2020 IEEE REGION 10 CONFERENCE (TENCON)
Osaka, Japan, November 16-19, 2020

Improved Viseme Recognition using Generative Adversarial Networks

Jayanth Shreekumar
Electronics and Communication
PES University
Bengaluru, India
jayanthshreekumar@gmail.com

Ganesh K Shet
Electronics and Communication
PES University
Bengaluru, India
ganeshkshet1998@gmail.com

Vijay P N
Electronics and Communication
PES University
Bengaluru, India
vijay.pn1998@gmail.com

Preethi S J
Electronics and Communication
PES University
Bengaluru, India
preethisj@pes.edu

Niranjana Krupa
Electronics and Communication
PES University
Bengaluru, India
bnkrupa@pes.edu

***Abstract*—The proliferation of convolutional neural networks (CNN) has resulted in increased interest in the field of visual speech recognition (VSR). However, while VSR for word-level and sentence-level classification has received much of this attention, recognition of visemes has remained relatively unexplored. This paper focuses on the visemic approach for VSR as it can be used to build language-independent models. Our method employs generative adversarial networks (GANs) to create synthetic images that are used for data augmentation. VGG16 is used for classification both before and after augmentation. The results obtained prove that data augmentation using GANs is a viable technique for improving the performance of VSR models. Augmenting the dataset with images generated using the Progressive Growing Generative Adversarial Network (PGGAN) model led to an average increase in test accuracy of 3.695% across speakers. An average increase in test accuracy of 2.59% was achieved by augmenting the dataset using images generated by the conditional Deep Convolutional Generative Adversarial Network (DCGAN) model.**

***Index Terms*—Generative Adversarial Network, Conditional Generative Adversarial Network, Convolutional Neural Network, Visual Speech Recognition, Viseme Recognition**

I. Introduction

Visual speech recognition (VSR) [1], [2], popularly called lip reading, is the process of understanding speech uttered by an individual by interpreting the movement of the mouth region. It is widely used by people with hearing impairments to better perceive speech, and involuntarily by everyone in environments where audio speech recognition is hampered by noise. Phonemes are the smallest distinct parts of speech that help distinguish between words of a language. A viseme is a generic snapshot of the face that is made when a phoneme is uttered. Visual speech recognition deals with utilizing visemes to aid in speech recognition. It is a computer vision task as the machine utilizes image frames from a video to capture facial signals and emotions, which will provide it with additional information, thereby increasing recognition accuracy of the phoneme being uttered. These systems, called audio-visual speech recognition systems [3], [4], have already been implemented and tested over a wide range of tasks.

The lack of a standard database has been cited as a major bottleneck for the comparison of audio-visual recognition algorithms by Potamianos et al. [5]. Furthermore, a major deficit of any database for VSR is that visemes are unequally represented as expected of natural speech patterns. As a consequence, it becomes inherently harder to train a convolutional network to recognize lesser-used visemes. Data augmentation techniques are usually used to balance the dataset prior to training and increase its size to prevent overfitting.

Generative adversarial networks [6] are a relatively new type of neural network architecture. They consist of two competing neural networks, the generator and the discriminator, taking part in a game. The generator network creates images using a latent space of noise. The discriminator network is fed either a generated image or a real image and attempts to classify whether the image is real (a true data sample) or fake (generated by the generator). The gradients of both the networks are then used to propel them in the right direction. After extensive training, the generator becomes capable of producing images which have a distribution indistinguishable from the distribution of the images used for training the discriminator. Conditional GANs are an extension of the GAN architecture introduced by Mirza and Osindero in [7] in which they provide the label of each image to both the networks that is used as conditioning information. This can later be used to generate images belonging to a specific class.

The main objective of this paper is two-fold. Firstly, it focuses on the use of generative adversarial networks in the field of lip reading, as a novel technique for data augmentation to improve viseme recognition, and secondly, it compares the ability of two different GANs, the PGGAN, and the DCGAN, to generate images using the same training dataset. Images of two different speakers, namely, speaker 01M, a male, and speaker 11F, a female, are extracted from videos in the TCD-Timit audio-visual corpus of continuous speech [8] and used for all experiments. Data pre-processing is done to obtain the lip region of each image to create a lip image dataset. VGG16 CNN [9] is used to perform classification on this dataset to establish initial viseme classi-

978-1-7281-8455-5/20/$31.00 ©2020 IEEE

fication accuracies. Then, the conditional deep convolutional generative adversarial network (DCGAN) [10], conditioned on the viseme class labels, is trained on the dataset. The generator is used to synthesize 2400 synthetic images (200 for each class), which are used to augment the original dataset. VGG16 is again used for classification after augmentation. A similar procedure is adopted using the progressive growing generative adversarial network (PGGAN) [11]. However, the major difference is that the PGGAN was not conditional and 12 different models are used to generate images belonging to the 12 viseme classes. Finally, the results obtained are compared and discussed.

