Artificial Neural Networks Generate Human-like Continuous Speech Perception

Gasser Elbanna
Speech and Hearing Bioscience and Technology
Harvard University
Boston, MA 02115
gelbanna@mit.edu

Josh H. McDermott
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139
jhm@mit.edu

Abstract

Humans have a remarkable ability to convert acoustic signals into linguistic representations. To advance toward the goal of building biologically plausible models that replicate this process, we developed an artificial neural network trained to generate sequences of American English phonemes from audio processed by a simulated cochlea. We trained the model with phoneme transcriptions inferred from text annotations of speech corpora. To compare the model to humans, we ran a behavioral experiment in which humans transcribed non-words, and evaluated the model on the same stimuli. While humans slightly outperformed the model, the model exhibited human-like patterns of phoneme confusions for consonants (r=0.91) and vowels (r=0.87). Additionally, the recognizability of individual phonemes was highly correlated (r=0.93) between humans and the model. These results suggest that human-like speech perception emerges from optimizing for phoneme recognition from cochlear representations.

1 Introduction

The core computational challenge of speech perception is the absence of consistent one-to-one mappings between the acoustic signal and the sub-lexical units (such as phonemes) that make up speech (1; 2). Despite substantial acoustic variability across speakers, speaking rates, and environmental conditions, humans exhibit remarkable accuracy in perceiving speech sounds. Various theoretical ideas have been proposed to account for our abilities. These include acoustic invariances (3), categorical perception (4), and the motor theory of speech perception (5), among others. While these theories have guided research, they are not sufficiently specified to fully explain how listeners achieve such robust speech recognition.

In parallel, computational models of speech perception have been developed to explain aspects of speech perception. Traditional models often operated on abstract representations or handcrafted features, and were generally unable to account for human recognition of real-world speech (6; 7; 8; 9). In recent years, artificial neural networks (ANNs) have emerged as a powerful alternative, learning directly from raw acoustic input. These models are often able to approach human levels of performance, unlike traditional models, and have helped explain human-like perceptual abilities in other domains of audition (10; 11). However, most available speech ANN models lack biological plausibility and often do not exhibit the same patterns of performance as human listeners (12; 13; 14).

Some work has attempted to bridge this gap by developing biologically plausible speech models, for instance using a simulated cochlear representation as a front-end for a neural network (15; 16). These models exhibit human-like patterns of speech intelligibility in some conditions but do not perform continuous speech recognition because they were trained to recognize a set of words substantially smaller than the full vocabulary of English, and thus cannot account for many aspects of human speech perception.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

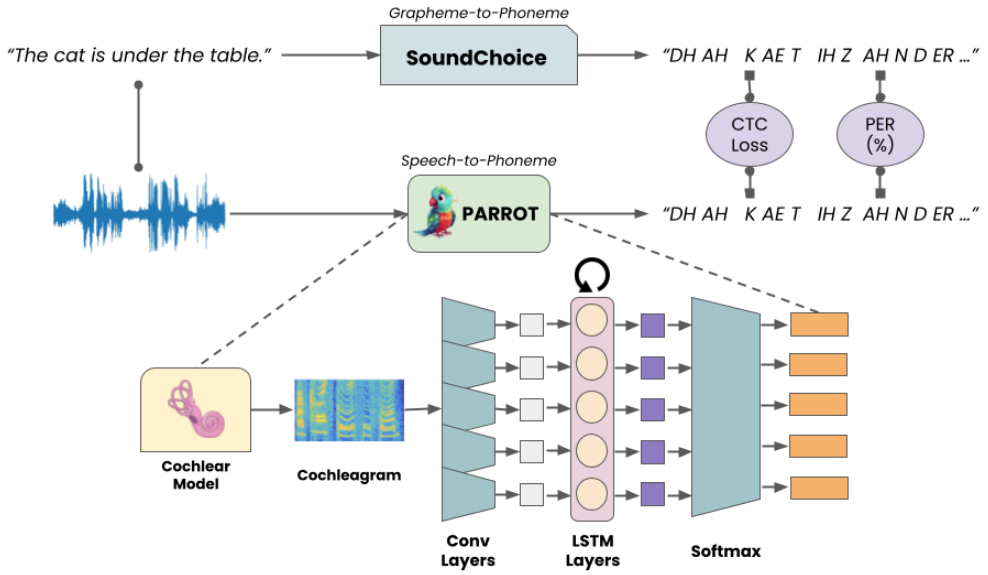

Figure 1: PARROT architecture and training pipeline.

