First steps in simulating semantic and phonological impairments in aphasia with a computational model of speech processing

Ihintza Malharin (i.malharin@bcbl.eu)[1,2], María Belén Saavedra (belen.saavedra.39@gmail.com)[2], Linkai Peng (fie24002@uconn.edu)[4], Simona Mancini (s.mancini@bcbl.eu)[1,3] & James S. Magnuson (j.magnuson@bcbl.eu, james.magnuson@uconn.edu)[1,3,4]

[1]BCBL. Basque Center on Cognition, Brain & Language, 69 Mikeletegi 20009 Donostia-San Sebastián. Gipuzkoa, Spain
[2]UPV/EHU Euskal Herriko Unibertsitatea, Donostia-San Sebastián, Gipuzkoa, Spain
[3]Ikerbasque. Basque Foundation for Science, Bilbao, Spain
[4]Psychological Sciences, University of Connecticut, Storrs, Connecticut 06269 USA

Abstract

EARSHOT is a model of human speech processing that learns to map real speech to semantic representations (Magnuson et al., 2020). We report first steps towards simulating aspects of aphasia with EARSHOT by damaging increasing proportions of randomly selected weights in different model layers. Although the model is purely receptive, we simulated naming/identification tasks by presenting spoken words to damaged models and measuring the cosine similarity of the output to every word in the lexicon, with the model's naming/identification response operationalized as the word with highest cosine similarity. For errors, we evaluated phonetic and semantic similarity of the response to the target vector. We were interested in how robust EARSHOT would be to damage, and whether we might observe systematic patterns of phonetic vs. semantic deficits following damage to different model components. As expected, lower-level damage impaired phonology most, and higher-level damage most affected semantics. Further development could turn EARSHOT into a valuable tool for enhancing understanding of aphasia.

Keywords: computational modeling;aphasia;semantic similarity;phonological similarity;error types

Introduction

A productive approach to learn about how we think, understand and communicate, is to compare healthy systems with impaired ones. Human language has been the topic of decades of research spanning multiple domains. Whether we are interested in learning more about its smallest grammatical variations, the brain areas recruited when we hear a funny story, or how an adult learns a second language differently from a child, a healthy-impaired comparative approach has potential to illuminate typical processing and development by revealing key or fragile component processes or representations.

Aphasiology, the study of language impairments usually resulting from brain damage, started in the 19th century when Broca studied the now famously known patient 'Tan', named after the only syllable that they could pronounce. Broca discovered post-mortem that this patient's brain had been badly damaged, and suggested that the damaged area in the frontal lobe was responsible for the patient's impaired speech (Maudsley, 1868). *Broca's area* was thus discovered and with it a new way of studying and understanding the brain. The primarily production impairments that accompanied damage to Broca's area were termed *Broca's aphasia*, and were quickly followed by other types of aphasia such as Wernicke's aphasia (associated with temporal lobe damage; Le et al., 2024) for language comprehension impairments.

Since then, aphasia has typically been described as an acquired language disorder where people with aphasia (PWA) present an impairment to comprehend and/or produce language following brain damage to language areas, typically resulting from a stroke. The Wernicke-Lichtheim-Geschwind model of language, or classical model, led to multiple subtypes of aphasia being proposed to account for lesions to different areas and different symptoms, such as conduction aphasia or transcortical aphasia (Le et al., 2024). Conduction aphasia was thought to result from a disconnection of the white matter tracts connecting Broca's and Wernicke's area, particularly of the arcuate fasciculus. Recent neuroimaging evidence has challenged the role of the arcuate fasciculus in conduction aphasia, as authors have described conduction aphasia cases with cortical lesions and preserved integrity of the arcuate fasciculus (Ardila, 2010). However, there is converging evidence that white matter integrity is linked with preserved language functions, with correlations between the loss of white matter integrity and language function impairments in PWA (Ivanova et al., 2016; J. Zhang et al., 2021) such as auditory comprehension (Xing et al., 2017) or speech fluency (Basilakos et al., 2014). Authors have also suggested using white matter integrity as a biomarker for post-stroke aphasia (B. Zhang et al., 2025),

However, the field of aphasia research is evolving toward a paradigm shift, as researchers have suggested that the classical model was based on outdated brain anatomy that cannot integrate many new findings. Indeed, an online survey of 159 professionals within the neurobiology of language field showed an absence of consensus regarding the anatomical location of both Broca and Wernicke's areas (Tremblay & Dick, 2016), suggesting limits to anatomy-based classification.

