Novel Algorithm Utilizing Deep Learning for Enhanced Arabic Lip Reading Recognition

Doaa Sami Khafaga[1], Hanan A. Hosni Mahmoud[2], Norah S. Alghamdi[3], Amani A. Albraikan[4]
Department of Computer Sciences, College of Computer and Information Sciences
Princess Nourah bint Abdulrahman University, Riyadh, 11047, KSA

***Abstract*—Computerized lip reading is the science of translating visemes and oral lip reading into written text, where visemes are lip movement without sound. Video processing is applied for the recognition of those visemes. Previous research developed automated systems to for computerized lip reading recognition to be hearing-impaired aid. Many challenges face such an automation process, including insufficient training datasets. Also, speaker-dependency is one of the challenges that are faced. Real-time applications which respond within a specific time period are also widely required. Real-time human computer interaction are systems which require real time response. Response time for human computer interaction is measured by number of elapsed video frames. Video processing of lip reading necessitates real-time implementation. There are applications for viseme recognition, as an aid for deaf people, video games with human computer interaction, and surveillance systems. In this paper, a real-time viseme recognition system is introduced. In order to enhance existing methods to overcome these pitfalls, this paper proposed a computerized lip reading technique based on feature extraction. We utilized blocks arrangement techniques to reach a near-optimal appearance feature extraction technique. A Deep Neural Network is utilized to enhance recognition. The benchmark dataset SAVE, for Arabic visemes, is employed in this research, and high viseme recognition accuracies are achieved. The described computerized lip reading recognition technique is advantageous for the hearing impaired and for other speakers in noisy environments.**

***Keywords*—Lip reading; deep CNN; deep learning; recognition; viseme**

I. Introduction

Computerized lip reading is the viseme understanding of lips movements and convert it to written text for both the hearing impaired and for other speakers in noisy environment [1]. Image and video processing techniques have been employed in such applications. For example, computerized video systems were built for automated lip reading systems. In this research, an approach is presented which examines feature extraction of the computerized lip reading models. Classification is also investigated. Features extraction techniques are applied to lip reading images and discriminative ones are chosen. In the classification phase, a Deep Neural Network is utilized. Deep learning (DNN) usually utilize raw features for their input stage, our work is engineered on extracted features as the DNN input to yield a discriminative DNN.

Real-time viseme recognition systems are very crucial in different paradigms such as surveillance, as a hearing aid utilized in video conferencing also for video games with lip reading interfaces especially in noisy environments. Most of these applications anticipate a specific response time, usually real time. Real-time viseme recognition systems is driven by the lip reading movement recognition using video processing techniques.

Real time viseme recognition could help as a communication linguistic for people with disabilities. It should be noted that different languages have different visemes for their alphabet. Also, viseme recognition can be utilized as an interface for playing a video game effectively.

Viseme recognition uses computer vision methodologies. Where a video sequence, is obtained as an input and the corresponding alphabet or word are produced as output. Statistical modelling, pattern recognition, machine learning techniques are also widely utilized. In this paper, we study the performance of employing spatial and temporal video sequence segmentation for feature extraction.

A viseme can be defined as a sequence of lip reading video frames. In order to identify a viseme, the lip area from the main frame is segmented, then work on the segmented portion of the whole viseme video sequence. Real-time requirements make it difficult since response time must fall in a strict time interval for the procedure to be satisfactory. On the other lip, to identify lip area, we might need skin color information. Skin color information enables the identification process. It can help in determining the lip regions in a satisfactory amount of time. Skin color segmentation can differentiate between lip area and other areas in the video frames. We can apply it to determine lip-like pixels in an image which can be a binary classification problem.

This paper is divided as follows: Related work is discussed in Section 2. Section 3 presents the dataset and the methodology. Experimental results are depicted in Section 4. Conclusions are discussed in Section 5.

