One-Vs-Rest Neural Network English Grapheme Segmentation: A Linguistic Perspective

Anonymous EMNLP submission

Abstract

Grapheme-to-Phoneme (G2P) correspondences form foundational frameworks of tasks such as text-to-speech (TTS) synthesis or automatic speech recognition. The G2P process involves taking words in their written form and generating their pronunciation. In this paper, we critique the status quo definition of *grapheme*, currently a forced alignment process relating a single character to either a phoneme or a blank unit, that underlies the majority of modern approaches. We develop a linguistically-motivated redefinition from simple concepts such as vowel and consonant count and word length and offer a proof-of-concept implementation based on a multi-binary neural classification task. Our model achieves state-of-the-art results with a 31.86% Word Error Rate on a standard benchmark, while generating linguistically meaningful grapheme segmentations.

1 Introduction

Segmenting words into graphemes is crucial for accurate and reliable text-to-speech systems (Le et al., 2020; Taylor, 2022; Ying et al., 2024), as well as providing a tokenisation framework for training language models for use by varied segments of society (Raškinis et al., 2019; Basher et al., 2023). The currently predominant approach to G2P, which extracts phonemes from a list of graphemes, is one of forced alignment (Williams et al., 2024; Gao et al., 2024; Cheng et al., 2016; Rao et al., 2015). In this approach, a grapheme is defined as a single character that either does or does not have a respective phoneme when using G2P correspondences. This process is illustrated in Table 1 (a) with blank units denoted as φ. However, from a linguistic perspective, a grapheme is not just a single character, but a representation of a phoneme, consisting of up to four characters (Brooks, 2019). Redefining the notion of grapheme could therefore change sub-word tokenisation, allowing for models to be trained on a set of compound graphemes in addition to providing a more linguistically correct method to split words into phonemes. This is shown in Table 1 (b).

The contributions of this paper are as follows:

- We redefine the concept of graphemes in G2P segmentation, aligning it with *Referential Conception* theory (Kohrt, 1986).
- We present a novel twin-staged method for (a) G2P segmentation and (b) phoneme correspondences that equals state-of-the-art approaches on a standard CMUDict benchmark.
- We release a new dataset to the community, *EngGraph*, a subset of CMUDict, with 9,641 pre-transcribed graphemes to enable future grapheme segmentation research.

2 Related Work

LSTM-based G2P Significant advances in LSTM models for G2P have commonly relied on a one-to-one mapping between graphemes and phonemes. Rao et al. (2015) introduced a unidirectional LSTM with output delays, achieving a word error rate (WER) of 25.8% on the CMUDict benchmark by ensuring 1:1 phoneme-grapheme alignment (e.g., "google" transcribed to g, u, g, @, l, ϕ, where ϕ is a placeholder). Mousa and Schuller (2016) addressed the many-to-many alignment issue with a bidirectional LSTM (BLSTM), achieving a 23.23% WER on the same task by adding a linear projection layer, splicing window, and decoding beam to a 4-layer BLSTM network to improve alignment. Yao and Zweig (2015) achieved a 23.55% WER with a BLSTM and character-to-phoneme alignment that allowed for single, multiple, or no corresponding phonemes (e.g., "tangle" transcribed to T, AE, NG, G, AH: L, NULL).

Attention-based G2P Recent advances in attention mechanisms and transformers have largely kept to the same definition of a grapheme. Toshniwal and Livescu (2016)'s early ensemble model

Word	Grapheme Transcription (a)	Phoneme Transcription (a)	Grapheme Transcription (b)	Phoneme Transcription (b)
accuse	a-c-c-u-s-e	@-k-φ-U-z-φ	a-cc-u-se	@-k-U-z
commercial	c-o-m-m-e-r-c-i-a-l	k-ah-m-φ-e-r-s-h-ah-l	c-o-mm-er-ci-a-l	k-ah-m-er-sh-ah-l
boulevard	b-o-u-l-e-v-a-r-d	b-φ-uh-lə-φ-v-φ-ar-d	b-ou-le-v-ar-d	b-ou-lə-v-ar-d

Table 1: Current (a) and proposed (b) linguistic Grapheme transcription examples

with global attention achieved a 20.24% WER on the CMUDict task, struggling with foreign names, a common issue in G2P models (Waxmonsky and Reddy, 2012). Řezáčková et al. (2021)'s Text-to-Text Transfer Transformer showed a 0.96% WER, but similarly struggled with unseen words, increasing errors to 33.8%. Dong et al. (2022)'s BERT model had a 23.36% WER on Dutch due to English complexities, making it a less comparable baseline.

