INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Mitigating the Exposure Bias in Sentence-Level Grapheme-to-Phoneme (G2P) Transduction

Eunseop Yoon[1,*], *Hee Suk Yoon*[1,*], *Dhananjaya Gowda*[2], *SooHwan Eom*[1], *Daehyeok Kim*[1], *John Harvill*[3], *Heting Gao*[3], *Mark Hasegawa-Johnson*[3], *Chanwoo Kim*[2], *and Chang D. Yoo*[1†]

[1]Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
[2]University of Illinois at Urbana-Champaign, USA, [3]Samsung Electronics, Republic of Korea

{esyoon97, hskyoon, sean1105, kimshine, cd_yoo}@kaist.ac.kr,
{d.gowda, chanw.com}@samsung.com, {harvill2, hgao17, jhasegaw}@illinois.edu

Abstract

Text-to-Text Transfer Transformer (T5) has recently been considered for the Grapheme-to-Phoneme (G2P) transduction. As a follow-up, a tokenizer-free byte-level model based on T5 referred to as ByT5, recently gave promising results on word-level G2P conversion by representing each input character with its corresponding UTF-8 encoding. Although it is generally understood that sentence-level or paragraph-level G2P can improve usability in real-world applications as it is better suited to perform on heteronyms and linking sounds between words, we find that using ByT5 for these scenarios is nontrivial. Since ByT5 operates on the character level, it requires longer decoding steps, which deteriorates the performance due to the exposure bias commonly observed in auto-regressive generation models. This paper shows that the performance of sentence-level and paragraph-level G2P can be improved by mitigating such exposure bias using our proposed loss-based sampling method.

Index Terms: grapheme-to-phoneme conversion, phonetic transcription, ByT5, exposure bias

1. Introduction

Grapheme-to-phoneme (G2P) transduction is essential for various applications that require phonetic lexicon, including automatic speech recognition (ASR) [1] and text-to-speech synthesis (TTS) [2]. There are two primary types of G2P tasks: word-level and sentence-level. In word-level G2P, the goal is to predict the pronunciation of a single word, whereas sentence-level G2P[1] involves predicting the pronunciations of all the words in a sentence. The latter is a more challenging task as it requires modeling context-dependent pronunciation variations of words (i.e., heteronyms) and linking sounds between words, making it more suitable for real-world applications.

With the recent advancements in deep learning, transformer-based encoder-decoder language models, such as the Text-to-Text Transfer Transformer (T5) [3], have emerged as a powerful tool for the G2P conversion [4, 5, 6] where they learn to map input sequences (i.e., graphemes) to their corresponding output sequences (i.e., phonemes). Particularly, ByT5, a byte-level model based on T5, has been introduced as a token-free language model representing each character in a sequence with its corresponding UTF-8 encoding. Previous research shows that ByT5 gives promising results on word-level G2P conversion [6], outperforming traditional token-based models. In this paper, we take a step towards extending ByT5 to the sentence-level G2P conversion.

*Equal contribution

†Corresponding author

[1]For simplicity, sentence-level refers to both the sentence and paragraph-level (i.e., more than one sentence) G2P transduction.

However, despite their remarkable performance, transformer-based models are affected by *exposure bias*, which is a fundamental issue in auto-regressive generation models [7]. This problem arises due to the discrepancy between the maximum likelihood training (also referred to as teacher forcing [8]) and the generation procedure during inference [9]. As a result, errors can accumulate and propagate throughout the generation process, leading to a significant decrease in performance as the decoded sequence becomes longer.

We show that due to the aforementioned exposure bias, the performance of ByT5 significantly deteriorates when applied to sentence-level G2P tasks due to the model's character-level operation, which leads to longer decoding sequences. This paper proposes a loss-dependent sampling method that builds upon a previously proposed two-pass decoding strategy [10] used to mitigate the exposure bias in Natural Language Processing (NLP). Our method involves identifying positions in the sequence with a high probability of wrong prediction by calculating the cross-entropy loss for each position. The positions with higher loss values are then sampled more frequently during training, in which their predictions are replaced into the ground truth phoneme sequence before inputting into the decoder. By sampling the positions in the sequence where errors are likely to occur during training, the proposed technique allows the model to learn from these errors and improve its ability to correct them. During this process, we use an adaptive sampling ratio determined by the Phoneme Error Rate (PER) of the previous epoch to determine the number of desired replacements adaptively.

