Personalized One-Shot Lipreading for an ALS Patient

Bipasha Sen*[1]
bipasha.sen@research.iiit.ac.in

Aditya Agarwal*[1]
aditya.ag@research.iiit.ac.in

Rudrabha Mukhopadhyay[1]
radrabha.m@research.iiit.ac.in

Vinay Namboodiri[2]
vpn22@bath.ac.uk

C V Jawahar[1]
jawahar@iiit.ac.in

[1] CVIT, IIIT Hyderabad,
Telangana, India

[2] University of Bath,
Claverton Down,
Bath, England

Abstract

Lipreading or visually recognizing speech from the mouth movements of a speaker is a challenging and mentally taxing task. Unfortunately, multiple medical conditions force people to depend on this skill in their day-to-day lives for essential communication. Patients suffering from 'Amyotrophic Lateral Sclerosis' (ALS) often lose muscle control, consequently their ability to generate speech and communicate via lip movements. Existing large datasets do not focus on medical patients or curate personalized vocabulary relevant to an individual. Collecting large-scale dataset of a patient, needed to train modern data-hungry deep learning models is however, extremely challenging. In this work, we propose a personalized network to lipread an ALS patient using only one-shot examples. We depend on synthetically generated lip movements to augment the one-shot scenario. A Variational Encoder based domain adaptation technique is used to bridge the real-synthetic domain gap. Our approach significantly improves and achieves high top-5 accuracy with 83.2% accuracy compared to 62.6% achieved by comparable methods for the patient. Apart from evaluating our approach on the ALS patient, we also extend it to people with hearing impairment relying extensively on lip movements to communicate.

1 Introduction

Lipreading is the skill of recognizing speech visually from a person's lip movements. Humans naturally rely on lipreading to discern speech, especially in crowded and noisy environments [25]. It is the fundamental mode of communication for many people, such as (1) those suffering from medical conditions such as Amyotrophic Lateral Sclerosis (ALS) - leading them to lose their voice [20, 52], or (2) those with hearing impairment - making it difficult for them to produce proper voice. In such cases, talking to a person without voice may need you to lipread them to understand the spoken words.
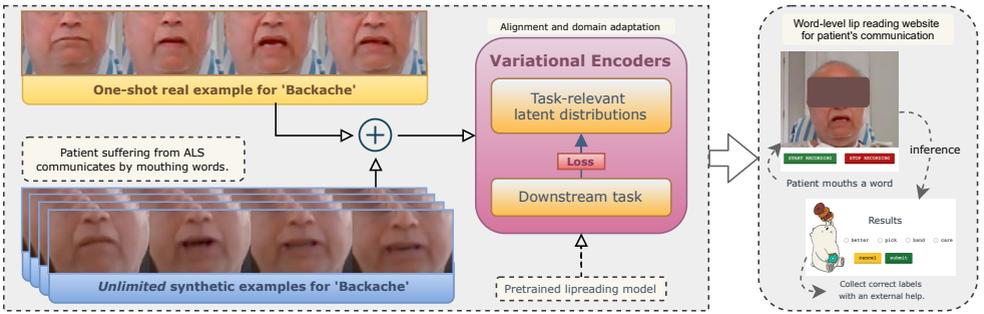
---

*These authors contributed equally to this work.

Figure 1: We propose a personalized one-shot lipreading framework to tackle a real-world medical challenge of lipreading a patient suffering from ALS[1]. The patient communicates primarily by mouthing the words. In our case of limited available real data, we use a combination of synthetic data augmentation technique and a domain adaptation technique called Variational Encoders to build a robust word-level lipreading model for the patient.

Lipreading is mentally taxing and can affect the communication quality. For instance, speakers with hearing impairment may lack holistic audio feedback [29] and ALS patients may have lesser control over their mouth muscles [6, 18]. This may cause them to have irregular and unreliable mouth movements making it difficult for people to lipread them. Applications and automated algorithms capable of lipreading a person can thus significantly improve the day-to-day communication of people dependent on lipreading. Motivated by this need, we tackle the real-world challenge of lipreading a patient suffering from ALS and people with hearing impairment. ALS is a progressive nervous system disease that affects nerve cells in the brain and spinal cord, causing loss of muscle control [32]. An ALS patient may lose their voice and rely solely on mouth movements for communication [2, 14].

Current Works and Limitations: Current deep learning techniques are inherently data-hungry. Collecting large amounts of data specifically from a patient is, however, not an option. Mouthing words is a tiring maneuver for people suffering from ALS, and thus a patient undergoes physical and mental stress during such data collection exercises. Manually labeling words mouthed by a person is time-consuming. It is thus crucial to use the minimum amount of manually labeled data to build lipreading models that can work well on a person.