II. Related Work

Data augmentation is not a new topic and has been an area of research for many decades. Affine transformations are the most commonly used methods of data augmentation and are used to synthetically increase the size of the dataset by performing basic image manipulations such as translation, scaling, rotation, shearing etc. on the images in the original dataset. These techniques have been successful in preventing overfitting and thus increasing classification accuracies in convolutional neural networks. A naive way to balance the dataset across classes was to simply replicate the images of the minority class. The Synthetic Minority Over-sampling Technique (SMOTE), proposed by Chawla et al. [12] in 2002, incorporated a combination of over-sampling the minority classes while simultaneously under-sampling the majority classes to improve classification accuracy.

The advent of generative adversarial networks has enabled new methods of data augmentation through generative modelling. They have become popular in the field of medical imaging where datasets are small in size. Wu et al. [13] proposed the use of a conditional infilling GAN to produce high-resolution synthetic mammogram patches. Frid–Adar et al. [14] used DCGAN to generate synthetic images of liver lesions. Their dataset was very limited and consisted of 182 computed tomography images. They recorded an improved sensitivity of 85.7% and specificity of 92.4% on the dataset augmented using GANs as opposed to a sensitivity of 78.6% and a specificity of 88.4% on the dataset augmented using classic data augmentation techniques. Chuquicusma et al. [15] proposed the use of DCGANs to generate images of lung nodules. Their dataset consisted of 1018 lung cancer screening thoracic computed tomography (CT) scans. Two radiologists took part in 18 visual Turing tests to check whether the generated images were of high quality and were good enough to pass as real samples. They recorded that their generated images managed to fool radiologist 1 in 67% of their experiments while they managed to fool radiologist 2 in all experiments.

In the field of lip reading, Oliveira et al. [16] utilize the Pix2Pix GAN to map between views of the mouth taken from random angles to the frontal view. Their generator receives an image taken from a random angle and generates the corresponding image as seen from the frontal view. The discriminator is trained on pairs of images generated by the generator (that consist of a real image and its corresponding generated frontal view) and on original pairs, so that it learns to classify between them. They then performed data augmentation on the original dataset by concatenating real and generated views and fed it to Xception architecture [17] for classification. This method of utilizing multiple views for viseme classification resulted in an average increase in accuracy of 5.9% on the GRID corpus [18].

III. Method

The methodology followed in this paper has three major parts: data pre-processing, image generation using GANs, and finally, training the VGG16 convolutional neural network for viseme classification. A block diagram of the implementation is given in Fig. 1. A Lip image dataset is created by performing lip detection and extraction on each frame of the videos in the TCD-Timit database belonging to two different speakers. Classification is done on the original dataset. Next, both GANs are used to generate synthetic images and DCGAN and PGGAN-based data augmentation is carried out to obtain two augmented datasets. Finally, classification is performed on these augmented datasets and the results are compared.

A. The Dataset

The TCD-Timit audio-visual corpus of continuous speech [8] is a publicly available video database of 62 speakers uttering a total of 6913 phonetically diverse sentences, of which 3 are professionally trained lip speakers. Videos were recorded at a resolution of 1920×1080-pixel frames at 30 fps from two different angles – frontal (0 degrees) and 30 degrees. All experiments conducted in this paper were performed on images extracted from a total of 196 videos (98 videos each) of 2 different speakers, 01M, a male, and 11F, a female.

