Your Multimodal Speech Model Says I Have a Face for Radio

Anonymous ACL submission

Abstract

As large neural models have become better at language tasks, researchers are increasingly building multi- and omnimodal models that handle more modalities of data. One example is the expansion of speech recognition models to audio-visual data for noise mitigation and multimodal subtitling. While performance and bias have been studied extensively in the single-modality regime, it is unknown how new modalities affect this, even though they produce biases in humans. We therefore propose the first bias evaluation of multimodal speech recognition, where we create videos pairing different faces with the same audio, and measure changes in speech transcription accuracy. We find large quality-of-service differences across mWHISPER-FLAMINGO and GEMINI models, with drops of up to 4.05 word error rate points, across self-declared gender, ethnicity, and their intersection. Our findings point to a priority for developers to evaluate, fix, and communicate such limitations, as providing more signals through additional modalities is not necessarily better, and may even lead to biased outcomes.

1 Introduction

As large neural models improve on single modalities of language data like text or audio, it is becoming increasingly common to incorporate multiple modalities into single "multimodal" or "omnimodal" models, such as images and text (Radford et al., 2021; Liu et al., 2023; Bai et al., 2024), or video and audio (Shi et al., 2022; Rouditchenko et al., 2024, 2025; Shu et al., 2025). Within audio-based research, the expansion to multimodality has focused on audio-visual speech recognition, with applications in noisy environments (Rouditchenko et al., 2024, 2025), automatic video captioning, subtitling (El Idrissi et al., 2025), and more.

Auditory and visual processes are also implicated in human speech perception (McGurk and MacDonald, 1976), where it similarly helps us deal with noisy environments (Sumby and Pollack, 1954). However, visual perceptions of the speaker can sometimes worsen our auditory processing of their speech, a phenomenon termed *reverse linguistic stereotyping* (Kang and Rubin, 2009). In other words, nonlinguistic factors such as visually-perceived speaker ethnicity affect speech perception and processing (Rubin and Smith, 1990; Rubin, 1992; Zheng and Samuel, 2017; Bhuvanagiri et al., 2019; Kutlu, 2023). No study thus far has examined whether such multimodal raciolinguistic biases also appear in audio-visual speech recognition systems, despite a long history of study in single-modality settings in vision (Buolamwini and Gebru, 2018) and in speech (Koenecke et al., 2020).

The lack of such studies is largely due to the difficulty of collecting adequate data: Videos must be *audio-matched* to eliminate confounding factors, as in the matched-guise setup from sociolinguistics. We propose a synthetic setup that allows us to evaluate whether the ethnicity and gender of different faces affects audio-visual speech recognition performance at scale. We construct a dataset of $75,000$ videos where English audio samples that vary in auditory characteristics (accent and noise) are systematically paired with lip-synced faces that vary in visual characteristics (ethnicity and gender). Using this dataset, we evaluate audio-visual speech recognition with two models: mWHISPER-FLAMINGO MEDIUM (Rouditchenko et al., 2025) and GEMINI-2.5-FLASH (Comanici et al., 2025).

We show that our synthetic setting is a valid testbed for bias, as our lip-synced videos help audio-visual speech recognition (AVSR) performance under noisy conditions, just like natural video. Both with and without noise, AVSR models show single-axis and intersecting ethnicity and gender biases when transcribing the exact same set of audios with different faces. Our findings thus provide evidence of reverse linguistic stereotyping in AVSR models. We hope that our work spurs

developers to fix bias in contexts where biases may come from multiple modalities, towards better and fairer multimodal speech recognition.[1]

Bias statement. In response to calls for clearer articulations of bias in language technology (Blodgett et al., 2020; Wang et al., 2022), we conceptualize bias in our work as differences in quality-of-service with audio-visual speech recognition, based purely on a change in visual cues such as race and gender that should be irrelevant to transcription of the exact same audio. Our definition of bias is parallel to reverse linguistic stereotyping in humans, and we measure it by testing for statistically significant changes in word error rate (WER), which can have detrimental effects on users of speech recognition technologies (Mengesha et al., 2021; Markl, 2022).

2 Related Works

Multimodal speech perception in humans. The exact same English audio gets different intelligibility ratings from listeners when paired with pictures of White versus South Asian faces (Kutlu, 2023), or Asian faces (Rubin, 1992). However, this effect seems to disappear with video (Zheng and Samuel, 2017). Notably, these works use a very small sample size (three unambiguous words, presented on an accent continuum in Zheng and Samuel (2017)), whereas we test at a much larger scale with automated and computational methods. Our methods are inspired by this literature (in particular the matched-guise test setup), but we evaluate speech recognition systems rather than human listeners, and focus on transcription rather than speaker attitudes as in some of these studies.

