Intelligent Methods in Engineering Sciences 3(1): 001-006, 2024

IMIENS

INTELLIGENT METHODS
IN ENGINEERING SCIENCES

https://www.imiens.org

International
Open Access

March, 2024

e-ISSN 2979-9236

Research Article

https://doi.org/10.58190/imiens.2024.81

Detection of Emergency Words with Automatic Image Based Lip Reading Method

Beyza ULKUMEN [a,*], ***Ali OZTURK*** [a,b]

[a] *KTO Karatay University, Konya, TURKIYE*
[b] *Havelsan A.S, Konya, TURKIYE*

ARTICLE INFO

Article history:
Received 21 December 2023
Accepted 06 March 2024

Keywords:
lip reading,
convolutional neural networks,
SSD

ABSTRACT

Lip reading automation can play a crucial role in ensuring or enhancing security at noisy and large-scale events such as concerts, rallies, public meetings, and more by detecting emergency keywords. In this study, the aim is to automatically detect emergency words from the lip movements of a person using images extracted from silent video frames. To achieve this goal, an original dataset consisting of silent video images in which 8 emergency words were spoken by different 14 speakers was used. The lip regions of the images obtained from the videos in the dataset were labeled through relevant region detection. Labeled data were then evaluated using the SSD (Single Shot MultiBox Detector) deep learning method. Subsequently, subsets of labeled data with 8, 6, and 4 classes were created. The SSD algorithm was evaluated separately for each of these subsets. During the training of the SSD algorithm, weight initialization methods such as 'he,' 'glorot,' and 'narrow-normal' were used, and their performances were compared. Additionally, the SSD algorithm was trained with two different values of the maxepochs parameter, which were 20 and 30, respectively. According to the results, the lowest accuracy value was found for the 8-class subset, with an accuracy of 42% obtained using 20 epochs of training and the 'narrow-normal' weight initialization method. The highest accuracy value was achieved for the 4-class subset, with an accuracy of 76% obtained using the 30 epochs of training and the 'glorot' weight initialization method.

1. INTRODUCTION

Lip reading, also known as visual speech recognition, refers to deciphering the content of spoken text based on visual information related to the movement of the speaker's lips. Speech recognition, public safety, intelligent human-computer interaction [1–3], visual synthesis, and various other applications use it. When looking at the development history of lip-reading technology, it is evident that both the software and hardware for lip synchronization have been continuously and rapidly advancing not only in theoretical research but also in practical applications.

Lip reading technology advancements, particularly in scenarios where audio signals are weak or completely absent, have the potential to enhance coping mechanisms. The ability of lip reading technology to accurately and swiftly identify words from lip movements can be a significant advantage in situations where there are disruptions in sound-based communication. In instances of communication breakdowns in emergency situations, the challenges faced by lip reading technology become crucial in addressing compromised or non-existent audio signals. Especially in the context of emergency word identification, lip reading technology plays a critical role in tackling situations where sound signals are compromised or unavailable.

Emphasizing the critical role of automatic lip reading in situations where audio signals are compromised or non-existent highlights how important this technology can be in emergency scenarios. In situations with limitations in sound-based communication, the potential of lip reading technology to facilitate accurate communication can be a significant factor in enhancing the effectiveness of emergency communication. Therefore, efforts to overcome the challenges of lip reading technology, especially in emergency scenarios, are of great importance in supporting accurate and effective communication.

Lip-reading technology was first proposed as an idea by Sumby in 1954 [4]. Sumby described lip movements as a visual information source used to understand speech. Lip reading is the process of interpreting lip movements. Additionally, some information about facial expressions and emotions can also be gathered from lip movements.

In 1994, Hidden Markov Models (HMMs) were first

* **Corresponding Author:** beyzaulkumen.98@gmail.com

utilized as a novel lip-reading technology, specifically for a speaker-dependent lip-reading system. HMMs were among the first methods used in traditional lip-reading automation [5, 6]. The use of HMMs is generally explained as a visual- auditory system known as "lip reading" aimed at improving automatic speech recognition. One of the most crucial aspects is the integration of a sound-based automatic lip- reading system, particularly to enhance accuracy under degraded acoustic conditions. Additionally, various feature extraction techniques have been utilized in traditional lip-reading methods, including Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Discrete Cosine Transforms (DCTs), and Active Appearance Models (AAMs), among others [7–10].

