A Deep Neural Network Trained on Congruent Audiovisual Speech Reports the McGurk Effect

Haotian Ma, Zhengjia Wang, Xiang Zhang, John F Magnotti, Michael S Beauchamp

Department of Neurosurgery, Perelman School of Medicine, University of Pennsylvania,

Philadelphia, PA

Abstract

In the McGurk effect, incongruent auditory and visual syllables are perceived as a third, illusory syllable. The prevailing explanation for the effect is that the illusory syllable is a consensus percept intermediate between otherwise incompatible auditory and visual representations. To test this idea, we turned to a deep neural network known as AVHuBERT that transcribes audiovisual speech with high accuracy. Critically, AVHuBERT was trained only with *congruent* audiovisual speech, without exposure to McGurk stimuli or other incongruent speech. In the current study, when tested with congruent audiovisual "ba", "ga" and "da" syllables recorded from 8 different talkers, AVHuBERT transcribed them with near-perfect accuracy, and showed a human-like pattern of highest accuracy for audiovisual speech, slightly lower accuracy for auditory-only speech, and low accuracy for visual-only speech. When presented with incongruent McGurk syllables (auditory "ba" paired with visual "ga"), AVHuBERT reported the McGurk fusion percept of "da" at a rate of 25%, many-fold greater than the rate for either auditory or visual components of the McGurk stimulus presented on their own. To examine the individual variability that is hallmark of human perception of the McGurk effect, 100 variants of AVHuBERT were constructed. Like human observers, AVHuBERT variants was consistently accurate for congruent syllables but highly variable for McGurk syllables. Similarities between the responses of AVHuBERT and humans to congruent and incongruent audiovisual speech, including the McGurk effect, suggests that DNNs may be a useful tool for interrogating the perceptual and neural mechanisms of human audiovisual speech perception.

Introduction

Humans perceive speech by integrating auditory and visual information from the face and voice of the talker. The powerful influence of visual information on speech perception is demonstrated by an illusion known as the McGurk effect, in which the pairing of visual “ga” with auditory “ba” results in the percept of “da” (McGurk & MacDonald, 1976). The illusory “da” percept is referred to as a "fusion" percept because it corresponds to neither of the unisensory components (“ga” or “ba”). In the decades since its discovery, the McGurk effect has become a staple of psychology teaching and research, with the original paper cited nearly 10,000 times.

Most existing models of the McGurk effect adopt a Bayesian, causal inference approach (Aller & Noppeney, 2019; Magnotti & Beauchamp, 2015, 2017; Olasagasti & Giraud, 2020; Parise, 2024; Parise & Ernst, 2023). In these models, observers make decisions about the likelihood of the incongruent auditory and visual speech arising from a single talker and integrate the modalities accordingly. However, the models are applicable only to a restricted stimulus set (such as the syllables “ba”, “da” and “ga”) and require assumptions about the relationship between the three syllables, such as a representational space in which “da” is intermediate between “ba” and “ga”.

In contrast, deep neural networks (DNNs) take a completely different approach to modeling speech perception. Instead of being restricted to small stimulus sets, DNNs are trained to decode arbitrary speech. Instead of requiring a pre-specified representational space, DNNs learn the properties of speech through training with hundreds of hours of audiovisual speech, a process that is, in some respects, similar to the way that human children learn language (Hosseini et al., 2024).

DNNs are the state-of-the-art for automatic speech recognition (Roger et al., 2022) and a DNN known as AVHuBERT (Audiovisual Hidden-unit Bidirectional Encoder Representations from Transformers) is able to transcribe audiovisual speech with a 1.4% word error rate (Shi et al., 2022). AVHuBERT has a number of properties that make it well-suited for modeling the McGurk effect. Unlike models of the McGurk effect that rely on abstract descriptions of speech (such as the identity of a particular input as "auditory ba" or "visual ga"), DNNs are stimulus computable: any speech stimulus that can be presented to a human observer can also be processed by a speech DNN. While most speech DNNs are auditory-only, AVHuBERT considers both auditory and visual speech, meaning that, in principle, it receives the information necessary to form the illusory fusion percept from McGurk syllables.