Single-modality bias in vision and speech. Prior to our work, quality-of-service differentials based on race and gender have been shown in image classification (Buolamwini and Gebru, 2018), image captioning (Hendricks et al., 2018; Wolfe and Caliskan, 2022), and English speech recognition (Tatman, 2017; Koenecke et al., 2020). In contrast to our work, none of these papers study multi-modal speech recognition biases. Additionally, work on bias in speech recognition tends to focus on acoustically-perceptible differences (accent, dialect, etc.), in contrast to our work, where identical audio is paired with different *visual* input.

Multi- and omnimodal models. Analyses of multimodal models focus overwhelmingly on vision-language models that take images and text as input. For instance, Fu et al. (2025) show that these models tend to overlook their vision input representations, Kamruzzaman et al. (2025) analyze how race, gender, and skin tone affect these models' attribution of emotions, Bui et al. (2025) evaluate multimodal, multicultural hate speech, and Kim et al. (2025) evaluate models' cultural stereotypes when visual ethnicity information is perturbed. Although this last paper's methodology is similar to ours in terms of systematically perturbing ethnicity features, our work focuses on video and speech modalities, which none of the above consider. Closest to our work, Dai et al. (2024) evaluate the robustness of audiovisual speech recognition to dropped frames, while we look at robustness to variance in the visual component of the input.

3 Methodology

The dominant signal for speech recognition is the audio. In order to disentangle the impact of visual biases from audio-level performance differences in AVSR, we require a controlled setup with rich demographic information. No such audio-visual dataset is publicly available. Therefore, we create a usefully-close approximation through modifying and combining existing datasets. The result allows us to compare the performance of AVSR models with varying visuals (faces of varying ethnicities and genders) against the *same* audio signal.

3.1 Dataset

To evaluate multimodal speech recognition bias, we create a dataset of videos that systematically combine faces and audio files, corresponding to experiments in the matched-guise setup with humans (Kutlu, 2023; Zheng and Samuel, 2017).

Faces. For high-resolution, standardized photographs of racially diverse faces, we use the Chicago Face Database (Ma et al., 2015) and the India Face Set (Lakshmi et al., 2021). From these datasets, we select a total of 500 faces that are marked to have a neutral expression. The faces are equally balanced across five ethnicities. Using their terminology, these are: Asian American, Black, Indian Asian, Latino/a, and White.[2] Within each ethnicity, the faces we select are equally balanced across Female and Male, the only two gender labels in the datasets. Ethnicity and gender in these

[1] We release code at **REDACTED FOR REVIEW**.

[2] In the rest of the text, we use the labels Asian, Black, Indian, Latinx, and White, for brevity.

datasets are self-reported, and we exclude faces of those who chose not to report. We ensure that each individual's face is only included once.

Audio. We use English audio from the Common-Voice 17.0 dataset (Ardila et al., 2020), and select samples corresponding to the following high-level accent labels: UK, US, and Indian. As accents in this dataset are self-identified, each grouping contains a mixed set of accent profiles, resulting in a sample that reflects some of the diversity of these coarse labels. We describe accent selection in Appendix A. Since all audio samples are self-recorded under diverse recording conditions, audio quality varies widely. We use a heuristic based on audio ratings from CommonVoice to obtain the top 50 audio samples per accent, as outlined in Appendix B. As many participants do not indicate gender, we ignore gender labels and focus on quality. With 50 samples each from three accents, we have a total of 150 distinct audio files. As CommonVoice is not designed to collect the same text for different accents or speakers, there are no repeated texts.

Pairing faces and audio. All faces are combined with all audios, leading to a total of 500 faces $\times$ 3 accents $\times$ 50 audios $= 75,000$ data points. For each combination, we create an audio-video file where the video is lip-synced to the audio with WAV2LIP + GAN (Prajwal et al., 2020), using the clean audio data as input. All videos are first downscaled to a height of 256, maintaining the original aspect ratio, to fit the lip-sync model. Our use of lip-synced videos reflects realistic and intended usage of audio-visual models, in contrast to most sociolinguistic work where still faces are used. We additionally provide results with videos of static faces in Appendix G, for reference.

As AVSR mostly benefits from visual input in the presence of noise, we also generate noisy versions of the audio-video files by adding "babble" noise with a Signal-to-Noise Ratio (SNR) of 5 dB to the audio track. We obtain this noise from Rouditchenko et al. (2025), which in turn is sampled from the MUSAN dataset (Snyder et al., 2015).

3.2 Experimental Setup

Models. We ran our experiments using MWHISPER-FLAMINGO MEDIUM[3] (Rouditchenko et al., 2025) and GEMINI-2.5-FLASH (Comanici et al., 2025). As input, MWHISPER-FLAMINGO takes audio and greyscaled video, cropped to the mouth region, separately. As for GEMINI, the full video is encoded in Base64. We show prompts for GEMINI in Appendix D. We also report results with an automatic speech recognition (ASR) baseline obtained by turning off the video input, which both models support.

