11th ISCA Speech Synthesis Workshop (SSW 11)
26-28 August 2021, Budapest, Hungary

Liaison and Pronunciation Learning in End-to-End Text-to-Speech in French

Jason Taylor[1], *Sébastien Le Maguer*[2], *Korin Richmond*[1]

[1] The Centre for Speech Technology Research, The University of Edinburgh.
[2]Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland

jason.taylor@ed.ac.uk, lemagues@tcd.ie, korin@cstr.com

Abstract

Sequence-to-sequence (S2S) TTS models like Tacotron have grapheme-only inputs when trained fully end-to-end. Grapheme inputs map to phone sounds depending on context, which traditionally is handled by extensive preprocessing in the TTS front-end. However, French orthography does not provide a clear one-to-one mapping between graphemes and sounds, and in English, which similarly has rather non-phonetic orthography, pronunciations are a significant cause of error in S2S-TTS with grapheme-inputs. In this paper, we test implicit pronunciation knowledge where graphemes do not map directly to phones. Implicit pronunciation knowledge learnt in S2S-TTS is similar to a standalone grapheme-to-phoneme (G2P) model, which makes explicit phone predictions at the sequential level. We find grapheme-input S2S-TTS makes implicit pronunciation errors similar to explicit G2P models - notably for foreign names. In a traditional front-end pipeline, there are also post-lexical rules which modify G2P output at the sequential level. In French, post-lexical rules require a deep knowledge of linguistic structure in a process called *Liaison*. Without explicit rules, we find S2S-TTS with grapheme-inputs over-inserts *Liaison* sounds, leading to a significant preference for a phone-based equivalent. By testing with linguistically-motivated stimuli, we observe differences that would otherwise go undetected.
Index Terms: Text-to-Speech, Phoneme, *Liaison*, *Enchaînement*

1. Introduction

Neural text encoders enable text-to-speech synthesis from raw text-audio pairs without extensive text normalisation and/or linguistic preprocessing such as lexicon and G2P model lookups. Traditionally these initial steps, formulated in the front-end, ensured correct pronunciations and provided useful information for modules further down the text-to-speech pipeline. With the rise of end-to-end (E2E) TTS with Tacotron [1] and subsequent text encoders [2, 3], the extent and need for a front-end for TTS is in question.

In [4], implicit pronunciation knowledge learned in a grapheme-based Tacotron was framed as a G2P model trained on the text from training datasets in English such as LJ [5] and VCTK [6]. Implicit G2P models were poorer than lexicon-based G2P models, being unable to pronounce place names and foreign names - especially those with non-phonetic orthography in English.

French also has non-phonetic orthography. In [7], the use of graphemes and phones were analysed as input features. The authors visualised embedded grapheme-input with t-SNE, observing single graphemes in context can map to multiple phone sounds. The authors sampled 50 sentences from the SIWI dataset in a MUSHRA comparing graph and phone input. Listeners were also asked to rate the pronunciation of the samples on a scale from 1-5 in a MOS-style test. Grapheme and phone-input performed with no-significant difference in these tests. In addition, tongue twisters were tested to measure pronunciation learning abilities, also with no significant difference found. The authors noted that both grapheme and phone-input based systems produced errors in the pronunciation of *Liaison*, but did not formally test this difficulty.

In this paper, we compare grapheme and phone input for French E2E-TTS using linguistically motivated stimuli. We first target stimuli to test the implicit G2P model and disallowed cases of *Liaison*. Under *Liaison*, phones may be inserted between word boundaries (me**s** amis, mo**n** amour). The post-lexical rules governing *Liaison* derive from linguistic information such as part-of-speech (POS) tags and semantic roles (subject, object, etc). We think phonetic control of *Liaison* is handled more reliably when using phones as a representation.

We proceed to add syllable boundaries to input to test another supra-segmental process in French known as *Enchaînement*. In French, syllables span word-boundaries so that consonants are not left at the end of syllables (eg, mon cher ami - mon. che. **r**ami). We test using stimuli containing examples of *Enchaînement*.

Overall, we find there are definite differences in pronunciations between grapheme- and phone-inputs in French, and these differences are revealed when using linguistically targeted stimuli.