Recent years have seen much progress in word-level [10, 19, 28] and sentence-level [5, 31] lipreading. Oxford's Visual Geometry Group released large-scale in-the-wild datasets such as Lipreading in the Wild (LRW) [7] and Lipreading sentences (LRS) [3, 4] consisting of 1000+ speakers. LRW, the most relevant dataset to our task, is a word-level lipreading dataset made of 1000 examples for 500 English words but turns out to be somewhat limiting: (1) The speakers in the dataset do not have any speaking disability thus making perfect mouth movements. (2) It is curated by cropping words from long speech segments resulting in fast-paced speech with co-articulation in the videos. (3) It contains a large amount of head motion and variations like the different characteristics of the mouth region, both of which are unnecessary for lipreading specific medical patients. Thus, SOTA models like LipReading without Pains (LRwP) [10] and Lipreading using Temporal Convolutional Networks (LTCN) [19] trained on LRW do not directly adapt to speakers with speaking disabilities.

---

[1] The ALS patient and his family have consented to the use of his pictures in this work.

Figure 2: Speakers in our study. In left to right order, the first speaker suffers from ALS and relies solely on lip movements. The next two speakers primarily use sign language while making imperfect lip movements. Next speaker uses deaf speech along with sign language for daily communication. The following speaker is the 46th president of the USA, Joe Biden.

LRW only supports a limited pre-curated vocabulary missing out on medically essential words like 'nauseous' or 'backache'. It also lacks a personalized vocabulary relevant to a person's daily communication. Deploying systems to enable a persons' communication would need highly accurate models on their specific lip movements for their particular vocabulary. The problem of personalized lip reading has also been explored in [22]. According to [22], lip movements vary across speakers. Observing a single speaker for an extended period could lead to better speaker-specific lipreading models. They collect ~20 hours of data per speaker to train a personalized lipreading model generating speech purely from an individual speaker's lip movements. Collecting such a large dataset is, however, not always an option.

Lack of medical data has been studied [11, 13, 26] widely in the past. Taking inspiration from these, we formally tackle one-shot lipreading in a personalized setting. We first synthetically generate data using a SOTA lipsync network [23]. We then use a SOTA lipreading network [10] as our backbone and use synthetically generated data along with very limited real examples to train a word-level lipreading network. Our approach includes an important domain adaptation step using a novel network – Variational Encoders – a modified VAE for bringing a vast number of synthetic examples closer to the real domain using only one real instance per class. We specifically tackle the use case of an ALS patient and also explore the same for four other speakers (refer to Fig. 2) including the 46th President of the United States, Joe Biden, as an additional example to show that our approach can easily be extended to speakers with no disability. Our contributions in this work are threefold:

1. We tackle a real-world medical challenge of lipreading speakers with ALS and hearing impairment by developing highly accurate personalized models for each speaker.

2. We propose Variational Encoders, a novel network-based on VAE. Instead of autoencoding, they exploit the loss of the downstream task for generating task-relevant latent distributions. The learned distributions are then used for domain adaptation.

3. To the best of our knowledge, we are the first to propose lipreading in one-shot setting. In this vein, we curate a medical dataset involving speakers with medical conditions.

2 One-shot personalized lipreading framework

As mentioned previously we aim to build a personalized lip-reading model for each speaker using only single real examples. Fig. 3 presents the pipeline for personalized lipreading.
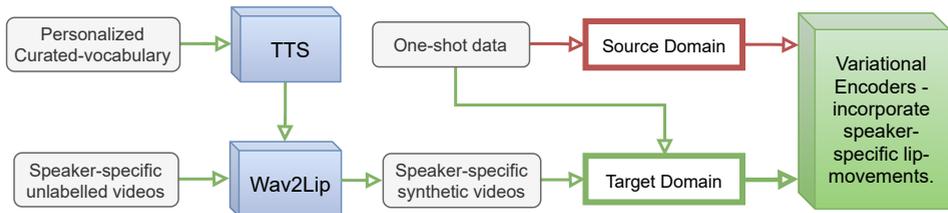
Figure 3: Personalized lipreading - We start by curating personalized vocabulary and collect unlabeled videos for each speaker. TTS models are used to generate speech utterances for the curated words. The speaker's unlabeled videos and the generated speech utterances are given as input to Wav2Lip that generates synthetic data to augment one-shot data. Variational Encoders then use the synthetic and the one-shot data to train the model for lipreading.

