LipMOOC: A Modular Synthetic MOOC for Lipreading

Snigdha Agarwal[1[0009-0008-7262-5022]], Vinay Namboodiri[2[0000-0001-5262-9722]], and C.V. Jawahar[1[0000-0001-6767-7057]]

[1] International Institute of Information Technology Hyderabad, Telangana 500032, India snigdha.a@research.iiit.ac.in, jawahar@iiit.ac.in
[2] University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom vpn@bath.ac.uk

Abstract. For people with hearing impairments, lipreading is a crucial skill, as it bridges the gap between spoken language and understanding, improving the ability to communicate effectively. However, the creation of lipreading resources is pervasive and labor-intensive, leading to a scarcity of high-quality, structured, and widely accessible educational materials. To address this, we developed LipMOOC, a Massive Open Online Course (MOOC) tailored for lipreading education. Our approach leverages synthetically generated content, reducing the dependence on human-recorded videos and building an accessible and effective learning option. Using the state-of-the-art models, we created a comprehensive lipreading course and conducted a study to evaluate its effectiveness. We statistically proved that (1) Synthetic videos can replace human-recorded videos without affecting the learning outcomes, and (2) Learner performance on a structured course, i.e., LipMOOC, is much better than a standard course. A total of 120 participants with hearing impairment and no prior lipreading experience volunteered. These results highlight the potential of AI-driven content generation and scientific instructional design.

Keywords: MOOC · AI Teaching · Inclusion · Learning Pedagogy

1 Introduction

According to the World Health Organization (WHO) [9], more than 430 million people worldwide experience hearing loss, a number expected to rise to over 700 million by 2050. Despite its importance, lipreading education still faces significant challenges. There are platforms like *lipreading.org* [3] and *lipreadingpractice* [4] which provide basic online resources for lipreading. However, most available learning materials require a fee, are unstructured, or do not include the diversity of speech patterns, accents, and ethnic representations required for thorough learning. To overcome these constraints, we created an AIED system that can

[3] https://www.lipreading.org/
[4] https://lipreadingpractice.co.uk/

offer structured, scalable, and accessible learning opportunities. Moreover, such systems can mass-produce content that can be made widely available, ensuring equal accessibility to individuals from different socio-economic backgrounds.

To leverage benefits of AI-driven technology for societal benefit, we made the following contributions through this work:

- A beginner-friendly and self-paced lipreading course (LipMOOC) with structured design & high-quality content, to help improve learning outcomes.
- We evaluated LipMOOC with 120 participants, showing that its varied question type and modular design significantly improve learning outcomes.

We use statistical analysis to prove two key hypotheses.

(H1) Lipreading performance remains consistent whether users view real or synthetic videos generated through our pipeline.

(H2) A structured course design, as implemented in LipMOOC, leads to better learning outcomes compared to traditional lip-reading courses.

2 Related Work

Lipreading education has been a longstanding area of research, with various methodologies explored to enhance learning effectiveness such as, video-based lessons [6], and systematic lipreading training [13, 4]. AI-driven solutions like, deep learning-based assistive technology to improve audio-visual speech recognition [8], voice-based smart devices [2], generating real-time synthetic talking faces [10] have been proposed.

Agarwal et al., 2023 [1] introduced the use of synthetic talking heads for lipreading MOOCs, but their work primarily serves as a proof-of-concept rather than a structured educational framework. Furthermore, our work extends beyond proof-of-concept by conducting empirical assessments to validate the effectiveness of structured AI-driven lipreading training.

3 Methods & Data

3.1 LipMOOC

We developed a MOOC to teach lipreading by initiating the learning process at the phoneme level. This foundational approach was particularly chosen to ensure that learners focus on the smallest units of speech, making the task of lipreading more manageable and less overwhelming. Thus, learners can develop a strong understanding of basic visual patterns of speech, which serves as a critical building block for more complex lipreading tasks. The course content, featuring videos with no audio, focuses on different phonemes and how to spot and articulate them along with examples. The question types are - Single Word (SW), Phrase Level (PL), Fill in the Blanks (FIB), and Sentence Level (SL) items The course is divided into three distinct levels: Beginner, Intermediate, and Expert. Each level contain same ten lessons with increasing question difficulty to ensure progressive skill development. This structured approach is designed to ensure a clear learning path and more personalized pacing. For every lesson covered, there are 4 practice questions of each type 10x4=40 at each level. The course has a total of 120 practice questions (40x3=120). This approach ensured that learners remained engaged and motivated, avoiding potential frustration or discouragement by offering progressively more demanding tasks. This was hosted on a website[5] and supported various devices PC, laptop, mobile, etc.

