Hybrid ASR-T5 Architecture for Robust French Medical Speech Recognition: Dual-Encoder Design for Production Systems

Mohammad Alzahrani[1] **Damien Forest**[2]

[1]*Jeddah, Saudi Arabia* [2]*Paris, France* *PraxySanté (praxy.ai)*

Abstract

We present a dual-encoder hybrid ASR-T5 architecture for French speech recognition that leverages both the encoder and decoder of pre-trained T5 language models — unlike existing approaches using only encoders (Whisper) or only decoders. A FastConformer acoustic encoder is combined with T5's full encoder-decoder via a lightweight adapter (~4M params). Controlled comparison on identical 10k-hour training data yields an 83.3% win rate over our production NeMo baseline across 18 evaluation categories. General-domain training achieves 3.9% validation WER with strong zero-shot generalization to the medical domain (22.78% WER), demonstrating architectural robustness rather than dataset overfitting.

Index Terms: automatic speech recognition, hybrid ASR, dual-encoder, T5, French ASR, medical speech, LoRA, silence handling

1. INTRODUCTION

Automatic Speech Recognition (ASR) has evolved from Hidden Markov Models to modern end-to-end neural approaches. While Conformers [6] and Transducers achieve strong performance, they tightly couple acoustic and language modeling. Recent work explores hybrid approaches combining acoustic encoders with pre-trained language models [9, 11].

Existing hybrid models exploit only one LM component: Whisper [11] uses its encoder for feature extraction while the decoder generates; decoder-only approaches append acoustic features directly to decoder inputs. Our key insight is that utilizing *both* encoder *and* decoder of a pre-trained LM yields superior linguistic modeling via a dual-encoder design.

As a company deploying production ASR, we ask: can a dual-encoder hybrid ASR-T5 architecture outperform a strong production baseline under identical training conditions?

Contributions.

- Dual-encoder architecture combining FastConformer [7] with T5's complete encoder-decoder [4] via a lightweight adapter (~4M params).
- Two controlled configurations: general-domain SOTA comparison (3.9% val. WER) and 10k-hour mixed-data architectural validation.
- Three-pronged silence handling: VAD preprocessing, pure silence training (100h), and silence token injection (250k samples).
- 83.3% win rate (15/18) over our production NeMo baseline; 18.6% relative WER reduction on 11 proprietary medical test sets.

2. RELATED WORK

2.1 End-to-End ASR

Modern ASR is dominated by Listen Attend and Spell [13], RNN-Transducers, and Conformers [6]. FastConformer [7] introduces linearly scalable attention. NeMo [10] provides production-grade multi-language implementations.

2.2 Hybrid Architectures

Whisper [11] demonstrated large-scale weakly-supervised encoder-decoder ASR but uses its encoder primarily for feature extraction. Canary [8] and Voxtral [9] are current SOTA trained on large datasets. Both use encoder + decoder but never treat the encoder as an independent linguistic enrichment stage. Preliminary experiments with single-encoder hybrid designs — acoustic encoder connected directly to a decoder LM — yielded WER of 50–100% with systematic hallucinations on all evaluation sets, making such architectures impractical for production and motivating the dual-encoder approach presented here.

2.3 Transfer Learning

Pre-trained LMs (T5 [4], BERT, GPT) offer strong transfer. LoRA [5] enables efficient adaptation with minimal additional parameters while preserving pre-trained knowledge.

3. METHOD

3.1 Architecture Overview

The pipeline has three components. The **ASR FastConformer Encoder** processes 16 kHz spectrograms producing 512-dim acoustic features (NeMo [10]). The **Hybrid Adapter** (~4M params) aligns them to T5's 768-dim input space using Expansion Mode (1.25× hidden dim, dual LayerNorm, dropout 0.05). The **T5 Encoder-Decoder** (French T5-base) adds 12 layers of linguistic enrichment before auto-regressive generation.

