

Harnessing GANs for Zero-Shot Learning of New Classes in Visual Speech Recognition

Yaman Kumar*
Adobe
ykumar@adobe.com

Dhruva Sahrawat*
NUS, Singapore
dhruva@comp.nus.edu.sg

Shubham Maheshwari*
MIDAS Lab, IIT-Delhi
shubham14101@iiitd.ac.in

Debanjan Mahata
Bloomberg, USA
dmahata@bloomberg.net

Amanda Stent
Bloomberg, USA
astent@bloomberg.net

Yifang Yin
NUS, Singapore
idsyin@nus.edu.sg

Rajiv Ratn Shah
MIDAS Lab, IIT-Delhi
rajivrtn@iiitd.ac.in

Roger Zimmermann
NUS, Singapore
rogerz@comp.nus.edu.sg

Abstract

Visual Speech Recognition (VSR) is the process of recognizing or interpreting speech by watching the lip movements of the speaker. Recent machine learning based approaches model VSR as a classification problem; however, the scarcity of training data leads to error-prone systems with very low accuracies in predicting unseen classes. To solve this problem, we present a novel approach to zero-shot learning by generating new classes using Generative Adversarial Networks (GANs), and show how the addition of unseen class samples increases the accuracy of a VSR system by a significant margin of 27% and allows it to handle speaker-independent out-of-vocabulary phrases. We also show that our models are language agnostic and therefore capable of seamlessly generating, using English training data, videos for a new language (Hindi). To the best of our knowledge, this is the first work to show empirical evidence of the use of GANs for generating training samples of unseen classes in the domain of VSR, hence facilitating zero-shot learning. We make the added videos for new classes publicly available along with our code¹.

1 Introduction

In Visual Speech Recognition (VSR), also known as automated lip reading, a system has to recognize words spoken by a human in a silent video, by primarily focusing on the region of the speaker’s lips. VSR has a wide variety of potential applications such as in cybersecurity, and as assistive/augmentative technology (Schmidt et al. 2013). For example, if a person has a laryngectomy or voice-box cancer, dysarthria, or works in very loud environments such as a factory floor, VSR can be enabling.

VSR is treated as a classification problem. The training data can have single or multiple views of different speakers, paired with utterance labels (e.g. digits, phrases or sentences). Given a video displaying lip movements of a speaker, the model outputs a probability distribution over the utterance

labels. The final prediction is the utterance label with the highest probability score. The performance of humans in lip reading is poor, with the best lip readers achieving an accuracy of just around 21% (Easton and Basala 1982), so VSR is a challenging task.

Modeling VSR as a classification problem limits the predictions to a fixed number of utterances. This significantly hinders model generalization on predicting *out-of-vocabulary* (OOV) classes, i.e. new utterances, new poses or new speakers. Additionally, preparing datasets for training VSR models is a challenging task by itself as it requires collecting hours of video recordings of utterances from multiple speakers and multiple poses. Recently, significant effort has gone into collecting high-quality VSR datasets, yet VSR is still both more complex than and has far less available training data than its close relative, “classic” audio speech recognition. Therefore in functionality VSR is basically where audio speech recognition was in the late 1980s - isolated words/phrases, limited speakers.

Due to the unique aspects of VSR as a task, the challenges associated with it and its potential applications, it forms an appropriate down-stream task for low-shot OOV problem settings. The challenge of automatically learning new classes and to generalize from very few labeled examples is often called *low-shot* (or *few-shot*) learning (Wang et al. 2018); or *zero-shot* learning (Socher et al. 2013), in situations where there are no examples of certain classes at all. Two main techniques applied to solve such machine learning problems are: (i) *data driven approaches* (Krizhevsky, Sutskever, and Hinton 2012), which rely on adding more data either by collecting data from external sources or by augmenting existing datasets by synthesizing similar data points, and (ii) *parameter driven approaches* (Elhoseiny et al. 2017), which rely on regularization techniques in order to constrain the number of parameters to be learnt by the model.

In 2014, Goodfellow et al. presented the idea of *Generative Adversarial Networks* (GANs) (2014). Since then, there has been significant research on GANs and their applications (Creswell et al. 2018). In their seminal work, Goodfellow et al. mention data augmentation as one of the applications of GANs, and indeed there is a growing literature focusing on it (Antoniou, Storkey, and Edwards 2017;

*These authors *contributed equally*. Author order was determined by a coin flip.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹The code as well as the supplementary file can be found at <https://github.com/midas-research/DECA/>. The videos can be accessed through this YouTube playlist link: <https://bit.ly/336geft>.

Wang et al. 2018). However, to date, the primary use of GANs for image/video data augmentation has been as a proxy to traditional approaches such as rotation, reflection, cropping, translation, scaling and adding Gaussian noise (Krizhevsky, Sutskever, and Hinton 2012), and augmenting training data for a classification problem by generating similar data items (Antoniou, Storkey, and Edwards 2017). Recently, GANs have been used to generate unseen images from text descriptions (Zhu et al. 2018).

However, to the best of our knowledge, no prior work uses GANs to augment VSR datasets with data items representing *new classes* and performing zero-shot learning.

In this paper, we describe how we generate lip movement videos of *unseen* utterances using *Temporal Conditional Generative Adversarial Networks* (TC-GANs) (Vougioukas, Petridis, and Pantic 2018) and a *viseme-concatenation* approach (Huang, Cosatto, and Graf 2002) (Section 5). As illustrated in Figure 1, we use the generated data for training a deep learning model for VSR. We empirically show the usefulness of the augmented data in increasing the accuracy of the VSR model while predicting unseen utterances. Additionally, we also show that our solution does not only cater to unseen utterances of the same language but also works well for a new language.