DNNs adjust their model weights during training on examples of the material of interest; in the case of AVHuBERT, the training data consisted of many hours of congruent audiovisual speech (Shi et al., 2022). Hypothetically, this training could produce representations of auditory and visual features in AVHuBERT that would result in fusion responses to McGurk stimuli, similar to the processes occurring during human perception. It is important to emphasize that only *congruent* speech was used to train AVHuBERT: it would be uninformative to study a model that had been explicitly trained to produce a fusion response to McGurk stimuli. If they do report the McGurk effect, DNNs might provide a useful modeling framework to understand human perception of the illusion. For instance, different human participants vary dramatically in their susceptibility to the illusion (Basu Mallick et al., 2015; Dong et al., 2025; Magnotti et al., 2024, 2025; Strand et al., 2014) and it has been proposed that variation among similar DNNs might serve as a model for individual variability among human participants (Mehrer et al., 2020).

Methods

Speech stimuli

There were 32 total stimuli, 4 stimuli each from 8 different talkers (4 male, 4 female). There were three congruent stimuli: auditory “ba” paired with visual “ba” (AbaVba); AgaVga; AdaVda; and one McGurk stimulus (AbaVga).

Response scoring

Each speech stimulus was presented to AVHuBERT using the default processing pipeline described in the AVHuBERT documentation (*https://github.com/facebookresearch/av_hubert*). For each stimulus, AVHuBERT generated five predictions and the likelihood of each one. As in Basu Mallick et al (2015), each response was classified as one of four response categories: "fusion"; "auditory"; "visual" and "other". Classification was done automatically based on a phonetic description provided by the Carnegie Mellon pronouncing dictionary, as described in (Yu et al., 2024). Only the first phoneme in the response was considered. Initial phonemes “D”, “T”, “DH” and “TH” were classified as fusion responses, initial phonemes “B” and “P” were classified as auditory responses, initial phonemes “G” and “K” were classified as visual responses, and any other initial phoneme was classified as “other”. The response likelihoods were then converted to percentages. For instance, for a representative visual-only “ga” stimulus, AVHuBERT produced the following five predictions and associated likelihoods:

```
aha (0.43), ah (0.43), khah (0.40), huh (0.37), aha (0.31)
```

Converting these to percentages resulted in:

```
Fusion (0%), Auditory (0%), Visual (21%), Other (79%)
```

In an additional analysis, the “other” responses were divided into two subcategories based on whether the first phoneme in the response did (or did not) contain an initial consonant.

Because AVHuBERT selects its responses from a dictionary of 1000 possible sub-tokens, the chance rate was determined by randomly selecting sub-tokens from the dictionary. Mapping each sub-token to one of the four response categories resulted in base rates of 14% for “da”, 11% for “ba”, 12% for “ga" and 63% for “other”.

AVHuBERT model and model variants

The original, published AVHuBERT-large model (Shi et al., 2022) was downloaded from GitHub (*https://github.com/facebookresearch/av_hubert*, fixed version archived to read-only on Sept 1, 2024) and installed on an Apple Mac Studio with 192 GB of RAM.

The AVHuBERT architecture consists of four functional parts: separate auditory and visual feature extractors; 24-layers of transformer encoders that combine the auditory and visual features; and a sequence-to-sequence decoder that produces transcribed output in the form of text tokens.

As a substitute for the resource-intensive process of retraining AVHuBERT with hundreds of hours of video to create model variants, slight perturbations were introduced into the first six transformer layers (the model layers most important for integrating auditory and visual speech information). The perturbations were restricted to the linear output layer (“out_proj”) of the transformer encoders (Vaswani et al., 2017). A normal distribution with mean of zero and standard deviation of 0.1 was used as the noise generator, resulting in only small changes to the model weights. One hundred unique random number seeds were used to create 100 model variants: this assured that each model variant contained different model weights, but that the variants could be uniquely identified and reproduced if necessary.

