Exploring Pretrained Transformer Phonotactics Through IPA based G2P Finetuning

Mario Peng Lee
Stanford University Online
University of California Los Angeles
mariopeng@ucla.edu

Abstract

Phonotactics, despite being a core component in the field of Linguistics, remains relatively unexplored as compared to semantics and syntax in natural language processing. Traditional G2P (Grapheme to Phoneme) models are constructed by linguist defined phonotactics, which in turn provides little unprecedented insights to the field. Meanwhile, deep learning requires real world natural language examples and no set of defined rules, allowing undefined rule learning, which is more representative modeling of human knowledge acquisition and could in term provide new discoveries to the field of phonotactics.

Current SOTA (state of the art) NMT (Neural Machine Translation) models allow G2P to be an easy task. By fine tuning pretrained transformer language models on G2P tasks, we achieved considerably good results in just a few training epochs. Other approaches of model architectures and data augmentation are presented and explored. With these experiments, we first showcase G2P fine tuning on transformers as an easy approach to creating phonotactic models, then we briefly analyze machine phonotactic knowledge acquisition. Ultimately we provide useful insights and future research directions for human phonotactic modeling through deep learning, along with two Seq2Seq G2P datasets to encourage future G2P deep learning research.

1 Introduction

NLP (Natural Language Processing) technology went through a rapid development in the past decades, going from simple bags of words to transformer LMs (language models). (Harris, 1954; Vaswani et al., 2017) A LM is essentially a probability distribution over words or sentences, this technology can be applied to a wide variety of different tasks, from text generation (Brown et al., 2020) and machine translation (Raffel et al., 2020) to seemingly unrelated tasks such as image recognition (He et al., 2015) and text-to-image generation (Ramesh et al., 2021). This raises the question of what is being encoded into these language models, as well as what that information can teach us about the tasks we assign them to complete. (Sajjad et al., 2022) Founded on language, LMs fundamentally contribute to NLP and linguistic research over all its domains.

In particular, phonotactics studies what phonetic rules are acquired by speakers and how this knowledge is formed. (Chomsky and Halle, 1968) This field of study remains relatively outdated as compared to semantics and syntax in NLP, despite being a core component in the field of Linguistics. The long history of computational linguistics has already studied phonotactic patterns intensively. (Bisani and Ney, 2008) With both the emergence of probabilistic phonotactic models – such as n-gram based models – and later on RNN (Recurrent Neural Network) phonetic models, linguists uncovered unprecedented findings about phonetics and phonotactic knowledge. (Mayer and Nelson, 2020) This advancement followed the path of language modeling development, with probabilistic n-grams that transitioned to RNNs. (Sherstinsky, 2020)

The last major advancement in NLP has been the transformer architecture, after the introduction of attention heads, (Vaswani et al., 2017) a powerful technology that propelled language model performance in both syntactic and semantic knowledge. (Chowdhery et al., 2022) Even though acoustic modeling and TTS (text to speech) technology has widely caught up to this technology, there is a gap to fill for phonotactics, and exploration needs to be done. (Rao et al., 2015)

G2P (Grapheme to Phoneme) is a process in which a word's grapheme, which is its textual rep-

resentation, is converted into phonemes, which encode its pronunciation, as observed in figure 1. G2P is significant because graphemes are representations of phonemes and they encode information such as phonotactic structures, which are crucial for language understanding. It is not trivial task, since the same grapheme can map to multiple different phonemes, resulting in a complex world of rules that need to be explored; see figure 4 for an example. (Chomsky, 2006) Essentially G2P is a direct way to study phonetic understanding in text-based machines. Traditional G2P models were constructed by linguist-defined phonotactics, such as parsers and FSAs (Finite State Automatas). (Deri and Knight, 2016) On the other hand, deep learning AI requires real world natural language examples and does not necessarily need a set of defined rules for improved performance, allowing a more representative modeling of human phonotactics. (Sajjad et al., 2022)

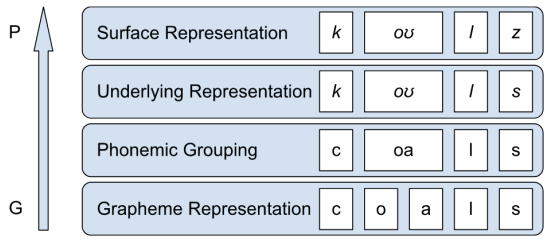

Figure 1: Pictorial representation of the G2P process.