II. Background

Lip movement Shape and viseme features extraction are two groups of challenges facing computerized lip reading recognition. In [2], feature enhancement processes of each speaker with normalization, and hierarchical feature arrangement are used to decrease the effects of speaker differences. In [3], a CNN architecture is proposed, using the audio decoding process and isolated viseme pronunciation of audio commands. Accuracy results are computed as visemes of Russian vowels are recognized. They noted that the dependence of the recognition accuracy uses photo features, and the used camera is investigated. The accuracy of specific speaker recognition is computed as 83% for a specific camera, where first and second moments are applied using a support vector machine. Research in [4], investigated the utilization of deep neural network classifiers in visual feature extraction. The authors in [5] coupled a discrete likelihood transform followed by adaptive pre-training in the visual feature extraction. Validation is performed utilizing the Gaussian probabilistic model recognizers, and word repetition rates for twenty different speakers. The recognition accuracy is 55.5% on average.

The computerized lip reading recognition system in [6], uses a CNN with feature extraction process. The network is pre-trained with the lip area videos coupled along with its text labels. Seven speakers are utilized in the validation process, with seven independent CNN architectures. Their system achieved an average viseme recognition of 59%.

The authors in [7], proposed lip viseme modalities through multimodal learning methods. In [8], two deep CNNs are trained using text labels and video frames and their final layers are fused to extract mutual features to be classified by a deep network. Accuracy of 65% is achieved through this bi-model. In [9], the computerized lip reading is performed using a processing pipeline of neural networks. In [10], Short Memory is coupled with long memory to form a uni-model. They utilized raw lip images as the CNN inputs, the performance of this uni-modal is validated against Support Vector Machine with accuracy averaging over different speakers to be 69%.

Geometric features are extracted from the lip regions like lip width, height and orientation [11]. The authors in [12] proposed a skin color segmentation process to differentiate between mouth and non-mouth areas utilizing convex hull algorithm. Lip features including ratio between height and width and color properties were extracted utilizing multi-dimension time warping technique [13]. Several geometric properties are defined to analyze the lip shape in the research performed by the authors in [14]. These properties are depicted as follows: lip height, lip width, lip area, long diagonal height, short diagonal width, nose to lip distance, lip to chin distance. The active contour method was utilized to find the lip contour to extract the geometric properties. These properties can extract the mouth height and width that encompass the information needed for lip reading [15]. In supervised lip reading video analysis, the coordinates are defined on different face points in each video frame of the video taken for each individual in the training set. Useful Geometric features for lip reading were defined by these coordinates. The height of the lips curvature posed a challenge for recognition due to the low similarity of symmetric sides [16]. Lip tracking models faces several challenge such as the robustness of the model in terms of testing with a small sample of individuals [17]. Authors in [18] utilized two public audio-visual databases to increase robustness and accuracy of their method. The databases are listed in [19-20] as AV-CM [19] and AVLetters [20].

The method in [21] performed better than the Hidden Markov classifier when utilizing the CMU dataset to compute statistical mouth contour algorithm. Spatiotemporal features such as viseme visual features, the AVLetters dataset will perform up to better accuracy as compared with other models. UNR dataset was collected by authors in [22] and compiled lip shape features and is recognized with better accuracy than achieved by the HMM classifier. The results of the research found in [23-24] using deep learning method is depicted to be better across all three databases.

The authors in [25-26] proposed an image dataset with two partitions, a training partition of eight subjects with 5000 viseme images. The other partition contains 2000 images from five subjects. The images are of resolution is 256×256 pixels. The experiments were performed to compute the accuracy of the lip reading model to utilize the Arabic language as a test language.

The current research exhibits that the extracted lip visemes are mostly speaker-dependent when executing lip reading recognition with the lip movement style of an a specific utterer is different from the utterers in the training set. This drawback can lead to recognition accuracy to drop greatly. Hence, in this research the training using deep learning will extract independent features that are relevant to the viseme features that are independent to the variations of the speakers. Materials and Methods.

The proposed computerized lip reading model is depicted in this section. The model has two major phases of viseme feature extraction and CNN training and classification. The first phase utilizes video frames to extract visual features. The second phase utilizes a deep CNN network with the parameters tuned and examined.

A. Viseme Feature Extraction

Viseme features are mainly appearance features [27-28]. Past research investigated those features and pruned their dimensionally through a reduction procedure and can be employed as the inputs to the deep CNN to enhance classification. In this research shape and appearance features were extracted and were precisely examined. The introduced lip-reading method, has the principle phase of viseme feature extraction (FE) and viseme recognition (VR). Video frames are extracted from the viseme video sequence and the viseme features are extracted from them. The second phase which is the VR utilize an optimized CNN architecture with tuned parameters.

