Automatic Prediction of Prominence and Boundary Strength from Text

Pauline Mas[a,b], Kevin Vythelingum[a], Jonathan Chevelu[b], Marion Ouédraogo[a], Damien Lolive[c], Olivier Rosec[a]
[a] Voxygen, Pleumeur-Bodou, France
[b] University of Rennes, IRISA, CNRS, France
[c] University of South Brittany, IRISA, CNRS, France
firstname.lastname@voxygen.fr, firstname.lastname@irisa.fr

Abstract

In Text-to-Speech synthesis (TTS), the prediction of prosodic information from text is a difficult challenge, since it requires information related to the context that may not be present in the text. Previous studies have shown that prosodic annotations from an oracle benefit TTS models and improve their prosodic rendering as well as their controllability. In this paper, we investigate different strategies to automatically predict prominence and boundary strength from text. We compare three prediction strategies on a French audiobook dataset: dedicated predictors jointly trained in a TTS model, a BERT-informed Prosody Predictor (BIPP) and its auto-regressive counterpart, both benefiting from semantic text embeddings. BIPP exhibits the best performance in our experiments, indicating that using phonetized syllables as complementary information to the semantic embedding provided by a BERT-like model is the best strategy to predict prosodic events.

Keywords: prosody, emphasis, TTS

1. Introduction

Often summarized as "all [suprasegmental] aspects of speech unrelated to the identification of segments" (Vaissière, 2020), prosody encompasses a variety of acoustic phenomena. Its manifestation in the signal may be observed in pauses, variations of the fundamental frequency (f0), voice quality, controlled variations of sound duration or of its intensity. Prosody contributes at different levels and in different forms to the transmission of intentions as well as linguistic content.

As one of the challenges faced by text-to-speech (TTS) systems when generating audio from a text, the prediction of prosodic aspects of speech must rely on various approximations and indirect hints obtained from text (Chien and Lee, 2021). Prosodic contour, for instance, is conditioned on the type of sentences and related to the grammatical structure, or choice of words, to reflect a particular intention. Prosody or style transfer prove to be all the more challenging, as it ideally requires a good disentanglement of speaker-dependent variations from prosody-related modulations. As observed by Sigurgeirsson and King, 2023, prosody transfer is usually handled with a dedicated prosody encoder incorporated in (and jointly learned with) a TTS model. The resulting representation has a strong impact on the generation but is hardly interpretable and may also model undesired speaker-related characteristics.

Various propositions have been made in previous works to characterize prosody in a humanly understandable manner, either to categorize prosodic events (Silverman et al., 1992; Grabe et al., 1998) or to characterize the prosodic structure (Dai et al., 2022). The TTS framework of Zou et al., 2021 relies on a front-end model that produces ToBI annotations for the input text. Its predictions are given to a Tacotron model and guides the prosody of the synthesized speech. Their experiments demonstrate that prosodic annotations can be learned from text and give control over prosodic features in synthetic speech generation. However, the ToBI annotation of training data is a non-trivial constraint in the implementation of their method, since it requires expert knowledge.

The wavelet prosody toolkit of Suni et al., 2017 is an unsupervised method manipulating f0, energy and duration features from `wav` audio files to automatically assign to each audio unit (*e.g.* word, syllable) a pair of labels designed as a salience and structural boundary annotation. The positive impact of these events on both the control and reproduction of prosodic patterns in TTS is demonstrated in a separate study (Suni et al., 2020), where the labels computed by their toolkit on audio references are introduced as an oracle in a tacotron model. To the best of our knowledge, no alternative has been developed so far to predict prosodic salience and boundary annotation (refered to as prominence and prosodic boundary strength) exclusively from the input text.

In this paper, we investigate different strategies to automatically predict prosodic events from text in French. More specifically, we seek to under-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5588–5596
11–16 May 2026.

stand how a model benefiting from semantic embeddings performs compared to predictors jointly trained with a TTS model on audio-text pairs. Our contributions can be summarized as follows: 1) two BERT-Informed Prosody Predictor (BIPP) models trained to predict continuous prominence and prosodic boundary strength from textual content, 2) the use of a Masked Language Model (MLM) to enhance these models with enriched textual representations, 3) a modified FastSpeech2 model capable of predicting or accepting prosodic event information, 4) an evaluation of the prediction methods in a multi-speaker scenario, and 5) an analysis of the performance focused on salient prosodic events.