3.2 Lipreading Data Generation

This pipeline (1) Uses face images from Synthetic Faces High Quality (SFHQ) [3], encompassing a wide range of ethnicities, pose, etc. (2) Processes the images in which lips are fully visible and filter out the rest (side view angles). (3) Generates a synthetic speech utterance from the provided text, adding buffers at the start and end. (4) Selects the face and the utterance and gives it as input to a talking head-generation model, EchoMimic [5], that generates synthetic talking heads using audio and a reference face, keeping the background intact.

Text-to-Speech System For our experiments we used Bark[6] to generate speech. The TTS model used in this pipeline is a configurable plug-and-play module which can be replaced by any TTS of choice.

Synthetic Talking Head Videos We tested a lot of talking head generation models [14–17, 11, 12, 5] and consequently, we adopt EchoMimic [5] in our pipeline.

The images from SFHQ [3] are used to ensure that we do not use real individual faces without permission as we will use it as content for our LipMOOC on lipreading. This approach helps maintain ethical standards and ensures that we comply with intellectual property laws. Images with face and lips clearly visible with a front view angle are then filtered from the dataset. This image is passed as a reference image along with the generated audio from Bark to our pipeline. Thus, creating realistic-looking synthetic talking faces.

3.3 Test Design

Experiment 1 For our first experiment, the primary goal was to test **(H1)**. For this study 50 (22 female and 28 male) participants with varying degrees of hearing loss (40 db to >90 db) volunteered. The average age of participants was 25 years, ranging from 13 to 35 years. Participants included eighth-grade students to Master's degree students. This validation seeks to establish that synthetic videos maintain a comparable level of clarity and effectiveness, ensuring their utility in lipreading tasks without compromising the accuracy of visual information. Each participant participates in two quizzes corresponding to two datasets made for (i) Real and (ii) Synthetic talking head videos. In total every user attempted 2 quizzes hosted on Google Forms, and participants reported taking the quiz from various devices PC, laptop, mobile, etc.

[5] https://ilocr.iiit.ac.in/lipreading
[6] https://github.com/suno-ai/bark

Data Analysis Plan To validate **(H1)** the difference in the mean score should be statistically insignificant. For that evidence of the null hypothesis is required, just the absence of evidence is not enough to support the hypothesis **(H1)**. For that we can conduct a Bayesian Equivalence Analysis using the **Bayesian Estimation Supersedes the t-test (BEST)** [7] and **95% HDI** to help validate potential of our pipeline to develop video content for a large-scale lipreading platform.

Experiment 2 Our second experiment aimed to validate **H2**. Our study used three different lipreading training courses, each designed to investigate specific aspects of lipreading instruction: (1) Standard Course (Lipreading.org) (SC): We utilized the established "Lipreading with Alphabet" course from Lipreading.org as our control condition. (2) Non-Modular Synthetic Course (NMS): The content for the LipMOOC we developed was presented in a nonmodular structure. (3) Modular Synthetic Course (MSC) i.e. LipMOOC. All three courses focused on teaching lipreading through a phonetic structure approach. To ensure comparability, we maintained consistency in the core lesson content across all courses, with variations primarily in the practice pattern (NMS) and course structure (MSC) to further assess the effects of these changes in learning. A total of 120 (52 female and 68 male) participants with moderate hearing loss (40-60 dB) to profound hearing loss (>90 dB) volunteered in this study. None of the participants had any previous experience in lipreading. All of them came from a low-income background. The study was conducted among sixth to eighth-grade school students following a special educational curriculum made for students with special needs. The age range of participants was 12 to 17 years. SC was conducted on the lipreading.org website, NMS was hosted on articulate 360 platform, and for MSC a course website[7] was created. All these courses supported various devices PC, laptop, mobile, etc. Implementation characteristics for Experiment 2 are shown in Table 1.