The use of synthetic data to augment low data in the medical domain [1, 13, 26] has gained traction with improving generative models. Similarly, we augment the one-shot examples collected for each speaker by generating synthetic data for each of them. To achieve this, we use a SOTA talking face generation model, Wav2Lip [23] pretrained on the large-scale LRS2 [8] dataset. Given a speaker's video, Wav2Lip preserves the speaker's pose, facial expressions, and mouth characteristics like beard and skin color while modifying the speaker's lip movements according to a guiding speech. We generate word-level speech utterances using SOTA TTS models FastSpeech2 [24] and GlowTTS [16] as a replacement for the speaker's voice. Using these TTS models allows us to generate variations in the speech in terms of the speed of the spoken word, pitch, and energy. Additionally, we collect unlabeled face videos for each speaker which, along with the generated speech utterances, is used to create 1 hour of speaker-specific synthetic data on an average. The augmented dataset is then used to train an LRwP and LTCN based architecture for the classes curated per speaker.

A combination of the speaker-specific synthetic videos and a single real video per class are used to train our model. The synthetic data helps the model learn the general underlying word-level characteristics for the new classes. However, the properties of personal style of lip-movements for a word – could be because of the medical condition – is not captured in the synthetic dataset. We utilize the one-shot examples for introducing the properties of personal speaking style in the model. Single examples per class are however, not enough to capture the underlying style variations of a speaker. A person may not utter the same word, exactly, each time. To tackle this, we use our novel approach – Variational Encoders.

2.1   Variational Encoders: Mapping words into distributions

Deep learning suffers from the fundamental challenge of source-target domain shift - a model trained on a given dataset (source domain) performs poorly on the test examples (target domain). The target domain may lack the amount of labeled data needed for training or fine-tuning a model. Recent techniques include adversarial networks to generate domain invariant features using adversarial losses [12, 21, 30] and Variational Auto Encoders (VAE) [15, 27] to generate a joint latent distribution across domains with KL divergence and reconstruction loss. Unlike adversarial loss [12] that may quickly become lopsided [9], VAEs use a distance-based metric to incorporate domain invariance. The reconstruction loss in VAE ensures that
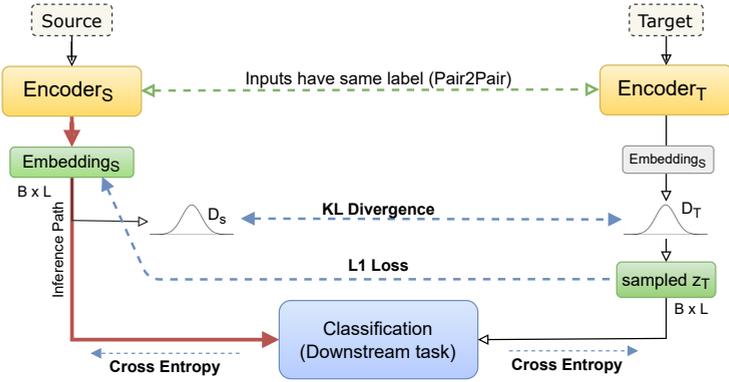
Figure 4: Variational Encoders – $Encoder_S$ and $Encoder_T$ denote the source and target encoders respectively. $Encoder_T$ generates a latent distribution using the target examples. The latent distribution is trained against the downstream classification task. $Encoder_S$ aims to learn robust domain invariant features by minimizing the distance between its learned embedding and points randomly sampled from the generated target distribution.

latent representations preserve important domain characteristics. However, using a decoder to reconstruct the input accurately is a non-trivial task, especially for videos needing spatial and continuous temporal reconstruction. We propose Variational Encoders, a modified VAE that uses the loss of the downstream task to generate task-relevant latent distributions.

How do Variational Encoders differ from a standard VAE? Similar to VAEs, Variational Encoders (refer to Fig. 4) generate latent distributions given a sample. A key difference with a standard VAE is - instead of autoencoding, it uses the final multi-class classification loss of the downstream task to generate a task-relevant latent distribution that represents the "class" of input instead of a generic input feature. The learned distribution is used to sample a variation of the input example. We hypothesize that this sample introduces the missing speaker-specific style variation for the input class. The sampled variation encourages the encoder to see and align to the potential interpretation of the input class and learn robust representations. In summary, Variational Encoder retains all the benefits of a VAE, like generating domain invariant features while removing the complicated video reconstruction loss.

2.1.1 Network Architecture

Given a source domain S and a target domain T, domain adaptation aims to bridge the gap between the two domains by generating domain invariant features. We denote S and T as,

$$S = \{(x_1,y_1),(x_2,y_2),...(x_{N_S},y_{N_S})\} \text{ and } T = \{(\bar{x}_1,\bar{y}_1),(\bar{x}_2,\bar{y}_2),...(\bar{x}_{N_T},\bar{y}_{N_T})\} \quad (1)$$

where N_S is the number of source samples and N_T is the number of target samples.