The process of our viseme appearance feature extraction model deploys each video as a set of frames, the pre-processing phase is done to extract the mouth area in the video frames. We utilized the Viola-jones technique depicted in [29-30] The Viola-jones is utilized to detect the face and the lips region, which is defined as the region of interest (ROI).

The SAVE dataset includes recorded videos of Arabic visemes using the NTST standard of 30 frames per seconds [30-31]. The Audio files are also included for validation of the output of the lip reading recognition system. The extracted features will be stored in the appearance vectors. Video frames are extracted from the viseme video sequence and the viseme features are extracted from them. The second phase which is the VR utilize an optimized CNN architecture with tuned parameters.

Viseme features are categorized as appearance features. Other researchers utilized feature dimensions pruning process and are utilized as the inputs to CNN [9]. In this paper, feature reducing process were extracted and applied to appearance features.

The proposed process of our viseme feature extraction model is depicted in Fig. 1. Each video is transformed to a set of frames, and a pre-processing phase is performed to extract region of interest of the lip area. The lip area is up-scaled to size of 30×40 pixel-image, and represented with a 1200-element vectors. Component analysis (CA) will reduce the vector dimensionality reduction to 64 dimensions with the highest probability feature elements. According to the experimental results we found that 64 element dimensions will achieve better performance than reducing dimensions to 32, 96, and 128 features using the CA algorithm. The whole scheme is depicted in Fig. 1.

In the SAVE dataset [19-20], the videos are recorded with the Pal standard of 30 frames per seconds. Audio files are converted and labelled as text of the visemes for the supervised learning process. The mapping of the labelled text and the viseme are saved. We performed up-sampling of the feature vectors using both the first and second derivatives. Normalization is the carried using the z-score process presented in [22-24]. Then we combined 13 successive Z-normalized feature vectors to produce an 832 single hyper-feature-vector. These successive frames are defined as six frames before and six frames after the central frame. The combined hyper-feature-vector, is defined as the appearance feature vector.

There are two main norms of classifying visemes: according to the face appearance, the form and shape of the lips, teeth clarification during the pronunciation of particular phonological units, and according to the sounds with a matching visual illustration.

The second description is particularly popular since it expedites the learning phase of training and testing data.

B. Deep CNN Network Classifier

The deep CNN is a neural network, where units are connected through weighted connections using Gaussian distributions [12]. Pre-training layers are utilized to compute the initial weights [12], and a fine-tuning methodology is added to compute the supervised cross-entropy. Deep CNNs are usually utilized as a generative process but with some modifications, it can be utilized in classification [13].

After extracting the video frames from the input video we start the feature extraction phase, a matrix, with samples and features are represented in the rows and columns respectively. A vector representing a video file with label of classes is made, where classes are numbered 0 to 18, and a fine-tuning methodology is added to compute the supervised cross-entropy. Deep CNNs are usually utilized as a generative process but with some modifications, it can be utilized in classification. The classes are labeled by those Arabic phonemes {‘اه’, ‘با’, ‘تا’, ‘را’, ‘اوه’, ‘وا’, ‘اي’, ‘نا’, ‘تا’, ‘اوه’, ‘يا’, ‘ووه’, ‘ما’, ‘مي’, ‘فا’, ‘في’, ‘سا', ‘سو’, ‘كا’}. Those vectors create a super-vector.