2. Previous Work

2.1. Linguistic Features in Tacotron

The TTS front-end consists of a pipeline of processes to normalise input text and generate a linguistic specification for use by neural encoders, duration/prosody models, and vocoders. E2E TTS is an approach that aims to simplify the traditional modular TTS pipeline. The first Tacotron paper demonstrated high quality E2E-TTS was possible with grapheme-input, although the authors noted pronunciation errors were common and performance was enhanced with a front-end [1].

Some pronunciation issues derive from text normalisation. For instance the string '3' may be ordinal or cardinal, or abbreviations such as stock-ticker symbols can have ambiguous pronunciations. Traditionally, such errors have been averted using rule-based verbalisers. While the general performance of RNN-based verbalisers is accurate, some errors are irrecoverable and unacceptable for deployed systems [8]. RNN-based errors require an FST filter, a core problem presented in the Kaggle-hosted Text Normalisation Challenge [9], where the hosts noted the high degree of manual rule-writing for the top performing systems [10]. There is a recent drive to verbalisation that shares a unified representation across ASR and TTS [11] enabling swift rollout of FST filters to low resource languages using a template-based questionnaire [12].

10.21437/SSW.2021-34

Relatedly, pronunciation errors may also derive from a lack of deeper linguistic knowledge learned implicitly from text-audio pairs in the dataset. Increasingly, research demonstrates augmenting E2E-TTS with linguistic features improves quality in English, such as with phones [13] or with morphemes [14]. Pronunciation correction is also possible when mixing input representations between graphemes, phones and syllables [15, 16]. For non-alphabetic languages such as Japanese and Chinese, phones are preferred to characters to avoid large character sets. In these languages, the implicit pronunciation model does not learn pitch or other prosodic information meaningfully. Contextual linguistic features such as the mora [17] and pitch accents [18] are helpful, although such contextual features must be compact to be beneficial [19]. Such features were used in [20] with simplified alignments.

Recently in English the field has also used linguistic features to improve prosody: using syllabic stress [21], semantic and syntactic features [22, 23] and pre-trained language model embeddings [24, 25]. Clockwork RNNs were also used to hierarchically encode linguistic features at varying levels in [26], a hierarchical encoder having previously helped in DNN-based TTS [27, 28].

2.2. French Pronunciation

Recently, grapheme and phone inputs were tested in a French Tacotron model [7], with the authors finding no significant difference between the two inputs in a MUSHRA listening test. They chose samples from a random test set, however, which can mask subtle but important differences between systems. It was proposed in [29] for instance that listening test samples should instead be chosen containing large differences in acoustic mismatch. Tongue twisters were also tested in [7], with no significant difference found between grapheme and phone inputs. While the rapid repetition of certain articulations are difficult for humans to pronounce, we posit the grapheme-to-sound relations contained in tongue twisters are usually unambiguous and thus not an appropriate way to test implicit pronunciation learning. Instead, we target test stimuli to evaluate particular G2P and post-lexical challenges for E2E-TTS in French: G2P error words, *Liaison* and *Enchaînement*.

With grapheme-input, the text-encoder learns pronunciations implicitly while learning acoustic features. In TTS, data driven G2P models are typically trained with more than 100,000 entries from a pronunciation lexicon. While G2P conversion is regular in French, the training data is restricted in vocabulary covering fewer words than in a lexicon and G2P relations of foreign words. Figure 1 shows that the full size of the SIWI and CSS10 French datasets have limited word coverage. In [4], the authors demonstrated explicit G2P models trained on words in TTS training data underperformed G2P models trained on a full lexicon in English. They also showed G2P error words were mispronounced by grapheme-input E2E-TTS. Likewise here, we test the pronunciation of grapheme- and phone-input models using stimuli containing words with inaccurate G2P conversion.