If we extrapolate the development of language modeling, we could expect that G2P tasks on transformer language models should yield significantly better results along with interesting insights about phonetics. (Kaplan et al., 2020; Řezáčková et al., 2021) Studying a LM's NLU (Natural Language Understanding) might sound unrelated to human psycholinguistics. However, deep learning is becoming increasingly more similar to human inference learning than naive probabilistic models, to the point of being possible to consider them for human sample simulations. (Argyle et al., 2022) In addition to that, phonotactics — and language rules in general — are known to mutate over time, and there is no better way to analyze a language than looking at examples of the language itself, (Chomsky, 2006) which is exactly how neural networks learn as opposed to parser algorithms or FSA, therefore we should expect interesting phonotactic findings from example-based learning. (Sajjad et al., 2022)

In this paper, we first experiment with transfer learning on pretrained transformer models with simple G2P uniword tasks. This establishes the benchmark and motivates further fine tuning. Then, data augmentation techniques are applied to the original data; two Seq2Seq English to IPA labelled datasets are developed, `ipa_dict_seq2seq`[1] and `OpenWebIPA`[2]. Finally, the finalized model is tested on different phontactic hypotheses such as the Wug Test, and the results are recorded and analyzed.

We show that T5 fine tuned for G2P tasks passed the Wug Test, despite few training epochs. It seems to be the case that the larger the dataset, the more phonotactic patterns are acquired, similar to how human's vocabulary correlates to their morphemic and phonotactic understanding of a language. (Sajjad et al., 2022). Mainly, we set up a building foundation for future research on the understanding of human phonetic knowledge.

2 Prior Literature

2.1 Groundwork for Analysis

According to Marr (1982), there are three levels of analysis for a given system's operational algorithm, in descending level of abstraction: the computational level, which is the high level overview of the system's motivation and action plan; the algorithmic level, which explains the step-by-step process the algorithm takes; and finally the implementation level, which explains how the 'hardware' functions to make the algorithm happen. This paradigm was created to understand human psychology and cognosis, and later on was used by many cognitive scientists and computational linguistics researchers on their work, relating machine algorithms to human reasoning.

2.2 Phonotactics

Linguists revealed that a native speaker of a language can easily classify unprecedented words into either 'possible' or 'not possible' in their language. A common example being 'brick' and 'bnick', where most native English speakers would classify the former as 'English' and the latter as 'non English'. (Chomsky and Halle, 1968) At the time, this knowledge was understood but little information was available on how this knowledge was acquired or how it was represented in human psychology.

[1] https://huggingface.co/datasets/mariopeng/openIPAseq2seq
[2] https://huggingface.co/datasets/mariopeng/openwebIPA

In the Wug Test, (Gleason, 2014) the acquisition of morphological and phonological knowledge on children was explored. This experiment is a morphological evaluation, by presenting children one object and then two or more of the same object, it tests whether they can infer the object's plural form. In the core task in the experiment, an imaginary creature is presented and taught to be 'a wug', right after, two wugs would be presented, and the child is expected to say 'two wugs'.

wug + s	wugs
/wəg/ + /s/	/wəgz/

The correct answer would be pronouncing the 's' as the voiced sound /z/, as in bugs, this is due to English phonotactics, where a voiced consonant would be followed by a /z/ instead of a voiceless /s/, as in bats. To further test linguistic understanding, other question/answer pairs are included.

Q: 'a man who zibs is a?' / A: 'zibber'

All these examples were composed of made up words to ensure the right answer was not due to memorization but through inference from internalized linguistic rules. This is similar to any learning outside of the training set.

The particular rules that govern a language's phonemic composition is the language's phonotactics. A speaker's phonotactic knowledge is directly influenced by their lexical knowledge. (Frisch and Brea-Spahn, 2010) This was proved in an experiment, where the vocabulary size of a speaker was directly correlated to their performance on non-words that were well-formed. Additionally, it has also been observed in bilingual speakers that the individual knowledge of each language they know does not interfere with each other, and performance is independent per vocabulary.

2.3 Computational Approaches to G2P

Although CL VC (cross-lingual voice cloning) with IPA has not been thoroughly explored, general cross lingual IPA interpretation using G2P models have been around for decades. One method was to create phoneme and language distance metrics on high resource languages, to create models that would work for low-resource languages, which rarely have enough G2P training and evaluation dataset. (Deri and Knight, 2016) Multilingual data expansion proved to be one of the most effective ways of data augmentation for phonemic tasks.