[3]We ran preliminary experiments with MWHISPER-FLAMINGO SMALL and WHISPER-FLAMINGO, but exclude those as they perform much worse.

Model	Accent	Average	ASR
No noise (SNR= ∞)			
MWHISPER-F	Indian	24.13$\pm$.43	21.53$\pm$7.33
MWHISPER-F	US	13.47$\pm$.25	12.94$\pm$4.72
MWHISPER-F	UK	19.78$\pm$.35	17.77$\pm$6.90
GEMINI	Indian	10.84$\pm$.24	7.24$\pm$3.13
GEMINI	US	8.43$\pm$.18	6.67$\pm$3.60
GEMINI	UK	11.73$\pm$.28	8.78$\pm$4.12
Babble noise (SNR= 5)			
MWHISPER-F	Indian	58.31$\pm$.52	49.90$\pm$9.63
MWHISPER-F	US	34.75$\pm$.43	31.76$\pm$8.10
MWHISPER-F	UK	46.49$\pm$.54	47.97$\pm$15.64
GEMINI	Indian	39.16$\pm$.53	43.05$\pm$11.61
GEMINI	US	23.60$\pm$.31	27.25$\pm$7.96
GEMINI	UK	34.29$\pm$.49	34.26$\pm$9.44

Table 1: Average per-accent WER scores of AVSR systems compared to audio-only ASR versions thereof. MWHISPER-F = MWHISPER-FLAMINGO MEDIUM; GEMINI = GEMINI-2.5-FLASH.

Metrics. We assess model performance with WER: The corpus-level average number of errors weighted by document length. Implementation notes are provided in Appendix E. We also compute WER standard deviations as a weighted average of the per-document difference to the mean, and estimate standard errors of the mean WER using the normal assumption.

For statistical testing of significant pairwise quality-of-service differences (Table 2), we apply $10,000$ iterations of a one-sided permutation test, stratified over audio files. To correct for multiple comparisons within each row, we apply the *max-T* method (Westfall and Young, 1993). To correct for the number of rows, we use the Bonferroni method.

4 Main Findings

Table 1 provides an overview of WERs on average, compared to an ASR-only (i.e., no visual input) baseline. See Appendix C for a full breakdown by gender and ethnicity conditions. Overall, GEMINI models perform better than MWHISPER-FLAMINGO models in all settings (despite some hallucinations and deviations from instructions -

Model	Accent	Δ WER (Max−Min) Ethnicity		Gender		Intersection	
No noise (SNR= ∞)							
MWHISPER-F	Indian	2.06	(Indian − Asian)	1.14	(M − F)	4.05*	(Indian male − Asian male)
MWHISPER-F	US	0.72	(Black − Asian)	0.43	(M − F)	0.98	(Black male − Latinx female)
MWHISPER-F	UK	0.51	(Indian − Asian)	1.34*	(M − F)	2.27*	(Indian male − Latinx female)
GEMINI	Indian	0.77	(Indian − White)	0.25	(F − M)	1.19*	(Indian male − White male)
GEMINI	US	0.13	(Indian − White)	0.08	(M − F)	0.26	(Asian male − Asian female)
GEMINI	UK	0.49	(Indian − Black)	0.35	(M − F)	0.90	(Indian male − Black female)
Babble noise (SNR= 5)							
MWHISPER-F	Indian	1.01	(Indian − Asian)	0.10	(F − M)	1.71	(Indian male − Asian male)
MWHISPER-F	US	1.75*	(Indian − Asian)	1.45*	(M − F)	3.95*	(Indian male − Asian female)
MWHISPER-F	UK	0.91	(Black − White)	0.47	(F − M)	2.84*	(Indian male − Latinx male)
GEMINI	Indian	1.49*	(Asian − White)	0.51	(M − F)	2.73*	(Asian male − White male)
GEMINI	US	0.53	(Black − White)	0.20	(F − M)	0.95*	(Latinx female − White male)
GEMINI	UK	1.03	(Black − White)	0.43	(M − F)	1.62*	(Black male − White female)

Table 2: Largest AVSR quality-of-service differences between faces, across ethnicity, gender, and their intersection. Asterisk (*) denotes statistical significance ($p < 0.005$) according to a permutation test (Section 3.2). MWHISPER-F = MWHISPER-FLAMINGO MEDIUM; GEMINI = GEMINI-2.5-FLASH. M = Male; F = Female.

see Appendix F). We caution, however, that GEMINI's better performance could be due to train-test contamination. US accents are transcribed better than UK accents, which in turn are transcribed better than Indian accents. These performance differences could also come from differences in audio quality, however, they pattern similarly to previously-studied accent biases in ASR models (Chan et al., 2022; DiChristofano et al., 2023).