Variational Encoder is a paired domain adaptation network that assumes the target domain is labeled. However, the number of examples in the target domain is expected to be at max k where $k \approx 1$. Thus, we assume $N_S >> N_T$. The adaptation network is made of two encoders - source and target encoders - and a single classifier. For a given input $\hat{x}_i$ belonging to either domain, the encoder learns to generates an embedding $e_{\hat{x}_i}$ for the input. A distribution

$p(\hat{x}_i)$ is then generated for the input denoted by $\mu_{\hat{x}_i}$ and $\sigma_{\hat{x}_i}$ from the learnt embedding $e_{\hat{x}_i}$. A random point is sampled from the generated distribution denoted as $z_{\hat{x}_i} = \mu_{\hat{x}_i} + \sigma_{\hat{x}_i} \odot \varepsilon$.

The encoder network for both domains are identical. Each encoder generates the embedding e, and the latent distribution p. To introduce domain invariance, we use KL divergence and L1 loss. The KL divergence is computed between the two generated distributions $p(x_i) = (\mu_{x_i}, \sigma_{x_i})$ and $p(\bar{x}_i) = (\mu_{\bar{x}_i}, \sigma_{\bar{x}_i})$. In addition, a point $z_{\bar{x}_i}$ is randomly sampled from the target distribution $p(\bar{x}_i)$. L1 loss is then applied between $z_{\bar{x}_i}$ and the source embedding e_{x_i}. This minimizes the distance between the source embedding against several randomly sampled points from the target distribution thus acting as a pseudo for multiple target examples.

Gradient stopping: The L1 Loss between the source embedding and the sampled target embedding will force the target distribution down to a single point, losing the essence of a distribution. To prevent that, we stop the gradient from flowing back through the target encoder. This way, only the source encoder is regularized against the variations sampled from the learnt target distribution, while leaving the target encoder unaffected. We denote the sampled target embedding as $z_{\bar{x}_i detach}$, where $_{detach}$ denotes that the sample $z_{\bar{x}_i}$ does not have any gradient. The combined loss to bridge the two domains is given as,

$$\Delta_{dist} = |e_{x_i} - z_{\bar{x}_i detach}| - \beta \cdot D_{KL}(p(x_i)||p(\bar{x}_i)). \tag{2}$$

where β is a hyper-parameter and i denotes both the samples belong to the same class.

For the downstream task of classification, we employ a common classifier that is trained on both domains at the same time. A single classifier allows both encoders to have a common base-loss helping them generate features relevant to the common downstream task. The downstream classifier also trains on large amounts of variations and learns robust representations. The classifier receives a randomly sampled point $z_{\bar{x}_i}$ from the learnt distribution $p_{\bar{x}_i}$ as input for the target domain. For the source encoder, however, the classifier receives the learnt embedding e_{x_i} instead of a randomly sampled latent point z_{x_i}. The motivation behind a target distribution is to encourage the target encoder to hallucinate variations of the single real examples available for each class. However, during inference, we want to obtain a definite point for the input to avoid any uncertainties. The classification loss is given as,

$$\Delta_{entropy} = g(e_i, y) + g(\bar{z}_i, y) \tag{3}$$

where g can be any classification loss such as cross entropy or negative log likelihood. y denote the same label for both the source x_i and target $\bar{x}_i$ input.

The combined loss for the entire network is then given as (see Fig. 4),

$$\Delta_{ve} = \alpha \cdot (g(e_i, y) + g(\bar{z}_i, y)) + \gamma \cdot (|e_{x_i} - z_{\bar{x}_i detach}| - \beta \cdot D_{KL}(p(x_i)||p(\bar{x}_i))). \tag{4}$$

where α and γ are hyper-parameters. All the network components are trained end-to-end.

3 Experiments

Dataset: We first collect a set of unlabeled videos for each speaker that are used later to generate the speaker-specific synthetic data using Wav2Lip. Unlabeled videos of the ALS patient are recorded without any manual intervention. For other speakers, the videos are randomly selected from their respective YouTube channels. As the next step, we curate the personalized vocabulary for each of the speakers. For the ALS patient, we curate a list of

200 words with the help of his family. We modify the existing list for the remaining speakers by removing irrelevant words and adding the most occurring keywords in the transcription of the collected videos. Splits of the curated dataset for each speaker is presented in Table 1.