We advocate for a precise linguistic definition of graphemes, as accurate G2P conversion is vital for natural and clear speech synthesis. Mousa and Schuller (2016)'s models adopt a many-to-many alignment, but still miss the essential graphemic units of trigraphs (e.g., "ear" in "research" for the /ɛ:/ phoneme), quadgraphs (e.g., "ough" in "thought" for the /c:/ phoneme), and split digraphs, a non contigous two character grapheme, (e.g., "a.e" in "rationale" for the /eɪ/ phoneme).

3 Linguistic Definitions of Graphemes

A grapheme is defined as a single character, with G2P models aligning each character with a phoneme or a blank unit. There are two main linguistic theories on graphemes. Referential conception (Kohrt, 1986) defines a grapheme as the smallest written unit corresponding with phonemes, like "ph" in "phonetics" for the /f/ phoneme. This theory suggests writing depicts speech. The analogical concept (Lockwood, 2000) uses minimal pairs to show phoneme differences based on spelling, such as "t" and "k" in "parts" and "parks," arguing that writing and speech should be studied separately.

G2P correspondences balance these theories by viewing graphemes as influencing pronunciation but also as distinct from phonemes in TTS research. This hybrid approach presents challenges. Given the focus on TTS in G2P models, we propose adopting the referential conception for computational linguistic applications in English as in these purposes writing is being used to mimick and create spoken language. We rely on Brooks (2019), who conducted a detailed analysis of British English spelling, identifying 284 graphemes: 89 in the 'main system' and 195 in the 'extended system,' corresponding to 543 phonemes. Brooks notes that while the number of graphemes remains the same in American English, phoneme correspondences differ to reflect pronunciation differences.

Grapheme Length	Main System	Extended System
Single Character	26	0
2 Characters	53	118
3 Characters	10	57
4 Characters	0	20

Table 2: Grapheme lengths for the main and extended system (Brooks, 2019)

Analysing grapheme lengths highlights flaws in current G2P models, see Table 2. Current models, which use only single or digraph graphemes, fail to handle the complexities of English, leading to mispronunciations. For instance, without recognising trigraphs, TTS systems can add an extra phoneme in the G2P stage, such as an additional /d/ in "acknowledge." Proper grapheme segmentation transcribes the word as "a-ck-n-o-w-le-dge" with the "dge" grapheme represented with a single /g/ sound with the d being silent, enhancing pronunciation accuracy for simple and complex words.

4 Case Study

4.1 Data Analysis

The initial task of this project was to compile a comprehensive corpus of English words along with their grapheme transcriptions. The Oxford English Dictionary states that the 7,000 most common English words account for 90% of word use (Oxford Dictionaries, 2023), which we used as lower bound of coverage for our resource. Given that there are no existing resources with our intended grapheme transcriptions, we selected a large set of common English words, specifically, the 10,000 most Googled British English words (WorldlyWisdom, 2021) as a basis for a new resource. All words were transcribed into grapheme form based on the guidelines in Brooks (2019). All words in this new British English dataset are also found in the American English CMUDict benchmark.

Our new corpus *EngGraph* includes 9,641 words annotated with grapheme transcriptions, grapheme counts, and basic linguistic features such as word length, vowel and consonant count.[1] Figure 1 illustrates the number of graphemes against characters, consonants and vowels. While the feature counts plotted approximate Gaussian distributions, some grapheme distributions exhibit significant skew and overlap. These deviations pose challenges for mathematical models by distorting data representation and complicating decision boundaries. Specifically, skewness results in asymmetric distributions, affecting membership function evaluations, while overlap makes class distinction difficult, leading to less precise classification and increased ambiguity.

4.2 One-vs-Rest (OvR) Model

As our key aim is to evaluate the effectiveness of our new linguistically-motivated definition of grapheme, we opt for a simple, easy-to-replicate One-vs-Rest (OvR) architecture: a set of ten identical binary feedforward neural networks, see Fig. 2. Each network has three inputs (word length, vowel count, consonant count), two dense layers with 128 units, and 30% dropout, with a binary output. The models were trained for 150 epochs with ADAM optimisation, learning rate of 0.001, a batch size of 8, and early stopping with a patience of 20 epochs.

