

4. failure ofIntra-Speaker VariabilityIntra-speaker variability, which includes changes in lip shape and movement patterns for the same speaker across different utterances, is frequently underrepresented in the being datasets. This limitation hampers the model's capability to fete speech directly when speakers parade natural variations in their lip movements.

5. shy Noise and Environmental Variability The being datasets frequently warrant the objectification of environmental noise and hindrance, making it grueling to develop VSR systems that perform reliably in noisy or adverse conditions.

VI. Proposed System

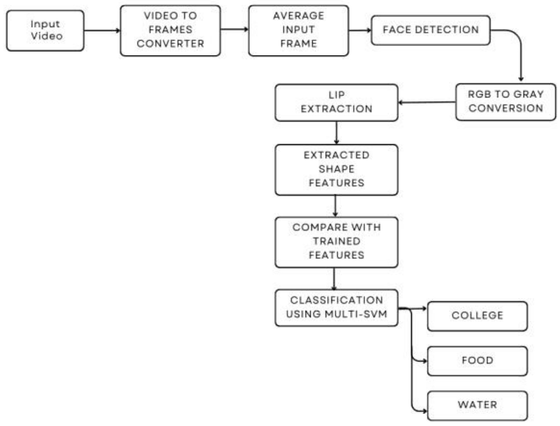
In the being system of Visual Speech Recognition(VSR) using lip movement, several significant challenges and limitations persist, hindering the system's overall effectiveness. One of the most burning issues is the lack of a comprehensive and different dataset for training and testing VSR models. The problem can be described as follows inadequate Dataset Diversity and Size The being VSR systems heavily calculate on training datasets that are frequently limited in terms of diversity and size. This limitation poses several critical problems

1. conception Issues numerous VSR models are trained on fairly small datasets that contain a limited range of speakers, languages, accentuations, and speech surrounds. As a result, these systems struggle to generalize effectively to a broader population, leading to reduced recognition delicacy when faced with speakers, accentuations, or languages not adequately represented in the training data.

2. LimitedCross-Cultural connection The lack of different datasets in terms of artistic backgrounds and languages impedes the development of VSR systems that can be applied encyclopedically. Being systems tend to perform well for certain languages or societies but may perform inadequately for others, limiting theircross-cultural connection.

3.shy Representation of Real- World Variability VSR models trained on small datasets frequently fail to capture the full range of real- world variability in lip movements during speech. Factors similar as varying lighting conditions, facial expressions, and speaker movements are not adequately addressed, performing in reduced robustness in real- world scripts. .

VII.Flowchart Of Proposed System



VIII. Algorithm

	DESCRIPTION
1	Data Collection: Gather a dataset of video clips or images of people speaking with synchronized audio, and annotate the corresponding transcripts or phonetic representations.
2	Preprocessing: Normalize the data by resizing and cropping the lip region in each frame to ensure uniformity.
3	Lip Region Extraction: Detect and extract the lip region from each frame using face detection or lip-tracking techniques.
4	Feature Extraction: Extract relevant features from the lip region in each frame, such as optical flow, lip shape, or texture information.
5	Model Selection: Choose a suitable machine learning model for lip movement recognition, often based on Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs).
6	Training: Train the selected model on the feature representations and their associated transcripts using labeled data.
7	Inference: Apply the trained model to new lip movement data to predict the spoken words or phonemes, typically frame by frame.
8	Post-processing and Evaluation: Combine the frame-level predictions into meaningful words or sentences and evaluate the system's accuracy using metrics like word error rate (WER) or character error rate (CER).
9	Conclusion: Visual speech recognition using lip movement is a challenging but valuable technology with applications in various fields, such as accessibility for the hearing impaired and security. This algorithm provides a high-level overview of the process, from data collection and preprocessing to feature extraction, model training, and evaluation. Continuous refinement and optimization of the model and features are essential to enhance the system's accuracy and real-world applicability.