Comparison with human responses

For comparison with the AVHuBERT responses to McGurk stimuli, we reanalyzed data from human participants collected for the open choice portion of Experiment 2 of Basu Mallick et al. (2015). Identical McGurk stimuli were tested in humans and in AVHuBERT and AVHuBERT model variants, and the exact same response scoring procedure was applied.

Results

Syllables were presented to the original AVHuBERT model and one hundred variants. The original model and the variants identified each of the congruent audiovisual syllables (AbaVba, AgaVga and AdaVda) stimuli from 8 different talkers with near-perfect accuracy (Figure 1A). Averaging across syllables and then the 100 model variants, the mean accuracy was 99 ± 0.05% (mean ± SEM across variants). Accuracy was significantly lower for unisensory syllables (Figure 1B), with an accuracy of 94% ± 0.06% for auditory-only syllables and 43% ± 0.3% for visual-only syllables (paired *t*-tests: auditory-only *vs.* audiovisual, 94% *vs.* 99%, $t_{99} = 60$, $p < 10^{-16}$; visual-only *vs.* audiovisual, 43% *vs.* 99%, $t_{99} = 210$, $p < 10^{-16}$; auditory-only *vs.* visual-only, 94% *vs.* 43%, $t_{99} = 191$, $p < 10^{-16}$).

When presented with McGurk syllables consisting of AbaVga, the original AVHuBERT model reported the fusion percept of *da* at a rate of 25%. Across model variants, the mean rate of fusion percepts was 21% ± 0.8% (mean ± SEM across variants) significantly above the chance rate of fusion responses assuming random responding (one-sample t-test, $t_{99} = 12$, $p < 10^{-16}$). The rate of fusion responses to the McGurk syllable was greater than to either unisensory component (Figure 1C). Auditory-only "ba" evoked the fusion percept at a rate of 0.2% and visual-only "ga" evoked the fusion percept at a rate of 6%, both significantly less than the 21% rate of fusion responses to the combined AbaVga syllable (0.2% *vs.* 21%, $t_{99} = 28$, $p < 10^{-16}$; 6% *vs.* 21%, $t_{99} = 21$, $p < 10^{-16}$).

Variability

There was significant variability in the rate of fusion percepts across AVHuBERT model variants, ranging from 6% to 50% (Figure 1D). There was also high variability in the efficacy of the different McGurk stimuli, ranging from 0% for the weakest stimulus to 83% for the strongest

stimulus. A linear mixed-effect model with % “da” responses as the dependent measure, stimulus as fixed effect, and model variant as random effect showed a significant effect of the stimulus factor ($X^2_7 = 2198$, $p < 10^{-16}$), indicating that stimulus differences were consistent across model variants.

Comparison with human responses to the same McGurk stimuli

In Basu Mallick et al (2015), one hundred human participants were presented with the same McGurk stimuli as used in the present study. Data from the human participants were re-analyzed using the same processing pipeline developed for AVHuBERT.

For identical McGurk stimuli, AVHuBERT model variants had a lower mean rate of fusion percepts than humans (21% *vs.* 41%, $p = 10^{-7}$ from unequal variance *t*-test) and a lower standard deviation of fusion rates across model variants and human participants (8% *vs*. 34%, $p < 10^{-16}$ from Levene’s test). Across the 8 different McGurk stimuli, the correlation between the rate of fusion responses for AVHuBERT model variants and humans was low ($r = 0.18$, $p = 0.67$).

The rate of auditory responses to McGurk stimuli was similar between AVHuBERT and humans (41% *vs*. 44%) and visual responses were rare for both (0.3% *vs*. 7%). AVHuBERT had much higher rates of “other” responses to the McGurk stimuli than human observers, 38% *vs*. 8% ($p < 10^{-16}$). Interestingly, for both AVHuBERT and humans, these “other” responses often consisted of vowels without any consonant, such as “ah”. This was true for 64% of “other” responses for AVHuBERT and 49% of "other" responses in humans.

