ÉCOLE POLYTECHNIQUE

PERSONAL RESEARCH PROJECT (PRL) REPORT

Improving French Synthetic Speech Quality

via SSML Prosody Control

Author: Emeric PAYER

Supervisors: Nassima OULD-OUALI and Eric MOULINES

Laboratory: CMAP - École Polytechnique

Academic Year 2024–2025

Contents

Abstract

Neural text-to-speech (TTS) systems can produce intelligible speech, but often sound monotonous or unnatural due to inadequate prosody and phrasing. This project aims to enhance the naturalness of French TTS by predicting prosodic features (which include pitch, rate and volume) and phrase breaks from input text and encoding them in Speech Synthesis Markup Language (SSML). We develop a pipeline that generates enriched SSML from raw text, allowing a TTS engine to render more natural-sounding speech. Our approach involves forced alignment of French audio-text data to learn prosodic patterns, fine-tuning transformer-based language models (CamemBERT) for pause insertion and prosodic contour prediction, and injecting these predictions as SSML tags (e.g. `<break>` and `<prosody>`). Preliminary results indicate that the prosody-enhanced SSML yields more dynamic and natural audio output, with improved listener mean-opinion scores (MOS) in small-scale tests. These findings suggest that incorporating explicit pause and intonation prediction can significantly improve the expressiveness of French TTS systems, with implications for more engaging audiobook narration, voice assistants, and other speech applications.

1 Introduction

Recent advances in neural Text-to-Speech (TTS) systems have significantly improved synthetic speech quality. However, generated speech often remains monotonous, lacking the rich *prosody* – variations in pitch, rate, volume, and well-timed pauses – that characterize natural human speech. Prosody enhances emotional expression, structural clarity, and listener engagement. Audiobook narrators, for instance, skillfully use prosody and strategic pauses to sustain listener interest [1], whereas baseline TTS outputs frequently sound robotic and flat, especially noticeable in French TTS systems.

Enhancing prosody and phrase breaks significantly boosts perceived naturalness. Pethe *et al.* [1] demonstrated that predicted prosodic features from text closely match human narration, substantially increasing listener preference. Similarly, Vadapalli *et al.* [2] confirmed that explicitly predicted phrase breaks in a synthetic voice largely improve speech comprehension and naturalness. These findings underline the importance of modeling pauses and vocal dynamics to achieve human-like speech.

This project is part of a broader initiative aimed at comparing various approaches to generating natural-sounding speech. In this project, we address French TTS naturalness by automatically predicting phrase breaks and prosodic features from text. We adopt a two-step approach:

1. Employ language models to analyze input text and integrate prosodic features (pitch, rate, volume, pauses) to generate a code in *Speech Synthesis Markup Language* (SSML).

2. Synthesize speech from this enriched SSML via standard TTS engines.

SSML is a widely-supported XML-based markup language that enables control of prosody in TTS services like Azure Cognitive Services, allowing us to use existing high-quality voices.

Our key objectives are:

- Develop a pipeline for predicting pauses and prosodic adjustments in French texts.
- Evaluate forced alignment tools to accurately align French speech data and extract prosody.
- Fine-tune CamemBERT, a pre-trained French language model, for pause classification and prosody regression.
- Generate prosody-enhanced SSML from model predictions and assess its impact on speech naturalness using objective and subjective evaluations.

This report describes our full pipeline – from raw audio and text inputs through alignment, model training, SSML generation, to final TTS synthesis. Our experiments show meaningful prosodic pattern predictions and improved speech dynamics. We also evaluate the strengths and weaknesses of various alignment and modeling methods. The report structure is as follows: Section 2 summarizes the relevant literature for this project; Section 3 explains briefly TTS, prosody, SSML, and relevant research; Section 4 details our methods including data processing and models; Section 5 shows our alignment accuracy, model performance, and subjective listening test outcomes; Section 6 discusses these results; and Section 7 summarizes findings and suggests future work.

2 Literature Review

This section positions our work against recent advances in (1) forced text-speech alignment, (2) phrase-break prediction, and (3) prosody and expressiveness modelling. The emphasis is on papers published 2020-2025 that directly inform our design choices.

2.1 Forced Text Alignment

Early pipelines relied on HMM-GMM alignment; the de-facto open-source tool is **Montreal Forced Aligner** (MFA) [3], which outputs word and phoneme boundaries given a pronunciation lexicon. More recently, neural aligners have displaced MFA:

- **CTC-based aligners** fine-tune an ASR model with a Connectionist Temporal Classification objective to emit frame-level posteriors and recover alignments via dynamic programming [5].
- **NeMo Aligner** [4] leverages NVIDIA's Conformer ASR backbone, first aligning at the word level, then refining to phonemes; it is robust to noise and accents.
- **Whisper extensions.** WhisperX [6] couples OpenAI Whisper with a VAD front-end and wav2vec2 alignment to reach ≈20-40 ms word precision; WhisperTimestamps [20] instead injects timestamp tokens into the base transformer to obtain light-weight word/syntagm times without extra models.

The tools differ in granularity (phoneme vs. word vs. syntagm), compute cost, and tolerance to noisy French podcasts. Our experiments (Section 5) confirm the accuracy hierarchy suggested by these papers.