The primary methods for model-based feature extraction include Active Shape Models (ASM) and AAMs. The Snake algorithm, originally proposed by Kass and others [11], is also known as the Active Contour Model (ACM). ACM primarily detects key points of the lips based on specific conditions. It then uses these points to create a curve that determines the shape and smoothness of the lips. This process ensures that the lips take on the desired form through the interaction of constraints and energy coefficients. According to a study by Dinh and Milgram [12], the difficulty in accurately detecting lips in noisy environments can arise due to frequent trapping in local minima under such conditions. The suggested approach to overcome this problem is a multi-featured ASM that combines different features such as the normal contour, gray blocks, and Gabor wavelets, rather than relying on a single feature.

Developments in visual speech recognition systems have brought about a significant transformation, particularly in recent years with the utilization of deep learning networks. The growing interest in feature extraction and classification tasks has propelled advancements in this technology to a pioneering position. The effective and customized use of deep learning networks has strengthened the capability of visual speech recognition systems to address the increasing complexity effectively. Ngiam and colleagues initially proposed a deep visual- auditory speech recognition system based on Restricted Boltzmann Machines (RBMs) [8, 13]. This has replaced traditional methods used in lip-reading automation, such as PCA, with the use of neural networks.

The use of Convolutional Neural Networks (CNN) is a significant milestone in computer vision tasks, and the study conducted by Krizhevsky, A. and Sutskever, I., & Hinton, G.E. indicates a notable performance improvement in this field [14]. CNN, particularly by performing convolution operations on image data, is effectively employed in visual recognition tasks. This has contributed to achieving higher accuracy and precision in tasks such as object detection and classification.

In the field of object detection, methods such as Single Shot Multibox Detector (SSD), YOLO (You Only Look Once), Faster-RCNN are widely preferred and successful models in computer vision. SSD, aiming to perform object detection quickly and effectively, consolidates the process into a single deep network, successfully combining feature maps at different scales, demonstrating the ability to detect objects of various sizes [15].

The ability of deep learning networks to combine feature extraction and classification tasks in the field of object detection has been emphasized in the study conducted by Ren, S., He, K., Girshick, R., et al. [16]. This approach has led to the emergence of models like Faster R-CNN, significantly enhancing the performance of object detection. Models such as Faster R-CNN have made groundbreaking progress in object detection, offering higher accuracy and lower error rates.

In the realm of automatic lip reading, geometric-based, image-based, model-based, and motion-based methods are utilized [17, 18]. Geometric-based approaches analyze lip structures through mathematical models and geometric features, while image-based methods identify lips using visual features such as color, brightness, and contrast. Model-based techniques detect lip structures using learned data or statistical models, while motion-based methods aim to represent lips by analyzing lip movements. The combination of these methods allows for more robust and accurate results in the field of lip reading.

In this research, an original lip-reading system capable of detecting emergency words in Turkish has been developed. The system can accurately recognize eight distinct emergency words, such as "Bomb," "help," "blood," "plan," "police," "weapon," "injury," and "aid." These words encompass expressions commonly used in emergency and security contexts in the Turkish language.

The main highlight of the study is the presence of various Turkish words in the specially created Turkish dataset. This emphasizes the potential of lip-reading technology to provide a more effective solution to the challenges of limited datasets in the Turkish language. The developed lip-reading system stands out as an effective tool, especially in crowded environments and large events, with its high accuracy in recognizing Turkish emergency words.

This lip-reading system can play a significant role in enhancing security measures in Turkish-speaking regions, particularly in critical areas such as counterterrorism and crime detection. The high accuracy rates obtained indicate that the system can be effectively utilized in real-time emergency detection.

2. Automated Lip Reading

Automatic lip-reading systems follow a series of processing steps, including visual input, pre-processing,

feature extraction, classification, and understanding meaningful speech, for feature extraction.

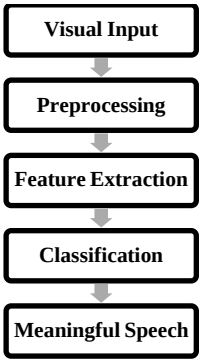

Figure 1. Process Steps for Automated Lip Reading

For a recorded video of a speaking person, an automatic lip-reading system first needs to sample the video into individual frames. Frames containing the speakers are then cut from the video and converted into black and white image frames, completing the preprocessing stage. After the video is sampled, it is necessary to extract only the lips of the speaker using Region of Interest (ROI) detection. Once the lip region is identified and labeled in every frame of the video containing a human face, a dataset is obtained.