We also test *Liaison* which is a process where linking sounds are inserted between words. Traditionally, it occurs during the "post lexical" module of a TTS front-end, after an initial phone string has been obtained from a lexicon lookup or G2P model. The plural possessive '*mes*' before a following consonant has no pronunciation corresponding to the 's' grapheme: mes chats - [me . ʃa]. But before a following vowel, the 's' grapheme corresponds to the pronunciation [z]: mes amis - [me. **z**a. mi] The rules governing *Liaison* operate at a deep linguistic level which are difficult to model. For instance, *Liaison* is disallowed after a singular noun. While data modelling of *Liaison* has been tested with decision trees [30] and templates [31], the process is complicated further because its use is often stylistic and optional [32], consequently hand-written rules are often used for TTS. [7] notes that grapheme-input Tacotron does insert *Liaison* sounds but does not learn when to use it appropriately. Their phone-input model also made *Liaison* errors, but their front-end used a low-accuracy rule-based G2P system and did not use post-lexical *Liaison* rules. We re-evaluate grapheme and phone-based *Liaison* using a test set of disallowed *Liaisons*.

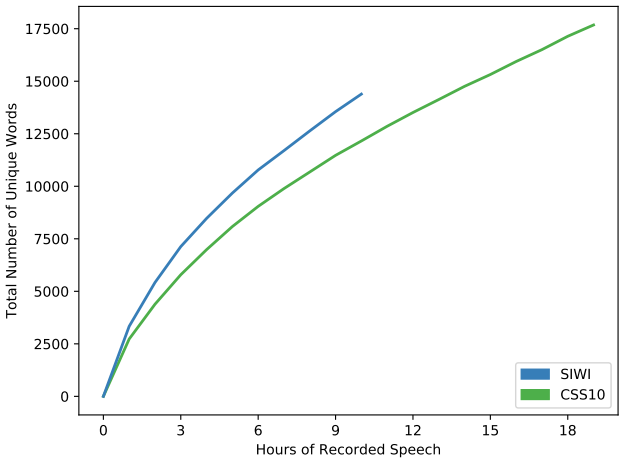

Figure 1: *Total unique words in SIWI and CSS10 French TTS Datasets. The datasets cover fewer unique words than standard pronunciation lexica which typcially contain more than 100,000 entries. Unusual G2P relations not covered in the training data may not be predicted accurately, such as in foreign names.*

Enchaînement occurs when the final sound of one word transfers to the first syllable of the next word. For instance, in *mon cher ami* the final rhotic of the word 'cher' is the onset to the syllable of the next word *ami* - [mõ . ʃɛ . **ʁ**a . mi] A multi-task G2P with syllabic boundaries included in output was shown to improve G2P performance in 14 languages [33], although French was not included in their reported results. As noted above, contextual phone information has been helpful in mora-based languages such as Japanese.

3. Methods

3.1. Tacotron Model

The Tacotron model we use for our experiments here [34], has a pre-net and CBHG module to encode a series of one-hot input characters into a single representation. Unlike previous DNN-based systems, a sequential text encoder and attention mechanism align input text to audio directly, enabling grapheme-based input. We used Location Sensitive Attention (LSA) to reduce instability in output speech as recommended in [2]. Each Tacotron was trained for 350k training steps. We use a WaveRNN vocoder based on [35], trained using Tacotron's predicted outputs up to 2000k steps, and synthesised samples in batch-mode. We used a sampling rate of 16kHz.

3.2. Data

3.3. Front-End

For our phone-based systems, we used the French front-end from MaryTTS [36], with its default lexicon and G2P model. The lexicon is based on the database Lexique [37] and each word has been phonetized as well as syllabified using *LIA PHON* [38] whose Phone-Error-Rate is 1.3%. However, in contrast to *LIA PHON*, MaryTTS doesn't provide post-lexical rule-based phonetization such as *Liaison*. Therefore, we manually wrote *Liaison* post-lexical rules based upon the guide available in [39]. Since POS tagging was a core input attribute we used the Stanford POS tagger [40] to ensure as high accuracy as possible.

3.4. Experiments

We ran AB preference tests on 10 sentences held-out from the CSS10 dataset between: i) graphemes (G) and phones (P) as input; and ii) phones (P) and phones enriched with syllable boundaries (S). The general AB tests complement the targeted AB test results.

To test the implicit knowledge of French pronunciation in the grapheme-based Tacotron, we applied the method used in [4] to test implicit pronunciation learning of grapheme-based Tacotron in English: train a G2P model using the TTS training data, identify and synthesise G2P error words with the Tacotron model. We used OpenNMT [41] for G2P modelling. We placed 10 problematic words in carrier sentences and synthesised them using the G and *Liaison* P systems.

