Improving French Synthetic Speech Quality

via SSML Prosody Control

Emeric PAYER

Academic Year 2024–2025

Personal Research Project (PRL) Defense
CMAP - École Polytechnique
Supervisors: Nassima OULD-OUALI and Eric MOULINES

Outline

Introduction and Motivation

Introduction to Text-to-Speech (TTS) Systems

- TTS systems convert written text into audio
- Modern systems are built with deep neural networks
- TTS has many applications:
  - Audiobooks and voice assistants
  - Customer service bots
  - Help for visually impaired people
  - Language learning

Project Context

- This work is part of a broader research project:
  "Improving French Synthetic Speech Quality via SSML Prosody Control".
- Goal: Enhance the expressiveness and naturalness of French TTS through better prosody modeling
- Several pipelines are being explored in parallel, differing in:
  - How prosody is predicted (rule-based, BERT, LLMs, etc.)
  - Alignment tools and data preprocessing methods
  - Integration with TTS engines (via SSML or direct conditioning)
- **We focus on one such pipeline** using forced alignment and CamemBERT-based prosody predictors

The Problem with Current TTS Systems

- Neural TTS systems produce **intelligible** speech
- But often sound **monotonous** and **unnatural**
- Missing rich prosody: variations in pitch, rate, volume, and pauses
- Especially noticeable in French TTS systems

Key Challenge

How can we make synthetic French speech sound more natural and expressive?

Importance of Prosody

Prosody enhances:

- Emotional expression - conveying feelings and attitudes
- Structural clarity - marking syntactic boundaries
- Listener engagement - sustaining attention

Examples:

- Rising pitch signals questions
- Strategic pauses indicate sentence structure
- Pitch resets after major breaks

Our Approach

Two-Step Solution

1. Use **language models** to predict prosodic features from text
2. Generate **SSML markup** to control existing TTS engines

Key Advantages:

- Leverages existing high-quality TTS voices
- No need to retrain acoustic models
- Modular and flexible architecture

Background and Related Work

Speech Synthesis Markup Language (SSML)

SSML = W3C standard XML-based markup for TTS control

Example:

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

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

Supported by: Amazon Polly, Microsoft Azure, Google Cloud TTS

Literature Review: Forced Alignment

Traditional Tools:

- Montreal Forced Aligner (MFA) - HMM-GMM based
- Requires clean audio and accurate transcripts

Neural Aligners:

- CTC-based aligners - ASR model with CTC loss
- NeMo Aligner - NVIDIA's Conformer backbone
- WhisperX - Whisper + wav2vec2 alignment
- WhisperTimestamps - Direct timestamp extraction

Literature Review: Prosody Prediction

BERT-based Approaches:

- Kenter et al.: BERT embeddings + RNN-TTS for English
- Context-aware prosody prediction

Phrase Break Prediction:

- Futamata et al.: Japanese BERT + BiLSTM (F1 = 90%)
- Vadapalli: End-to-end English TTS with breaks
- Transformer classifiers outperform rule-based methods

Our Contribution: First open baseline for French TTS with CamemBERT

Methodology

System Architecture Overview

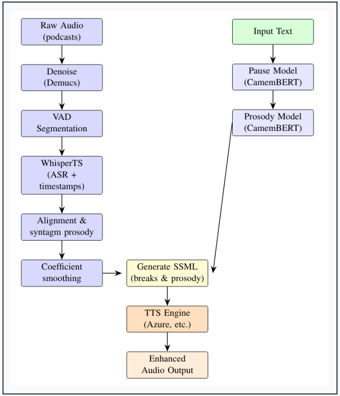


Figure 1: Full pipeline from raw audio (left) and input text (right)

Data Collection and Preprocessing

Dataset: Majelan X (ex ETX Majelan) French podcast collection

- 20 hours of speech
- Multiple speakers (male/female)
- Natural, expressive speech

Preprocessing Steps:

1. Audio cleaning - Demucs for source separation
2. Segmentation - VAD for 30-60s clips
3. Transcript verification - Whisper + manual correction

Forced Alignment Tool Comparison

Evaluation Metrics:

- Average alignment error (ms)
- Percentage within 50ms threshold
- Tested on Multilingual LibriSpeech (MLS) French

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.