Lip-synced input is useful for AVSR under noise. In the absence of noise, visual input degrades recognition performance. However, when we add noise at an SNR of 5 dB, compared to ASR alone, MWHISPER-FLAMINGO improves for UK accents, and GEMINI improves in all settings. This aligns with the motivation and results in MWHISPER-FLAMINGO (Rouditchenko et al., 2025), and suggests that our synthetic data setup resembles realistic conditions. Audio signals are often enough for ASR, as they are in some sense captured more faithfully than video (where mouths may be badly cropped or undetected, or faces may be off-camera). Nevertheless, the fact that video sometimes *worsens* rather than maintains performance under no noise points to the infancy of these models.

Models show ethnicity and gender biases when transcribing the same audio with different faces. As Table 2 shows, there are several statistically significant quality-of-service differences in transcribing the exact same set of audios with different groupings of faces, across both models and all accent classes. While some single-axis ethnicity and gender differences may not be significant, they hide *intersecting* biases as in other settings in both vision (Buolamwini and Gebru, 2018) and language (Rodríguez et al., 2025): Indian male faces consistently worsen MWHISPER-FLAMINGO's transcription, and White male faces consistently get the best WERs with GEMINI. In fact, results even seem to worsen under noise, where we have seen that models rely more on their visual input as well. These results provide evidence of the reverse linguistic stereotyping effect in models, and point to an important priority for model developers to fix.

5 Conclusion

We contribute a framework to study biases in audio-visual speech recognition by creating videos that present the same audio with lip-synced faces that vary along the dimensions of ethnicity and gender. WERs with MWHISPER-FLAMINGO and GEMINI-2.5-FLASH confirm that our synthetic setup, while inferior to real videos of people speaking, simulates it well enough, as it improves transcription over audio-only ASR under noise. We then show that models show single-axis and intersecting ethnicity and gender biases when transcribing the same audio with different faces. Our results provide evidence of reverse linguistic stereotyping in AVSR models, and suggest that synthetic data with lip-sync might be a useful avenue to mitigate these quality-of-service differences. Fixing such biases that emerge from the combination of multiple modalities is an important priority for the future of multi- and omni-modal language technology and research.

Limitations

The primary limitation of this work is that we evaluate with stand-ins for real video: Lip-synced videos in the main text, and videos with static faces in Appendix G. Second, we work with a limited set of accents and faces, as well as only in English, while results are sure to differ in other languages and with a broader variety of facial and dialectal differences. Finally, we use self-declared ethnicity, gender, and accent labels, which is good for participant autonomy and dignity, but we note that *perceptions* of people may differ from identity, and this is likely to play a larger role in bias in computational systems.

Ethics Statement

Ethnicity and gender are multi-faceted aspects of identity, and our operationalization choices affect the conclusions we can draw (Steidl and Werum, 2019). Ethnicity, gender, and accent are all self-declared in the data we use. Although this supports participant autonomy and dignity, perceptions and behaviour (from both humans and machines) may differ, which can cause harm to individuals. Our results show quality-of-service differentials in AVSR that replicate structural power differentials in society and are an example of algorithmic oppression, with harms to users of these technologies with certain faces. We use all datasets in compliance with their terms of use and licensing restrictions, i.e., we do not identify the owners of the faces or voices, we use the face data only for non-commercial scientific research, and we use CommonVoice for evaluating speech recognition models.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond.

Deepa Bhuvanagiri, Anna Grahlherr, Pamela Beardsell, Jaishree Chittoor, and Donald Rubin. 2019. The role of race in the stereotyping of a speaker's accent as native or non-native. *Journal of Emerging Investigators*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025. Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

May Pik Yu Chan, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole Holliday. 2022. Training and typological bias in ASR performance for world Englishes. In *Interspeech 2022*, pages 1273–1277.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Yusheng Dai, Hang Chen, Jun Du, Ruoyu Wang, Shihao Chen, Haotian Wang, and Chin-Hui Lee. 2024. A study of dropout-induced modality bias on robustness to missing video frames for audio-visual speech recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27445–27455.

Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2023. Global performance disparities between english-language accents in automatic speech recognition. *Preprint*, arXiv:2208.01157.

Zakaria El Idrissi, Mohamed-Amine Chadi, and Samar Mouchawrab. 2025. Deep learning for video captioning: A review. In *International Conference on Advances in Communication Technology and Computer Engineering*, pages 335–346. Springer.