Our Contributions - The key contributions we make are summarized as follows:

- We introduce zero-shot learning using GANs for augmenting VSR datasets with data for unseen utterances.
- We show that our solution results in improved accuracies of up to 27% in prediction of unseen utterances at test time.
- We show results for the cold-start problem (no training data at all for training the VSR model) on lipreading for the first-time in the literature (Amplayo et al. 2018).
- We further show that this method can work for zero-shot learning for an unseen language, Hindi.
- We expand the OuluVS2 dataset (Anina et al. 2015) by adding videos for 10 new Hindi phrases and 25 new English phrases generated using our model². We plan to release this dataset as well as the code for generating it. But due to space constraints, we only conduct experiments on 3 new Hindi phrases and the 10 phrases present in OuluVS2.

2 Related Work

2.1 Visual Speech Recognition

Most early approaches to VSR such as (Ngiam et al. 2011) treat it as a single-word classification task. Other researchers such as (Goldschen, Garcia, and Petajan 1997) use methods borrowed from acoustic speech recognition like HMMs. However, most work has focused on predicting utterances from limited lexicons. Furthermore, these techniques heavily rely upon hand-engineered feature pipelines.

With advancements in deep learning and the increasing availability of large lip reading datasets, there have been approaches such as (Wand, Koutník, and Schmidhuber 2016;

Zhou et al. 2019; Salik et al. 2019) addressing lip reading using deep learning based algorithms such as CNNs and LSTMs. Further, there have been works such as (Lee, Lee, and Kim 2016; Kumar et al. 2018b; Uttam et al. 2019; Kumar et al. 2019) extending lip reading from single view to multi-view settings by incorporating videos of mouth sections from multiple views together. Multi-view lip reading has been shown to improve performance significantly as compared to single view. Although the deep learning-based methods don't require hand-engineered features, most still treat lip reading as a classification task.

2.2 Few- and Zero-Shot Learning

In few-shot learning, one learns a model that can generalize to classes that have few examples in the training set (Lake, Salakhutdinov, and Tenenbaum 2015). The few-shot learning problem has been studied from different perspectives such as optimization (Ravi and Larochelle 2017), metric learning (Vinyals et al. 2016), similarity-matching (Koch, Zemel, and Salakhutdinov 2015), and hierarchical graphical models (Salakhutdinov, Tenenbaum, and Torralba 2012). However, generalizing a classifier on *unseen* classes is mostly uninvestigated. We show that GANs can generate data for classes unseen during training, and that this generated data can give improved accuracy for the downstream VSR task.

This problem falls under the meta-learning umbrella (Thrun 1998; Shrivastava et al. 2019). High-level visual concepts like camera-pose or brightness, or even higher-level concepts like articulation (in the case of VSR) are shared across computer vision problems. Contemporary meta-learners, however, largely ignore these cross-task structures (Wang et al. 2018). Using augmentation strategies like cropping and shifting, researchers have tried to make meta-learners learn some of these high-level visual concepts. However, high-level concepts like articulation require a different type of augmentation than simply changing brightness or camera angle; one needs a way to break down the high-level concept into decomposable elements (Zhu et al. 2018). In the case of (Zhu et al. 2018), the decomposable elements are keywords. In the case of VSR, the decomposable basic elements of articulation are *visemes*, the visual equivalents of phonemes.

Several research studies have shown that by knowing the probability distribution on transformations, one can generate examples from an unseen class by applying those transformations (Hariharan and Girshick 2017; Wang et al. 2018; Shah and Zimmermann 2017). Some authors (Hariharan and Girshick 2017), try to first learn transformations from some examples of a known class and then apply those transformations to some *seed* examples of a new class. This approach increases the data size of the novel class and hence helps in boosting accuracy. In our case, the transformations are the conditions which we impose on temporal-conditional GANs to make them generate new visemes. We present these conditions and transformations in Section 5.

3 Dataset

We use the OuluVS2 dataset (Anina et al. 2015), a multi-view audio visual dataset. OuluVS2 has been used in several

²More details are present in supplementary materials

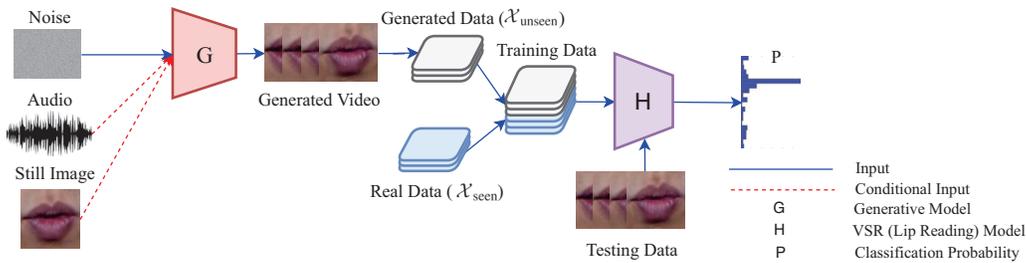


Figure 1: Overall pipeline demonstrated with the generator G as the TC-GAN architecture

previous VSR studies (Chung et al. 2017; Kumar et al. 2018b; 2018a). It contains data from 52 speakers. Each speaker is filmed speaking three types of utterance:

- *Oulu-digits* - digit sequences like $1, 9, 0, 7, 8, 8, 0, 3, 2, 8$.
- *Oulu-phrases* - ten short common English phrases such as “Excuse me” and “see you”.
- *Oulu-sentences* - randomly chosen TIMIT sentences (Garofolo 1993), such as “Basketball can be an entertaining sport” and “Are holiday aprons available to us?”.