CamemBERT Models

Two Fine-tuned Models:

1. **Pause Prediction Model**
   - **Task:** Token classification (3 classes)
   - **Classes:** Small ($<$300ms), Medium (300-600ms), Large ($>$600ms)
   - **Architecture:** CamemBERT + classification head

2. **Prosody Regression Model**
   - **Task:** Predict pitch, rate, volume at phrase level
   - **Features:** F0 deviation, intensity (dB), speech rate
   - **Architecture:** CamemBERT + regression head

Prosodic Feature Extraction

From aligned audio, we extract:

- **Pitch level** - F0 deviation from speaker average
- **Intensity** - Relative loudness (dB)
- **Speech rate** - Syllables per second

Tools used:

- Parselmouth (Praat backend) for feature extraction
- Normalization per speaker
- Smoothing across segments

SSML Generation Pipeline

Rule-based conversion from predictions:

1. **Pause Insertion** - Add `<break>` tags
   - Medium pause: `time="400ms"`
   - Large pause: `time="800ms"`
2. **Text Segmentation** - Split by breaks
3. **Prosody Adjustment** - Wrap segments with `<prosody>`
   - Rate: $\pm 20\%$ maximum
   - Pitch: ± 4 semitones maximum

SSML Example Output

```
<speak>

<prosody pitch="+2.01%" volume="+10.00%" rate="-3.10%">
Il y a dans la parole ce qu’on appelle la voix d’implication.</prosody>
<break time="500ms"/>

<prosody pitch="+2.73%" volume="+10.00%" rate="-2.18%">
Lorsque je vous parle actuellement,</prosody>
<break time="360ms"/>

<prosody pitch="+1.97%" volume="+10.00%" rate="-2.26%">
je fais un effort particulier pour moduler ma voix.</prosody>

</speak>
```

Experimental Results

Alignment Tool Results

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

Winner: WhisperTimestamps - best accuracy and robustness

Prosody Prediction Results

Architecture Comparison (MSE Loss):

Architecture	Val. MSE	Performance
Linear Head	**24.71**	**Best**
1-layer Nonlinear	24.77	Slightly worse
2-layer Nonlinear	24.91	Risk of overfitting

Prosody Prediction Results

Key Finding: Simpler architectures prevent overfitting.

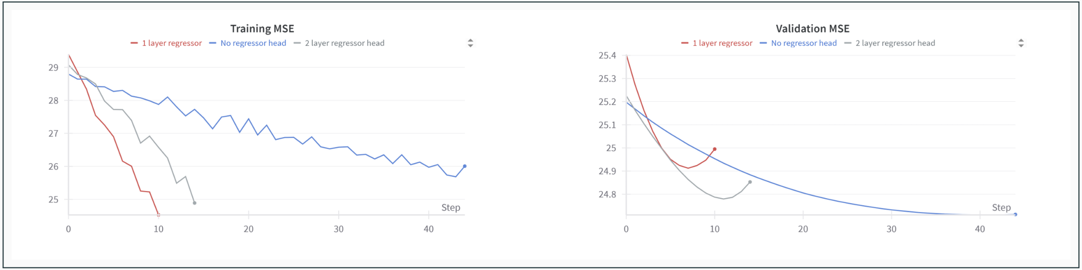


Figure 2: MSE loss curves for training (left) and validation (right) sets across epochs.

Pause Prediction Results

Architecture Comparison:

Architecture	Val. Loss	F0.5 Score	Remarks
Linear Head	1.062	-	Simplest
1-layer Nonlinear	1.055	-	Slight improvement
2-layer Nonlinear	**1.036**	**0.46**	**Best model**

Pause Prediction Results

Key Finding: The 2-layer nonlinear head model outperformed other variants.