Additionally, recent research has shown discrepancies between the clinical diagnoses following these aphasia subtypes, lesion locations, and behavioural profiles of patients. For example, Kasselimis et al. (2017) investigated the validity and clinical utility of the conventional syndrome-based approach by classifying 65 aphasic patients with their lesion location and aphasia type following diagnosis. A large proportion (26.5%) of the aphasic profiles remained unclassified at the

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

end of the study, and most patients (63.5%) did not show the canonical lesion-to-syndrome syndrome correspondence where, for example, a lesion in Broca's area (i.e., left inferior frontal gyrus) leads to Broca's aphasia (i.e., expressive aphasia characterized by production deficits), and a lesion in Wernicke's area (i.e., left posterior superior temporal gyrus) leads to Wernicke's aphasia (i.e., receptive aphasia characterized by semantic deficits).

Given this difficulty of linking lesion location and aphasia subtype, researchers have started using data-driven methods on PWA to identify the cognitive systems supporting language processing and their neural bases. Instead of using aphasia diagnosis to infer lesion location and thus the impaired language functions, these researchers analyze the performance of PWA in multiple standardized tests and group them based on these results. The conjunction between the constructs (i.e., language dimensions) evaluated by the tests and the performance of each of the groups would indicate the language dimensions impaired in aphasia. In a review, Mirman and Thye (2018) show that most studies on the topic identified phonology, semantics, fluency and executive functioning as being central to language processing as they were the main impaired dimensions in PWA. This approach differs from the classical model in that it does not focus only on the symptoms of the patients (e.g., inability to repeat sentences, disfluency of speech or difficulties in comprehension), but rather attempts to define the behavioral profiles of PWA along the language dimensions established by cognitive models.

These data-driven methods can also be used to classify PWA according to their behavioral profiles to provide them with better tailored therapy and make more homogeneous groups for research purposes. Landrigan et al. (2021) used data-driven methods coupled with machine learning classification algorithms to classify the behavioral deficits profiles and group patients together. They observed that both their patients' behavioral and neural data clustered along semantic and phonological processing, instead of their symptom-defined aphasia subtype. Indeed, their analysis revealed three clusters of behavioral performance: individuals with milder deficits, individuals with phonological processing deficits, and individuals with semantic cognition deficits. When comparing these clusters to the individuals' aphasia subtypes, individuals with Broca's aphasia were distributed evenly within the three clusters, while individuals with conduction aphasia were evenly distributed between the first two clusters, and individuals with Wernicke's aphasia were evenly distributed within the second and third clusters (Landrigan et al., 2021).

The authors conclude that these results illustrate how useful data-driven and machine learning tools are in the study of post-stroke aphasia. Indeed, the combination of these techniques allowed them to group the PWA based on their behavioral data, which was also correlated with their neural data, thus showing a conjunction between the two types of data that has been missing from the classical diagnostic of aphasia for decades. The use of computational models, and especially of the community detection analysis, allowed the natural links between the PWA to emerge without having to fit categories previously determined by the human hand. In a way, the use of data-driven methods with computational modeling provided a better account of the different impairments found in aphasia, by stepping away from the constraints of the classical model of aphasia.

While it is obvious that behavioral and neural assessments are essential for studying language impairment, computational models can provide tools to help us make sense of complex data patterns, and possibly even simulate different potential interventions. In aphasiology, computational models have been shown to be useful tools to emulate healthy and impaired human processes (Guest et al., 2020). Damaging computational models of language processing allows one to simulate impairments present in some disorders such as aphasia. In aphasia research, computational models have been used to simulate specific deficits such as aggrammatism and anomia (Gordon & Dell, 2003) or speech production errors (Walker & Hickok, 2016), as well as broader behaviors in bilingual aphasia rehabilitation (Grasemann et al., 2021; Kiran et al., 2013). However, most of these models have been created with the intent of simulating impaired language processing, rather than starting with well-validated models of typical language processing. We propose that we can derive deeper insights by starting with validated models of healthy development and processing. Here, we explore the potential utility of a recent model of human speech processing for simulating aphasia.