Stephanie Fu, tyler bonnen, Devin Guillory, and Trevor Darrell. 2025. Hidden in plain sight: VLMs overlook their visual representations. In *Second Conference on Language Modeling*.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Computer Vision – ECCV 2018*, page 793–811, Cham. Springer International Publishing.

Mahammed Kamruzzaman, Amanda Cercas Curry, Alba Cercas Curry, and Flor Miriam Plaza-del Arco. 2025. Seeing race, feeling bias: Emotion stereotyping in multimodal language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7317–7351, Suzhou, China. Association for Computational Linguistics.

Okim Kang and Donald L. Rubin. 2009. Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, 28(4):441–456.

Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. 2025. WHEN TOM EATS KIMCHI: Evaluating cultural awareness of multimodal large language models in cultural mixture contexts. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 143–154, Albuquerque, New Mexico. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Ethan Kutlu. 2023. Now you see me, now you mishear me: Raciolinguistic accounts of speech perception in different english varieties. *Journal of Multilingual and Multicultural Development*, 44(6):511–525.

Anjana Lakshmi, Bernd Wittenbrink, Joshua Correll, and Debbie S. Ma. 2021. The india face set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in Psychology*, Volume 12 - 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4):1122–1135.

Nina Markl. 2022. Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 521–534, New York, NY, USA. Association for Computing Machinery.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence*, Volume 4 - 2021.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (Demonstrations)*, pages 48–53.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. 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, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Elisa Forcada Rodríguez, Olatz Perez-de Vinaspre, Jon Ander Campos, Dietrich Klakow, and Vagrant Gautam. 2025. Colombian waitresses y jueces canadienses: Gender and country biases in occupation recommendations from LLMs. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 182–194, Vienna, Austria. Association for Computational Linguistics.

Andrew Rouditchenko, Yuan Gong, Samuel Thomas, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass. 2024. Whisper-Flamingo: Integrating Visual Features into Whisper for Audio-Visual Speech Recognition and Translation. In *Interspeech 2024*, pages 2420–2424.

Andrew Rouditchenko, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. 2025. mwhisper-flamingo for multilingual audio-visual noise-robust speech recognition. *IEEE Signal Processing Letters*, 32:2144–2148.

Donald L. Rubin. 1992. Nonlanguage factors affecting undergraduates' judgments of nonnative english-speaking teaching assistants. *Research in Higher Education*, 33(4):511–531.

Donald L. Rubin and Kim A. Smith. 1990. Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative english-speaking teaching

assistants. *International Journal of Intercultural Relations*, 14(3):337–353.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*.

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2025. Audio-Visual LLM for Video Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4246–4255.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Christina R. Steidl and Regina Werum. 2019. If all you have is a hammer, everything looks like a nail: Operationalization matters. *Sociology Compass*, 13(8):e12727. E12727 SOCO-1495.R2.

W. H. Sumby and Irwin Pollack. 1954. Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2):212–215.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring representational harms in image captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 324–335, New York, NY, USA. Association for Computing Machinery.

Peter H Westfall and S Stanley Young. 1993. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, 1 edition. John Wiley & Sons. Wiley Series in Probability and Statistics.

Robert Wolfe and Aylin Caliskan. 2022. Markedness in visual semantic ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1269–1279, New York, NY, USA. Association for Computing Machinery.

Yi Zheng and Arthur G. Samuel. 2017. Does seeing an asian face make speech sound more accented? *Attention, Perception, & Psychophysics*, 79(6):1841–1859.

A Accents in CommonVoice

CommonVoice includes self-declared accent labels. First, we lowercase all labels. All audio files that include "india" in the label, and that exclude "united states" and "england english" are selected as Indian accented audio files for our dataset. We select all files that include "england english" in the label as UK-accented data. Finally, all labels that include "united states english" are counted as US accents. Through manually inspecting the remaining labels, we confirmed that no other accents were declared.

B Quality Filtering for CommonVoice

CommonVoice data includes audio ratings consisting of up-votes and down-votes. We designed a custom metric to use these ratings to select audios with the highest quality.

For a given accent, we first weight the up-votes and down-votes. We define U_i as the up-votes for sample i and U_{max} as the max up-votes for the accent. Equivalently, we define D_i as the down-votes for sample i and D_{max} as the max down-votes for the accent.

Finally, we define $\bar{U}_i$ as the weighted up-votes for a given sample, and $\bar{D}_i$ as the weighted down-votes for a given sample, which we calculate as follows:

$$\bar{U}_i := \frac{U_i}{U_{\text{max}}} \tag{1}$$

$$\bar{D}_i := \frac{D_{\text{max}} - D_i}{D_{\text{max}}}. \tag{2}$$

A higher weighted score represents more up-votes and fewer down-votes, respectively.