Discussion

The responses of AVHuBERT to speech stimuli consisting of single syllables showed many similarities to human observers. For AVHuBERT, receiving visual information about the face of the talker increased accuracy: recognition was most accurate for audiovisual syllables, less accurate for auditory-only syllables, and least accurate for visual-only speech. This same pattern has been observed in dozens of studies of human speech perception, likely reflecting a fundamental property of speech: while auditory speech cues convey more information about speech content than visual speech cues, the modalities are not redundant so that combining them enhances perception (Peelle & Sommers, 2015; Ross et al., 2007; Sumby & Pollack, 1954).

The main finding of the study was that when presented with the incongruent syllable pairing of auditory "ba" and visual "ga", AVHuBERT reported the fusion percept of "da" that defines the McGurk effect (McGurk & MacDonald, 1976). The rate of fusion percepts was much greater than chance, and greater than the rate of fusion percepts for either auditory "ba" or visual "ga" presented on their own. Different model variants of AVHuBERT were created to serve as a model for different human participants (Mehrer et al., 2020), and like human participants, the variants differed in their susceptibility to the effect, while showing consistently high accuracy at identifying *congruent* audiovisual speech.

Explaining the McGurk effect with Bayesian models vs. DNNs

AVHuBERT is not the only model capable of predicting the McGurk effect. Most current models of the McGurk effect operate in the framework of Bayesian causal inference (Aller & Noppeney, 2019; Magnotti & Beauchamp, 2015, 2017; Olasagasti & Giraud, 2020; Parise, 2024; Parise & Ernst, 2023) making it worthwhile to compare the benefits of the two types of models.

Interpretable parameters vs. connection weights: Compared with DNNs, Bayesian causal inference models have the advantage that they rely on a small number of interpretable

parameters, such as a parameter that describes the disparity between the auditory and visual cues in a particular stimulus, and another parameter for the disparity threshold of a particular observer (Magnotti & Beauchamp, 2015). In contrast, DNNs have millions of parameters (the connection weights between model nodes) that are not easily interpretable. Bayesian causal inference models that have been fitted to human data can be applied to untested stimuli or observers to predict perception, and are effective across cultures, languages and patient populations (Magnotti et al., 2024, 2025; Stropahl et al., 2017). In contrast, DNNs are limited by the domain of their training data (for instance, AVHuBERT was only trained on English-language videos and hence is not effective in other languages) and it is not clear how DNNs could be used to predict perception of a particular stimulus by a particular observer.

Explicit vs. implicit representational spaces: One disadvantage of Bayesian causal inference models is they require explicit assumptions about the representational space used by humans during speech perception. The original description of the McGurk effect proposed an explanation grounded in perceptual processing: “da” is pronounced with an open mouth, compatible with the visual properties of “ga”, while in the auditory domain, the acoustics of “da” are similar to “ba”. Thus, the fusion percept represents a compromise between conflicting sensory cues (McGurk & MacDonald, 1976). Bayesian models instantiate this idea by assuming a joint audiovisual representational space in which “da” is intermediate between “ba” and “ga” (Magnotti & Beauchamp, 2017). In contrast, AVHuBERT constructs its own representational space during the training process and does not require any explicit instruction about the identity or relationship between different speech elements. The benefit of this approach is that an open-choice response format may be used. In our data, AVHuBERT often reported vowel-only responses, such as "ah", to McGurk stimuli. These responses are also reported by human participants in open-choice

testing environments. In contrast, Bayesian models constructed with explicit representations of "ba", "da" and "ga" can never predict such responses.