Sometimes G2P is used as a joint tool for other tasks such as TTS (text to speech), JoinAP (Joining of Acoustics and Phonology) is a new method that does that, (Zhu et al., 2021) it encodes each phone in the IPA (International Phonetic Alphabet) to its own phonological vector. This vector representation is the phone embedding that is used by the model for multilingual speech recognition, yielding outstanding results on the CommonVoice multilingual dataset – which includes German, French, Spanish, and Italian – and the AISHLL-1 dataset – which includes Mandarin Chinese.

That proved that IPA is a good source of information for phonological analysis, arguably the most complete phonetic alphabet that exists, however, there are not enough studies of it on cross-lingual text-to-speech (CL TTS). By using IPA characters as input, researchers also explored building a a cross-lingual TTS model for CL VC. and the results were optimal. (Zhang et al., 2021) Through the experiments, it is clear that incorporating phoneme characters such as those from the IPA is relevant for increased performance on any form of acoustic tasks, which entails a more complete information encoding.

On the other hand, to enhance G2P tasks, researchers moved from traditional computational models to neural network based architectures, following the trend of language modeling. (Mayer and Nelson, 2020) Those experiments not only proved that neural models are the correct direction for advancement – by showing their advantage on unprecedented G2P tasks – but also revealed interesting insights on phonotactic learning and generalization representations.

As of most recently, in the work of Řezáčková et al. (2021), the large pretrained language model T5 (Raffel et al., 2020) was finetuned on English and Czech G2P tasks, yielding SOTA G2P results on uniword and homograph tasks on the sampa alphabet. Later, Zhu et al. (2022) performed G2P tests on ByT5 for increased G2P performance. However, none of them analyzed the phonological behavior of the language models they developed.

2.4 Transformer Architectures

The neural models used to test state of the art G2P were RNN (recurrent neural networks) and LSTMs (Long Short Term Memory) architectures. However, current state of the art results on NLP tasks are all achieved by pre-training a transformer archi-

tecture on a large corpus and then fine-tuning it on domain-specific tasks.

The understanding that the more complex the machine the more similar it is to humans has been around for decades. With the rapid advancement of the transformer architecture, researchers now expect the use of language models as proxies to human samples, to the extent of being able to model different sub-populations for social science research by mitigating bias. (Argyle et al., 2022)

One of the most popular transformer architectures is GPT-3, (Brown et al., 2020) by drawing comparison with humans, who require only a handful of examples to complete unprecedented tasks, they developed an autoregressive model that could perform in few-shot settings with no finetuning or domain adaptation whatsoever. GPT-3 performed strongly on a wide array of tasks, including translation, QA (question-answering), and other non NLP tasks such as simple arithmetic and step-by-step reasoning. This few shot learning behavior allowed researchers to thoroughly analyze GPT-3's semantic and syntactic understanding, by providing series of input-output pairs.

In BERT, (Devlin et al., 2019) the pretraining of bidirectional representations is explored as an alternative to traditional transformer representations by analyzing an unlabeled text through reading both left and right context in all layers. One of the highlights of pretrained BERT is that, at the time of fine tuning, it only needs one additional output layer to achieve state-of-the-art performance on a wide variety of tasks. Aggregating on BERT's original paper, NMT (neural machine translation) using BERT was explored. (Zhu et al., 2020) Due to the lack of previous work, the researchers had to start from the beginning. First they tried to initialize downstream models and fine-tuning them afterwards. This method turned out unsuccessful yielding no significant improvement. The second method was to use BERT representations as context-aware embeddings for other downstream models. This led to the exploration of the BERT-fused models, which were the main finding of their paper: BERT is used to obtain the representations of the input sequence, and these representations are fused with each layer of the NMT encoder-decoder transformer model through attention mechanisms. These publications show the power of raw pretrained transformers and how much can be achieved by simple finetuning or transfer learning.

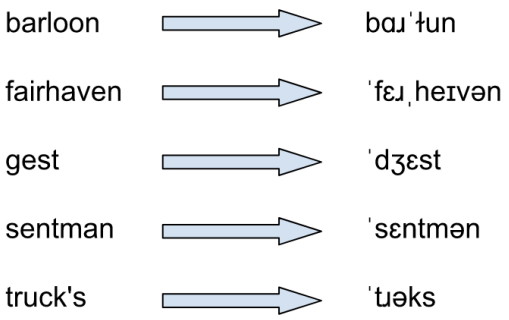

Figure 2: a random sample from the `ipa_dict en_US`, powered mostly by the CMU Pronunciation Dictionary.

