Towards MOOCs for Lip Reading: Using Synthetic Talking Heads to Train Humans in Lipreading at Scale

Anonymous WACV 2023 APPLICATIONS TRACK submission

Paper ID 720

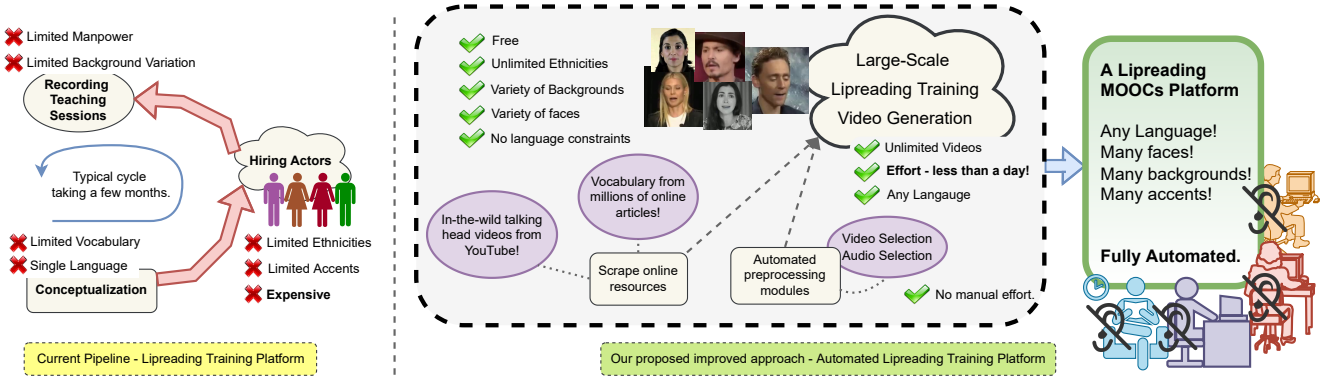

Figure 1: Lipreading is a primary mode of communication for people with hearing loss. The United States of America alone is home to 48 million people who are dealing with some form of hearing loss. Despite these staggering stats, online resources for lipreading training is scarce and are available for only a handful of languages. However, hosting new lipreading training platforms is an extensive ordeal that can take months of manual effort. We propose a fully-automated approach towards building large-scale lipreading training platforms. Our approach enables any language, any accent, and unlimited vocabulary on any identity! We envision a lipreading MOOCs platform to enable millions of people with hearing loss across the globe. In this work, we provide a thorough analysis of the viability of such an approach.

Abstract

Many people with some form of hearing loss consider lipreading as their primary mode of day-to-day communication. However, finding resources to learn or improve one's lipreading skills can be challenging. This is further exacerbated in COVID19 pandemic due to restrictions on direct interactions with peers and speech therapists. Today, online MOOCs platforms like Coursera and Udemy have become the most effective form of training for many kinds of skill development. However, online lipreading resources are scarce as creating such resources is an extensive process needing months of manual effort to record hired actors. Because of the manual pipeline, such platforms are also limited in the vocabulary, supported languages, accents, and speakers, and have a high usage cost. In this work, we investigate the possibility of replacing real human talking videos with synthetically generated videos. Synthetic data can be used to easily incorporate larger vocabularies, variations in accent, and even local languages, and many speakers. We propose an end-to-end automated pipeline to develop such a platform using state-of-the-art talking heading video generator networks, text-to-speech models, and computer vision techniques. We then perform an extensive human evaluation using carefully thought out lipreading exercises to validate the quality of our designed platform against the existing lipreading platforms. Our studies concretely point towards the potential of our approach for the development of a large-scale lipreading MOOCs platform that can impact millions of people with hearing loss.

1. Introduction

Communication is a crucial ingredient that makes Humans the most intelligent species on the planet. While other animals also have different forms of communication, human language is more advanced in several orders of magnitudes. But, we are not inherently born with these skills! Then, how do we acquire them? Most of us learn linguis-

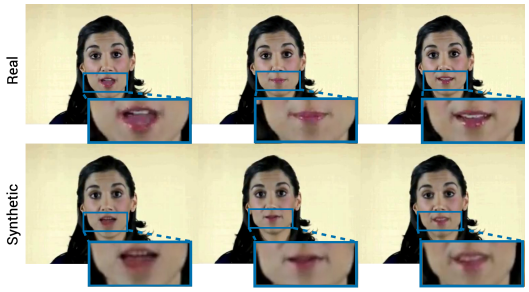

Figure 2: Talking-face video generated using our pipeline.

tic skills through a formal education system consisting of schools, universities, and other organizations related to education. While this is still the most trusted & popular way of imparting education, the 21st century has seen an exponential rise in online forms of education like the Massive Open Online Courses (MOOCs). Online courses are generally designed to cover hundreds of topics in various domains, including language, and are often available free of cost. MOOCs have several advantages over the physical form of education. They are more accessible, cheap, and reachable to a broader audience. In today's world, it is quite natural to learn a whole new language from the comfort of your home by attending a high-quality MOOCs course.

Unfortunately, every person does not get the chance to learn linguistic skills like we usually do. Hearing loss is a very common form of disability that can become a massive barrier to education! According to organizations like WHO[1] and Washington Post[2], over 5% of the world's population (432 million adults and 34 million children) and at least 48 million Americans are deaf with some form of hearing loss. About $500,000$ Americans have a disabling hearing loss that noticeably disrupts communication.