In this work, we present a novel speech model called PARROT that is trained to generate sequences of English phonemes from simulated cochlear input. We directly compare the model's performance to that of human listeners using a non-word transcription task. This task allows us to analyze and compare patterns of successes and failures of phoneme recognition, shedding light on the similarities and differences between human and model performance and providing a first step towards more realistic models of continuous speech perception.

2 Methods

2.1 Model Architecture and Task Objective

The first stage of the architecture is a cochlear model adapted from prior work (17) that processes speech waveforms sampled at 16 kHz. The simulated cochlea applies gamma-tone filters with center frequencies between 40 Hz and 20 kHz, with tuning and spacing intended to replicate that of the human ear. The output from the filter banks is half-wave rectified and low-pass filtered with a cutoff frequency of 4 kHz, simulating the upper limit of phase-locking (18). The filtered output is then downsampled to a sampling rate of 8 kHz and raised to the 0.3 power to replicate the nonlinear amplification of the ear (19).

Cochlear representations are fed into six 2-dimensional convolutional layers with 512 channels, which further downsample the signal from 8 kHz to 50 Hz, encoding 20 ms of speech per frame. After each convolutional layer, batch normalization is applied, followed by a ReLU non-linear activation function. The resulting latent representations are passed through six bi-directional Long Short-Term Memory (LSTM) layers with hidden size of 512, which capture temporal dependencies across frames. Finally, the LSTM hidden states are projected into a 40-class phoneme space (consisting of 39 phonemes and a blank class) using a linear fully connected layer. The logits are converted into a probability distribution via a softmax function.

The model was trained to map the probability distribution of phoneme classes to tokens using a Connectionist Temporal Classification (CTC) loss (20) (see Section A.1). Phoneme Error Rate (PER) was calculated after aligning predicted and ground truth phonemes using the Levenstein distance algorithm (21). The percentage of phoneme errors were computed as in Eq. 1.

$$PER(\%) = \frac{No.of Insertions + No.of Deletions + No.of Substitutions}{Total No.of Ground Truth Phonemes} \quad (1)$$

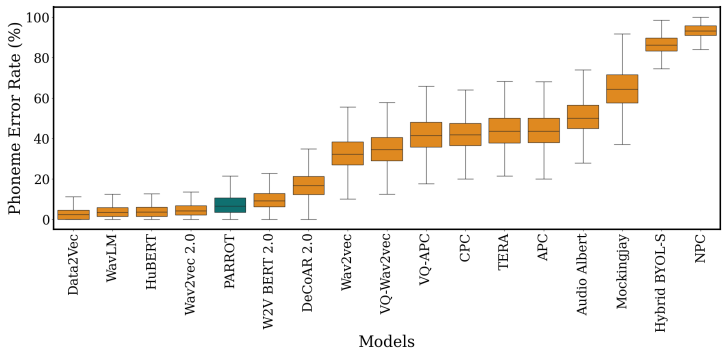

Figure 2: Phoneme Error Rate (%) of speech models on Phoneme Recognition task from SUPERB Benchmark. PARROT result is shown in teal. The lower the value, the better the model.

2.2 Experimental Setup and Evaluation

Because there is little available phoneme-labeled speech data, we used a pseudo-supervised training approach, employing a Grapheme-to-Phoneme model called SoundChoice (22) to transcribe phonemes from text annotations of large-scale speech corpora. We transcribed around 6 Million utterances from open-source corpora including GigaSpeech (23), Librispeech (24), VCTK (25), LJSpeech (26), Speech commands (27), FSDD (28), and TIMIT (29), yielding around 10,000 hours of training data. We used the Librispeech dev set for validation and the Librispeech test set for testing which were not part of the training data. Also, we benchmarked the model on SUPERB phoneme recognition task (30).