B. Dataset Pre-processing

Pre-processing was done on videos taken from the frontal view of the 2 speakers. Every frame in each video was extracted and stored in the corresponding phoneme folder with the help of the phoneme transcription label files provided with the TCD-Timit database. There were a total of 38 phonemes including the silence /sil/ phoneme. However, the /sil/ phoneme was not included in the experiments. as it would lead to further ambiguity among /m/, /b/, /p/ and /sil/ phonemes. The Dlib toolkit [19] was used to code the lip detection and extraction algorithms in Python [20], which

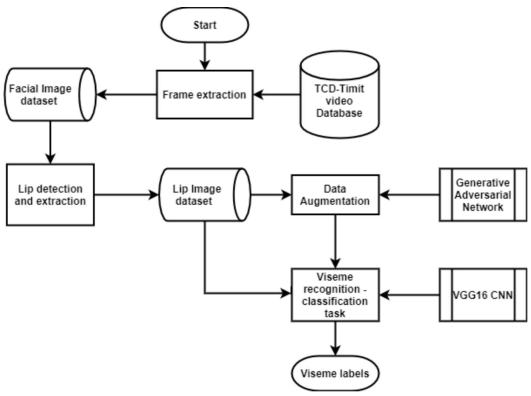


Fig. 1. Our proposed methodology

were used to extract only the lip region of each image to obtain a new image dataset consisting of the lip region for each speaker. All lip images were further resized to 256×256 dimensions.

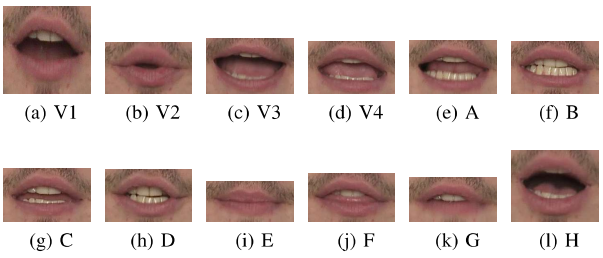
Finally, Neti's phoneme-viseme mapping [3], shown in Table I, was used to create viseme class folders by merging similar phoneme class folders. The facial configuration made by speaker 01M for each of the 12 viseme classes in the dataset is shown in Fig. 2. There were a total of 7845 frames that were extracted from the videos of speaker 01M and a total of 8710 frames that were extracted from the videos of speaker 11F. Note that the datasets were highly skewed. For example, viseme class V1 of speaker 01M had more than 1200 images while the viseme class F had just 95. This skewness was preserved as it is an important feature of real-world samples.

C. Deep Convolutional Generative Adversarial Network

Deep Convolutional Generative Adversarial Networks [10] were proposed by Radford et al. in an attempt to generate high resolution images using deeper generative models. The DCGAN architecture used in this work was slightly modified as it did not have batch normalization layers [21] in the generator and used Leaky ReLU [22] in the generator along with Gaussian noise layers. Another important note here is that image generation was done using an unconditional DCGAN at first, where 12 different DCGANs were used to generate images corresponding to the 12 viseme classes. This did not yield good results as the DCGANs that were trained on minority classes (having few training samples) went into mode collapse, where they generated a small set of images repeatedly instead of learning new features. Conditional GANs have been shown to alleviate this problem [23]. Therefore, a single conditional DCGAN, conditioned on viseme class labels was used for training. This model was later used to generate images belonging to a specific viseme class as required using these labels. This architecture was used to generate images of dimensions $256 \times 256 \times 3$.

D. Progressive Growing Generative Adversarial Network

The PGGAN architecture [11] proposed by Karras et al. described an innovative training methodology for GANs in which they proposed to train both the generator and the discriminator progressively, starting by training it on images of low resolution and gradually increasing the resolution to later learn the finer details. The PGGAN was implemented as an alternative to the DCGAN as it was expected to generate images having better quality. However, the major difference is that the PGGAN was not conditioned on class labels, as the unconditional PGGAN was capable of generating images without suffering mode collapse. Therefore, 12 PGGAN models were implemented to generate images belonging to their respective viseme classes.



(a) V1 (b) V2 (c) V3 (d) V4 (e) A (f) B

(g) C (h) D (i) E (j) F (k) G (l) H

Fig. 2. Sample frames belonging to each of the 12 viseme classes of speaker 01M before resizing.

E. Convolutional Neural Networks

The VGG16 convolutional neural network, pre-trained on the ImageNet [24] database was trained to classify the images into different viseme classes. Classification was performed both before and after data augmentation for comparison. Additionally, classification was also done solely on the generated images to understand how well each GAN had captured the features of the 12 viseme classes.

