Multimodal Integration in Audio-Visual Speech Recognition How Far Are We From Human-Level Robustness?

Marianne Schweitzer[1] **Anna Montagnini**[2] **Abdellah Fourtassi**[1] **Thomas Schatz**[1]
[1]Aix Marseille Univ, CNRS, LIS [2]Aix Marseille Univ, CNRS, INT, Inst Neurosci Timone
asouvleris2000@gmail.com anna.montagnini@univ-amu.fr
abdellah.fourtassi@gmail.com thomas.schatz@univ-amu.fr

Abstract

This paper introduces a novel evaluation framework, inspired by methods from human psychophysics, to systematically assess the robustness of multimodal integration in audiovisual speech recognition (AVSR) models relative to human abilities. We present preliminary results on AV-HuBERT [Shi et al., 2022a,b] suggesting that multimodal integration in state-of-the-art (SOTA) AVSR models remains mediocre when compared to human performance and we discuss avenues for improvement.

1 Introduction

Humans are capable of near-optimal multimodal integration, seamlessly combining auditory and visual information to perceive speech robustly, even in challenging environments featuring large and unpredictable moment-to-moment changes in the reliability of the information coming from different modalities [Alais and Burr, 2004, Ernst and Banks, 2002, Bejjanki et al., 2011]. While building artificial systems capable of human-level speech recognition had long seemed out of reach, deep learning approaches have enabled spectacular progress in recent years, including in audio-visual (AVSR) settings. For example, on the LRS3 large-vocabulary AVSR benchmark, from 2018 to 2024, word error rate (WER) has dropped from 7.2% to 0.9% in the audio-visual condition and from 55.1% to 19.1% in the vision-only condition (lip-reading) [Afouras et al., 2018b,a, Shillingford et al., 2018, Ma et al., 2023]. Is there much room left for further progress? In this paper, we ask whether SOTA AVSR systems have already reached—or even exceeded—the multimodal integration capabilities displayed by humans in challenging real-world environments. We introduce a novel evaluation metric, inspired by methods from the field of human psychophysics, to characterize a model's multimodal integration capabilities. The general idea is to create an audio-visual continuum between two speech syllables belonging to different phonetic categories (for this abstract, we focused on /ba/ vs /da/) and have the model categorize the elements from the continuum based (i) solely on audio, (ii) solely on video and (iii) based on both audio and video. Using the observed performances in the two unimodal conditions and a simple parametric model of categorical perception, we are then able to predict the expected performance in the bimodal condition for a model performing *optimal* multimodal integration, to which we compare the observed performance of the model in the bimodal condition. This metric provides a useful complement to more classical, broader metrics like WER. Although, unlike WER, it does not directly measure performance in a practical application, it enables focusing on a model's multimodal integration capabilities by disentangling them from the quality of the model's unimodal processing. Furthermore, it allows explicitly situating a system's performance relative to optimal multimodal integration. In the following, we present our method, as well as preliminary results on the AV-HuBERT model [Shi et al., 2022a,b] and a comparison with human performance. Although important controls—which we describe—need to be carried out before a

NeurIPS 2024 Workshop on Behavioral Machine Learning.

definitive conclusion can be reached, these results suggest that SOTA AVSR systems are still a far cry from human-level of robustness and, in particular are not robust to noise in the visual modality. We conclude by briefly discussing the implications of our results for the design of future AVSR systems.

2 Related work

On one hand, multimodal integration in humans has been extensively studied, beginning with the seminal works of Alais and Burr [2004], Ernst and Banks [2002]. On the other hand, recent advances have been made in the development of artificial audiovisual speech recognition systems [Afouras et al., 2018a, Shillingford et al., 2021, Shi et al., 2022a,b, Ma et al., 2023]. However, to the best of our knowledge, no direct comparisons have yet been made between the multimodal integration capabilities of these new models and those of humans.

3 Method

We first describe the probing stimuli we use, then we explain how we measured multimodal integration accuracy in humans and, finally, we describe how we adapted the procedure to evaluate multimodal integration in machines.