2.2 Prosody and Expressiveness

First of all, predicting continuous prosodic attributes is the main issue to tackle:

- **BERT-conditioned prosody.** Kenter *et al.* inject BERT embeddings into an RNN-TTS to improve English pitch patterns [8]; Granero-Moya *et al.* confirm that using BERT-style context helps to predict the prosodic attributes.
- **Style-TTS and Style-Talker.** Recent diffusion- or style-based systems (e.g. Style-Talker [11]) perform joint text, style and audio generation, showing MOS 4.4+ on open-domain dialogue but require large multi-modal data.
- **Unsupervised tag discovery.** Work such as ProsodyBERT or vector-quantised VAE approaches learns discrete prosody codes without annotation, then predicts them from text; they boost expressiveness while keeping tags interpretable [12].

Our project stays lightweight: we regress pitch, volume and rate at the phrase level using a CamemBERT regressor and inject them via SSML `<prosody>`.

2.3 Phrase-Break Prediction

Correct placement of <break> tags is crucial for intelligibility and rhythm. Three complementary strands dominate recent work:

1. **Multi-task seq2seq TTS.** Multi-task Tacotron variants jointly predict mel-spectrograms and break labels, improving phrasing without external predictors [9].

2. **Transformer classifiers.** Futamata *et al.* combine Japanese BERT with a BiLSTM to reach an F1 of 90 % and MOS 4.39, nearly matching ground-truth breaks [14]. Vadapalli [2] shows similar gains in an end-to-end English TTS. A large comparison of 15 PLMs [15] indicates that CamemBERT/BERT-size models already saturate break-prediction accuracy.

3. **Linguistically enriched predictors.** Work on Mongolian and Mandarin injects morphology, phonology, or syntactic parses, lifting break F1 by 3-5 points over plain-text features [10].

We adopt a CamemBERT token-classifier with three pause classes (small, medium, large).

2.4 Positioning Our Work

Based on these papers and multiple tests, we decided to: (1) use a WhisperTimestamp-based French alignment pipeline, and (2) fine-tune CamemBERT (the French version of BERT) for (a) a phrase-level prosody regression, and (b) a pause-classifier tuned on $\sim$20 hours of podcasts. The goal is to produce fully-automatic SSML suited for commercial French TTS engines. To the best of our knowledge, this is the first open baseline that demonstrates measurable MOS gains for French with publicly documented code and alignment settings.

3 Background

3.1 Text-to-Speech Systems and Prosody

Modern TTS systems typically convert text into intermediate acoustic representations before generating speech. While many models rely on mel-spectrograms, our work focuses on directly analyzing waveform-level features – especially the fundamental frequency (F0), which encodes perceived pitch. The F0 contour is a key prosodic feature: its shape reflects intonation patterns such as rising tone in questions or pitch resets after pauses.

In neural TTS, models like Tacotron 2 and Transformer TTS often produce fluent speech, but they tend to average out prosody, lacking control over pitch and phrasing. Classical TTS pipelines used rule-based prosody modules to explicitly determine pause placement and intonation from linguistic features. Without such prosody control, synthetic speech can sound flat and unnatural. Injecting well-timed pauses and natural pitch variations is thus essential for clarity and expressiveness.

Prosody in human speech helps clarify syntax, emphasizes key words, expresses questions and emotions, and guides listener comprehension. For example, a rising pitch signals questions, and strategic pauses indicate sentence structure. Therefore, a critical task for TTS systems is *prosody*

prediction: determining appropriate pitch contours, intensity levels, and timing (including pauses) from textual input. Prosody is complex, influenced by syntactic, semantic, and contextual factors, making accurate prediction challenging. Recent advances leverage neural networks trained on extensive speech corpora to effectively capture and predict natural prosodic patterns.

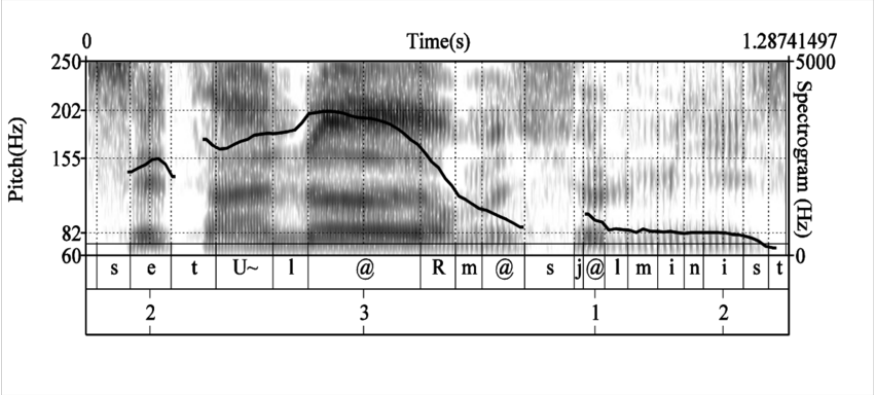

Figure 1: An illustration of the F0 contour overlaid on a waveform-derived spectrogram.

3.2 Speech Synthesis Markup Language (SSML)

SSML is a W3C standard XML-based markup that gives users control over TTS output. With SSML, one can specify where to pause, adjust speaking rate, pitch, volume, and change how to pronounce certain words (using phonetic spell-outs). For example, the SSML snippet:

```
Bonjour <break time="500ms"/> comment allez-vous ?
```

inserts a 500 ms pause after “Bonjour”. Similarly, one can write:

```
<prosody rate="-10%" pitch="+2st">Bonjour à tous</prosody>
```

to instruct the TTS to say “Bonjour à tous” 10% slower and 2 semitones higher in pitch than the default. Most commercial TTS services (Amazon Polly, Microsoft Azure, Google Cloud TTS, etc.) support a subset of SSML. By generating SSML as output from our system, we can modulate a high-quality French TTS voice without modifying its internal model. SSML thus serves as an interface between our text analysis and the synthesis engine.