Relationship to neural architecture: There is no simple correspondence between neural structures and the high-level equation(s) that characterize Bayesian models. In contrast, the architecture of DNNs was inspired by biological neural systems, and therefore there is a possible correspondence between them (*e.g.* model units ~ neurons, model weights ~ synaptic strengths). Examining the internal structure of AVHuBERT might give clues to the neural substrates of the McGurk effect, as has been successfully demonstrated for DNNs and convolutional neural networks (CNNs) trained to accomplish visual, auditory, and language tasks (Kell et al., 2018; Kumar et al., 2024; Li et al., 2023; Saddler & McDermott, 2024; Yamins et al., 2014; Yamins & DiCarlo, 2016). Of course, the many possible methodological confounds in comparisons between DNN/CNNs and human brains warrant careful consideration (Dujmovic et al., 2024; Hadidi et al., 2025) and emphasize the diversity of possible neural network architectures (Rideaux et al., 2021).

In summary, Bayesian causal inference models and DNN models both provide insight into the intricacies of the McGurk effect and other aspects of human speech perception. An example of their complementary natures is provided by representational spaces. While Bayesian models require assuming a representational space in which "da" is intermediate between "ba" and "ga", AVHuBERT constructs such a space on its own from the training data, illustrating that it is reasonable that human observers could do the same using the statistical regularities in human auditory and visual speech.

Future Directions: Training individual DNNs to understand human variability

AVHuBERT was selected for our analysis because it is a state-of-the-art audiovisual speech recognizer and is freely available as an archived "frozen" version downloadable by anyone for exact replication. One limitation of the present study is that, due to the immense computational resources (graphical processing units, GPUs) required to train DNNs, AVHuBERT model variants were created by adding noise to the weights of the original model, rather than retraining them *de novo* (Mehrer et al., 2020). However, as GPUs become more widely available, training models *de novo* could provide additional insight in the causes of the McGurk effect.

Retraining AVHuBERT with different forms of audiovisual speech could explain how developmental exposure shapes perception, experiments that are difficult or impossible to conduct in humans (Hosseini et al., 2024). For instance, in the present study, AVHuBERT variants displayed lower rates of fusion percepts and less variability, than human participants. However, all the variants were derived from the original AVHuBERT model, trained on a particular set of audiovisual speech containing only one talker. In contrast, humans often experience environments with multiple talkers, and individual children are exposed to very different language environments. The experimental prediction would be that training AVHuBERT with videos containing multiple faces and a single voice (or vice-versa) would be expected to increase the rate of fusion responses, while training different AVHuBERT variants with different source materials would produce higher variability in McGurk responding.

Implications for theories of human perception of the McGurk illusion

In their paper titled "*Rethinking the McGurk effect as a perceptual illusion*", Getz and Toscano propose that, rather than being due to perceptual processing, reports of the fusion percept are primarily attributable to experimenter expectations, experimental context, and other top-down cognitive factors (Getz & Toscano, 2021). Because AVHuBERT is a purely bottom-up model, its

reports of the McGurk effect cannot be due to the "task demands, emotions and intentions" cited by Getz and Toscano. Getz and Toscano also suggest that the high variability between human participants be taken as evidence for a top-down origin of the fusion percept. However, both Bayesian models (Magnotti et al., 2024; Magnotti & Beauchamp, 2015; Stropahl et al., 2017) and AVHuBERT produce high variability without top-down mechanisms, demonstrating that variability on its own is not evidence for a cognitive explanation for the McGurk effect. Additional evidence for a perceptual basis of the McGurk effect is demonstrated by the discovery that synesthetes experience color based on the integrated fusion percept, not on the unisensory components of the McGurk stimulus (Bargary et al., 2009).