		Patient	Spk-1	Spk-2	Spk-3	Joe Biden	Total
	Classes	200	75	70	80	75	-
Train	Real	200	75	70	80	75	-
	Synth	17×Real	22×Real	22×Real	22×Real	25×Real	-
	Aug.	2×Real	2×Real	2×Real	2×Real	2×Real	-
Test	Real	320	120	80	90	90	710

Table 1: Split up for the datasets curated for each of the 5 speakers.

We obtain two sets of manually curated real data from the ALS patient's family. We use the first set for model training and the second set as the intermediate test set. We collect additional data by deploying a website that records the patient's word-level mouthings and displays the inference on our best model (see Fig. 1). The website needs external help to start and stop recording the patient. The helper can then either select one of the displayed words as the correct label or manually assign the correct label. Through this exercise, we collect an additional 320 examples. Out of these, we use 200 data points, in addition to the original 200 train data points, to train a model in a two-shot setting. We report our test results of one-shot and two-shot models on the combined set of the intermediate test set and the remaining additional 120 examples (see Table 1). To simulate the same setting, we maintain one real example for training the rest of the speakers. To generate the train and test examples, we use the transcriptions with timestamps available on YouTube for the selected videos.

Preprocessing: Our preprocessing steps are similar to LRwP. The lip landmarks are first detected using dlib [7]. The lip is then cropped out such that it is horizontally and vertically centered in the cropped image. The image is converted to gray-scale and resized to a fixed dimension of 88×88. A maximum sequence length of 64 frames is used. A batch size of 16 on a multi-GPU NVIDIA GeForce RTX 2080 setup is used. The models are trained up to 200 epochs using a cosine scheduler and Adam optimizer, with a $3e-5$ learning rate and a weight decay of $1e-4$. The encoders in the experiments are adopted from LRwP that uses Resnet18 and BiGRU, and LTCN that uses Resnet18 and Temporal Convolutional Networks.

3.1 Training strategy

Table 2 presents a comprehensive overview of all the experiments conducted on the 5 speakers. Spk-1 uses deaf-speech and sign language, Spk-2 and Spk-3 use sign language as their primary mode of communication while mouthing words with imperfect lip movements. Table 3 presents additional experiments performed on the dataset of the ALS patient. We initialize our models with the weights of LRwP or LTCN, both of them pretrained on LRW. We observed that using the pretrained weights leads to faster convergence.

Baseline (Exp-$cl-r$): We begin by training our model directly on the one-shot examples. Since we are the first to perform one-shot lipreading on a personalized vocabulary, we treat this model as our baseline. Table 2 Exp-$cl-r$ presents the performance of the baseline model on each of the 5 speakers. The average accuracy of the speakers at top-1 and top-5 is only 33.0% and 51.6% respectively. The accuracy of the current SOTA lipreading model on the LRW dataset is 88.5% at top-1. This presents us with a huge scope for improvement.

Experiment	Patient top1	Patient top5	Spk-1 top1	Spk-1 top5	Spk-2 top1	Spk-2 top5	Spk-3 top1	Spk-3 top5	Joe Biden top1	Joe Biden top5	Avg. top1	Avg. top5
LRwP												
cl−r	49.3	62.6	32.4	52.1	23.1	42.5	26.7	49.3	33.5	51.5	33.0	51.6
cl−r+s	53.6	68.1	32.3	49.8	**37.2**	45.1	32.4	58.8	57.6	69.3	42.6	58.7
ve−r+s	66.4	81.4	**48.6**	**68.5**	34.4	**59.6**	41.6	71.4	54.1	72.4	49.0	**70.4**
cl−r+s+aug	61.3	75.2	34.6	51.2	33.4	40.4	51.2	63.6	**64.5**	75.3	49.4	61.3
ve−r+s+aug	**68.1**	**83.2**	44.3	62.6	31.2	50.8	**53.4**	**73.5**	61.4	**77.1**	**52.2**	69.4
LTCN												
cl−r+s	55.2	67.1	36.5	52.0	34.8	44.8	34.4	59.6	55.4	67.3	43.3	58.2
ve−r+s	64.5	80.2	**49.3**	**71.1**	**35.7**	**57.9**	46.4	73.2	56.3	71.1	50.4	**70.7**
cl−r+s+aug	61.9	74.2	33.2	49.8	32.4	39.9	52.5	66.1	**62.1**	**76.2**	48.4	61.2
ve−r+s+aug	**66.8**	**81.7**	41.7	60.4	32.4	48.8	**54.9**	**75.2**	61.9	75.3	**51.5**	68.3

Table 2: Evaluation of our models against each speaker reported in %. All metrics are evaluated on the curated test set made of only real-data. Spk-1 uses a combination of deaf-speech and sign language, Spk-2 and Spk-3 use sign language for communication. Patient denote the ALS patient in our study. r, s, and aug indicate real, synthetic, and augmented-real datapoints. cl indicates standard classification while VE is the proposed technique.