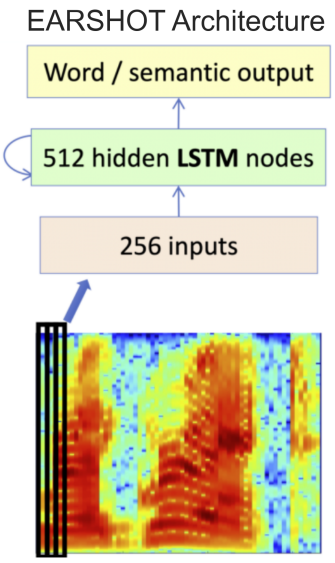

Figure 1: The EARSHOT model maps 10ms spectral slices (black rectangles on spectrogram) to semantic vectors via a single hidden layer of *long short-term* (LSTM) nodes.

EARSHOT: *Emulation of Auditory Recognition of Speech by Humans Over Time*

EARSHOT (Magnuson et al., 2020) is a two-layer recurrent network that learns to process real English speech from multiple talkers. EARSHOT was the first shallow model that has been able to process real speech, including a sizable lexicon ($\tilde{1}$000 words) produced by multiple (9) talkers. To do so, it uses long short-term memory (LSTM) nodes as a hidden layer (see Fig 1). The LSTM layer is recurrent, which means that every node of that layer is connected to every other node of the same layer. This recurrence allows the model to learn temporal contingencies in its input.

EARSHOT is intentionally situated between conventional cognitive models (e.g., McClelland & Elman, 1986; Norris & McQueen, 2008), which distill theoretical principles into fairly small, low-complexity models, but sacrifice realism to do so (neither Shortlist B nor TRACE, for example, work on real speech), and 'industrial-strength' deep learning and/or transformer models, which process real speech and are used for real-world, but are too complex to give much theoretical insight to cognitive scientists. The goal with EARSHOT was to develop the simplest model possible that could operate on real speech inputs. Minimizing complexity is intended to keep the model as understandable and analytically tractable as possible.

EARSHOT has $\tilde{1}$.7 million parameters. Despite being $\tilde{2}$5% smaller than TRACE and many orders of magnitude smaller than deep learning models, it is still quite complex. Despite this complexity, Magnuson et al., 2020 found that EARSHOT achieved high accuracy on its training lexicon, moderate generalization to untrained tokens and talkers, and was able to simulate the fine-grained timecourse of human phonological competition. Despite not being trained on phonetic targets, an internal phonetic code emerges, such that hidden-layer responses closely resemble human cortical responses to speech (cf. Mesgarani et al., 2014). Using cortical speech tracking methods, Brodbeck et al., 2025 found that over-time hidden layer activity in EARSHOT accounts for significant variance in human MEG responses to speech beyond that accounted for by conventional auditory and information-theoretic predictors.

EARSHOT's ability to emulate fine-grained aspects of human speech processing behavior and to relate to neural measures of speech processing, as well as its relative simplicity, make it a good candidate for exploring receptive deficits following damage. We report first steps towards simulating aspects of aphasia with EARSHOT by damaging increasing proportions of randomly selected weights in different model layers. We aimed to observe if systematic patterns of phonetic vs. semantic deficits (Landrigan et al., 2021) would emerge following damage to different model components. Systematic patterns could provide a foundation for future investigations comparing damaged models to data from PWA.

Method

Model Architecture

EARSHOT maps 256-dimensional spectral slices to 300-dimensional semantic vectors (here, Skip-gram Mikolov et al., 2013) via 512 hidden nodes (*long short-term memory* or LSTM nodes; Hochreiter & Schmidhuber, 1997). The LSTMs develop sensitivity to short and long scales, which is allows the network to develop sensitivity to local and distant contingencies in speech and accurately recognize words.