Figures

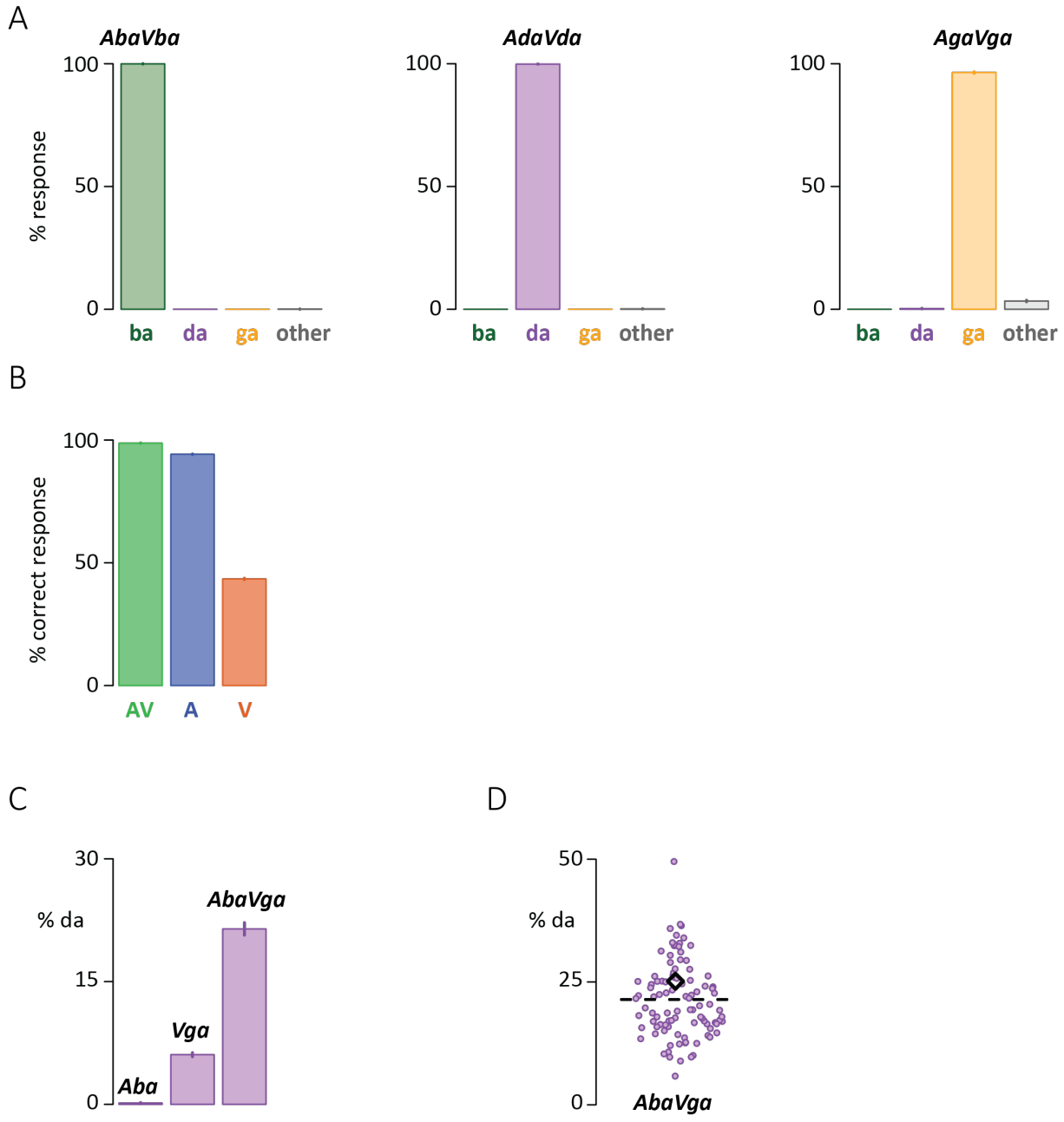

Figure 1

A. Left: One hundred AVHuBERT model variants were presented with congruent audiovisual syllables consisting of auditory "ba" paired with visual "ga" (AbaVba) recorded by 8 different talkers. The responses of AVHuBERT were classified into four different categories based on the

initial phoneme of the response ("ba", "da", "ga", and "other"). The height of the bar shows the percentage of responses in each category (error bar shows SEM across model variants). Green bar shows percent of correct responses, corresponding to accuracy for that stimulus.

Middle: AVHuBERT model variant responses for AdaVda. Purple bar shows rate of correct responses.

Right: AVHuBERT model variant responses for AgaVga. Yellow bar shows rate of correct responses.