2.1. Dataset

In this study, a dataset obtained from 14 different speakers was used. Among these speakers, there are 6 females and 8 males. This dataset consists of silent video frames, and the speakers have uttered various emergency words (e.g., bomb, help, blood, plan, police, gun, wound, help).

Figure 2. Frames Extracted from Speakers' Videos

Instantaneous photos were cropped from the speakers' videos and transformed into black and white visual frames as illustrated in Figure 2.

The lip regions on the human faces in each video frame were labeled with bounding boxes, as shown in Figure 3, and assigned to their respective classes.

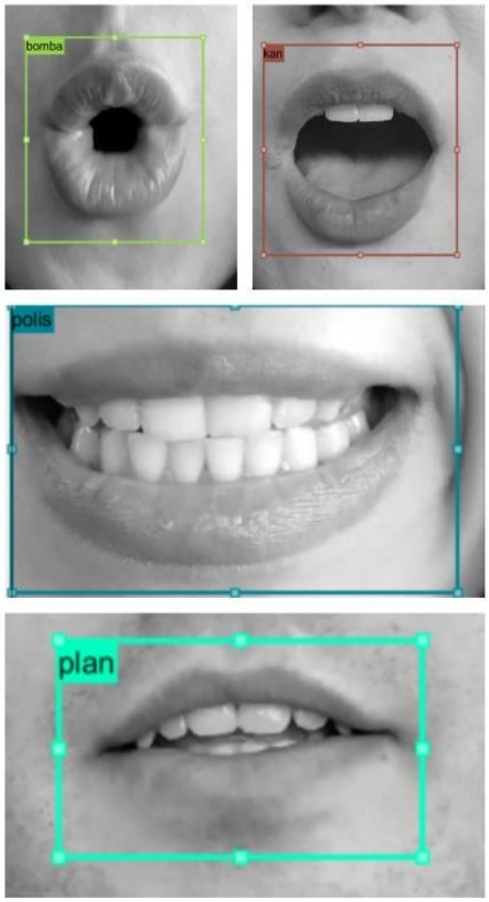

Figure 3. Examples of Bounding Boxes for Emergency Words

2.2. The SSD Model

The Single Shot Detector (SSD) algorithm is a leading model in the field of deep learning, particularly in object detection and image processing. SSD stands out for its ability to detect objects in a single shot, making it suitable for essential tasks in image analysis and object recognition. SSD [15] is constructed on top of a "base" network that ends (or is cut off at the end) with some convolutional layers. Each added layer, along with some of the previous base network layers, is used to predict scores and distances for predefined bounding boxes. This model has the capability to accurately determine the precise locations of objects within an image by detecting objects in the image. By performing both object detection and localization in a single algorithm, it offers an effective and efficient solution. SSD is considered a significant development in the field of deep learning and is successfully employed in numerous application areas, playing a crucial role in this domain.

The working principle of SSD starts with extracting feature maps from the input image. Feature maps are data representations that capture the characteristics of objects in desired regions of the image. In this study, the lips present in the image serve as the data representing our feature maps. These feature maps contribute to our object recognition process by emphasizing the characteristics of the lips.

Following the feature map extraction phase, SSD generates proposal boxes (anchor boxes) on these feature maps. These proposal boxes are configured to suit user preferences in various sizes and aspect ratios on the image. This customizability allows for more efficient and faster performance in tasks such as automatic image-based lip reading and object detection while conserving computational resources.

Subsequently, SSD successfully accomplishes two fundamental tasks for each proposal box. The first and crucial step involves a regression process to determine the precise location of the object, aiming to predict the exact position of the lips in the image using coordinate transformations. The second step is the classification of the lip image. In the classification stage, the lips inside the proposal boxes are assigned to specific categories based on visual data.

SSD performs object detection using a multi-stage process that includes feature extraction, proposal box generation, and classification/regression phases. This multi-stage approach of SSD enhances its success in object detection, providing effective results in terms of both accuracy and speed.