The architecture was trained on curated subsets of our EngGraph corpus, ensuring all elements are also present in the CMUDict benchmark dataset for comparability. We generated 10 balanced data subsets by selecting all examples with a specified number of graphemes (from 1 to 10) and augmenting each subset with an equal number of examples featuring a different number of graphemes. For instance, the subset for one grapheme includes all records with one grapheme, alongside an equal number of randomly selected records with 2-10 graphemes. This approach ensures an equal distribution of true and false records for each OvR model, with a random 30% of the data reserved as a testing set. Earlier experiments with a single multi-class architecture failed with low accuracy, arguably due to complexities discussed in Figure 1.

Following the classification of grapheme counts, we developed an iterative word-to-grapheme mapping method to established word error rates. This method uses the OvR classifier with the highest confidence to identify grapheme mappings. If no mapping of graphemes with the classified number of graphemes is possible, the next highest confidence classification is used, until a valid grapheme mapping and phonetic transcription is achieved. This approach was validated against the ground truth phonetic transcriptions, resulting in a Word Error Rate (WER) of 31.86%, comparable to state-of-the-art models in Sec. 2. This indicates a significant opportunity for future refinements to enhance the accuracy of G2P transcriptions using our proposed new redefinition of graphemes in NLP.

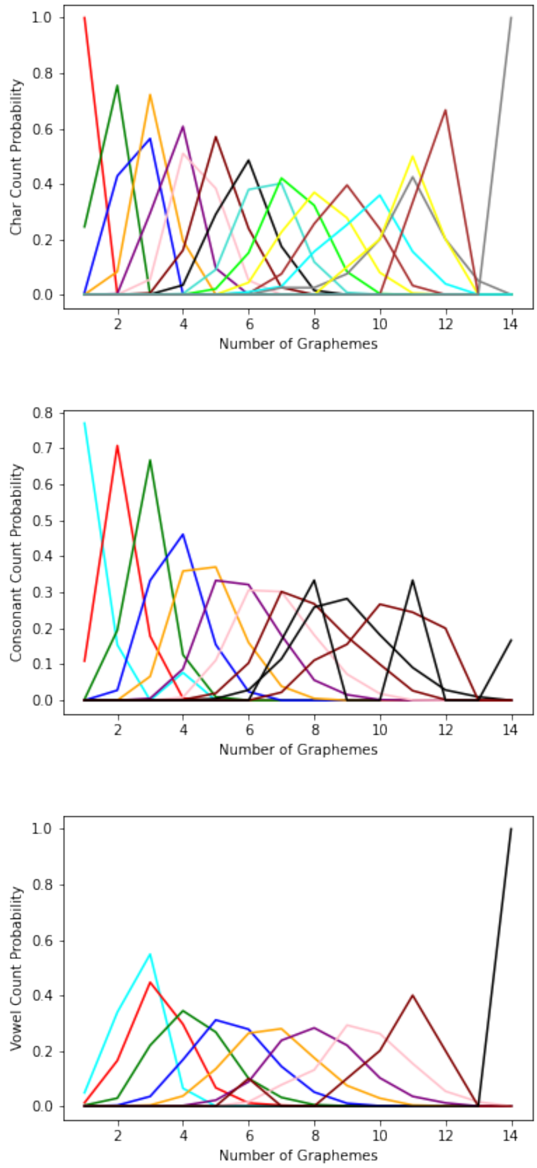

Figure 1: Character, consonant, and vowel count distributions for different numbers of graphemes.

[1]This dataset will be made freely available to the research community via GitHub upon publication.

Word	Input	One-vs-Rest Neural Network Outputs										Grapheme Count
		One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten	
labelled	[8,3,5]	0.0002	0.0001	0.0217	0.3746	0.6110	0.8101	0.1190	0.0237	0.0097	0.0057	6
ribbon	[6,2,4]	0.0009	0.0028	0.1512	0.7157	0.7987	0.2549	0.0023	0.0021	0.0016	0.0018	5
study	[5,1,4]	0.0054	0.0129	0.5885	0.8390	0.3667	0.0003	0.0001	0.0002	0.0006	0.0006	4
strengthen	[10,2,8]	0.0001	0.0000	0.0001	0.0085	0.0611	0.4718	0.8796	0.8622	0.4068	0.6171	6

Table 3: One-vs-Rest Networks Input and Output Examples

OvR Model	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
Accuracy	0.9531	0.9238	0.8744	0.7976	0.7844	0.7855	0.8402	0.8878	0.9255	0.9466
F1-Score	0.95	0.93	0.86	0.82	0.81	0.80	0.85	0.89	0.93	0.95
Recall	0.94	0.99	0.88	0.91	0.89	0.89	0.96	0.97	0.94	0.98
Precision	0.97	0.88	0.85	0.74	0.74	0.73	0.77	0.83	0.92	0.93

Table 4: One-vs-Rest Neural Network classification results, where n equals the number of graphemes.

