A Two-Step Approach for Data-Efficient French Pronunciation Learning

Hoyeon Lee[1]   Hyeeun Jang[2]   Jong-Hwan Kim[1]   Jae-Min Kim[1]

[1]NAVER Cloud   [2]Université de Strasbourg
yeon.lee@navercorp.com

Abstract

Recent studies have addressed intricate phonological phenomena in French, relying on either extensive linguistic knowledge or a significant amount of sentence-level pronunciation data. However, creating such resources is expensive and non-trivial. To this end, we propose a novel two-step approach that encompasses two pronunciation tasks: grapheme-to-phoneme and post-lexical processing. We then investigate the efficacy of the proposed approach with a notably limited amount of sentence-level pronunciation data. Our findings demonstrate that the proposed two-step approach effectively mitigates the lack of extensive labeled data, and serves as a feasible solution for addressing French phonological phenomena even under resource-constrained environments.

1 Introduction

Phonetic information plays a crucial role in text-to-speech systems, improving the clarity and naturalness of synthetic speech. Grapheme-to-phoneme (G2P) relationships are typically modeled using a sizeable set of phonetic transcriptions to predict the pronunciation of out-of-vocabulary words. However, pronunciation learning in French remains challenging due to its intricate phonetic structure and phonological phenomena such as *Linking (Enchaînement)* and *Liaison*.

These phenomena mediate between words by modifying phonemes and their placement (Adda-Decker et al., 1999; Bybee, 2001). *Linking* is the articulation of a consonant-final word and its re-syllabification with the following vowel-initial word (Gaskell et al., 2002; Fougeron et al., 2003). For example, when a consonant-final word *une* - [yn] precedes a vowel-initial word *amie* - [a.mi], the phoneme "n" is resyllabified and positioned adjacent to the primary phoneme of the following word, as illustrated in the example below:

une [yn] *amie* [a.mi] → *une amie* [y.**n**a.mi]

Liaison refers to the pronunciation of a silent consonant-final word to its phoneme or another when it is followed by a vowel-initial word (Bybee, 2001; Gaskell et al., 2002). For instance, when the determiner *mes* is an initial word, its pronunciation is altered depending on the following word. If the following word, such as *frères*, begins with a consonant, the "s" grapheme remains silent in the corresponding pronunciation.

mes [me] *frères* [fʁɛʁ] → *mes frères* [me.**f**ʁɛʁ]

However, when it is followed by a vowel-initial word like *amis*, the "s" is pronounced as [z].

mes [me] *amis* [ami] → *mes amis* [me.**z**a.mi]

These phonetic modifications are influenced by various factors and contexts, leading to numerous exceptions. Such variability further adds to the complexity of addressing phonological phenomena[1].

Recent studies have focused on modeling these phonological phenomena through two main approaches: post-lexical rules (PLR) (Tzoukermann, 1998) and data-driven methods (Pontes and Furui, 2010; Taylor et al., 2021; Comini et al., 2023). While both PLR and data-driven methods achieve decent performance, they demand deeper phonological/linguistic knowledge and a substantial amount of human-annotated sentence-level data, respectively. Unfortunately, existing approaches without such expensive resource commitments have not been satisfactory.

In this paper, we propose a novel two-step approach to address the challenge of learning French pronunciation with limited resources. Specifically, we explicitly decompose the intricate and comprehensive pronunciation task into two sub-tasks: G2P conversion and post-lexical processing. First, we leverage a large amount of easily accessible word-level pronunciation data to train the autoregressive transformer (ART)-based G2P model (Co-

---

[1]Further elaboration and examples of the complexities are described in Appendix A.
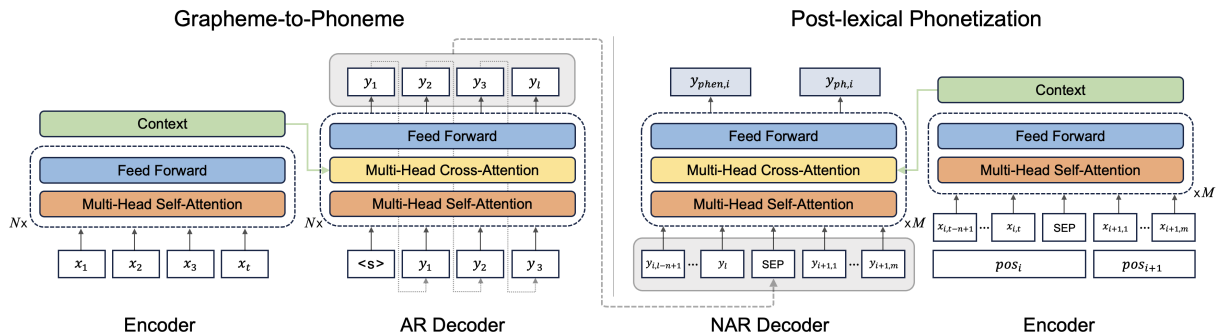
Figure 1: An overview of our proposed architecture.