In particular, T5 (Raffel et al., 2020) was designed in the context of being an encoder-decoder pretrained purely on a text-to-text format for transfer learning. The general model outperformed many domain-specific models on a wide range of NLP tasks and even achieved state of the art results on most of them, including machine translation, summarizing, QA, and text classification.

2.5 This Paper

Thus, G2P tasks are effective experiments for phonotactic knowledge, which is an interesting proxy for linguistic understanding, which explains a big part of human psychology. This paper thus investigates the study of phonotactics through a deep learning approach, building on top of the work by previous G2P models, analyzing phonotactics at a computational level of abstraction. By intervening with language models, we relate the computational approaches to G2P to human phonotactic reasoning, since it is presumed that more advanced the machine, the more similar it is to human reasoning. (Argyle et al., 2022) Previous G2P work is catching up to current language model technology, this leaves an unexplored gap in NLP that needs to be strengthened.

3 Data

G2P in many ways is similar to machine translation, where a word or sequence of words are mapped to their respective pronunciation, in this case denoted by the IPA. In addition to existing data, the lack of English to IPA datasets motivated us to introduce our own datasets for this work.

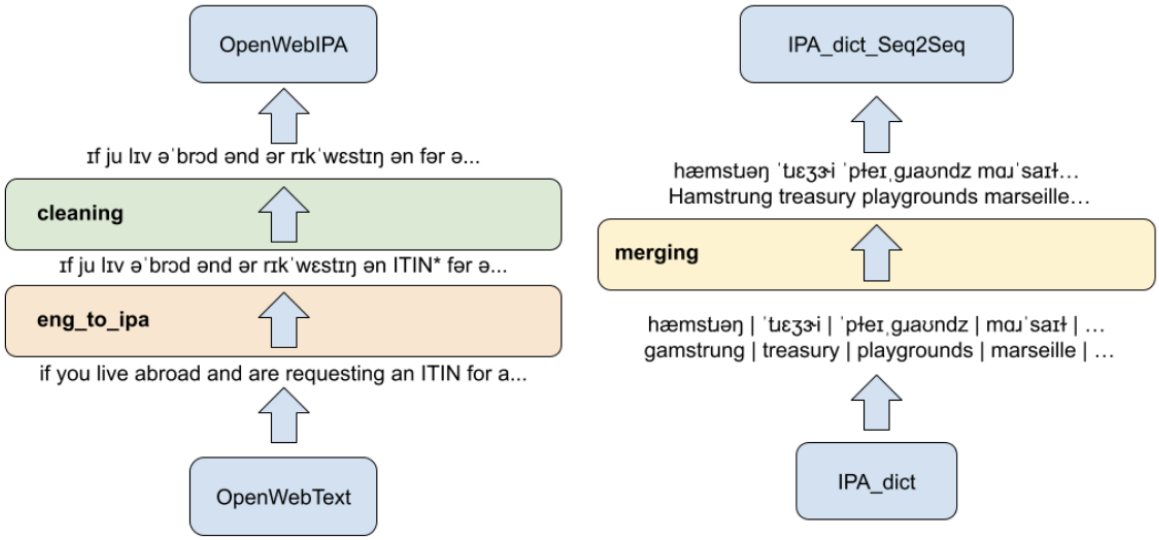

Figure 3: the data preprocessing pipeline represented, along with an example entry from each dataset.

3.1 Single Word Dictionary

We used the `ipa_dict` American English (`en_us`) dataset for the main G2P task. In essence it is a dictionary dataset with words as keys and pronunciation transcribed in IPA as values. (CMU, 2015) Depicted in figure 2. It has 125927 entries, the dataset was randomly shuffled with a constant seed and split using the 90/10 rule into a training set of 113328 entries and an evaluation set of 12599 entries, with an average of 7.49 characters per word and 8.05 characters per label. Taking words as prompts and IPA transcriptions as labels, this dataset was ideal for the models to learn primitive G2P translation. `ipa_dict` offers a variety of languages, but only English was used, this was due to better understand monolingual behavior before advancing into bilingual or multilingual in future work.