2.3 Non-word Recognition Experiment

To compare model and human speech perception, we designed a non-word recognition task where participants transcribed synthesized non-words. The non-words were generated with Wuggy (31), a pseudo-word generator that produces non-word variants from real words while abiding by English phonotactics rules. We generated around 15,000 non-words and selected a subset of 5,000 that maximized representation of the least common phonemes. Non-words were then synthesized using the MeloTTS text-to-speech model (32).

We ran an online behavioral experiment on Prolific. Each of the 100 participant transcribed 200 non-words randomly chosen from the pool of 5000 non-words. Participants had to first pass a headphone check (33). We converted each text string response into phonemes using the same G2P model used to generate training data labels. The same non-words were presented to the model.

3 Results

3.1 Model Evaluation

The model achieved a median PER of 6.8% on on Librispeech test set, and performed competitively on a standard phoneme recognition benchmark (30) (Figure 2).

3.2 Human-Model Comparison

We computed the PER for each non-word in the experiment. The model was slightly worse than humans (median PER of 33% vs 29%; Figure 3.a). However, individual phonemes varied in the accuracy with which they were recognized, and the phoneme-wise accuracy was highly correlated between humans and the model ($r=0.92$; $p<0.01$; Figure 3.b-c). The correlation remained high when calculated separately for consonants ($r=0.97$, $p<0.01$) and vowels ($r=0.86$, $p<0.01$).

To assess whether the model replicated the pattern of confusions exhibited by humans, we compared human and model confusion matrices, separately for consonants and vowels (Figures 3.d-i). The off-diagonal matrix entries were also strongly correlated between humans and PARROT for both consonants ($r=0.91$, p value<0.01) and vowels ($r=0.87$, p value<0.01). The same phoneme was