Words and Training

We trained EARSHOT on 997 high-frequency English words spoken by one talker, using backpropagation of error for 8000 epochs on a model with randomly-initialized weights. Each epoch was one pass through the lexicon, with words in random order and presented with no intervening silence. We simulated a naming/identification task where our model had to activate the defined semantic vector (SkipGram) for the current target at each time frame. At each step, we computed semantic similarity between the model's output layer vector and every word in the lexicon, operationalizing EARSHOT's naming/identification response as the word with peak cosine similarity. A limitation of this approach that we discuss later is that a model response is *always* defined, even when the cosine similarity is very low.

We tested undamaged and damaged model instances by presenting all the words and assessing accuracy and error patterns by measuring semantic and phonemic similarity of responses and targets. For a specific word, accuracy was equal to 1 if the semantic vector output by the model was closest to the semantic vector of the target word. Otherwise, accuracy was equal to 0. The undamaged model achieved 95% accuracy. Semantic similarity was computed as the cosine similarity of the words in the word2vec embedding space. Phonemic/phonological similarity was based on phonemic transcriptions of each word. After identifying the 'named' word (again, defined as the word with highest semantic cosine similarity to the output), we compared the phonemic transcriptions of that word and the target. We defined phonetic distance as 1 – Levenshtein edit distance normalized by the length of the longest word. The initial edit distance could range from 0 (exact correspondence) to 1 (a word mismatching at every phoneme). By subtracting this value from 1, higher values indicate greater similarity. Thus, for both semantic and phonetic distance, a higher value indicated larger similarity of the two items in the word pair.

Simulating Damage

Before simulating damage, we duplicated the trained model 10 times to have 10 'simulated patients'. This ensured that our observations would not result from the specific weights deleted through damage, but would reflect the general impact of a particular degree of damage at the chosen location.

To simulate damage, we randomly removed increasing proportions of connections between input and hidden, hidden and hidden, or hidden to output layers. At a given location

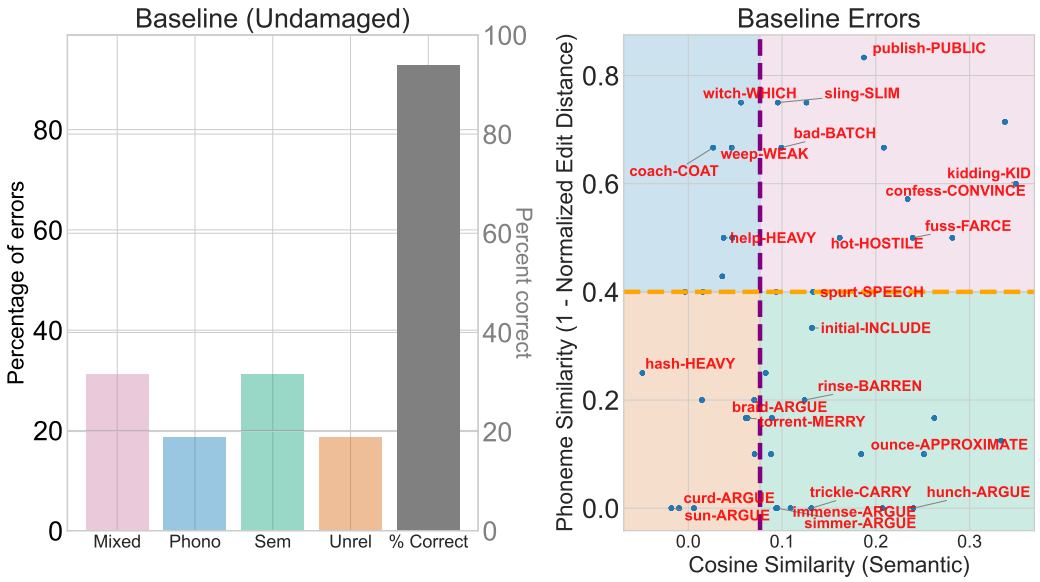

Figure 2: Types of errors and error pairs for the trained model (target-MODEL'S CHOICE). Error types are color-coded similarly across the two panels: mixed errors are pink (top right quadrant of right panel), phonological errors are blue (top left quadrant of right panel), semantic errors are green (bottom right quadrant of right panel) and unrelated errors are orange (bottom left quadrant of right panel).