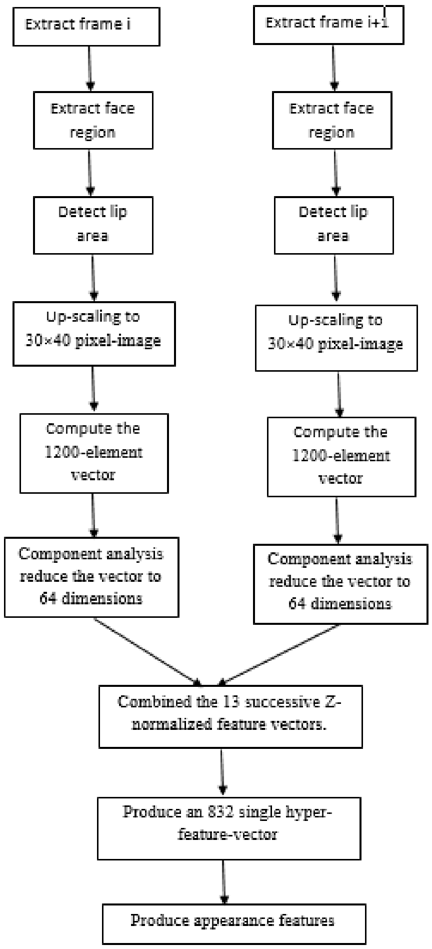

Extract frame i
Extract frame i+1
Extract face region
Extract face region
Detect lip area
Detect lip area
Up-scaling to 30×40 pixel-image
Up-scaling to 30×40 pixel-image
Compute the 1200-element vector
Compute the 1200-element vector
Component analysis reduce the vector to 64 dimensions
Component analysis reduce the vector to 64 dimensions
Combined the 13 successive Z-normalized feature vectors.
Produce an 832 single hyper-feature-vector
Produce appearance features

Fig. 1. The Appearance Features Extraction Process.

That is utilized in the training stage. Both the matrix and the super-vector are utilized in the input phase and the output phase of the deep CNN. The deep CNN is depicted in Fig. 2.

To extract the viseme features, a Shape Model is utilized and eighty landmarks are marked to represent the coordinates of different face points. Twenty of those coordinates define the lips contour region, and are utilized to define a 42 feature vector.

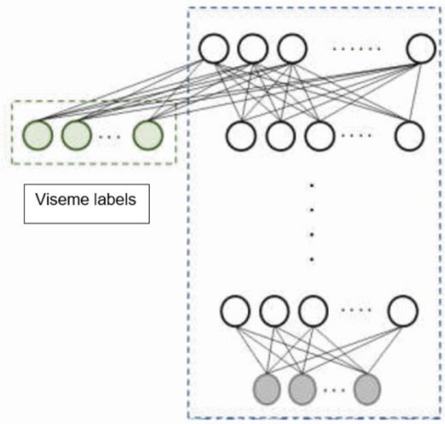


Fig. 2. The CNN Architecture of the Lip-reading Model.

To employ the deep CNNs, we create a window of twelve successive frames that are represented by feature vectors. The CNN output layer has n Softmax classifiers. The deep CNNs are trained using a stochastic gradient method with a batch of size 64 cases with fifty epochs with rate of 0.1. Momentums of values 0.6 down to 0.1 and are utilized in various settings. The momentum has a maximum value of 0.6 decreasing to 0.1 after six epochs of the training stage. Errors are computed for each frame as a probability over different possible labels for each video frame. The accuracies are computed on the viseme-level.

The CNN architecture of the lip-reading model has the first block describes the features extraction phase and the second block represents the deep CNN. The final block is the decoder. The viseme high feature extraction model and the deep CNN classifier validations are performed on an 8 Core CPU and 64 Gigabyte RAM, and a single GTX X graphic unit with 16 Gigabyte Graphics memory.

III. Experimental Result

The used dataset over the base method, and the sets of the extracted features for the computerized lip reading are discussed. To validate the strength of our method we included the proposed viseme features in the base model.

A. *Dataset*

The experiments are performed using the SAVE dataset [20]. SAVE is a corpus of thirty speakers pronouncing 8000 connected visemes. In this research, the experiments are performed using the speakers uttering of the corpus excluding the profile of the speakers. This is performed with only the front view of the lips area.

B. *Base Lip Reading Recognition Model*

Hidden Markov Models (HMM) are studied for the base lip reading recognition phase. In this research, the HTK software toolkit is utilized on the frames of the lip movements. It is the base model that applies several context-independent HMMs. For each viseme model, a 3-state HMM with covariance GMM over five units is utilized. The introduced viseme features extraction model coupled with the discrete cosine transform and shape model are employed in the base model.

To extract the viseme features, a Shape Model is utilized and eighty landmarks are marked to represent the coordinates of different face points. Twenty of those coordinates define the lips contour region, and are utilized to define a 42 feature vector.

To compute the Geometric viseme features, high level features, including lip contour region height, width, and area are computed. The DCT coefficients of the lip area are computed. The lip area is down-sampled into a 16×16 intensity blocks, and is sampled into one vector using zigzag reading style. The up-triangle six coefficients per each frame are attained.