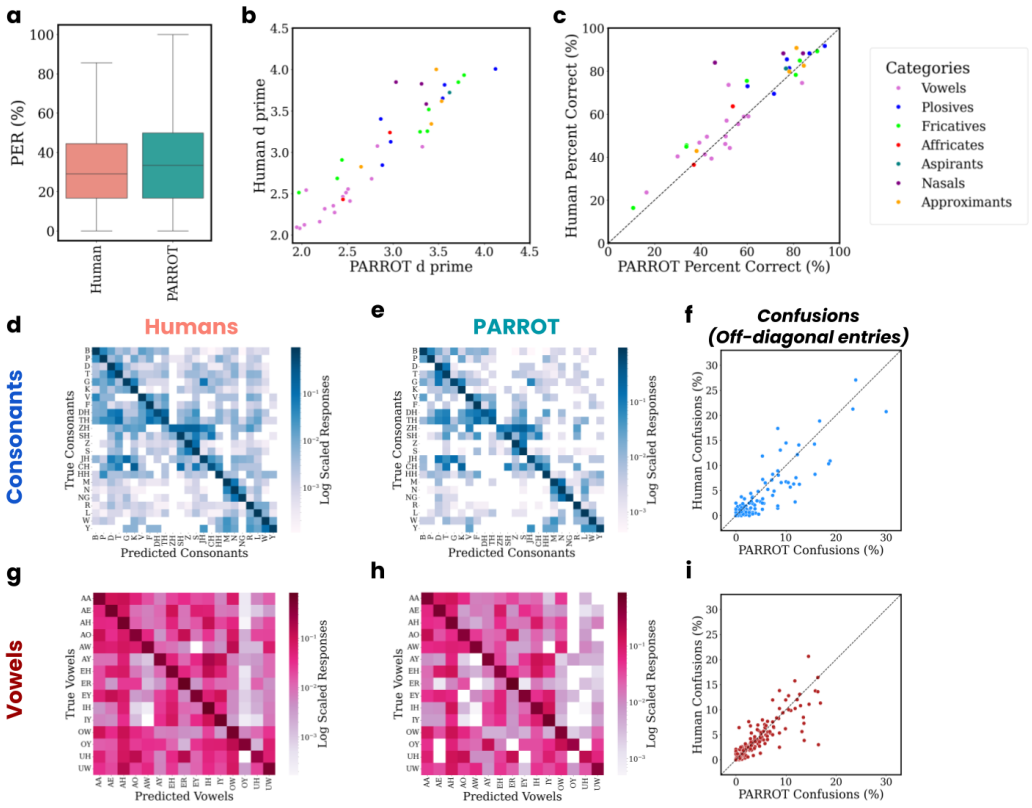

Figure 3: Human-Model comparison on non-word recognition task. (a) PER distribution for Humans and PARROT across non-word stimuli. (b) Human vs. model recognition accuracy for individual phonemes, expressed as d prime. Colors indicate manner of articulation. (c) Same as b, but plotting percent correct. (d) and (g) Phoneme confusions in humans for consonants and vowels, respectively. (e) and (h) Phoneme confusions in PARROT for consonants and vowels, respectively. (f) and (i) Off-diagonal correlation between humans and PARROT for consonants and vowels, respectively.

confused the most for both humans and PARROT (/ZH/, as in "measure"). Similarly, the same phoneme was confused the least for both (/K/ as in "cat").

4 Discussion

Compared to existing automatic speech recognition systems, the model demonstrated competitive performance on unseen data and various transcription methods (see Figure 2). Humans performed slightly better than the model on the non-word recognition task. However, at the phoneme level, the model exhibited a similar pattern of phoneme confusions as humans, both for consonants (r=0.91) and for vowels (r=0.87). The recognizability of individual phonemes was also highly correlated between humans and the model (r=0.93), highlighting the model's alignment with human perception.

5 Conclusion

We developed a novel deep learning model that was trained to recognize phonemes using data transcribed by an existing Grapheme-to-Phoneme model. The model performed competitively on the task and showed human-like patterns of phoneme confusions. The findings collectively suggest that aspects of human-like speech perception emerges by optimizing for phoneme recognition from cochlear representations. In future work, we plan to identify the key model components driving this alignment. The results provide a first step towards building biologically-plausible models that replicate and explain human speech representations.

References

[1] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *The Journal of the acoustical society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[2] J. S. Perkell and D. H. Klatt, *Invariance and variability in speech processes*. Psychology Press, 2014.

[3] S. E. Blumstein and K. N. Stevens, "Phonetic features and acoustic invariance in speech." *Cognition*, 1981.

[4] A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith, "The discrimination of speech sounds within and across phoneme boundaries." *Journal of experimental psychology*, vol. 54, no. 5, p. 358, 1957.

[5] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[6] B. P. Lowerre and B. R. Reddy, "Harpy, a connected speech recognition system," *The Journal of the Acoustical Society of America*, vol. 59, no. S1, pp. S97–S97, 1976.

[7] J. Wolf and W. Woods, "The hwim speech understanding system," in *ICASSP'77. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1977, pp. 784–787.

[8] V. Lesser, R. Fennell, L. Erman, and D. Reddy, "Organization of the hearsay ii speech understanding system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 11–24, 1975.

[9] J. L. McClelland and J. L. Elman, "The trace model of speech perception," *Cognitive psychology*, vol. 18, no. 1, pp. 1–86, 1986.

[10] A. Francl and J. H. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments," *Nature Human Behaviour*, vol. 6, pp. 111–133, 2022.

[11] M. R. Saddler, R. Gonzalez, and J. H. McDermott, "Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception," *Nature Communications*, vol. 12, p. 7278, 2021.

[12] L. Weerts, S. Rosen, C. Clopath, and D. F. Goodman, "The psychometrics of automatic speech recognition," *bioRxiv preprint bioRxiv:2021.04.19.440438*, 2022.

[13] F. Adolfi, J. S. Bowers, and D. Poeppel, "Successes and critical failures of neural networks in capturing human-like speech recognition," *Neural Networks*, vol. 162, pp. 199–211, 2023.