(input-hidden, hidden-hidden, hidden-output), we randomly set a proportion of connection weights to 0. This allowed us to model white matter disconnection with an impaired transfer of information. As mentioned earlier, white matter integrity has been shown to support language functions, with temporal lobe tracts being associated to language comprehension (Ivanova et al., 2016) and particularly with auditory comprehension (Xing et al., 2017). The exact percentages of damage varied by layer: while initially all layers were damaged in 10% increments (e.g., 10%, 20%, 30%...), the impact on layers was quite different. So instead we sought series of damage values for each damage location that would lead to a similar range of post-damage accuracy levels.

Error Types

We classified errors as mixed (both semantic and phonological similarities high), phonological (high phonological, low semantic), semantic (low phonological, high semantic), or unrelated (both similarities low). Examples of errors include coach-COAT (phonological), downstairs-CEILING (semantic), distant-DISTANCE (mixed), and simmer-ARGUE (unrelated). All errors of the baseline undamaged model are shown in Figure 2, with a random subset of 60% error pairs labeled for readability.

The threshold for classifying an error as phonological followed a common definition used in clinical studies where an error is phonological if two words share at least two phonemes and differ in one or two phonemes. In the absence of an objective measure of semantic error in the clinical literature, we set one by selecting the mean cosine similarity in the word2vec embeddings as the semantic threshold. Word pairs above this threshold were considered to be highly semantically similar, while words below the threshold were less semantically similar. Note also that human subjective ratings of semantic similarity do not apply directly to this small lexicon of 997 words. The mean semantic threshold is relative to the limited relations among words in this small lexicon.

Hypotheses

We expected to see the accuracy of the model decrease with increasing damage. Regarding the error types and location of damage, we expected to see an intuitive pattern: (1) more phonological errors following damage to the input to hidden connections as these connections are closer to the auditory input, (2) a mixed pattern of errors following damage to the recurrent hidden connections, and (3) more semantic errors following damage to the hidden to output connections as these connections are closer to the semantic output.

Results

Training results

After training for 8000 epochs and before damage, the model reached 95% accuracy. Fig. 2 shows that out of the few errors made by the model at this stage, most of them are mixed

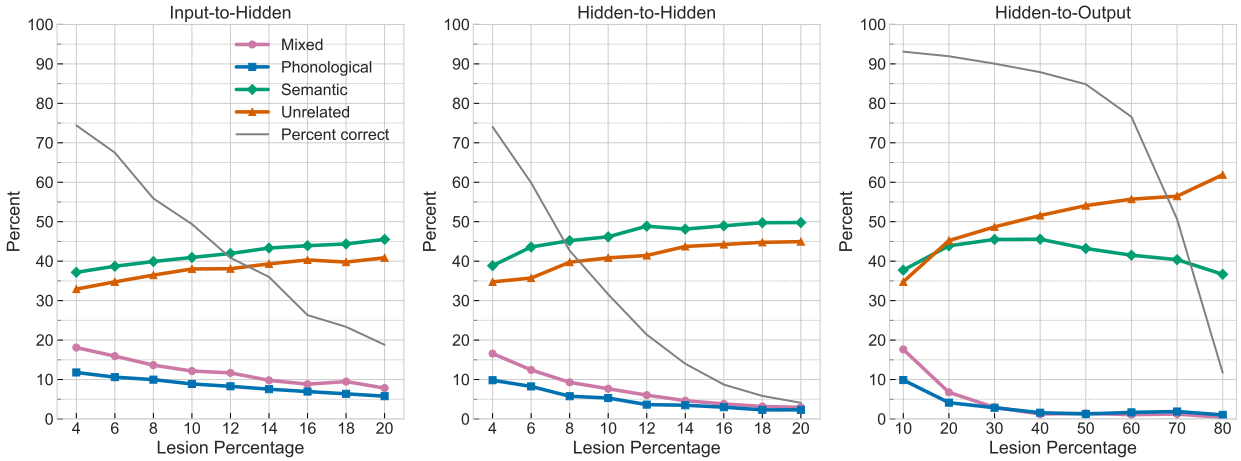

Figure 3: Types of errors for each level and layer of damage. Color-coding is consistent across panels.

