What does an Audio-Visual Speech Recognition Model know about Visemes?

Aristeidis Papadopoulos, Naomi Harte

Sigmedia Group, School of Engineering, Trinity College Dublin, Ireland

papadoar@tcd.ie, nharte@tcd.ie

Introduction

Phonemes are the basic speech unit with numerous studies exploring the inner workings of end-to-end transformer-based speech models, but they have mainly focused on Audio Speech Recognition (ASR). These studies have shown that there is significant phoneme capturing and encoding within the encoder layers. The methodologies found in the literature include probing [1] and the use of similarity measures [2], among others.

Considerably less investigation into the interpretability of Audio-Visual Speech Recognition (AVSR) models has been done. In particular, no work has explored what AVSR models learn about visemes, the visual equivalent of phonemes. Our work therefore utilizes the concepts developed for ASR and applies them to AV-HuBERT, where a thorough analysis is performed to establish what the model learns about visemes.

Methodology

For our probing experiment, the LRS3 [3] dataset was used. To match the phonemes to visemes, the Montreal Forced Aligner (MFA) was employed, as the timestamps and the phonemes for each utterance were needed. Then, the mapping proposed by Lee [4] was applied to the phonemic transcription to yield visemes. To reduce the impact of co-articulation, the first and last third of the frames for a given phoneme were discarded. The remaining frames were averaged, resulting in one feature vector per viseme. Three experiments were conducted, with features being extracted with AV input, video input, and AV input with babble noise at -5dB and used to train the probes. The base AV-HuBERT, fine-tuned for AVSR, was used for the feature extraction. The layout of the probes is inspired by the work of [1], with a hidden layer of size of 200 units, learning rate equal to 0.001 and with early stopping enabled.

Results

Fig.1 presents the viseme classification accuracy of our probing experiment, for the different types of input we tested. A consistent, rising trend is observed for all inputs, indicating that viseme information is encoded well in the extracted features. The inclusion of audio introduces helpful information, even in the case of noisy input, highlighting the complementary nature of the two modalities. Notably, further analysis of our results reveals that the least visible visemes benefit the most from the presence of audio. This effect is presented in Fig. 2, where the F1-Scores for two visemes are compared, 'P', (/p/, /b/, /m/), which is a highly visible viseme, and 'K', (/k/, /g/, /ng/, /n/, /l/, /y/, /hh/), one of the least visible.

Conclusion

By probing into the features of AV-HuBERT, we have shown that there is significant viseme presence in the hidden embeddings. We will expand our work by visualizing the extracted features and by performing an in-depth analysis of our results from our probing experiment.

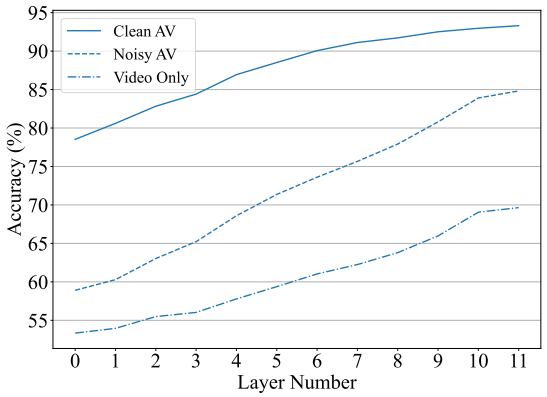

Figure 1: *Viseme Accuracy per Layer for the LRS3 Test Set*

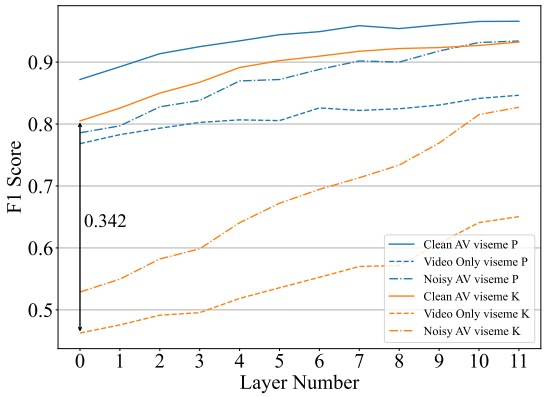

Figure 2: *F1 Scores for visemes 'P' and 'K'*

1. References

[1] P. Cormac English, J. D. Kelleher, and J. Carson-Berndsen, "Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features," in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Seattle, Washington: Association for Computational Linguistics, 2022, pp. 83–91.

[2] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-Wise Analysis of a Self-Supervised Speech Representation Model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Cartagena, Colombia: IEEE, Dec. 2021, pp. 914–921.

[3] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," 2018.

[4] S. Lee and D. Yook, "Audio-to-visual conversion using hidden markov models," in *PRICAI 2002: Trends in Artificial Intelligence*. Berlin: Springer, 2002, pp. 563–570.