Speakers are filmed from five different poses: $\{0^\circ, 30^\circ, 45^\circ, 60^\circ, 90^\circ\}$.

We describe our training, testing, and validation splits in Section 6, along with the relevant experiments.

4 Problem Formulation

In this section, we describe the data-driven approach we followed in this paper.

In general, a supervised classifier can be depicted as learning a conditional distribution $p(y_i|x_i)$, s.t. $(x_i, y_i) \in \mathcal{S}_{\text{train}}$, the set of training samples. The set $\mathcal{S}_{\text{test}}$ is used for testing the model. The set of all the labels present in $\mathcal{S}_{\text{train}}$ is given by $\mathcal{Y}_{\text{train}}$ (also referred as $\mathcal{Y}_{\text{seen}}$ later in the text), and $\mathcal{Y}_{\text{test}}$ gives the set of all the labels present in $\mathcal{S}_{\text{test}}$. Similarly, $\mathcal{X}_{\text{train}}$ represents the set of all the inputs present in $\mathcal{S}_{\text{train}}$ (also referred as $\mathcal{X}_{\text{seen}}$ later), and $\mathcal{X}_{\text{test}}$ denotes those in $\mathcal{S}_{\text{test}}$.

In our problem setting, unseen classes are present in the test dataset i.e., $\mathcal{Y}_{\text{test}} - \mathcal{Y}_{\text{train}} \neq \phi$. Let $\mathcal{Y}_{\text{test}} - \mathcal{Y}_{\text{train}} = \mathcal{Y}_{\text{unseen}}$. Let the inputs corresponding to $\mathcal{Y}_{\text{unseen}}$ be denoted as $\mathcal{X}_{\text{unseen}}$. Thus, $\mathcal{Y}_{\text{unseen}}$ represents all the *unseen classes*, the classes which the model does not get to see while training.

Our main objective in this work is to enable the model to correctly predict the labels present in $\mathcal{Y}_{\text{unseen}}$, given that $\mathcal{Y}_{\text{unseen}}$ is available but no $x_i \in \mathcal{X}_{\text{unseen}}$, is available for training. In other words, we aim to increase accuracy on $\mathcal{X}_{\text{unseen}}$ using the knowledge available at train time ($\mathcal{S}_{\text{train}}, \mathcal{Y}_{\text{unseen}}$), while not decreasing accuracy on $\mathcal{X}_{\text{train}}$.

To solve this problem, we “hallucinate” $\mathcal{X}_{\text{unseen}}$ using $\mathcal{S}_{\text{train}}$ and the labels present in $\mathcal{Y}_{\text{unseen}}$. We do this by learning a function $f: \mathcal{Y} \rightarrow \mathcal{X}$. For approximating $\mathcal{X}_{\text{unseen}}$ with f using machine learning, we only have $\mathcal{S}_{\text{train}}$ which, notably does not have any information about $\mathcal{X}_{\text{unseen}}$. For solving this problem, we introduce an intermediate representation \mathcal{Z} such that $g: \mathcal{Y} \rightarrow \mathcal{Z}$ and $h: \mathcal{Z} \rightarrow \mathcal{X}$.

In the case of lipreading, \mathcal{X} represents videos of speakers and \mathcal{Y} represents the utterance labels. \mathcal{Z} represents the *audio* representation of the utterances. The function g takes an utterance label and converts it into its corresponding audio. The function h is approximated in this work by a TC-GAN model or a viseme-concatenation approach. The function h takes in audio as input and produces a corresponding video as output. The input-output relation between audio and video is not one-to-one since videos of different speakers in different poses have the same audio. In order to make this mapping one-to-one, we introduce speaker and pose information \mathcal{I} ; h becomes $h: \mathcal{Z} \times \mathcal{I} \rightarrow \mathcal{X}$. The pair of (audio, speaker image in the specified pose angle) uniquely identifies a video.

The overall data hallucination pipeline is as follows:

- The utterances present in the set $\mathcal{Y}_{\text{unseen}}$ are projected onto the audio space \mathcal{Z} (Sections 5 and 6).
- A speaker’s image chosen randomly from the training subset of *Oulu-phrases* and the audio projections of utterance z_i (where $z_i \in \mathcal{Z}$) help the generator models (i.e. TC-GAN and viseme-concatenation) generate videos for the set $\mathcal{X}_{\text{unseen}}^{\text{TC-GAN}}$ and $\mathcal{X}_{\text{unseen}}^{\text{viseme-concat}}$ (Section 5).
- The sets $\mathcal{X}_{\text{unseen}}^{\text{TC-GAN}}$ and $\mathcal{X}_{\text{unseen}}^{\text{viseme-concat}}$ are then included individually in $\mathcal{S}_{\text{train}}$ in different experiments. These different experiments allow us to gauge the efficacy of the two approaches of generating unseen videos separately (Section 6).

The model thus trained is then tested on the full $\mathcal{S}_{\text{test}}$ which contains original (rather than hallucinated) videos of both seen and unseen classes (Section 6).