Lesioned layer	Phonological error	Semantic error	Mixed error	Unrelated error
Baseline	18.8%	31.2%	31.2%	18.8%
Input-to-Hidden	8.4%	41.9%	11.7%	37.9%
Hidden-to-Hidden	4.9%	46.6%	7.4%	41.2%
Hidden-to-Output	3.0%	41.8%	4.1%	51.1%

Table 1: Proportion of each type of error at each layer averaged over all damage levels.

or semantic errors, with fewer unrelated and phonological errors. As noted above, some word pairs might not seem very semantically similar for human readers, but the model only 'knows' 997 words. As such, two words that are rather different within a lexicon of one hundred thousand words might be very similar within a lexicon of a thousand words.

Damage results

Fig. 3 shows the proportion of error types per damage level for each of the three damaged layers. For clarity purposes, we report the proportion of errors of each type for the baseline undamaged model, as well as the three damaged layers in Table 1. Due to space limitations, the proportions of error are averaged over all levels of damage.

Accuracy Grey lines show the accuracy of models across damage levels for each of the three lesion locations. Accuracy drops sharply in the lower damage levels (4-20%) for the input-hidden and hidden-hidden layers, reaching 20% accuracy for input-hidden at 20% damage, and between 0 to 5% accuracy for hidden-hidden at 20% damage. In contrast, accuracy is around 80% for hidden-output until 60% damage, where accuracy suddenly decreases to reach 15% at 80% damage. This suggests that the hidden-output layer is more resistant to damage, while input-hidden and hidden-hidden are more sensitive to damage.

Note that while the accuracy patterns for input-hidden and hidden-hidden seem similar and distinct from the pattern for hidden-output, this is at least partly due to the levels of damage for each layer. As we noted above, we shifted to the 0-20% range for input-hidden and hidden-hidden because accuracy approaches floor levels by 20% damage. A damage range for hidden-output of perhaps 60-90% might yield an accuracy 'curve' more like those for input-hidden and hidden-hidden. We will explore this possibility in our next steps. In the meantime, we must be cautious in interpreting apparent qualitative differences that may be parameter-driven (e.g., by selected damage percentage levels).

Error types Across all damage-locations, semantic and unrelated errors were the most prevalent, while phonological and mixed errors were less frequent. Input-hidden and hidden-hidden shared the same profile, with semantic errors as the most frequent, followed by unrelated errors, mixed and phonological errors. Hidden-output shared a similar distribution of error types, with the notable exception of unrelated errors being the most prevalent, followed by semantic errors, mixed and phonological errors.

In line with our hypothesis, the highest proportion of phonological errors was found following input-hidden damage, across all damage-levels (see Table 1 for the averaged proportion of errors across all damage levels). On the other hand, the highest proportion of semantic errors was observed in hidden-hidden for higher levels of damage, and hidden-

output for lower levels of damage. This confirms intuitive expectations (but not forgone conclusions) for the model architecture (that disrupting the input level would drive more form errors than semantic errors, and damaging higher levels would drive more semantic than form errors), and provides a foundation for more detailed analyses we will pursue in our continuing work.

Discussion

Data-driven approaches have previously identified phonology and semantics to be two main language dimensions that can be found to be impaired in PWA (Landrigan et al., 2021; Mirman & Thye, 2018). Additionally, white matter tract impairments have been linked to language deficits in PWA (Ivanova et al., 2016; J. Zhang et al., 2021). Our study aimed at simulating semantic and phonological impairments resulting from white matter disconnection in a speech processing model previously shown to emulate key aspects of human speech processing (Magnuson et al., 2020). As expected, the patterns of impairment varied as a function of the location of damage.