mini et al., 2023; Yolchuyeva et al., 2019; Yu et al., 2020; Zhu et al., 2022) and generate correct pronunciations for corresponding words. Second, we adopt a shallow non-autoregressive transformer (NART) (Gu et al., 2018; Sun et al., 2019) as the post-lexical phonetization model to process phonological phenomena between the pronunciations of individual words. This model is trained on our manually constructed dataset, comprising a modest number of sentence-level examples.

Accordingly, we assess whether the proposed approach can effectively leverage a limited set of sentence-level examples to overcome the challenges, and further analyze how varying the size of these resources impacts performance. The experimental results reveal that the proposed approach successfully addresses intricate phonological phenomena, utilizing only around 2k examples, with even 1.5k examples proving somewhat effective.

2 Related Work

PLR is one of the essential modules used to address phonological phenomena in the French text-to-speech front-end, yet it requires extensive phonological and linguistic knowledge to construct a comprehensive set of hand-crafted post-lexical rules. Although only a few studies[2] (Tzoukermann, 1998) provide initial guidelines for manually constructing post-lexical rules, the implementation of their intricate interactions still necessitates the deep knowledge and substantial efforts of linguistic experts.

To alleviate this burden, data-driven approaches have also been proposed (Pontes and Furui, 2010; Taylor et al., 2021; Comini et al., 2023). Unlike PLR, these approaches demonstrate performance improvements by leveraging large-scale sentence-level pronunciation datasets, even without linguistic knowledge. Comini et al. (2023) reported using

a pronunciation dataset for training, comprising about 33.7k sentences and 800k words, with the sentence-level pronunciations generated via internal front-ends. Nevertheless, they also highlighted that, even with 33.7k sentence-level phonetic transcriptions, the absence of data specifically designed for post-lexical processing may still cause certain contexts to be missed.

3 Proposed Method

In this section, we describe our two-step approach to French pronunciation learning. Our approach addresses the extensive and complex pronunciation task by explicitly decomposing it into two key sub-tasks: G2P conversion and post-lexical processing. The overall architecture is illustrated in Figure 1.

3.1 Grapheme-to-Phoneme

To generate pronunciation for a given word, we employ a vanilla ART architecture (Vaswani et al., 2017). Following the autoregressive model-based sequence-to-sequence paradigm applied in G2P (Milde et al., 2017; Peters et al., 2017; Yolchuyeva et al., 2019; Yu et al., 2020; Zhu et al., 2022; Comini et al., 2023), the encoder transforms the grapheme sequence $x = \{x_1, x_2, ..., x_t\}$ into contextual information, and the decoder generates the corresponding phoneme sequence $y = \{y_1, y_2, ..., y_l\}$ based on the encoder's output. We train the ART G2P model on word-level pronunciation data, which includes <*word, pronunciation*> pairs like <*enfant, ɑ̃fɑ̃*>. During training, we use cross-entropy (CE) loss between the generated phonemes $\hat{y}$ and ground truth phonemes y.

3.2 Post-lexical Phonetization

The key intuition underlying our approach is to adopt a separate post-lexical processing module rather than directly predicting phoneme sequences

---

[2] http://research.jyu.fi/phonfr/20.html

covering the phonological phenomena. The hypothesis is that learning French pronunciation, including the post-lexical phenomena, is particularly challenging when relying on a limited sentence-level dataset. This challenge is compounded by the fact that each example contains merely a few words affected by the phenomena. This dedicated module targets post-lexical phonetization, effectively leveraging a modest number of sentence-level examples.