With extensive experiments, we show that our loss-dependent sampling method improves the overall performance on our curated English G2P benchmark and quantitatively shows the reduction of exposure bias using the recently proposed metric [9]. Moreover, we present some examples showing improved phoneme prediction due to the mitigation of exposure bias and improved heteronyms prediction over the word-level G2P ByT5.

2. Related work

2.1. Grapheme-to-Phoneme (G2P) Conversion

Early G2P models were based on a pronunciation dictionary, which looked up the corresponding pronunciation of the letter sequence. However, the size of the dictionary had to be very large and costly, and even then, they always had limited coverage with a finite-sized dictionary. Another early approach was the rule-based model, which determines the pronunciation of each letter or the subsequence based on a set of pre-defined phonetic rules [11, 12]. However, designing phonetic rules was often difficult, and still had difficulties capturing irregularities or complex rules, which are frequent in natural languages. To

 10.21437/Interspeech.2023-2336

overcome the limitations of pre-defined phonetic rules, a data-driven stochastic approach that can mitigate complicated phonetic rules from the large dataset was widely used [13, 14, 15]. These approaches usually involve weighted finite state transducers (WFST) [16, 17].

Over time, the G2P conversion studies have shifted to deep learning-based methods [4, 5, 6, 18, 19, 20]. Recently, transformer-based language models have emerged as a powerful tool for the G2P conversion [4, 5, 6] due to their effectiveness in modeling long-term dependencies in sequential data. Notably, the Text-to-Text Transfer Transformer (T5) [3], which has gained attention due to its impressive performance on various NLP tasks, has been widely adapted for the G2P task [5].

2.2. Token-Free Language Model

Recently, token-free models that do not use word tokens have been emerging in NLP, and ByT5 [21] has shown significantly better performance than traditional T5 models by using UTF-8 encoding as input. Because it does not use tokens for corresponding words or subwords, token-free models can process any language, even with unknown vocabulary. This makes it possible to handle a wide variety of languages with high performance, compared to mT5 [22] using multilingual vocabulary embedding. Zhu et al. [6] has used this advantage for multilingual G2P conversion using ByT5.

2.3. Exposure Bias

Despite their significant achievements, transformer-based models suffer from exposure bias, a fundamental problem of auto-regressive natural language generation models [7]. Exposure bias is explained to occur due to the discrepancy between training and test-time generation. During training, the model is trained based on the ground-truth data distribution. On the other hand, the model generates the next token during test time based on the prefix sequences sampled from the model itself. As a result, the error can propagate and accumulate throughout the generation process, leading to a critical degradation in performance, especially when the decoding sequence becomes longer [23]. There have been several attempts to resolve exposure bias in auto-regressive sequence generation model. One way to tackle the problem is to replace or perturb the ground-truth sequence [10, 24]. Some authors claim that the error accumulation can be solved with better decoding methods [25, 26]. Other attempts found the fault in maximum likelihood estimation (MLE) method and took non-MLE based approaches including reinforcement learning [27, 28] or generative adversarial networks [29], while He et al. [30] questions the impact of exposure bias in MLE training.

While it has been an active discussion on whether exposure bias really matters or not, Arora et al. [9] have brought up a quantifiable definition of exposure bias and showed the accumulation of error due to the exposure bias actually exists during natural language generation.

3. Loss-based Sampling

G2P is a sequence-to-sequence task where the input sequence (i.e., grapheme) $\mathbf{X} = [x_0, x_1, x_2, ..., x_n]$ is transformed to an output phoneme sequence $\mathbf{Y} = [y_0, y_1, y_2, ..., y_t]$ through a mapping model p_θ, parameterized by θ, in which case we consider using the ByT5 [6].