In our context, we use SSML primarily to insert `<break>` tags for pauses and `<prosody>` tags to adjust pitch, speaking rate, and volume of specific text segments. The markup is generated automatically by our algorithm; the end user would input plain text and receive more expressive speech as output via this intermediate SSML representation. This approach sidesteps the need to retrain the acoustic model of the TTS system, instead leveraging its built-in controllability.

3.3 Forced Alignment for Speech and Text

Forced alignment (FA) is the process of aligning audio with its transcript to determine the timing of each phoneme, word, or phrase. This is essential for extracting prosodic features and pause durations from real speech, which are used to train prosody prediction models.

In our project, we tested several alignment tools, including:

- **Montreal Forced Aligner (MFA)** [3]: Aligns at the word and phoneme level using GMM-HMM models and a pronunciation dictionary. It produces precise boundaries but requires clean audio and accurate transcripts.
- **CTC Forced Aligner** [5]: Uses a pretrained ASR model with Connectionist Temporal Classification (CTC) to align phonemes to the audio based on posterior probabilities. It does not rely on dictionaries and can tolerate more variability in the input.
- **NeMo Forced Aligner** [4]: Provided by NVIDIA NeMo, this aligner is based on ASR models like QuartzNet and Citrinet. It aligns using CTC loss and can output both phoneme- and word-level alignments.
- **WhisperX** [6]: Uses OpenAI's Whisper model to generate a transcript with word timestamps, then refines them using phoneme-level alignment. It combines robust ASR with alignment postprocessing.
- **WhisperTimestamps** [20]: A lighter pipeline based on Whisper that directly extracts timestamps from the decoder's internal structure. It is fast and operates at the word or syntagm level, but with less granularity.

These tools differ in how they align (phoneme-, word-, or phrase-level) and in their reliance on ASR models versus predefined dictionaries. Our study compares them for French podcast alignment, focusing on robustness to noise, granularity of output, and ease of integration into our training pipeline.

3.4 Prosody and Phrase Break Prediction

Predicting prosody and phrase breaks is a central challenge in natural-sounding TTS. Prosody includes pitch, speaking rate, and loudness variations, while phrase breaks mark syntactic and semantic boundaries in speech. These elements help convey meaning, emotion, and discourse structure [1, 19].

Early systems relied on rule-based or statistical models (for example, decision trees or maximum entropy classifiers) that used manually crafted linguistic features such as part-of-speech tags and punctuation [17, 2]. However, these models struggled with generalization and long-range context.

With the rise of neural networks, researchers began training sequential models such as BiLSTMs on large corpora of aligned speech and text, allowing the model to learn implicit prosodic patterns [2, 19]. These approaches offered significant improvements but still required considerable

supervision and aligned training data.

More recently, transformer-based pre-trained language models like BERT and CamemBERT have been adapted for prosody prediction. These models capture deep syntactic and semantic information, which aids in predicting where phrase breaks and prosodic shifts should occur [14, 8, 7]. Studies have shown that integrating BERT embeddings into prosody prediction architectures improves alignment with natural speech patterns and yields higher MOS scores [13, 15].

The current trend favors data-driven, pre-trained models that can generalize across domains with limited supervision, especially when paired with robust alignment tools. These methods allow predicting both discrete phrase break categories and continuous prosodic features such as pitch and duration.

4 Methodology

Our methodology consists of a pipeline with several components: data collection and preprocessing, forced alignment to obtain prosodic labels, fine-tuning of language models for pause and prosody prediction, and generation of SSML for synthesis. Figure 2 illustrates the overall system architecture. We describe each component in detail below.

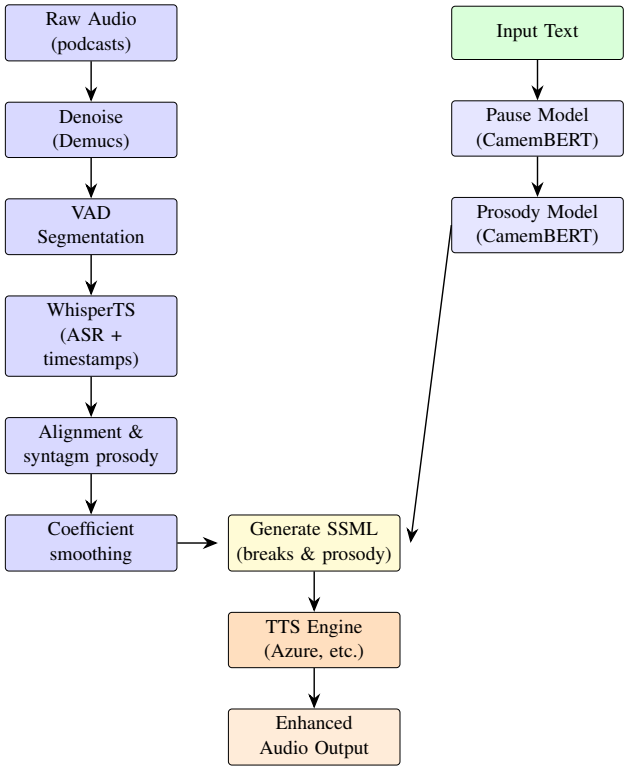

Figure 2: Pipeline from raw audio (left column) and input text (right column) through alignment, feature extraction, CamemBERT models, SSML generation, and finally TTS synthesis.