	Experiment	top-1	top-3	top-5
One-shot (classification)	only fs tts	56.83	68.46	71.92
	only glow tts	57.56	67.73	71.55
	combined	**61.38**	**73.63**	**75.26**
Two-shot	classification	64.33	74.69	82.16
	variational encoders	**71.64**	**76.54**	**89.36**

Table 3: Accuracy of additional experiments performed on the ALS patient in %. The test dataset used for both, One-shot and Two-shot experiments is the same. FastSpeech2 and GlowTTS are represented by fs tts and glow tts, respectively. All experiments are conducted using LRwP as the backbone on the combined real, synthetic, and augmented real datasets.

Data Augmentation using Synthetic Data (Exp-*cl−r+s*): We augment the one-shot examples with a potentially unlimited number of synthetic examples. We train our models with varying amounts of synthetic data. We plot a graph of accuracy against the combined synthetic and one-shot examples to determine the optimal amount of synthetic data needed for each speaker against the one-shot examples as shown in the supplementary, Fig. 1. The optimal number of synthetic examples per speaker is reported in Table 1.

As shown in Table 2 Exp-*cl−r+s*, we observe a significant jump in the accuracy consistently for every speaker with an overall improvement of 9% and 7% at top-1 and 5, respectively on LRwP, and 7% and 12% at top-1 and 5, respectively on LTCN. Although the accuracy improved over the baseline model, we observe that the model overfits on the synthetic data after a few epochs. Since we use only one example per class for the real domain, the model cannot foresee the target (real) variations that it may encounter during testing. We, therefore, try to introduce variations using Variational Encoders.

Variational Encoders (Exp-*ve−r+s*): To train the network, we use the one-shot examples as the source domain. For the target domain, we combine the real and synthetic dataset.
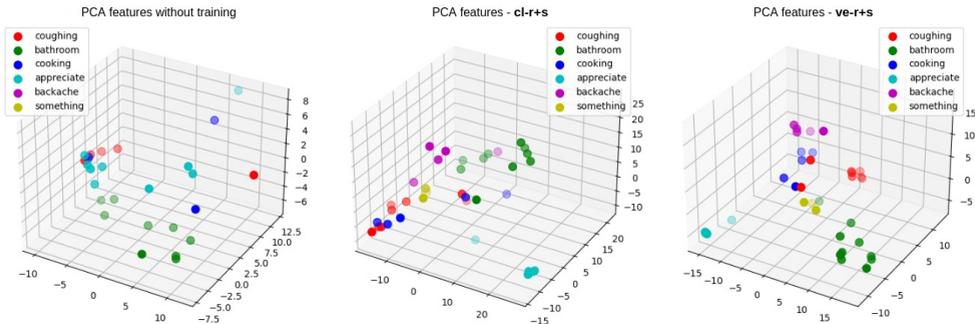
Figure 5: PCA visualization for the embeddings generated by the feature extraction layer of LRwP based encoder. (left) before training, (middle) trained on $cl-r+s$, and (right) trained on $ve-r+s$. Samples from the ALS patient's test set are used for visualization

This lets the source encoder see the existing real examples and also become robust against the additional pseudo examples sampled from the target distribution. We observed that using only the synthetic data in the source encoder makes the training highly unstable and the network does not converge. Instead, we allow the source encoder to first fit and then improve.

As seen from Table 2 Exp-$ve-r+s$, Variational Encoders consistently achieves the highest accuracy at top-5 across LRwP and LTCN. There is an overall improvement of $\sim 19\%$ over the baseline and $\sim 12\%$ over Exp-$cl-r+s$. The improvement at top-1 accuracy is comparatively much lower, $\sim 15\%$ over the baseline and only $\sim 6\%$ over Exp-$cl-r+s$. We use PCA visualizations to analyze the latent representations learnt by the LRwP based source encoder, as shown in Fig. 5 (right). We observe better separation for each class compared to the separation of Exp-$cl-r+s$ (middle). We also observe better disentanglement between non-homophenes such as 'bathroom', 'appreciate', and 'coughing' compared to Exp-$cl-r+s$. For Varitional Encoders we observe that homophenes 'coughing', 'cooking', 'something' are closer together. This suggests that the model trained on Variational Encoders can get confused for homophenes bringing the accuracy at top-1 down and at top-5 higher. For Exp-$cl-r+s$ however, all the classes seem equally apart.