The paper is organized as follows. We begin with a definition of the prosodic events and their origin (Section 2), before describing the systems and models tested in Section 3. Section 4 details the data used in our experiments. We report on the overall results of our experiments in 5.1 before investigating how well salient events are reproduced 5.2 and proposing our analysis in Section 5.3. We finally discuss our work and conclude in Section 6.

2. Prosodic Events

As previously stated, the wavelet prosody toolkit relies on three audio features: F0, energy and duration. They can be either provided or computed at a frame level in a preliminary step of the procedure. These features are normalized and summed to obtain a single signal that is then decomposed by the CWT defined as follows:

$$W_s(\sigma, \tau) = \sigma^{-1/2} s * \psi_{\tau,\sigma} \quad (1)$$

where s is a one-dimensional signal, σ is a positive scale and τ a temporal translation. $*$ symbolizes a convolution while $\psi_{\tau,\sigma}$ is the Ricker mother wavelet "translated by τ and dilated by σ " (Suni et al., 2017).

The CWT provides an analysis on various scales, from which lines of **maximum amplitude (LoMA)** and lines of **minimum amplitude (LomA)** are created. These lines correspond respectively to the continuous unit level prominence strength and prosodic boundary strength, hereafter called "prominence" and "boundaries". The former quantifies the relative salience of speech units within the signal (lexical accent, local focus...), while the latter represents the different levels of border in the hierarchical structure of speech (Suni et al., 2017). Both are discretized into categories ranging from 0 to 2 at the final stage of the process, using arbitrary intervals.

We choose to manipulate both prosodic descriptors in their continuous form instead of relying on the toolkit's labeling. The resulting representation is expected to be closer to the corresponding speech signal than discrete labels. We also compute prominence and boundary on a syllable level instead of on a word level.

3. System and Model Description

Unlike Suni et al., 2017 and Zou et al., 2021, we choose to develop a baseline from FastSpeech2 (Ren et al., 2021), as its architecture includes a variance predictor tasked with predicting the same signal properties that determine prominence and boundary strength indices. We adapt this architecture by incorporating a dedicated predictor for each prosodic event (prominences and boundaries) upstream of any other predictors, and thus obtain our baseline model. To investigate whether semantic embeddings provide useful information for this task, we also develop an independent predictor. This second model follows an encoder-decoder transformer architecture (Vaswani et al., 2017), in which a pre-trained MLM serves as encoder. We opt for a pre-trained camemBERT model dedicated to the French language (Martin et al., 2020) to fulfill this role.

3.1. Prosody Prediction Model

The independent predictor, named **BIPP** (Bert-Informed Prosody Predictor), is trained to generate a pair of prominence and boundary values **for each syllable of the input sequence**. As mentioned above, it is composed of a pre-trained text encoder, a syllable convolutional encoder, as well as a transformer decoder and two parallel prediction modules.

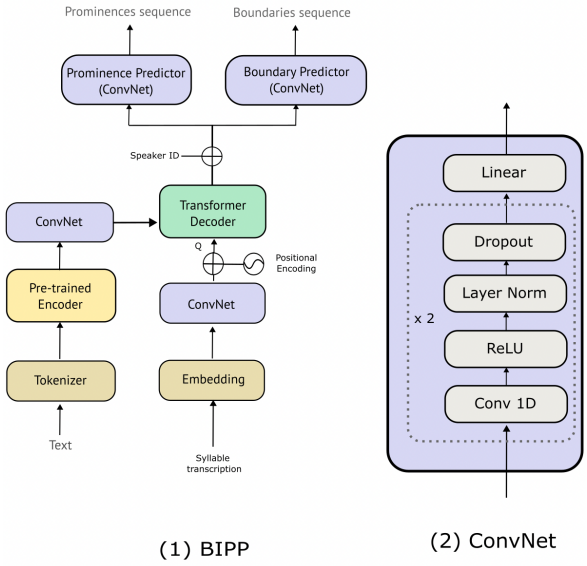


Figure 1: Overview of the Bert-Informed Prosody Predictor's architecture (BIPP)

Several studies have been devoted to analyzing the knowledge acquired by MLM models of the BERT family (Devlin et al., 2019), and the encoding of this knowledge through successive attention layers (Jawahar et al., 2019; Mohebbi et al., 2021; Lin et al., 2019; De Vries et al., 2020). Following the premise that syntax and semantics are determinants of prosody, Kakouros and O'Mahony, 2023 set out to determine whether BERT encodes textual information relevant for prominence labelling. Their results suggest that BERT does contribute to the task, relying on structural information widespread through its layers.