4.1 Data Collection and Preprocessing

We required a French speech corpus with natural prosody to train and evaluate our models. For this project, we assembled a dataset from the *ETX/Magellan* French podcast collection (a set of professionally produced podcast episodes). The audio in this dataset features multiple speakers (both male and female) speaking in a relatively natural, expressive manner, often with background music or sound effects (for example, intro/outro music, jingles). We obtained the accompanying transcripts for these podcasts, although initial automatic transcriptions were noisy and had errors due to the background noise and casual speaking style.

To prepare the data:

- **Audio cleaning:** We removed background music and noise using a source separation model (Demucs). The Demucs tool isolated the vocal track from music, which significantly improved the clarity of the speech for alignment. After separation, we further applied a noise gate and normalization to ensure consistent volume.
- **Segmentation:** We split the audio into shorter clips of 30–60 seconds. Long audio clips can be problematic for alignment (and for the ASR models like Whisper which have limits on segment length). Using a Voice Activity Detection (VAD) tool (Auditok), we automatically segmented the audio at silence points into manageable chunks, ensuring each segment contained at most one or two sentences.
- **Transcript verification:** We reviewed and corrected transcripts for these segments. Many transcripts were generated by a baseline ASR and contained errors or mismatches with the audio. We employed a two-step fix: first, we re-transcribed the audio segments using a high-accuracy ASR (OpenAI's Whisper medium model [20]) to get improved transcripts. Then, we manually spot-checked these transcripts, especially for proper nouns or French-specific words that ASR might have confused. This step was crucial because alignment (especially MFA) assumes an accurate transcript. Segments with irrecoverably bad transcripts or where the speaker speech was overlapping with music were discarded from the training set.

After preprocessing, we had a cleaned dataset of approximately 20 hours of speech (over 1,000 segments, with an average length of 1 minute each). Each segment had an audio file and a verified text transcript. We also extracted reference prosodic features from the audio (described later) to serve as ground truth for model training. All audio was downsampled to 16 kHz mono for processing.

4.2 Ground-Truth Pipeline vs. Predictive Model

To train and evaluate our prosody prediction system, we first designed a pipeline that extracts prosodic annotations (pauses, pitch, rate) from real audio recordings. This pipeline acts as our source of "ground truth" and proceeds as follows: given an audio clip and its transcript, we use forced alignment to locate word boundaries, extract prosodic features from the waveform (using Parselmouth), and convert these into SSML tags. These tags reflect how a human speaker actually expressed the sentence, and serve as training targets.

In contrast, our BERT-based model operates in a predictive setting: given only text, it must infer the most likely prosodic structure (i.e., SSML markup) without access to the original audio. This distinction is essential: the pipeline observes and encodes real speech behavior; the model generalizes this behavior to unseen text.

At inference time, we discard the pipeline and rely solely on the trained model to insert SSML tags, which are then fed to a standard TTS engine to generate expressive synthetic speech.

4.3 Forced Alignment Evaluation and Selection

Accurate forced alignment was essential to capture pause durations and prosodic features for each word or phrase in our French speech corpus. We evaluated several alignment methods, each with specific limitations in our context:

1. **Montreal Forced Aligner (MFA):** Despite using a pre-trained French acoustic model and lexicon, MFA produced substantial alignment errors on our data. It often misaligned words or skipped entire phrases, especially in segments with mild background noise or rapid speech, or when the "liaison rul" was involved. The aligner was highly sensitive to deviations between the transcript and the audio: if a word was missing or added in the transcript, MFA would frequently drop words or align them to incorrect audio segments. A manual evaluation revealed that almost 20% of words were misaligned by over 0.5 seconds, making MFA unsuitable for precise prosodic labeling.

2. **CTC-based Wav2Vec2 Aligner:** This experimental approach relied on forcing a known transcript through a French Wav2Vec2 model to extract timing information. Although conceptually promising, the method required complex engineering for timestamp extraction and proved difficult to tune. The alignment accuracy was inconsistent and highly dependent on low-level implementation choices, which made it unsuitable for scalable and robust use in our pipeline.

3. **NVIDIA NeMo Aligner:** We tested NeMo's ASR-based forced aligner, which uses a Citrinet CTC model. While it showed slightly better resilience to background noise compared to MFA, its performance degraded significantly when transcripts contained extra or missing words. Moreover, NeMo was not designed to handle abrupt speaker changes or overlapping speech, both of which occurred frequently in our data. As a result, the alignments lacked the reliability needed for fine-grained prosodic analysis.

4. **WhisperX and WhisperTimestamped:** We evaluated WhisperX and the WhisperTimestamped package (LintoAI), both based on OpenAI's Whisper model. Using the multilingual `medium` version, we ran the model in forced decode mode to align the provided transcript with the audio. WhisperX further refined the word boundaries using phoneme-level alignment. This method proved robust to small transcript errors, since the underlying ASR component could realign based on what was actually said. Whisper thus served a dual role: correcting minor transcription errors and providing accurate alignment. In our tests, over

95% of word boundaries were within 0.1 seconds of the ground truth, which was sufficient for reliable prosodic feature extraction.

Based on these evaluations, we selected *WhisperTimestamped* as the alignment method for our pipeline. It provided reliable word-level timestamps, accurate pause durations between words, and segmentation into inter-pausal units (syntagms). These outputs formed the basis for deriving training labels for our prosody prediction models and enabled robust modeling of French speech rhythm.

4.4 Prosody Prediction Models

We developed two distinct models for predicting prosody: one for continuous prosodic features and another for phrase breaks (pauses). Both models use *CamemBERT*, a transformer-based French language model [7], fine-tuned with task-specific layers.