To test *Liaison*, we hand-crafted a test set of 10 sentences, each containing disallowed *Liaisons*. As noted in [7], disallowed cases of *Liaison* are problematic for Tacotron - for example where an s is inserted before an aspirated-h as in *les_haricots*. We submitted the G and *Liaison* P systems to a forced choice test for preference.

To test *Enchaînement*, we hand-crafted a test set of 10 sentences, each containing cases where the word-final consonant becomes the onset of the following word-initial syllable. We did augment the G model here as syllable strings could only be derived from phone-based systems. We synthesised samples from the *Liaison* phone-input model (containing word boundaries) and the *Enchaînement* phone-input systems for an AB preference test.

We built the AB preference tests in Qualtrics. Due to social distancing policies, we held our listening test online using the Prolific platform. We used 30 participants. Participants were paid £5 per 30 minutes of their time. Participants were native French speakers and had no known hearing difficulties. We did not allow participants to take the test on their mobile phones - forcing them to use a desktop. For the general and targeted preference tests the accompanying question on each screen was: *Which clip has better pronunciation?/ (Quel clip a la meilleure prononciation?)*[1]

3.5. Results

3.6. CSS10 Test Stimuli

The results from the general AB listening test are shown in Figure 2. No significant differences were found between the G and P systems, nor between the P and S systems.

[1]We encourage the reader to listen to samples using this link: http://homepages.inf.ed.ac.uk/s1649890/fren/

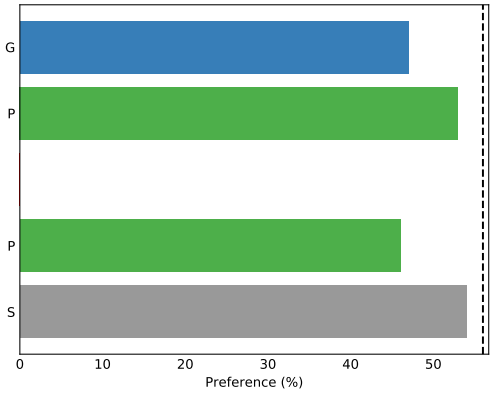

Figure 2: *Results from preference tests using CSS10 stimuli. No significant differences were observed between grapheme-input (G), phone-input (P) and phones enriched with syllable boundaries (S). The significance level at p = 0.05 is shown by the black dotted line at x=57.*

3.7. Targeted stimuli

The results from the targeted AB listening test are shown in Figure 4.

3.7.1. Words of inaccurate G2P

The phone-input models had accurate phone labels for this targeted preference test. Listeners significantly preferred the phone-based model over the grapheme-based model. Some incorrect pronunciations by system G are shown in Figure 3.

The words contain unusual G2P relations in French missing from the TTS training data. Representation mixing [16, 15] may correct pronunciations provided the reader has a large enough pronunciation lexicon to label a sufficient quantity of training data.

3.7.2. Liaison *stimuli*

Listeners significantly preferred the phone-based system. The French language has a highly active normative body called the Academy (l'Académie Française) who maintain a strict standard form of the language prohibiting insertion of *Liaison* sounds in certain contexts, such as before the aspirated h in combinations like *les **h**aricots* or *les **h**érissons*. While speakers do not strictly obey all rules, the prescribed norm of correct pronunciation remains, and incorrect *Liaison* insertion was perceived by listeners.

Word	G (Incorrect)	P (Correct)
Miguel de Cervantès	[**d**igɛl də sɛʁvã**tz**]	[migɛl də sɛʁvãtɛz]
Les Coopers	[**te sko**pə]	[le kypɛ]
Monica Lewinsky	[**pw**anika lew**ẽs**i]	[monika lywinski]
Rio de Janeiro	[**tʁ**io də ʒanero]	[ʁio də ʒanero]
McLaren	[**klaʁno**]	[məklaʁɛn]

Figure 3: *IPA transcriptions of words of inaccurate G2P included in preference test. Mispronunciation of names by the G model are highlighted in bold.*