Assuming, in turn, that MLM representations hold relevant information to predict prosodic events, and that French MLM models possess similar knowledge to their English counterpart, we integrate a **frozen camemBERT** (Martin et al., 2020) as the text encoder in the predictor. The final embedding produced by the last layer of camemBERT is passed to the decoder after going through two convolution layers, as illustrated in Figure 1. According to our preliminary experiments, the last layer embedding leads to better performances than a max-pooling of the outputs of all hidden layers for this task.

The phonetized syllable sequence goes through an embedding layer before being encoded by a convolution network (consisting of 2 convolution layers with a final layer output dimension of 256) to match camemBERT's embedding latent space. The transformer decoder (Shazeer, 2019) takes this encoded output added to a positional encoding. It serves as a query, while the key and value are derived from camemBERT's embeddings. The cross-attention mechanism involved at this stage conveniently spares us the trouble of explicitly aligning our syllable with camemBERT's tokens (Kudo and Richardson, 2018) while selecting relevant information for the last modules. We fix the number of attention heads to 2. Prominence and boundary sequences are then predicted separately by parallel predictors as described in Figure 1 (2). Both predictors receive the output of the decoder added to a speaker embedding, and produce sequences of continuous values at syllable level.

To verify whether the syllable sequences help in orienting the attention on the relevant aspects of the text's representation, we also develop an auto-regressive variant of BIPP alongside the "standard" BIPP model. In this auto-regressive variant, during the training phase, we project the prosodic target in camemBERT's embedding latent space with a convolutional encoder. These embeddings are gradually shown to the decoder and replace the syllable embedding as query. At inference time, the prosodic values are predicted recursively from past estimations.

3.2. FastSpeech2 Adaptation

FastSpeech2 is a non-autoregressive model that predicts audio features from phonetized sequences (Ren et al., 2021). Its main components include a transformer encoder, a transformer decoder and a variational adaptor. This last component produces intermediary predictions of duration, pitch and energy at a phoneme level that are added to the decoder's input. While these three features need to be extracted from the ground truth signal during training, they are predicted by the model from the encoded textual input at inference time.

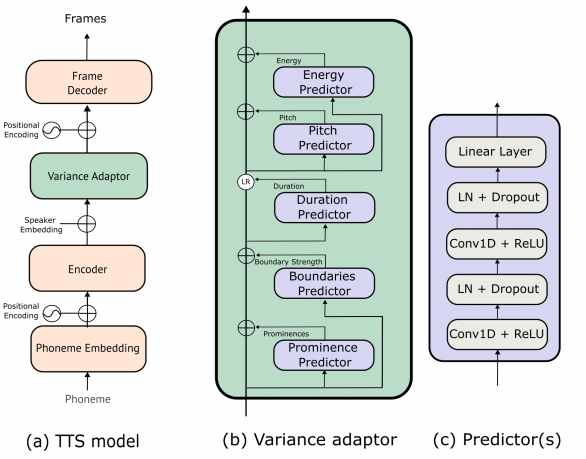


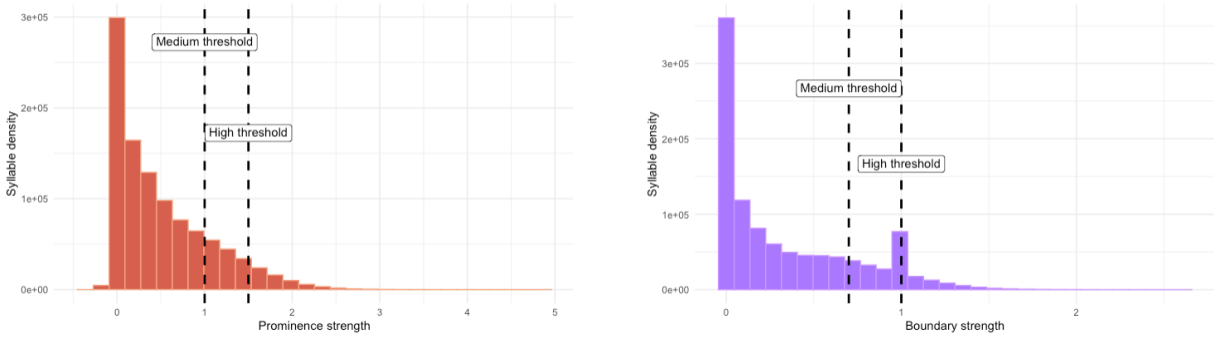
Figure 2: Overview of the adapted architecture of FastSpeech2

In order to integrate the prominence and boundary information in FastSpeech's architecture, we suggest incorporating the prominence and boundary values in parallel and upstream of the duration predictor. Both predictors are designed in the same way as the pre-existing feature predictors, as shown in Figure 2. The underlying idea behind this integration choice is that the prominence and boundary features guide the following duration, pitch and energy predictors.