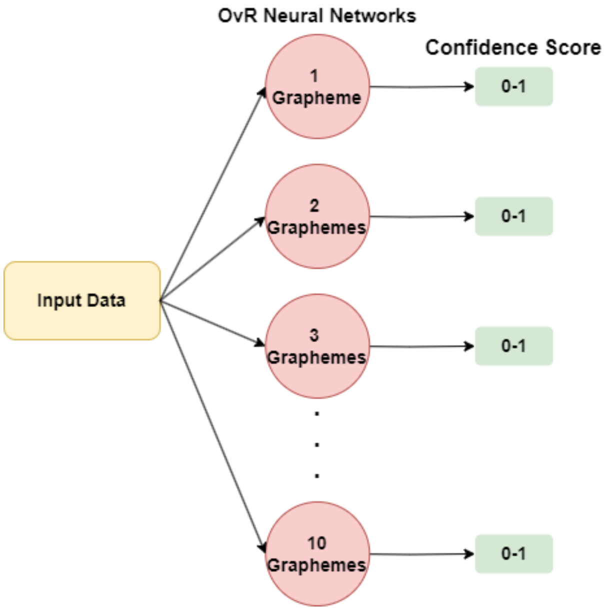

Figure 2: One-vs-Rest multi-binary MLP architecture.

4.3 Results and Discussion

The performance of these ten networks is notably high, see Table 4, and comparable to state-of-the-art WERs, despite our simple architecture. The system is computationally efficient despite maintaining ten neural networks. Our OvR format ensures each model is trained on a balanced dataset, distinguishing the characteristics of words with a specified number of graphemes, which adds transparency to grapheme analyses. Our multi-network system is easily extendable, e.g. new datasets can accommodate longer, more linguistically complex words, and more complex neural architectures may further enhance classification performance. Table 3 shows examples of network inputs and outputs, where 3/4 predictions matched the correct grapheme count, while the fourth was off by one.

5 Conclusion

Our redefinition of *graphemes*, inspired by the referential conception theory of linguistics, has profound implications for the task of G2P. Already matching state-of-the-art methods using a simple architecture, our research challenges current methodologies, highlighting the limitations of single-character graphemes, and offering a more inclusive framework for text representation and semantic research. This shift paves the way for more accurate, culturally-sensitive language processing systems. This paper advances NLP research by advocating for hybrid graphemes, addressing critical gaps in existing methods. It provides practitioners with tools to improve the performance and adaptability of their applications, and encourages exploration of the phonetics-semantics connection, influencing text tokenisation, segmentation, and feature extraction in NLG applications. Additionally, the application of hybrid graphemes will aid in speech recognition tasks, such as differentiating homophones, and modelling dialect differences in English, reflecting true linguistic diversity and additionally allowing for more culturally sensitive models.

6 Limitations

Our study has several limitations that should be noted. The dataset, while comprehensive, includes only 9,641 words and focuses on British English pronunciation, potentially limiting its applicability to other English dialects and languages. In addition, while all elements of EngGraph are present in the standard CMUDict Dataset, our study is looking at British English compared to American English and additionally our dataset is not at expansive as the CMUDict dataset which has over 134,000 words with their phonetic transcription. The preprocessing steps and basic feature set, including word length, vowel count, and consonant count, may not fully capture the nuances required for accurate grapheme segmentation, particularly for irregular, slang, borrowed, or complex words. Additionally, the model's simple architecture, though computationally efficient, may not perform as well as more advanced architectures like transformers.

The use of Word Error Rate (WER) as the primary evaluation metric, while standard, does not fully reflect linguistic accuracy, particularly for partial matches. Ethical considerations include potential biases in the dataset, which overlooks regional dialects and minority languages, impacting accessibility and fairness in applications. Furthermore, our study has not been extensively tested in real-world scenarios, which may present challenges not accounted for in controlled experiments. Future work should explore more advanced architectures, a wider range of linguistic features, and larger, more diverse datasets, as well as extend the approach to other languages, English dialects, and real-world applications.