The training of p_θ is usually done via teacher forcing. The basic idea behind teacher forcing is to train the decoder to predict the next phoneme in a sequence given the previous phonemes (context) in that sequence. More formally, the model is trained by minimizing the negative log-likelihood for a given batch B,

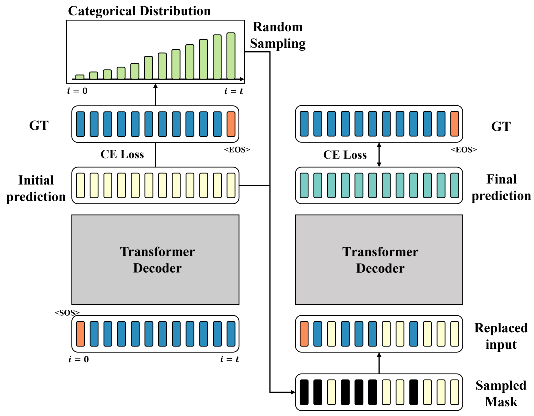

Figure 1: *Our proposed loss-dependent sampling technique. The sampling method identifies positions in the sequence where errors are more likely to occur based on their cross-entropy loss values. By emphasizing these positions during training, the model is encouraged to pay more attention to them, resulting in improved accuracy of the predicted sequence during testing.*

$$\theta^* = \arg\min_{\theta} \frac{-1}{|B|} \sum_{(\mathbf{X}, y_0^t) \in B} \sum_{i=0}^{t} \log p_\theta(y_i | y_0^{i-1}; \mathbf{X}), \quad (1)$$

where y_i is the phoneme generated at step i, and y_0^{i-1} is the context at time step i. During inference, however, the simplest strategy for generating a target sequence is to auto-regressively sample a sequence. That is, at each step i, pick the most probable phoneme $\hat{y}_i = \arg\max p_\theta(\cdot|\hat{y}_0^{i-1}; \mathbf{X})$. The process carries on until either the sequence length limit is attained, or a unique token representing the end of the sequence (EOS) is produced.

Due to this discrepancy between the training objective and inference generation procedure, also referred to as exposure bias, auto-regressive generation models show degraded performance especially as the sequence of the decoding gets longer.

3.1. Loss-Based Sampling for Exposure Bias Mitigation

Previous work proposed a sampling-based training which involves a two-pass decoding strategy where they randomly mix the gold target sequence with the predicted sequence to mitigate the exposure bias [10]. However, to give positions in the sequence where there is a high probability of wrong prediction during training, we propose a loss-based sampling method (Figure 1).

The method involves several steps, starting with inputting the ground truth phoneme sequence and obtaining the predicted phoneme probability distribution for each time step $i \in [0, t]$,

$$[p_\theta(\cdot|y_0^{-1}; \mathbf{X}), p_\theta(\cdot|y_0^{0}; \mathbf{X}), ..., p_\theta(\cdot|y_0^{t-1}; \mathbf{X})]. \quad (2)$$

We then calculate the cross-entropy loss between each of the probability distributions with the one-hot encoding of the ground truth phoneme (note that we do not collect gradients during this step).

$$[\mathrm{H}(p_o, p_\theta)_0, \mathrm{H}(p_o, p_\theta)_1, ..., \mathrm{H}(p_o, p_\theta)_t], \quad (3)$$

where $\mathrm{H}(p_o, p_\theta)_i = \mathrm{CrossEntropy}(\mathrm{onehot}(y_i), p_\theta(\cdot|y_0^{i-1}; \mathbf{X}))$.

Next, we normalize the cross-entropy loss to obtain a categorical distribution. From the distribution, we randomly sample a specific number of time steps without replacement, creating a mask to determine whether to replace each ground truth input with its corresponding prediction. During this sampling process, we use an *adaptive sampling ratio* specified by the Phoneme Error Rate (PER) of the previous epoch to determine the number of desired replacements adaptively. The replaced sequence obtained from the loss-based sampling is once again used as the decoder input. Because the second prediction is generated based on its prediction randomly sampled, the resulting output can reflect the auto-regressive behavior. The cross-entropy loss calculated from the second prediction is our final loss and is used for backpropagation.