The post-lexical phonetization model follows a similar architecture as the G2P model of the first sub-module. The key distinction here is using a fairly shallow, non-autoregressive architecture. The encoder compresses the concatenated grapheme sequences of the partial grapheme sequence of the word x_i and the following word x_{i+1}, along with the part-of-speech (POS) tags extracted from a pre-trained POS model. Given that post-lexical phenomena are related to the graphemes between the immediately surrounding words, we use the final n graphemes of the word x_i and the initial m graphemes of the following word x_{i+1}. A [SEP] token is inserted between each partial grapheme sequence. When the number of graphemes in each word is fewer than the pre-defined values of n and m, a [PAD] token is added to the beginning or end of each sequence, respectively. This approach allows the encoder to effectively capture contextual information essential for processing post-lexical phenomena, leveraging both grapheme sequences and POS tags of adjacent words.

Conditioned on the contextual information extracted from the encoder, the decoder then predicts whether a phonological phenomenon occurs y_{phen} and the resulting phoneme y_{ph}. We utilize word-level phoneme sequences predicted by the pre-trained G2P model as input for the decoder. Similar to the encoder input, the decoder's input comprises the concatenation of partial phoneme sequences, encompassing the final n phonemes of the word x_i and the initial m phonemes of the subsequent word x_{i+1}. [SEP] and [PAD] tokens are appended in the same manner as the encoder.

We then introduce two loss terms to train the phonetization model. The first loss term ($\mathcal{L}_{phen}$) is the weighted binary cross-entropy (WBCE) loss used to identify the presence of the phonological phenomenon in class-imbalanced settings:

$$\mathcal{L}_{phen} = \text{WBCE}\left(\hat{y}_{phen}, y_{phen}\right) \tag{1}$$

where y_{phen} is the ground truth of phonological phenomenon occurrence, and $\hat{y}_{phen}$ is the predicted

	N	SD
Examples	2,645	-
Examples w/ phonological phenomena	2,107	-
Examples w/o phonological phenomena	538	-
Avg. words	12.27	2.96
Avg. *Liaison*	0.60	0.76
Avg. *Linking*	0.83	0.90

Table 1: Statistics of sentence-level pronunciation dataset. The number of examples and the average number of words and occurrences of phonological phenomena per example.

probability of the phonological phenomenon. The second loss term ($\mathcal{L}_{ph}$) is the CE loss, used for predicting pronunciation variation as follows:

$$\mathcal{L}_{ph} = y_{phen} \cdot \text{CE}\left(\hat{y}_{ph}, y_{ph}\right) \tag{2}$$

where y_{ph} is the ground truth of the phoneme at locations where phonological phenomenon may occur. The key characteristic is adding a simple selector that is selectively activated based on the occurrence of a phenomenon, as follows:

$$y_{phen,i} = \begin{cases} 1, & \text{if post-lexical phenomenon} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Overall, the loss of the phonetization model is the sum of $\mathcal{L}_{ph}$ and $\mathcal{L}_{phen}$, as follows:

$$\mathcal{L}_{plp} = \mathcal{L}_{ph} + \mathcal{L}_{phen} \tag{4}$$

4 Experiment

4.1 Datasets

First, we collect pronunciation data at both the word and sentence levels. Following Comini et al. (2023), we gather word-level pronunciation data using an internal pronunciation dictionary, which contains 106,857 unique entries, each with phonetic transcriptions denoted using X-SAMPA notation. Furthermore, we collect 2,645 sentence-level data from various domains, including news, social media posts, and the Multilingual LibriSpeech (MLS) (Pratap et al., 2020) dataset. Each example is manually annotated by a French linguistic expert using the same X-SAMPA phonetic transcription as the word-level data, including phonological phenomena like *Liaison* and *Linking*[3]. Each sentence consists of an average of 12.27 (SD = 2.96) words, and the average frequency of phonological phenomena occurrences per sentence is 1.43 (SD =

---

[3]The details of the phonetic transcription and example are provided in Appendix B.1

1.13). Table 1 reports the detailed statistics of the sentence-level dataset.

In addition to resource creation, we apply only essential preprocessing steps and curation criteria to avoid examples that may negatively impact the performance of our proposed method, while preserving the original context and natural structure of the data. The overall preprocessing techniques and curation criteria applied to the dataset are described in Appendix B.2.

4.2 Experimental Setup

We split the word-level dataset into training (85%), validation (5%), and test (10%) sets for the G2P model training. For the post-lexical phonetization model, the number of training instances k varies from 2,045 to 512, decreasing in 25% intervals, with 300 examples for both the validation and test sets. To evaluate the G2P and post-lexical phonetization models, we use the following metrics: phoneme error rate (PER), word error rate (WER), and Acc_{plp} (Appendix B.3).