IV. Experiments and Results

The experiments conducted in this paper were all speaker dependent and fall into two major categories: training the GANs to generate synthetic images, and training and testing the VGG16 CNN to classify visemes. Firstly, the CNN was trained and tested on the dataset consisting of only original images. Then the original dataset was used to train the two different GANs to generate images of size $256 \times 256 \times 3$ belonging to each viseme class. Finally, the original dataset was augmented with these synthetic images and VGG16 was once again trained and tested on these augmented datasets. All of this was done on images belonging to two different speakers, 01M and 11F.

A. Synthetic Image Generation using Conditional DCGAN

The input to the discriminator of the conditional DCGAN was the resized lip images having dimensions $256 \times 256 \times 3$, conditioned on the class labels that were one-hot encoded vectors. Adam optimizer [25] was used for both networks. The learning rate for the discriminator was 4×10^{-3} and the learning rate for the generator was 1×10^{-3}. Advanced training techniques such as label smoothing, noisy labels, input normalization and Gaussian weight initialization were implemented. A sample of the images generated by the conditional DCGAN is shown in Fig. 3. A total of 2400 synthetic images, 200 per viseme class, were generated for augmentation. It was observed that the generated images improved in quality and showed more diversity for the first 60 epochs. After this, the quality of images did not improve and GAN collapse occurred.

TABLE I
Neti's Phoneme-Viseme Mapping

Viseme	Phonemes
V1	/ao/, /ah/, /aa/, /er/, /oy/, /aw/, /hh/
V2	/uw/, /uh/, /ow/
V3	/ae/, /eh/, /ey/, /ay/
V4	/ih/, /iy/, /ax/
A	/l/, /el/, /r/, /y/
B	/s/, /z/
C	/t/, /d/, /n/, /en/
D	/sh/, /zh/, /ch/, /jh/
E	/p/, /b/, /m/
F	/th/, /dh/
G	/f/, /v/
H	/ng/, /k/, /g/, /w/

B. Synthetic Image Generation using PGGAN

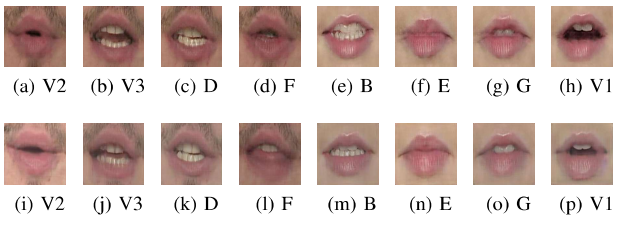
The dataset that was used to train the conditional DCGAN was also used to train the PGGAN. However, unlike the training of the former, 12 different PGGAN models were implemented, corresponding to each of the 12 different viseme classes to generate images. Therefore, while the total size of the dataset used for training was the same, smaller parts of the whole dataset, corresponding to different visemes, were used to train the 12 PGGAN models. Again, Adam optimizer was used for both networks. A learning rate of 2.8×10^{-6} was used for the discriminator and a learning rate of 2.5×10^{-6} was used for the generator. Advanced training techniques such as label smoothing, noisy labels, input normalization and Gaussian weight initialization were also implemented. Similar to the DCGAN, a total of 2400 images were generated using the 12 PGGAN models, 200 per model. A sample of the images generated by the PGGAN is shown in Fig. 3.

C. CNN Viseme Classification

Classification using VGG16 was performed on the original lip datasets of the two speakers. A train-validation-test split of 80%–10%–10% was used. Transfer learning technique and Adam optimizer were used throughout the experiments with a learning rate of 1×10^{-5} along with two dropout [26] layers having rates of 0.3 and 0.7. Additionally, batch normalization [21] layers were used and an early stopping mechanism was implemented that stopped training when the validation loss did not decrease for 10 consecutive epochs. Initial accuracies obtained on the original datasets of each speaker before data augmentation are displayed in Table II.

To get an idea of how well the DCGAN had captured the features of each class, a dataset consisting of only the 2400 generated images was fed to VGG16 for classification. Similarly, a dataset consisting of only 2400 images generated using the PGGAN was also classified. The results clearly show that both GANs have captured the features quite well and are summarized in Table II. However, the PGGAN performs better as the CNN was able to classify all fake test images correctly.