A question that generally gets asked is how the viseme-concatenation and TC-GAN approaches are able to generate videos of unseen classes just by taking audio as input. The viseme-concatenation approach uses a manual mapping of phonemes to visemes. The TC-GAN approach, by contrast, learns to maps phonemes to visemes. The TC-GAN approach *also* learns inter-visemic frames, i.e., the frames which occur between two visemes. These are intermediate mouth formations that stitch together articulations. The importance of these inter-visemic frames can be judged by the fact that the viseme-concatenation approach does not result in as good a performance as the TC-GAN approach (Section 6).

5 Methodology

In this paper, we apply two techniques for zero-shot learning to VSR; first, a Generative Adversarial Network (GAN) approach, and second, a viseme-concatenation approach. Both approaches synthesize examples belonging to classes unseen at training time ($\mathcal{Y}_{\text{unseen}}$). The overall process is depicted in Figure 1. Our pipeline is primarily composed of two components - a *generator* \mathbf{G} (Sections 5.1 and 5.2), and a *classifier network* \mathbf{H} , whose main objective is to perform the task of VSR as described in Section 5.3. \mathbf{G} 's task is to hallucinate lip movement videos for utterances not present in the training data ($\mathcal{Y}_{\text{unseen}}$), as if they were uttered by a given speaker in a given pose. \mathbf{H} is then trained on the videos generated by \mathbf{G} as well as on real videos, hopefully making it a robust model which performs well on both seen and unseen class labels ($\mathcal{Y}_{\text{unseen}} \cup \mathcal{Y}_{\text{seen}}$). Next, we describe the individual components of our pipeline.

5.1 Temporal Conditional GAN

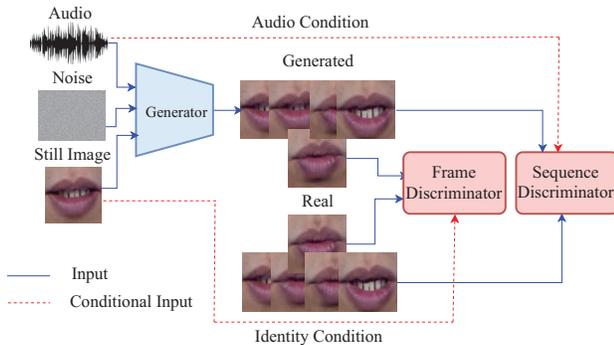


Figure 2: Temporal Conditional GAN Architecture

Our TC-GAN model relies heavily on the recently published work of Vougioukas, Petridis and Pantic (2018), who generated lip movement animations using *Temporal Conditional GANs* (TC-GANs). We reimplemented the model from this paper, and trained it on *Oulu-sentences*. The TC-GAN is composed of three main components as shown in Figure 2, comprising a single *generator*, and two discriminators - (a) a *frame discriminator*, and (b) a *sequence discriminator*. Here, we briefly describe these components³.

Generator: The Generator takes three inputs: a) Gaussian noise, b) a still image of a speaker’s lips from a particular pose angle, and c) audio of the utterance being spoken. The Generator is conditioned on the audio and the still image. The image encoding component of the popular segmentation algorithm U-Net (Ronneberger, Fischer, and Brox 2015) is used to encode the identity of the still image into a latent vector (L_I). The audio file is divided into small chunks of overlapping samples, passed through a 1-D CNN, and then fed to a 2-layer GRU to generate an audio latent vector (L_A)

³Please refer to the original work (Vougioukas, Petridis, and Pantic 2018) for a detailed description and also refer to the supplementary materials to get a detailed list of the we parameters used in training our TC-GAN model.

for each corresponding time step. A Gaussian noise vector is generated and passed through a 1-Layer GRU at each time step producing the noise latent vector (L_N). The three latent vectors are concatenated to produce a single latent vector which is then passed through the decoder component of U-Net to produce the generated frame at that time step. The process is repeated temporally to generate animated frames corresponding to each chunk of the audio sample.

Frame Discriminator: The Frame Discriminator is a 6-Layer CNN. It takes in an input frame, which can be real or hallucinated, concatenated channel-wise with the still image of the speaker’s lips (the same still image as the one fed to the Generator) and is trained to discriminate whether the input frame is real or hallucinated.

Sequence Discriminator: The real or hallucinated video segment at each time step is passed through a 2-Layer GRU and encoded into a latent vector. The same is done for the audio sample at each time step. The two latent vectors produced are then concatenated and passed through a 2-Layer CNN that is trained to discriminate between real and hallucinated videos.

$$\begin{aligned} \mathcal{L}_{adv}(D_{img}, D_{seq}, G) = & \mathbf{E}_{x \sim P_d} [\log D_{img}(S(x), x_1)] + \\ & \mathbf{E}_{z \sim P_z} [\log (1 - D_{img}(S(G(z)), x_1))] + \\ & \mathbf{E}_{x \sim P_d} [\log D_{seq}(x, a)] + \\ & \mathbf{E}_{z \sim P_z} [\log (1 - D_{seq}(G(z), a))] \quad (1) \end{aligned}$$

Equation 1 represents the adversarial loss function of the model where D_{img} is the frame discriminator, D_{seq} is the sequence discriminator, G is the frame generator, x_1 and a are the still image and audio used as conditional input, and x is the sequence of frames in the real video.

$$\mathcal{L}_{L_1} = \sum_{p \in [0, W] \times [0, H]} |F_p - G_p| \quad (2)$$

The pixel-wise L_1 reconstruction loss function on each real frame F and generated frame G , both of dimension $W \times H$, is represented by Equation 2.