Ad-hoc Data Augmentation (Exp-$cl-r+s+aug$ and Exp-$ve-r+s+aug$): In addition to the implicit variations introduced by Variational Encoders during the model training, we introduce explicit real-domain variations by augmenting the one-shot dataset. First, we use moviepy [⬚] library to speed up and speed down the one-shot videos by a factor of $1.2\times$ and $0.8\times$. The videos in our work are unconstrained, that is, the actual mouthing could be spread across several frames placed temporally anywhere in the video. Thus, we increase the video frames sequence length to 85 and add temporal variations during training by padding the videos with random number (between 0 to 20) of empty frames at the start and end.

We observe improvement in the performance across speakers for both, classification (Exp-$cl-r+s+aug$) and Varitional Encoders (Exp-$ve-r+s+aug$). We observe that for Spk-1 and Spk-2, the performance degrades. Upon further analysis, we find that the videos maintain constant pace and have fewer overall variations. Thus, the added variations during training behave as noise driving the performance down. The overall performance improves by 7% and 3% at top-1 on Exp-$cl-r+s+aug$ and Exp-$ve-r+s+aug$ over Exp-$cl-r+s$ and Exp-$ve-r+s$ respectively. We observe the best performance on one-shot setting with these ad-hoc

additions for classification. For Variational Encoders, the improvements are less significant, especially at top-5, indicating the technique itself makes up for these ad-hoc additions.

Additional Experiments on the ALS Patient's Dataset: To observe the affect of using different TTS models for generating synthetic data, we perform an ablation by eliminating one TTS model at a time for data generation and compare its performance against the data generated by combining both the TTS models. As shown in Table 3, the performance with both the TTS combined gives us the best performance. This indicates that the variations introduced by different TTS models are important for generalization.

Lastly, we train an additional model with the additional data obtained from the patient's family to evaluate the performance difference between one-shot and two-shot setting. As seen from Table 3, the performance improves by $\sim 7\%$ at top-5 against the best performing one-shot model on both training methodologies. Specifically for Variational Encoders, the performance improves to 71.64% and 89.36% at top-1 and top-5. This suggests that the network learns a better real-domain latent distribution using the extra real examples.

Evaluation on Joe Biden: Although Joe Biden represents speakers without disabilities, he speaks in an American accent, while LRW is composed of British speakers. As can be observed from Table 2 Exp-$cl-r$, the performance for Joe Biden on the baseline model is similar to the performance for the deaf speakers. A different accent can also be thought of as a different style of speaking. Thus the pretrained model does not directly adapt to Joe Biden. We observe in his case, adding synthetic data leads to significant improvements in the accuracy over the baseline with an average gain of 21% at top-1 and top-5 (Exp-$cl-r+s$). Variational Encoders fail to bring expected improvements compared to the other speakers. With LRwP, at top-1, the accuracy drops by 3%, and we see a marginal improvement at top-5. We note that the TTS used to generate synthetic data is of an American accent. This suggests that the synthetic data captures the variations of the real domain exceptionally well in his case. Variational Encoders, on the other hand, adds more noise than valuable variations.

4   Conclusion and Future Work

In this work, we leap from previous lipreading approaches and propose a one-shot personalized lipreading framework to aid patients suffering from ALS or hearing disabilities. Due to the extreme scarcity of personalized data available for a medical patient, we generate synthetic data to augment our training process. We train our network with Variational Encoders, a domain adaptation technique, to bridge the gap between the synthetically generated examples and the available one-shot real examples for each class. Our method proves to be highly effective, and we achieve over 83% top-5 accuracy for the ALS patient. We also report the performance of speaker-specific models trained for multiple speakers with hearing impairment and a speaker with no disability. In the future, we would like to improve the accuracy of our model while also increasing the vocabulary it can handle. We believe our work achieves essential milestones for the lipreading community and can enhance the communication between people dependent on lipreading. It encourages research in the direction of few-shot lipreading that is an important real-world challenge that can have far-reaching applications including and beyond medical lipreading. Although, Variational Encoders has been proposed for few-shot lipreading, its utility can be explored in other areas with a few-shot settings.

References

[1] Creating and exporting video clips¶. URL https://zulko.github.io/moviepy/getting_started/videoclips.html.

[2] Reasons for tiredness in deaf children: Reasons for fatigue. URL https://www.ndcs.org.uk/information-and-support/parenting-and-family-life/parenting-a-deaf-child/tiredness-in-deaf-children.

[3] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. In *arXiv:1809.02108*, 2018.