Detailed precision (P), recall (R) and F1 scores (F1) obtained for the experiments on speaker 01M visemes (V) before data augmentation are shown in Table III. As expected, classification of images belonging to Viseme F (which had just 95 frames in the original dataset for 01M) was extremely poor. The viseme class B had over 1200 images and hence, has very high performance metrics. This is also due to the fact that viseme B has a very distinct configuration of the mouth region, and is easily distinguished from other visemes.



(a) V2 (b) V3 (c) D (d) F (e) B (f) E (g) G (h) V1

(i) V2 (j) V3 (k) D (l) F (m) B (n) E (o) G (p) V1

Fig. 3. Sample images generated using the conditional DCGAN models: 3a, 3b, 3c and 3d are generated using the 01M model and 3e, 3f, 3g and 3h are generated using the 11F model. Sample images generated using the PGGAN models: 3i, 3j, 3k and 3l are generated using 4 of the 12 01M models while 3m, 3n, 3o and 3p are generated using 4 of the 12 11F models.

TABLE II
ACCURACIES (%) OBTAINED FOR CLASSIFICATION OF ORIGINAL AND GENERATED IMAGES USING VGG16

Speaker	Original		DCGAN		PGGAN	
	Val	*Test*	*Val*	*Test*	*Val*	*Test*
01M	51.13	50.70	85.65	87.08	100.00	100.00
11F	50.32	49.65	92.59	92.08	100.00	100.00

An interesting detail to note here is that, even though the number of images belonging to viseme classes D, E, and G was lesser than 500 each, all of them have reasonably good performance metrics. This can again be attributed to the fact that all of them have very unique mouth configurations, which leads to easier classification. Although the viseme classes V1, A, and C had around a thousand images each (V1 had 1347), they do not have high metrics. This anomaly might be due to the nature of the visemes, and can be explained by looking at the phoneme-viseme mapping as shown in Table I. Notice that the lip and teeth positions for all the three classes are very similar. The only feature that changes among them is the position of the tongue, and as the tongue is not easily visible, classification amongst them becomes harder, leading to errors. The performance metrics for speaker 11F before augmentation are very similar to those of 01M and are shown in Table IV. Clearly, the extra male features, the moustache and the beard, do not help in improving classification accuracy.

Next, images generated using the conditional DCGAN were used to augment the training dataset. The new training dataset consisted of original images as well as the generated images while the test dataset consisted of the original images only. Classification was once again performed on this modified dataset and and an average test accuracy of 52.765% across speakers was recorded, which is an increase of about 2.59% compared to the original accuracy.

The procedure described above was used to create a separate, modified dataset that consisted of the original dataset along with the images generated by the PGGAN. An average test accuracy of 53.87% across speakers was

TABLE III
PERFORMANCE METRICS OBTAINED FOR CLASSIFICATION OF ORIGINAL AND GENERATED IMAGES OF SPEAKER 01M VISEMES ROUNDED TO 2 DECIMAL DIGITS

V	Original			DCGAN			PGGAN		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
V1	0.53	0.50	0.51	0.73	0.61	0.67	1.0	1.0	1.0
V2	0.33	0.45	0.39	1.0	1.0	1.0	1.0	1.0	1.0
V3	0.49	0.62	0.55	1.0	0.96	0.98	1.0	1.0	1.0
V4	0.46	0.48	0.47	0.92	0.96	0.94	1.0	1.0	1.0
A	0.39	0.34	0.37	0.53	0.45	0.49	1.0	1.0	1.0
B	0.70	0.80	0.75	0.92	1.0	0.96	1.0	1.0	1.0
C	0.47	0.37	0.42	1.0	0.92	0.96	1.0	1.0	1.0
D	0.55	0.57	0.56	0.94	0.94	0.94	1.0	1.0	1.0
E	0.50	0.83	0.62	1.0	1.0	1.0	1.0	1.0	1.0
F	0.00	0.00	0.00	0.95	0.91	0.93	1.0	1.0	1.0
G	0.35	0.76	0.48	1.0	1.0	1.0	1.0	1.0	1.0
H	0.49	0.24	0.32	0.52	0.68	0.59	1.0	1.0	1.0

TABLE IV
PERFORMANCE METRICS OBTAINED FOR CLASSIFICATION OF ORIGINAL AND GENERATED IMAGES OF SPEAKER 11F VISEMES ROUNDED TO 2 DECIMAL DIGITS