The loss-based sampling technique we proposed allows the model to focus during training on positions in the sequence where errors are more likely to occur. By doing so, it can learn from these errors and improve its ability to correct them. This results in improved performance on longer inputs during testing, demonstrating the effectiveness of our approach.

3.2. Evaluation of Exposure Bias

Arora et al. [9] suggested using the time accumulation of the expected per-step prediction losses. During the evaluation, the decoder generates the predicted phoneme probability distribution $p_\theta(\hat{y}_i|\hat{y}_0^{i-1}; \mathbf{X})$ at each step i, based on the previous phoneme sequence $\hat{y}_0^{i-1}$ produced by the decoder. The predicted distribution is compared with the true distribution $p_o(\hat{y}_i|\hat{y}_0^{i-1}; \mathbf{X})$ to yield the per-step auto-regressive prediction loss:

$$l_i^{AR}(\mathbf{X}) = D_{\mathrm{KL}}(p_o(\hat{y}_i|\hat{y}_0^{i-1}; \mathbf{X}) \,||\, p_\theta(\hat{y}_i|\hat{y}_0^{i-1}; \mathbf{X})) \quad (4)$$

$$= \sum_{\hat{y}_i \in P} p_o(\hat{y}_i|\hat{y}_0^{i-1}; \mathbf{X}) \log \frac{p_o(\hat{y}_i|\hat{y}_0^{i-1}; \mathbf{X})}{p_\theta(\hat{y}_i|\hat{y}_0^{i-1}; \mathbf{X})}, \quad (5)$$

where P is the set of phonemes. The loss is averaged over the evaluation set D_e to approximate the expected per-step auto-regressive loss:

$$L_i^{AR} \approx \frac{1}{|D_e|} \sum_{(\mathbf{X}, y_0^t) \in D_e} l_i^{AR}(\mathbf{X}). \quad (6)$$

Similarly, the per-step teacher forcing losses are obtained by replacing $\hat{y}_0^{i-1}$ with the ground-truth phoneme sequence y_0^{i-1}, thereby assuming the absence of exposure bias,

$$L_i^{TF} \approx \frac{1}{|D_e|} \sum_{(\mathbf{X}, y_0^t) \in D_e} l_i^{TF}(\mathbf{X}, y_0^t), \quad (7)$$

$$l_i^{TF}(\mathbf{X}, y_0^t) = D_{\mathrm{KL}}(p_o(\hat{y}_i|y_0^{i-1}; \mathbf{X}) \,||\, p_\theta(\hat{y}_i|y_0^{i-1}; \mathbf{X})). \quad (8)$$

Then [9] proposed the metric $\mathrm{AccErr}_{\leq}(l)$, defined as the accumulation of the expected auto-regressive prediction loss relative to the expected teacher forcing loss until step l,

$$\mathrm{AccErr}_{\leq}(l) = l \times \frac{\sum_{i=1}^{l} L_i^{AR}}{\sum_{i=1}^{l} L_i^{TF}}. \quad (9)$$

The metric is known to have values between l and l^2 under the proper assumption. The value of l implies the absence of exposure bias for sequences of lengths up to l. On the other hand, a high deviation of $\mathrm{AccErr}_{\leq}(l)$ from l implies a high exposure bias for sequences of length l. In the worst-case scenario, the metric grows quadratically with the phoneme sequence length.