Lipreading is a primary mode of communication for people with hearing loss. The Scottish Sensory Censor (SSC)[3] quotes "whatever the type or level of hearing loss, a child is going to need to lipread some of the time". However, learning to lipread is not an easy task! Lipreading can be thought of being analogous to "learning a new language" for people without hearing disabilities. People needing this skill undergo formal education in special schools and involve medically trained speech therapists. Other resources like daily interactions also help understand and decipher language solely from lip movements. However, these resources are highly constrained and inadequate for the large number of patients suffering from hearing disabilities.

Inspired by the boom in online courses available for virtually every topic, we envision a MOOCs platform for **L**ip **R**eading **T**raining (LRT) for the hearing disabled.

[1] Deafness and Hearing Loss | WHO
[2] As wearing masks becomes the norm, lip readers are left out!
[3] Factors which help or hinder lipreading | SSC

Current Online Platforms for Lip Reading Training Platforms like lipreading.org[4] and lipreadingpractice[5] provide basic online resources to improve lipreading skills. These platforms allow users to learn limited levels of lip reading constrained by resources. Unfortunately, the amount of vocabulary systematically covered during the exercises is extremely narrow. The videos also have minimal real-world variations in head-pose, camera angle, and distance to a speaker, making it difficult for a lipreader to adapt to the real world. Finally, since these resources are all available only in American or British-accented English, it becomes challenging for people from other regions to adapt to their local accents and languages. All the above factors severely limit the quality of human training. Therefore, we believe it is quintessential to scale the current lipreading training platforms to incorporate extensive vocabulary and introduce variation in videos, languages, and accents. However, recording videos is a costly affair. It requires expensive camera equipment, studio environments, professional editors, and a substantial manual effort from the perspective of a speaker whose videos are being recorded.

To resolve this issue, we approach this from a different angle and ask: "Can we replace real talking head videos used for training people suffering from hearing loss with synthetic versions of the same?" A synthetic talking head with accurate lip synchronization to a given text or speech signal can enable the scaling of LRT platforms to more identities, accents, languages, speed of speech, etc., making the training process more rigorous. We take advantage of the massive progress made by the computer vision community on synthetic talking head generation and employ a state-of-the-art (SOTA) algorithm [23] as mentioned below.

We propose a novel approach that can be used to automatically generate a large-scale database for the development of a LRT MOOCs platform. We use SOTA text-to-speech (TTS) models [7] and talking head generators like Wav2Lip [23] to generate training examples automatically. Wav2Lip [23] requires driving face videos and driving speech segments (generated from the TTS in our case) to generate lip-synced talking head videos according to the driving speech. It preserves the head-pose, background, identity, and distance of the person from the camera while modifying only the lip movements as shown in Fig. 2.

Our approach can lead to an exponential increase in the amount of online content present on the LRT platforms in an automated and cost-effective manner. It can also seamlessly increase the vocabulary and the number of speakers in the database. We investigate the implications of our system for a range of deaf users and perform multiple experiments to show its effectiveness in replacing the current manually recorded LRT videos. We show through statistical analysis

[4] lipreading.org
[5] lipreadingpractice.co.uk

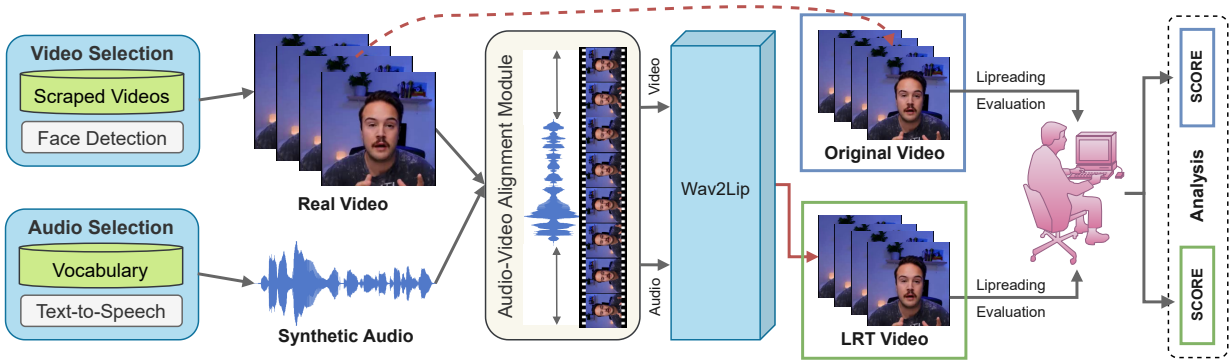

Figure 3: Proposed pipeline for generating large-scale lipreading training platform: **(a) Video Selection**: Videos are scraped from various online sources (such as YouTube) and invalid videos are filtered out. **(b) Audio Selection:** Synthetic speech utterances are generated using vocabulary curated from various online articles. **(c) Audio-Visual Alignment Module**: A video and a speech utterance is selected and aligned on each other such that the speech utterance overlaps with the region in the video with lip movements. **(d) Wav2Lip:** A state-of-the-art talking head generation model that modifies the lip movements of the video according to the speech utterance. **(e) User Evaluation:** A validation step to ensure that users perform comparably on real videos and synthetic videos generated using our approach.

that (1) the performance of the users on lipreading videos is not significantly different when switching from 'real' to 'generated' videos, and (2) the benefit of lipreading platforms in one's native accent through an extensive user study. We believe our approach towards generating fully synthetic videos is the first step towards developing a LRT MOOCs platform to benifit millions of users with hearing loss.