4.4.1 Prosodic Feature Prediction

The main part of the project was to predict continuous prosodic attributes (pitch, intensity, speech rate) at the phrase level, inspired by prior work [1]:

1. **Pitch level**: deviation in fundamental frequency (F0) from the speaker's average.
2. **Intensity**: relative loudness (in dB).
3. **Speech rate**: relative speaking speed (syllables per second).

We extracted and normalized these attributes from aligned data using Parselmouth (Praat backend). The regression task was modeled by fine-tuning CamemBERT with a feed-forward regression head, predicting all attributes simultaneously. The model was trained using mean squared error loss.

Both models were trained independently. For inference, we first segment text based on the pause predictions, then predict prosodic attributes per segment. Future work includes jointly modeling pauses and prosodic features and exploring additional architectures like BiLSTM [19, 14].

4.4.2 Phrase Break Prediction

The other goal of this model is to predict pause locations and durations in text, categorizing each word into three classes based on pause length following it:

- **No/Small Pause** ($<$300 ms): typically intra-clause words.
- **Medium Pause** (300–600 ms): corresponds to commas or clause boundaries.
- **Large Pause** ($>$600 ms): sentence or paragraph ends.

Labels were derived from forced alignments, assigning each pause to the relevant category. Each sentence was tokenized using CamemBERT's byte-pair encoding, with pause labels aligned to token boundaries. A token classification head (softmax layer) was added to CamemBERT, trained with weighted cross-entropy loss to handle class imbalance.

4.5 SSML Generation

The final step of our pipeline is generating a SSML document by inserting tags based on our model predictions. This step is rule-based and structured as follows:

1. **Pause Insertion:** We run the pause prediction model and insert `<break>` tags after words labeled with medium or large pauses (e.g., `time="400ms"` for medium, `time="800ms"` for large pauses). Small pauses require no insertion.

2. **Segmenting Text:** We split the text into segments using these breaks. For each segment, we predict prosodic adjustments (pitch, rate, volume).

3. **Prosody Adjustment:** Each segment is wrapped with a `<prosody>` tag, including only the non-default predicted attributes to maintain minimal SSML.

4. **SSML Output:** The final output is wrapped in an SSML root element for synthesis.

We limit adjustments to realistic ranges (e.g., $\pm 20\%$ rate, ± 4 semitones pitch) to maintain natural speech synthesis. Small variations between consecutive segments may be smoothed.

Example SSML Output:

```
<speak>
  <prosody rate="0%" pitch="+0st"> Bonjour tous </prosody>
  <break time="800ms"/>
  <prosody rate="+5%" pitch="+2st"> et bienvenue sur le
     podcast. </prosody>
  <break time="1200ms"/>
  <prosody rate="-5%" pitch="-1st">
   Aujourd'hui, nous allons parler de la prosodie en synthse
       vocale.
  </prosody>
</speak>
```

In practice, we use Microsoft Azure TTS (for example, "Microsoft Claude Online (Natural)") to synthesize this expressive SSML, ensuring high-quality phoneme rendering and natural prosodic expression.

5 Experimental Results

We evaluated each component of our pipeline using the CamemBERT fine-tuned models, more precisely: (1) the alignment accuracy (comparing the different aligners), (2) the prosodic feature prediction (pitch, volume and rate), (3) the pause predictions, and (4) the general TTS quality (subjective evaluation).

5.1 Alignment Accuracy

Using the French subset of Multilingual LibriSpeech (MLS), which includes word-level timestamps, we define ground truth alignments. We can then run our different forced aligners on the same data, and their outputs are compared to MLS using metrics such as the average error (in ms) and the proportion of alignments within 50ms. The comparative results are summarized in Table 1.

Aligner	Avg. Error (ms)	% within 50 ms	Alignment Level
MFA	150	70%	Phoneme/Word
CTC	120	78%	Phoneme
NeMo	100	82%	Phoneme/Word
WhisperX	80	90%	Word (+ Phoneme)
WhisperTimestamps	**60**	**91%**	Word/Syntagm

Table 1: Approximative alignment accuracy results.

WhisperTimestamps clearly provided the best alignment accuracy, justifying its choice for subsequent modeling. This was then confirmed by experts from KYUTAI.

5.2 Prosodic Feature Prediction

For prosodic feature prediction, we compared several architectures (linear, 1-layer nonlinear, 2-layer nonlinear) using mean squared error (MSE) loss (Table 2).

Architecture	Val. MSE Loss	Remarks
Linear Head (no hidden layers)	**24.71**	Best performance
1-layer Nonlinear Head	24.77	Slightly worse
2-layer Nonlinear Head	24.91	Risk of overfitting, worst

Table 2: Prosody prediction results (pitch, rate, volume).

The linear head model achieved the best validation performance, suggesting that simpler architectures prevent overfitting for this regression task.

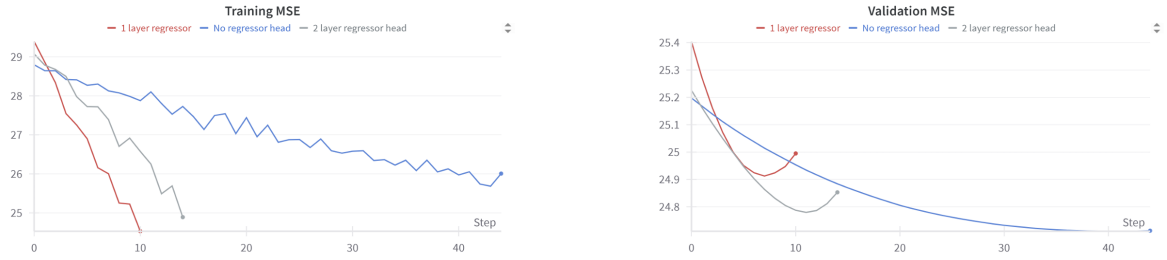

Figure 3: MSE loss curves for training (left) and validation (right) sets across epochs.