For generating a video of a particular speaker speaking a given utterance, we follow these steps: First, if the phrase is present in *Oulu-phrases* then the audio file present in the dataset are used; otherwise, an off-the-shelf Text-To-Speech (TTS) software⁴ is used for generating the audio file. The audio file is sub-sampled into small overlapping chunks. Each chunk, along with the still image of the speaker, is given as input to the generative part of our model, which then outputs a video frame (an image) corresponding to that chunk. The resulting sequence of video frames is then concatenated to obtain the generated video, as shown in Figure 1.

In essence, the model learns how to map phonemes to visemes. Our model is trained on *Oulu-sentences*, which contains phonetically rich TIMIT sentences in the English language. Therefore, our model sees a rich set of phonemes for the English language during training. Based on this, we assume that it can map the majority of the phonemes present in the English language into their corresponding visemes. However, our model is not limited to only a fixed set of visemes; it also generates inter-visemic frames, as shown in Figure 3a. These inter-visemic frames cannot be mapped to a distinct

⁴<https://ttsmp3.com>

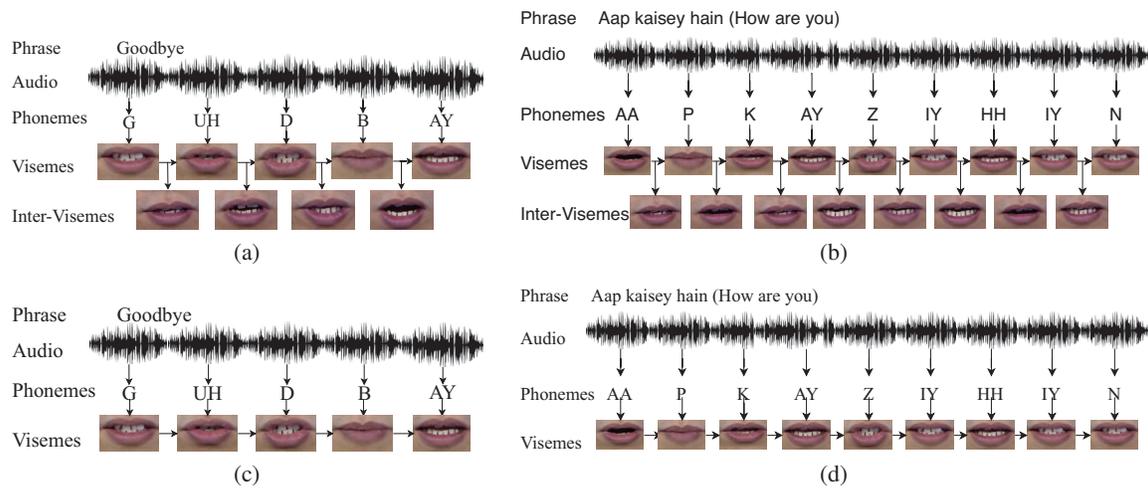


Figure 3: TC-GAN output for an English phrase (a) and a Hindi phrase (b). Viseme-concatenation output for an English phrase (c) and a Hindi phrase (d).

viseme; they are the images of intermediate mouth movements present between two different visemes. The generation of inter-visemic frames by the TC-GAN model ensures smooth and continuous mouth movements in the entire generated video. If the inter-visemic frames are removed, anyone viewing the video would not be able to understand what is being spoken and the whole video would appear unnatural and discontinuous.

We also generate utterances in the Hindi language, which is an entirely different language from that which the model is trained on. The generation process of an example Hindi phrase *Aap kaisey hain* (How are you) is shown in Figure 3b. The phonemes present in the Hindi phrase are ['AA', 'P', 'K', 'AY', 'Z', 'IY', 'HH', 'IY', 'N'], and are also present in the TIMIT sentences. Since the phonemes are common and our model has learned to map phonemes to visemes, it can also generate a video of a speaker speaking the given Hindi utterance.

5.2 Concatenation of Visemic Frames

The other technique we use for data generation is a Concatenative Visual Speech Synthesis technique (Huang, Cosatto, and Graf 2002), which works by using visemes to generate video. Each viseme is just an image of a mouth movement of a speaker corresponding to a particular sound. In this technique, video of lip movements for a speaker speaking a given utterance is generated by selecting appropriate viseme frames from a set of visemes belonging to that speaker and then concatenating them. This technique requires a speaker-to-viseme database which contains the set of all the visemes for every speaker whose video we want to generate. We construct this database⁵ by parsing through all the videos present in *Oulu-sentences* and manually annotating the set of visemes for every speaker present in the dataset.

An example of the generation process using this technique for a speaker speaking the English utterance *Goodbye* is

⁵It is present in the supplementary file.

shown in Figure 3c. For generating a video of a speaker speaking the given English utterance, first, we convert the utterance into a sequence of phonemes using the CMU Pronouncing Dictionary⁶. The resulting sequence of phonemes ['G', 'UH', 'D', 'B', 'AY'] is then matched to a sequence of visemes. This is done by using the Jeffers phoneme-to-viseme map (Jeffers and Barley 1971)⁷, and our speaker-to-viseme database. Note that the *Oulu-sentences* segment of the dataset contains all the phonemes present in the English language. Therefore it also contains all the visemes which can be mapped from any phoneme present in the English language. The resulting sequence of visemes is then concatenated to obtain the generated video.

We also generate phrases in the Hindi language. An example is the Hindi phrase *Aap kaisey hain* (How are you) shown in Figure 3d. Since the CMU Pronouncing Dictionary does not work for Hindi phrases, we use *phonemizer*⁸, which is a multi-lingual text-to-phoneme converter, for converting an input Hindi phrase into a sequence of phonemes. Since most of the phonemes present in Hindi are also present in *Oulu-sentences*, we can use the Jeffers phoneme-to-viseme map and our speaker-to-viseme database to map the sequence of phonemes into a sequence of visemes. The resulting sequence of visemes is then concatenated to obtain the generated video.