2. Related Work

The usefulness of MOOCs as a medium of education has been accepted [24] worldwide. Surveys like [11] analyze various aspects of the impact of MOOCs and helps us understand their positives and negatives. MOOCs are shown to increase the audience and offer viable alternatives to the traditional form of education in [19]. The increasing demand for content has also led to improvement in student engagement [10, 15]. The requirement for MOOCs and other forms of online education has skyrocketed since the beginning of the COVID19 pandemic[6]. We believe this trend to continue and impact different types of education required by people with special needs. Our work also aligns with assistive technology where Digital media has historically played an important role. Much of these efforts have been invested in improving the communication skills of certain groups. In 2006, [22] published their work on "Baldi", a computer-animated tutor to teach children with autism. Following this, another work [6] has focused on generating 3D animated tutors for autism-affected children to improve their communication skills. Research aimed at improving the communication skills of the hearing impaired is also popular. [4] developed a computer-assisted vocabulary for educating the deaf to communicate orally. Special courses [20] are designed to help people with limited hearing abilities. Human-computer interaction interfaces [5, 1] targeted for similar groups are also prevalent. Recently a landmark work [9] targeted to create a home assistant for people hard of hearing. The main focus of their work was to incorporate sign language based commands (replacing speech commands) into a personal assistant. Similar efforts were made for automatic lipreading in [25, 21].

[6]The rise of online learning during the COVID-19 pandemic

3. Synthetic Talking Head Database

Our lipreading training database generation pipeline: (1) Scrapes a set of face videos automatically from the internet. This helps us in covering a large number of identities, background variations, lip shapes, etc. (2) Post-processes the scraped videos to filter out invalid faces (such as drastic pose change) (3) Automatically curates a vocabulary made of many words and sentences from various online sources. (4) Generates synthetic speech utterances on the curated vocabulary. (5) Selects a driving face video and a speech utterance to generate synthetic talking head videos using a SOTA talking head generation model, Wav2Lip in our case. Wav2Lip modifies the lip movements of the driving video according to the speech utterance. The rest of the video (background, pose, etc.) is retained. These synthetic videos (with or without the speech) are used to train humans in lipreading. The overall pipeline is illustrated in Fig. 3.

Text-to-Speech System We evaluate several TTS models: Fastspeech2 [7], Real time voice cloning [13], Glow-tts [16], and Tacotron2 [26] trained on LibriTTS [28] and LJSpeech [12]. We evaluate them on different speeds - 1×, 1.5×, 1.7×, 2×, pitch, and volume variations. We collect qualitative feedback from 30 participants without any

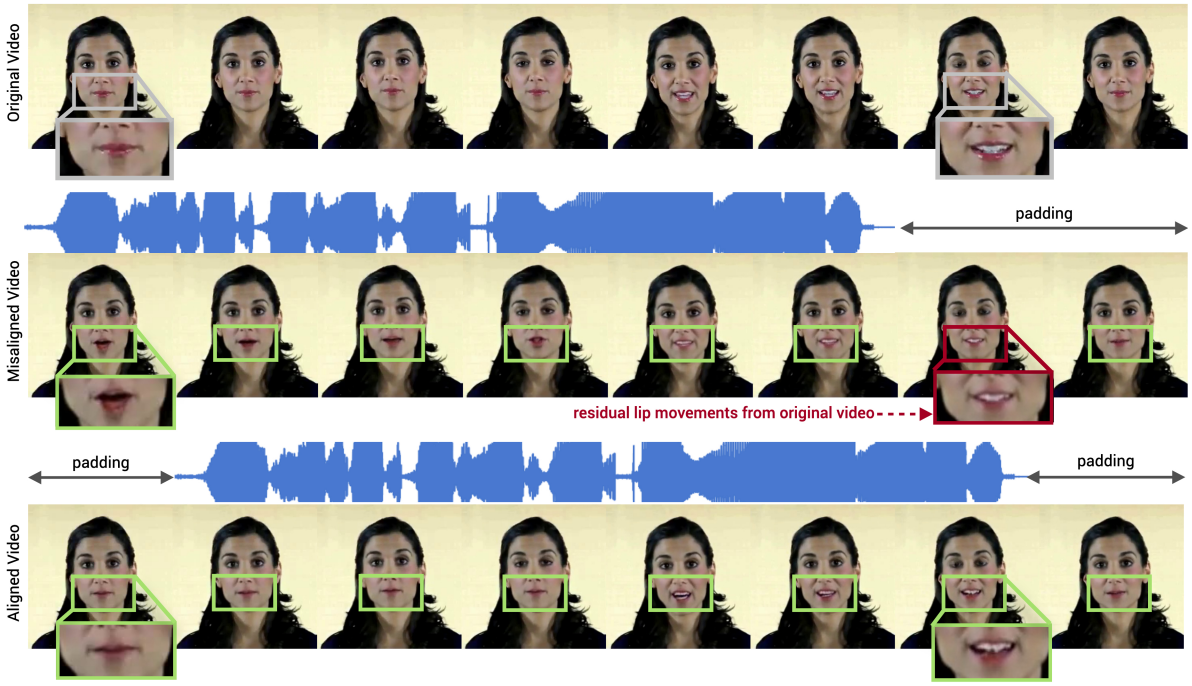

Figure 4: Audio-Video Alignment Module: Lip-sync models such as Wav2Lip modify the lip movements of an 'Original Video' (driving video) according to a given speech utterance. However, naively aligning the audio and video before passing through Wav2Lip can result in a 'Misaligned Video' with residual lip movements as indicated in red-boxes. We design an audio-video alignment module that detects the mouth movements in the original video. We then align the speech utterance on the region with the mouth movements and add silence around the aligned utterance. Wav2Lip then generates an 'Aligned Video' without any residual lip movements as indicated in green boxes.