V	Original			DCGAN			PGGAN		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
V1	0.58	0.40	0.48	0.87	0.82	0.85	1.0	1.0	1.0
V2	0.34	0.60	0.44	0.95	1.0	0.98	1.0	1.0	1.0
V3	0.53	0.64	0.58	1.0	0.93	0.96	1.0	1.0	1.0
V4	0.40	0.47	0.43	0.71	1.0	0.83	1.0	1.0	1.0
A	0.25	0.16	0.2	0.81	0.81	0.81	1.0	1.0	1.0
B	0.55	0.59	0.57	0.9	0.9	0.9	1.0	1.0	1.0
C	0.51	0.49	0.5	0.89	0.89	0.89	1.0	1.0	1.0
D	0.69	0.87	0.77	1.0	1.0	1.0	1.0	1.0	1.0
E	0.45	0.82	0.58	1.0	1.0	1.0	1.0	1.0	1.0
F	0.00	0.00	0.00	1.0	1.0	1.0	1.0	1.0	1.0
G	0.64	0.78	0.7	1.0	1.0	1.0	1.0	1.0	1.0
H	0.41	0.27	0.32	0.95	0.76	0.84	1.0	1.0	1.0

recorded, which is 3.695% greater than the original accuracy. This clearly shows that the PGGAN, on the whole, has created images of better quality. The accuracies obtained after data augmentation are summarized in Table V.

Detailed precision (P), recall (R) and F1 scores (F1) obtained for the experiments after data augmentation are shown in Tables VI and VII. The metrics obtained on the generated datasets in Table III seem to indicate that both GANs have been able to do considerably well in generating images belonging to viseme F. However, upon closer inspection of the performance metrics that were obtained after data augmentation, it is clear that the PGGAN has not been able to successfully capture the features of viseme F for speaker 01M as the metrics have not improved. The DCGAN-based data augmentation method, although not significant, has led to an improvement. In contrast, the PGGAN-based data augmentation has improved the performance metrics of viseme F for 11F, but the DCGAN approach has failed to do so.

As expected, the performance metrics for viseme classes that had around 1000 or more original images (V1, A, B, and C) generally improve after data augmentation, the sole exception being that of viseme V1 of speaker 01M. This is because, the GANs are able to extract relevant features and generate realistic images using the relatively greater number of images, which in turn leads to improved performance after augmentation. Another trend that is noticed is that the performance metrics for visemes that have very distinct mouth configurations, namely V2, V3, D, E,and G, generally improve after data augmentation. This is because the GANs easily capture the unique features, and reproduce images faithfully.

Our work achieves a maximum improved accuracy of 53.87% using only the VGG16 CNN trained on the PG-GAN augmented dataset. Other works performed on the TCD-Timit database that are similar to our experiments include [27] in which the performances of different DNN networks were compared. They report a best accuracy of 66.27%, achieved using a hybrid CNN-BiLSTM network. A five-fold mean accuracy high of 69.58% was achieved in [28] for viseme classification on speaker independent tests in which they used the ResNet [29] CNN-SVM system. Zimmermann [30] used a custom CNN network and achieved a best accuracy of 44.11% for speaker independent experiments.

TABLE V
ACCURACIES (%) OBTAINED FOR CLASSIFICATION OF IMAGES AFTER DATA AUGMENTATION USING VGG16

Speaker	DCGAN		PGGAN	
	Validation	*Test*	*Validation*	*Test*
01M	52.23	53.12	55.03	53.50
11F	51.89	52.41	53.79	54.24

TABLE VI
PERFORMANCE METRICS OBTAINED FOR CLASSIFICATION OF SPEAKER 01M IMAGES AFTER DATA AUGMENTATION

V	DCGAN			PGGAN		
	P	*R*	*F1*	*P*	*R*	*F1*
V1	0.515	0.413	0.458	0.439	0.535	0.482
V2	0.522	0.387	0.444	0.542	0.371	0.441
V3	0.486	0.667	0.562	0.507	0.617	0.556
V4	0.36	0.34	0.35	0.5	0.377	0.43
A	0.435	0.42	0.428	0.515	0.495	0.505
B	0.715	0.79	0.751	0.656	0.879	0.751
C	0.451	0.525	0.485	0.471	0.378	0.419
D	0.762	0.593	0.667	0.7	0.7	0.7
E	0.594	0.837	0.695	0.694	0.782	0.735
F	0.25	0.167	0.2	0.0	0.0	0.0
G	0.63	0.68	0.654	0.658	0.806	0.725
H	0.481	0.312	0.379	0.355	0.19	0.247