We then compute the score of an audio as the harmonic mean of the weighted votes, and take the top 50 with the highest score per accent.

$$\text{Score} := \frac{2 \cdot \bar{U}_i \cdot \bar{D}_i}{\bar{U}_i + \bar{D}_i}. \tag{3}$$

C Full Results

All WER results with confidence intervals across all splits are shown in Table 3.

D Prompting GEMINI

We pass two text prompts to GEMINI together with the video or audio: A system prompt and an instruction. The system prompt was "You are a helpful transcription agent."

The instruction for video transcription was "Please transcribe everything that is said in the following video into English. Output only the transcription. Prioritize accuracy in transcription, including punctuation and sentence structure, to reflect natural speech flow. The audio quality may vary from good to quite noisy, please do your best to isolate the main speaker." For audio transcription, we replace the word "video" with "audio".

Model	Accent	Gender	Asian	Black	Indian	Latinx	White	Average
MWHISPER-F	Indian	Female	23.30±1.28	23.97±1.24	23.37±0.99	23.68±1.17	23.46±1.02	23.56±0.51
		Male	22.97±1.05	24.85±1.33	27.03±1.77	23.80±1.05	24.83±2.14	24.70±0.68
		Average	23.14±0.83	24.41±0.91	25.20±1.01	23.74±0.79	24.14±1.18	24.13±0.43
MWHISPER-F	US	Female	13.04±0.67	13.91±1.07	13.07±0.68	13.01±0.67	13.25±0.69	13.26±0.34
		Male	13.42±0.70	13.99±1.07	13.96±0.75	13.58±0.72	13.47±0.69	13.68±0.36
		Average	13.23±0.48	13.95±0.76	13.52±0.50	13.29±0.49	13.36±0.49	13.47±0.25
MWHISPER-F	UK	Female	19.37±1.04	19.08±1.05	19.27±1.06	18.79±1.00	19.04±1.02	19.11±0.46
		Male	19.94±1.06	20.31±1.17	21.06±1.26	20.64±1.13	20.29±1.12	20.45±0.51
		Average	19.65±0.74	19.70±0.79	20.16±0.82	19.72±0.76	19.66±0.76	19.78±0.35
GEMINI	Indian	Female	11.29±0.79	10.88±0.76	11.03±0.76	10.95±0.76	10.68±0.76	10.97±0.34
		Male	10.52±0.71	10.86±0.78	11.30±0.78	10.79±0.75	10.11±0.69	10.72±0.33
		Average	10.90±0.53	10.87±0.54	11.17±0.55	10.87±0.53	10.40±0.51	10.84±0.24
GEMINI	US	Female	8.33±0.53	8.34±0.54	8.58±0.56	8.34±0.56	8.34±0.55	8.39±0.25
		Male	8.59±0.59	8.44±0.57	8.40±0.56	8.51±0.55	8.38±0.55	8.47±0.25
		Average	8.46±0.40	8.39±0.39	8.49±0.39	8.42±0.39	8.36±0.39	8.43±0.18
GEMINI	UK	Female	11.70±0.87	11.19±0.86	11.96±0.88	11.42±0.84	11.53±0.85	11.56±0.38
		Male	11.98±0.89	11.87±0.91	12.09±0.91	11.95±0.90	11.67±0.89	11.91±0.40
		Average	11.84±0.62	11.53±0.63	12.02±0.63	11.69±0.62	11.60±0.61	11.73±0.28
MWHISPER-F	Indian	Female	57.87±1.63	58.35±1.63	58.19±1.60	58.42±1.64	58.94±1.64	58.35±0.73
		Male	57.62±1.57	58.04±1.62	59.33±1.71	58.33±1.80	57.98±1.61	58.26±0.74
		Average	57.74±1.13	58.20±1.15	58.76±1.17	58.38±1.22	58.46±1.15	58.31±0.52
MWHISPER-F	US	Female	33.28±1.08	34.81±1.21	34.07±1.13	34.58±2.10	33.40±1.10	34.03±0.62
		Male	34.52±1.24	35.67±1.39	37.23±1.60	34.96±1.29	35.00±1.30	35.48±0.61
		Average	33.90±0.82	35.24±0.92	35.65±0.98	34.77±1.23	34.20±0.85	34.75±0.43
MWHISPER-F	UK	Female	46.78±1.58	47.41±1.65	46.00±1.57	47.07±2.25	46.36±1.56	46.73±0.78
		Male	45.65±1.57	46.64±1.67	47.99±1.90	45.15±1.59	45.86±1.62	46.26±0.75
		Average	46.21±1.11	47.03±1.17	47.00±1.23	46.11±1.38	46.11±1.13	46.49±0.54
GEMINI	Indian	Female	39.02±1.46	39.08±1.45	38.89±1.47	38.76±1.46	38.76±1.49	38.90±0.66
		Male	40.39±2.58	38.87±1.48	40.29±1.56	39.84±1.86	37.66±1.45	39.41±0.82
		Average	39.70±1.48	38.97±1.04	39.59±1.07	39.30±1.18	38.21±1.04	39.16±0.53
GEMINI	US	Female	23.74±0.96	23.69±0.97	23.64±0.96	23.97±1.00	23.47±0.95	23.70±0.43
		Male	23.51±0.95	23.85±0.97	23.56±1.04	23.56±0.95	23.02±0.91	23.50±0.43
		Average	23.62±0.68	23.77±0.68	23.60±0.71	23.76±0.69	23.25±0.66	23.60±0.31
GEMINI	UK	Female	34.30±1.56	34.31±1.54	34.00±1.47	34.07±1.47	33.67±1.48	34.07±0.67
		Male	34.20±1.49	35.30±1.80	34.71±1.53	34.42±1.53	33.87±1.53	34.50±0.71
		Average	34.25±1.08	34.81±1.18	34.36±1.06	34.24±1.06	33.77±1.06	34.29±0.49