hearing loss on the clarity in speech of the generated audios and report the Mean Opinion Scores (MOS) in the supplementary. For the purpose of our experiments conducted on American-accented English, we use Fastspeech2 with 1× speed configuration pretrained on LJSpeech. For Indianised English accent, we use an online TTS at[7] with qualitatively similar performance to the speech generated by FastSpeech2. Please note that the TTS models used in our pipeline are configurable plug-and-play modules and can be easily replaced with any other TTS. This allows scalability and variations with little to no manual effort.

Synthetic Talking Head Videos Since 2015, talking head generation models, that modify the lip movements according to a given speech utterance, has gained much traction in the computer vision community [18, 8, 27]. While some of these works generate accurate lip-sync, they are trained for specific speakers requiring large amounts of speaker-specific data. [2] can be remodeled for generating talking heads but require far more manual intervention limiting their use in our approach. Recent advances like LipGAN [14] and Wav2Lip [23] are perfect for our approach since they work for any identity without requiring any speaker-specific data. Consequently, we adopt Wav2Lip in our pipeline. Wav2Lip takes a face video of any identity (driving face video) and an audio (guiding speech) as inputs. The model then modifies the lip movements in the original video to match the guiding speech. Rest of the video features such as the background, identity, pose, of the face is preserved. We use Wav2Lip for generating the synthetic data because of its ability to generate highly accurate lip-synced talking head videos as shown in Fig. 2 on any language and voice. The algorithm also works for synthetic TTS-generated speech segments essential in our case.

[7] http://ivr.indiantts.co.in/en/

3.1. Data Generation Pipeline

Data Collection Module: Random videos are first collected from various online sources such as YouTube. These random videos introduce real-world variations a lipreader encounters in the real life, such as, the variations in head-pose of the speaker, speaker's distance from the camera (lipreader), speaker's complexion, and lip structure. We post-process these videos with a face-detection model to detect valid videos. Valid videos are single-identity front-facing talking head videos with no drastic pose changes. Speech utterances are generated using TTS models on vocabulary curated automatically from online sources.

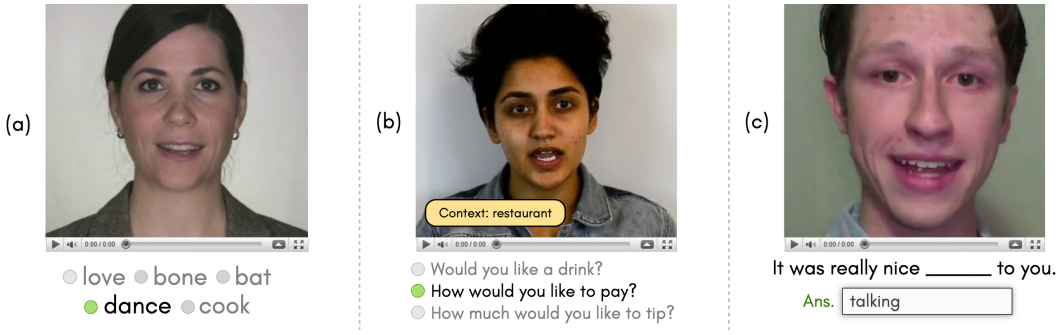

Figure 5: Examples of different protocols used for our user study. (a) lipreading isolated words (WL): the speaker mouths a single word and the user is expected to select one of the multiple choices presented. (b) lipreading sentences with context (SL): the speaker mouths an entire sentence. The user is presented with the context of the sentence and is expected to select one of the sentences present in multiple choices, and (c) lipreading missing words in sentence (MWIS): the speaker mouths an entire sentence. The user is presented with a sentence with blanks (masked words), the user needs to identify the masked word from the video and sentence context and answer in text format.

Audio-Video Alignment Module: In our next step, we randomly select a pair of driving speech and a face video. To generate lip-synced videos using Wav2Lip, we match the video and speech utterance length by first aligning them and then padding the speech utterance with silence. Naively aligning the speech utterance on the driving video can lead to residual lip movements, as shown in Fig. 4, 'Misaligned Video' row. Wav2Lip does not modify the lip movements in the driving video in the silent region. As a result, the output contains residual lip movements (indicated in the red box) from the original video. This can confuse and cause distress to the user learning to lipread. Our audio-video alignment module aligns the speech utterance on the video region with lip movements. This way, Wav2Lip naturally modifies the original mouth movements to correct speech-synced mouth movements, while keeping the regions with no mouth movements untouched. Here, we ensure that the silent areas of speech coincide with parts in the driving video with no lip movements as shown in Fig. 4, 'Aligned Video' row. We use lip-landmarks and the rate of change of the lip-landmarks between a predefined threshold of frames to detect mouth movements in the face videos. Once we have detected lip movements, we align the audio on the detected video region and add silences around the speech.