3. Experimental Results

When training the SSD model, the selection of the 'weightsInitializerValue' parameter used to determine the initial weights is a critical decision. This choice can influence the model's learning process inception and its outcomes. In this study, the effects of using three different initializers named 'he', 'glorot', and 'narrow-normal', along with both 'MaxEpochs (MaxEpochs = 20, MaxEpochs = 30)' values, on the success and application time are evaluated. These initializer types set the initial weights of the model in different ways and can affect the model's learning capability and success rate. The 'MaxEpoch' parameter is a crucial factor that influences both the learning time and the accuracy performance of the SSD model.

The "He" initialization method sets the weights randomly but adjusts these initial values based on the number of inputs to the layer, making the weights more stable during the training process. The "Glorot" initialization, on the other hand, provides a general-purpose approach compatible with different activation functions but has a narrower variance range compared to "He." The "Narrow- Normal" initialization is used in specific cases and aims to initialize the weights within a particular range.

When comparing the results, the performance of all three initializer types and both 'MaxEpochs' values on success criteria and training times were evaluated. Performance on the original dataset was analyzed to determine which initializer and 'MaxEpochs' value provided the most successful results. Additionally, the impact of these parameters on computation time was assessed. The selection of which initializer ('he', 'glorot', and 'narrow- normal' initializer types) and 'MaxEpochs' ('MaxEpochs=20' and 'MaxEpochs=30') value was determined by comparing the best results and the most suitable training time for the original dataset.

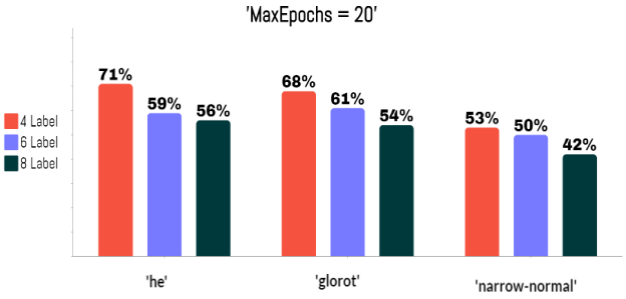

Figure 4. Results for 'MaxEpochs = 20' for Emergency Words

Figure 4 displays the accuracy results for 'MaxEpochs = 20'. When the weight initializer type is 'narrow-normal', it achieved the lowest accuracy with 53%, 50%, and 42% for all datasets, respectively. The highest accuracy for the 4-label dataset was achieved with 71% using the 'he' weight initializer. For the 6-label dataset, the highest accuracy of 61% was obtained through training with the 'glorot' weight initializer. In the case of the 8-label dataset, the highest accuracy of 58% was achieved with the 'he' weight initializer.

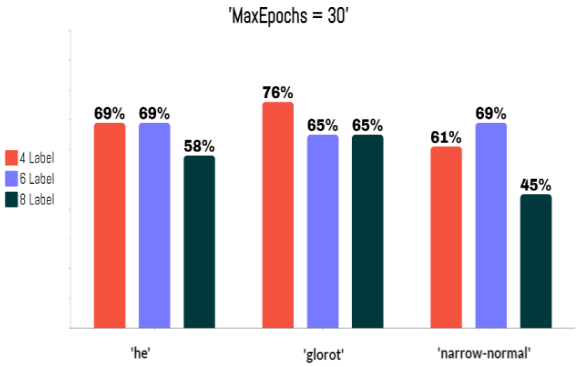

Figure 5. Results for 'MaxEpochs = 30' for Emergency Words

Figure 5 shows the accuracy results for 'MaxEpochs = 30'. When the weight initializer type is 'narrow-normal', it achieved the lowest accuracy with 61%, 69%, and 45% for all datasets, respectively. The highest accuracy for the 4-label dataset was achieved with 76% using the 'glorot' weight initializer. For the 6-label dataset, the highest accuracy of 69% was obtained through training with both the 'he' and 'narrow-normal' weight initializers. In the case of the 8-label dataset, the highest accuracy of 65% was achieved with the 'glorot' weight initializer.

