Wednesday, April 29, 2026

TReC 2026 Project Proposal Submission Form

Submit your project proposal for the 7th TRAIL Research Camp (August 24th - September 4th, 2026, Lausanne, Switzerland). Please complete all required sections and submit your proposal before April 30th, 01:00 PM (CET).

Administrative Data

Full Name of Team Leader Olivier Ratenon

Contact Email Olivier.RATENON@umons.ac.be

Project Information

Project Title When the Model Should Hesitate: Uncertainty-Aware ASR/NLP for Early Language Disorder Screening in Children

Profile of the Team Leader(s) & Expected Team Composition

Olivier Ratenon is an AI PhD researcher at UMONS, working on the WEAVE project to analyze early language development through natural speech measures. Bridging STEM and SSH, he holds a Linguistics degree, an NLP Master's, and an ML certification. Previously an Applied AI Researcher at Multitel, he specialized in Trustworthy AI and uncertainty quantification. This dual background perfectly positions him to lead the NLP components of this clinical speech processing project.

For this project, we are looking to assemble a small multidisciplinary team able to cover the full chain from audio processing to clinically meaningful interpretation. The ideal team would include:

- A speech/ASR profile with experience in end-to-end speech recognition and acoustic perturbation.
- A deep learning profile with expertise in uncertainty quantification, calibration and selective prediction.
- A NLP or computational linguistics profile able to define and implement transcript-derived language measures.
- If possible, a speech-language pathology, developmental linguistics, or digital health profile to help ground the technical choices in clinically meaningful screening needs.

The goal is to combine strong technical implementation with domain relevance, so that the camp produces not only a working proof of concept, but also a credible roadmap toward later validation with real pediatric clinical data.

Have you already identified potential team members for your project? Yes

List the team members you have identified and briefly describe their profiles/roles (e.g., expertise, affiliation, expected contribution).

Bertrand Braeckeveldt specializes in the calibration and uncertainty quantification of deep learning

architectures at Multitel. His contributions include the development of a modular library designed to streamline uncertainty estimation across classification, regression, and spatial tasks like depth estimation. His research emphasizes selective classification frameworks, aimed at improving the decision-making integrity of models in uncertain environments.

Domain of Application Healthcare

Scientific Theme Trustworthy AI

Proposal Content

Abstract

This project aims to develop an uncertainty-aware Automatic Phoneme Recognition (APR) pipeline for early language-disorder screening. Standard Automatic Speech Recognition (ASR) systems and Generative Error Correction (GER) pipelines map acoustic inputs to normative semantic text, actively erasing the atypical phonetic substitutions and elisions that serve as primary clinical biomarkers. To preserve this diagnostic value, we want to reframe the task strictly as phoneme-level sequence classification. By using the encoder portion of self-supervised acoustic models with a Connectionist Temporal Classification (CTC) objective, we will extract unadulterated phonetic representations. Then by applying parameter-efficient Uncertainty Quantification (UQ) methodologies to output calibrated confidence scores, the system will be able to safely flag ambiguous acoustic segments for human review rather than blindly auto-correcting itself. Establishing this proof of concept on English proxy data will create a robust, reusable pipeline ready for future clinical datasets.

Background Information & Problem Statement

Children's speech significantly deviates from adult acoustic training data, inducing a severe distribution shift [1]. In early language-disorder screening, clinical reasoning relies on fine-grained articulatory and phonological deviations. Unfortunately, standard ASR systems optimize for fluent orthographic transcripts, often hiding exactly the segmental deviations that matter clinically. Modern techniques attempting to improve robustness, such as GER [2], exacerbate this by using Large Language Models (LLMs) to map noisy hypotheses to standard linguistic norms. GER effectively "auto-corrects" clinical mispronunciations, destroying their diagnostic value.

Our project pivots from error correction to uncertainty quantification. To prevent semantic auto-correction, we strictly isolate the acoustic pipeline from linguistic priors. We discard autoregressive generative decoders, relying exclusively on the encoder portion of self-supervised models (e.g., Whisper encoder, Wav2Vec2, HuBERT) [3], [4], [5]. Decoders use learned language structures to predict sequences, which inherently biases the output toward standard speech. Instead, we use a linear classification head optimized with a CTC objective Connectionist temporal classification mathematically aligns unsegmented acoustic frames directly to phonemes without manual boundary annotations [6]. By relying purely on the encoder and CTC, we hope to guarantee the model transcribes exactly what it hears, preserving raw acoustic evidence of speech disorders [7]. Because we frame this as a classification problem rather than generative modeling, the uncertainty layer will no longer act at the word-transcript level, but directly at the phoneme-classifier level.

