الجمـهورية الجزائرية الديمقراطية الشعبيـة
وزارة التعليم العالي والبحث العلمي
جـامعـة سعــيـدة د. مـولاي الطـاهـر
كلية الرياضيات و الإعلام الآلي و
الاتصالات السلكية و اللاسلكية
قسم: الإعلام الآلي

Mémoire de Master en informatique

Spécialité : Modélisation Informatique des Connaissances et du Raisonnement

T h è m e

Assistant IA pour l'amélioration orale en anglais
Plateforme EvalEnglish

- **Présenté par :**

Hamri Malika Firdaws

- **Dirigé par :**

M. Ahmed Zahaf

Année universitaire 2025-2026

Dédicace

---

ma très chère mère,
qui m'a offert son amour inconditionnel, sa patience et son soutien
tout au long de ce parcours.
Tu es ma force et mon inspiration.

mes enseignants,
en particulier mon encadrant ***M. Ahmed Zahaf****,*
pour sa guidance précieuse, sa disponibilité et ses conseils éclairés
qui ont rendu ce travail possible.

mes chers amis,
qui ont partagé avec moi les moments de doute et de joie,
et qui ont toujours cru en moi même quand je n'y croyais plus.

toute personne qui un jour a rêvé de créer quelque chose
et qui a eu le courage de commencer.

— Hamri Malika Firdaws

Remerciements

Je tiens à exprimer ma sincère gratitude à toutes les personnes qui ont contribué, de près ou de loin, à l'élaboration de ce mémoire.

Mes premiers remerciements vont à mon encadrant, **Monsieur Ahmed Zahaf**, pour avoir accepté de diriger ce travail, pour ses orientations pertinentes, sa rigueur scientifique et sa disponibilité tout au long de la réalisation de ce projet.

Je remercie également tous les **enseignants du département d'Informatique** de l'Université de Saïda – Dr. Moulay Tahar, pour la qualité de la formation dispensée durant les années de Master, et pour les connaissances transmises avec générosité et dévouement.

Un remerciement tout particulier à **ma mère**, pour son amour sans limite, son soutien moral infaillible et ses encouragements constants qui ont été pour moi une source d'énergie inépuisable.

Je remercie chaleureusement mes **amis et camarades de promotion**, pour les échanges d'idées, l'entraide mutuelle et les moments de complicité qui ont rendu ce parcours universitaire inoubliable.

Enfin, je remercie **les membres du jury** pour l'honneur qu'ils me font en acceptant d'évaluer ce travail et pour le temps précieux qu'ils y consacrent.

Résumé

Ce mémoire présente **EvalEnglish**, une plateforme web intelligente dédiée à l'évaluation automatique des compétences orales en langue anglaise. Face à la problématique du manque d'outils accessibles et objectifs pour évaluer la production orale selon le référentiel européen CEFR (A1–C2), nous proposons une solution basée sur cinq modèles d'intelligence artificielle complémentaires : deux pipelines en cascade (wav2vec2 → LLaMA et Whisper → LLaMA), deux modèles de type Speech-LLM (Qwen2-Audio en zéro-shot et fine-tuné), et un modèle de prononciation par GOP (Goodness-of-Pronunciation).

Le modèle Qwen2-Audio fine-tuné avec QLoRA sur le corpus SAN-DI/Speak&Improve 2025 atteint un PCC de 0,393 et un SRC de 0,462, doublant les corrélations du modèle zéro-shot. La plateforme intègre un test CEFR oral guidé en 3 parties, une bibliothèque de 18 cours A1–C2, un tableau de bord apprenant et un espace enseignant.

Mots-clés : *évaluation orale automatique, CEFR, LLM, Qwen2-Audio, QLoRA, fine-tuning, wav2vec2, Whisper, prononciation, plateforme e-learning.*

Abstract

This thesis presents **EvalEnglish**, an intelligent web platform for the automatic assessment of English oral skills. To address the lack of accessible and objective tools for evaluating speaking performance according to the CEFR framework (A1–C2), we propose a solution based on five complementary AI models: two cascade pipelines (wav2vec2 → LLaMA and Whisper → LLaMA), two Speech-LLM models (Qwen2-Audio zero-shot and fine-tuned), and a GOP pronunciation model.

The Qwen2-Audio model fine-tuned with QLoRA on the SANDI/Speak&Improve 2025 corpus achieves a PCC of 0.393 and an SRC of 0.462, doubling the zero-shot correlations. The platform integrates a 3-part guided oral CEFR test, a library of 18 A1–C2 courses, a learner dashboard, and a teacher workspace.

Keywords: *automatic oral assessment, CEFR, LLM, Qwen2-Audio, QLoRA, fine-tuning, wav2vec2, Whisper, pronunciation, e-learning platform.*

ملخّص

تقدّم هذه المذكّرة منصّة EvalEnglish، وهي منصّة وِب ذكيّة مخصّصة للتقييم الآلي لمهارات التعبير الشفهي في اللغة الإنجليزية. ولمعالجة نقص الأدوات المتاحة والموضوعية لتقييم الأداء الشفهي وفق الإطار الأوروبي المرجعي المشترك للّغات CEFR (من المستوى A1 إلى C2)، نقترح حلًّا يقوم على خمسة نماذج ذكاء اصطناعي متكاملة: مسارَين تعاقبيّين (wav2vec2 وLLaMA، ثمّ Whisper وLLaMA)، ونموذجَين من نوع Speech-LLM (نموذج Qwen2-Audio في وضع zero-shot وفي وضع التحسين الدقيق)، ونموذج لتقييم النطق يعتمد تقنية GOP.

يحقّق نموذج Qwen2-Audio بعد تحسينه الدقيق بتقنية QLoRA على متن SANDI/Speak&Improve 2025، معامل ارتباط بيرسون (PCC) قدره 0.393 ومعامل ارتباط سبيرمان (SRC) قدره 0.462، أي ما يقارب ضعف نتائج النموذج في وضع zero-shot. تدمج المنصّة اختبارًا شفهيًّا موجّهًا من ثلاثة أجزاء، ومكتبةً تضمّ 18 دورة من A1 إلى C2، ولوحة تحكّم للمتعلّم، وفضاءً مخصّصًا للأستاذ.

الكلمات المفتاحية: التقييم الشفهي الآلي، CEFR، LLM، Qwen2-Audio، QLoRA، التحسين الدقيق، wav2vec2، Whisper، تقييم النطق، منصّة التعلّم الإلكتروني.

Hamri Malika Firdaws

Saïda, 2026

Contents

List of Figures

List of Tables

Introduction Générale

L'Intelligence Artificielle et l'éducation

L'intelligence artificielle (IA) représente aujourd'hui l'une des révolutions technologiques les plus profondes de notre époque. Depuis les premières formalisations théoriques d'Alan Turing dans les années 1950 jusqu'aux systèmes modernes capables de générer du texte, d'analyser des images ou de comprendre la parole humaine, l'IA a connu une progression exponentielle qui transforme en profondeur tous les secteurs de la société [1].

Le domaine de l'éducation n'échappe pas à cette transformation. On parle désormais d'*Intelligence Artificielle en ducation* (AIED – Artificial Intelligence in Education), un champ interdisciplinaire qui explore comment les systèmes intelligents peuvent améliorer les processus d'enseignement et d'apprentissage. Les systèmes tutoriels intelligents (*Intelligent Tutoring Systems*), les environnements d'apprentissage adaptatifs, les outils de correction automatique et les plateformes d'évaluation constituent autant d'applications concrètes qui redéfinissent la relation entre l'apprenant et le savoir [2].

L'IA au service de l'apprentissage des langues

L'apprentissage des langues étrangères constitue l'un des domaines où l'IA a trouvé ses applications les plus emblématiques. Des plateformes comme Duolingo, Busuu ou Babbel ont démontré que l'apprentissage gamifié et personnalisé, soutenu par des algorithmes adaptatifs, pouvait engager des millions d'apprenants à travers le monde [3]. Cependant, malgré ces avancées, l'évaluation de la production orale – la compétence la plus difficile à automatiser – reste largement sous-exploitée.

Parler une langue est fondamentalement différent de la lire ou de l'écrire. La production orale mobilise simultanément la phonologie, la syntaxe, le vocabulaire, la fluidité et la cohérence discursive. Or, évaluer cette compétence de façon objective, instantanée et à grande échelle constitue un défi technique et pédagogique majeur. Les examens oraux traditionnels (IELTS, TOEFL, DELF) requièrent des examinateurs humains formés, ce qui engendre des coûts élevés, des délais importants et une inévitable subjectivité [4].