The appearance viseme feature extracted set is also applied to the base HMM model to validate their strength. These extracted features are used before applying of the context video frames. The removal block is employed to be consistent with the size of the HMM frames. Results show that a viseme recognition accuracy of 76.4% is attained while utilizing the proposed viseme features. This accuracy validates the strength of the utilized features, in spite of the usage of a shallow HMM classifier.

C. *Deep CNN Network using Appearance Viseme Features*

Deep CNNs with various layers are examined and the experiments were deployed with all speakers in the SAVE dataset. An 8-fold validation model is utilized. The database includes two different videos for each speaker. We divided the database into three folds of 24 speakers for training, 6 speakers for testing and 6 speakers as the development set. The proposed viseme features are used as an input to the CNN, and various CNN models and layers are inspected. We develop the deep CNN is as 1056 input layers, 1024 intermediate hidden layers, 2000 last hidden layers, and 20 are the output layers. This network is altered and structured parametrically in two tasks. The whole classification method is depicted in Fig. 3.

The effect of the network width is investigated in the first task, where 512, 1024 and 2048 layers are considered for the intermediate deep CNN, and 1052, 2000 and 3000 hidden units are considered for the upper CNN. The second task, we investigated the effect of the CNN depth is examined, where different layers are established from 4 to 7. The accuracies are depicted in Table I and Table II.

We should understand that the results accuracies are attained after the pre-training process through fine tuning of the deep CNN, where it is unrolled to its DNN architecture, and the decoder is used to convert the outputs to the appropriate classes. We utilized a corpus of thirty speakers pronouncing 8000 connected visemes. In this research, the experiments are performed using the speakers uttering of the corpus excluding the profile of the speakers.

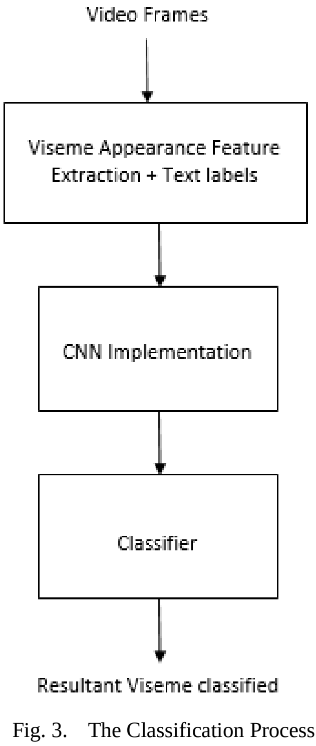

Video Frames
Viseme Appearance Feature Extraction + Text labels
CNN Implementation
Classifier
Resultant Viseme classified

Fig. 3. The Classification Process.

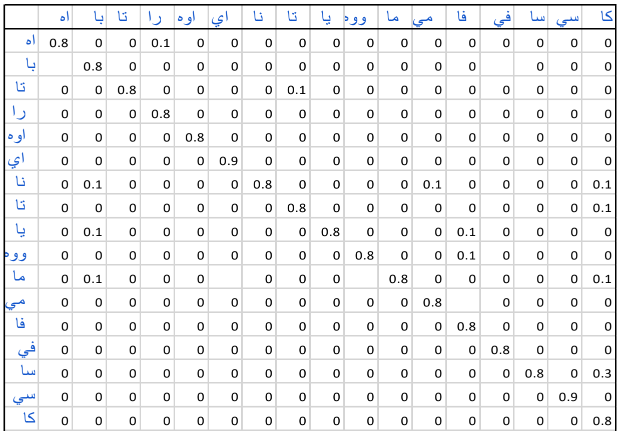
TABLE I. CLASSIFICATION ACCURACY BASED ON FEATURE SELECTION POLICY

Classification	**Viseme Feature Type**	**Classification Accuracy**
Appearance features	Our proposed technique	93.3%
	DCT	79.5%
Shape and color features	Geometric	82.5%
	Both geometric and color	86.2%

TABLE II. CLASSIFICATION ACCURACY OF DIFFERENT CNN ARCHITECTURES

CNN Architecture	**Input to the CNN**	**Layers**	**Accuracy**
CNN	Appearance features	1056 - 512 (4 layers)- 1052-20	87.3%
		1056- 1024 (4 layers)- 2000 -20	90.5%
		1056- 2048 (4 layers)- 3000-20	95.3%

Looking at the results depicted in Table I and Table II, we found that deep CNN with the following layers (1056 - 2048 (4 layers) – 3000 - 20) is the best in accuracy results with average accuracy of 95.3%. The confusion matrix of the accuracies of the Arabic visemes dataset is depicted in Fig. 4.