3.2 Seq2Seq

For Seq2Seq fine tuning, we needed a dataset that contained sequences of plain text in English with their respective sequences of IPA as labels. Two approaches were taken for two new datasets, preprocessing can be visualized in figure 3. The first dataset was created by concatenating from 15 to 30 random words from the `ipa_dict` dataset, this created nonsense sentences but correctly allowed the model to perform multi-word English to IPA translation. The second dataset was `openwebipa`. This dataset was crafted by using the python module eng-to-ipa [3]– which is based on the CMU pronunciation dictionary (CMU, 2015) – on a subset of the `openwebtext` corpus, resulting in a 300k entries corpus of internet text with IPA labels. Open web text is a collection of 8013769 passages from the web. (Gokaslan and Cohen, 2019) Some words that are not in the dictionary parser module were deleted by a linear algorithm. This resulted in both datasets containing nonsense sentences, which due to the research scope of this paper, is not a true limitation. However, one limitation of this dataset is that there are no interesting phonotactic interactions across words since every word is converted to IPA individually. This could be explored in the future.

4 Model

For all tasks, we utilized T5 encoder-decoder architecture. (Raffel et al., 2020) We used this language model as a baseline for transfer learning because T5 was pretrained on multilingual NMT tasks, which in theory should facilitate G2P because T5 has been shown to be easy to fine tune to other NMT tasks on both Seq2Seq and character level, as depicted in figure 4. The three models of T5 evaluated were 1) T5-small with 60M parameters and 6 layers, 2) T5-base with 220M parameters, 12-layers, and 3) T5-large, with 770M parameters and 24 layers, all pretrained on the same corpus. We used the Pytorch implementation of T5 from the HuggingFace

[3]https://pypi.org/project/eng-to-ipa/

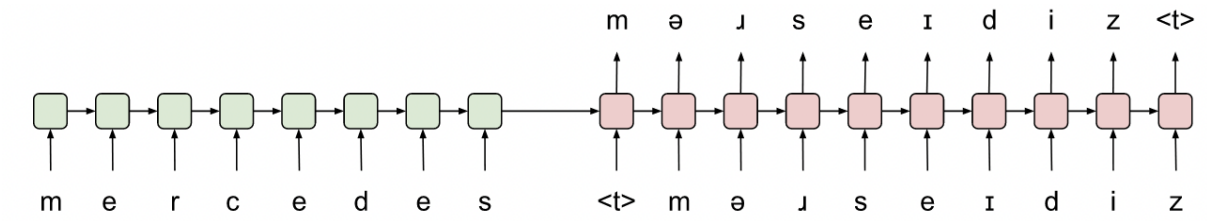

Figure 4: Overview of the G2P process in an encoder-decoder transformer architecture such as T5. As observed, the single grapheme /e/ can be 'mapped' to 3 different phonemes depending on the surrounding context.

Transformers library.[4] This simplified the training process. Details about training hyperparameters are found in appendix 8 table 3.

Since the base models of T5 do not contain IPA symbols in their vocabulary, all special IPA characters contained in the labels of the datasets where added to the tokenizer and the model dimensions were adjusted accordingly. These characters can be found in appendix 8, figure 5.

5 Methods

5.1 G2P Performance

G2P takes a plain text word or sequence as input and returns the corresponding IPA translation. In a theoretical perspective, G2P performance should be comparable to phonotactic understanding. By nature of being a machine translation and pseudo-speech-recognition task, we use CER (character error rate) to see how accurate the predictions are in contrast to the gold labels.

Concretely, Character Error Rate is the sum of the nunmber of substitutions, deletions, and insertions, over the sum of the number of substitutions, deletions, and correct characters from the label. The denominator is equivalent to the number of characters in the reference. See equation 1.

$$CER = \frac{S + D + I}{S + D + C} = \frac{S + D + I}{N} \quad (1)$$

5.1.1 Single Word Performance

We evaluated CER on uniword G2P tasks across 3 different sizes of T5 finetuned on the `ipa_dict en_us` dataset, this gave a broad perspective on how parameter size related to phonotactic learning.

5.1.2 Seq2Seq Performance

Due to lack of computational resources, we could only train an instance of `t5-small` on the Seq2Seq dataset `ipa_dict_seq2seq` and no model on `openwebipa`.

[4] https://huggingface.co/docs/transformers/index

5.2 Single-shot Phonotactic Experiments

In order to test whether the model acquired certain hypothesized phonotactics, a series of single word prompts were tested and their results registered. This was meant more as an structured exploratory experiment rather than a performance evaluation. All the prompt words for this part were ensured to not be in the dataset, to avoid pattern matching or memorization.