Table 1. Summary of implementation characteristics across courses. In Question type *Varied* refers to all SW, PL, FIB, and SL question type and *Single* refers to only SW question type.

	Standard Course (SC)	**Non-Modular Synthetic Course (NMS)**	**Modular Synthetic Course (MSC)**
Video Content	Human recorded	Synthetic	Synthetic
Duration	18 days	18 days	18 days
No. of students	38	42	40
Question Type	*Single*	*Varied*	*Varied*
Structure Type	Non Modular	Non Modular	Modular
No. of Questions	110	120	120

[7] https://ilocr.iiit.ac.in/lipreading

Data Analysis Plan Every participant took the same pre-test, which was used as the deciding factor to assign participants to different groups. This would avoid giving any unfair advantage to one group and prevent any imbalance as each group starts with a similar skill level. By ensuring homogeneous groups, we reduce variance and increase the ability to detect meaningful effects. To validate **(H2)**, we performed two-way mixed ANOVA to determine the effect, one for Question type (Single Vs. Varied) and another for Structure type (Non-Modular vs Modular).

4 Results

To validate this **(H1)**, we report the BEST statistics on Real and Synthetic videos for all four types of lipreading question patterns in Table 2. We also report

Table 2. We perform Bayesian Equivalence Analysis using BEST to calculate the 95% HDI range. The p-value and t-statistic were calculated from a standard t-test for comparison.

Measure	**Mean**	**MGD**	**P-value**	**T-statistic**	**95% HDI**
SW	0.9333	0.941	0.1019	1.6621	[-0.1488, 2.0155]
FIB	0.6000	0.624	0.2651	1.1253	[-0.4275, 1.6275]
PL	0.6667	0.701	0.2399	1.1874	[-0.4153, 1.7487]
SL	0.5000	0.513	0.2876	1.0732	[-0.3978, 1.3978]

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 four question types 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.

Table 3 shows descriptive statistics from Experiment 2, suggesting both question and structure types positively affect learning outcomes. To rigorously test (H2), we conducted two-way mixed ANOVAs.

Table 3. Experiment 2 descriptive statistics with averages for pre-test and post-test scores.

Group	Course	Questions Type	Structure Type	Pre-test	Post-test
Group 1	SC	Simple	Non-Modular	7.7	9
Group 2	NMS	Varied	Non-Modular	8.025	11.35
Group 3	MSC	Varied	Modular	7.8	14.58

Question Type A two-way mixed ANOVA (Table 4) between Groups 1 and 2 examined the impact of Question Type on post-test scores. Results showed a highly significant main effect, $F(1, 78) = 143.37, p = 2.39 \times 10^{-19}$, with large

Table 4. Two way mixed ANOVA Analysis of Question Type and test score between Group 1 and 2.

Source of Variation	DF	F-Statistic)	p-unc	np2
Between-subjects				
questions type	78	143.37	2.39e-19	0.648
Within-subjects				
time	78	17.82	6.50e-05	0.186
Interaction				
questions type:time	78	134.95	1.09e-18	0.634

effect size ($\eta_p^2 = 0.648$), indicating that varied questions lead to greater gains. A strong interaction effect ($F(1, 78) = 134.95, p = 1.09 \times 10^{-18}, \eta_p^2 = 0.634$) showed improvement over time depended on question type.

Structure Type A separate ANOVA (Table 5) between Groups 2 and 3 assessed the impact of Structure Type. There was a significant main effect, $F(1, 78) = 95.59, p = 3.39 \times 10^{-15}$, with large effect size ($\eta_p^2 = 0.551$), showing structure influences learning. The interaction effect ($F(1, 78) = 102.41, p = 7.43 \times 10^{-16}, \eta_p^2 = 0.568$) revealed that modular formats enhance learning more effectively over time. Thus, validating **(H2)** that a synthetically generated lipreading course with varied question types and a modular structure yields superior learning outcomes compared to a standard lipreading course.

Table 5. Two way mixed ANOVA Analysis of Structure Type and test score between Group 2 and 3.