After decoding, calibrated posterior distributions will be used to compute frame-wise and segment-level uncertainty scores, primarily predictive entropy over phoneme posteriors. This allows us to use state-of-the-art UQ methods adapted for classification. We will focus on LoRA-based methods, adapted to transformer architectures. It includes LoRA-Ensemble [8] Laplace LoRA [9], SWAG and B-LORA-XS [10]. Recent conformal methods that provide statistical guarantees will also be considered in the benchmark [11]. The pipeline validation requires controlled shifted dataset testing. The first regime will use English public child-language corpora

from CHILDES (TalkBank) [12], with priority given to phonology-oriented resources.

Crucially, a second regime will use public adult speech transformed to approximate child-like and ecologically mismatched recording conditions through noise, reverberation, pitch and speed perturbation, and related acoustic mismatch. This robust dual-regime strategy allows us to systematically benchmark whether calibrated phoneme-level uncertainty correlates with actual model failure, ensuring the system safely defers to clinicians when operating outside its competence zone [13].

Project Objectives & Concrete Implementation

The primary objective of this two-week research camp is to develop, implement, and validate a Trustworthy Artificial Intelligence (TrustAI) phoneme recognizer capable of quantifying its own uncertainty on pediatric speech. To preserve raw acoustic evidence of childhood speech disorders, we must strictly isolate the acoustic pipeline from linguistic priors. Standard generative error correction methods actively overwrite these phonetic anomalies. Therefore, we discard GER and frame automatic phoneme recognition as a strict classification problem, allowing us to evaluate state-of-the-art uncertainty quantification methods. The project will be executed through four concrete technical objectives.

End-to-end audio-to-phoneme classification baseline

We will leverage the encoder portion of pre-trained self-supervised acoustic models, such as Wav2Vec 2.0, HuBERT, or the Whisper encoder. We will deliberately discard auto-regressive generative decoders to prevent the model from using learned language structures to auto-correct the output. Instead, we will fine-tune a lightweight linear classification head on top of the encoder using a connectionist temporal classification loss. By relying purely on the acoustic encoder and a CTC head, we should guarantee that the model transcribes exactly what it hears, preserving clinical mispronunciations rather than forcing semantic coherence.

Computationally-efficient uncertainty modeling

Because our architecture reduces automatic phoneme recognition to a sequence-classification problem, we can directly apply advanced UQ methods designed for classification. Standard CTC models output overconfident softmax probabilities. To capture epistemic uncertainty effectively under the severe pediatric distribution shift, we will implement and compare several parameter-efficient UQ techniques. Specifically, we will consider Low-Rank Adaptation (LoRA) techniques. Low-Rank Adaptation is a fine-tuning mechanism that freezes the pre-trained model weights and injects small, trainable rank-decomposition matrices into the Transformer architecture [14]. It exists to drastically reduce the computational and memory footprint of adapting large foundation models. We use LoRA because it makes complex Bayesian inference and ensembling mathematically tractable. It allows us to quantify model uncertainty without requiring the prohibitive memory overhead of hosting multiple full copies of the heavy backbone. In this family of methods, we will target and compare:

- LoRA-Ensemble [8] by training multiple independent, lightweight LoRA modules within a single frozen backbone to simulate a deep ensemble, capturing predictive variance with minimal memory overhead.
- Laplace LoRA [9] by applying a post-hoc Laplace approximation to the converged LoRA weights, estimating a Gaussian posterior distribution around the local optimum to capture weight uncertainty analytically.
- SWAG LoRA [10] by tracking the trajectory of LoRA parameters during the training phase to construct a Stochastic Weight Averaging Gaussian (SWAG) [15] approximation of the

posterior, enabling highly efficient stochastic sampling.
- B-LORA-XS [10], a Bayesian adaptation that places priors on the low-rank matrices to infer their posterior distributions, designed specifically for robust epistemic UQ in large Transformer architectures.

These methods will be integrated into the Transformer blocks and the final CTC projection layer. This enables the extraction of frame-wise and segment-level uncertainty scores—primarily predictive entropy over phoneme posteriors—without the prohibitive memory overhead of full deep ensembles

Post-hoc Calibration and Conformal Error Bounding

Raw neural confidence is often miscalibrated, especially under distribution shifts. We will implement post-hoc calibration, notably temperature scaling [16], on a held-out validation split to improve Expected Calibration Error (ECE). Furthermore, we will apply Inductive Conformal Prediction (ICP) [11] to the frame-level phoneme logits. ICP guarantees a user-defined maximum error rate by converting uncalibrated point estimates into statistically valid prediction sets. For instance, if a child's distorted production of a phoneme is highly ambiguous, the model will output a set of probable phonemes and flag the segment for human review rather than forcing an arbitrary, overconfident guess.