4. Dataset

All models are trained and tested on a subset of a multi-speaker corpus of open-source French audiobooks called **Mufasa** (Sini et al., 2022). We selected 9 speakers (5 women and 4 men) based on the quality of their recording as reported in the corpus documentation. As shown in Table 4, the dataset is unbalanced in terms of sample distribution. Almost a third of the audios are attributed to two female speakers, whereas Alain (male) appears much less frequently than the others.

The phonetized syllable sequences were generated along with segment duration and fundamental frequency values using an in-house toolkit. We filter



(a) Distribution of prominence values

(b) Distribution of boundary strength values

Figure 3: Distribution of prominence (a) and boundary strength (b) values in our dataset, breaks included. Black dotted lines indicate the thresholds used to identify salient events (see section 5.2). They respectively correspond to threshold values set at 80% and 90% of the cumulative density.

Speaker	Gender	Samples	All data points[a]	Syllabes
Alain	Male	2 600	52 993	45 653
Cocotte	Female	5 626	132 876	115 850
Damien	Male	6 099	124 647	108 299
Daniel	Male	4 740	78 991	66 503
Ezwa	Female	10 093	168 145	142 907
Nadine	Female	9 787	170 400	144 225
Menager	Male	4 840	107 427	92 014
Orangeno	Female	6 057	98 797	84 398
Victoria	Female	7 400	155 283	129 825
Total	–	57 242	1 089 559	929 674

[a]Counting one point of data per silence segment and syllable

Table 1: Corpus description after filtering on the transcription quality

out sample pairs whose text does not correspond to Whisper's transcription (Radford et al., 2022), as well as the pairs whose phonetized sequence mismatches the phonetized speech recognition output. The prosodic targets are finally computed on the remaining audio samples with the help of the wavelet prosody toolkit. Overall, 57k samples are left at our disposal, corresponding to approximately 56 hours of speech. We assume all sample pairs to be unique, as we expect each person to produce a distinctive prosodic pattern in the event of two speakers pronouncing the same text. Table 4 and Figure 3 describe the resulting filtered dataset in terms of volume and distribution. We randomly select 10k samples ($\sim$ 11 hours) that we split evenly between the development subset and the test subset, leaving about 47k samples ($\sim$ 45 hours) in the training dataset.

5. Experiments

The FastSpeech model is trained for 500k steps on our *train* set from the dataset described above, while the BIPP models require about 60k steps with a dual MSE loss (i.e. a combination of MSE losses computed separately for each predictor). Preliminary experiments on the hyper-parameters of BIPP's training allowed us to select the best model based on the results on our *development* set. It is used to generate prominence and boundary values over all samples of the *test* set.

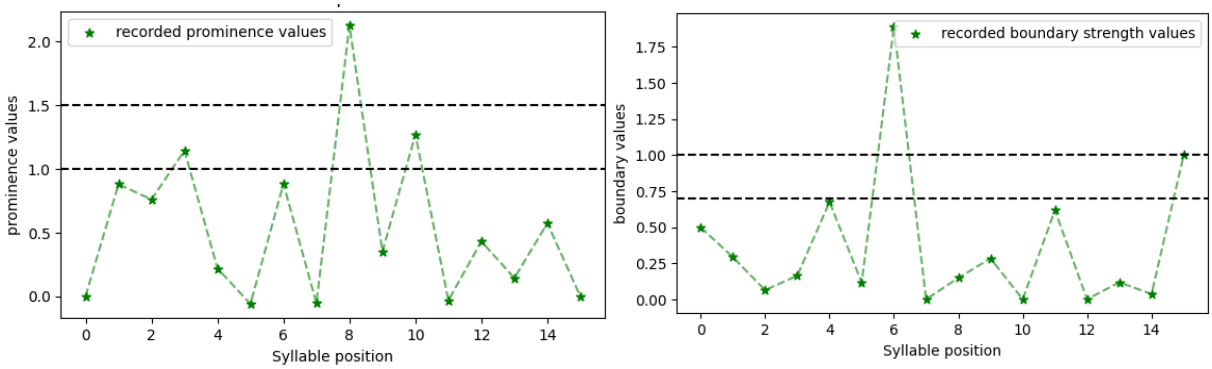
We then set out to determine the extend to which prosodic events are predictable from the text, and whether camemBERT's representations are relevant for the prediction of prosodic events. To do so, we set up an evaluation in two phases. We first evaluate the global modeling of prosodic events by comparing the full predictions of the three models with the reference values computed from signal samples. In the second strategy, we only consider the stronger prominence and boundary values to see how important prosodic events are predicted.