V. CONCLUSION

The aim of the experiments conducted in this paper was to improve viseme recognition using CNN models by implementing a novel data augmentation technique using generative adversarial networks. The results obtained indicate that GAN-based data augmentation is a viable technique to improve viseme recognition accuracy. An average test accuracy of 52.765% across speakers was recorded when images generated by a conditional DCGAN were used in the data augmentation process, which is an improvement of 2.59% over the baseline accuracy. An average test accuracy of 53.87% across speakers was recorded when images generated by a PGGAN were used for data augmentation, which is an improvement of 3.695% over the baseline accuracy. Therefore, the results also indicate that out of the two,

TABLE VII
PERFORMANCE METRICS OBTAINED FOR CLASSIFICATION OF SPEAKER 11F IMAGES AFTER DATA AUGMENTATION

V	DCGAN			PGGAN		
	P	*R*	*F1*	*P*	*R*	*F1*
V1	0.519	0.473	0.495	0.597	0.512	0.551
V2	0.545	0.419	0.474	0.533	0.593	0.561
V3	0.515	0.639	0.57	0.542	0.755	0.631
V4	0.469	0.435	0.451	0.529	0.474	0.5
A	0.466	0.333	0.388	0.436	0.286	0.345
B	0.682	0.529	0.596	0.545	0.609	0.575
C	0.462	0.671	0.547	0.531	0.613	0.569
D	0.63	0.586	0.607	0.7	0.609	0.651
E	0.644	0.691	0.667	0.486	0.686	0.569
F	0.0	0.0	0.0	1.0	0.167	0.286
G	0.647	0.815	0.721	0.8	0.69	0.741
H	0.476	0.294	0.364	0.432	0.275	0.336

the PGGAN is capable of generating images having better quality.

For further research involving the use of GANs for data augmentation in lip reading, different types of GAN architectures can be explored that can yield more quality and diversity in generated images. Another natural follow-up is to generate images that are related temporally, that is, to use GANs to generate a sequence of coherent images that make up a viseme.

REFERENCES

[1] B. Shillingford, Y. M. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett, M. Mulville, B. Coppin, B. Laurie, A. W. Senior, and N. de Freitas, "Large-scale visual speech recognition," *CoRR*, vol. abs/1807.05162, 2018. [Online]. Available: http://arxiv.org/abs/1807.05162

[2] G. J. Wolff, K. V. Prasad, D. G. Stork, and M. Hennecke, "Lipreading by neural networks: Visual preprocessing, learning, and sensory integration," in *Advances in neural information processing systems*, 1994, pp. 1027–1034.

[3] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, and A. Mashari, "Audio visual speech recognition," IDIAP, Tech. Rep., 2000.

[4] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE transactions on multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[5] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[7] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014. [Online]. Available: http://arxiv.org/abs/1411.1784

[8] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[11] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[13] E. Wu, K. Wu, D. Cox, and W. Lotter, "Conditional infilling gans for data augmentation in mammogram classification," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Springer, 2018, pp. 98–106.

[14] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *CoRR*, vol. abs/1803.01229, 2018. [Online]. Available: http://arxiv.org/abs/1803.01229

[15] M. J. Chuquicusma, S. Hussein, J. Burt, and U. Bagci, "How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 240–244.

[16] D. A. B. Oliveira, A. B. Mattos, and E. da Silva Morais, "Improving viseme recognition using gan-based frontal view mapping," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2018, pp. 2229–22 297.

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[18] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[19] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[20] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[22] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3.

[23] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Systems with Applications*, vol. 91, pp. 464 – 471, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417306346

[24] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[27] G. Sterpu, C. Saam, and N. Harte, "Can dnns learn to lipread full sentences?" in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 16–20.

[28] J. Burton, D. Frank, M. Saleh, N. Navab, and H. L. Bear, "The speaker-independent lipreading play-off; a survey of lipreading machines," *CoRR*, vol. abs/1810.10597, 2018. [Online]. Available: http://arxiv.org/abs/1810.10597

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[30] M. Zimmermann, "Visual speech recognition: from traditional to deep learning frameworks," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, 2018.