5.2.1 Wug Test

The Wug Test was conducted by feeding the following three words in respective order: "wuks", "wugs", "musips", "absapps", "appsabs". These made up words explore phonotactic understanding of the allophones of the /s/ phoneme.

5.2.2 Loan Words

A collection of 10 loan words and 3 extra words were compiled and tested on each model, comparing their understanding of English phonotactics.

6 Results

When given a single word, the best model, T5 large, had a CER of 0.45. This is not SOTA but is considerably good for only 10 training epochs. When given a sequence of words as opposed to a single word, performance dropped significantly for all models except `t5-small-s2s`, which was only trained on Seq2Seq tasks.

All models passed the Wug Test by correctly articulating the plural form of 'wug' as expected by the original experiment. Results seen in table 2. The models differed greatly in loan word responses. Although there is no evaluation metric for the loan word, native speakers should easily identify if the articulated response is correct, this aspect of the paper could be analyzed further in future Linguistic research. Results are found in appendix 8, table 4.

CER performance		
Model	Uniword	Seq2Seq
t5-small	51.89	92.25
t5-small-s2s	116.13	91.83
t5-base	46.08	92.27
t5-large	**44.77**	**92.08**

Table 1: Character Error Rate. As observed in the equation 1, character error rate can exceed 1.0, this is likely the case when the number of extra incorrect insertions are larger than the number of correct characters.

7 Analysis

Overall, it is difficult to understand whether the performance of the model was good or bad due to lack of previous work in this field. However, we can affirm with high certainty that the results are neither SOTA nor the limit of G2P performance on language models.

G2P performance was high for single words but dropped considerably for sequence of words, this was an interesting finding which most probably is due to T5 being pretrained with heavy positional embedding encoding. If future work performs interpretability on the attention heads and layers, we could hypothesize that mostly those layers or blocks governing the first word in a sequence were being trained during the uniword G2P fine tuning while the other ones were not.

Uniword G2P results correspond to the scaling laws, in which performance increases with model size. Kaplan et al. (2020) However, a curious behavior can be seen in `t5-small` and `t5-base`'s results. Despite a `6%` difference in Uniword performance, there is negligible difference in Seq2Seq performance. Furthermore, `t5-large` has an approximately `7%` performance difference with `t5-small` on uniword tasks, yet their difference on Seq2Seq tasks is only `0.17%`. This seems to align with the interpretation that the models tend to predict the first word correctly and proceed to fail on subsequent words. This seems to align with the fact that each entry in the dataset has on average 22.5 words per sequence, considering the first word of each of those sequences is exactly right, the average CER should be around `95%`, but since no model has `100%` accuracy, it is expected that `95%` is the maximum possible performance if only the first word is being properly translated.

Due to lack of a computational evaluation method for the Wug Test and loan word experiments, we can only analyze this section's results through a linguistics perspective.

The Wug Test provided useful information on phonotactic acquisition in language models. The results suggest that the plural morpheme 's' is correctly voiced whenever preceded by a voiced phoneme. However, the fact that this behavior was correct on 'appsabs' yet not on 'absapps' suggests that positional information of a character is affecting the model's phonotactic understanding. This can be seen in table 2.

The loanword experiments showcase that larger models seem to acquire phonotactic rules at a faster rate, and is able to generalize to excpetion words such as 'chow mein' and 'coyote', while smaller models struggle to do so. An interesting prompt is 'cocoa', since English phonotactics disallow the /oa/ diphthong, we can see that the models all applied a voiced consonant insertion in their answers between /o/ and /a/, e.g. [kəkoʊdə]. This suggests a partial understanding of restricted diphthongs in the language. However, the fact that this is not exhibited in `t5-large` could potentially imply that it learns to generalize better, as a true human speaker would, by breaking the rules. We could draw the comparisong between the behavior of the smaller models to FSA, where exceptions are 'squeezed' into the existing phonotactics. When given the prompt 'tsunami', none of the models got the /ts/ right, and curiously that particular consonant cluster is not in the English phonotactics. Furthermore, models seem to be able to correctly replicate Spanish and Italian words such as 'tacos' and 'fresco' with high similarity as human speakers would, as noted in the CMU dictionary. Lastly, for the prompt 'receipt', only `t5-large` was close to the correct answer, showing great phonotactic knowledge. Overall, all the examples, included in appendix 8 table 4, exhibit similarities to human phonetic reasoning and further analysis with a better trained model should be conducted for more interesting behavior. (Chomsky, 2006; Gleason, 2014)