Data Generation: The aligned speech utterance and the face video are passed through Wav2Lip. Wav2Lip modifies the lip movements in the original video and preserves the original head movements, background, and camera variations, thus allowing us to create realistic-looking synthetic videos in the wild. Overall pipeline is illustrated in Fig. 3.

4. Human Lipreading Training

Lipreading is an involved process of recognizing speech from visual cues - the shape formed by the lips, teeth, and tongue. A lipreader may also rely on several other factors, such as the context of the conversation, familiarity with the speaker, vocabulary, and accent. Thus, taking inspiration from lipreading.org and readourlips.ca[8], we define three lipreading protocols for conducting a user study to evaluate the viability of our platform - (1) lipreading on isolated words, (2) lipreading sentences with context, and (3) lipreading missing words in sentences. These protocols rely on a lipreader's vocabulary and the role that semantic context plays in a person's ability to lipread.

4.1. Lipreading on isolated Words (WL)

The ability to disambiguate different words through visual lip movements helps shape auditory perception and speech production. In **w**ord-**l**evel (WL) lipreading, the user is presented with a video of an isolated word being spoken by a talking head along with multiple choices and one correct answer. When a video is played on the screen, the user is required to respond by selecting a single answer from the provided multiple choices. Visually similar words (homophenes) are placed as options in the multiple choices to increase the difficulty of the task. The difficulty can be further increased by testing for difficult words - difficulty associated with the word to lipread, e.g., uncommon words are harder to lipread. For the purpose of our study, we test the users only on the commonly known words. The multiple answer choices have been fixed to 5 options. An example of word-level lipreading is shown in Fig. 5 (a).

4.2. Lipreading Sentences with Context (SL)

In **s**entence-**l**evel (SL) lipreading, the users are presented with (1) videos of talking heads speaking entire sentences and (2) the context of the sentences. The context acts as an additional cue to the mouthing of sentences and is meant to simulate real conversations in a given context. Accord-

[8] https://www.readourlips.ca/

ing to [3] the context of the sentences can improve a person's lipreading skills. Context narrows down the vocabulary and also helps in the disambiguation of different words. We evaluate our users on two contexts - A) Introduction - 'how are you?', 'what is your name?', and B) Lipreading in a restaurant - 'what would you like to order?'. Like WL lipreading, we provide the user with a fixed number of multiple choices and one correct answer. Apart from context, no other information is provided to the participants regarding the length or semantics of the sentence. Fig. 5 (b) shows an example of sentence-level lipreading with context.

4.3. Lipreading missing words in sentences (MWIS)

According to[9], an expert lipreader can discern only 40% of a given sentence or $4-5$ words in a 12 words long sentence. In this protocol, we try to emulate such an experience by **m**asking **w**ords **i**n the **s**entence (MWIS). The participants watch videos of sentences being spoken by a talking head with a word in the sentence masked as demonstrated in Fig. 5 (c). Unlike lipreading sentences mentioned in Sec. 4.2, the users are not provided with any additional context of the sentence. Lip movements are an ambiguous source of information due to the presence of homophenes, thus, this exercise aims to use the context of the sentence to disambiguate between multiple possibilities and guess the correct answer. For instance, given the masked sentence "a cat sits on the {masked}", a lipreader can disambiguate between homophenes 'mat', 'bat', and 'pat' using the sentence context to select 'mat'. The user is required to enter the input in text format for the masked word as shown in Fig. 5 (c). Minor spelling mistakes are accepted.

5. User Study

In this section, we explain the collective background of our participants, the types of videos used for the study, and the design of our testing platform.

5.1. Participants

We perform our study on 50 participants with varying degrees of hearing loss with 32 male and 18 female participants. The average age of the participants in this study is 35 years, ranging from 29 years to 50 years. Participants in this study reside in the Indian states of Maharashtra and Rajasthan. 29 participants have a Master's degree while the remaining 21 have a Bachelor's degree. All the participants in the study report having sensorineural hearing loss[10] and use hearing aids in their daily life along with lipreading and oral deaf speech as their primary mode of communication.

[9] Speech Reading, Hearing Loss in Children | CDC
[10] What is Sensorineural Hearing Loss?

Task	Real	Synthetic	
	American	American	Indian
WL	80	800	800
SL	60	600	600
MWIS	70	700	700
Total	210	2100	2100

Table 1: No. of examples curated for each protocol in different English accents (American / Indian).

5.2. Dataset

We scrape real videos from lipreading.org and generate our synthetic videos on them. Lipreading.org videos allow us to (i) make a direct comparison between the real lipreading training videos and our synthetically generated videos and (ii) lipreading.org provides the correct answer of the video; this provides the correct ground truth label for the real videos later used for quantitative analysis.

Primarily, we aim to compare a user's performance on the synthetic videos generated using our proposed pipeline against the real videos on lipreading.org. We use the three protocols explained in Sec. 4 for this purpose. Our synthetic videos are divided into: (1) non-native **A**merican-accented **E**nglish (AE) videos and (2) native **I**ndian-accented **E**nglish (IE) videos. Our users are of Indian origin.