Contexte et motivation du projet

C'est dans ce contexte que s'inscrit notre projet de fin d'études : **EvalEnglish**, une plateforme web intelligente d'évaluation des compétences orales en anglais. Notre motivation principale découle d'un constat simple mais éloquent : des millions d'apprenants algériens et francophones souhaitent améliorer leur anglais oral, mais n'ont pas accès à des outils fiables et abordables pour connaître leur niveau réel et obtenir des retours détaillés sur leur production.

Les grands modèles de langage (*Large Language Models* – LLMs) et les modèles de traitement de la parole ont atteint un niveau de maturité technologique suffisant pour aborder ce défi. En combinant des modèles de transcription automatique (wav2vec2, Whisper), des modèles de langage de grande taille (LLaMA) et des Speech-LLMs multimodaux (Qwen2-Audio), il devient possible de construire un système d'évaluation automatique robuste, transparent et accessible.

Structure du mémoire

Ce mémoire est organisé en quatre chapitres principaux :

- **Chapitre 1 – Fondements théoriques des LLMs :** présente les bases du deep learning, les techniques de représentation du langage,

l'architecture Transformer et les différentes familles de grands modèles de langage, y compris les Speech-LLMs.

- **Chapitre 2 – état de l'art :** dresse un panorama des travaux existants en évaluation orale automatique, des référentiels linguistiques comme le CEFR, et des systèmes concurrents (IELTS, TOEFL, ELSA, etc.).
- **Chapitre 3 – Conception et méthodologie :** détaille l'architecture technique de la solution, les cinq modèles déployés, le processus de fine-tuning et les résultats expérimentaux obtenus.
- **Chapitre 4 – Réalisation et présentation de la plateforme :** présente l'implémentation concrète de la plateforme EvalEnglish, ses fonctionnalités et son interface utilisateur.

Chapter 1

Fondements Théoriques des Grands Modèles de Langage (LLMs)

Introduction du chapitre

Ce premier chapitre pose les bases théoriques indispensables à la compréhension de notre travail. Nous y explorons les concepts fondamentaux du deep learning, les techniques de représentation du langage, l'architecture révolutionnaire des Transformers, et les différentes familles de modèles de langage qui constituent le socle technologique de la plateforme EvalEnglish.

1.1 Deep Learning

1.1.1 Définition et principes fondamentaux

Le *deep learning* (apprentissage profond) est une sous-branche de l'apprentissage automatique (*machine learning*) qui repose sur l'utilisation de réseaux de neurones artificiels à plusieurs couches cachées. Inspiré du fonctionnement du cerveau humain, il permet aux machines d'apprendre des représentations hiérarchiques des données de manière automatique, sans nécessiter d'extraction manuelle de caractéristiques [5].

Un réseau de neurones profond est composé de :

- Une **couche d'entrée** qui reçoit les données brutes (texte, image, son)
- Plusieurs **couches cachées** qui transforment progressivement les représentations

- Une **couche de sortie** qui produit la prédiction finale

Mathématiquement, chaque couche l effectue la transformation :

$$\mathbf{h}^{(l)} = f\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right) \tag{1.1}$$

où $\mathbf{W}^{(l)}$ est la matrice de poids, $\mathbf{b}^{(l)}$ le vecteur biais et f la fonction d'activation (ReLU, sigmoid, tanh, etc.).

1.1.2 L'apprentissage par rétropropagation

L'entraînement des réseaux de neurones profonds repose sur l'algorithme de *rétropropagation du gradient* (*backpropagation*). L'objectif est de minimiser une fonction de perte $\mathcal{L}$ en ajustant iterativement les paramètres du réseau dans la direction opposée au gradient [6] :

$$\mathbf{W} \leftarrow \mathbf{W} - \eta \nabla_{\mathbf{W}} \mathcal{L} \tag{1.2}$$

où η est le taux d'apprentissage (*learning rate*).

1.1.3 Architectures clés du deep learning

Plusieurs architectures ont marqué l'évolution du deep learning :

Table 1.1: Principales architectures de deep learning

Architecture	**Application principale**	**Année**
Perceptron multicouche (MLP)	Classification générale	1986
Réseau convolutif (CNN)	Vision par ordinateur	1998
Réseau récurrent (RNN/LSTM)	Séquences temporelles, NLP	2014
Transformer	NLP, vision, audio	2017
Modèles diffusion	Génération d'images	2020

1.2 Représentation du langage : la vectorisation des mots

1.2.1 Le problème de la représentation

Les algorithmes d'apprentissage automatique ne peuvent traiter que des nombres. Pour appliquer le deep learning au texte, il est donc nécessaire de convertir les mots en vecteurs numériques. Deux approches classiques ont précédé les méthodes modernes : le *Bag of Words* et le *One-Hot Encoding*.

1.2.2 Bag of Words (BoW)

Le modèle *Bag of Words* représente un document comme un vecteur de fréquences de mots, sans tenir compte de leur ordre[7]. Soit un vocabulaire $V = \{w_1, w_2, \ldots, w_n\}$, un document d est représenté par :

$$\mathbf{v}(d) = [tf(w_1, d),\ tf(w_2, d),\ \ldots,\ tf(w_n, d)] \tag{1.3}$$

où $tf(w_i, d)$ est la fréquence du terme w_i dans le document d.

Avantages : simple, interprétable, efficace pour la classification de textes courts.

Limites : ne capture pas le contexte ni la sémantique ; vecteurs creux de grande dimension.

1.2.3 One-Hot Encoding

Le *One-Hot Encoding* représente chaque mot par un vecteur binaire de taille $|V|$ où seule la position correspondant au mot est à 1, toutes les autres étant à 0.

$$\text{chat} \rightarrow [0, 0, 1, 0, 0, \ldots, 0] \tag{1.4}$$

Limites majeures : aucune information sémantique ("chat" et "chien" sont aussi différents que "chat" et "ordinateur") ; dimensionnalité très élevée ; aucune notion de similarité entre mots.

1.3 Word Embedding

1.3.1 Principe des embeddings

Les *word embeddings* (plongements lexicaux) sont des représentations vectorielles denses qui encodent la sémantique des mots dans un espace de faible dimension. Contrairement au One-Hot Encoding, deux mots sémantiquement proches auront des vecteurs proches dans cet espace [8].

La propriété fondamentale des embeddings est illustrée par l'analogie célèbre :

$$\overrightarrow{\text{roi}} - \overrightarrow{\text{homme}} + \overrightarrow{\text{femme}} \approx \overrightarrow{\text{reine}} \tag{1.5}$$

1.3.2 Word2Vec

Proposé par Mikolov et al. [8] en 2013, Word2Vec est un modèle neuronal entraîné à prédire :

- **CBOW** (*Continuous Bag of Words*) : prédire un mot à partir de son contexte
- **Skip-gram** : prédire le contexte à partir d'un mot

1.3.3 GloVe et FastText

GloVe (Global Vectors for Word Representation) [9] exploite les statistiques de co-occurrence globale du corpus pour apprendre des embeddings.

FastText [10] développé par Facebook, enrichit les embeddings par une décomposition en sous-mots (*subword*), ce qui permet de mieux traiter les mots rares et les néologismes.

1.4 L'architecture Transformer

1.4.1 Motivations et contexte

L'architecture **Transformer**, introduite par Vaswani et al. en 2017 dans l'article fondateur *"Attention Is All You Need"* [11] a provoqué une révolution dans le traitement du langage naturel. Elle surmonte les limites des RNNs – notamment l'incapacité à paralléliser l'entraînement et le problème des

gradients qui disparaissent sur de longues séquences – en remplaçant la récurrence par un mécanisme d'**attention**.

1.4.2 L'attention multi-têtes

Le mécanisme d'attention calcule une représentation pondérée de la séquence en fonction de la pertinence relative de chaque élément. Pour chaque position, on calcule une *Query* Q, une *Key* K et une *Value* V :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1.6}$$

où d_k est la dimension des clés. Le facteur $\sqrt{d_k}$ évite les gradients trop faibles lors du softmax.