5.3 VSR Model

For VSR, we implement and train the architecture from (Petridis et al. 2017). The architecture consists of three parts: an encoder, two BiLSTM layers, and a fully connected layer. The encoder is composed of 3 fully connected layers. The resultant encoded information is concatenated with its first and second derivative and fed into the first BiLSTM layer. The first BiLSTM layer captures the temporal information associated with each view, and the second BiLSTM layer is

⁶<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

⁷More details are present in the supplementary materials.

⁸<https://github.com/bootphon/phonemizer>

used to fuse the information from different views. The last fully-connected layer produces the final output class probabilities. We use PyTorch (Paszke et al. 2017), and release our implementation as part of the supplementary material.

6 Experiments

6.1 Experimental Setting

In this section, we present experimental results:

- comparing the TC-GAN and viseme-concatenation approaches for adding new (unseen) classes to $\mathcal{Y}_{\text{unseen}}$ (Section 6.2).
- assessing the extent to which a VSR model can harness hallucinated data for improving accuracy on $\mathcal{Y}_{\text{unseen}}$ while maintaining accuracy for $\mathcal{Y}_{\text{seen}}$.
- evaluating the ability of the overall system to adapt to a new language (Section 6.3).

In order to conduct our experiments we train three different model configurations:

- **Model 1** - a model trained only on *Oulu-phrases*.
- **Model 2** - a model trained on *Oulu-phrases* and on videos generated using the TC-GAN model.
- **Model 3** - a model trained on *Oulu-phrases* and on videos generated using viseme-concatenation.

Table 1: Overall top-3 accuracy (%) for each model on $\mathcal{Y}_{\text{unseen}} \cup \mathcal{Y}_{\text{seen}}$.

S. No.	Seen classes	Unseen classes	Model 1	Model 2	Model 3
1	9	1	90	98.3	86.4
2	8	2	79.4	90	76.4
3	7	3	70	84.4	67.8
4	6	4	60	72.2	60.5
5	0	10	30.8	70	34.7

Table 2: Overall top-1 accuracy (%) for each model on $\mathcal{Y}_{\text{unseen}} \cup \mathcal{Y}_{\text{seen}}$.

S. No.	Seen classes	Unseen classes	Model 1	Model 2	Model 3
1	9	1	87.2	89.7	81.1
2	8	2	77.2	81.4	70.8
3	7	3	69.2	72.8	62.8
4	6	4	58.9	65.3	54.7
5	0	10	13.3	40	17.5

As suggested by (Petridis et al. 2017; Anina et al. 2015; Kumar et al. 2018b), we take 35 speakers for training, 5 speakers for validation, and the remaining 12 speakers for testing the different models. For *Oulu-phrases*, which is used to train the VSR model, this results in a total of 5250, 750, and 1800 videos belonging to train, test, and validation data respectively. Similarly, for *Oulu-sentences*, which is used to train the TC-GAN model, this results in a total of 875, 125, and 300 videos belonging to train, test, and validation data

Table 3: The best view combination ranked according to their score on validation split.

Experiment	S. No.	Model 1	Model 2	Model 3
Set-1	1	{0°,30°,45°,60°,90°}	{30°,60°,90°}	{0°}
	2	{0°,45°,90°}	{0°,45°,60°,90°}	{0°}
	3	{0°,60°,90°}	{45°,90°}	{0°}
	4	{0°,30°,60°,90°}	{45°,60°,90°}	{0°}
	5	{0°}	{0°,30°,45°,60°}	{0°}
Set-2	1	{45°,60°,90°}	{0°,30°,45°,90°}	{0°}

respectively. It is to be noted that the VSR model and the data generation models share the same training, validation and test speaker splits, ensuring that no data belonging to the test speakers is seen by the data generating models during training and vice versa.

In our experiments, Model 1 is our baseline. The major differences between Models 2 and 3 are that the videos belonging to Model 2, on account of being generated using TC-GANs, have inter-visemic frames. Thus, Model 2’s videos have articulatory continuity while frames in videos for Model 3 are non-continuous.

We report top-1 and top-3 accuracy. Top-3 accuracy denotes whether the true utterance occurs in the top-3 predicted utterances ranked by confidence. Top-1 and Top-3 results follow the same trend.

Note that since our dataset has multiple views (pose angles), we have a separate TC-GAN model for each view and the pose of the input speaker lip region image corresponds with the TC-GAN model used. For handling different view videos as input, we have a different VSR model corresponding to each view combination. During evaluation, we report the results on the test split for the view combination VSR model which has the highest score on the validation split. For all the different sets of experiments, the best view-combination is given in Table 3.

6.2 Experiment-Set 1 : Experiments on Unseen Classes

In this set of experiments, we test our models on *unseen classes* in the same language ($\mathcal{Y}_{\text{unseen}}$). The unseen classes are utterances not present in the real training data. Model 1 therefore contains no instances of these utterances. Model 2 contains instances of these utterances hallucinated by the TC-GAN model, while model 3 contains instances of these utterances hallucinated by viseme-concatenation.

Table 4: Overall top-3 accuracy (%) of each model on $\mathcal{Y}_{\text{unseen}}$ and separately on $\mathcal{Y}_{\text{seen}}$ (separated by /).