We create our synthetic dataset using 10 driving videos on 5 speakers. For WL lipreading protocol, we scrape 80 labels from lipreading.org's single-word lipreading quiz. Using these we generate $80 \times 10 = 800$ talking head videos - 10 variations per word. For SL lipreading, we scrape 60 questions from lipreading.org's sentence-level quiz across two contexts: introductions and lipreading in a restaurant. Using these sentences, we generate $60 \times 10 = 600$ talking head videos - 10 variations for each sentence. Lastly, we scrape 70 sentences from lipreading.org's missing words in sentences task and generate $70 \times 10 = 700$ talking head videos for the MWIS protocol. We generate these videos once using American-accented TTS and the second time using Indian-accented TTS. As shown in Table 1, we generate a total of 4200 synthetic videos and collect 210 real videos from lipreading.org across the protocols.

5.3. Test Design

Our primary goal is to validate that the synthetic talking head videos generated using our pipeline can replace real videos in terms of visual quality and ease to discern.

Each participant participates in all 3 protocols. For each protocol, the user takes 3 quizzes corresponding three datasets: (1) Real AE, (2) **Synth**etic **AE** (Synth AE), and (3) **Synth**etic **IE** (Synth IE). In total, user attempts 9 quizzes. Quizzes are delivered through a web-based platform that we developed. Our users report taking the quizzes from a plethora of personal devices like PCs, laptops, Android and

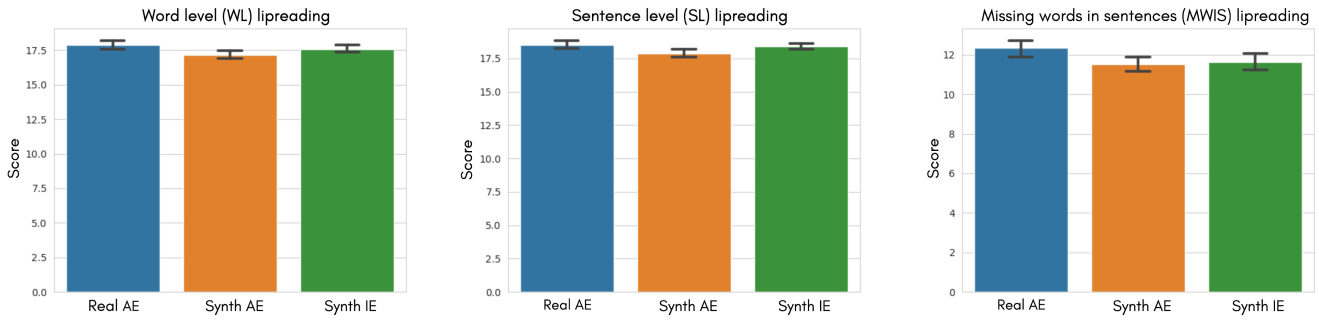

Figure 6: Mean user performance on the three lipreading protocols. Error bars are the standard errors of the mean.

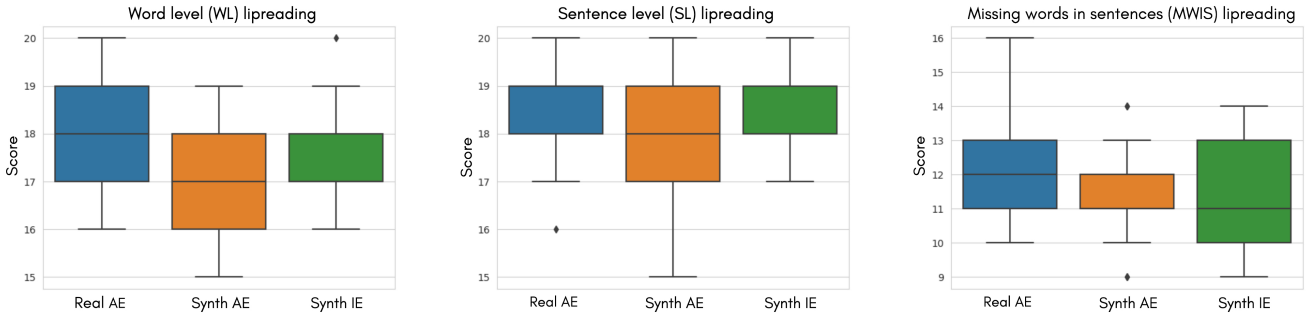

Figure 7: Box plots depicting the distribution of scores on the three lipreading protocols. Horizontal lines within the rectangles represent median scores, and the top and bottoms of the rectangles correspond to the first and third quartiles; the horizontal lines at the ends of the vertical "whiskers" represent the minimum and maximum scores, and the diamonds represent scores outside this range.

iPhone mobile devices, and tablets. The number of days taken to complete a test is left at the user's discretion to prevent the user from feeling fatigued as lipreading is an involved process and can be mentally taxing. The longest time taken by any user to complete our test is four days.

For each quiz, the user is presented with 20 questions/videos. For each question, a word/sentences is first randomly sampled from the database. One of the 10 variations of the sampled word/sentence present in the database is then randomly chosen. The audio is removed from the videos before displaying to the users. We ensure that words/sentences are not repeated across the quizzes in a single protocol to prevent bias by familiarization. We also ensure that the difficulty of lipreading across all the datasets and protocols is kept consistent. For each correct attempt, the user is rewarded 1 point and the score is computed out of 20. We expect the user to finish a single test in one sitting. For a fair comparison, we do not inform the user if they are being tested on real or synthetic data.