L'attention **multi-têtes** exécute h têtes d'attention en parallèle, chacune capturant des relations différentes :

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{1.7}$$

1.4.3 Architecture complète

Le Transformer est composé d'un **encodeur** et d'un **décodeur**, chacun constitué de blocs empilés comprenant :

1. Une couche d'attention multi-têtes (auto-attention ou cross-attention)
2. Une normalisation de couche (*Layer Normalization*)
3. Un réseau *feed-forward* à deux couches
4. Des connexions résiduelles (*residual connections*)

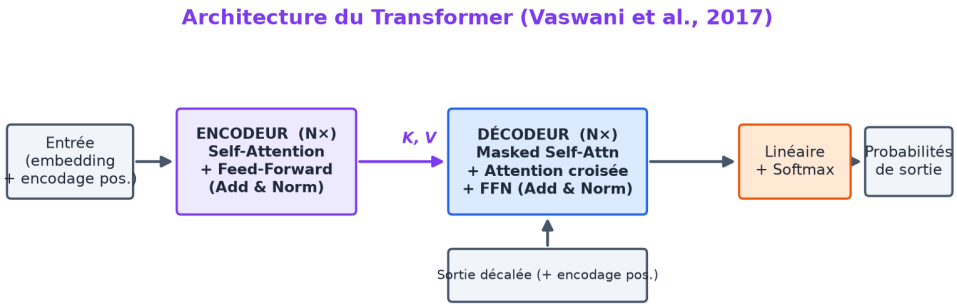

Figure 1.1: Architecture générale du Transformer [11]

1.4.4 Encodage positionnel

Contrairement aux RNNs qui traitent les séquences dans l'ordre, le Transformer traite tous les tokens simultanément. Il utilise donc un **encodage positionnel** pour injecter l'information de position :

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{1.8}$$

1.5 Modèles de Langage Masqués (Masked Language Models)

1.5.1 BERT : Bidirectional Encoder Representations from Transformers

Introduit par Devlin et al. [12]en 2018, **BERT** est un modèle de langage pré-entraîné basé sur l'encodeur du Transformer. Il utilise deux tâches de pré-entraînement originales :

1. **Modélisation du langage masqué (MLM) :**

15% des tokens de l'entrée sont masqués aléatoirement, et le modèle doit les prédire. Contrairement aux modèles auto-régressifs, BERT considère le contexte *bidirectionnel* (gauche et droite) :

$$P(\mathbf{x}_{mask} \mid \mathbf{x}_{\setminus mask}) = \prod_{i \in M} P(x_i \mid \mathbf{x}_{\setminus mask}) \tag{1.9}$$

2. **Prédiction de la phrase suivante (NSP) :**

Le modèle apprend à prédire si deux phrases sont consécutives dans le corpus.

1.5.2 Variantes de BERT

Table 1.2: Variantes notables de BERT

Modèle	Innovation	Paramètres
BERT-base	Modèle original	110M
RoBERTa	Entraînement plus long, sans NSP	125M
DistilBERT	Version distillée (plus rapide)	66M
CamemBERT	BERT pour le français	110M
wav2vec 2.0	BERT pour la parole	317M

1.5.3 wav2vec 2.0 : BERT pour la parole

wav2vec 2.0 [13]est l'adaptation du paradigme MLM au domaine de la parole. Il encode le signal audio brut en représentations latentes, masque certaines d'entre elles, et s'entraîne à les reconstruire. Cette approche permet d'apprendre des représentations riches de la parole sans avoir besoin de grandes quantités de données annotées.

Dans notre plateforme, wav2vec 2.0 est utilisé comme modèle de transcription automatique dans le pipeline en cascade (Volet 1).

1.6 Modèles de Langage Causaux (Causal Language Models)

1.6.1 Principe de la modélisation causale

Contrairement aux MLMs qui voient le contexte dans les deux directions, les **modèles causaux** (aussi appelés auto-régressifs) prédisent chaque token en fonction uniquement des tokens précédents :

$$P(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} P(x_t \mid x_1, x_2, \ldots, x_{t-1}) \tag{1.10}$$

Cette propriété les rend naturellement adaptés à la **génération de texte**.

1.6.2 GPT et ses successeurs

La famille **GPT** (*Generative Pre-trained Transformer*), développée par OpenAI, a progressivement défini l'état de l'art en génération de texte [14] :

Table 1.3: volution de la famille GPT

Modèle	Capacités	Paramètres	Année
GPT-1	Génération basique	117M	2018
GPT-2	Génération cohérente longue	1.5B	2019
GPT-3	Few-shot learning	175B	2020
GPT-4	Multimodal, raisonnement	∼1T	2023

Le modèle **LLaMA** (Meta AI) [15]est un modèle causal open-source comparable à GPT en termes de performances. Dans notre travail, LLaMA est utilisé comme évaluateur de niveau CEFR après transcription du discours oral.

1.7 Le Fine-tuning des LLMs

1.7.1 Paradigme pré-entraînement / fine-tuning

Le paradigme moderne du NLP repose sur deux étapes [16]:

1. **Pré-entraînement** (*pre-training*) : le modèle est entraîné sur de très grandes quantités de texte non annoté pour apprendre des représentations générales du langage.
2. **Fine-tuning** : le modèle pré-entraîné est adapté à une tâche spécifique sur un dataset annoté plus petit.

Cette approche est extrêmement efficace : un modèle pré-entraîné sur des milliards de tokens peut être fine-tuné en quelques heures sur quelques milliers d'exemples.

1.7.2 PEFT : Parameter-Efficient Fine-Tuning

Le fine-tuning complet d'un LLM de plusieurs milliards de paramètres nécessite des ressources GPU considérables. Les méthodes PEFT (*Parameter-Efficient Fine-Tuning*) permettent d'adapter le modèle en ne modifiant qu'une petite fraction des paramètres [17]

1.7.3 LoRA : Low-Rank Adaptation

LoRA [17]est la méthode PEFT la plus populaire. L'idée centrale est que la mise à jour des poids ΔW peut être factorisée en deux matrices de faible rang :

$$W' = W_0 + \Delta W = W_0 + BA \tag{1.11}$$

où $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, et $r \ll \min(d, k)$ est le rang. Seules les matrices A et B sont entraînées.

QLoRA [18]combine LoRA avec la *quantification en 4 bits* (NF4), réduisant encore la consommation mémoire. C'est précisément cette approche que nous utilisons pour fine-tuner Qwen2-Audio dans notre projet.

1.7.4 Instruction Tuning et RLHF

Le **fine-tuning par instructions** (*instruction tuning*) consiste à entraîner le modèle sur des paires (instruction, réponse) pour améliorer sa capacité à suivre des directives en langage naturel.

Le **RLHF** (*Reinforcement Learning from Human Feedback*) [19]ajoute une étape d'optimisation par renforcement guidée par des préférences humaines, produisant des modèles comme InstructGPT et ChatGPT.

1.8 Les Speech LLMs

1.8.1 De la parole au texte : l'ASR classique

La reconnaissance automatique de la parole (*Automatic Speech Recognition* – ASR) a longtemps reposé sur des modèles hybrides HMM-DNN. Les modèles modernes comme **Whisper** [20](OpenAI) et **wav2vec 2.0** [13] (Meta) ont considérablement amélioré la qualité de transcription grâce au pre-training auto-supervisé.

1.8.2 Les modèles de langage audio multimodaux

La frontière suivante est l'intégration directe de l'audio dans les LLMs, créant des **Speech LLMs** capables de comprendre et d'analyser la parole sans étape de transcription intermédiaire.

Qwen2-Audio [21]développé par Alibaba, est l'un des Speech LLMs les plus avancés. Il combine :

- Un **encodeur audio** basé sur Whisper-large-v2, qui extrait des représentations du signal sonore
- Un **décodeur LLM** (Qwen-7B) qui raisonne sur ces représentations pour produire une réponse textuelle

Cette architecture permet à Qwen2-Audio d'analyser non seulement *ce qui est dit* (le contenu), mais aussi *comment c'est dit* (le rythme, les hésitations, l'intonation) – une propriété cruciale pour l'évaluation de la compétence orale.

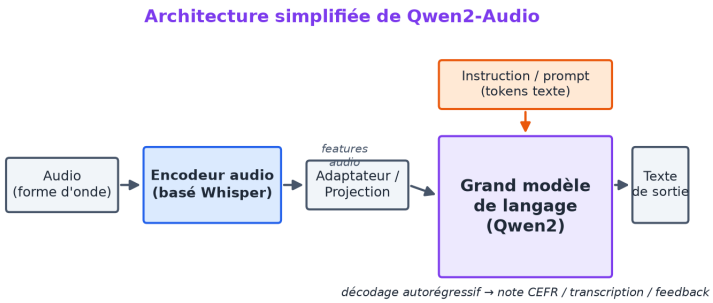

Figure 1.2: Architecture simplifiée de Qwen2-Audio [21]

1.8.3 Comparaison des approches

Table 1.4: Comparaison cascade vs. Speech-LLM

Critère	**Approche Cascade**	**Speech-LLM**
Pipeline	ASR → Transcription → LLM	Audio → LLM (direct)
Informations capturées	Contenu lexical uniquement	Contenu + prosodie + hésitations
Robustesse au bruit	Limitée par l'ASR	Meilleure
Coût computationnel	Moindre	Plus élevé
Explicabilité	Via la transcription	Boîte noire
Exemple	wav2vec2 → LLaMA	Qwen2-Audio

Conclusion du chapitre

Ce chapitre a posé les bases théoriques de notre travail en couvrant l'ensemble de la chaîne technologique qui constitue le socle d'EvalEnglish : du deep learning classique aux architectures Transformer, en passant par les représentations de mots, les MLMs, les modèles causaux, les techniques de fine-tuning efficace et les Speech LLMs. Ces connaissances seront directement mobilisées dans les chapitres suivants pour concevoir et évaluer notre système.