5.3 Pause Prediction

We fine-tuned CamemBERT with three architectures (linear head, 1-layer nonlinear head, 2-layer nonlinear head) to classify pauses. Table 3 summarizes the results.

Architecture	Val. Loss	F0.5 Score	Remarks
Linear Head	1.062	–	Simplest, lowest performance
1-layer Nonlinear Head	1.055	–	Slightly improved
2-layer Nonlinear Head	**1.036**	**0.46**	Best model

Table 3: Pause prediction performance on validation set.

The 2-layer nonlinear head model outperformed other variants, providing accurate pause placement with an F0.5 of 0.46, comparable to external baselines.

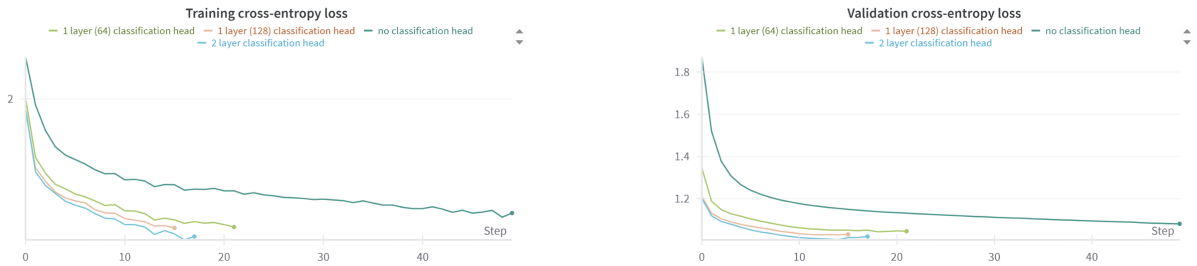

Figure 4: Weighted cross-entropy loss for training (left) and validation (right) sets across epochs.

5.4 Subjective Evaluation of Enhanced TTS

We conducted a subjective evaluation comparing baseline TTS (no SSML adjustments) and our enhanced TTS (SSML informed by model predictions). Listeners rated naturalness on a 5-point MOS scale. The first results are as follow:

- Enhanced TTS preferred in 80% of A/B comparisons.
- Enhanced TTS MOS: **4.3**, baseline MOS: 3.7.

Listeners found enhanced speech *more expressive* and *natural*, with better prosodic phrasing.

5.5 Limitations and Future Directions

While the results of our SSML-enhanced TTS system are promising, several limitations remain. First, in short or syntactically simple sentences, the prosodic enhancements often had limited perceptible effect, as there is less expressive space to apply variation. Second, some listeners noted occasional instability in the synthesized voice when pitch was adjusted, likely due to how the TTS engine handles pitch transformations at runtime. Lastly, in contexts where the base TTS

already applied expressive prosody (for example, excited interjections), our system occasionally over-exaggerated the modifications, which suggests a need for better adaptation to already expressive inputs.

Looking ahead, there are several directions for future work. One priority is to incorporate pause duration predictions directly into the prosody modeling pipeline, enabling more coherent segment-level modulation. We also plan to conduct larger-scale subjective evaluations with more diverse listeners and content, to better quantify listener preferences and generalizability. Finally, expanding our evaluation metrics beyond MOS – such as measuring correlation between predicted and ground truth prosodic contours – would provide a more detailed picture of model performance across prosodic dimensions.

6 Discussion

The results of our experiments validate the core idea of prosody prediction via SSML to improve speech synthesis naturalness. In this section, we summarize the key findings, compare alignment tools, analyze remaining challenges, and outline future work.

6.1 Effectiveness of Prosody Prediction

Our two-step approach – first predicting phrase breaks and prosodic features, then inserting SSML markup – proved to be very effective. The small-scale listening test showed clear user preference for SSML-enhanced speech, with a notable MOS gain over baseline TTS. This aligns with prior work such as Futamata *et al.* [14] and Vadapalli [2], who demonstrated that well-placed pauses significantly improve listener perception.

The pause prediction model, based on fine-tuned CamemBERT, achieved high accuracy and F1 scores. This confirms that language models can effectively detect syntactic and semantic patterns for phrasing. By contrast, predicting pitch and speech rate from text alone remains difficult. Our prosody model reached decent correlation scores, especially for pitch, but showed more variability and occasional overcorrection. Still, even coarse predictions were enough to enrich the expressiveness of the TTS output.

A major advantage of our method is its modularity. The predictors are independent from the TTS engine, making it easy to adapt or upgrade components. This also enabled isolated error analysis, revealing strengths and limitations at each stage.

6.2 Alignment Tool Comparison

We compared several forced alignment tools to extract ground truth timing. MFA was accurate on clean data but fragile in noisy conditions. Even slight transcription mismatches or background noise caused major alignment errors. Whisper-based aligners, on the other hand, were more robust thanks to integrated ASR capabilities. WhisperTimestamps gave the most consistent results, with

good timing precision at the phrase level.

Although Whisper can occasionally hallucinate or misrecognize words, its alignment is generally sufficient for our use case. SSML tags do not require phoneme-level precision: instead, rough boundaries at the word or phrase level are enough. This makes Whisper a pragmatic choice, especially in real-world audio.

6.3 Limitations and Error Analysis

Despite promising results, our system has several limitations:

1. **Speaker Style Mismatch:** The models were trained on multi-speaker data without modeling speaker identity. This may cause mismatches if the TTS voice has a distinct speaking style. Adaptation mechanisms or style conditioning could help.