As illustrated by the curves of Figure 4, high peaks are rare in both prominence and boundary strength sequences. However, these salient prosodic events bear important information, and may signal primary emphasis or a strong delimitation in the grammatical structure of the speech sample. These relative "maxima" are all the more relevant for the rendering of a speaker's prosody. Therefore, we further focus our analysis on these salient events and try to determine if they are correctly reproduced in our predictions.

5.1. Global Evaluations of Prosodic Events

Our global evaluation of the modeling of prosodic events is centered on two metrics: the **Mean Square Error (MSE)** that severely penalizes values as they stray from their corresponding reference point, and the **Mean Directional Accuracy (MDA)** to quantify the changes in curve tendencies. Results of this evaluation are reported in Table 2.

Table 2a presents the performances on prominence prediction, along with the confidence intervals. We observe that BIPP outperforms the



(a) Example of prominence sequence

(b) Example of boundary strength sequence

Figure 4: Example of syllable-level prominence **(a)** and boundary strength **(b)** values computed from the reference audio of the segment "*ses plaisanteries grossières, sa vulgaire obscénité,*". The dotted lines correspond to the thresholds used to define "peaks" in section 5.2, with the lowest being the "medium" threshold and the highest the "high" threshold. Data point touching the line *are* included in the peak category.

Predictors	MSE ↓	MDA ↑
FastSpeech	0.704 ±0.012	0.348 ±0.002
BIPP	**0.491 ±0.008**	**0.802 ±0.003**
Auto-reg. BIPP	0.985 ±0.011	0.534 ±0.004

(a) Prominences

Predictors	MSE ↓	MDA ↑
FastSpeech	1.079 ±0.010	0.293 ±0.002
BIPP	**0.601 ±0.010**	**0.680 ±0.004**
Auto-reg. BIPP	0.839 ±0.010	0.538 ±0.003

(b) Boundaries

Table 2: Overall performance metrics (with 95% confidence intervals) of prominence and boundary strength predictions. Bold values correspond to the best result for their corresponding metric.

other models on both metrics, whereas its auto-regressive version ranks below FastSpeech's predictors in terms of MSE. The auto-regressive model may somehow follow a better curve tendency than FastSpeech, but fails at predicting the amplitude of variation between values.

As for boundary values, Table 2b shows that both BIPP models outperform the predictor of FastSpeech regardless of the metrics, with the syllable-informed version performing better than the auto-regressive one. The BIPP predictors appear to globally benefit from camemBERT's representation when predicting this type of events. Querying the semantic embedding using syllables instead of past boundaries values visibly further helps at modeling the sample's structure.

The Figure 5 gives an example of prominence and boundary predictions for the same segment as Figure 4. We see how poorly the predictions of the auto-regressive BIPP fit the expected curve of data in this specific case, setting it far from the other two models. By contrast BIPP achieves to locate the boundary and prominence patterns here, although it fails to reach the highest values. Furthermore, FastSpeech seems to struggle with the end of the boundary sequence.