[14] S.-E. Kim, B. R. Chernyak, O. Seleznova, J. Keshet, M. Goldrick, and A. R. Bradlow, "Automatic recognition of second language speech-in-noise," *JASA Express Letters*, vol. 4, no. 2, 2024.

[15] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott, "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," *Neuron*, vol. 98, no. 3, pp. 630–644, 2018.

[16] M. R. Saddler and J. H. McDermott, "Models optimized for real-world tasks reveal the task-dependent necessity of precise temporal coding in hearing," *bioRxiv preprint bioRxiv:2024.04.21.590435v2*, 2024.

[17] J. Feather, G. Leclerc, A. Mądry, and J. H. McDermott, "Model metamers reveal divergent invariances between biological and artificial neural networks," *Nature Neuroscience*, vol. 26, no. 11, pp. 2017–2034, 2023.

[18] J. Rose, J. Brugge, D. Anderson, and J. Hind, "Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey," *Journal of Neurophysiology*, vol. 30, pp. 769–793, 1967.

[19] M. Ruggero, "Responses to sound of the basilar membrane of the mammalian cochlea," *Current Opinion in Neurobiology*, vol. 2, pp. 449–456, 1992.

[20] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[21] F. P. Miller, A. F. Vandome, and J. McBrewster, "Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? levenshtein distance, spell checker, hamming distance," 2009.

[22] A. Ploujnikov and M. Ravanelli, "Soundchoice: Grapheme-to-phoneme models with semantic disambiguation," *arXiv preprint arXiv:2207.13703*, 2022.

[23] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[25] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2019.

[26] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[27] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[28] Z. Jackson, C. Souza, J. Flaks, Y. Pan, H. Nicolas, and A. Thite, "Jakobovski/free-spoken-digit-dataset: v1.0.8," Aug. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1342401

[29] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[30] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[31] E. Keuleers and M. Brysbaert, "Wuggy: A multilingual pseudoword generator," *Behavior research methods*, vol. 42, pp. 627–633, 2010.

[32] W. Zhao, X. Yu, and Z. Qin, "Melotts: High-quality multi-lingual multi-accent text-to-speech," 2023. [Online]. Available: https://github.com/myshell-ai/MeloTTS

[33] K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, pp. 2064–2072, 2017.

[34] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*. Ieee, 2018, pp. 1–2.

A Appendix / supplemental material

A.1 Training Details

Training examples varied in duration, so we padded each example in a batch to the length of the longest duration in the batch. We used batches of size 4 with 2 gradient accumulation steps. We trained the model on 8 A100 GPUs yielding an effective batch size of 4*2*8 = 64. The model was trained on a total of 400,000 gradient steps per GPU translating into a total of 5 epochs. We use Adam optimizer (34) with weight decay of 0.01 and a warming up the learning rate (LR) linearly for the first 10,000 steps reaching a peak of 0.001. LR is fixed to peak value until training reaches 200,000 steps then LR decreases using cosine annealing for the second half of training reaching min LR of 0.00001.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims about human-model alignment in the abstract are clearly and accurately described in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

Justification: Limitations are not included since this is still on-going work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theoretical results are provided in this paper, only experimental.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Details about training the model and its architecture are mentioned in the main body as well as Appendix.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: No data or code is shared since this is an ongoing-work.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: These details are mentioned in Methods section and Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report p Value for all correlational analyses mentioned in the paper. Also, we show box plots for performance demonstrating the distribution of PER.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We mention in the APpendix that the model was trained on 8 A100 GPUs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the Code of Ethics provided by NeurIPS.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work doesn't exhibit any potential risk or societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our proposed model doesn't show a high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite open-source systems used in our experiments.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets are released in this submission.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

Justification: Details for online behavioral experiment are mentioned in the Methods section. For page limit, we don't show screenshots of the experiment instructions. We recruited participants from prolific and all were paid the suggested amount by the platform for their contribution.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The online study was run under an approved IRB protocol.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.