Table 3: All WER results with confidence intervals across all splits.

E Word Error Rate Calculations

We calculate Word Error Rate with the Fairseq (Ott et al., 2019) `WerScorer`, which we initialize as follows:

```
scorer = WerScorer(
  WerScorerConfig(wer_tokenizer="13a",
  wer_remove_punct=True,
  wer_char_level=False,
  wer_lowercase=True))
```

F Hallucinations and Refusals

Here, we provide some examples of model hallucinations (output unrelated to input) in transcriptions, (partial) refusals to transcribe, and other unrequested output.

Table 4 provides an example of a hallucination.

Sometimes GEMINI refused to transcribe part of the input, replacing parts of the text with some variation of [unintelligible]. Table 5 provides an example.

GEMINI also refused to transcribe some audios in full. Table 6 provides some examples. Sometimes, these descriptions lead to a longer output than the reference text, resulting in a WER higher than 100.

Table 7 provides examples of GEMINI adding sound descriptions on top of the transcribed text.

Table 8 provides an example of GEMINI repeating part of the prompt in the output.

Finally, Table 9 provides an example of an unacceptable refusal.

G Results with Static Faces

Table C provides an overview of the quality-of-service differences for static faces.

Reference	Each Bantu group changes musical style with every change of language.
Hypothesis	This is my, okay. My name is Olanrewaja. I just want to tell you about the future. The future about getting crypto to convert cash to crypto. Okay, a lot of people like converting their crypto into cash and they don't want to convert it like through any exchange or any, okay, anything. Because some exchanges have very high prices and some exchanges have very low prices, so a lot of people just want to get to someone that will convert their crypto and convert it into cash and give it to them, or convert the cash to crypto, and they'll send it to them. That's a good business because you can earn on this, a lot. Thank you.
Reference	Potatoes are less space-saving, durable and cheap than pasta.
Hypothesis	Cheetos R S P C E D N J R 1 2 3 N S M B Q R S.

Table 4: GEMINI-2.5-FLASH hallucinations.

Reference	Historians and Ma Chao's contemporaries have a generally negative view of him.
Hypothesis	Historians have [unintelligible] contemporaries [unintelligible] a [unintelligible] negative view of [unintelligible].

Table 5: GEMINI-2.5-FLASH partial refusals.

Short description	"[No discernible speech from the main speaker. Background chatter is present.]", "[no audible speech]", "[Muffled voices and background chatter]", "[No discernible speech]", "[Indistinct chatter]"
Long description	"[The main speaker's speech is unintelligible due to overwhelming background noise.]", "The audio quality for the main speaker is very poor, making it difficult to isolate words clearly due to significant background chatter. The main speaker's speech is almost entirely unintelligible, consisting mostly of mumbles or very soft sounds that are drowned out. Therefore, I cannot accurately transcribe what the main speaker is saying."

Table 6: GEMINI-2.5-FLASH full refusals.

Reference	Chickens lay so many eggs a week, it wouldn't be possible to breed them all.
Hypothesis	[Sound of firework/firecracker] Chickens lay so many eggs a week, it wouldn't be possible to breed them all. [Loud bang] Chickens lay so many eggs a week, it wouldn't be possible to breed them all. [EXPLOSION] Chickens lay so many eggs a week, it wouldn't be possible to breed them all. [Sound of fireworks/gunshot] Chickens lay so many eggs a week, it wouldn't be possible to breed them all.

Table 7: GEMINI-2.5-FLASH sound descriptions.