At all damage-location combinations, unrelated errors dominated, with few mixed errors. Relative proportions of semantic and phonological errors varied by location. 'Early' damage (input-hidden) yielded the most phonological errors, which decreased at higher levels. 'Late' damage (hidden-output) yielded the most semantic errors. This reflects EARSHOT's internal division-of-labor; as Magnuson et al., 2020 document, it primarily encodes patterns phonologically in its hidden layer (especially in the early timecourse of processing a word), and our results suggest the hidden-output transformation is primarily responsible for decoding semantics.

Limitations and future directions

Here, we began an effort to simulate phonological and semantic impairments in post-stroke aphasia using a model developed to simulate healthy spoken word recognition. We characterized phonological errors by setting a phonological threshold based on clinical studies. However, in the absence of a clear consensus on the nature and limitations of a semantic error, we set an objective measure of semantic error by selecting the 90th percentile of cosine similarity in the word2vec embeddings. As such, the thresholds used in the study and the resulting proportions of errors are largely dependent on the variation inherent to the word2vec embeddings for the specific words used. Using other types of semantic representations on a different lexicon might lead to somewhat different results. We believe that this does not strictly constitute a limitation but rather opens up possibilities to explore the distribution of language impairments in the EARSHOT model depending on the lexicon and semantics used.

A limitation of this model for studying aphasia is its architecture. As EARSHOT is a receptive model of human speech processing, it does not generate phonological output. We instead treated its semantic output as the basis for a response: we operationalized the response as the closest word (in semantic space) to the model's semantic output. Therefore, the answers that the model provides are limited: unlike a human patient, our model cannot produce neologisms or answers that are absent from the trained lexicon, and cannot respond "I don't know" (though we could operationalize the latter when the cosine similarity between the model's output and the closest word is below a threshold). One next step will be to add a hidden-layer decoder to provide phonological readout, which could substantially expand the scope of our approach.

Additionally, we only trained EARSHOT on single words. In unpublished work from our group, we are training EARSHOT on continuous speech inputs, which could substantially change some outcomes, even when the model is tested on isolated words following training. Future work damaging this expanded version of EARSHOT could offer additional insights, including possibly discourse-level impairments.

We are also expanding EARSHOT as a model of bilingual development and processing. One aspect of our ongoing work is using the bilingual model as a platform for studying aphasia in bilingual PWA, including work with languages that are rarely studied in the aphasiology literature (e.g., Basque).

Conclusion

EARSHOT's division-of-labor intuitively suggests lower-level damage impairs phonology most, and higher-level damage most affects semantics.

A few limitations must be addressed to enhance clinical relevance. First, EARSHOT lacks phonological output, and so cannot generate nonwords/neologisms. While operationalizing EARSHOT's output as its naming or identification 'response' is a reasonable first step, in our ongoing work, we are decoding hidden layer activity as a phonological readout to expand our ability to probe effects of damage. Second, this version of EARSHOT is only trained on single words. As a next step, we are planning to train EARSHOT on continuous speech before testing the model on single words, in order to simulate PWA more realistically. Finally, this work has focused on monolingual aphasia. Our next aim is to expand EARSHOT as a bilingual model to investigate aphasia in bilingual individuals.

Despite these limitations, EARSHOT offers a promising platform for simulating aphasia and testing intervention strategies, such as retraining after damage. With further development, it could become a valuable tool for enhancing understanding of aphasia.

Acknowledgments

This research was supported by the Basque Government through the BERC 2022-2025 program and Funded by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020001010-S-21-1, grant ref. "PRE2021-097232; funded by MCIN/AEI/10.13039/501100011033 and FSE+", and the grant ref. PID2024-159519OB-I00 funded by AEI.

References

Ardila, A. (2010). A review of conduction aphasia. *Current neurology and neuroscience reports*, *10*(6), 499–503.