Within our proposed method, the ART G2P model is mainly implemented following the setup described in (Zhu et al., 2022). The transformer encoder and decoder consist of 8 layers each, with 8 self-attention heads, 512-dimensional embeddings, and 2048 feed-forward dimensions, resulting in 58.9M parameters. For the post-lexical phonetization model, we employ a shallow NART architecture. This architecture consists of just 2 transformer layers, each with 8 self-attention heads, and 512-dimensional embeddings, resulting in 14.8M parameters only. The total number of parameters for the proposed architecture is around 73.7M. Additional details of other configurations are provided in Appendix B.4.

Building on this foundation, we compare our proposed method to the ART-based G2P model, which serves as our baseline due to its renowned superior performance in existing approaches. This model is trained in three distinct settings: using word-level data (Yolchuyeva et al., 2019), sentence-level data, and a combination of both (Comini et al., 2023). To ensure a fair comparison with the proposed method, we employ a 10-layer ART G2P model and follow the same procedure to identify optimal hyperparameters in all settings.

4.3 Results and Analysis

We evaluate the proposed approach alongside several baseline models, by breaking down the analy-

Annotation	Phonological Case	$Acc_{plp}\uparrow$	PER $\downarrow$	WER $\downarrow$
Word	Whole sentence	**84.92**	**10.64**	**24.68**
	Phonological phen.	0.00	**12.40**	**29.03**
	Liaison	0.00	**11.87**	**28.66**
	Linking	0.00	**14.58**	**33.65**
Sentence	Whole sentence	14.37	136.48	95.04
	Phonological phen.	13.23	135.71	95.07
	Liaison	12.68	133.60	95.45
	Linking	13.69	136.10	95.37
Word/Sentence	Whole sentence	77.85	29.91	34.39
	Phonological phen.	**55.38**	32.79	37.21
	Liaison	**67.32**	34.00	37.24
	Linking	**45.23**	33.10	38.72

Table 2: Evaluation of the baseline models with different types of pronunciation datasets.

sis into four distinct cases: Whole sentence, Phonological phenomena, *Liaison*, and *Linking*.

Can the baseline models address phonological phenomena? We train the naive ART G2P model on three distinct sets of pronunciation data, with the experimental results detailed in Table 2. As expected, the model trained solely on word-level phonetic transcription data was completely unable to address any phonological phenomena, resulting in an Acc_{plp} of 0% in all phonological phenomena cases. Despite this, when leveraging around 100k entries of large-scale word-level data for training, the model showed adequate performance in PER and WER, aligning with the findings of previous studies. For the model trained on sentence-level pronunciation data, we can observe a slight improvement in addressing phonological phenomena with an Acc_{plp} of 13.23%. However, due to the limited number of training examples, the overall performance substantially declined. In contrast, compared to using sentence-level data alone, combining a large amount of word-level data with a small amount of sentence-level data led to a considerable improvement. PER and WER significantly decreased in all cases, reaching 29.91% and 34.39%, respectively. Acc_{plp} also improved relative to other baselines, reaching 55.38% in the Phonological phenomena case; however, it remains markedly inferior compared to the Whole sentence.

How effective is the two-step approach in addressing phonological phenomena? Initially, we evaluate the G2P model of the proposed approach to generate correct word-level pronunciations that serve as inputs for the subsequent post-lexical phonetization model. Interestingly, the G2P model demonstrates performance on par with the

	25% of *Full*		50% of *Full*		75% of *Full*		*Full*	
	Acc_{plp} ↑	PER/WER ↓	Acc_{plp} ↑	PER/WER ↓	Acc_{plp} ↑	PER/WER ↓	Acc_{plp} ↑	PER/WER ↓
Whole sentence	93.80(+16.0)	5.67/14.40	93.93(+16.1)	5.69/13.54	94.52(+16.7)	5.27/12.80	**95.52(+17.7)**	**4.79/11.47**
Phonological phen.	69.51(+14.1)	6.20/15.57	71.30(+15.9)	6.24/14.55	80.99(+25.6)	5.64/13.46	**83.86(+28.5)**	**5.18/12.11**
Liaison	63.73(-3.6)	6.06/15.49	67.16(-0.2)	6.20/15.17	68.16(+0.8)	**5.42**/13.71	**77.56(+10.2)**	5.53/**13.06**
Linking	74.38(+29.2)	6.64/16.92	74.79(+29.6)	6.52/15.42	91.74(+46.5)	5.73/13.81	**89.21(+44.0)**	**5.11/12.11**