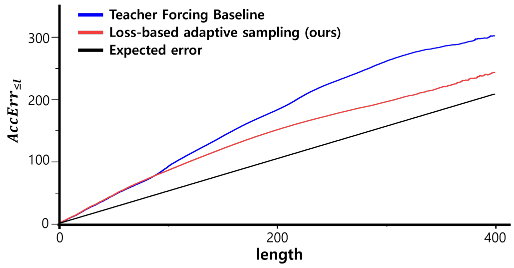

Figure 2: *Plot of accumulated error up to length l ($AccErr_{\leq}(l) = l$) w.r.t. l. The Expected Error curve represents the ideal case with no exposure bias.*

4. Experimental Settings

4.1. Dataset and Model

TIMIT [31] is a widely used corpus for speech processing tasks such as automatic speech recognition and text-to-speech. TIMIT contains 6300 spoken sentences, 10 by each of the 630 speakers from 8 major dialect regions of US English. TIMIT provides time-aligned orthographic and phonetic transcription, which can be used for labels in G2P conversion. In order to train and evaluate on the sentence-level G2P, we randomly concatenate up to 3 sentences during training. For the test set, we create two subsets, $\mathrm{Testset}_{short}$ and $\mathrm{Testset}_{long}$, which contain sentences concatenated up to 3 and 5, respectively. For the model selection, we used ByT5, a token-free model among variants of T5 [21]. For all experiments, we finetune the pre-trained word-level G2P models[2] [6] based on `ByT5-small`, which has about 300M parameters.

4.2. Training Details

We use AdamW optimizer [32] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay $\lambda = 5 \times 10^{-3}$. The learning rate is set to 10^{-5}, and the batch size is 32 for all experiments. We use gradient clipping with a 5.0 maximum gradient. The fixed sample ratio is chosen by grid search {0,1.0.3, 0.6, 0.9}, and the best ratio is independently selected for each experiment. The best model is selected by the lowest loss on the validation set. All experiments were done on NVIDIA Quadro RTX 8000.

5. Results

5.1. Experiment Results

Table 1 shows the Phoneme Error Rate (PER) and Word Error Rate (WER) of various sampling methods on different decoding strategies (greedy search and beam search) on the two groups of test set we created (i.e., $\mathrm{Testset}_{short}$ and $\mathrm{Testset}_{long}$). The naive two-pass decoding-based sampling method (uniform sampling) performs better than the teacher-forcing baseline. Our loss-based sampling method gives better results than the uniform sampling for both the fixed and adaptive sample ratio settings. Our loss-based sampling with adaptive sample ratio on $\mathrm{Testset}_{long}$ obtained 21.99% with greedy search and 21.50% with beam search, outperforming teacher forcing baseline by 2.11%p and 2.15%p, respectively.

In Figure 2, we analyze the exposure bias on teacher forcing baseline and our loss-based adaptive sampling method by plotting the accumulated error with the expected error base on

[2] `https://github.com/lingjzhu/CharsiuG2P`

Model \Decoding Strategy	Testset$_{short}$ (1-3 sentences)				Testset$_{long}$ (4-5 sentences)			
	Greedy search		Beam Search		Greedy search		Beam Search	
	PER	WER	PER	WER	PER	WER	PER	WER
Teacher forcing(baseline)	17.70	49.42	17.33	48.54	24.10	55.66	23.65	54.85
Training with sampling method								
fixed sample ratio								
· Uniform sampling	16.07	48.69	15.95	47.97	22.25	54.40	21.98	53.50
· Loss-based sampling(ours)	15.77	47.75	15.72	47.31	22.65	55.81	21.73	53.73
adaptive sample ratio								
· Uniform sampling	16.18	47.94	15.71	46.90	22.43	54.52	22.15	53.67
· Loss-based sampling(ours)	**15.64**	**47.62**	**15.56**	**46.50**	**21.99**	**54.29**	**21.50**	**53.41**

Table 1: *Comparison of PER (%) and WER (%) (lower is better) on the created test sets using TIMIT (Testset$_{short}$, Testset$_{long}$). For beam search, the number of beams is set to 3. The values reported are the average of three runs with different random seeds.*