Reference	He also won the South African National Time Trial Championships at junior level.
Hypothesis	You on the all right. You on the all right to go out there when you saw that happen on the field, and that was just an unbelievable moment, and I think a very emotional one for that whole team.The user wants a transcription of the audio. I will listen carefully to the speaker's words and transcribe them, focusing on accuracy and natural speech flow. I need to be aware of any background noise and try to isolate the main speaker.

Table 8: GEMINI-2.5-FLASH prompt repetition.

Reference	Historians and Ma Chao's contemporaries have a generally negative view of him.
Hypothesis	The audio appears to be in Spanish. The speaker says: "Historia con el más chachi, contemporánea, apagándole la vida a la gente que y uno va a". This translates to: "History with the coolest, contemporary, turning off the lives of people who and one goes to." Please confirm if you would still like this to be transcribed in English, or if you meant a transcription of the Spanish spoken.

Table 9: GEMINI-2.5-FLASH unacceptable refusal.

Model	**Accent**	**Average**	**ASR**
No noise (SNR= ∞)			
MWHISPER-F	Indian	28.26$\pm$.48	21.53$\pm$7.33
MWHISPER-F	US	14.23$\pm$.25	12.94$\pm$4.72
MWHISPER-F	UK	19.65$\pm$.34	17.77$\pm$6.90
GEMINI	Indian	9.84$\pm$0.19	7.24$\pm$3.13
GEMINI	US	8.21$\pm$0.17	6.67$\pm$3.60
GEMINI	UK	9.85$\pm$0.22	8.78$\pm$4.12
Babble noise (SNR= 5)			
MWHISPER-F	Indian	70.98$\pm$1.28	49.90$\pm$9.63
MWHISPER-F	US	40.43$\pm$0.85	31.76$\pm$8.10
MWHISPER-F	UK	49.92$\pm$0.53	47.97$\pm$15.64
GEMINI	Indian	37.97$\pm$0.47	43.05$\pm$11.61
GEMINI	US	24.06$\pm$1.02	27.25$\pm$7.96
GEMINI	UK	41.93$\pm$14.21	34.26$\pm$9.44

Table 10: Static-face (no lip-sync) absolute WER scores by groupings of faces across gender, ethnicity, and accent. MWHISPER-F = MWHISPER-FLAMINGO MEDIUM; GEMINI = GEMINI-2.5-FLASH.

H Compute

For our experiments with the WHISPER-FLAMINGO model family, we used two NVIDIA RTX A6000 GPUs for approximately 500 hours. For our experiments with GEMINI-2.5-FLASH, we used approximately 1B input and output tokens and an Intel(R) Xeon(R) CPU E5-2673 v3 @ 2.40GHz for approximately 400 hours.

I Use of AI Assistants

We used Deepseek to look up statistical analysis references, which were read, corroborated, and verified by hand. We also used Copilot and Genie for coding support, which was also manually checked.

Model	Accent	Δ WER (Max−Min) Ethnicity		Gender		Intersection	
No noise (SNR= ∞)							
MWF	Indian	3.22*	(Latinx − Indian)	0.36	(M − F)	4.81*	(White female − Indian female)
MWF	US	0.82	(Latinx − Indian)	0.22	(M − F)	1.46*	(Latinx male − Indian female)
MWF	UK	0.58	(Asian − Indian)	0.15	(M − F)	0.87	(Asian male − Indian male)
GEMINI	Indian	0.41	(Indian − White)	0.37	(F − M)	0.77	(Latinx female − White male)
GEMINI	US	0.17	(Black − White)	0.00	(n/a)	0.30	(Indian female − Indian male)
GEMINI	UK	0.31	(Asian − White)	0.19	(F − M)	0.57	(Asian female − White female)
Babble noise (SNR= 5)							
MWF	Indian	8.46*	(White − Black)	0.54	(M − F)	12.01*	(White male − Black male)
MWF	US	1.75*	(Latinx − Asian)	0.18	(F − M)	5.29*	(Black female − Black male)
MWF	UK	2.12*	(Latinx − Indian)	0.00	(n/a)	3.07*	(Latinx female − Indian female)
GEMINI	Indian	0.55	(Latinx − Indian)	0.15	(M − F)	0.77	(Latinx male − Indian female)
GEMINI	US	3.31*	(Asian − White)	1.35*	(F − M)	6.22*	(Asian female − White female)
GEMINI	UK	31.18*	(Asian − White)	13.16*	(F − M)	63.40*	(Asian female − Asian male)

Table 11: Static-face (no lip-sync) largest AVSR quality-of-service differences between faces, across ethnicity, gender, and their intersection. Asterisk (*) denotes statistical significance ($p < 0.005$) according to a permutation test (Section 3.2). MWF = MWHISPER-FLAMINGO MEDIUM; GEMINI = GEMINI-2.5-FLASH. M = Male; F = Female.