Chapter 2

tat de l'Art : valuation Orale Automatique et Référentiels Linguistiques

Introduction du chapitre

L'évaluation de la compétence orale en langue étrangère est un domaine de recherche actif qui mobilise des expertises en linguistique, psychométrie, traitement automatique de la parole et intelligence artificielle. Ce chapitre dresse un panorama des travaux existants, des référentiels utilisés et des systèmes déployés dans le domaine.

2.1 Les compétences orales en langue anglaise

2.1.1 Définition de la compétence orale

La compétence de communication orale (*speaking competence*) est généralement décomposée en plusieurs sous-compétences [24]:

- **Fluidité (Fluency) :** capacité à produire un discours spontané sans pauses excessives, hésitations ou reformulations
- **Précision grammaticale (Grammar) :** maîtrise des structures syntaxiques de la langue cible
- **Richesse lexicale (Vocabulary) :** étendue et précision du vocabulaire utilisé

- **Cohérence et cohésion (Coherence) :** capacité à organiser logiquement les idées et à les articuler avec des connecteurs appropriés
- **Prononciation (Pronunciation) :** clarté phonétique, accentuation, intonation et rythme

Ces dimensions sont précisément celles que notre plateforme EvalEnglish évalue de façon automatique.

2.1.2 Les difficultés de l'évaluation orale

L'évaluation de la production orale pose des défis spécifiques par rapport à l'écrit [23] :

1. **L'éphémérité :** la parole est fugace et doit être captée en temps réel
2. **La subjectivité :** les évaluateurs humains varient dans leurs jugements (intra- et inter-évaluateur)
3. **Le coût :** un entretien oral avec un examinateur certifié représente un coût élevé
4. **Le passage à l'échelle :** impossible de faire passer des examens oraux à des milliers d'apprenants simultanément

2.2 Le référentiel CEFR

2.2.1 Présentation du Cadre Européen Commun de Référence pour les Langues

Le **CEFR** (*Common European Framework of Reference for Languages*), élaboré par le Conseil de l'Europe [22] est le référentiel international le plus utilisé pour décrire les compétences langagières. Il définit six niveaux organisés en trois grandes catégories :

Table 2.1: Les six niveaux du CEFR et leurs descripteurs pour la production orale

Niveau	**Catégorie**	**Descripteur oral**
A1	Débutant	Peut se présenter et utiliser des expressions simples pour décrire son lieu d'habitation et ses relations
A2	lémentaire	Peut décrire sa famille, ses conditions de vie avec des phrases courtes
B1	Intermédiaire	Peut s'exprimer sur des sujets familiers, raconter des expériences et donner des opinions
B2	Intermédiaire supérieur	Peut s'exprimer spontanément et couramment sur des sujets complexes
C1	Avancé	Peut s'exprimer spontanément avec précision et fluidité
C2	Maîtrise	Peut participer sans effort à toute conversation avec une maîtrise native

2.2.2 Le CEFR dans la notation numérique

Pour l'entraînement de nos modèles, nous utilisons le corpus **SANDI/Speak&Improv 2025** qui encode les niveaux CEFR sur une échelle numérique continue :

$$\text{Score numérique} \in [2.0, 6.0] \leftrightarrow \text{Niveaux CEFR} \in \{A1, A2, B1, B2, C1, C2\} \tag{2.1}$$

Cette représentation permet d'utiliser des métriques de régression standard (RMSE, corrélation de Pearson) pour évaluer la performance des modèles.

2.2.3 Les critères d'évaluation de la production orale au CEFR

Les grilles d'évaluation du CEFR pour la production orale s'organisent autour de quatre critères principaux, auxquels notre système attribue un score sur 100 :

Table 2.2: Critères d'évaluation orale et poids dans EvalEnglish

Critère	**Description**	**Poids (%)**
Grammaire	Maîtrise des structures syntaxiques	25%
Vocabulaire	Richesse et précision lexicale	25%
Fluidité	Spontanéité, débit, absence de blocages	25%
Cohérence	Organisation et articulation des idées	25%

2.3 Les examens oraux d'anglais : IELTS et TOEFL

2.3.1 L'IELTS (International English Language Testing System)

L'**IELTS** est l'un des examens de langue anglaise les plus reconnus au monde, administré conjointement par British Council, IDP et Cambridge Assessment English [4].

Structure de la partie speaking de l'IELTS :

- **Partie 1 – Interview** (4–5 min) : questions sur des sujets familiers (famille, travail, loisirs)
- **Partie 2 – Long turn** (3–4 min) : monologue sur un sujet donné avec 1 minute de préparation
- **Partie 3 – Discussion** (4–5 min) : discussion abstraite liée au thème de la partie 2

Cette structure est directement inspirée du test CEFR oral que nous avons intégré dans EvalEnglish.

Critères d'évaluation IELTS :

- Fluidité et cohérence (*Fluency & Coherence*)
- Richesse lexicale (*Lexical Resource*)
- Précision et variété grammaticale (*Grammatical Range & Accuracy*)
- Prononciation (*Pronunciation*)

Score : sur une échelle de 0 à 9 bandes (*bands*), correspondant grossièrement aux niveaux CEFR.

Table 2.3: Correspondance IELTS – CEFR

Score IELTS	Niveau CEFR
8.5 – 9.0	C2
7.0 – 8.0	C1
5.5 – 6.5	B2
4.0 – 5.0	B1
3.0 – 3.5	A2

2.3.2 Le TOEFL iBT (Test of English as a Foreign Language)

Le **TOEFL iBT** est administré par Educational Testing Service (ETS) [29]. Sa partie Speaking comprend 4 tâches :

- **Independent task** : donner son opinion sur un sujet familier (45 sec de réponse)
- **Integrated tasks** (x3) : répondre après lecture/écoute d'un document (60 sec de réponse)

Depuis 2019, le TOEFL utilise un système de **notation automatique partielle** basé sur l'IA (SpeechRater™ d'ETS [26]), ce qui constitue l'une des inspirations directes de notre projet.

2.3.3 Limites des examens officiels

Les examens IELTS et TOEFL présentent plusieurs limitations que notre plateforme cherche à surmonter :

Table 2.4: Limites des examens officiels vs. EvalEnglish

Critère	**IELTS / TOEFL**	**EvalEnglish**
Coût	200–280$ par passage	Gratuit (freemium)
Délai des résultats	3–13 jours	Instantané
Fréquence	1 à 4 fois par an	Illimité
Feedback détaillé	Limité	Par critère + par phonème
Accessibilité	Centres agréés	En ligne, partout

2.4 Les systèmes d'évaluation orale automatique existants

2.4.1 ELSA Speak

ELSA (*English Language Speech Assistant*) [27]est une application mobile spécialisée dans l'amélioration de la prononciation en anglais. Elle utilise des modèles ASR et phonétiques pour identifier les erreurs de prononciation à l'échelle du phonème.

Forces : feedback phonétique très détaillé, gamification efficace.
Limites : se concentre uniquement sur la prononciation, ne fournit pas d'évaluation du niveau CEFR global.

2.4.2 Duolingo English Test (DET)

Le **Duolingo English Test** [28]propose une évaluation adaptative de l'anglais en 45 minutes, reconnue par plusieurs universités. Il utilise des algorithmes de scoring automatique pour la production orale.

Forces : accessible en ligne, coût modéré (59$), résultats rapides.
Limites : boîte noire, peu de feedback sur les points spécifiques à améliorer.

2.4.3 SpeechRater (ETS)

SpeechRater™ [26]est le système de notation automatique développé par ETS pour le TOEFL. Il extrait des features acoustiques et linguistiques pour prédire un score.

Forces : utilisé en production dans un examen officiel.

Limites : propriétaire, non transparent, non accessible au grand public.

2.4.4 Speak & Improve (Cambridge)

Speak & Improve est une plateforme de Cambridge English qui permet aux apprenants de pratiquer leur expression orale et d'obtenir un feedback automatique. Le corpus SANDI/Speak&Improve 2025 que nous utilisons pour fine-tuner nos modèles provient de cette plateforme.

2.5 Les travaux de recherche en évaluation automatique du CEFR

2.5.1 Approches par features acoustiques et linguistiques

Les premières approches d'automated speaking assessment extractaient manuellement des features comme le débit de parole (*speech rate*), le nombre de disfluences, le ratio type/token (TTR) et des features prosodiques, puis les soumettaient à un classifieur (SVM, régression linéaire) [29]

2.5.2 Approches end-to-end deep learning

Les approches récentes utilisent des réseaux profonds pour apprendre directement des représentations end-to-end [30] Les Speech LLMs comme Qwen2-Audio représentent l'état de l'art de cette tendance.

