## Towards MOOCs for Lipreading: Using Synthetic Talking Heads to Train Humans in Lipreading at Scale (Supplementary)

Aditya Agarwal\* IIIT Hyderabad

aditya.ag@research.iiit.ac.in

Rudrabha Mukhopadhyay IIIT Hydearbad

radrabha.m@research.iiit.ac.in

## 1. Statistical Analysis

We present the graphs of the user study conducted on the three lipreading protocols - (1) lipreading isolated words (WL), (2) lipreading sentences with context (SL), and (3) lipreading missing words in sentence (MWIS). Fig. 1 represents the mean user performance on the three protocols against standard deviation, and Fig. 2 represents the mean user performance against the 95% confidence interval (Eqn. 2) of the mean. Standard error indicated in the main paper is computed using Eqn. 1. The blue bars indicate scores on real videos with American-accented English, the orange bars indicate synthetic American-accented English. The synthetic data is generated using our pipeline.

standard error 
$$=\frac{\sigma}{\sqrt{n}}$$
 (1)

$$\mathbf{c}\mathbf{i} = \hat{x} \pm z \left(\frac{s}{\sqrt{n}}\right) \tag{2}$$

As mentioned in the main paper, we perform the Bayesian Estimation Supersedes the t-test (BEST) [4] for comparing the lipreading scores of the users across the real and synthetically generated videos. BEST estimates the difference in means between the two groups and yields a probability distribution over the difference. From the distributions, the mean credible value as the best guess of the actual difference and the 95% Highest Density Interval (HDI) as the range where the actual difference is with 95% credibility are computed. We validate if the ideal difference between the two groups lies in the 95% HDI. If it does, the difference between the two groups is not statistically significant; otherwise, the difference is statistically significant.

Bipasha Sen\* IIIT Hyderabad

bipasha.sen@research.iiit.ac.in

Vinay Namboodiri University of Bath

vpn@bath.ac.uk

C V Jawahar IIIT Hyderabad jawahar@iiit.ac.in

In Fig. 3, we show the graph of the distribution of the difference of means between the real and synthetically generated American-accented English (AE) videos for the three lipreading protocols. Please note that the ideal mean difference for all three lipreading tasks lies in the 95% HDI, indicating that the difference in the lipreading scores across the synthetic and real datasets are statistically insignificant. The graphs denote the distribution of the difference of means for the three protocols - WL, SL, and MWIS.

To validate if lipreading native-accented videos affect lipreading performance, we conduct a statistical analysis of the user's performance on the synthetically generated Indian-accented English (IE) and American-accented English (AE). Since the participants of our user study are from India, our expectation is that their lipreading scores on IE should be better than their scores on the test with AE, even though the users are comfortable with both accents. We conduct the two-sample z-test as our sample size is large (>30) for comparing the scores of the users across the synthetically generated IE and AE and plot the graph of the z-statistic for the 90% confidence interval. The 90% confidence interval is the acceptable region from -1.96 to +1.96and is represented by the green region, and the region lying outside in red is the rejection region. The graph of z-statistic for the three lipreading protocols is shown in Fig. 4. The zstatistic is given by the formula:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
(3)

where  $\bar{x_1}$  and  $\bar{x_2}$  denote the sample average of Indian and American accents.  $s_1$  and  $s_2$  denote the standard deviation of the two groups, and  $n_1$  and  $n_2$  represent the sample size of the two groups. The population's average is represented using  $\mu_1$  and  $\mu_2$ .

For the difference in the means of the two groups to be

<sup>\*</sup>Equal contribution







Figure 2: Mean user performance on the three lipreading protocols. The error bars are the 95% confidence interval of the mean.

statistically insignificant, the expected difference  $(\mu_1 - \mu_2)$  between the population's average should be 0. Consequently, our null hypothesis is **H0**: the difference in the mean scores between Synth IE and Synth AE is not statistically significant, i.e.,  $\mu_1 - \mu_2 = 0$ . Our alternate hypothesis is **H1**: the difference in the mean scores between Synth IE and Synth AE is statistically significant. If the z-statistic lies outside the acceptable range, H0 is rejected in favor of H1, indicating that the difference in the mean scores of the two groups is statistically significant. From the graphs 4, we observe that the z-statistic lies outside the acceptable range for two tasks - (1) lipreading words (WL) and (2) lipreading sentences (SL). Consequently, the p-value is also lower than the significance value ( $\alpha = 0.1$ ).

## 2. Quantitative and Qualitative Results

Even though all the individual modules in our pipeline can be replaced with other equivalent modules, we provide quantitative and qualitative metrics of the modules used in our pipeline and other similar modules. We evaluate the choice of the TTS model by conducting a user study comprising 30 participants. We compare three recent SOTA TTS works - (1) FastSpeech2 [1], (2) Tacotron2 [6], and (3) Glow-TTS [3] by performing Mean Opinion Score

Method	MOS
Tacotron2	$3.85\pm0.08$
Glow-TTS	$3.96\pm0.06$
FastSpeech2	$3.98\pm0.04$
GT	$4.53\pm0.07$

Table 1: Mean Opinion Scores (MOS) evaluations of different TTS models with 95% confidence interval.

Method	LSE-D
Wav2Lip	6.902
GT	6.718

Table 2: LSE-D metric computed for the synthetically generated videos using Wav2Lip model against real videos.

(MOS) [2] evaluation to evaluate the perceptual quality of the models. The MOS evaluation scores for the 95% confidence interval are provided in Table 1.

We provide quantitative scores for comparing the real and synthetically generated videos using the lipsync model. For quantitative comparison, we report the LSE-D [5] scores for comparing the lipsync performance of real videos against videos generated synthetically using Wav2Lip [5].



Figure 3: Distribution of difference of means performed using the Bayesian Analysis on real and synthetically generated Americanaccented English (AE). The graphs for the protocols are displayed in the following order: (1) lipreading words (WL), (2) lipreading sentences (SL), (3) lipreading missing words in sentence (MWIS). The 95% HDI interval is represented using the horizontal red line.



Figure 4: z statistic for the 90% confidence interval computed using the two-sample z-test for synthetically generated Indian-accented English (IE) and American-accented English (AE) videos. The critical z-value corresponding to the 90% confidence interval is  $\pm 1.96$ .

## References

- 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 multispeaker multi-style text-to-speech, 2021.
- [2] Min Chu and Hu Peng. An objective measure for estimating mos of synthesized speech. pages 2087–2090, 01 2001.
- [3] 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.
- [4] John Kruschke. Bayesian estimation supersedes the t test. *Journal of experimental psychology. General*, 142, 07 2012.
- [5] 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.
- [6] 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.