According to the results of the trials, as shown in Figure 5, the combination of 'MaxEpochs = 30' and the 'glorot' weight initializer type has achieved the highest success compared to all other trial combinations in the field of automatic image-based lip reading. These findings

indicate that when the object detection model undergoes a longer training period, it can yield superior results, and the 'glorot' initializer, by initializing the weights more evenly, significantly contributes to enhancing this success. These findings underscore the importance of fine-tuning deep learning methods and the model's long-term training in achieving more effective results in complex visual tasks.

Based on the information obtained from all of this research, the most successful weight initializer ranking is as follows: "Glorot," "He," and "Narrow-Normal."

These results indicate that the "Glorot" method is the preferred choice when initializing deep learning models. Additionally, the value of "MaxEpochs=30" has led to better results due to providing more training time, although it significantly increased the computational time.

The obtained results strongly emphasize the importance of optimizing parameters during the initialization and training stages of deep learning models. Carefully tuning parameters such as the choice of weight initializer and the number of epochs is shown to be a critical step in achieving higher levels of success in complex visual tasks.

Table 1 shows a comparison of the performance of different deep learning algorithms on various datasets in the literature with our study. According to this table, when compared to similar studies, our study achieved above-average success with a 76% accuracy rate in recognizing emergency words.

To further enhance the performance, the next step in this study in the field of image-based lip reading for emergency words is to improve the quality and diversity of the original dataset. This will be achieved by focusing on data collection and processing processes. Enriching and diversifying our dataset will lead to better training of our model and further improve its performance.

Additionally, enhancing the quality of the dataset will contribute to making the model more reliable, thus helping achieve more effective results in real-world applications. These strategies are crucial steps in making a more significant impact in the field of automatic image-based lip reading.

Table 1. Comparison of Our Study with the Literature

Author	Dataset	Techniques	Performance
Jake Burton [19]	TCD-TIMIT	HMM, CNN, LSTM	69.58 %
Fenghour,S. [20]	BBCLRS2	Viseme classifiers& attention based transformer.	64.6 %
Fung, I.,& Mak, B[21]	Ouluvs2 10 phrasetask	CNN in addition to BLSTM witha fusion of max out kindling units	87.6 %
Wand,M [22]	GRID corpus	LSTM	85 % to 95 %
Kastaniotis,D [23]	CUAVE and VPR are used inthis study	PR, HMM, ASR, RBM, VPR, HMM,ASR, RBM, DBN	45.63 %
Abderrahim Mesbaha [24]	AVLetters, Ouluvs2, BBC LRW	Heterogeneous Convolutional Neural Networks	90.86 %
N. Puviarasan[25]	The recorded video of speakers	HMM withDCT and DWT features	DCT –91 % DWT –97%
Apurva H. Kulkarni [26]	IBMVI A VOICECUAVE AVLletters2	CNN+BLSTM, DNN and LSTM	87.6 %
Our Study	***Recorded emergency words video of speakers***	***SSD***	***76 %***

4. Conclusion

In this study, a custom dataset was created with the specific purpose of automatically detecting lip movements in human speech to identify emergency words. This dataset was designed not only to include words associated with emergency situations but also to provide a detailed examination of lip movements and the speech process. The dataset comprises examples that reflect various variables such as different age groups, genders, language accents, and speech rates.

This unique dataset was enriched with images obtained from videos that recorded lip movements and gestures to understand how people use emergency words in different speech scenarios.

This enables the understanding of the relationship between lip movements and emergency words and demonstrates the potential for developing more effective emergency detection systems.

The study's 76% success rate highlights the potential for significant improvement in the learning capability and performance of the SSD model by optimizing parameters such as initialization methods and epoch count. Subsequent research will focus on enhancing the quality and diversity of the original dataset while incorporating various deep learning methods in a hybrid approach.

Moreover, future studies will emphasize the integration of user authentication and feedback features into lip reading systems, aiming to further optimize this integration for increased accessibility and effectiveness. In-depth analysis of user feedback will provide valuable insights, allowing for the development of tailored recommendations to align the system with individual preferences. These technical advancements are anticipated to contribute positively to the ongoing evolution of lip-reading systems.

Acknowledgments

The authors equally contributed to this study and they

declare that they have no conflicts of interest.

This study has been presented in 7th International Conference on Engineering Technologies (ICENTE 2023), 23-25 November 2023, Konya/Turkey.

References

[1] Potamianos, G.; Neti, C.; Gravier, G.; Garg, A.; Senior, A.W. Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE 2003*, 91, 1306–1326.