2.5.3 Positionnement de notre travail

Notre contribution se distingue par :

1. La **comparaison systématique** de 5 modèles hétérogènes (cascade, Speech-LLM, GOP)
2. L'**intégration dans une plateforme complète** (assess → test → cours → dashboard)

3. Le **fine-tuning accessible** : Qwen2-Audio-7B fine-tuné sur Colab gratuit avec QLoRA

4. La **transparence** : tous les modèles sont comparés côte à côte avec leurs métriques

Conclusion du chapitre

Ce chapitre a permis de situer notre travail dans le paysage de l'évaluation orale automatique. Nous avons vu que les outils existants sont soit coûteux, soit peu transparents, soit limités à la prononciation. EvalEnglish se positionne comme une solution open, multi-modèle et intégrée, s'appuyant sur les référentiels CEFR reconnus internationalement.

Chapter 3

Conception et Évaluation

Introduction du chapitre

Ce chapitre présente l'architecture technique de la plateforme EvalEnglish, les cinq modèles d'IA déployés, le corpus utilisé, le processus de fine-tuning de Qwen2-Audio et l'analyse détaillée des résultats expérimentaux.

3.1 Architecture générale du système

3.1.1 Vue d'ensemble

EvalEnglish est une application web full-stack composée de trois niveaux :

1. **Frontend** : interface utilisateur en Next.js 16 + React 19 + Tailwind CSS v4, déployée sur Vercel
2. **API Gateway** : routes Next.js qui proxifient les requêtes audio vers les serveurs de modèles
3. **Model Servers** : serveurs d'inférence hébergeant les 5 modèles d'IA

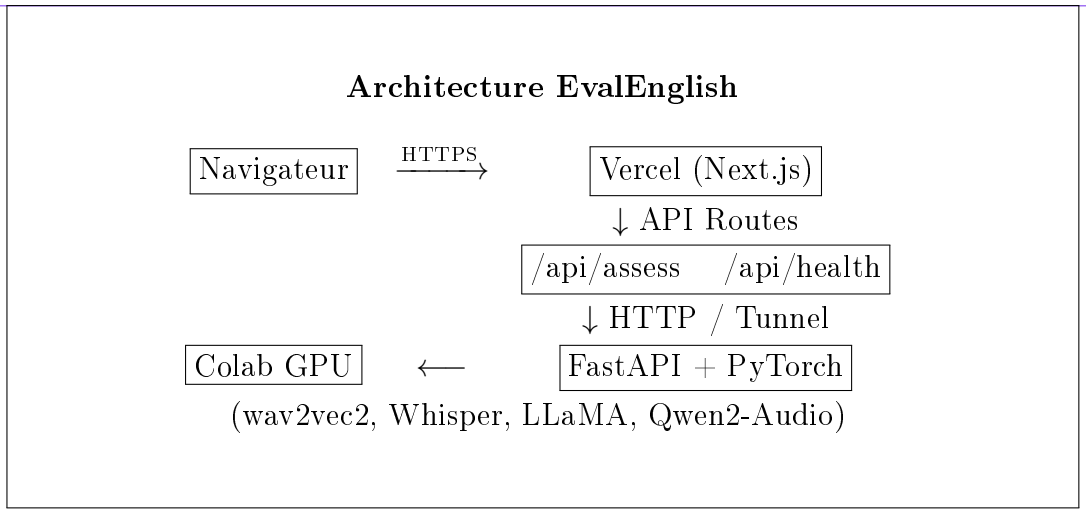

Figure 3.1: Architecture technique d'EvalEnglish

3.1.2 Stack technologique

Table 3.1: Stack technologique complet d'EvalEnglish

Couche	Technologie	Rôle
Frontend	Next.js 16, React 19	Interface utilisateur, routing
Styling	Tailwind CSS v4, Framer Motion	Design responsive, animations
Déploiement	Vercel	Hosting frontend
Backend	FastAPI (Python)	API d'inférence des modèles
ML Framework	PyTorch, transformers (HuggingFace)	Inférence et fine-tuning
Fine-tuning	peft, bitsandbytes	QLoRA sur Colab T4
Audio	librosa, torchaudio	Traitement du signal
ASR	openai-whisper	Transcription automatique
Tunnel	cloudflared	Exposition Colab sur internet
Stockage	cookies, localStorage, IndexedDB	Sessions, données utilisateur

3.2 Les cinq modèles d'évaluation

3.2.1 Présentation des deux approches (volets)

Notre système repose sur deux grandes approches de scoring CEFR :

Volet 1 – Approche en cascade (Cascade) :
Le signal audio est d'abord transcrit en texte par un modèle ASR, puis le texte est soumis à un LLM qui prédit le niveau CEFR. Cette approche est explicable (via la transcription) mais limitée : les informations prosodiques sont perdues.

Volet 2 – Speech-LLM (end-to-end) :
Le modèle reçoit directement le signal audio et prédit le niveau CEFR sans transcription intermédiaire. Cette approche capture des informations suprasegmentales (rythme, hésitations, débit) inaccessibles au Volet 1.

3.2.2 Modèle 1 : wav2vec2 → LLaMA (Volet 1)

- **Transcription :** wav2vec2-large-960h (Meta) transcrit le signal audio en texte
- **Scoring :** LLaMA reçoit le texte et un prompt d'instruction pour prédire le score CEFR
- **Prompt utilisé :**

Listing 3.1: Prompt d'évaluation LLaMA

```
prompt = f"""
You are an expert CEFR examiner.
Analyze this English transcript and provide a CEFR score
between 2.0 (A1) and 6.0 (C2).

Transcript: {transcript}

Respond with only a number between 2.0 and 6.0.
CEFR Score:"""
```

Limitation majeure : wav2vec2 peine sur la parole spontanée non académique et produit des transcriptions bruitées, ce qui dégrade la prédiction du LLaMA.

3.2.3 Modèle 2 : Whisper → LLaMA (Volet 1)

Même pipeline que le Modèle 1, mais avec **Whisper-large-v2** comme ASR.

Avantages de Whisper vs. wav2vec2 :

- Entraîné sur 680,000 heures de parole multilingue
- Bien plus robuste aux accents, bruits de fond et parole spontanée
- Détection automatique de la langue (utilisée pour notre validation d'entrée)

Limitation : Whisper peut halluciner du texte sur les segments silencieux, d'où notre mécanisme de validation qui rejette les enregistrements trop courts ou silencieux.

3.2.4 Modèle 3 : Qwen2-Audio en zéro-shot (Volet 2)

Qwen2-Audio-7B-Instruct est utilisé sans fine-tuning, en *zéro-shot* : on lui soumet directement l'audio avec un prompt d'instruction.

Listing 3.2: Utilisation de Qwen2-Audio en zéro-shot

```
messages = [
    {"role": "system",
     "content": "You are a CEFR oral English examiner."},
    {"role": "user",
     "content": [
        {"type": "audio",
         "audio_url": audio_path},
        {"type": "text",
         "text": "Score this English speech from 2.0 (A1)
                  to 6.0 (C2). Reply with a number only."}
     ]}
]
```

Résultats : RMSE = 0.702, PCC = 0.205, SRC = 0.214. Les corrélations proches de zéro indiquent que le modèle sans fine-tuning est essentiellement en train de deviner .

3.2.5 Modèle 4 : Qwen2-Audio fine-tuné avec QLoRA (Volet 2)

Il s'agit du modèle le plus avancé de notre système. Nous fine-tunons Qwen2-Audio-7B-Instruct sur le corpus SANDI avec QLoRA.

Le corpus SANDI / Speak & Improve 2025

Table 3.2: Caractéristiques du corpus SANDI

Propriété	Valeur
Source	Speak & Improve (Cambridge English)
Taille totale	Plusieurs milliers d'enregistrements
Sous-ensemble utilisé	255 clips (195 train + 60 test)
Scores	Continus, 2.0 (A1) à 6.0 (C2)
Stratification	quilibrée par bande CEFR
Licence	Non-commercial (recherche)

Configuration du fine-tuning

Table 3.3: Hyperparamètres du fine-tuning QLoRA

Hyperparamètre	Valeur
Modèle de base	Qwen2-Audio-7B-Instruct
Méthode	QLoRA (LoRA + quantification 4-bit NF4)
Rang LoRA (r)	16
Alpha LoRA	32
Dropout LoRA	0.05
Encodeur audio	Gelé (frozen)
Modules ciblés	q_proj, v_proj (LLM uniquement)
Données d'entraînement	195 clips stratifiés
poques	2
GPU	Colab T4 (gratuit)

Bogue critique détecté et corrigé

Durant l'entraînement, un bogue majeur a été identifié : une mise à jour de la bibliothèque `transformers` avait renommé l'argument audio

du processeur (`audios` → `audio`), causant l'entraînement du modèle *en mode sourd* (sans audio). Le modèle produisait un score constant quelle que soit l'entrée.