Unseen classes	Model 1	Model 2	Model 3
1	0 / 100	86.1 / 99.7	5.5 / 95.4
2	0 / 99.3	52.8 / 99.3	1.4 / 95.1
3	0 / 100	49.1 / 99.6	0.0 / 96.8
4	0 / 100	31.2 / 99.5	8.3 / 95.4

The top-3 and top-1 accuracies of each model on $\mathcal{Y}_{\text{seen}} \cup \mathcal{Y}_{\text{unseen}}$ are given in Tables 1 and 2. To make sure that training on new data for $\mathcal{Y}_{\text{unseen}}$ does not negatively impact the

Table 5: Overall top-1 accuracy (%) of each model on $\mathcal{Y}_{\text{unseen}}$ and separately on $\mathcal{Y}_{\text{seen}}$ (separated by /).

Unseen classes	Model 1	Model 2	Model 3
1	0 / 96.9	33.3 / 96	0 / 90.1
2	0 / 96.5	23.6 / 95.8	0 / 88.5
3	0 / 98.8	21.3 / 94.8	0 / 89.7
4	0 / 98.1	16 / 98.1	0.7 / 90.7

accuracy for $\mathcal{Y}_{\text{seen}}$, we also report top-3 and top-1 accuracies on $\mathcal{Y}_{\text{unseen}}$ and $\mathcal{Y}_{\text{seen}}$ separately in Tables 4 and 5.

We ran multiple experiments varying the number of unseen classes from 10% to 40%. Model 2 always performs better than Models 1 and 3 on unseen classes. The top-3 accuracy cost for Model 2 on seen classes is 0%-0.5%, depending on the number of unseen classes. Model 3 performs far worse than Model 2 on unseen classes, while incurring a bigger accuracy cost on seen classes.

As an extreme case, we also show the performance of the models when 100% of the classes are unseen, in the last row in Tables 1 and 2. This represents the cold-start problem for VSR systems (Amplayo et al. 2018). On this type of problem, Model 1 gets a top-3 accuracy of just 30.8% (an accuracy which is still better than randomly predicting any class) while Model 2 gets more than double that accuracy. This indicates that data hallucination can be used for cold-start for VSR.

6.3 Experiment-Set 2 : Experiments on New Language

In this set of experiments, our unseen classes are classes from a different language, Hindi. The utterances we use for this analysis are { *Aap kaisey hai* (How are you), *Kyaa chal raha hai* (What is going on), *Shubh raatri* (Good night) }. We test the models by retaining English Oulu-phrases $\mathcal{Y}_{\text{seen}}$ and introducing Hindi phrases as $\mathcal{Y}_{\text{unseen}}$. The results on the test split pertaining to this setting are given in Table 6.

Both Model 2 and Model 3 perform better than the Model 1. This is expected since Model 1 has never seen any Hindi phrase. Model 2 performs better than Model 3 by a margin of 22.5%. It is also interesting to note that Model 2 has a 100% top-3 accuracy on the newly introduced Hindi phrases ($\mathcal{Y}_{\text{unseen}}$), while Model 3 only has a 32.4% top-3 accuracy. Additionally, the top-1 accuracy of Model 2 and Model 1 is almost the same when evaluated on the seen classes, while we see a drop in top-1 accuracy of more than 2% in the case of Model 3. Both of these results again demonstrate the importance of the presence of inter-visemic frames in the generated video.

Since, for Model 2, new data has a negligible affect on seen class accuracy, and the difference between Model 1 and Model 2 is only the addition of the data generated by the TC-GAN model for unseen classes, we can safely assume that we can extend the number of classes significantly without adversely affecting the performance on seen classes. These results clearly show that Model 2 is able to clearly distinguish between Hindi and English phrases and is able to predict both with a very high degree of certainty compared to both Model 1 and Model 3.

Table 6: Overall accuracy of each model on unseen classes from a different language (Hindi)

Metrics	Model 1	Model 2	Model 3
Top-1 Accuracy (Acc)	72.6	95.1	72.6
Top-3 Acc	76.7	99.6	83.9
Unseen Class Top-1 Acc	0	96.3	7.4
Unseen Class Top-3 Acc	0	100	32.4
Seen Class Top-1 Acc	94.4	94.7	92.2
Seen Class Top-3 Acc	99.7	99.4	99.4

7 Conclusions and Future Work

In this paper, we introduced a GAN-based method to deal with the zero-shot learning problem for VSR. We demonstrated that this method is more effective than viseme-concatenation for data augmentation for unseen classes, and that it does not cause appreciable degradation in accuracy for seen classes. We show that this method can be used to help a VSR model generalize to unseen classes and to a new language.

Although in this work, we explored the phonetically close language pair of English and Hindi, it would be interesting to explore it on phonetically-far language pairs, like English and Hopi, where new visemes must also be hallucinated. Additionally, we would like to explore TC-GAN data augmentation for continuous VSR, as opposed to the classification-based VSR approach we used in this paper.

Acknowledgement

This research has been supported in part by Singapore Ministry of Education Academic Research Fund Tier 2 under MOE’s official grant number MOE2018-T2-1-103.

Rajiv Ratn Shah is partly supported by the Infosys Center for AI, IIIT Delhi and ECRA Grant (ECR/2018/002776) by SERB, Government of India.