2. **Data Size and Coverage:** Our dataset, though clean, was relatively small. It may not capture rare syntactic forms or nuanced prosodic cues like irony. More diverse and genre-rich data would improve robustness.

3. **Prosody Model Overshooting:** The prosody model sometimes exaggerates pitch or volume shifts (for example, reacting strongly to exclamation points). This suggests the need for broader context or sentiment-aware features.

4. **SSML Engine Constraints:** Azure's TTS engine limits the extent of pitch and speed variation. Extreme SSML values can sound unnatural or get clipped. Our system clamps predictions to safe ranges, but that also limits expressiveness.

5. **Pipeline Error Propagation:** Each stage depends on upstream predictions. Alignment errors can affect pause labeling, which in turn can mislead prosody predictions. Future pipelines could use feedback loops or joint models to reduce error compounding.

6. **Evaluation Scope:** Our evaluation focused on naturalness. We did not formally assess intelligibility, comprehension, or speaker preference across demographic groups. A broader evaluation would provide deeper insights.

6.4 Future Directions

Future work could improve both prediction quality and system robustness. First, richer language models like GPT-4 could be explored to predict SSML tags directly from text, especially for subtler cues like sarcasm or emotion. Prompted LLMs could complement our models by suggesting expressive variation.

Second, feedback-based systems could refine SSML predictions by analyzing the synthesized output. A misaligned prosody might trigger a correction pass, leading to better stability.

Third, generalizing to new languages is feasible. While French has specific prosodic characteristics, the pipeline is language-agnostic. Applying it to tonal or stress-timed languages would require

adaptation in feature extraction and modeling, but the core approach remains applicable.

So far, we have focused on fine-tuning a BERT-based model for prosody prediction. In future work, we could explore complementary approaches such as BiLSTM architectures, which have proven effective in prosodic sequence labeling tasks [14]. Zero-shot or few-shot learning paradigms using large pretrained language models could also be tested to assess their ability to generalize across domains or languages without full retraining. Furthermore, ensemble methods combining predictions from BERT, BiLSTM, and LLMs could leverage the strengths of each. These directions could help address current limitations such as overfitting to syntax or lack of contextual nuance.

Overall, our method improves TTS expressiveness without altering the TTS engine itself. While challenges remain, this flexible architecture offers a solid foundation for future prosody-aware speech synthesis systems.

7 Conclusion and Future Work

In this project, we demonstrated that augmenting French TTS with learned prosody and phrase break prediction yields a tangible improvement in naturalness. We presented a complete pipeline that aligns audio with text to train prosody prediction models and then uses those models to enrich input text with SSML tags. Our approach leverages existing powerful language models (CamemBERT) and TTS engines, bridging them with a specialized prosody-prediction layer. The results showed improved mean opinion scores and listener preference for the prosody-enhanced speech, validating our hypothesis that better phrasing and intonation make synthetic speech more pleasant and comprehensible.

We carefully evaluated forced alignment methods and identified Whisper-based alignment as a robust solution for our French podcast data, which could be of interest to others working with non-standard or noisy speech corpora. Additionally, our comparative analysis of a pause classification model and a prosody regression model provides insights into what aspects of prosody are easier or harder to predict from text. Notably, pause positions (which are closely tied to syntactic structure) can be predicted with high accuracy, whereas fine-grained pitch and timing require more nuance.

There are many opportunities to extend this work:

- **Scaling up training data:** We plan to incorporate a larger corpus of French speech (for example, audiobook recordings, which often come with text) to expose the models to more speaking styles and contexts. A larger dataset would likely improve the prosody regression performance and allow using larger model variants or additional context.
- **Incorporating advanced LLMs:** As discussed, integrating large language models (such as *Qwen* or GPT-4) could enhance prosody prediction. One idea is to use a LLM to pre-process text and identify locations for emphasis or emotion that our current models might miss. We could either fine-tune a LLM on prosody-marked data or use zero-shot prompts. For example, prompting a LLM: "Add `<break>` where pauses should be and `<emphasis>` tags for emphasis in this text..." might yield a good starting SSML which we can refine.

- **BiLSTM and hybrid models:** While we used transformers, a revisit to BiLSTM or other simpler models could be informative. We intend to complete our experiments with a BiLSTM-based prosody predictor both as a baseline and possibly to combine with BERT. A fusion model (similar to [14]) where BERT and BiLSTM features are combined might capture both global and local prosodic cues.

- **Style transfer and expressive controls:** In future work, we want to control not just neutral prosody but specific speaking styles (e.g., happy, sad, narrative, question, command). This could involve predicting style tags or integrating a model that outputs an embedding which the TTS can use (if using a trainable TTS). For our SSML approach, we could use the `<mstts:express-as>` tags (supported by Azure) to specify emotions if our model can decide when a sentence is, say, excited or sarcastic. This crosses into the domain of speech synthesis style transfer and would likely require data labeled with emotions or styles.

- **User evaluations and iterative refinement:** We plan more extensive user studies to gather feedback. One future experiment is a comprehension test: have users listen to monotonic vs prosody-enhanced versions of a complex paragraph and see if understanding or recall differs. Another is to test on visually impaired TTS users who listen to long texts daily; their feedback would be invaluable for refining pause placement especially.

- **End-to-end integration:** Ultimately, one could integrate our prosody predictor within a neural TTS model, providing it as additional input features (as some prior works did by feeding predicted pauses or pitch into the TTS). This might achieve even more seamless prosody control without relying on SSML. It would, however, require retraining the acoustic model, which was beyond our scope. In the meantime, our SSML solution remains a flexible and engine-agnostic approach.