Grapheme	Stacked heels are also popular on dressy or tailored shoes. She had your dark suit in greasy wash water all year. This can be done to a high degree of accuracy by existing surveying techniques. The small boy put the worm on the hook. Please dig my potatoes up before frost.
Ground Truth Phoneme	stæk hil̩ z ɑɪ olsoʊ pɑpjɨlɝ on dɹɛsi oɹ teɪlɝd ʃʉz ʃi ɦɛddʒ dʒɝ dɑɪk sʉt n̩ gɹisi wɑʃ worɚ ol jiɝ ðɪs kɪn bi dʌn tu ə ɦaɪ digɹi ɨv ækɚsi baɪ ɛgzɪstɨŋ sɝveɪiŋ tɛkniks ðə smol bɔɪ pʊt ðə wɝm ʌn ðə ɦʊk **pliz dɪg maɪ pɨteɪroʊz ʌp bifoɹ fɹost**
Phoneme prediction (baseline)	stæk hilz ɑɪ olsə pɑpjɨlɚ on dɹɛsi oɹ teɪlɚd ʃʉz ʃi ɦæddʒ dʒɚ dɑɪk sʉt ʔɨn gɹisi woʃ worɚ ol jɪɚ ðɪs kɨn bi dʌn tu ə ɦaɪ dɨgɹi ɨv ækjɝsi baɪ ɨgzɪstɨŋ sɝveɪŋ tɛkniks ðə smol bɔɪ pʊt ðə wɝm on ðə hʊk **<EOS>**
Phoneme prediction (ours)	stæk hilz oɑɹ olsə pɑpjɨlɝ on dɹɛsi ɚ teɪlɚd ʃʉz ʃi ɦæd jɚ dɑɪk sʉt ʔɨn gɹisi woʃ worɚ ol jɪɚ ðɪs kɨn bi dʌn tʉ eɪ aɪ dɨgɹi ɨv ækjɝɨsi baɪ ɨgzɪstɨŋ sɝveɪɨŋ tɛkniks ðə smol bɔɪ pʊt ðə wɝm on ðə hʊk **pliz dɪg maɪ pɑteɪroʊz ʌp bifoɹ fɹost**
Grapheme	Increasing tempo of new product development. He will allow a rare lie. She had your dark suit in greasy wash water all year. She had your dark suit in greasy wash water all year.
Ground Truth Phoneme	ʔɪnkɹisiŋ tɛmpoʊ əv nʉ pɹɑrɨk dɨvɛləpmɨnt hi wl̩ əlaʊ ɚ ɹɛɹ laɪ ʃi ɦæd jɚ dɑɪk sʉɨt ʔɨn gɹisi wɑʃ worɚ ol jɪɚ ʃi ɦɛd jɚ dɑɪk sʉt ʔɨn gɹisi wɑʃ worɚ ol jɪɝ **<EOS>**
Phoneme prediction (baseline)	ɪnkɹisɨŋ tɛmpoʊ əv nʉ pɹɑdʌk dɨvɛləpmɨnt hi wl̩ əlaʊ ə ɹɛɹ laɪ ʃi ɦæddʒ dʒɚ dɑɪk sʉt n̩ gɹisi wɑʃ worɚ ol jɪɚ ʃi ɦæddʒ dʒɚ dɑɪk sʉt n̩ gɹisi wɑʃ worɚ ol jɪɚ **ʃi ɦæddʒ dʒɚ dɑɪk sʉt n̩ gɹisi wɑʃ worɚ ol jɪɚ ʃi ɦæddʒ dʒɚ dɑɪk sʉt n̩ gɹisi wɑʃ worɚ ol jɪɚ ʃi ɦæddʒ dʒɚ dɑɪk sʉt n̩ gɹisi wɑʃ worɚ ol jɪɚ ʃi ɦæddʒ dʒ**
Phoneme prediction (ours)	ɪnkɹisɨŋ tɛmpoʊ əv nʉ pɹɑdʌk dɨvɛləpmɨnt hi wl̩ əlaʊ ə ɹɛɹ laɪ ʃi ɦæd jɚ dɑɪk sʉt n̩ gɹisi wɑʃ worɚ ol jɪɚ ʃi ɦæd jɚ dɑɪk sʉt n̩ gɹisi wɑʃ worɚ ol jɪɚ **<EOS>**

Figure 3: *Examples of G2P conversion using greedy decoding for baseline and our method.*

Section 3.1. The difference with the expected error is significantly reduced compared to the baseline due to our loss-based sampling method effectively mitigating the exposure bias.