[4] T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. In *arXiv preprint arXiv:1809.00496*, 2018.

[5] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2019. ISSN 1939-3539. doi: 10.1109/tpami.2018.2889052. URL http://dx.doi.org/10.1109/TPAMI.2018.2889052.

[6] David Beukelman, Susan Fager, and Amy Nordness. Communication support for people with ALS. *Neurology Research International*, 2011:1–6, 2011. doi: 10.1155/2011/714693. URL https://doi.org/10.1155/2011/714693.

[7] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2016.

[8] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[9] Nanqing Dong and Eric Xing. Domain adaption in one-shot learning. pages 573–588, 01 2019. ISBN 978-3-030-10924-0. doi: 10.1007/978-3-030-10925-7_35.

[10] Dalu Feng, Shuang Yang, Shiguang Shan, and Xilin Chen. Learn an effective lip reading model without pains, 2020.

[11] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification¶. *Neurocomputing*, 321:321–331, Dec 2018. ISSN 0925-2312. doi: 10.1016/j.neucom.2018.09.013. URL http://dx.doi.org/10.1016/j.neucom.2018.09.013.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks, 2016.

[13] Prashnna Kumar Gyawali, Sandesh Ghimire, Pradeep Bajracharya, Zhiyuan Li, and Linwei Wang. Semi-supervised medical image classification with global latent mixing, 2020.

[14] Chris Hopwood. Concentration fatigue, Feb 2021. URL https://deafblind.org.uk/concentration-fatigue/.

[15] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation, 2017.

[16] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search, 2020.

[17] Davis E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10: 1755–1758, December 2009. ISSN 1532-4435.

[18] Tanja Makkonen, Hanna Ruottinen, Riitta Puhto, Mika Helminen, and Johanna Palmio. Speech deterioration in amyotrophic lateral sclerosis (ALS) after manifestation of bulbar symptoms. *International Journal of Language & Communication Disorders*, 53(2):385–392, November 2017. doi: 10.1111/1460-6984.12357. URL https://doi.org/10.1111/1460-6984.12357.

[19] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks, 2020.

[20] P. Masrori and P. Van Damme. Amyotrophic lateral sclerosis: a clinical review. *European Journal of Neurology*, 27(10):1918–1929, July 2020. doi: 10.1111/ene.14393. URL https://doi.org/10.1111/ene.14393.

[21] Saeid Motiian, Quinn Jones, Seyed Mehdi Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation, 2017.

[22] K R Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C V Jawahar. Learning individual speaking styles for accurate lip to speech synthesis, 2020.

[23] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 484–492, 2020. doi: 10.1145/3394171.3413532.

[24] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech, 2021.

[25] Lawrence D. Rosenblum, Rachel M. Miller, and Kauyumari Sanchez. Lip-read me now, hear me better later. *Psychological Science*, 18(5):392–396, May 2007. doi: 10.1111/j.1467-9280.2007.01911.x. URL https://doi.org/10.1111/j.1467-9280.2007.01911.x.

[26] Nima Tajbakhsh, Yufei Hu, Junli Cao, Xingjian Yan, Yi Xiao, Yong Lu, Jianming Liang, Demetri Terzopoulos, and Xiaowei Ding. Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data, 2019.

[27] Ryuhei Takahashi, Atsushi Hashimoto, Motoharu Sonogashira, and Masaaki Iiyama. Partially-shared variational auto-encoders for unsupervised domain adaptation with target shift, 2020.

[28] Amirsina Torfi, Seyed Mehdi Iranmanesh, Nasser Nasrabadi, and Jeremy Dawson. 3d convolutional neural networks for cross audio-visual matching recognition. *IEEE Access*, 5:22081–22091, 2017. ISSN 2169-3536. doi: 10.1109/access.2017.2761539. URL http://dx.doi.org/10.1109/ACCESS.2017.2761539.

[29] Zerrin Turan. Deaf children with additional disabilities: Description and research. 2015:148–151, 01 2015.

[30] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation, 2017.

[31] Kai Xu, Dawei Li, Nick Cassimatis, and Xiaolong Wang. Lcanet: End-to-end lipreading with cascaded attention-ctc, 2018.

[32] Sara Zarei, Karen Carr, Luz Reiley, Kelvin Diaz, Orleiquis Guerra, PabloFernandez Altamirano, Wilfredo Pagani, Daud Lodin, Gloria Orozco, and Angel Chinea. A comprehensive review of amyotrophic lateral sclerosis. *Surgical Neurology International*, 6(1):171, 2015. doi: 10.4103/2152-7806.169561. URL https://doi.org/10.4103/2152-7806.169561.