IX. ADVANTAGES OF PROPOSED SYSTEM

1. Bettered Availability The system can enhance availability for individualities with hail impairments, furnishing them with a means to understand spoken language through visual cues.
2. Enhanced Security It can be employed in security and authentication systems, adding a biometric subcaste grounded on lip movement for access control and surveillance operations.
3. Accurate in Noisy surroundings Visual speech recognition is effective in noisy surroundings where traditional audio- grounded speech recognition systems may struggle. This makes it precious in scripts like busy public spaces and artificial settings.
4. Multimodal Communication When combined with audio- grounded speech recognition, it can produce a more robust and accurate recognition system, making it suitable for a wide range of operations.
5. Human- Computer Interaction The proposed system can ameliorate mortal- computer commerce, making it more natural and intuitive, particularly in virtual reality, stoked reality, and other interactive technologies.
6. Language Learning and Pronunciation It has operations in education by aiding in language literacy, pronunciation enhancement, and speech remedy, furnishing precious feedback to learners and preceptors.
7. Healthcare Support In the healthcare sector, the system can help in monitoring and assessing cases with speech diseases, abetting in speech remedy, and enhancing communication between cases and healthcare providers.
8. Emotion and Sentiment Analysis It can fete emotional and sentiment cues during spoken language, which is precious in colorful operations similar as client feedback analysis, request exploration, and happy recommendation.
9. Multilingual Support The system can be extended to support multiple languages, making it a precious tool for language restatement andcross-linguistic communication.
10. Robotic operations It can be integrated into robotic systems, enabling robots to more understand and respond to mortal communication, making them more useful in colorful disciplines, including client service and healthcare.
11. Real-time communication the system’s capacity to deliver information in real-time makes it indispensable in circumstances requiring quick feedback and action.
12. In dynamic circumstances relationships benefit from inclusive communication bridge barriers and promote inclusive communication

X. CONCLUSION

In summary, lip movement-based visual speech recognition is a rapidly developing field that has the potential to fundamentally alter how we communicate and engage with technology. Its operations are wide-ranging and promising, ranging from improving accessibility for the injured caused by hailstorms to refining security and surveillance, education, healthcare, and online retail. The technology's implied ability to lubricate emotion and sentiment analysis, as well as its capacity to condense audio- based speech recognition in noisy environments, underscore its significance. However, as the field of visual speech recognition advances, ethical questions arise, notably with regard to data operation and sequestration. To handle these businesses and ensure responsible development and execution, certain rules and regulations must be put in place. Utilizing the full potential of this technology will require cooperation between experimenters, masterminds, and subject matter specialists, improving its accuracy, adaptability, and accessibility. With continued development and innovation, lip movement-based visual speech recognition is expected to have a substantial impact on a wide range of fields and help ensure that communication and computer-mediated commerce remain secure, efficient, and inclusive in the future. As this technology continues to evolve, it also brings forth ethical considerations, particularly regarding sequestration and data operation. It's imperative to establish robust regulations and guidelines to address these enterprises and insure responsible development and perpetration. Collaboration between experimenters, masterminds, and sphere experts will be necessary in employing the full eventuality of this technology, making it more accurate, protean, and accessible. Also, visual speech recognition has the implicit to revise educational styles by furnishing real- time feedback on pronunciation and language literacy. In the healthcare sector, it can help in speech remedy, enabling therapists to more understand and support their cases. also, its operations in the entertainment and gaming diligence are measureless, offering further immersive and interactive gests . In the realm of security and surveillance, visual speech recognition can significantly ameliorate the delicacy of relating individualities, offering an added subcaste of authentication. This technology's integration into law enforcement and public safety could give inestimable backing in working crimes and precluding security breaches. In summary, with ongoing advancements and invention, visual speech recognition using lip movement is poised to significantly impact a multitude of diligence and contribute to a further inclusive, effective, and secure future of communication and humancomputer commerce. As we continue to explore and expand the boundaries of this technology, its implicit to compound and transfigure our diurnal lives remains both instigative and promising. still, it's imperative that we address ethical considerations and unite across disciplines to insure that these advancements are made responsibly and with a focus on serving humanity as a whole.