5.2. Comparison of G2P Quality

Figure 3 shows a comparison of the phoneme predictions for a given Grapheme between a baseline and our model. In the first example, the baseline model fails to complete the prediction and terminates early, while our proposed model accurately predicts the full sequence. This demonstrates the superiority of our model in handling long input sequences and producing more accurate predictions. In the second example, the baseline model continues generating phonemes beyond the true length of the sequence while our model produces the correct number of phonemes. This further confirms the effectiveness of our proposed model by avoiding unnecessary phoneme generation.

5.3. Heteronyms: Context Dependency

Heteronyms are words that are spelled the same but have different pronunciations and meanings depending on the context. Word-level G2P systems cannot understand the context in which the word is used, limiting their capability of disambiguating heteronyms. On the other hand, sentence-level G2P systems can analyze the whole input sequence to determine the correct pronunciation for heteronyms. Figure 4 illustrates some examples of heteronym predictions in which it demonstrates that the previously proposed word-level ByT5 [6] is unable to disambiguate the phoneme in contrast to our sentence-level ByT5.

Heteronym	IPA	Example Sentence	Word-level G2P Model	Ours
wind	/ˈwɪnd/	The **wind** is strong at the top of the mountain.	/ˈwɪnd/	/ˈwɪnd/
	/ˈwaɪnd/	The road seems to **wind** through the mountains.	/ˈwɪnd/	/ˈwaɪnd/
read	/ɹid/	She will **read** the book that her older sister read last year.	/ɹed/	/ɹid/
	/ɹed/	She will read the book that her older sister **read** last year.	/ɹed/	/ɹed/

Figure 4: *Examples of heteronyms predictions using word-level and sentence-level G2P system (Ours).*

6. Conclusions

We find that ByT5 G2P performance for longer decoding sequences in sentence or paragraph-level G2P tasks is sub-optimal due to exposure bias. However, the proposed loss-based sampling method has been found to mitigate this problem and improve the performance of sentence and paragraph-level G2P using ByT5. These findings highlight the potential for further improvements in G2P conversion using transformer architectures.

7. Limitation

Our research on sentence and paragraph-level G2P conversion was limited by the need for high-quality phoneme-labeled data incorporating contextual information and linking sounds. The TIMIT dataset has human-labeled phonemes for text utterances but may reflect pronunciation specific to particular dialects. Despite these limitations, our study proposes a novel approach to sentence-level G2P that can address some of these challenges and provides valuable insights for future research.

8. Acknowledgements

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-01381) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics).

9. References

[1] M. A. Hasegawa-Johnson, L. Rolston, C. Goudeseune, G.-A. Levow, and K. Kirchhoff, "Grapheme-to-phoneme transduction for cross-language asr," in *International Conference on Statistical Language and Speech Processing*, 2020.

[2] Z. Hong, J. Wang, X. Qu, J. Liu, C. Zhao, and J. Xiao, "Federated Learning with Dynamic Transformer for Text to Speech," in *Proc. Interspeech 2021*, 2021, pp. 3590–3594.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *CoRR*, vol. abs/1910.10683, 2019. [Online]. Available: http://arxiv.org/abs/1910.10683

[4] S. Yolchuyeva, G. Németh, and B. Gyires-Tóth, "Transformer based grapheme-to-phoneme conversion," in *Interspeech*, 2019.

[5] M. Rezácková, J. Svec, and D. Tihelka, "T5g2p: Using text-to-text transfer transformer for grapheme-to-phoneme conversion," in *Interspeech*, 2021.

[6] J. Zhu, C. Zhang, and D. Jurgens, "Byt5 model for massively multilingual grapheme-to-phoneme conversion," in *Interspeech*, 2022.

[7] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *CoRR*, vol. abs/1511.06732, 2015.

[8] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989.