Listing 3.3: Correction du bogue de compatibilité transformers

```python
# Detection automatique du nom de l'argument
import inspect
sig = inspect.signature(processor.__call__)
audio_key = "audio" if "audio" in sig.parameters else "audios"

# Utilisation du bon argument
inputs = processor(
    text=texts,
    **{audio_key: audio_list},
    return_tensors="pt"
)
```

Impact : après correction, la perte d'entraînement est passée de 2.13 à 0.59 (réduction de 72%).

3.2.6 Modèle 5 : GOP – Goodness-of-Pronunciation

Le modèle de prononciation utilise la technique **GOP** (*Goodness-of-Pronunciation*) pour évaluer chaque phonème individuellement.

Pipeline GOP :

1. **G2P** (*Grapheme-to-Phoneme*) : conversion texte → phonèmes avec le lexique CMU
2. **Alignement forcé** : alignment de la séquence de phonèmes sur le signal audio (Montreal Forced Aligner)
3. **Score GOP** : pour chaque phonème p aligné sur le segment $[t_1, t_2]$:

$$\text{GOP}(p) = \log \frac{P(\text{audio}[t_1 : t_2] \mid \text{phonème} = p)}{\max_q P(\text{audio}[t_1 : t_2] \mid \text{phonème} = q)} \tag{3.1}$$

Sortie : chaque phonème reçoit une couleur (vert / jaune / rouge) et un score global sur 10.

3.3 Validation d'entrée

Avant de soumettre un enregistrement aux modèles, notre système effectue une validation :

1. **Détection du silence :** l'énergie RMS du signal est calculée avec librosa. Si elle est inférieure à un seuil, l'enregistrement est rejeté.
2. **Détection de la langue :** Whisper identifie la langue de l'audio. Si ce n'est pas de l'anglais, un message d'erreur explicite est affiché.

3.4 Résultats expérimentaux et analyse

3.4.1 Métriques d'évaluation

Trois métriques sont utilisées pour évaluer les modèles (n=60 clips de test) :

- **RMSE** (*Root Mean Square Error*) : erreur quadratique moyenne entre le score prédit et le score humain (plus bas = meilleur)
- **PCC** (*Pearson Correlation Coefficient*) : mesure la corrélation linéaire avec les scores humains (plus proche de 1 = meilleur)
- **SRC** (*Spearman Rank Correlation*) : corrélation de rang, robuste aux valeurs aberrantes

3.4.2 Résultats comparatifs

Table 3.4: Comparaison des performances des 5 modèles (n=60)

#	Modèle	RMSE ↓	PCC ↑	SRC ↑
1	wav2vec2 → LLaMA	1.346	–	–
2	Whisper → LLaMA	1.346	–	–
3	Qwen2-Audio zéro-shot	0.901	0.205	0.214
4	**Qwen2-Audio fine-tuné**	**0.684**	**0.393**	**0.462**
5	GOP Prononciation	–	–	–
Baseline officielle SANDI		0.440	0.726	–

3.4.3 Analyse des résultats

Modèles 1 et 2 (cascade) : le RMSE élevé de 1.346 (soit environ 1,3 niveau CEFR d'erreur) et l'absence de corrélation significative s'expliquent par la qualité insuffisante des transcriptions sur la parole spontanée. La chaîne ASR → LLM amplifie les erreurs.

Modèle 3 (zéro-shot) : avec un RMSE de 0.901 et des corrélations proches de 0.2, Qwen2-Audio sans fine-tuning n'est guère mieux qu'un hasard raisonné. Le modèle n'a pas été entraîné pour la tâche spécifique d'évaluation CEFR.

Modèle 4 (fine-tuné) : le fine-tuning avec seulement 195 clips double les corrélations (PCC : +92%, SRC : +116%) et réduit le RMSE de 24%. C'est la démonstration principale de notre travail : *un petit dataset stratifié suffit à apprendre le scoring CEFR.*

3.4.4 Analyse de la courbe de perte

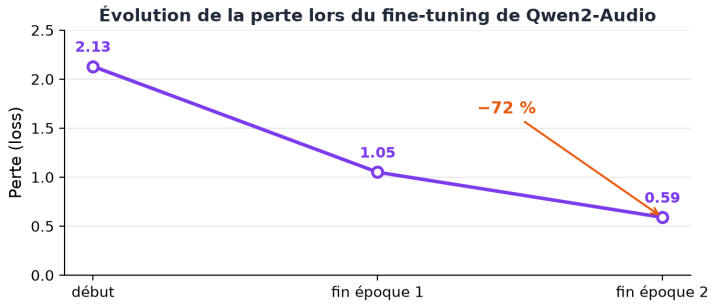

Figure 3.2: Évolution de la perte lors du fine-tuning de Qwen2-Audio

La décroissance continue et régulière de la perte confirme que l'entraînement s'est correctement déroulé après correction du bogue de compatibilité.

3.4.5 Limites et honnêteté scientifique

Notre travail présente des limites importantes qu'il convient de mentionner honnêtement :

- **Corrélations modérées :** PCC = 0.393 reste loin de la baseline officielle (0.726). Notre modèle est un prototype, pas un système de production.
- **Petit dataset :** 195 clips d'entraînement est insuffisant pour généraliser à tous les accents et contextes.
- **Biais de corpus :** SANDI est surreprésenté en niveaux B1-B2, ce qui compresse la gamme de prédiction.
- **Infrastructure fragile :** les tunnels cloudflared sur Colab ne sont pas fiables en production.

Conclusion du chapitre

Ce chapitre a présenté l'architecture technique complète d'EvalEnglish et le processus expérimental ayant conduit à nos résultats. Le fine-tuning de Qwen2-Audio avec QLoRA constitue la contribution technique principale de ce travail : il démontre qu'un Speech-LLM peut apprendre à évaluer le niveau CEFR oral à partir d'un petit corpus stratifié, avec des ressources computationnelles gratuites.

Chapter 4

Réalisation et Présentation de la Plateforme EvalEnglish

Introduction du chapitre

Ce dernier chapitre présente la réalisation concrète de la plateforme EvalEnglish. Nous décrivons les choix technologiques du frontend, les fonctionnalités implémentées et nous illustrons chaque module avec des captures d'écran de l'application.

4.1 Technologies frontend

4.1.1 Next.js 16 et React 19

Le frontend d'EvalEnglish est développé avec **Next.js 16**, un framework React fullstack qui offre le rendu côté serveur (SSR), la génération statique (SSG) et les routes API [31] **React 19** apporte les dernières améliorations en termes de performance et de concurrent rendering.

4.1.2 Tailwind CSS v4 et Framer Motion

Tailwind CSS v4 permet un design system cohérent basé sur des classes utilitaires. La palette de couleurs d'EvalEnglish est centrée sur le violet (`#7c3aed`), créant une identité visuelle moderne et distinctive.

Framer Motion gère toutes les animations de l'interface : transitions de pages, animations des barres de progression, effets d'apparition des

composants.

4.2 Authentification et gestion des utilisateurs

4.2.1 Page d'inscription

La page d'inscription permet à l'utilisateur de créer son espace personnel. Elle propose deux types de profils :