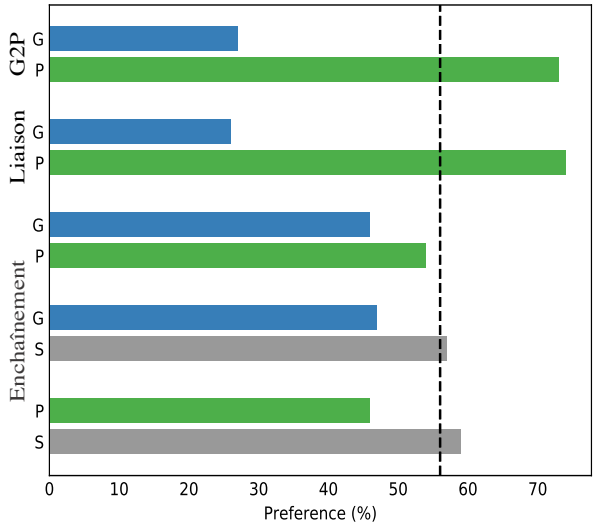

Figure 4: *Results from targeted preference Test. The first tier shows G2P results, the second tier shows* Liaison*. The last 3 tiers show results from the test with* Enchaînement *stimuli.*

3.7.3. Enchaînement *stimuli*

No significant differences were observed between G and P, but there was a preference for S over P. With syllable boundaries replacing word-boundaries, prosodic breaks occurred between syllables and less so at word boundaries.

4. Discussion

To compare grapheme and phone inputs, consider that phone inputs result from a pipeline of complex processes in the front-end. The final quality of phone labels depends on processes such as the pronunciation lexicon, the G2P model and post-lexical rules. Error propagation from these processes may contribute to phone-label inaccuracies, as was noted in [7] where *Liaison* errors were observed in the phone-based system. However, phones are preferred where graphemes do not offer the same level of control. Thus, we highlight the importance of linguistically motivated stimuli to observe the differences in pronunciation of G2P error words and *Liaison* for phones and graphemes.

5. Conclusion

We investigated pronunciation learning with a Tacotron model's text-encoder when using grapheme inputs in French. Grapheme inputs from raw or minimally normalised text reduce preprocessing required to build TTS voices. However, graphemes

Input	Labels
G	Les **h**aricots pousseront plus efficacement en plein air. Il a mi**s** **u**ne chemise.
P	[le aʁiko puzəʁõ plys efikazəmã ã plẽn ɛʁ] [il a mi yn ʃəmiz]

Figure 5: Liaison *inserts sounds at word boundaries according to complex rules, but inadequate insertion such as after aspirated-h or between a past participle and a determiner was dispreferred. Inaequate* Liaisons *are highlighted in bold.*

Input	Labels
G	Le <> ciel <> est <> bleu <> et <> la <> mer <> aussi Les <> sept <> enfants <> ont <> raconté <> une <> histoire <> amusante
P	lə <> sjɛl <> ɛ <> blø <> e <> la <> mɛʁ <> osi le <> sɛt <> ãfa˜ <> õ <> ʁakõte <> yn <> istwaʁ <> amyzãt
S	lə . sjɛ **.** **l**ɛ . blø . e . la . m**ɛ** **.** **ʁ**o . si le . s**ɛ** **.** **t**ã . fã . õ . ʁa . kõ . te . **y** **.** **n**i . stw**a** **.** **ʁ**a . my . zãt

Figure 6: *Input string differences with syllable boundaries. '<>' denote word boundaries, '.' denote syllable boundaries. The boundaries in the S system cross the word boundaries between 'ciel-est', 'mer-aussi', 'sept-enfants' and 'histoire-amusante'.*

are not accurate phonetic labels so the text encoder learns an implicit, data-driven G2P model. Previous work had found implicit G2P models to be weaker than explicit data-driven G2P models trained on pronunciation lexica. The paucity of Tacotron's implicit G2P model was observed when synthesising problematic words identified by dedicated G2P models.

We used AB preference tests to compare listener opinions on pronunciation. Using sentences from the speaker dataset we find no significant differences between grapheme or phone-input. When we use sentences containing G2P "error words" we find the grapheme-based system makes mispronunciations and the phone-based model is preferred.