Source of Variation	DF	F-Statistic	p-unc	np2
Between-subjects				
structure type	78	95.59	3.39e-15	0.551
Within-subjects				
time	78	760.84	5.54e-42	0.907
Interaction				
structure type:time	78	102.41	7.43e-16	0.568

5 Conclusion

This work introduces LipMOOC, a scalable AIED system designed to teach lipreading through a structured and self-paced approach. We demonstrate that lipreading can be effectively taught with AI-generated visual speech, making the learning process more accessible. Our empirical evaluation with 120 participants confirms that structured lipreading training leads to improved learning outcomes and that synthetic talking faces can serve as a viable alternative to real-world speakers. LipMOOC has the potential to bridge accessibility gaps and transform the way individuals with hearing impairments develop and refine their communication skills.

References

1. Agarwal, A., Sen, B., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: Towards moocs for lipreading: Using synthetic talking heads to train humans in lipreading at scale. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2217–2226 (2023)
2. Battina, D.S., Surya, L.: Innovative study of an ai voice-based smart device to assist deaf people in understanding and responding to their body language. SSRN Electronic Journal **9**, 816–822 (2021)
3. Beniaguev, D.: Synthetic faces high quality - text 2 image (sfhq-t2i) dataset (2024). https://doi.org/10.34740/kaggle/dsv/9548853, https://github.com/SelfishGene/SFHQ-T2I-dataset
4. Bernstein, L.E., Jordan, N., Auer, E.T., Eberhardt, S.P.: Lipreading: A review of its continuing importance for speech recognition with an acquired hearing loss and possibilities for effective training. American Journal of Audiology **31**(2), 453–469 (2022)
5. Chen, Z., Cao, J., Chen, Z., Li, Y., Ma, C.: Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. arXiv preprint arXiv:2407.08136 (2024)
6. Dodd, B., Plant, G., Gregory, M.: Teaching lip-reading: The efficacy of lessons on video. British journal of audiology **23**(3), 229–238 (1989)
7. Kruschke, J.K.: Bayesian estimation supersedes the t test. Journal of Experimental Psychology: General **142**(2), 573 (2013)
8. Kumar, L.A., Renuka, D.K., Rose, S.L., Wartana, I.M., et al.: Deep learning based assistive technology on audio visual speech recognition for hearing impaired. International Journal of Cognitive Computing in Engineering **3**, 24–30 (2022)
9. Olusanya, B.O., Davis, A.C., Hoffman, H.J.: Hearing loss grades and the international classification of functioning, disability and health. Bulletin of the World Health Organization **97**(10), 725 (2019)
10. Palanivel, N., Balamurugan, G., Kumar, S.S., Hassan, S., Thirumurugan, C., Thulasinathan, R.: Visual recognition system for hearing impairment using internet of things. In: AIP Conference Proceedings. vol. 3075. AIP Publishing (2024)
11. Peng, Z., Hu, W., Shi, Y., Zhu, X., Zhang, X., Zhao, H., He, J., Liu, H., Fan, Z.: Synctalk: The devil is in the synchronization for talking head synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 666–676 (2024)
12. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: 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. pp. 484–492 (2020)
13. Santana, R.S., Menezes, E.C., Ralin, V., Givigi, S., Batorowicz, B.: Lipreading as a communication strategy to enhance speech recognition in individuals with hearing impairment: a scoping review. Disability and Rehabilitation: Assistive Technology pp. 1–12 (2025)
14. Song, L., Liu, P., Yin, G., Xu, C.: Adaptive super resolution for one-shot talking-head generation. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4115–4119 (2024). https://doi.org/10.1109/ICASSP48485.2024.10446837
15. Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., Zhu, S.: Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. arXiv preprint arXiv:2406.08801 (2024)

16. Ye, Z., Zhong, T., Ren, Y., Yang, J., Li, W., Huang, J., Jiang, Z., He, J., Huang, R., Liu, J., et al.: Real3d-portrait: One-shot realistic 3d talking portrait synthesis. arXiv preprint arXiv:2401.08503 (2024)
17. Zhang, Y., Liu, M., Chen, Z., Wu, B., Zeng, Y., Zhan, C., He, Y., Huang, J., Zhou, W.: Musetalk: Real-time high quality lip synchronization with latent space inpainting. arXiv preprint arXiv:2410.10122 (2024)