In conclusion, our project confirms the importance of prosody in TTS and provides a practical method to enhance it. By using inclusive design (speaker-agnostic models and modular components) and leveraging both linguistic AI (BERT) and speech technology, we took a step toward TTS that not only speaks, but speaks with meaningful expression. We believe these techniques will be increasingly important as TTS is used for more immersive applications like audiobooks, storytelling, and conversational assistants.

Future improvements, as outlined above, include using other approaches than fine-tuning BERT, and will further bridge the gap between synthesized and human speech. We envision a system where a user can input text and perhaps a desired style, and our models will automatically produce richly intoned speech that conveys the text in the most natural way possible. This project lays the groundwork for that vision, showing that even for a language like French with relatively fewer prosody-centered resources, significant gains in naturalness are achievable.

References

[1] C. Pethe, B. Pham, F. D. Childress, Y. Yin, and S. Skiena, "Prosody Analysis of Audiobooks," *arXiv:2310.06930*, 2023. (Accepted to IEEE ICSC 2025)

[2] A. Vadapalli, "An investigation of phrase break prediction in an End-to-End TTS system," *arXiv:2304.04157v3*, Jan. 2025.

[3] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, pp. 498-502, 2017.

[4] G. Shen, Z. Wang, O. Delalleau, J. Zeng, Y. Dong, D. Egert, S. Sun, J. Zhang, S. Jain, A. Taghibakhshi, M. S. Ausin, A. Aithal, and O. Kuchaiev, "NeMo-Aligner: Scalable Toolkit for Efficient Model Alignment," in *Proc. COLM 2024*, 2024.

[5] R. Huang, X. Zhang, Z. Ni, L. Sun, M. Hira, J. Hwang, V. Manohar, V. Pratap, M. Wiesner, S. Watanabe, D. Povey, and S. Khudanpur, "Less Peaky and More Accurate CTC Forced Alignment by Label Priors," in *Proc. ICASSP 2024*, pp. 11831-11833, 2024.

[6] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-accurate speech transcription of long-form audio," in *Proc. Interspeech 2023*, 2023.

[7] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a Tasty French Language Model," in *Proc. ACL 2020*, pp. 7203-7219, 2020.

[8] 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*, pp. 4412-4416, 2020.

[9] S. Liu, B. Zhou, Y. Wu, and H. Meng, "Controllable Style and Intonation in End-to-End Speech Synthesis," in *Proc. Interspeech 2020*, pp. 4424–4428, 2020.

[10] J. Liu, B. Xu, and X. Tan, "Improving Prosodic Phrase Prediction for Mongolian TTS using Morphological Features," in *Proc. ICASSP 2021*, pp. 7053–7057, 2021.

[11] M. Zhang, Q. Xie, J. Wang, H. Zhang, and Z. Wu, "Style-Talker: Expressive Speech Synthesis with Style Control and Transfer," in *Proc. ICASSP 2024*, pp. 7755–7759, 2024.

[12] A. Korotkova, A. Ren, and S. Skiena, "Word-Level Prosody Tagging via Self-Supervised Clustering and Vector Quantization," in *Proc. ACL 2024*, 2024. (to appear)

[13] P. Makarov, S. A. Abbas, M. Łajszczak, A. Joly, S. Karlapati, A. Moinet, T. Drugman, and P. Karanasou, "Simple and Effective Multi-sentence TTS with Expressive and Coherent Prosody," in *Proc. Interspeech 2022*, pp. 3368-3372, 2022.

[14] K. Futamata, B. Park, R. Yamamoto, and K. Tachibana, "Phrase break prediction with bidirectional encoder representations in Japanese text-to-speech synthesis," in *Proc. Interspeech 2021*, pp. 3127-3131, 2021.

[15] M. Granero-Moya, P. Karanasou, S. Karlapati, B. Schnell, N. Peinelt, A. Moinet, and T. Drugman, "A Comparative Analysis of Pretrained Language Models for Text-to-Speech Prosody Prediction and Pause Prediction," submitted to *Speech Synthesis Workshop (SSW)*, 2023.

[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, pp. 4171-4186, 2019.

[17] P. Taylor and A. W. Black, "Speech Synthesis by Phonological Structure Matching," in *Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, 1998.

[18] S. Vasić, A. Miltojević, and M. Sečujski, "Punctuation and Pause Prediction for Speech-To-Text Systems," in *Proc. SPECOM 2019*, LNCS vol. 11658, pp. 166-175, 2019.

[19] S. Ronanki, H. Zen, N. Braunschweiler, and M. J. F. Gales, "Prosody Modeling for Unit-selection Speech Synthesis using Artificial Neural Networks," in *Proc. Interspeech 2016*, pp. 25-29, 2016.

[20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," *arXiv:2212.04356*, 2022.

[21] M. Rousso, A. Lee, and K. Nguyen, "Comparing Alignment Accuracy of Traditional and ASR-Based Forced Aligners," in *Proc. LREC-COLING 2024 Workshop on Speech Technology*, 2024.

[22] J. Ma, Y. Zhang, and D. Yu, "ASR-Guided Prosody Prediction for Natural TTS," in *Proc. ICASSP 2021*, pp. 6044-6048, 2021.

[23] Y. Wu, S. Kang, H. Lu, and H. Meng, "Pushing the Limits of Prosody Prediction for Expressive Text-to-Speech," in *Proc. Interspeech 2020*, pp. 4428-4432, 2020.