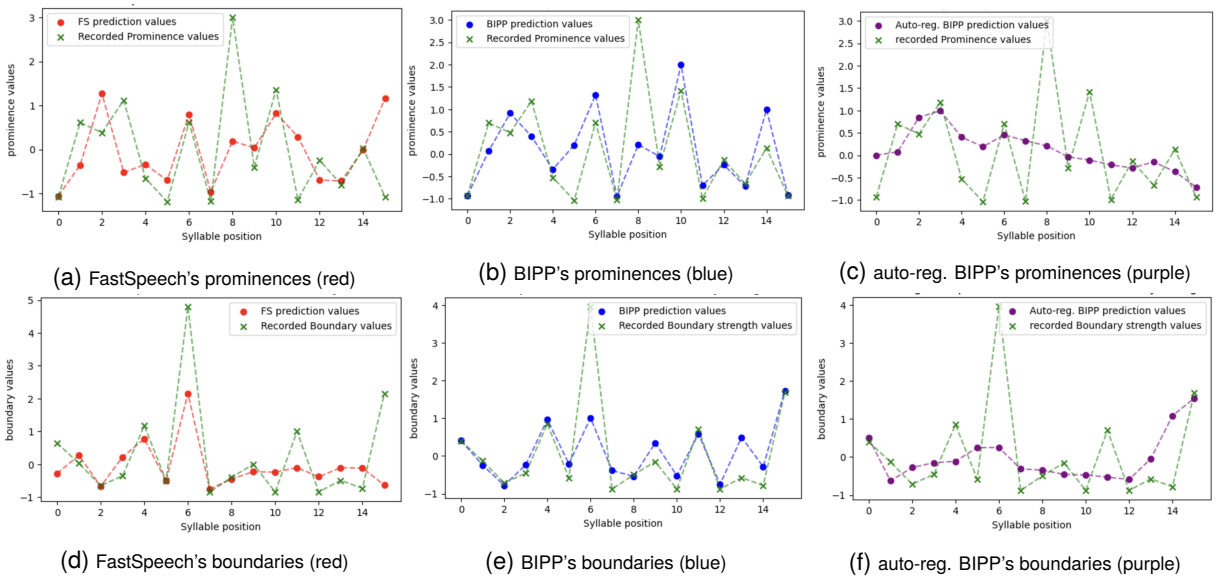
5.2. Focus on Salient Events

Such observations motivated further investigations to determine how the predictors perform on particularly salient events. To evaluate this more objectively, we convert our reference and predicted sequences into binary values where null values mark the absence of a peak for the corresponding syllables, and "1" signals the presence of a peak exceeding a given threshold. The **accuracy, precision and recall** are computed on the resulting binary sequences, along with their confidence intervals.

Threshold level	Percentile	Prominence	Boundary
Medium	80%	1.0	0.7
High	90%	1.5	1.0

Table 3: Threshold values defined to filter "medium" and "high" peaks of prosodic events

We apply this binary transformation twice: once using an "upper" threshold above which we consider the values to be particularly "high", and a second time to include points of data down to a "medium" threshold. We thus obtain a first classification where only very high values are associated



(a) FastSpeech's prominences (red) (b) BIPP's prominences (blue) (c) auto-reg. BIPP's prominences (purple)

(d) FastSpeech's boundaries (red) (e) BIPP's boundaries (blue) (f) auto-reg. BIPP's boundaries (purple)

Figure 5: Examples of prominence and boundary strength curves for the text segment "*ses plaisanteries grossières , sa vulgaire obscénité ,*" :
- Comparison of FastSpeech's estimation (red curve) with the reference computed from the audio (green curve) for prominences **(a)** and boundaries **(d)**;
- Comparison between BIPP's estimation (blue curve) and the reference computed from the audio (green) from prominence **(b)** and **(e)**.
- Comparison between the **autoregressive** BIPP's estimation (blue curve) and the reference computed from the audio (green) from prominence **(c)** and **(f)**.
All data points of each graph are centered and standardized according to the data preprocessing procedure of the corresponding model. Note that silent segments associated to punctuation symbols appear with their own data points.

to the "peak" class, and a second version of this classification where this same class encompasses a larger group of values. In both cases, we face an unbalanced classification problem, where null values outnumber the "peak" category instances by far. Table 3 details the threshold values for each prosodic event type, chosen to respectively exclude 90% and 80% of the distributions given in Figure 3. We then compute the classification metrics to see whether all events were predicted at the right timing. Additionally, we introduce a **recall MSE** to measure how closely the amplitude of **expected events** is reproduced in the predictions. In other words, this last metrics quantifies the error of amplitude between the expected value of prominence (or boundary) and the corresponding predicted value for all reference data point exceeding the threshold. Results for all these metrics are reported in Table 4.

5.2.1. Major Prosody Events

Table 4a and 4b summarize the results obtained when conditioning the definition of the "peak" category on the highest thresholds. We first notice that the accuracy exceeds 90% for prominences and boundaries across all our models, which sets it apart from the other metrics. This seemingly high performance may be explained by the overwhelming proportion of "non-peak" prominence and boundary values "correctly" predicted as such. Most data samples at our disposal include two to three values exceeding the "high" threshold at most, making the "peak" category quite rare compared to its counterpart.

The precision and recall, however, provide us with better insights of how the three models stand on high prominence and boundary values. Both precision and recall are globally low, although BIPP does reach a precision of 0.6 and a recall of 0.4 for boundary peaks (far from the other two models lying under 0.2 on both metrics). Another global tendency puts the auto-regressive BIPP in last position, as it is not able to match either FastSpeech or BIPP, with the exception of the boundary recall MSE, for which it performs similarly to FastSpeech's predictor. As for prominence peaks, FastSpeech outperforms BIPP, even though the difference is tighter when comes precision. BIPP does not seem to improve the prediction of very salient prominence events compared to the FastSpeech model, especially in terms of position in the sequence.