	اه	با	تا	را	اوه	اي	نا	تا	يا	ووه	ما	مي	فا	في	سا	سي	كا
اه	0.8	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0
با		0.8	0	0	0	0	0	0	0	0	0	0	0		0	0	0
تا	0	0	0.8	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0
را	0	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0
اوه	0	0	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0	0
اي	0	0	0	0	0	0.9	0	0	0	0	0	0	0	0	0	0	0
نا	0	0.1	0	0	0	0	0.8	0	0	0	0	0.1	0	0	0	0	0.1
تا	0	0	0	0	0	0	0	0.8	0	0	0	0	0	0	0	0	0.1
يا	0	0.1	0	0	0	0	0	0	0.8	0	0	0	0.1	0	0	0	0
ووه	0	0	0	0	0	0	0	0	0	0.8	0	0	0.1	0	0	0	0
ما	0	0.1	0	0	0		0	0	0		0.8	0	0	0	0	0	0.1
مي	0	0	0	0	0	0	0	0	0	0	0	0.8		0	0	0	0
فا	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0	0	0	0
في	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0	0	0
سا	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0	0.3
سي	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9	0
كا	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8

Fig. 4. The Confusion Matrix for the Arabic Viseme Test Data.

We developed two experiments:

Experiment 1: The Deep-learning CNN is verified with the datasets DS1 and DS2. The accuracy of the classifier is confirmed through runs of convolution layers, max-pooling and classification layers.

Experiment 2: The Deep learning CNN is verified with the same datasets. The CNN is united with transfer learning. The proposed CNN is trained with the DS1 and DS2 with transfer learning from AlexNet. The flow diagram of Experiment 2 is illustrated in Fig. 5.

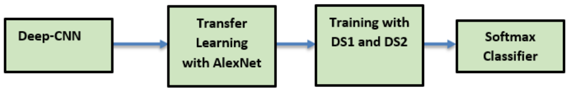


Fig. 5. Block Diagram of Experiment 2.

We performed our experiments applying our proposed model and performed comparison with recent models for lip reading in literature. The experimental results and the comparison are depicted in Table III and Table IV.

TABLE III. EXPERIMENTAL RESULTS OF OUR PROPOSED MODEL COMPARED TO OTHER LIP READING MODELS

Model	**Methodology**	**Accuracy**
Experiment 1: Our proposed model without transfer learning	Deep learning CNN	90.3%
Experiment 2: Our proposed model with transfer learning	CNN with transfer learning	96.7%
Model in [2]	Geometry based and CNN	90.06%
Model [14]	Contour extraction	93.56%
Model in [15]	Optical flow estimation	92.11%
Model in [13]	Deep learning	87.78%
Model in [11]	Active shape models	83.6%

TABLE IV. EXPERIMENTAL RESULTS OF OUR PROPOSED TECHNIQUE COMPARED TO OTHER MODELS

Model	**Precision**	**Recall**	**F-Measure**
Experiment 1: Our proposed model without transfer learning	0.95	0.945	0.94
Experiment 2: Our proposed model with transfer learning	0.98	0.975	0.97
Model in [2]	0.825	0.90	0.89
Model [14]	0.92	0.93	0.935
Model in [15]	0.89	0.90	0.896
Model in [13]	0.90	0.91	0.90
Model in [11]	0.91	0.92	0.915

IV. DISCUSSION

We devised two experiments, where the training and validation in the first experiment were done using the datasets DS1 and DS2. The accuracy was 95% on average using two hundred different runs with recall of 94.5% and with 25 different speakers. This exhibit the great independence of our model on the speaker features. In the second experiment 2, the Deep learning CNN is validated with the same datasets coupled with transfer learning using AlexNet. The accuracy was enhanced to 98% due to transfer learning, the experiments still was done for 200 runs with 25 different speakers.