Benchmarking on English Proxy Data

To validate the pipeline during the two-week camp, we will utilize English proxy datasets. Establishing this proof of concept in English ensures a stable technical baseline and avoids immediate data bottleneck issues. To capture uncertainty, we will extract the predictive entropy computed from the posterior predictive distribution over phonemes. For calibration quality, we will rely on standard metrics such as the Expected Calibration Error (ECE) which measures the alignment between confidence and empirical accuracy [16]. We will also rely on the Brier score, which evaluates the mean squared error of the probabilistic forecasts.

To assess the system's ability to flag incorrect predictions globally, we will formulate error detection as a binary classification task [17]. Based on the resulting confusion matrices indicating whether UQ methods correctly flag erroneous predictions, we will compute the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC). Finally, we will evaluate selective prediction through risk-coverage curves [18], plotting the error rate as predictions are progressively discarded starting with the most uncertain. We will compare our UQ methods against an oracle baseline that perfectly rejects true errors first.

Do you plan to deliver, as an outcome of your project, a reusable "brick" for the TRAIL Factory (https://factory.trail.ac/en/home_page) that could later be transferred and converted into a company process?

Yes

Briefly describe what the brick would be and its intended users.

The brick will be a reproducible codebase with a demo that incorporates the following considerations:

- Input: Audio file upload/recording.

- Processing: Execution of the audio through the trained CTC baseline coupled with the selected UQ method.
- Output: A phoneme-level transcript featuring per-phoneme uncertainty scores. The interface will color-code outputs by confidence (e.g., flagging high-uncertainty segments in red) and feature a method selector to compare different UQ outputs.

Project Dataset

Because clinical recordings are not yet available, the project will use a dual proxy-data strategy.

The first dataset regime will rely on public child-language corpora from CHILDES [12], which provide authentic child speech and interaction data suitable for technical benchmarking. These resources are particularly relevant because they offer real child-language material and associated annotations that can support proof-of-concept evaluation.

The second regime will rely on public adult speech recordings that will be transformed to emulate child-like or ecologically mismatched conditions. Planned perturbations include additive household noise, reverberation, speed and pitch modification, and related acoustic mismatch. The purpose is not to simulate clinical reality perfectly, but to generate controlled distribution shift conditions under which the reliability layer can be evaluated systematically.

From a data-format perspective, the project expects audio files in standard waveform formats and transcript or annotation files in corpus-standard text/alignment formats, depending on the selected subsets. Metadata of interest include speaker identity when available, age group when available, recording condition, transcript quality, and perturbation regime.

This proxy-data strategy is appropriate for a two-week proof of concept because the main objective is to validate the uncertainty and calibration layer under realistic mismatch, not yet to estimate final clinical sensitivity or specificity on the target cohort. It also ensures that the work remains reproducible and shareable within the camp framework.

Detailed Work Plan

The work plan is structured around three parallel tracks corresponding to the team's multidisciplinary profiles: Automatic Speech Recognition (ASR), Uncertainty Quantification (UQ), and Natural Language Processing (NLP). To ensure seamless collaboration during the intensive two-week camp, dataset curation and the core pipeline architecture will be established prior to the workshop.

Pre-Workshop Preparation

- Select and download CHILDES sub-corpora, prioritizing phonologically annotated subsets.
- Prepare the perturbation scripts for the adult speech regime (e.g., additive noise, reverberation, pitch/speed shifts) to simulate out-of-distribution conditions.
- Implement the abstract CTC training loop, consisting of the frozen acoustic encoder, linear classification head, and CTC loss.
- Standardize audio formats and annotation files across all datasets.
- Set up the shared repository, experiment tracking, and compute environment.

Baseline Establishment

The first phase aims to deliver a trained CTC baseline and a functional LoRA adapter layer, ready for advanced UQ method integration.

- Day 1 (Whole team): Review the pre-built pipeline, proxy datasets, and evaluation metrics together.
- Days 1–3 (ASR team): Train the CTC baseline on the CHILDES dataset using the pre-built loop and validate the Phoneme Error Rate (PER) on held-out data.
- Days 1–3 (UQ team): Integrate the base LoRA adapters into the Transformer encoder architecture.
- Days 1–3 (NLP team): Define the clinically relevant phoneme error taxonomy and implement transcript-level phoneme scoring functions.

Initial UQ Implementation

The second phase focuses on producing calibrated, phoneme-level uncertainty scores using the LoRA-Ensemble approach.