Predictor	Recall MSE ↓	Accuracy ↑	Precision ↑	Recall ↑
FastSpeech	0.210 ±0.008	91.2 ±0.2	**23.3 ±1.1**	**16.8 ±0.9**
BIPP	**0.163 ±0.006**	**93.7 ±0.2**	20.0 ±1.1	14.6 ±0.9
Auto-reg. BIPP	0.301 ±0.009	93.1 ±0.2	8.3 ±0.8	5.9 ±0.6

(a) Prominences above "higher" threshold.

Predictor	Recall MSE ↓	Accuracy ↑	Precision ↑	Recall ↑
FastSpeech	0.334 ±0.013	92.2 ±0.2	18.0 ±1.0	11.0 ±0.7
BIPP	**0.207 ±0.007**	**92.7 ±0.2**	**60.5 ±1.3**	**41.9 ±1.1**
Auto-reg. BIPP	0.328 ±0.008	89.6 ±0.2	11.8 ±0.9	9.3 ±0.7

(b) Boundaries strength above "higher" threshold.

Predictor	Recall MSE ↓	Accuracy ↑	Precision ↑	Recall ↑
FastSpeech	0.352 ±0.009	80.8 ±0.3	56.6 ±0.9	**51.3 ±0.9**
BIPP	**0.267 ±0.007**	**85.1 ±0.2**	**58.6 ±01.1**	43.3 ±0.9
Auto-reg. BIPP	0.481 ±0.009	81.4 ±0.2	29.8 ±01.2	13.4 ±0.6

(c) Prominences above "medium" threshold

Predictor	Recall MSE ↓	Accuracy ↑	Precision ↑	Recall ↑
FastSpeech	0.538 ±0.014	83.2 ±0.3	53.7 ±0.1	37.9 ±0.8
BIPP	**0.325 ±0.008**	**86.8 ±0.2**	**81.1 ±0.7**	**47.2 ±0.7**
Auto-reg.BIPP	0.494 ±0.008	81.6 ±0.2	44.0 ±1.2	22.3 ±0.7

(d) Boundaries strength above "medium" threshold

Table 4: Performance metrics (with 95% confidence intervals) of prominence and boundary strength values exceeding the defined threshold (either "high" or "medium"). The accuracy, precision and recall are expressed in percentage (%). Bold values correspond to the best result for their corresponding metric .

5.2.2. Intermediary and Major Events

With a looser definition of "peak" that includes values down to our "medium" thresholds (about 20% of all data points), the accuracy globally drops to about 80% while the precision and recall increase. As it happens, the accuracy is still strongly influenced by the majority of "non-peak" values correctly categorized as such, although the three models seem to better categorize the enlarged "peak" category.

Looking into the details of table 4c, BIPP outperforms the other models in accuracy, precision and recall MSE but stands behind FastSpeech on the recall metric. In other words, FastSpeech predicts more peaks given this definition of the category, but sometimes place them in positions where the reference value remain under the threshold. The auto-regressive version is globally the worst of the three models, although its accuracy overlaps that of FastSpeech's predictor.

When comes boundary peaks, BIPP is again the best model out of the three models for all four metrics. It even reaches a precision of 0.81, thus being correct most of the time when placing a salient boundary value. However, it does not recall half of the peaks expected in the reference sequence. This time, the auto-regressive BIPP also performs better than FastSpeech in recall MSE, although it still stands behind on the classification criteria. It seems to generally predict closer values to the expected peaks but tries less often to actually predict a salient event.

In general, BIPP seems to better predict salient prosodic events than FastSpeech's predictors under this definition of the category.

5.3. Analysis

Overall, BIPP has a more conservative behavior than FastSpeech, which predicts prominence and boundary peaks more frequently, but not always correctly. BIPP is less prone to predicting peak values, but is closer to the reference when placing one. It is generally better than FastSpeech at predicting boundary strength, which could be explained by the connection between these events and the syntactic structure of the samples, supposedly captured in CamemBERT's embeddings. On the other hand, it struggles somewhat more with the prediction of prominences. These events, being closer to higher levels of linguistic analysis (such as semantic or pragmatic) and to the speaker's intention, are challenging. CamemBERT's embeddings, although helpful, may be insufficient to fully capture them. The poor performance of the auto-regressive BIPP show that considering semantic embeddings only leads to worse predictions. The type of query matters: in our case, conditioning the attention on the semantic embeddings using syllable sequences is better than using the history of previous event predictions. This is no surprise if we consider that the syllabified sequence contains more structural information on the text sample than the few past estimated prosodic events.