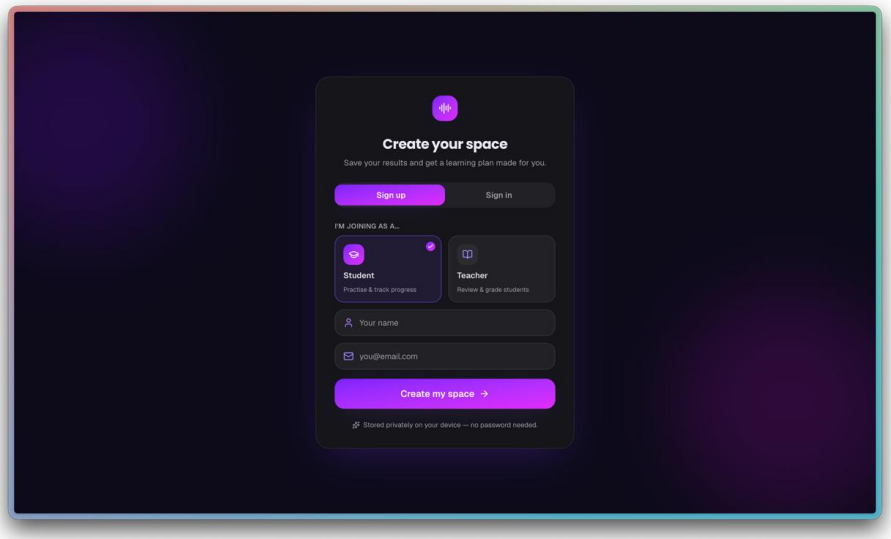

Figure 4.1: Page d'inscription – profil tudiant

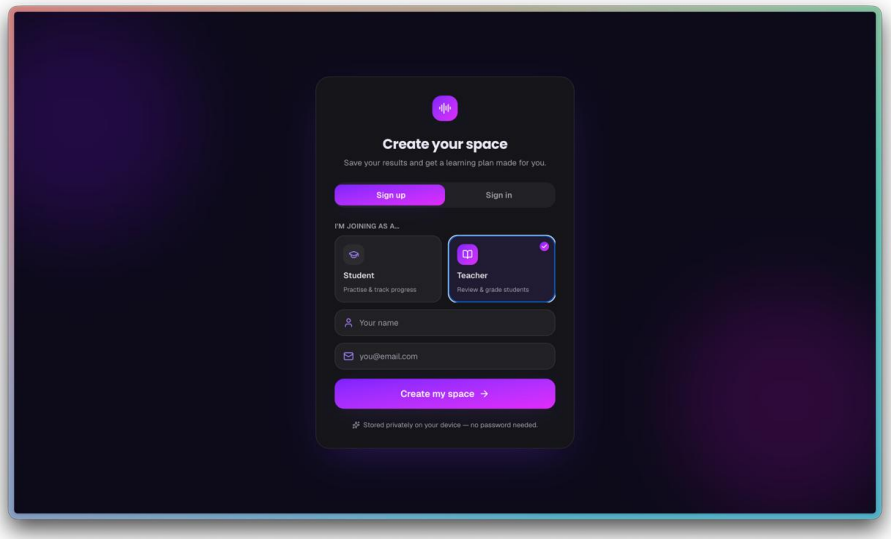

Figure 4.2: Page d'inscription – profil Enseignant

Fonctionnalités :

- Sélection du rôle : **tudiant** (pratique et suivi de progression) ou **Enseignant** (correction et notation des étudiants)
- Authentification sans mot de passe (*passwordless*) : les données sont stockées localement sur l'appareil
- Design épuré avec fond sombre et accents violets

4.2.2 Page de connexion

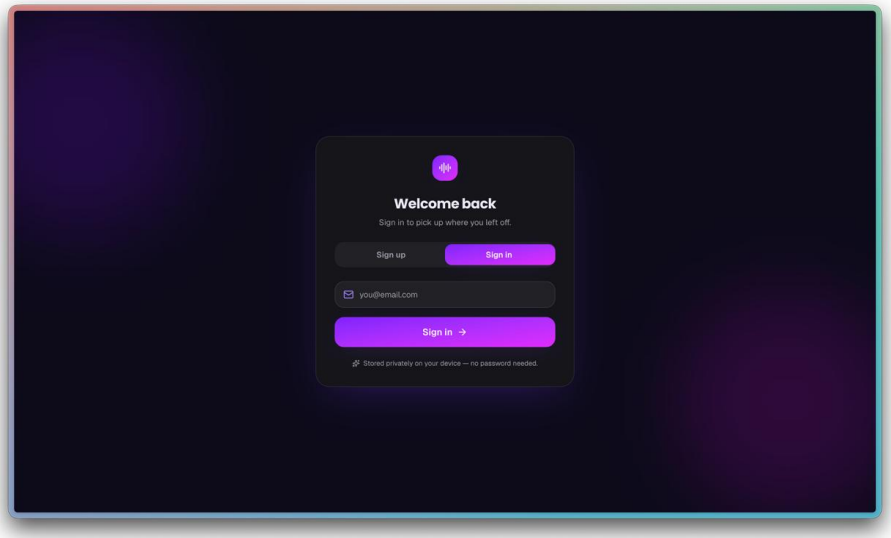

Figure 4.3: Page de connexion à l'espace personnel

La page de connexion permet aux utilisateurs existants de reprendre leur parcours là où ils l'avaient laissé. L'authentification repose uniquement sur l'adresse email, sans mot de passe.

4.3 Le test CEFR oral guidé

4.3.1 Page d'accueil du test CEFR

[Capture : Page CEFR Test – Find your CEFR level]
Sélection du niveau (A1 à C2) avec descripteurs

Figure 4.4: Page d'accueil du test CEFR – Sélection du niveau

L'utilisateur choisit son niveau présumé parmi les 6 niveaux CEFR avec leurs descripteurs complets. Cette page présente aussi les caractéristiques du test : 3 parties, 1–2 minutes, fonctionne hors-ligne en mode démo.

4.3.2 Partie 1 – Interview

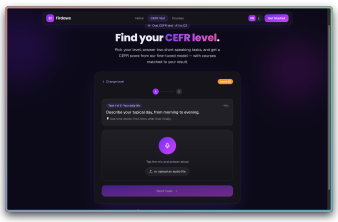

Figure 4.5: Test CEFR – Partie 1 : Interview (enregistrement en cours)

La Partie 1 est une auto-présentation guidée. L'utilisateur dispose de 45 secondes pour se présenter naturellement en anglais. Un indicateur de temps et un bouton d'enregistrement rouge clignotant guident l'utilisateur.

4.3.3 Partie 3 – Discussion

[Capture : Part 3 - Discussion]
"Some people prefer studying online..."

Figure 4.6: Test CEFR – Partie 3 : Discussion (sujet de débat)

La Partie 3 présente un sujet de débat plus complexe (ex: enseignement en ligne vs. présentiel). L'utilisateur doit exprimer son opinion, l'argumenter et présenter le point de vue opposé.

4.3.4 Résultats du test CEFR

[Capture : Résultats CEFR – Niveau B2, 59/100]
Breakdown : Grammar 57 / Vocabulary 56 / Fluency 60 / Coherence 58

Figure 4.7: Résultats du test CEFR – Niveau B2 avec breakdown par compétence

La page de résultats affiche :

- Le **niveau CEFR** obtenu (ex : B2 – Upper-Intermediate) avec le score global (/100)
- Une **barre visuelle** de progression de A1 à C2
- Le **breakdown par compétence** : Grammaire, Vocabulaire, Fluidité, Cohérence
- Un résumé **partie par partie** (Part 1, Part 2, Part 3)

4.4 La bibliothèque de cours

4.4.1 Vue catalogue des cours

[Capture : Bibliothèque de cours – Niveau C1]
Accent & Intonation / Academic Writing / Idioms & Natural Speech

Figure 4.8: Bibliothèque de cours – Vue niveau C1

La bibliothèque propose **18 cours** organisés de A1 à C2, avec :

- Filtrage par niveau (A1, A2, B1, B2, C1, C2) et par accès (Gratuit / Premium)
- Barre de recherche par sujet ou compétence
- Affichage de la note (), durée et prix pour chaque cours
- Un bandeau intelligent indiquant le niveau actuel de l'utilisateur

4.4.2 Page détail d'un cours

[Capture : Cours "First Words & Greetings" – A1 Beginner – Free]
Rating 4.8 / 16 lessons / 8h / 21.3k learners

Figure 4.9: Page détail d'un cours – First Words & Greetings (A1, Gratuit)

Chaque cours dispose d'une page détaillée présentant :

- Le titre, l'instructeur, le niveau et le type d'accès (Gratuit / Premium)
- Les statistiques : note, nombre de leçons, durée, nombre d'apprenants
- La description et les objectifs pédagogiques
- Le syllabus complet avec les chapitres et leur durée
- Un certificat à l'issue du cours

4.5 Le dashboard apprenant et enseignant

4.5.1 Dashboard apprenant

Le tableau de bord de l'apprenant centralise :

- Son **niveau CEFR actuel** et son score moyen
- L'**historique** de tous ses passages et évaluations
- Les **recommandations de cours** personnalisées selon ses points faibles
- La possibilité de **partager ses résultats** avec un enseignant

4.5.2 Espace enseignant

L'espace enseignant permet :

- De **recevoir les soumissions** des étudiants avec leurs scores automatiques
- De consulter le **détail** de chaque évaluation (breakdown par compétence)
- De gérer un **annuaire** d'étudiants avec autocomplétion

4.6 Synthèse des fonctionnalités

Table 4.1: Récapitulatif des fonctionnalités d'EvalEnglish

Module	**Fonctionnalités**	**Accès**
Authentification	Inscription étudiant/enseignant, connexion sans mot de passe	Gratuit
valuation (Home)	Enregistrement + 5 modèles côte à côte + breakdown	Gratuit
Test CEFR oral	3 parties guidées + score global + recommandations	Gratuit
Cours	18 cours A1–C2, chapitres, exercices, ressources	Free/Premium
Dashboard	Progression, historique, recommandations	Gratuit
Partage enseignant	Envoi de résultats, inbox enseignant	Gratuit
GOP Prononciation	Feedback phonème par phonème (couleurs)	Gratuit

Conclusion du chapitre

Ce chapitre a présenté la réalisation complète de la plateforme EvalEnglish. De l'authentification au test CEFR en passant par la bibliothèque de cours et les tableaux de bord, chaque module a été conçu pour offrir une expérience utilisateur fluide et moderne. La plateforme constitue un MVP (Minimum Viable Product) complet qui démontre la faisabilité technique et pédagogique de l'évaluation orale automatique basée sur les LLMs.