Liaison is a post-lexical insertion of consonant sounds that obeys complex rules. The rules governing correct *Liaison* insertion are complex and require deep linguistic labels. Knowledge about the etymology of a word may also be required in the case of disallowed *Liaisons* before the aspirated 'h'. Whilst speakers do not always obey strict *Liaison* rules, correct *Liaisons* from a phone-based model were preferred to *Liaison* over-insertion by the grapheme-based model.

We proceeded to test whether pronunciation of enchainement was improved by substituting word boundaries for syllable boundaries. We found that in sentences with word boundaries there were pauses at word boundaries where enchainement should occur. Listeners significantly preferred syllable boundaries in these sentences.

Overall, we find linguistically-motivated stimuli reveal differences in pronunciation learning between graphemes and phones which are not revealed when considering averaged scores from a held-out sample of TTS training data.

6. Acknowledgements

This work was supported by an ESRC doctoral training grant provided via the SGSSS.

7. References

[1] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. of Interspeech*, 2017, pp. 4006–4010.

[2] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[3] W. Ping *et al.*, "Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning," in *Proc. ICLR*, 2018.

[4] J. Taylor and K. Richmond, "Analysis of Pronunciation Learning in End-to-End Speech Synthesis," in *Proc. Interspeech*, 2019, pp. 2070–2074.

[5] K. Ito, "The LJ speech dataset," 2017, available: https://keithito.com/LJ-Speech-Dataset/.

[6] C. Veaux, J. Yamagishi, and K. MacDonald, "VCTK Corpus: English Multi-speaker Corpus," 2019. [Online]. Available: https://datashare.is.ed.ac.uk/handle/10283/2651

[7] A. Perquin, E. Cooper, and J. Yamagishi, "An Investigation of the Relation Between Grapheme Embeddings and Pronunciation for Tacotron-based Systems," in *Submission to Interspeech*, 2021.

[8] H. Zhang *et al.*, "Neural models of text normalization for speech applications," *Comput. Linguist.*, vol. 45, no. 2, pp. 293–337, Jun. 2019.

[9] R. Sproat and N. Jaitly, "RNN approaches to text normalization: A challenge," 2017. [Online]. Available: https://arxiv.org/abs/1611.00068

[10] R. Sproat and K. Gorman, "A brief summary of the Kaggle text normalization challenge," in *Medium Blog Post*, 2018. [Online]. Available: https://medium.com/kaggle-blog/a-brief-summary-of-the-kaggle-text-normalization-challenge-11\\797b7e696f

[11] S. Ritchie *et al.*, "Unified Verbalization for Speech Recognition & Synthesis Across Languages," in *Proc. Interspeech*, 2019, pp. 3530–3534.

[12] S. Ritchie *et al.*, "Data driven parametric text normalization: Rapidly scaling finite-state transduction verbalizers to new languages," in *Proc SLTU and CCURL*, 2020, pp. 218–225.

[13] J. Fong *et al.*, "Investigating the Robustness of Sequence-to-Sequence Text-to-Speech Models to Imperfectly-Transcribed Training Data," in *Proc. Interspeech*, 2019, pp. 1546–1550.

[14] J. Taylor and K. Richmond, "Enhancing Sequence-to-Sequence Text-to-Speech with Morphology," in *Proc. Interspeech*, 2020, pp. 1738–1742.

[15] J. Fong, J. Taylor, and S. King, "Testing the Limits of Representation Mixing for Pronunciation Correction in End-to-End Speech Synthesis," in *Proc. Interspeech*, 2020, pp. 4019–4023.

[16] K. Kastner, J. F. Santos, Y. Bengio, and A. Courville, "Representation mixing for TTS synthesis," in *Proc. ICASSP*, 2019, pp. 5906–5910.

[17] T. Fujimoto *et al.*, "Impacts of input linguistic feature representation on Japanese end-to-end speech synthesis," in *Proc. SSW*, 2019, pp. 166–171.

[18] Y. Lu, M. Dong, and Y. Chen, "Implementing prosodic phrasing in chinese end-to-end speech synthesis," in *Proc. ICASSP*, 2019, pp. 7050–7054.