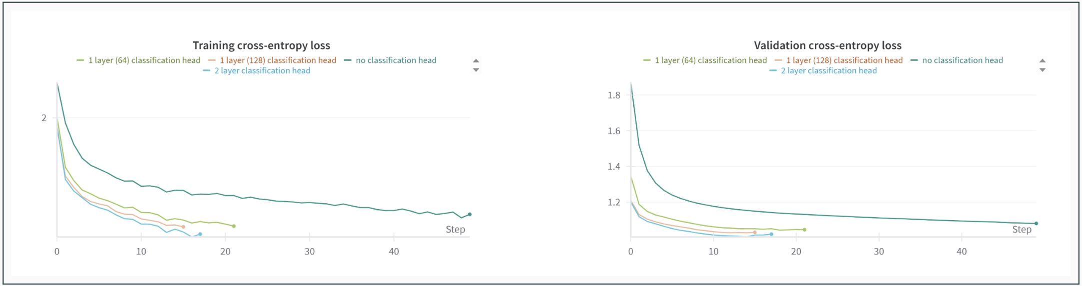


Figure 3: Weighted cross-entropy loss for training (left) and validation (right) sets across epochs.

Subjective Evaluation Results

A/B Listening Test:

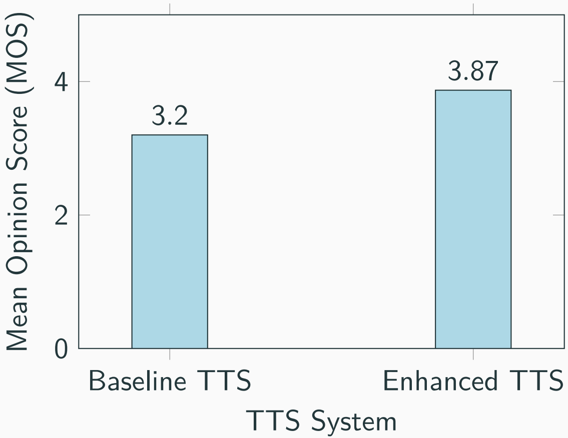
- Enhanced TTS vs. Baseline TTS
- 5-point Mean Opinion Score (MOS)

Results

- **Over 80% preference** for enhanced TTS
- **Enhanced MOS: 3.87** vs. Baseline: 3.20 (20% improvement)
- Listeners found speech more **expressive** and **natural**

Improvement: +0.67 MOS points significant for TTS evaluation

Performance Visualization



Discussion and Analysis

Key Findings

Strengths of Our Approach:

- **Modular design** - Independent components
- **Significant MOS improvement** - +0.67 points / 20% improvement
- **Robust alignment** - WhisperTimestamps excels
- **Effective pause prediction** - High accuracy on syntactic boundaries

Technical Insights:

- Simple architectures work better for prosody regression
- Pause prediction easier than fine-grained prosody
- SSML provides effective interface to existing TTS

Limitations and Challenges

Current Limitations:

1. **Speaker Style Mismatch** - No speaker modeling
2. **Limited Training Data** - Only 20 hours
3. **Prosody Overshooting** - Occasional over-exaggeration
4. **SSML Engine Constraints** - Azure TTS limitations

Error Analysis:

- Pipeline error propagation
- Sensitivity to transcript quality
- Limited evaluation scope (naturalness only)

Alignment Tool Trade-offs

Traditional (MFA):

+ High precision on clean data
+ Phoneme-level detail
- Sensitive to noise
- Requires perfect transcripts

Neural (Whisper):

+ Robust to noise/errors
+ Self-correcting transcripts
+ Easy integration
- Less granular output

Recommendation

WhisperTimestamps optimal for real-world applications

Future Directions

Future Directions and Improvements (1)

Unified End-to-End Model

- Investigate the feasibility of unifying the cascaded approach into a single end-to-end model.
- Aim to jointly predict prosodic structure and parameters for more efficient processing.