Quiz demo is provided in the supplementary video.

6. Results and Discussion

In this section, we conduct statistical analysis to verify **(T1)** If the lipreading performance of the users remain comparable across the real and synthetic videos generated using our pipeline. Through this, we will validate the viability of our proposed pipeline as an alternative to the existing online lipreading training platforms. **(T2)** If the users are more comfortable in lipreading in their native accent/language than lipreading in a foreign accent/language. This would validate the need for bootstrapping lipreading training platforms in multiple languages/accents across the globe.

Fig. 6 plots the standard errors of the mean. Fig. 7 presents the boxplot across the three lipreading protocols.

Synthetic videos as a replacement for real videos: To validate **(T1)**, the difference in the user scores across the real and synthetic videos should be statistically insignificant. Since our conclusion depends on the evidence for a null hypothesis (no difference between the categories), just the absence of evidence is not enough to support the hypothesis. Therefore, we perform a Bayesian Equivalence Analysis using the Bayeisian Estimation Supersedes the t-test (BEST) [17] to quantify the evidence in favour of our model. BEST estimates the difference in means between two distribution/groups and yields a probability distribution over the difference. Using this method, we compute (1) the mean credible value as the best guess of the actual difference between the two distributions and (2) the 95% Highest

Density Interval (HDI) as the range where the actual difference is with 95% credibility. For the difference in the two distributions to be statistically significant, the ideal difference in their mean scores should lie outside the 95% HDI.

We report the BEST statistics on Real AE and Synth AE studies for all three lipreading protocols in Table 2. We also report the t statistic and p-value using the standard two-tailed t-test. From Table. 2, it is clear that the BEST statistic lies within the acceptable 95% HDI for all the three protocols indicating that the difference in the scores between the two groups is statistically insignificant. This suggests that our pipeline is a viable alternative to the existing manually curated talking-head videos.

Native vs Non-native accented lipreading: To validate **(T2)**, the difference in the user scores between native and non-native accented English should be statistically significant. Since our participant pool is from India, we compare the user scores on Synth IE and Synth AE. We perform two-sample Z test since our sample size is large (> 30) to validate the statistical significance. To this end, we propose Null Hypothesis **H0**: the difference in the mean scores between Synth IE and Synth AE is statistically insignificant, and consequently the Alternate Hypothesis **H1**: the difference in the mean scores between the Synth IE and Synth AE is statistically significant. We compute the z statistics and report the p-value for the 90% confidence interval (significance value α=0.1) in Table 3 for the three protocols. We observe that the Z test statistic lies outside the 90% critical value accepted range for two tasks, WL and SL, indicating that the difference in their mean values is statistically significant in favor of IE and we reject **H0** in favor of **H1** for these protocols. For MWIS protocol, the p-value is > 0.1 and the z statistic falls within the acceptable 90% confidence interval, indicating that the difference in their mean scores is not statistically significant. Thus, we fail to reject **H0** in this case. The overall results support our claim that lipreading on native-accents makes much difference in the performance of a lipreader and they are more comfortable in lipreading native accents. Moreover, it reinforces the importance of our platform.

The development of lipreading training database for each new accent using real videos is a non-trivial, exhausting, and time-consuming task. Our platform could thus be easily adopted to add any new language/accent as long as a TTS for that language/accent is available.

Discussion: We note that the lipreaders score relatively higher for the SL protocol. The context of the sentence narrows the vocabulary space and helps in the disambiguation of homophenes. MWIS is the most challenging protocol as it involves retrieving the correct word from the user's own memory as opposed to classifying the given multiple choices. It also involves mapping the unmasked word from the sentences to the videos and recognizing the mouthing for missing word. Thus, they score relatively low on MWIS.

	95% HDI	Mean	MGD	t-value	p-value
WL	(-0.254, 1.63)	0.701	0.7059	1.676	0.1034
SL	(-0.226, 1.62)	0.671	0.6471	1.540	0.1333
MWIS	(-0.366, 1.98)	0.793,	0.8235	1.517	0.1390

Table 2: We perform BEST statistical analysis and compute the 95% HDI range of the difference in means of the real and synthetic distributions. Mean is the distribution of means. We also report the p-values and t-values from a standard t-test for comparison.

	p-value	accepted range	z statistic
WL	0.0786	(-1.645 : 1.645)	1.758816
SL	0.0171	(-1.645 : 1.645)	2.384995
MWIS	0.705	(-1.645 : 1.645)	0.378506

Table 3: Two-sample z-test on synthetic Indian-accented English and American-accented English videos. The significance level α is kept at 0.1. The null-hypothesis is rejected if the z statistic falls outside the 90% critical value accepted range. Consequently, the p-value is also less than the significance value α in that case.

As a conclusion of the user study, we present evidence that synthetic videos can potentially replace real videos. We show that the drop in user performance across Real AE and Synth AE is statistically insignificant across all the protocols. We also show that users are more comfortable lipreading in a native accent through paired z-test highlighting the dire need to bootstrap lipreading platforms in multiple languages/accents at scale.