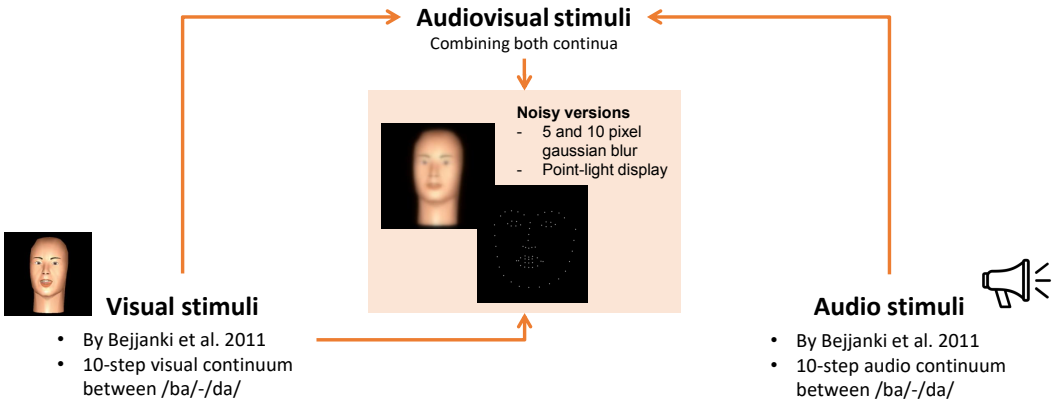

Figure 1: Probing stimuli used to evaluate AVSR models' multimodal integration capabilities. (The Point-Light Display stimulus' size has been increased for better visualization. The original size is the same as for the other stimuli.)

3.1 Probing stimuli

As illustrated in Figure 1, we adapted stimuli from Bejjanki et al. [2011] to create audio, visual, and audiovisual stimuli. Audio stimuli were sampled on a continuum ranging from /ba/ to /da/, synthesized from recordings by a male American English speaker. Visual stimuli also depicted a synthetic continuum from the visual pronunciation of /ba/ to /da/, animated based on facial feature parameters. Audio and visual continua were combined into 100 audiovisual stimuli after aligning them using audio onset detection. Two levels of gaussian blur (5 and 10 pixel radii) were applied to introduce visual uncertainty. We also created Point-Light Display (PLD) versions of the visual and audiovisual stimuli by drawing facial landmarks of the synthetic faces on a black background, to explore model responses to extreme stimulus degradation that are well-tolerated by humans [Johansson, 1973].

3.2 Human experiment

Five participants recruited on the Prolific platform [Prolific Academic] performed an online categorical judgment task on the stimuli described above (original and degraded versions). Since we tested an AVSR system trained on American English, we selected only native English speakers from the United States who grew up monolingual. We selected participants who were not dyslexic and had

normal or corrected vision and hearing. Some participants were excluded due to failing attention checks or exhibiting abnormal behavior during the experiments, such as unusually slow reaction times or extreme results (and are not counted in the five participants above). Although the number of participant was relatively small, we collected enough data (6500 categorical judgedments per participant on average) to enable within-subject analyses. Stimuli were presented multiple times in random order to estimate the average probability of a /ba/ answer for each participant and each stimulus. We measured participants' unimodal performances (audio-only and visual-only stimuli) and their bimodal performance (including with incongruent audiovisual stimuli). Participants' performance in the unimodal conditions was consistent with expectations, with average categorical thresholds around position 5-6 of the 10-step continuum in all conditions and considerable variance in threshold position from participant to participant, reflecting individual differences in sensory processing.