V. CONCLUSION

This research proposed an approach with two phases of viseme feature extraction and deep CNN classification. In the viseme feature extraction phase, we utilize the appearance features and for the deep CNN classifier is proposed. Experiment results are achieved using the SAVE dataset. Our investigated method is validated by comparing to the HMM classifier and outperformed it in the accuracy. viseme appearance features, were utilized in our system, resulting in the base algorithm of accuracy 69.8%. It should be noted that our deep CNN classifier outperforms the CNN model presented in [7], where accuracy of 76.6% is attained in [7], in spite of using large viseme dataset. In our work, the accuracy is increased by about 21%. Our deep CNN is pre-trained and tested using 700 visemes. The results accuracies are attained after the pre-training process through fine tuning of the deep CNN, where it is unrolled to its DNN architecture, and the decoder is used to convert the outputs to the appropriate classes. We found that deep CNN with the following layers (1056 - 2048 (4 layers) – 3000 - 20) is the best in accuracy results with average accuracy of 95.3%.

ACKNOWLEDGMENT

This research project was funded by the Deanship of Scientific Research, Princess Nourah bint Abdulrahman University, through the Program of Research Project Funding After Publication, grant No. (42-PRFA-P-52).

REFERENCES

[1] K. S. Talha, K. Wan, S. K. Za'ba, Z. Mohamad Razlan and A.B Shahriman, “Speech Analysis Based On Image Information from Lip Movement,” 5th International Conference on Mechatronics (ICOM‘19), Cairo, Egypt, pp. 1-8, 2019.

[2] M. Z. Ibrahim and D. J. Mulvaney, “Geometry based lip reading system using Multi Dimension Dynamic Time Warping,” 2012 Visual Communications and Image Processing, San Diego, CA, pp. 1-6, 2020.

[3] X. Zhang, C. C. Broun, R. M. Mersereau and M. A. Clements, “Automatic Speechreading with Applications to Human-Computer Interfaces,” EURASIP Journal on Advances in Signal Processing, vol. 2002, no. 11, pp. 1228-1247, 2019.

[4] E. Skodras and N. Fakotakis, “An unconstrained method for lip detection in color images,” 2021 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Paris, France, pp. 1013-1016, 2021.

[5] L. Rothkrantz, “Lip-reading by surveillance cameras,” 2019 Smart City Symposium Prague (SCSP), Prague, Czech, pp. 1-6, 2019.

[6] P. Sujatha and M. R. Krishnan, “Lip feature extraction for visual speech recognition using Hidden Markov Model,” 2020 International Conference on Computing Communication and Applications, London, England, pp. 1-5, 2020.

[7] M. H. Rahmani and F. Almasganj, “Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features,” 2019 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), Cleveland, Ohio, pp. 195-199, 2019.

[8] N. Eveno, A. Caplier and P.Y. Coulon, “Accurate and Quasi-Automatic Lip Tracking,” IEEE Transactions On Circuits And Systems For Video Technology, vol. 14, no. 5, pp. 706-715, 2020.

[9] M. Xinjun, Y. Long and Z. Qianyuan, “Lip Feature Extraction Based on Improved Jumping-Snake Model,” Proceedings of the 35th Chinese Control Conference, Xengian, China, pp. 6928-6933, 2020.

[10] L. D. Terissi, M. Parodi and J. C. Gomez, “Lip Reading Using Wavelet-Based Features and Random Forests Classification,” 22nd International Conference on Pattern Recognition, Alexandria, GA, pp. 791-796, 2019.

[11] Q. D. Nguyen and M. Milgram, “Multi Features Active Shape Models for Lip Contours Detection,” Proceedings of the 2018 International Conference on Wavelet Analysis and Pattern Recognition, Athens, Greece, pp. 172-176, 2018.

[12] A. G. Chitu and L. J. M. Rothkrantz, “Visual Speech Recognition Automatic System for Lip Reading of Dutch,” Information Technologies and Control, vol. 7, no. 1, pp. 2-9, 2019.

[13] L. L. Mok, W. H. Laut, S.H. Leung, S.L. Wang and H. Yan, “Person Authentication Using ASM Based Lip Shape and Intensity Information,” International Conference on Image Processing (ICIP), Lafayette, USA, pp. 561-564, 2020.