8 Conclusion

Phonotactic is a core component of Linguistics, yet it remains relatively outdated in terms of its language modeling analysis. In synthesis, computational tools for phonetics studies have been following NLP's footsteps – which mostly focuses on model semantics and syntax analysis – this opens up a research opportunity to take phonetic studies

Wug Test Input-Output Pairs					
Model	wuks	wugs	musips	absapps	appsabs
t5-small	[wəks]	[wəgs]	[məsəps]	[æbsəps]	[æpsəbz]
t5-base	[wəks]	[wəgs]	[mjuzɪps]	[æbsəps]	[æpsæbz]
t5-large	[wəks]	[wəgs]	[mjuzɪps]	[æbsəps]	[æpsæbz]

Table 2: the Wug Test results.

to the next step of deep learning evolution in the large pretrained transformers paradigm. This exploratory investigation builds upon recent research and shows that there is a research gap that might be promising for human phonetic understanding and further work needs to be done. The paper provides two English to IPA datasets for G2P tasks. Most importantly, future research directions are discussed.

Limitations

The main limitations of this work were the lack of computational resources and time. Due to the size of large pretrained models, training resulted to be resource intensive. Furthermore, the lack of monolingual open source IPA labeled data resulted in considerable effort put into creating new datasets for this specific G2P task. The Seq2Seq IPA labelled datasets were generated but not used.

In terms of validity, it is worth mentioning that only uniword behavior was studied thoroughly, and behavior may change drastically on sequences of words or with Seq2Seq English to IPA machines. Uniwords also forced heteronyms to be completely neglected. A richer dataset could resolve these issues. Finally, multilingual data augmentation could have been implemented theoretically without damaging the quality of monolingual phonetic understanding, however, due to the time constraints of this project, this angle was not explored.

Future Work

This research raises numerous questions that should be answered in future work. Due to the time constraints imposed on this project, three ideas were not implemented which would have built on top of this work; firstly, a phonetic model could be conjoined with a simple IPA to Acoustic dictionary, resulting in a cheaper text to speech with performance dependant on English to IPA CER; secondly, BERT could not be tested due to IPA characters not being easily added to the BERT tokenizer vocabulary, however, future work could explore BERT architecture performance on G2P; thirdly, a model trained on the Seq2Seq datasets will heoretically yield better performance on both uniword and Seq2Seq tasks.

In addition, bilingual or multilingual studies of this same project could yield results that align with the hypotheses of this paper and present new behaviors that are not present in monolingual models. This could be done by implementing multilingual data augmentation, which theoretically would instrumentally yield better performance on G2P tasks.

Another interesting research direction would be to create a large pretrained language model purely on phonological domains, either G2P or P2P (phoneme to phoneme) tasks. The knowledge representation of this model should be compared to traditional Text to Text models, we foresee promising interpretability research on these two different paradigms using tools such as those presented by Tenney et al. (2020). Further motivation includes few-shot prompting as introduced by Brown et al. (2020), which reveals interesting behavior about generalized knowledge. We expect rising interest in interdisciplinary interpretability research along with cognitive science and psycholinguistics to understand both human and deep learning reasoning and data representation.

In terms of IPA dataset, a state of the art G2P model could create a large English to IPA dataset for further phonetic analysis. Related to this, self-supervised learning (Hinton et al., 2006) should be explored to counteract the lack of IPA labelled data for G2P training. Masked training as seen in BERT is another alternative solution. (Devlin et al., 2019) However, machine generated data would resolve the issue of not having enough datasets to train inter-word phonetic interactions or new word/heteronym labelling, since unexpected interactions may occur as in the case of /e/ being mapped to multiple phonemes in figure 4.

Authorship Statement

Author worked individually, with no external collaborators for this project.

Acknowledgements

I want to thank the Fall 2022 XCS224U staff, particularly my Course Facilitator, Raul, and Professor Potts for this learning opportunity.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. Out of One, Many: Using Language Models to Simulate Human Samples. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862. ArXiv:2209.06899 [cs].

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Noam Chomsky. 2006. *Language and mind, 3rd ed*. Language and mind, 3rd ed. Cambridge University Press, New York, NY, US. Pages: xviii, 190.

Noam Chomsky and Morris Halle. 1968. THE SOUND PATTERN OF ENGLISH. page 242.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. ArXiv:2204.02311 [cs].