Table 3: Evaluation of the proposed approach with varying numbers of sentence-level instances. *Full* denotes using all sentence-level instances. Differences from the baseline, trained on word/sentence-level data, are in parentheses.

baseline model, achieving a PER of 9.78%, a WER of 24.95%, and an Acc_{plp} of 84.95% in the Whole sentence case. Based on the aforementioned G2P model, we evaluate our proposed approach, with the overall results illustrated in Table 3. As anticipated, by utilizing the entire dataset, the *Full* model achieved the best performance in all cases, with the only exception of the PER in the *Liaison* case, which used 75% of the *Full* dataset. Relative to the baseline, trained on sentence-level data capable of handling phonological phenomena to a certain extent, we can observe a considerable improvement in all metrics. Most notably, the Acc_{plp} for phonological phenomena cases showed a remarkable increase to 83.86%, up by an average of 28.5%, representing a substantial rise from Whole sentence. This result demonstrates the notable effectiveness of the two-step approach in addressing phonological phenomena, even with limited data.

How many sentence-level phonetic transcriptions are needed for our approach to work well? Reflecting the application in certain scenarios, possibly constrained by the more extreme scarcity of consistently labeled sentence-level data (Lee et al., 2023), we conduct a further empirical analysis exploring how resource size affects performance across all cases. As illustrated in Table 3, decreasing the number of training instances resulted in a progressive decline in performance across all cases, relative to the *Full* model. Although using less than half of the entire dataset led to improvements over the baseline in most cases, a marginal performance drop from 0.2% to 3.6% was observed in the *Liaison* case. This implies that relying on an excessively small number of examples may miss certain contexts, proving insufficient to capture the entire spectrum of phonological variation on contextual factors (De Mareüil et al., 2003; Kondo, 2012; Encrevé, 1983). In contrast, it is noteworthy that utilizing more than 75% of the entire dataset leads to a noticeable improvement, surpassing 80%

in the Phonological phenomena case. Specifically, the *Linking* case achieves around 90% in Acc_{plp}. Drawing from these results, we conjecture that addressing phonological phenomena requires a minimum of about 1.5k sentence-level examples.

5 Conclusion

In this paper, we present an effective two-step approach for French pronunciation learning. Our approach alleviates the burden of extensive resources by decomposing the intricate and comprehensive pronunciation task into two sub-tasks, thereby facilitating greater leverage of a modest number of sentence-level examples. The empirical analysis demonstrates the efficacy of our proposed approach in addressing phonological phenomena even in resource-constrained environments.

6 Limitations

This study has a few important limitations. We employed closed-source datasets. Although some open-source word-level pronunciation datasets exist, such as the one provided by Zhu et al. (2022), which contains approximately 250k French word-pronunciation pairs and is significantly larger than ours, we chose not to use it due to the presence of noise. Moreover, finding a publicly available dataset designed for post-lexical processing has been significantly challenging, as even sentence-level phonetic transcriptions reflecting phonological phenomena are not available.

Our manually constructed sentence-level dataset contains a smaller amount of annotated data compared to previous research. This limited size may not completely capture the generalizability of the entire spectrum of French phonological phenomena described in Appendix A. While this smaller dataset may result in some missing contexts, it represents our main contribution. Despite being a data-driven method, we achieved significant results using only less than 2k sentence-level pronunciation data.

The post-lexical phonetization model in our proposed method predicts a single resulting phoneme, which serves as the ground truth at locations where a phonological phenomenon may occur, though it cannot address the few cases involving more than one phonetic shift (including vowel changes) beyond the final grapheme of the first word in a sequence. For instance, in certain cases like *bon ami*, a phonetic shift occurs not at the final grapheme, but at the second-to-last grapheme:

bon [bɔ̃] *ami* [ami] → *bon ami* [b**o**.na.mi]

This represents a vowel shift, where the nasal vowel [ɔ̃] transitions to the oral vowel [o] during the *liaison* process.