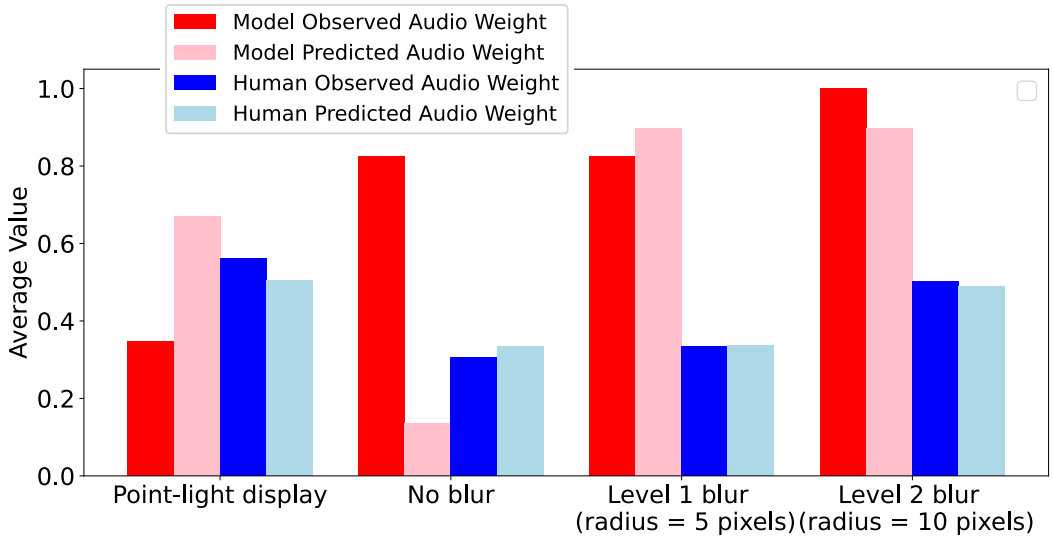

Figure 2: Predicted and observed audio weights across conditions (human and last layer of AV-HuBERT)

To assess the participants' audiovisual integration accuracy, we follow the methodology of Bejjanki et al. [Bejjanki et al., 2011], utilizing a simple parametric model of multimodal categorical perception that has been effective in evaluating human multimodal perception [Ernst and Banks, 2002]. This model operates under the assumption that noise in the auditory and visual modalities is independent. An optimal multimodal integrator, according to this model, weighs each modality (audio or visual) according to its sensory reliability. More specifically, the optimal weight for the audio modality is given by the following formula:

$$\text{Weight}_{\text{audio}} = \frac{\text{Reliability}_{\text{audio}}}{\text{Reliability}_{\text{audio}} + \text{Reliability}_{\text{visual}}}. \quad (1)$$

This formula expresses how much weight should be given to the audio input compared to the visual input, based on their respective reliabilities (the optimal weight for the visual modality is simply $\text{Weight}_{\text{video}}$ equal to $1 - \text{Weight}_{\text{audio}}$).

To compute the optimal audio weight for a given participant, we thus only need to know their sensory reliability for each modality. We obtain these unimodal reliabilities through the unimodal categorization tasks. For these tasks, we fit standard psychometric functions [Wichmann and Hill, 2001] to the categorical judgments, estimating parameters such as the mean, variance, guess rate, and lapse rate. These parameters then allow us to calculate the sensory reliability for each modality.

Next, to evaluate the participants' multimodal integration accuracy, we compare the optimal weight obtained for each participant to the 'observed' weight. The observed weight is derived from fitting a two-dimensional psychometric function to the categorical judgments obtained in the bimodal condition, which includes both audio and visual inputs. The psychometric function is extended to handle bi-dimensional input data and a parameter representing the observed weight assigned to the

audio in the integration process is introduced as in Bejjanki et al. [2011]. By comparing the optimal weight (calculated from unimodal reliabilities) with the observed weight (obtained from the bimodal condition), we can evaluate how accurately participants integrate audiovisual information.

3.2.1 Model probing

We can apply the analysis procedure outlined above for human data to an AVSR model as long as we are able to generate /ba/ vs /da/ judgments in the unimodal and bimodal conditions from the model. Although human judgments are probabilistic (i.e. we estimate the probability that a participant answers /ba/ for a given stimulus), models may only provide a deterministic representation for a fixed stimulus. To generate probabilistic categorical judgments comparable to those obtained from humans using AVSR models, we used representations obtained at a fixed layer of the target AVSR model to train logistic regression classifiers using a set of naturally spoken /ba/ and /da/ sounds. We then leverage Luce's celebrated choice model [Luce, 1959] to obtain probabilistic categorical judgements from these classifiers for the experimental stimuli from Section 3.1. To train the logistic classifiers, the AVSpeech dataset [Ephrat et al., 2018] was used, focusing on English utterances and ensuring clean audio for transcription. Audiovisual files were transcribed using Whisper [Radford et al., 2022] initial transcription and the Montreal Forced Aligner for precise phoneme alignment. From this process, a balanced dataset of 98 instances each of /ba/ and /da/ syllables was extracted.

4 Experiment and preliminary results

We present preliminary results obtained with the AV-HuBERT AVSR model [Shi et al., 2022a,b] and focusing on representation extracted from its last Transformer layer (we obtained similar results at other layers). We use the Noise-Augmented AV-HuBERT Large which is pre-trained in a self-supervised fashion on LRS3 and VoxCeleb2 datasets before being finetuned for AVSR on 433 hours of LRS3 data. This model has a WER of 1.4% on the LRS3 AVSR benchmark, the second best reported to date. As can be seen on Figure 2, the weight given to audio by human participants closely tracks the optimal weight, whereas the AV-Hubert model is less accurate, clearly putting too much weight on audio in the clean condition and too little in the PLD condition. Thus, while the AV-HuBERT model exhibits considerable robustness with traditional performance assessment methods, it falls short of human performance in psychophysical tasks assessing audiovisual integration performance.