References

- Amplayo, R. K.; Kim, J.; Sung, S.; and Hwang, S.-w. 2018. Cold-start aware user and product attention for sentiment classification. *arXiv preprint arXiv:1806.05507*.
- Anina, I.; Zhou, Z.; Zhao, G.; and Pietikäinen, M. 2015. OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis. In *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*.
- Antoniou, A.; Storkey, A.; and Edwards, H. 2017. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*.
- Chung, J. S.; Senior, A. W.; Vinyals, O.; and Zisserman, A. 2017. Lip reading sentences in the wild. In *Proceedings of CVPR*.
- Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine* 35(1):53–65.

- Easton, R. D., and Basala, M. 1982. Perceptual dominance during lipreading. *Perception & Psychophysics* 32(6):562–570.
- Elhoseiny, M.; Zhu, Y.; Zhang, H.; and Elgammal, A. 2017. Link the head to the “beak”: Zero shot learning from noisy text description at part precision. In *Proceedings of CVPR*.
- Garofolo, J. S. 1993. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*.
- Goldschen, A. J.; Garcia, O. N.; and Petajan, E. D. 1997. Continuous automatic speech recognition by lipreading. In *Motion-Based Recognition*. Springer. 321–343.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems, 2672–2680*.
- Hariharan, B., and Girshick, R. 2017. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of ICCV*.
- Huang, F. J.; Cosatto, E.; and Graf, H. P. 2002. Triphone based unit selection for concatenative visual speech synthesis. In *Proceedings of ICASSP*.
- Jeffers, and Barley. 1971. *Speechreading (Lipreading)*. Charles C Thomas Pub Ltd.
- Koch, G.; Zemel, R.; and Salakhutdinov, R. 2015. Siamese neural networks for one-shot image recognition. In *Proceedings of the ICML Deep Learning Workshop*.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.
- Kumar, Y.; Aggarwal, M.; Nawal, P.; Satoh, S.; Shah, R. R.; and Zimmermann, R. 2018a. Harnessing AI for speech reconstruction using multi-view silent video feed. In *Proceedings of the ACM Multimedia Conference*.
- Kumar, Y.; Jain, R.; Salik, M.; ratn Shah, R.; Zimmermann, R.; and Yin, Y. 2018b. Mylipper: A personalized system for speech reconstruction using multi-view visual feeds. In *2018 IEEE International Symposium on Multimedia (ISM)*, 159–166. IEEE.
- Kumar, Y.; Jain, R.; Salik, K. M.; Shah, R. R.; Yin, Y.; and Zimmermann, R. 2019. Lipper: Synthesizing thy speech using multi-view lipreading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2588–2595.
- Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–1338.
- Lee, D.; Lee, J.; and Kim, K.-E. 2016. Multi-view automatic lip-reading using neural network. In *Proceedings of the Asian Conference on Computer Vision*.
- Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of ICML*.
- Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. <https://openreview.net/pdf?id=BJJsrnfCZ>.
- Petridis, S.; Wang, Y.; Li, Z.; and Pantic, M. 2017. End-to-end multi-view lipreading. *arXiv preprint arXiv:1709.00443*.
- Ravi, S., and Larochelle, H. 2017. Optimization as a model for few-shot learning. In *Proceedings of ICLR*.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*.
- Salakhutdinov, R.; Tenenbaum, J.; and Torralba, A. 2012. One-shot learning with a hierarchical nonparametric Bayesian model. In *Proceedings of the ICML Workshop on Unsupervised and Transfer Learning*.
- Salik, K. M.; Aggarwal, S.; Kumar, Y.; Shah, R. R.; Jain, R.; and Zimmermann, R. 2019. Lipper: Speaker independent speech synthesis using multi-view lipreading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 10023–10024.
- Schmidt, C.; Koller, O.; Ney, H.; Hoyoux, T.; and Piater, J. 2013. Using viseme recognition to improve a sign language translation system. In *Proceedings of the International Workshop on Spoken Language Translation*.
- Shah, R., and Zimmermann, R. 2017. *Multimodal analysis of user-generated multimedia content*. Springer.
- Shrivastava, N.; Saxena, A.; Kumar, Y.; Shah, R. R.; Stent, A.; Mahata, D.; Kaur, P.; and Zimmermann, R. 2019. Mobivsr: Efficient and light-weight neural network for visual speech recognition on mobile devices. *Proc. Interspeech 2019* 2753–2757.
- Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. 2013. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, 935–943.
- Thrun, S. 1998. Lifelong learning algorithms. In *Learning to Learn*. Springer. 181–209.
- Uttam, S.; Kumar, Y.; Sahrawat, D.; Aggarwal, M.; Shah, R. R.; Mahata, D.; and Stent, A. 2019. Hush-Hush Speak: Speech Reconstruction Using Silent Videos. In *Proc. Interspeech 2019*, 136–140.
- Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 3630–3638.
- Vougioukas, K.; Petridis, S.; and Pantic, M. 2018. End-to-end speech-driven facial animation with temporal GANs. *arXiv preprint arXiv:1805.09313*.
- Wand, M.; Koutník, J.; and Schmidhuber, J. 2016. Lipreading with long short-term memory. In *Proceedings of ICASSP*.
- Wang, Y.-X.; Girshick, R.; Hebert, M.; and Hariharan, B. 2018. Low-shot learning from imaginary data. *arXiv preprint arXiv:1801.05401* 8.
- Zhou, P.; Yang, W.; Chen, W.; Wang, Y.; and Jia, J. 2019. Modality attention for end-to-end audio-visual speech recognition. In *Proceedings of ICASSP*.
- Zhu, Y.; Elhoseiny, M.; Liu, B.; Peng, X.; and Elgammal, A. 2018. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of CVPR*.