Basilakos, A., Fillmore, P. T., Rorden, C., Guo, D., Bonilha, L., & Fridriksson, J. (2014). Regional white matter damage predicts speech fluency in chronic post-stroke aphasia. *Frontiers in human neuroscience*, *8*, 845.

Brodbeck, C., Hannagan, T., & Magnuson, J. S. (2025). Recurrent neural networks as neuro-computational models of human speech recognition. *PLOS Computational Biology*, *21*(7), e1013244. https://doi.org/10.1371/journal.pcbi.1013244

Gordon, J. K., & Dell, G. S. (2003). Learning to divide the labor: An account of deficits in light and heavy verb production. *Cognitive Science*, *27*(1), 1–40.

Grasemann, U., Peñaloza, C., Dekhtyar, M., Miikkulainen, R., & Kiran, S. (2021). Predicting language treatment response in bilingual aphasia using neural network-based patient models. *Scientific Reports*, *11*(1), 10497.

Guest, O., Caso, A., & Cooper, R. P. (2020). On simulating neural damage in connectionist networks. *Computational brain & behavior*, *3*, 289–321.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Ivanova, M. V., Isaev, D. Y., Dragoy, O. V., Akinina, Y. S., Petrushevskiy, A. G., Fedina, O. N., Shklovsky, V. M., & Dronkers, N. F. (2016). Diffusion-tensor imaging of major white matter tracts and their role in language processing in aphasia. *Cortex*, *85*, 165–181.

Kasselimis, D. S., Simos, P. G., Peppas, C., Evdokimidis, I., & Potagas, C. (2017). The unbridged gap between clinical diagnosis and contemporary research on aphasia: A short discussion on the validity and clinical utility of taxonomic categories. *Brain and language*, *164*, 63–67.

Kiran, S., Grasemann, U., Sandberg, C., & Miikkulainen, R. (2013). A computational account of bilingual aphasia rehabilitation. *Bilingualism: Language and Cognition*, *16*(2), 325–342.

Landrigan, J.-F., Zhang, F., & Mirman, D. (2021). A data-driven approach to post-stroke aphasia classification and lesion-based prediction. *Brain*, *144*(5), 1372–1383.

Le, H., Lui, F., & Lui, M. Y. (2024). *Aphasia* [Updated 2024 Oct 29]. https://www.ncbi.nlm.nih.gov/books/NBK559315/

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabi, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., et al. (2020). Earshot: A minimal neural network model of incremental human speech recognition. *Cognitive science*, *44*(4), e12823.

Maudsley, H. (1868). Concerning aphasia. *The Lancet*, *92*(2361), 690–692.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive psychology*, *18*(1), 1–86.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, *343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mirman, D., & Thye, M. (2018). Uncovering the neuroanatomy of core language systems using lesion-symptom mapping. *Current Directions in Psychological Science*, *27*(6), 455–461.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Tremblay, P., & Dick, A. S. (2016). Broca and wernicke are dead, or moving past the classic model of language neurobiology. *Brain and language*, *162*, 60–71.

Walker, G. M., & Hickok, G. (2016). Bridging computational approaches to speech production: The semantic–lexical–auditory–motor model (slam). *Psychonomic Bulletin Review*, *23*, 339–352.

Xing, S., Lacey, E. H., Skipper-Kallal, L. M., Zeng, J., & Turkeltaub, P. E. (2017). White matter correlates of auditory comprehension outcomes in chronic post-stroke aphasia. *Frontiers in neurology*, *8*, 54.

Zhang, B., Schnakers, C., Wang, K. X.-L., Wang, J., Lee, S., Millan, H., Howard, M., Rosario, E., & Zheng, Z. S. (2025). Infratentorial white matter integrity as a potential biomarker for post-stroke aphasia. *Brain Communications*, *7*(3), fcaf174.

Zhang, J., Zhong, S., Zhou, L., Yu, Y., Tan, X., Wu, M., Sun, P., Zhang, W., Li, J., Cheng, R., et al. (2021). Correlations between dual-pathway white matter alterations and language impairment in patients with aphasia: A systematic review and meta-analysis. *Neuropsychology Review*, *31*(3), 402–418.