CMU. 2015. The carnegie mellon university pronouncing dictionary.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-Phoneme Models for (Almost) Any Language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 399–408, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Stefan A. Frisch and María R. Brea-Spahn. 2010. Metalinguistic judgments of phonotactics by monolinguals and bilinguals. *Laboratory Phonology*, 1(2):345–360. Publisher: De Gruyter Mouton.

Jean Berko Gleason. 2014. Wug Test and the Elicited Production Paradigm. In *Encyclopedia of Language Development*, pages 687–688. SAGE Publications, Inc., Thousand Oaks.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus.

Zellig S. Harris. 1954. Distributional Structure. *WORD*, 10(2-3):146–162.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. ArXiv:1512.03385 [cs].

Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. ArXiv:2001.08361 [cs, stat].

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA.

Connor Mayer and Max Nelson. 2020. Phonotactic learning with neural language models. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 291–301, New York, New York. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv:1910.10683 [cs, stat].

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. ArXiv:2102.12092 [cs].

Kanishka Rao, Fuchun Peng, Haşim Sak, and Françoise Beaufays. 2015. Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4225–4229. ISSN: 2379-190X.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Khan, and Jia Xu. 2022. Analyzing Encoded Concepts in Transformer Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3082–3101, Seattle, United States. Association for Computational Linguistics.

Alex Sherstinsky. 2020. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *Physica D: Nonlinear Phenomena*, 404:132306. ArXiv:1808.03314 [cs, stat].

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. ArXiv:2008.05122 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. ArXiv:1706.03762 [cs].

Haitong Zhang, Haoyue Zhan, Yang Zhang, Xinyuan Yu, and Yue Lin. 2021. Revisiting IPA-based Cross-lingual Text-to-speech. ArXiv:2110.07187 [cs, eess].

Chengrui Zhu, Keyu An, Huahuan Zheng, and Zhijian Ou. 2021. Multilingual and crosslingual speech recognition using phonological-vector based phone embeddings. ArXiv:2107.05038 [cs, eess].

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into Neural Machine Translation. ArXiv:2002.06823 [cs].

Markéta Řezáčková, Jan Švec, and Daniel Tihelka. 2021. T5G2P: Using Text-to-Text Transfer Transformer for Grapheme-to-Phoneme Conversion. In *Proc. Interspeech 2021*, pages 6–10.

A Appendices

Model Training Specifications				
Parameter	t5-small	t5-small-s2s	t5-base	t5-large
layer_norm_epsilon	1e-06	1e-06	1e-06	1e-06
feed_forward_proj	ReLU	ReLU	ReLU	ReLU
learning_rate	5e-5	5e-5	5e-5	5e-5
optimizer	AdamW	AdamW	AdamW	AdamW
drop_out_rate	0.1	0.1	0.1	0.1
batch_size	256	256	128	64
num_epochs	100	5	60	10

Table 3: Details of the different T5 model configurations for transfer learning on G2P tasks

Single Shot Loan Word Prompts			
prompt	t5-small	t5-base	t5-large
tacos	tækoʊz	tɑkoʊz	tɑkoʊz
brigade	bɹeɪgd	bɹɪgəd	bɹaɪgeɪd
coyote	kaɪət	kɔɪoʊt	koʊjoʊt
fresco	fɹɛskoʊ	fɹɛskoʊ	fɹɛskoʊ
zucchini	kəntʃiz	zətʃɛni	sətʃuni
Illinois	ɪɬəkənz	ɪɬəsəns	ɪɬɪnaʊ
chow mein	khaʊmaɪn	tʃaʊmaɪn	tʃaʊmaɪn
tsunami	taɪmnɑs	taɪndɹɑmən	təmsuoʊ
cocoa	kəkoʊdə	kəkɑmə	koʊkoʊə
gif	gɪf	dʒaɪf	dʒaɪf
mischievous	mɪstʃivəs	mɪʃivəs	mɪstʃivəs
wednesday	wɛndzdeɪ	wɛnzdeɪ	wɛndzdeɪ
receipt	ɹisɪpʃən	ɹisɛpʃən	ɹisit

Table 4: 10 Loan Word G2P prompts, borrowed from French, Spanish, Italian, Irenwa, Chinese, and Japanese. In addition to that, some extra commonly mispronounced English words were included.

Figure 5: All the special characters introduced in the datasets. This was used to update the tokenizer to avoid unknown tokens.