[9] K. Arora, L. E. Asri, H. Bahuleyan, and J. C. K. Cheung, "Why exposure bias matters: An imitation learning perspective of error accumulation in language generation," *arXiv preprint arXiv:2204.01171*, 2022.

[10] T. Mihaylova and A. F. T. Martins, "Scheduled sampling for transformers," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 351–356. [Online]. Available: https://aclanthology.org/P19-2049

[11] H. S. Elovitz, R. W. Johnson, A. McHugh, and J. E. Shore, "Automatic translation of english text to phonetics by means of letter-to-sound rules," NAVAL RESEARCH LAB WASHINGTON DC, Tech. Rep., 1976.

[12] R. M. Kaplan and M. Kay, "Regular models of phonological rule systems," *Comput. Linguist.*, vol. 20, no. 3, p. 331–378, sep 1994.

[13] R. I. Damper, Y. Marchand, M. J. Adamson, and K. Gustafson, "Comparative evaluation of letter-to-sound conversion techniques for english text-to-speech synthesis," in *Speech Synthesis Workshop*, 1998.

[14] M. Bisani and H. Ney, "Investigations on joint-multigram models for grapheme-to-phoneme conversion," *7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2002.

[15] M. Razavi, R. Rasipuram, and M. Magimai.-Doss, "Acoustic data-driven grapheme-to-phoneme conversion in the probabilistic lexical modeling framework," *Speech Commun.*, vol. 80, pp. 1–21, 2016.

[16] D. Caseiro, I. Trancoso, L. O. Inesc-Id, Ist, R. A. Redol, and C. Viana, "Grapheme-to-phone using finite-state transducers," *Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002.*, pp. 215–218, 2002.

[17] J. R. Novak, N. Minematsu, and K. Hirose, "WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding," in *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*. Donostia–San Sebastián: Association for Computational Linguistics, Jul. 2012, pp. 45–49. [Online]. Available: https://aclanthology.org/W12-6208

[18] S. Yolchuyeva, G. Németh, and B. Gyires-Tóth, "Grapheme-to-phoneme conversion with convolutional neural networks," *Applied Sciences*, vol. 9, p. 1143, 03 2019.

[19] K. Rao, F. Peng, H. Sak, and F. Beaufays, "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4225–4229.

[20] K. Yao and G. Zweig, "Sequence-to-sequence neural net models for grapheme-to-phoneme conversion," *CoRR*, vol. abs/1506.00196, 2015. [Online]. Available: http://arxiv.org/abs/1506.00196

[21] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel, "ByT5: Towards a token-free future with pre-trained byte-to-byte models," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 291–306, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.17

[22] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 483–498. [Online]. Available: https://aclanthology.org/2021.naacl-main.41

[23] F. Schmidt, "Generalization in generation: A closer look at exposure bias," in *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 157–167. [Online]. Available: https://aclanthology.org/D19-5616

[24] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 1171–1179.

[25] S. Welleck, I. Kulikov, J. Kim, R. Y. Pang, and K. Cho, "Consistency of a recurrent language model with respect to incomplete decoding," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 5553–5568. [Online]. Available: https://aclanthology.org/2020.emnlp-main.448

[26] R. Collobert, A. Y. Hannun, and G. Synnaeve, "A fully differentiable beam search decoder," *ArXiv*, vol. abs/1902.06022, 2019.

[27] Y. Chen, L. Wu, and M. J. Zaki, "Reinforcement learning based graph-to-sequence model for natural question generation," *ArXiv*, vol. abs/1908.04942, 2019.

[28] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. C. Courville, and Y. Bengio, "An actor-critic algorithm for sequence prediction," *ArXiv*, vol. abs/1607.07086, 2016.

[29] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, p. 2852–2858.

[30] T. He, J. Zhang, Z. Zhou, and J. R. Glass, "Exposure bias versus self-recovery: Are distortions really incremental for autoregressive text generation?" in *Conference on Empirical Methods in Natural Language Processing*, 2019.

[31] C. Lopes and F. Perdigão, "Timit acoustic-phonetic continuous speech corpus," 2012.

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7