XI. FUTURE SCOPE

The unborn compass of visual speech recognition using lip movement is both extensive and transformative, with farreaching counteraccusations across multitudinous disciplines. As technology develops, this sector is anticipated to have a significant impact on how accessibility for the hard of hearing is defined because it offers subscribe language restoration, real-time captioning, and improved communication skills. It implies that human-computer commerce ought to be reexamined, particularly in the context of enhanced reality and virtual reality operations, where interactions ought to be more smooth, natural, and immersive. Visual speech recognition is a subcaste of biometric authentication based on lip movements that can enhance security and surveillance systems' access control and monitoring capabilities. Similarly, its conjunction with traditional audio- based speech recognition is anticipated to significantly enhance delicacy and robustness, enabling more accurate and environment-aware speech recognition in scripts with possibly sparse or ambiguous audio cues as well as loud surroundings.

The healthcare sector is another area where visual speech recognition is poised to make a substantial impact, abetting in speech remedy, covering speech diseases, and enhancing case- croaker

communication. In education, it can help language literacy and pronunciation enhancement, furnishing inestimable feedback to learners and preceptors. Beyond these operations, visual speech recognition's implicit extends to emotion and sentiment analysis, enabling further nuanced understanding of communication. It also plays a part in robotic operations, making mortal- robot communication more intuitive and effective, especially in dynamic or noisy surroundings. Multilingual andcross-linguistic operations are on the horizon, contributing to language restatement and cross-cultural communication. As the technology evolves, it will spark important conversations about sequestration and ethics, challenging the development of strict regulations and guidelines regarding data operation and stoner concurrence. The cooperative sweats of experimenters, masterminds, and sphere experts will continue to shape the future of visual speech recognition, rendering it an decreasingly precious and protean tool with broad- ranging counteraccusations in availability, mortal- computer commerce, security, healthcare, education, and beyond.

XII.References

[1] Fenghour, S., Chen, D., Guo, K. and Xiao, P., 2020. Lip reading sentences using deep learning with only visual cues. IEEE Access, 8, pp.215516-215530.

[2] Wand, M., Koutník, J. and Schmidhuber, J., 2016, March.
Lipreading with long short-term memory. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6115-6119). IEEE.

[3] Potamianos, G., Neti, C., Luettin, J. and Matthews, I., 2004. Audio-visual automatic speech recognition: An overview. Issues in visual and audio-visual speech processing, 22, p.23.

[4] Cox, S.J., Harvey, R.W., Lan, Y., Newman, J.L. and Theobald, B.J., 2008, September. The challenge of multispeaker lip-reading. In AVSP (pp. 179-184).

[5] Hilder, S., Harvey, R.W. and Theobald, B.J., 2009, September. Comparison of human and machine-based lip-reading. In AVSP (pp. 86-89).

[6] Chung, J.S., Senior, A., Vinyals, O. and Zisserman, A., 2017, July. Lip reading sentences in the wild. In 2017 IEEE conference on computer vision and pattern recognition (CVPR) (pp. 3444-3453). IEEE

[7] P. Mohanaiah, P. Sathyanarayana and L. GuruKumar, Feature Extraction using GLCM", International Journal of Research Publications, Vol. 3, Issue 5, 2013.

[8] View Wen Chin, Li-Minn Ang and Kah Phooi Seng, "Lips Detection for Audio-Visual Speech Recognition System", International Symposium on
Intelligent Signal Processing and Communication Systems, 2008.

[9] Yong-Ki Kim, Jong Gwan Lim and Mi-Hye Kim, "Comparison of Lip Image
Visual Speech Recognition Based on Lip Movement for Indian Languages 2041
Feature Extraction Methods for Improvement of Isolated Word Recognition
Rate", Advanced Science and Technology Letters Vol. 107, pp. 57-61, 2015.

[10] Seman, N., Bakar, Z.A., Bakar, N.A., "An evaluation of endpoint detection
measures for Malay speech recognition of isolated words," Information
Technology (ITSim), 2010 International Symposium, Vol. 3, pp. 1628-1635,
2010.

[11] Cheang Soo Yee, Ahmad, A.M., "Malay language text-independent speaker
verification using NN-MLP classifier with MFCC," International Conference,
pp. 1-5, 2008.

[12] Matthews, I., Bangham, J.A., Cox, S., "Audiovisual speech recognition using
multi-scale nonlinear image decomposition," Spoken Language, ICSLP 96.
Proceedings. Fourth International Conference, Vol. 1, pp. 38-41, 1996.