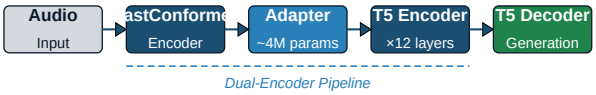

Fig. 1: *Dual-encoder pipeline. Acoustic features flow through FastConformer then T5 Encoder before T5 Decoder generation.*

3.2 Dual-Encoder Design

Two distinct encoding stages are defined:

$$h_{acoustic} = \text{ASR-Encoder}(audio) \quad (1)$$

$$h_{adapted} = \text{Adapter}(h_{acoustic}) \quad (2)$$

$$h_{linguistic} = \text{T5-Encoder}(h_{adapted}) \quad (3)$$

$$y = \text{T5-Decoder}(h_{linguistic}) \quad (4)$$

Unlike Whisper or Voxtral — where the encoder extracts features and the decoder generates — our T5 encoder is an *intermediate linguistic enrichment stage*, adding 12 layers of pre-trained semantic and syntactic knowledge and leveraging T5's encoder-decoder cross-attention from billion-token pre-training [4].

3.3 LoRA Integration

LoRA [5] with rank r=32, α=64, targeting {q, k, v, o, w_i0, w_i1, w_o}, dropout 0.01. ~4M additional trainable parameters enable efficient domain adaptation while preserving pre-trained knowledge.

3.4 Acoustic Silence Handling

Pre-trained T5 has no acoustic silence concept, causing hallucinations. Three complementary strategies address this:

- **VAD preprocessing:** removes silence before model input.

- **Pure silence training:** 100h silence audio paired with `<silence>` tokens.
- **Token injection:** 250k samples with silence tokens at 200–600ms pause positions [14].

3.5 Training Configuration

AdamW with differentiated learning rates: T5 Decoder 4×10^{-5}, T5 Encoder 3×10^{-5}, Adapter 5×10^{-5}, ASR Encoder 1.5×10^{-5}. 15 epochs on 8× NVIDIA A100 GPUs (GENCI Jean Zay HPC cluster), cosine LR (10% warmup), batch 16/GPU, duration filter 0.5–40 s. Inference: beam search (2 beams), repetition penalty 1.15.

4. EXPERIMENTAL SETUP

4.1 Training Configurations

Config 1 — General-domain: CommonVoice [2], Librispeech [1], FLEURS [3] only. Achieves 3.9% validation WER. Purpose: SOTA comparison and zero-shot out-of-domain generalization.

Config 2 — Mixed 10k hours: 50% proprietary synthetic medical speech + 50% real general-domain recordings. Both our hybrid model and production NeMo baseline [10] trained on identical data for controlled architectural validation.

Ethical note: all training data consists of open-source corpora or internally generated synthetic speech. No real patient recordings were used.

4.2 Evaluation and Baselines

Evaluation spans CommonVoice, Librispeech, FLEURS, 11 proprietary medical test sets, spontaneous speech (ESLO, Tcof), formal speech (Tedx, Voxpopuli [12]), and accented speech. External SOTA reference: Canary-1B [8] (85k hours) and Voxtral [9]. We additionally compare against **Whisper-large-v3 fine-tuned on identical Config 2 data** (LoRA, same 10k-hour mix) to provide a direct architectural comparison under equal training conditions [11].

5. RESULTS

5.1 Zero-Shot Generalization (Config 1)

Dataset	WER	CER	RTF
PxCorpus (Medical)	22.78	10.78	0.07
Tedx (Lectures)	16.01	11.21	0.08
Voxpopuli	13.23	9.39	0.11
African Accent	19.51	7.39	0.04
Voxforge	9.45	2.65	0.06
Average	16.20	8.28	0.07

Table 1: *Config 1: zero-shot generalization. RTF = Real-Time Factor on GPU L40.*

Despite zero medical training data, the model achieves 22.78% WER on PxCorpus, confirming architectural generalization. Average RTF = 0.07 on GPU L40 (0.7 s per 10-second sample), approximately 2x the RTF of our standalone FastConformer baseline — a trade-off justified by accuracy gains and significantly lower per-inference cost than cloud ASR APIs at clinical volume.