7. CONCLUSION

Lipreading is a widely adopted mode of communication for people with hearing loss. Online resource for lipreading training is however, scarce and limited in many factors such as vocabulary, speakers, languages. Moreover, launching a new platform in a new language is costly that would need months of manual efforts to record training videos on hired actors. In this work, we analyze the viability of using synthetically generated videos as a replacement for the real videos for lipreading training. We propose an end-to-end automated and cost-effective pipeline for generating lipreading videos and carefully design a set of protocols to evaluate the generated videos. We perform statistical analysis to validate that the difference in user performance on real and synthetic lipreading videos is statistically insignificant. We also show the advantage of lipreading in native accents thus highlighting the dire need for lipreading training in many languages and accents. In this vein, we envision a MOOCs platform for training humans in lipreading to potentially impact millions of people with some form of hearing loss across the globe.

References

[1] Deepali Aneja, Daniel McDuff, and Shital Shah. A high-fidelity open embodied avatar with lip syncing and expression capabilities. In *2019 International Conference on Multimodal Interaction*, ICMI '19, page 69–73, New York, NY, USA, 2019. Association for Computing Machinery. 3

[2] Deepali Aneja, Daniel McDuff, and Shital Shah. A high-fidelity open embodied avatar with lip syncing and expression capabilities. In *2019 International Conference on Multimodal Interaction*, ICMI '19, page 69–73, New York, NY, USA, 2019. Association for Computing Machinery. 4

[3] Spehar B, Goebel S, and Tye-Murray N. Effects of context type on lipreading and listening performance and implications for sentence processing. In *J Speech Lang Hear Res*. JJournal of Speech, Language, and Hearing Research (JSLHR), 2015. 6

[4] L. J. Barker. Computer-assisted vocabulary acquisition: The cslu vocabulary tutor in oral-deaf education. *Journal of Deaf Studies and Deaf Education*, 8(2):187–198, 2003. 3

[5] Hans-Heinrich Bothe. Human computer interaction and communication aids for hearing-impaired, deaf and deaf-blind people: Introduction to the special thematic session. In Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer, editors, *Computers Helping People with Special Needs"*, pages 605–608, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. 3

[6] Fei Chen, Lan Wang, Gang Peng, Nan Yan, and Xiaojie Pan. Development and evaluation of a 3-d virtual pronunciation tutor for children with autism spectrum disorders. *PLOS ONE*, 14(1):1–22, 01 2019. 3

[7] Chung-Ming Chien, Jheng-Hao Lin, Chien yu Huang, Po chun Hsu, and Hung yi Lee. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech, 2021. 2, 3

[8] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Trans. Graph.*, 38, 2019. 4

[9] Abraham Glasser, Matthew Watkins, Kira Hart, Sooyeon Lee, and Matt Huenerfauth. Analyzing deaf and hard-of-hearing users' behavior, usage, and interaction with a personal assistant device that understands sign-language input. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. 3

[10] Philip J. Guo, Juho Kim, and Rob Rubin. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, page 41–50, New York, NY, USA, 2014. Association for Computing Machinery. 3

[11] Nor Hafiza Haron and Yusof Hafidzan. The acceptance of mooc in teaching and learning process: A case study at malaysian public university. In Dragan Cvetković, editor, *MOOC (Massive Open Online Courses)*, chapter 4. IntechOpen, Rijeka, 2021. 3

[12] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. 3

[13] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 4485–4495, Red Hook, NY, USA, 2018. Curran Associates Inc. 3

[14] Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1428–1436, New York, NY, USA, 2019. Association for Computing Machinery. 4

[15] Aditya Kamath, Aradhya Biswas, and Vineeth Balasubramanian. A crowdsourced approach to student engagement recognition in e-learning environments. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016. 3

[16] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8067–8077. Curran Associates, Inc., 2020. 3

[17] John Kruschke. Bayesian estimation supersedes the t test. *Journal of experimental psychology. General*, 142, 07 2012. 7

[18] Rithesh Kumar, Jose M. R. Sotelo, K. Kumar, A. D. Brébisson, and Yoshua Bengio. Obamanet: Photo-realistic lip-sync from text. *ArXiv*, abs/1801.01442, 2018. 4

[19] Sarah R. Lambert. Do moocs contribute to student equity and social inclusion? a systematic review 2014–18. *Computers & Education*, 145:103693, 2020. 3

[20] L. Leeson and H. Sheikh. Signall: Developing online and blended deaf studies course content across eu borders. 2009. 3

[21] Pingchuan Ma, Yujiang Wang, Jie Shen, Stavros Petridis, and Maja Pantic. Lip-reading with densely connected temporal convolutional networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2857–2866, January 2021. 3

[22] Dominic W. Massaro and Alexis Bosseler. Read my lips. *Autism*, 10(5):495–510, 2006. 3

[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. 2, 4

[24] Harvard Business Review. Who's benefiting from moocs, and why, Sep 2020. 3

[25] Bipasha Sen, Aditya Agarwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C V Jawahar. Personalized one-shot lipreading for an als patient. *2021 British Machine Vision Conference (BMVC)*, 2021. 3

[26] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, 2018. 3

[27] Xin-Wei Yao, Ohad Fried, K. Fatahalian, and Maneesh Agrawala. Iterative text-based editing of talking-heads using neural retargeting. *ArXiv*, abs/2011.10688, 2020. 4

[28] Heiga Zen, Rob Clark, Ron J. Weiss, Viet Dang, Ye Jia, Yonghui Wu, Yu Zhang, and Zhifeng Chen. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech*, 2019. 3