- Days 4–5 (UQ team): Implement LoRA-Ensemble by training multiple independent LoRA modules within the frozen backbone to extract per-phoneme predictive entropy.
- Days 4–5 (ASR team): Process the perturbed adult speech regime through the trained baseline, validating that the distribution shifts are successfully detected by the uncertainty scores.
- Days 4–5 (NLP team) Extend phoneme scoring to support segment-level uncertainty annotation and link these scores to the error taxonomy for downstream failure analysis.

Calibration and Benchmarking

The third phase is dedicated to calibration and the comprehensive benchmarking of all targeted UQ methodologies.

- Days 6–7 (UQ Team): Implement the training pipelines for B-LORA-XS, Laplace LoRA, and SWAG LoRA.
- Days 8–9 (UQ Team): Apply temperature scaling and Inductive Conformal Prediction (ICP) on the held-out validation split to ensure rigorous error control.
- Days 6–7 (ASR and NLP teams): Run the complete pipeline on both the CHILDES and perturbed adult regimes, actively flagging segments that exceed the defined uncertainty thresholds.
- Days 8–9 (ASR and NLP teams): Conduct a comprehensive qualitative and quantitative failure analysis, classifying the flagged segments by error type using the established taxonomy.

Packaging and Demonstrator

The final phase focuses on packaging the deliverables, creating the final presentation, and deploying a reproducible software demonstrator.

- Days 9–10 (Whole Team): Build an interactive Streamlit/Dash application that acts as a reusable TRAIL Factory brick.

Bibliographic References

[1] S. Lee, A. Potamianos, and S. Narayanan, “Analysis of children’s speech: duration, pitch and formants,” in *Eurospeech*, 1997, pp. 473–476.

[2] C.-H. H. Yang, Y. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke, “Generative speech recognition error correction with large language models and task-activating prompting,” in *ASRU*, 2023, pp. 1–8.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *ICML*, 2023.

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *NeurIPS*, 2020.

[5] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *TASLP*, vol. 29, pp. 3451–3460, 2021.

[6] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *ICML*, 2006, pp. 369–376.

[7] E. Lyakso, O. Frolova, and A. Grigorev, “A comparison of acoustic features of speech of typically developing children and children with autism spectrum disorders,” in *SPECOM*, 2016, pp. 43–50.

[8] D. J. Mühlematter *et al.*, “LoRA-ensemble: Efficient uncertainty modelling for self-attention networks,” *TMLR*, 2026.

[9] A. X. Yang, M. Robeyns, X. Wang, and L. Aitchison, “Bayesian Low-rank Adaptation for Large Language Models,” in *ICLR*, 2023.

[10] P. Marszałek, K. Bałazy, J. Tabor, and T. Kuśmierczyk, “Minimal ranks, maximum confidence: Parameter-efficient uncertainty quantification for LoRA,” in *EMNLP*, 2025, pp. 1260–1271.

[11] F. Ernez, A. Arnold, A. Galametz, C. Kobus, and N. Ould-Amer, “Applying the conformal prediction paradigm for the uncertainty quantification of an end-to-end automatic speech recognition model (wav2vec 2.0),” in *COPA*, 2023, pp. 16–35.

[12] B. MacWhinney, *The CHILDES project: Tools for analyzing talk*, 3rd ed. Lawrence Erlbaum Associates, 2000.

[13] E. Begoli, T. Bhattacharya, and D. Kusnezov, “The need for uncertainty quantification in machine-assisted medical decision making,” *Nat. Mach. Intell.*, vol. 1, no. 1, pp. 20–23, 2019.

[14] E. J. Hu *et al.*, “LoRA: Low-Rank Adaptation of Large Language Models,” in *ICLR*, 2021.

[15] W. J. Maddox, T. Garipov, P. Izmailov, D. Vetrov, and A. G. Wilson, “A simple baseline for bayesian uncertainty in deep learning,” in *NeurIPS*, 2019.

[16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *ICML*, 2017, pp. 1321–1330.

[17] J. Gawlikowski *et al.*, “A survey of uncertainty in deep neural networks,” *Artif. Intell. Rev.*, vol. 56, pp. 1513–1589, 2023.

[18] Y. Geifman and R. El-Yaniv, “Selective classification for deep neural networks,” in *NeurIPS*,

2017, pp. 4885–4894.

Eligibility & Evaluation

Does the project include multidisciplinarity between STEM & SSH?

Yes

How?

The project combines speech processing, machine learning, uncertainty quantification, and NLP with speech-language pathology, developmental linguistics, and digital health considerations. The SSH dimension is essential because clinically meaningful interpretation and acceptable review thresholds cannot be defined from a purely technical perspective.

We confirm that the Team Leader will be present for the full duration of TReC'26 if the project is selected (August 24th - September 4th, 2026, Lausanne, Switzerland)

I/We agree and confirm