References

Mohammad Jahid Ibna Basher, Mohammad Raghib Noor, Sadia Afroze, Ikbal Ahmed, and Mohammed Moshiul Hoque. 2023. BnGraphemizer: A Grapheme-based Tokenizer for Bengali Handwritten Text Recognition. In *2023 IEEE 9th International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, pages 183–188. ISSN: 2837-8245.

Greg Brooks. 2019. *Dictionary of the British English Spelling System*. Open Book Publishers. Accepted: 2021-02-11T11:23:40Z.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561. Association for Computational Linguistics.

Lu Dong, Zhi-Qiang Guo, Chao-Hong Tan, Ya-Jun Hu, Yuan Jiang, and Zhen-Hua Ling. 2022. Neural Grapheme-To-Phoneme Conversion with Pre-Trained Grapheme Models. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6202–6206. ISSN: 2379-190X.

Heting Gao, Mark Hasegawa-Johnson, and Chang D. Yoo. 2024. G2PU: Grapheme-To-Phoneme Transducer with Speech Units. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10061–10065. ISSN: 2379-190X.

Manfred Kohrt. 1986. THE TERM 'GRAPHEME' IN THE HISTORY AND THEORY OF LINGUISTICS. In *THE TERM 'GRAPHEME' IN THE HISTORY AND THEORY OF LINGUISTICS*, pages 80–96. De Gruyter.

Duc Le, Thilo Koehler, Christian Fuegen, and Michael L. Seltzer. 2020. G2G: TTS-Driven Pronunciation Learning for Graphemic Hybrid ASR. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6869–6873. ISSN: 2379-190X.

David G. Lockwood. 2000. Phoneme and grapheme: how parallel can they be? *LACUS Forum*, 27:307–317. Publisher: Linguistic Association of Canada and the United States.

Amr El-Desoky Mousa and Björn Schuller. 2016. Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks for Grapheme-to-Phoneme Conversion Utilizing Complex Many-to-Many Alignments. In *Interspeech 2016*, pages 2836–2840. ISCA.

Oxford Dictionaries. 2023. Oxford English Corpus - Facts About the Language.

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.

Gailius Raškinis, Gintarė Paškauskaitė, Aušra Saudargienė, Asta Kazlauskienė, and Airenas Vaičiūnas. 2019. Comparison of Phonemic and Graphemic Word to Sub-Word Unit Mappings for Lithuanian Phone-Level Speech Transcription. *Informatica*, 30(3):573–593. Publisher: Vilnius University Institute of Mathematics and Informatics.

Jason Taylor. 2022. *Pronunciation modelling in end-to-end text-to-speech synthesis*. Ph.D. thesis, University of Edinburgh. Accepted: 2022-06-13T13:54:29Z Publisher: The University of Edinburgh.

Shubham Toshniwal and Karen Livescu. 2016. Jointly learning to align and convert graphemes to phonemes with neural attention models. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 76–82.

Sonjia Waxmonsky and Sravana Reddy. 2012. G2P Conversion of Proper Names Using Word Origin Information. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 367–371, Montréal, Canada. Association for Computational Linguistics.

Samantha Williams, Paul Foulkes, and Vincent Hughes. 2024. Analysis of forced aligner performance on L2 English speech. *Speech Communication*, 158:103042.

WorldlyWisdom. 2021. Google 10000 English.

Kaisheng Yao and Geoffrey Zweig. 2015. Sequence-to-Sequence Neural Net Models for Grapheme-to-Phoneme Conversion. ArXiv:1506.00196 [cs].

Zelin Ying, Chen Li, Yu Dong, Qiuqiang Kong, Qiao Tian, Yuanyuan Huo, and Yuxuan Wang. 2024. A Unified Front-End Framework for English Text-to-Speech Synthesis. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10181–10185. ISSN: 2379-190X.

Markéta Řezáčková, Jan Švec, and Daniel Tihelka. 2021. *T5G2P: Using Text-to-Text Transfer Transformer for Grapheme-to-Phoneme Conversion*. International Speech Communication Association. Accepted: 2022-03-28T10:00:27Z ISSN: 2308-457X.