Multimodal Audio Embeddings

- Incorporate multimodal audio embeddings to capture subtle speech characteristics beyond text-derived features.
- Enhance the model's ability to interpret and replicate nuanced prosodic elements.

Future Directions and Improvements (2)

Cross-Linguistic Generalizability

- Extend the methodology to additional languages to assess cross-linguistic generalizability and robustness.
- Adapt the model to handle diverse prosodic characteristics across different languages.

Enhanced Dataset

- Expand the dataset to include more diverse speech samples and a broader range of speakers.
- Incorporate more hours of annotated speech to improve model training and performance.

Advanced Research Directions

Large Language Models:

- GPT-4/Qwen for direct SSML generation
- Zero-shot prosody prediction
- Emotion and style detection

Expressive Controls:

- Style transfer (happy, sad, narrative)
- `<mstts:express-as>` tag prediction
- Multi-modal emotion recognition

End-to-end Integration:

- Direct neural TTS integration
- Continuous prosody embeddings

Broader Applications

Immediate Use Cases:

- **Audiobook narration** - More engaging storytelling
- **Voice assistants** - Natural conversational speech
- **Accessibility tools** - Better screen readers
- **Language learning** - Proper pronunciation modeling

Language Extension:

- Adapt pipeline to other Romance languages
- Explore tonal languages (Mandarin, Vietnamese)
- Cross-lingual prosody transfer

Conclusion

Project Summary

What we accomplished:

- **Complete pipeline** from text to expressive SSML
- **Robust alignment** evaluation and selection
- **Successful fine-tuning** of CamemBERT for French prosody
- **Measurable improvement** in speech naturalness

Technical contributions:

- First open French TTS prosody baseline
- Comprehensive alignment tool comparison
- Practical SSML generation approach

Key Takeaways

Main Results

- **+0.67 MOS improvement** with prosody-enhanced SSML
- **WhisperTimestamps** best for French alignment
- **Simple architectures** work well for prosody regression
- **Modular design** enables flexible improvements

Broader Impact:

- Demonstrates feasibility of prosody prediction for French
- Provides foundation for future expressive TTS research
- Applicable to various speech synthesis applications

Towards TTS that not only speaks, but speaks with meaningful expression

Next steps:

- Scale to larger, more diverse datasets
- Integrate advanced language models
- Develop user-controllable style parameters
- Bridge the gap between synthetic and human speech

References

Key References

- **Pethe et al. (2023):** "Prosody Analysis of Audiobooks," *arXiv:2310.06930*
- **Vadapalli (2025):** "Investigation of phrase break prediction in End-to-End TTS," *arXiv:2304.04157v3*
- **Martin et al. (2020):** "CamemBERT: a Tasty French Language Model," *Proc. ACL 2020*
- **Bain et al. (2023):** "WhisperX: Time-accurate speech transcription," *Proc. Interspeech 2023*
- **Futamata et al. (2021):** "Phrase break prediction with BERT in Japanese TTS," *Proc. Interspeech 2021*
- **Kenter et al. (2020):** "Improving RNN-based TTS Prosody with BERT," *Proc. Interspeech 2020*

Alignment Tools:

- Montreal Forced Aligner (MFA), NeMo Aligner, WhisperTimestamps

Thank you for your attention!

Any questions?

Appendix: SSML Tag Details

Break Tags:

- `<break time="Xms"/>` - Pause duration
- Supported: 0ms to 10000ms

Prosody Tags:

- `rate`: -50% to +100% (speech speed)
- `pitch`: -2st to +6st (semitones)
- `volume`: -50dB to +50dB (loudness)

Our Constraints:

- Rate: $\pm$20% maximum
- Pitch: $\pm$4 semitones maximum
- Volume: $\pm$10dB maximum