5.2 SOTA and Whisper Comparison

Model	Data	WER (%)
Ours — Hybrid (Config 1)	CV+LS+FL	3.9
Canary-1B [8]	85k hrs	4.47
Voxtral Mini [9]	Propr.	~4.5
Voxtral Mini-T [9]	Propr.	~4.7
Whisper-large-v3 (FT) [11]	10k hrs	~5.4
Voxtral Small [9]	Propr.	~6.4

Table 2: *French ASR. FT = fine-tuned on identical Config 2 training data.*

Our hybrid model outperforms Whisper-large-v3 fine-tuned on identical data (LibriSpeech: 3.04% vs 5.42%, DrugDB WER: 10.8% vs 13.7%, Medical avg: 8.29% vs 17.95%) while remaining competitive with SOTA models trained on 8-20x more data. This directly validates the architectural advantage of the dual-encoder over a strong fine-tuned encoder-decoder baseline.

5.3 Controlled Comparison (Config 2)

Dataset	NeMo	Hybrid	Rel. Δ
CommonVoice	5.73	5.04	–12.0%
ESLO	31.03	37.33	NeMo ▲
Emilia	12.50	10.69	–14.5%
FLEURS	8.95	7.72	–13.7%
Francais-parle	20.89	15.94	–23.7%
Librispeech	3.34	3.04	–9.0%
MediaSpeech	15.22	14.58	–4.2%
Ofrom	43.45	41.93	–3.5%
Synth-Short	16.96	11.26	–33.6%
Tcof	34.56	36.85	NeMo ▲
Tedx	16.54	17.02	NeMo ▲
Voxpopuli	8.23	6.63	–19.4%
African Acc.	8.38	7.93	–5.4%
Medical (11 avg)	10.19	8.29	–18.6%
Win rate	—	—	**15/18**

Table 3: *Config 2: identical 10k-hour training. Rel.Δ = relative WER reduction.*

Metric	NeMo	Hybrid
Total Char. Edits	15,671	15,418
Char. Score	0.8090	0.8223
Mean Edits/utt.	6.09±12.6	5.99±15.4
WER p-value	—	0.026 ✓

Table 4: *Character-level stability and significance.*

The hybrid model is statistically superior on WER (p=0.026). Character-level improvement (+1.61%) is consistent but not significant (p=0.30), suggesting the dual-encoder excels at word-level linguistic recovery. Medical domain: 18.6% relative reduction (10.19% → 8.29%).

6. ANALYSIS

6.1 Why Dual-Encoder Helps

(i) Decoupling: ASR encoder specializes in acoustics while T5 handles all linguistic processing. **(ii) Intermediate enrichment:** T5's 12 encoder layers add semantic and syntactic structure absent in Whisper-style models. **(iii) Transfer:** LoRA (r=32) preserves billion-token knowledge with only ~4M extra parameters. **(iv) Long-range context:** T5's cross-attention explains gains on longer datasets (Francais-parle: –23.7% rel.). **(v) Cost-quality trade-off:** at RTF = 0.07 on a single GPU L40, the architecture

delivers SOTA-competitive accuracy for monolingual French at inference cost well below commercial cloud ASR APIs — a decisive advantage for high-volume clinical deployment.

6.2 Silence Handling Impact

Without the three-pronged strategy, T5 hallucinated during silence — a fundamental seq2seq limitation. Combining VAD, pure silence training, and token injection eliminates this failure mode and encodes silence as a linguistic unit.

6.3 Where NeMo Wins

ESLO, Tcof, and Tedx share highly spontaneous speech: disfluencies, false starts, overlapping audio. Frame-synchronous CTC/transducer models may be more robust to such irregularities than auto-regressive seq2seq generation.