[2] Akhtar, Z.; Micheloni, C.; Foresti, G.L. *Biometric liveness detection: Challenges and research opportunities.* IEEE Secur. Priv. 2015, 13, 63–72.

[3] Rekik, A.; Ben-Hamadou, A.; Mahdi, W. Human machine interaction via visual speech spotting. In International Conference on Advanced Concepts for Intelligent Vision Systems; Springer: Berlin/Heidelberg, Germany, 2015; pp. 566–574.

[4] W. H. Sumby and I. Pollack, Erratum: *Visual contribution to speech intelligibility in noise*, J. Acoust. Soc. Am. 26(2) (1954) 212–215.

[5] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, ''A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments,'' in *Proc. 6th Int. Conf. Multimodal Interface (ICMI)*, 2004, pp. 235–242.

[6] S. Lee and D. Yook, ''Audio-to-visual conversion using hidden Markov models,'' in *Proc. 7th Pacific Rim Int. Conf. Artif. Intell., Trends Artif. Intell.*, 2002, pp. 563–570.

[7] I. Almajai, S. Cox, R. Harvey, and Y. Lan, ''Improved speaker independent lip reading using speaker adaptive training and deep neural networks,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2722–2726.

[8] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, ''Multimodal deep learning,'' in *Proc. 28th Int. Conf. Mach. Learn., (ICML)*, 2011, pp. 1–8.

[9] J. Huang and B. Kingsbury, ''Audio-visual deep learning for noise robust speech recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7596–7599.

[10] K. Thangthai, R. Harvey, S. Cox, and B.-J. Theobald, ''Improving lipreading performance for robust audiovisual speech recognition using DNNs,'' in *Proc. Int. Conf. Auditory-Visual Speech Process.*, Sep. 2015, pp. 127–131.

[11] M. Kass, A. Witkin, and D. Terzopoulos, ''Snakes: Active contour models,'' *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[12] Q. Dinh Nguyen and M. Milgram, ''Multi features active shape models for lip contours detection,'' in *Proc. Int. Conf. Wavelet Anal. Pattern Recognit.*, vol. 1, Aug. 2008, pp. 172–176.

[13] H. Lee, C. Ekanadham, and A. Y. Ng, ''Sparse deep belief net model for visual area V2,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 873–880.

[14] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Commun. ACM 2017, 60, 84–90. [Google Scholar] [CrossRef]

[15] W. Liu, D. Anguelov, D. Erhan, S. Christian, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: single shot multibox detector. In ECCV, 2016. 1, 3, 4, 6, 7, 8.

[16] S. Ren, K. He, R. Girshick, et al. (2016). "Faster R-CNN: Rapid Object Detection." Presented at IEEE Conference on Neural Information Processing Systems (NIPS), 2016.

[17] *S. Dupont and J. Luettin,* ''Audio-visual speech modeling for continuous speech recognition,'' *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000.

[18] *Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, ''*A review of recent advances in visual speech decoding,'' *Image Vis. Comput.*, vol. 32, no. 9, pp. 590–605, Sep. 2014.

[19] Burton, J., Frank, D., Saleh, M., Navab, N. and Bear,H.L., 2018, December. The speaker-independent lipreading play-off; a survey of lipreading machines. In 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS) (pp. 125- 130). IEEE.

[20] 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.

[21] Fung, I. and Mak, B., 2018, April. *End-to-end low-resource lip-reading with maxout CNN and LSTM*. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2511-2515). IEEE.

[22] 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.

[23] Kastaniotis, D., Tsourounis, D. and Fotopoulos, S., 2020, October. *Lip Reading modeling with Temporal Convolutional Networks for medical support applications.* In 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 366-371). IEEE.

[24] Mesbah, A., Berrahou, A., Hammouchi, H., Berbia, H., Qjidaa, H. and Daoudi, M., 2019. *Lip reading with Hahn convolutional neural networks.* Image and Vision Computing, 88, pp.76-83.

[25] Puviarasan, N. and Palanivel, S., 2011. *Lip reading of hearing-impaired persons using HMM.* Expert Systems with Applications, 38(4), pp.4477-4481.

[26] Kulkarni, A.H. and Kirange, D., 2019, July. Artificial Intelligence: *A Survey on Lip-Reading Techniques.* In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.