Conclusion Générale et Perspectives

Bilan du travail réalisé

Ce mémoire a présenté **EvalEnglish**, une plateforme web intelligente d'évaluation automatique des compétences orales en anglais selon le référentiel CEFR. Notre contribution s'articule autour de trois axes principaux :

1. Contribution théorique : Nous avons proposé une étude comparative systématique de deux approches de scoring CEFR (cascade vs. Speech-LLM) à travers cinq modèles complémentaires, offrant une vision transparente et reproductible des performances de chacun.

2. Contribution expérimentale : Nous avons démontré que le fine-tuning de Qwen2-Audio-7B avec QLoRA sur seulement 195 clips stratifiés permet de doubler les corrélations avec les scores humains (PCC : 0.205 $\rightarrow$ 0.393 ; SRC : 0.214 $\rightarrow$ 0.462). Ce résultat est obtenu sur un GPU gratuit (Colab T4), ouvrant la voie à une démocratisation du fine-tuning de Speech-LLMs.

3. Contribution applicative : Nous avons développé une plateforme complète et fonctionnelle qui va au-delà du simple modèle : test CEFR guidé en 3 parties, bibliothèque de 18 cours, feedback par compétence, feedback phonétique par GOP, dashboard apprenant et espace enseignant.

Limites

Notre travail présente des limites que nous reconnaissons pleinement : les corrélations obtenues (PCC = 0.393) restent inférieures à la baseline officielle (0.726) ; le dataset d'entraînement est de taille modeste ; l'infrastructure de déploiement (tunnels Colab) n'est pas adaptée à la production.

Perspectives

Les perspectives d'amélioration et d'extension d'EvalEnglish sont nombreuses :

- **Données :** augmenter le corpus d'entraînement (idéalement 2000+ clips) pour améliorer les corrélations
- **Modèles :** explorer des architectures plus récentes (Qwen2.5-Audio, Gemini Audio)
- **Infrastructure :** migrer vers des serveurs GPU dédiés pour une disponibilité 24/7
- **Fonctionnalités :** paiement réel, application mobile, support multilingue
- **Validation :** tests utilisateurs avec de vrais apprenants d'anglais et validation par des examinateurs CEFR certifiés
- **thique :** validation de l'équité du modèle sur différents accents et L1

EvalEnglish démontre qu'il est possible, avec des ressources accessibles, de construire un système d'évaluation orale automatique sérieux et multimodal. Ce travail constitue une première étape vers une plateforme commercialement viable qui pourrait bénéficier à des millions d'apprenants à travers le monde.

Bibliography

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ: Pearson, 2020.

[2] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning.* Boston, MA: The Center for Curriculum Redesign, 2019.

[3] M. Godwin-Jones, "Smartphones and language learning," *Language Learning & Technology*, vol. 21, no. 2, pp. 3–17, 2017.

[4] T. McNamara, *Language Testing.* Oxford: Oxford University Press, 2000.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* Cambridge, MA: MIT Press, 2016.

[7] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2–3, pp. 146–162, 1954.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. NIPS*, 2013, pp. 3111–3119.

[9] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

[10] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *TACL*, vol. 5, pp. 135–146, 2017.

[11] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020, pp. 12449–12460.

[14] T. Brown et al., "Language models are few-shot learners," in *Proc. NeurIPS*, 2020, pp. 1877–1901.

[15] H. Touvron et al., "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[16] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. ACL*, 2018, pp. 328–339.

[17] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.

[18] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Proc. NeurIPS*, 2023.

[19] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. NeurIPS*, 2022.

[20] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.

[21] Y. Chu et al., "Qwen2-Audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[22] Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press, 2001.

[23] L. F. Bachman, *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press, 1990.

[24] L. Taylor, Ed., *Examining Speaking: Research and Practice in Assessing Second Language Speaking*. Cambridge: Cambridge University Press, 2011.

[25] C. A. Chapelle and D. Douglas, *Assessing Language through Computer Technology*. Cambridge: Cambridge University Press, 2008.

[26] L. Chen, K. Zechner, S.-Y. Yoon, K. Evanini, X. Wang, A. Loukina, J. Tao, L. Davis, C. Lee, M. Ma, R. Mundkowsky, C. Lu, C. W. Leong, and B. Gyawali, "Automated scoring of nonnative speech using the SpeechRater v5.0 engine," in *Proc. INTERSPEECH*, 2018.

[27] V. Q. Tran, "ELSA: English language speech assistant," in *Proc. Interspeech*, 2020.

[28] J. LaFlair and M. Settles, "Duolingo English Test: Technical manual," Duolingo Inc., Pittsburgh, PA, Tech. Rep., 2021.

[29] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[30] C.-K. Lo and K.-F. Hew, "Grading of student speech using deep learning," *Computers & Education*, vol. 185, 2022.

[31] Vercel, "Next.js 16 Documentation," [Online]. Available: https://nextjs.org/docs. [Accessed: 2025].

ملخص

EvalEnglish منصة وب ذكية للتقييم الآلي لمهارات التعبير الشفهي في اللغة الإنجليزية وفق الإطار المرجعي الأوروبي المشترك CEFR (من A1 إلى C2). تجمع المنصة خمسة نماذج ذكاء اصطناعي متكاملة: مسارين تعاقبيين (wav2vec2 ثم LLaMA، و Whisper ثم LLaMA)، ونموذجي Speech-LLM (Qwen2-Audio في وضعي zero-shot والتحسين الدقيق)، ونموذج نطق يعتمد تقنية GOP. بعد تحسينه الدقيق بتقنية QLoRA على متن SANDI/Speak&Improve 2025، يحقق Qwen2-Audio معامل ارتباط بيرسون (PCC) قدره 0.393 ومعامل ارتباط سبيرمان (SRC) قدره 0.462. توفر المنصة اختبارا شفهيا موجها من ثلاثة أجزاء، ومكتبة من 18 دورة، ولوحة تحكم للمتعلم، وفضاء للأستاذ.

الكلمات المفتاحية: التقييم الشفهي الآلي، QLoRA، Qwen2-Audio، LLM، CEFR، التحسين الدقيق، Whisper، wav2vec2، تقييم النطق، منصة التعلم الإلكتروني.

Abstract

EvalEnglish is an intelligent web platform for the automatic, CEFR-aligned assessment of English speaking skills (A1–C2). It combines five complementary AI models — two cascade pipelines (wav2vec2 → LLaMA, Whisper → LLaMA), two Speech-LLMs (Qwen2-Audio, zero-shot and fine-tuned), and a GOP pronunciation model. Fine-tuned with QLoRA on the SANDI/Speak&Improve 2025 corpus, Qwen2-Audio reaches a PCC of 0.393 and an SRC of 0.462, roughly doubling the zero-shot correlations. The platform delivers a 3-part guided oral test, a library of 18 courses, a learner dashboard and a teacher workspace.

Keywords: *automatic oral assessment, CEFR, LLM, Qwen2-Audio, QLoRA, fine-tuning, wav2vec2, Whisper, pronunciation, e-learning platform.*

Résumé

EvalEnglish est une plateforme web intelligente d'évaluation automatique de l'expression orale en anglais, alignée sur le CEFR (A1–C2). Elle combine cinq modèles d'IA complémentaires : deux pipelines en cascade (wav2vec2 → LLaMA, Whisper → LLaMA), deux Speech-LLM (Qwen2-Audio en zéro-shot et fine-tuné) et un modèle de prononciation GOP. Fine-tuné avec QLoRA sur le corpus SANDI/Speak&Improve 2025, Qwen2-Audio atteint un PCC de 0,393 et un SRC de 0,462, doublant les corrélations du zéro-shot. La plateforme propose un test oral guidé en 3 parties, une bibliothèque de 18 cours, un tableau de bord apprenant et un espace enseignant.

Mots-clés : *évaluation orale automatique, CEFR, LLM, Qwen2-Audio, QLoRA, fine-tuning, wav2vec2, Whisper, prononciation, plateforme e-learning.*