The Importance of Noise in Audiovisual Learning: An Artificial Neural Network Simulation of the McGurk Effect

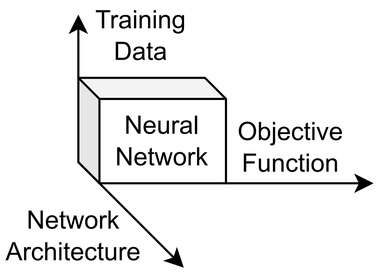
Lukas Grasse, Matthew Tata

Canadian Centre for Behavioural Neuroscience, University of Lethbridge

Introduction

Training ANNs to replicate human perception enables researchers to investigate why our perceptual mechanisms might behave in particular ways [3]. This study explores the McGurk effect: an auditory-visual illusion wherein incongruent inputs lead to a fused, but incorrect, auditory percept.

Three major aspects of ANNs can be modified to test whether certain conditions are required for them to exhibit human behaviour or dynamics. While manipulating these conditions in humans is not feasible, it is straightforward to do so with ANNs.



This study manipulated the *Training Data* dimension to test whether the presence of noise influences the fusion of auditory and visual cues during the McGurk effect.

Methodology

- Audiovisual dataset of nine word pairs based on [2]
- Tested humans and ANNs trained on audiovisual speech
- Human participants [n=14] selected perceived word in forced-choice task
- K-nearest neighbours classifier on ANN output embeddings for ANN forced-choice task
- Modulated noise level during CPC network training to test the impact of increased noise during *learning*

AudioVisual Networks

Deep AVSR (2018)[1]

- **Training Task:** Supervised Speech Recognition, Curriculum Learning
- **Training Noise:** Babble, *25% probability*

AV-HuBERT (2022)[4]

- **Training Task:** Self-Supervised Learning, Speech Recognition Fine-tuning
- **Training Noise:** Natural Sounds, Music, Babble, Speech, *25% probability*

AudioVisual CPC (2024, Our Network)

- **Training Task:** Self-Supervised Learning, Contrastive Predictive Coding
- **Training Noise:** Babble, *0-100% probability*

McGurk Forced-Choice Task Results

Network	Deep AVSR		AV-Hubert SSL		AV-Hubert SR	**AudioVisual CPC**	
Training Noise Condition	Clean	Noisy	Clean	Noisy	Noisy	Clean	Noisy
Accuracy	65.22%	80.47%	94.48%	90.81%	92.99%	95.97%	92.26%
Bootstrapped 95% CI	± 6.12%	± 4.39%	± 2.96%	± 3.48%	± 3.35	± 2.02	± 3.52%

Table 1. 10-Fold Cross-Validation Accuracy on Congruent Stimuli Training Set

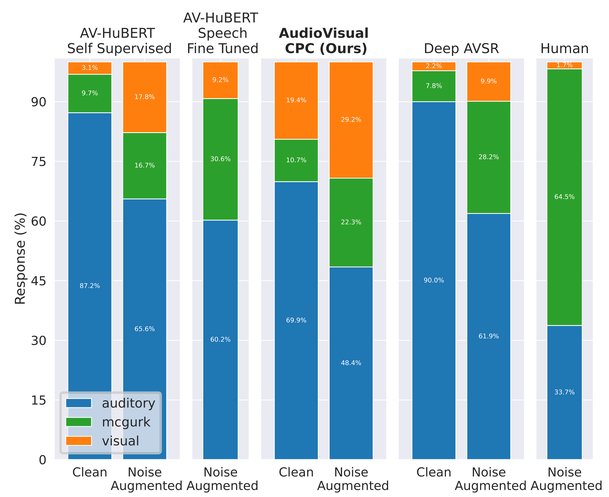


Figure 1. McGurk Stimuli Classification Results

Impact of Increased Noise during Training (not Testing!)

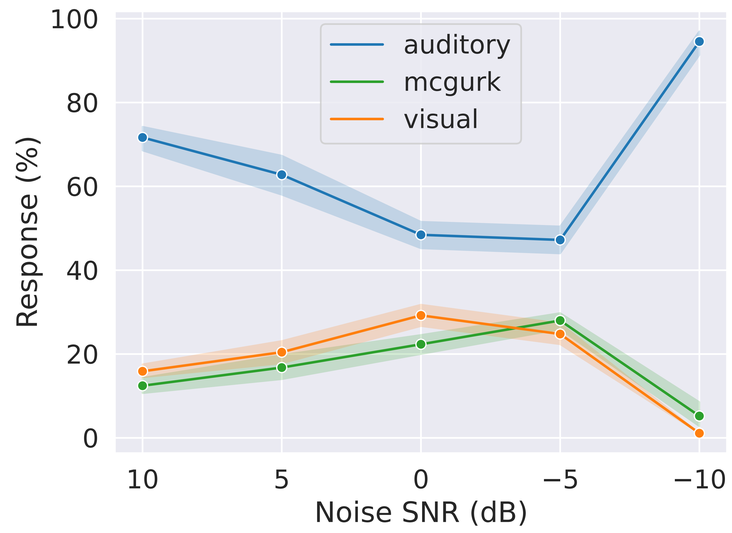


Figure 2. Audio-Visual CPC Network Performance Under Increased Levels of Training Noise

Discussion

- ANNs exhibit the McGurk effect under certain circumstances
- Training on audiovisual speech with noisy audio is important for replicating the illusion
- Biologically plausible self-supervised training most closely approached human performance on McGurk effect
- At very high levels of noise during training the McGurk effect disappears, suggesting that severe noise during development might impair cross-modal perception
- There is still a gap between ANN and human behaviour on the McGurk effect even with ANNs approaching human-level ability on audiovisual speech recognition tasks

References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *IEEE Trans. Pattern Analysis. Mach. Intell.*, 2018.

[2] D. J. Dekle, C. A. Fowler, and M. G. Funnell. Audiovisual integration in perception of real words. *Perception & Psychophysics*, 1992.

[3] N. Kanwisher, M. Khosla, and K. Dobs. Using artificial neural networks to ask 'why'questions of minds and brains. *Trends in Neurosciences*, 2023.

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