The performance on salient prominence and boundary values (peaks) drops for all three models tested. Given that they may be less related to the structure of the textual sample and more to the speaker's intentions while reading it, these salient events may be less predictable and thus harder to model.

6. Conclusion

In this work, we aimed at predicting prosodic events automatically, and at determining the extent to which this task can be solved using solely textual information. We chose the prominence and boundary strength of Suni et al., 2017 as a quantification for prosody, a dual representation that is both continuous and intuitive, making it possible to annotate datasets with minimal human intervention. Three models were developed and compared: a modified FastSpeech2 along with two versions of an inde-

pendent predictor, BIPP and its auto-regressive variant.

BIPP outperforms not only its auto-regressive version but also the predictors jointly trained with FastSpeech for the global prediction of prominence and boundary sequences. Even though BIPP fails to capture very salient prosodic events, our experiments show that predicting prosodic events solely from text is feasible, and that semantic representations queried by syllable information outperforms the other two methods tested. For instance, in the context of homogeneous speaking styles without "extremely" expressive patterns, the BIPP model seems to be a good candidate to locate moderate prosodic events. It may increase the variability in prosodic realizations compared to a standard TTS model such as FastSpeech, thus alleviating the perceived monotony of TTS systems also known as the TTS fatigue effect.

Future directions for this work include integrating BIPP into a TTS framework to test i) its impact on synthesized audio and ii) its performance in a style transfer setting. We can also consider its usage for manual prosodic control using a GUI, as the scale and principle behind prominence and boundaries are rather intuitive.

7. Acknowledgements

We would like to thank Laure Charonnat (Voxygen) and Philippe Martin (IRISA) for their technical support and helpful advices

8. Bibliographical References

Chung-Ming Chien and Hung-yi Lee. 2021. Hierarchical prosody modeling for non-autoregressive speech synthesis. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 446–453.

Ziqian Dai, Jianwei Yu, Yan Wang, Nuo Chen, Yanyao Bian, Guangzhi Li, Deng Cai, and Dong Yu. 2022. Automatic Prosody Annotation with Pre-Trained Text-Speech Model. In *Interspeech 2022*, pages 5513–5517.

Wietse De Vries, Andreas Van Cranenburgh, and Malvina Nissim. 2020. What's so special about BERT's layers? A closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esther Grabe, Francis Nolan, and Kimberley J. Farrar. 1998. Ivie - a comparative transcription system for intonational variation in english. In *5th International Conference on Spoken Language Processing (ICSLP 1998)*, page paper 0099.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Sofoklis Kakouros and Johannah O'Mahony. 2023. What does BERT learn about prosody?

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: Getting inside BERT's Linguistic Knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilehvar. 2021. Exploring the Role of BERT Token Representations to Explain Sentence Probing Results. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 792–806, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need.

Atli Thor Sigurgeirsson and Simon King. 2023. Do prosody transfer models transfer prosody? In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Kim Silverman, Mary Beckman, John Pitrelli, Mori Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. Tobi: a standard for labeling english prosody. In *2nd International Conference on Spoken Language Processing (ICSLP 1992)*, pages 867–870.

Antti Suni, Sofoklis Kakouros, Martti Vainio, and Juraj Šimko. 2020. Prosodic Prominence and Boundaries in Sequence-to-Sequence Speech Synthesis. In *Speech Prosody 2020*, pages 940–944, ISCA. ISCA.

Antti Suni, Juraj Šimko, Daniel Aalto, and Martti Vainio. 2017. Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, 45:123–136.

Jacqueline Vaissière. 2020. Chapitre IX - Prosodie. In *La Phonétique*, volume 4e éd. of *Que sais-je ?*, pages 96–117. Presses Universitaires de France, Paris cedex 14.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Yuxiang Zou, Shichao Liu, Xiang Yin, Haopeng Lin, Chunfeng Wang, Haoyu Zhang, and Zejun Ma. 2021. Fine-grained prosody modeling in neural speech synthesis using tobi representation. In *Interspeech 2021*, pages 3146–3150.

9. Language Resource References

Sini, Aghilas and Wadoux, Lily and Perquin, Antoine and Vidal, Gaëlle and Guennec, David and Lolive, Damien and Alain, Pierre and Barbot, Nelly and Chevelu, Jonathan and Delhay, Arnaud. 2022. *Techniques de synthèse vocale neuronale à l'épreuve des données d'apprentissage non dédiées : les livres audio amateurs en français*. ATALA.