5 Conclusion

Our preliminary results suggest that the AV-HuBERT model is far from performing multimodal integration optimally. Like other SOTA AVSR models [Ma et al., 2023], it is not explicitly trained to withstand visual degradation. This suggests testing whether AVSR models are also suboptimal for multimodal integration in the presence of audio noise (which they are explicitly trained to withstand [Shi et al., 2022a,b, Ma et al., 2023]) and whether including visual noise in existing training procedures for AVSR models is sufficient to obtain better results. If not, new training approaches may be required. Methods based on modality weighting through explicit dynamic estimation of modality reliability—which are inspired by observation of human abilities and were popular before the deep learning era [Abdelaziz et al., 2015] but do not currently lead to SOTA performance on standard AVSR benchmarks [Yu et al., 2022]—could be more robust to the moment-to-moment unpredictable visual degradation we consider, for example.

Future work will carry out important controls to confirm our preliminary results. We will add training examples to the logistic regressions, test additional contrasts beyond /ba/-/da/ and compare different AVSR models. We will also test contrasts obtained from ecological rather than synthetic stimuli to assess whether the synthetic nature of the probing stimuli can explain at least part of our results. Furthermore we will test if humans adapt to the probing stimuli by looking at the dynamic of human judgments throughout the experiment. If we do, we will leverage the possibility of online human experimentation to carry out a single trial experiment on a large number of participants to avoid any possibility of adaptation and provide a fairer comparison to the AVSR models. Finally, an important caveat is that although human participants appear to find optimal weights on average, they do not find them individually. This is a well-known result from the literature on human perception [Ernst and Banks, 2002]. A fair comparison with models therefore will require comparing human weights to

weights obtained from a population of AVSR models—for example models trained with different parameter initialization or on different sub-samples of a training set.

Acknowledgments and Disclosure of Funding

This work received support from the French government under the France 2030 investment plan, as part of the Initiative d'Excellence d'Aix-Marseille Université – A*MIDEX AMX-21-PEP-021. This work, carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government (France 2030), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

References

A. H. Abdelaziz, S. Zeiler, and D. Kolossa. Learning dynamic stream weights for coupled-hmm-based audio-visual speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(5): 863–876, 2015.

T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *arXiv preprint arXiv:1809.02108*, 2018a. doi: https://doi.org/10.1109/TPAMI.2018.2889052.

T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018b. doi: https://doi.org/10.48550/arXiv.1809.00496.

D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262, 2004.

V. R. Bejjanki, M. Clayards, D. C. Knill, and R. N. Aslin. Cue integration in categorical tasks: Insights from audio-visual speech perception. *PLOS ONE*, 2011. doi: https://doi.org/10.1371/journal.pone.0019812.

A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. doi: https://doi.org/10.1145/3197517.3201357.

M. O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973.

R. D. Luce. *Individual choice behavior*, volume 4. Wiley New York, 1959.

P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Panticn. Auto-avsr: Audio-visual speech recognition with automatic labels. *arXiv preprint arXiv:2303.14307*, 2023. doi: https://doi.org/10.48550/arXiv.2303.14307.

Prolific Academic. Prolific academic. `https://www.prolific.com/`. Accessed: 2024-07-15.

A. Radford, J. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv:2212.04356*, 2022. doi: https://doi.org/10.48550/arXiv.2212.04356.

B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022a.

B. Shi, W.-N. Hsu, and A. Mohamed. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*, 2022b.

B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, L. B. Kanishka Rao, M. Mulville, B. Coppin, B. Laurie, A. Senior, and N. de Freitas. Large-scale visual speech recognition. *arXiv preprint arXiv:1807.05162*, 2018. doi: https://doi.org/10.48550/arXiv.1807.05162.

B. Shillingford, I. A. Assael, M. Hoffmann, T. L. Paine, C. Hughes, U. Prabhu, and N. Freitas. Large-scale visual speech recognition. *Advances in Neural Information Processing Systems*, 34:7639–7649, 2021.

F. A. Wichmann and N. Hill. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63:1293–1313, 2001.

W. Yu, S. Zeiler, and D. Kolossa. Reliability-based large-vocabulary audio-visual speech recognition. *Sensors*, 22(15):5501, 2022.