B. One hundred AVHuBERT model variants were presented with congruent audiovisual (AV) syllables (AbaVba, AdaVda, AgaVga) from 8 talkers and the accuracy of the responses was calculated (equal to the average of the green, purple and yellow bars in the left, middle and right panels of **A**). The same process was repeated for auditory-only (A) syllables (Aba, Ada, Aga) and visual-only (V) syllables (Vba, Vda, Vga). Error bars shows SEM across model variants.

C. One hundred AVHuBERT model variants were presented with unisensory auditory ba (Aba), unisensory visual ga (Vga) and the pairing of the two (AbaVga) from 8 different talkers. The rate of fusion ("da") responses was calculated. Error bars shows SEM across model variants.

congruent audiovisual (AV) syllables (AbaVba, AdaVda, AgaVga) and the accuracy of the responses was calculated (equal to the average of the green, purple and yellow bars in the left, middle and right panels of **A**). The same calculation was performed for auditory-only (A) syllables (Aba, Ada, Aga) and visual-only (V) syllables (Vba, Vda, Vga).

D. One hundred AVHuBERT model variants were presented with McGurk syllables (AbaVga) from 8 different talkers and the average rate of fusion ("da") responses was calculated for each variant. Each symbol shows one model variant; black diamond shows value for original AVHuBERT model.

Declarations

Funding. This work was supported by the National Institutes of Health, NINDS (Grant No. 5R01NS065395) to MSB.

Conflicts of interest/Competing interests. The authors report there are no conflicts/competing interests to declare.

Ethics Approval. The study was approved by the University of Pennsylvania IRB.

Consent to participate. Informed consent was obtained from all participants included in the study.

Consent for publication. Participants gave written informed consent for publishing the overall results prior to the experiment.

Availability of data and materials. All data reported in this manuscript are included as supplemental material.

Code availability. All analysis code is included as supplemental material.

Open Practices Statement. All data and analysis code necessary to reproduce the results in this manuscript are included as supplemental material.

References

Aller, M., & Noppeney, U. (2019). To integrate or not to integrate: Temporal dynamics of hierarchical Bayesian causal inference. *PLoS Biol*, *17*(4), e3000210. https://doi.org/10.1371/journal.pbio.3000210

Bargary, G., Barnett, K. J., Mitchell, K. J., & Newell, F. N. (2009). Colored-speech synaesthesia is triggered by multisensory, not unisensory, perception. *Psychological Science*, *20*(5), 529–533. https://doi.org/10.1111/j.1467-9280.2009.02338.x

Basu Mallick, D., F. Magnotti, J., & S. Beauchamp, M. (2015). Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, *22*, 1299–1307. https://doi.org/10.3758/s13423-015-0817-4

Dong, C., Wang, Z., Li, R., Noppeney, U., & Wang, S. (2025). Variations in unisensory speech perception explain interindividual differences in McGurk illusion susceptibility. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-025-02697-3

Dujmovic, M., Bowers, J. S., Adolf, F., & Malhotra, G. (2024). IInferring DNN-brain alignment using representational similarity analyses can be problematic. *ICLR 2024 Workshop on Representational Alignment*.

Getz, L. M., & Toscano, J. C. (2021). Rethinking the McGurk effect as a perceptual illusion. *Attention, Perception & Psychophysics*, *83*(6), 2583–2598. https://doi.org/10.3758/s13414-021-02265-6

Hadidi, N., Feghhi, E., Song, B. H., Blank, I. A., & Kao, J. C. (2025). *Illusions of Alignment Between Large Language Models and Brains Emerge From Fragile Methods and Overlooked Confounds* (p. 2025.03.09.642245). bioRxiv. https://doi.org/10.1101/2025.03.09.642245

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2024). Artificial Neural Network Language Models Predict Human Brain Responses to Language Even After a Developmentally Realistic Amount of Training. *Neurobiology of Language (Cambridge, Mass.)*, *5*(1), 43–63. https://doi.org/10.1162/nol_a_00137

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630-644.e16. https://doi.org/10.1016/j.neuron.2018.03.044