We evaluated the proposed approach by comparing it with data-driven approaches, rather than a PLR method. The major challenge in comparing with a PLR method is that it requires significant expertise, as well as substantial temporal and financial costs for the manual construction of numerous post-lexical rules and their interactions. Therefore, we focused on comparisons with data-driven approaches, which may be more feasible solutions in resource-constrained environments.

References

Martine Adda-Decker, Philippe Boula de Mareüil, and Lori Lamel. 1999. Pronunciation Variants in French: Schwa & Liaison. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, pages 2239–2242.

Joan Bybee. 2001. Frequency Effects on French Liaison. *Frequency and the Emergence of Linguistic Structure*, pages 337–359.

Joan Bybee. 2005. La liaison: effets de fréquence et constructions 1. *Langages*, (2):24–37.

Giulia Comini, Manuel Sam Ribeiro, Fan Yang, Heereen Shim, and Jaime Lorenzo-Trueba. 2023. Multilingual context-based pronunciation learning for Text-to-Speech. In *Proceedings of INTERSPEECH*.

Benoit Crabbé and Marie Candito. 2008. Expériences d'analyse syntaxique statistique du français. In *15ème conférence sur le Traitement Automatique des Langues Naturelles-TALN'08*.

P Boula De Mareüil, Martine Adda-Decker, and Véronique Gendner. 2003. Liaisons in French: a corpus-based study using morpho-syntactic information. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 1329–1332.

Jacques Durand and Chantal Lyche. 2008. French liaison in the light of corpus data. *Journal of French Language Studies*, 18(1):33–66.

Pierre Encrevé. 1983. La liaison sans enchaînement. *Actes de la recherche en sciences sociales*, 46(1):39–66.

Cécile Fougeron, Odile Bagou, Alain Content, Muriel Stefanuto, and Ulrich Frauenfelder. 2003. Looking for acoustic cues of resyllabification in French. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 2257–2260.

M Gareth Gaskell, Elsa Spinelli, and Fanny Meunier. 2002. Perception of resyllabification in French. *Memory & cognition*, 30:798–810.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2018. Non-Autoregressive Neural Machine Translation. In *International Conference on Learning Representations*.

Nori Kondo. 2012. La liaison facultative dans l'interaction spontanée et les manuels de fle. *Flambeau*, 38:31–51.

Bernard Laks. 2005. La liaison et l'illusion. *Langages*, (2):101–125.

Hoyeon Lee, Hyun-Wook Yoon, Jong-Hwan Kim, and Jae-Min Kim. 2023. Cross-Lingual Transfer Learning for Phrase Break Prediction with Multilingual Language Model. In *Proceedings of INTERSPEECH*.

Benjamin Milde, Christoph Schmidt, and Joachim Köhler. 2017. Multitask Sequence-to-Sequence Models for Grapheme-to-Phoneme Conversion. In *Proceedings of INTERSPEECH*, pages 2536–2540.

Ben Peters, Jon Dehdari, and Josef van Genabith. 2017. Massively Multilingual Neural Grapheme-to-Phoneme Conversion. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*.

Josafá de Jesus Aguiar Pontes and Sadaoki Furui. 2010. Modeling Liaison in French by Using Decision Trees. In *Proceedings of INTERSPEECH*, pages 186–189.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Proceedings of INTERSPEECH*.

Zhiqing Sun, Zhuohan Li, Haoqing Wang, Di He, Zi Lin, and Zhihong Deng. 2019. Fast Structured Decoding for Sequence Models. *Advances in Neural Information Processing Systems*, 32.

Jason Taylor, Sébastien Le Maguer, and Korin Richmond. 2021. Liaison and Pronunciation Learning in End-to-End Text-to-Speech in French. In *Proceedings of the 11th ISCA Speech Synthesis Workshop (SSW11)*, pages 195–199.

Evelyne Tzoukermann. 1998. Text Analysis for the Bell Labs French Text-to-Speech System. In *Proceedings of the Fifth International Conference on Spoken Language Processing*, volume 5, pages 2039–2042.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems*, pages 5998–6008.

Sevinj Yolchuyeva, Géza Németh, and Bálint Gyires-Tóth. 2019. Transformer based Grapheme-to-Phoneme Conversion. In *Proceedings of INTERSPEECH*, page 2095–2099.