[19] Y. Yasuda, X. Wang, and J. Yamagishi, "Investigation of learning abilities on linguistic features in sequence-to-sequence text-to-speech synthesis," *Computer Speech & Language*, vol. 67, p. 101183, May 2021.

[20] M. Aso, S. Takamichi, and H. Saruwatari, "End-to-End Text-to-Speech Synthesis with Unaligned Multiple Language Units Based on Attention," in *Proc. Interspeech*, 2020, pp. 4009–4013.

[21] M. Elyasi and G. Bharaj, "Flavored tacotron: Conditional learning for prosodic-linguistic features," 2021. [Online]. Available: https://arxiv.org/abs/2104.04050

[22] S. Tyagi *et al.*, "Dynamic Prosody Generation for Speech Synthesis Using Linguistics-Driven Acoustic Embedding Selection," in *Proc. Interspeech*, 2020, pp. 4407–4411.

[23] H. Guo, F. K. Soong, L. He, and L. Xie, "Exploiting Syntactic Features in a Parsed Tree to Improve End-to-End TTS," in *Proc. Interspeech*, 2019, pp. 4460–4464.

[24] T. Kenter, M. Sharma, and R. Clark, "Improving the Prosody of RNN-Based English Text-To-Speech Synthesis by Incorporating a BERT Model," in *Proc. Interspeech 2020*, 2020, pp. 4412–4416.

[25] L. Zhao, J. Yang, and Q. Qin, "Enhancing prosodic features by adopting pre-trained language model in bahasa indonesia speech synthesis," in *Proc. ACAI*, 2020.

[26] T. Kenter *et al.*, "CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *Proc. ICML*, vol. 97, 2019, pp. 3331–3340.

[27] S. Ronanki, O. Watts, and S. King, "A hierarchical encoder-decoder model for statistical parametric speech synthesis," in *Proc. Interspeech 2017*, 2017, pp. 1133–1137.

[28] M. S. Ribeiro, O. Watts, and J. Yamagishi, "Parallel and cascaded deep neural networks for text-to-speech synthesis," in *Proc. SSW*, 2016, pp. 100–105.

[29] J. Chevelu *et al.*, "How to compare TTS systems: A new subjective evaluation methodology focused on differences," in *Proc. of Interspeech*, 2015, pp. 3481–3485.

[30] J. Pontes and S. Furui, "Predicting the phonetic realizations of word-final consonants in context – A challenge for french grapheme-to-phoneme converters," *Speech Communication*, vol. 52, no. 10, pp. 847–862, 2010.

[31] A. Greefhorst and A. Bosch, "Predicting liaison: An example-based approach," *Traitement Automatique des Langues*, vol. 57, pp. 13–32, Jan. 2016.

[32] J. Durand and C. Lyche, "French liaison in the light of corpus data," *Journal of French Language Studies*, vol. 18, no. 1, pp. 33–66, 2008.

[33] D. van Esch, M. Chua, and K. Rao, "Predicting pronunciations with syllabification and stress with recurrent neural networks," in *Proc. Interspeech*, 2016, pp. 2841–2845.

[34] Fatchord, "Tacotron implementation," 2020, available: https://github.com/fatchord/WaveRNN.

[35] N. Kalchbrenner *et al.*, "Efficient neural audio synthesis," in *Proc. ICML*, vol. 80, 2018, pp. 2410–2419.

[36] I. Steiner and S. L. Maguer, "Creating New Language and Voice Components for the Updated MaryTTS Text-to-Speech Synthesis Platform," in *Proc. LREC*, May 2018.

[37] B. New *et al.*, "Lexique 2: A new french lexical database," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 516–524, 2004.

[38] F. Béchet, "Lia phon: un systeme complet de phonétisation de textes," *Traitement automatique des langues*, vol. 42, no. 1, pp. 47–67, 2001.

[39] K. J. M., *Guide de prononciation française pour apprenants finnophones*. University of Jyväskylä, 2018. [Online]. Available: http://research.jyu.fi/phonfr/20.html

[40] K. Toutanova *et al.*, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proc. NAACL*, 2003, pp. 173–180.

[41] G. Klein *et al.*, "OpenNMT: Open-source toolkit for neural machine translation," in *Proc. ACL*, 2017, pp. 67–72.