7. FUTURE WORK

7.1 White Noise Hallucination — Primary Limitation

The principal architectural limitation of our seq2seq design is susceptibility to hallucination under white noise or non-speech audio. When the current binary VAD passes non-speech segments — background noise, equipment hum, or environmental sounds — the T5 decoder generates fluent but entirely fabricated text. This failure mode does not occur with CTC/transducer models and represents the most critical open problem for clinical deployment where ambient noise is unavoidable. The fundamental fix is a more discriminative upstream filter: rather than a binary speech/silence decision, a probabilistic VAD or speech confidence gate must block any audio that is not clearly voiced speech before it reaches the T5 decoder. Planned investigations:

- **Probabilistic VAD:** replace binary VAD with a confidence-scored speech detector; pass to the decoder only frames exceeding a tuned speech probability threshold, transmitting silence tokens otherwise.
- **Acoustic confidence gate:** compute a per-frame signal-to-noise estimate at the FastConformer output; gate decoder cross-attention on frames below a minimum SNR, preventing noise from activating the generative decoder.
- **Auxiliary CTC head:** CTC regularization on the acoustic encoder (ESPnet paradigm) penalizes blank-to-token transitions on noise frames, reducing hallucination risk.
- **Noise augmentation:** systematic SpecAugment + MUSAN/RIR training at SNR −5 to 20 dB and codec artifact simulation to build acoustic robustness.
- **Confidence-based output filtering:** token-level decoder scores to flag and suppress low-confidence spans before downstream clinical NLP pipelines.

7.2 Additional Directions

- **Spontaneous speech:** disfluency-aware objectives (ESLO/Tcof).
- **Streaming:** chunk-based causal decoding for real-time dictation.
- **Multilingual:** mT5 backbone for Arabic and English.
- **Adapter NAS:** architecture search over adapter topology.

8. CONCLUSION

We presented a dual-encoder hybrid ASR-T5 architecture leveraging both encoder and decoder of pre-trained language models — a design empirically motivated by the failure of single-encoder hybrids (50–100% WER, systematic hallucinations). Controlled benchmarking on identical 10k-hour data yields an **83.3% win rate** over our production NeMo baseline, **3.9% val. WER** competitive with SOTA, **18.6% medical domain improvement**, and a clear architectural advantage over Whisper-large-v3 fine-tuned on identical data. At RTF = 0.07 on a single GPU L40, the architecture delivers SOTA-competitive monolingual French accuracy at operational cost well below cloud ASR services. The primary open challenge is white noise hallucination, addressable through probabilistic VAD gating. All experiments used open-source corpora and synthetic speech — no real patient recordings.

Acknowledgements

The authors thank GENCI for access to the Jean Zay HPC cluster (IDRIS/CNRS), where all experiments were run on 8× NVIDIA A100 GPUs.

Ethical Data Statement

All models were trained on open-source speech datasets and internally generated synthetic speech only. No real patient recordings were used.

9. REFERENCES

[1] V. Panayotov et al., Librispeech, ICASSP 2015.
[2] R. Ardila et al., Common Voice, LREC 2020.
[3] A. Conneau et al., FLEURS, IEEE SLT 2023.
[4] C. Raffel et al., T5, JMLR 21(140), 2020.
[5] E. J. Hu et al., LoRA, ICLR 2022.
[6] A. Gulati et al., Conformer, Interspeech 2020.
[7] S. Rekesh et al., FastConformer, IEEE ASRU 2023.
[8] S. Rekesh et al., Canary, arXiv:2406.19674, 2024.
[9] X. Liu et al., Voxtral, arXiv:2501.04184, 2025.
[10] N. Kanda et al., NeMo, Interspeech 2021.
[11] A. Radford et al., Whisper, ICML 2023.
[12] C. Wang et al., VoxPopuli, ACL 2021.
[13] W. Chan et al., Listen Attend and Spell, ICASSP 2016.
[14] F. Seide et al., Decoder-only ASR, Interspeech 2024.
[15] error-align, PyPI v0.1.0b8, 2024.