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2024). Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, *15*(1), 5523. https://doi.org/10.1038/s41467-024-49173-5

Li, Y., Anumanchipalli, G. K., Mohamed, A., Chen, P., Carney, L. H., Lu, J., Wu, J., & Chang, E. F. (2023). Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, *26*(12), 2213–2225. https://doi.org/10.1038/s41593-023-01468-4

Magnotti, J. F., Basu Mallick, D., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2025). The McGurk effect is similar in native Mandarin Chinese and American English speakers. *Frontiers in Psychology*, *16*, 1531566. https://doi.org/10.3389/fpsyg.2025.1531566

Magnotti, J. F., & Beauchamp, M. S. (2015). The noisy encoding of disparity model of the McGurk effect. *Psychonomic Bulletin & Review*, *22*(3), 701–709. https://doi.org/10.3758/s13423-014-0722-2

Magnotti, J. F., & Beauchamp, M. S. (2017). A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech. *PLoS Comput Biol*, *13*(2), e1005229. https://doi.org/10.1371/journal.pcbi.1005229

Magnotti, J. F., Lado, A., & Beauchamp, M. S. (2024). The noisy encoding of disparity model predicts perception of the McGurk effect in native Japanese speakers. *Frontiers in Neuroscience*, *18*, 1421713. https://doi.org/10.3389/fnins.2024.1421713

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*(1), 5725. https://doi.org/10.1038/s41467-020-19632-w

Olasagasti, I., & Giraud, A.-L. (2020). Integrating prediction errors at two time scales permits rapid recalibration of speech sound categories. *eLife*, *9*, e44516. https://doi.org/10.7554/eLife.44516

Parise, C. V. (2024). *Spatiotemporal models for multisensory integration*. https://doi.org/10.1101/2023.12.29.573621

Parise, C. V., & Ernst, M. O. (2023). *Multisensory integration operates on correlated input from unimodal transients channels*. https://doi.org/10.1101/2023.07.07.548109

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006

Rideaux, R., Storrs, K. R., Maiello, G., & Welchman, A. E. (2021). How multisensory neurons solve causal inference. *Proceedings of the National Academy of Sciences*, *118*(32), e2106235118. https://doi.org/10.1073/pnas.2106235118

Roger, V., Farinas, J., & Pinquier, J. (2022). Deep neural networks for automatic speech processing: A survey from large corpora to limited data. *EURASIP Journal on Audio, Speech, and Music Processing*, *2022*(1), 19. https://doi.org/10.1186/s13636-022-00251-w

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Saddler, M. R., & McDermott, J. H. (2024). Models optimized for real-world tasks reveal the task-dependent necessity of precise temporal coding in hearing. *Nature Communications*, *15*(1), 10590. https://doi.org/10.1038/s41467-024-54700-5

Shi, B., Hsu, W.-N., Lakhotia, K., & Mohamed, A. (2022). *Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction* (No. arXiv:2201.02184). arXiv. http://arxiv.org/abs/2201.02184

Strand, J., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity. *Journal of Speech, Language, and Hearing Research : JSLHR*, *57*(6), 2322–2331. https://doi.org/10.1044/2014_JSLHR-H-14-0059

Stropahl, M., Schellhardt, S., & Debener, S. (2017). McGurk stimuli for the investigation of multisensory integration in cochlear implant users: The Oldenburg Audio Visual Speech Stimuli (OLAVS). *Psychon Bull Rev*, *24*(3), 863–872. https://doi.org/10.3758/s13423-016-1148-9

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (Version 7). arXiv. https://doi.org/10.48550/ARXIV.1706.03762

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yu, Y., Lado, A., Zhang, Y., Magnotti, J. F., & Beauchamp, M. S. (2024). Synthetic faces generated with the facial action coding system or deep neural networks improve speech-in-noise perception, but not as much as real faces. *Frontiers in Neuroscience*, *18*, 1379988. https://doi.org/10.3389/fnins.2024.1379988