[14] S. R. Chalamala, B. Gudla, B. Yegnanarayana and Sheela K Anita, “Improved Lip Contour Extraction for Visual Speech Recognition,” 2019 IEEE International Conference on Consumer Electronics (ICCE), Belin, Germany, pp. 459-462, 2019.

[15] C. Bouvier, P.Y. Coulon and X. Maldague, “Unsupervised Lips Segmentation Based on ROI Optimisation and Parametric Model,” International Conference on Image Processing (ICIP), New York, USA, pp. 301-304, 2020.

[16] S. S. Morade and S. Patnaik, “A novel lip reading algorithm by using localized ACM and HMM: Tested for digit recognition,” Optik, vol. 125, no. 1, pp. 5181-5186, 2021.

[17] A. B. A. Hassanat, “Visual passwords using automatic lip reading,” International Journal of Sciences: Basic and Applied Research (IJSBAR), vol. 13, no. 1, pp. 218-231, 2020.

[18] X. Hong, H. Yao, Y. Wan and R. Chen, “A PCA based Visual DCT Feature Extraction Method for Lip-Reading,” Proceedings of the 2019 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Napoli, Italy, pp. 1120-1127, 2019.

[19] Advanced Multimedia Processing Laboratory. Pittsburgh PA: Carnegie Mellon University, May 2018.

[20] I. Matthews, T. Cootes, J. A. Bangham, S. Cox and R. Harvey, “Extraction of visual features for lipreading,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, 2020.

[21] E. Patterson, S. Gurbuz, Z. Tufekci and J. Gowdy, “CUAVE: a new audio-visual database for multimodal human computer- interface research,” Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing, Madrid, Spain, pp. 2017-2020, 2019.

[22] S. Pathan and A. Ghotkar, “Recognition of Spoken English Phrases using Visual Features Extraction and Classification,” International Journal of Computer Science and Information Technologies, vol. 6, no. 4, pp. 3716-3719, 2020.

[23] A. Nasuha, F. Arifin, T. A. Sardjono, H. Takahashi and M. H. Purnomo, “Automatic Lip Reading for Daily Indonesian Words Based on Frame Difference and Horizontal-Vertical Image Projection,” Journal of Theoretical and Applied Information Technology, vol. 95, no. 2, pp. 393-402, 2017.

[24] M. Deypir, S. Alizadeh, T. Zoughi and R. Boostani, “Boosting a Multi-Linear Classifier with Application to Visual Lip Reading,” Expert Systems with Applications, vol. 38, no. 1, pp. 941-948, 2021.

[25] L. Lay, H.J. Yang, C.S. Lin and B.F. Lee, “Lip Language Recognition for Specific Words,” Indian Journal of Science and Technology, vol. 5, no. 11, pp. 3565-3572, 2012.

[26] L.V.S. Raghuveer and D. Deora, “Lip Localization and Visual Speech Recognition with Optical Flow in Hindi,” International Journal of Computer Sciences and Engineering (JCSE), vol. 5, no. 5, pp. 209-212, 2017.

[27] M. Z. Ibrahim and D. J. Mulvaney, “Robust Geometrical-Based Lip-Reading using Hidden Markov Models,” Eurocon 2020, Zagreb, pp. 2011-2016, 2020.

[28] A. A. Shaikh, D. K. Kumar, W. C. Yau and M. Z. Che Azemin, “cLip Reading using Optical Flow and Suppor Vector Machines,” 3rd International Congress on Image and Signal Processing, Amsterdam, Netherland, pp. 327-330, 2020.

[29] S. Sengupta, A. Bhattacharya, P. Desai and A. Gupta, “Automated Lip Reading Technique for Password Authentication,” International Journal of Applied Information Systems (IJAIS), vol. 4, no. 3, pp. 18-24, 2020.

[30] N. Rathee, “A Novel Approach for Lip Reading based on Neural Network,” International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), 2016.

[31] B. S. Lin, Y. H. Yao, C. F. Liu, C. F. Lien and B. S. Lin, “Development of Novel Lip-Reading Recognition Algorithm,” IEEE Access, vol. 5, pp. 794-801, 2017.