Mingzhi Yu, Hieu Duy Nguyen, Alex Sokolov, Jack Lepird, Kanthashree Mysore Sathyendra, Samridhi Choudhary, Athanasios Mouchtaris, and Siegfried Kunzmann. 2020. Multilingual Grapheme-to-Phoneme Conversion with Byte Representation. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8234–8238. IEEE.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. ByT5 Model for Massively Multilingual Grapheme-to-Phoneme Conversion. In *Proceedings of INTERSPEECH*, page 446–450.

A   Complexities of French Liaison

Historically rooted in the pronunciation of final consonants, French *liaison* has evolved into a complex linguistic feature influenced by socio-cultural factors (Adda-Decker et al., 1999). Consequently, it is categorized as obligatory, optional, or forbidden *liaison* based on the morpho-syntactic context (Durand and Lyche, 2008). For instance, while it is obligatory for the initial word *quand* with any subsequent vowel-initial word, the forbidden configuration in which *et* serves as the initial word indicates that regardless of the subsequent vowel-initial word, *liaison* will not occur. Additionally, the context of optional *liaison* may vary on subjective choices made by speakers, which depend on stylistic, socio-linguistic, and situational factors (Durand and Lyche, 2008; De Mareüil et al., 2003).

Exceptions arising from fixed expressions further compound the complexity of *liaison* (Laks, 2005; Bybee, 2005). For instance, in the fixed expression *accent aigu*, the final grapheme "t" of *accent* forms a *liaison* with the initial vowel of *aigu*:

$$accent \text{ [ak.sã] } aigu \text{ [e.gy] } \rightarrow$$
$$accent \ aigu \text{ [ak.sã.\textbf{t}e.gy]}$$

Despite the general rule that typically prevents *liaison*, such exceptions enforce the phonetic link, thus increasing the complexity of addressing *liaison*.

B   Experimental Details

B.1   Phonetic Transcription Example

The following is an example of a sentence-level phonetic transcription we collected. This example includes word-level phoneme sequences reflecting the phonological phenomena in the sentence.

- Sentence: *Un enfant innocent a oublié sa petite envelope.*

- Phonetic transcription: 9~ ˆ nA~fA~ / inOsA~ / a / ublije / sa / p@ti ˆ tA~vlOp

If no phonological phenomenon occurs between words, a slash (/) is inserted between the word-level phoneme sequences. However, if a phonological phenomenon is present, caret (ˆ) is added.

B.2   Data Preprocessing

To ensure our data retain their original meaning and naturalness, we apply only essential preprocessing steps: (i) character case folding, (ii) removing of special characters (e.g., HTML tags, links, emojis), and (iii) replacing punctuation marks with '#'. Crucially, apostrophes (') and hyphens (-) between words are preserved to avoid distorting the original meaning. Furthermore, our pronunciation datasets are composed of entries that meet the following criteria: (i) a complete sentence structure, (ii) a minimum of four words per sentence, (iii) words represented by no more than 32 graphemes or phonemes, and (iv) fewer than 192 total characters in the sentence.

B.3   Metrics

PER is the Levenshtein distance between the predicted and reference phoneme sequences, divided by the reference's length, and WER is the percentage of words with predicted phoneme sequences mismatched with the reference. For a focused evaluation of the post-lexical phonetization model's ability to address phonological phenomena, the Acc_{plp} metric is employed, representing the accuracy of phonemes at locations where phonological phenomena may occur.

B.4 Configuration Details

Hyperparameters Following Lee et al. (2023), we conducted a grid search across several hyper-parameters, exploring a variety of combinations to ensure optimal performance. Thus, based on the validation set performance, we selected the following settings: the AdamW optimizer with a learning rate of 1e-4, a dropout rate of 0.1, 100 epochs, and early stopping after 5 epochs.

Post-lexical Phonetization We set n=5 and m=5 for constructing input sequences of encoder/decoder, as preliminary experiments showed these values to be optimal for effective training and computational efficiency.

Part-of-Speech Tagging Drawing on insights from the previous studies (De Mareüil et al., 2003; Taylor et al., 2021), we adopt the open-source pre-trained French Part-of-Speech (POS) tagging model[4]. The model is fine-tuned on the Free French Treebank dataset[5] consisting of 29 POS tags (Crabbé and Candito, 2008). The French POS model extracts POS tags from the input text, which are then utilized as an auxiliary input for the post-lexical phonetization model. Its parameters are frozen during the training.

---

[4]https://huggingface.co/gilf/french-camembert-postag-model
[5]https://github.com/nicolashernandez/free-french-treebank