



# École Nationale des Sciences Appliquées de Tétouan

Filière : Sciences de Données et Intelligence Artificielle

## Rapport de stage PFA

### Application de l'Intelligence Artificielle à l'apprentissage de la langue arabe : amélioration de la prononciation

Réalisé par : Omaïma Younes

Encadrant : Mr. Fahd Seffar

Année universitaire : 2025 – 2026

# Dédicace

Je dédie ce travail

À ma mère, ma raison d'être, la lanterne qui éclaire mon chemin et m'illumine d'affection et d'amour, celle qui a cru en moi lorsque personne d'autre ne l'a fait.

A mon père , en signe d'amour, et de gratitude pour tous les soutiens et les sacrifices dont il a fait preuve à mon égard.

A mes chères sœurs et mon cher frère,  
En témoignage de mon affection Fraternelle, je vous souhaite, Fatima , Zakia , Rahma et Ahmed une vie pleine de bonheur et de succès et qu'Allah, le tout puissant, vous protège et vous garde !

...Omaïma YOUNES

# Remerciements

Ce travail n'aurait jamais pu se concrétiser sans l'aide et le soutien de plusieurs personnes que je souhaite vivement remercier et à qui je dédie ce travail.

Tout d'abord, je remercie Yafa Technologies, la startup qui m'a accueillie chaleureusement et m'a offert l'opportunité de mettre en pratique mes connaissances.

Je tiens également à exprimer ma profonde gratitude à Monsieur Fahd Seffar, mon encadrant au sein de Yafa Technologies, pour sa confiance, ses conseils avisés et son accompagnement tout au long de ce projet.

Je remercie aussi moi-même, pour mon engagement, ma persévérance et ma détermination à mener ce travail à bien.

Enfin, je remercie ma famille pour leur soutien indéfectible, leurs encouragements et leur présence à chaque étape de mon parcours.

# Résumé

Ce projet s'inscrit dans le domaine de l'intelligence artificielle et plus particulièrement de l'apprentissage automatique appliqué à la reconnaissance vocale. L'objectif principal était de concevoir et d'évaluer des modèles capables de traiter et de transcrire des signaux vocaux en texte afin d'aider à la correction de la prononciation des apprenants de la langue arabe.

Pour cela, deux approches ont été étudiées : d'une part, l'utilisation de réseaux de neurones convolutionnels (CNN) pour l'analyse et la correction de la prononciation des lettres arabes, et d'autre part, l'exploitation du modèle Whisper pour la transcription automatique des expressions courantes ainsi que pour leur correction. Chaque modèle a été entraîné et testé sur un ensemble de données vocales issues de locuteurs arabophones, permettant de comparer leurs performances en termes de précision et de robustesse.

Les résultats obtenus montrent que l'intégration de ces techniques permet de détecter et de corriger efficacement les erreurs de prononciation tout en mettant en évidence certaines limites liées à la variabilité des accents et des intonations. Ces travaux contribuent ainsi à l'amélioration des outils d'apprentissage des langues et ouvrent des perspectives pour des applications pédagogiques innovantes.

**Mots clés :** apprentissage automatique, reconnaissance vocale, CNN, Whisper, MFCC, traitement du signal, correction de prononciation.

# Abstract

This project falls within the field of artificial intelligence, and more specifically, machine learning applied to speech recognition. The main objective was to design and evaluate models capable of processing and transcribing speech signals into text in order to assist in correcting the pronunciation of Arabic language learners.

Two approaches were explored : firstly, the use of Convolutional Neural Networks (CNN) for analyzing and correcting the pronunciation of Arabic letters, and secondly, the use of the Whisper model for automatic transcription of common expressions as well as their correction. Each model was trained and tested on a dataset of Arabic-speaking speakers, allowing a comparison of their performance in terms of accuracy and robustness.

The results show that combining these techniques enables effective detection and correction of pronunciation errors, while highlighting certain limitations related to variations in accents and intonation. These findings contribute to the improvement of language learning tools and open up prospects for innovative educational applications.

**Keywords :** machine learning, speech recognition, CNN, Whisper, MFCC, signal processing, pronunciation correction.

# Table des matières

<b>Dédicace</b>	<b>1</b>
<b>Remerciements</b>	<b>2</b>
<b>Résumé</b>	<b>3</b>
<b>Abstract</b>	<b>4</b>
<b>Introduction générale</b>	<b>7</b>
<b>1 Contexte général du stage</b>	<b>8</b>
1.1 Présentation de l'entreprise et domaine d'activité . . . . .	8
1.2 Présentation du projet . . . . .	9
<b>2 Étude théorique et état de l'art</b>	<b>10</b>
2.1 Bases théoriques sur la voix et l'audition humaine . . . . .	10
2.1.1 Production de la voix humaine . . . . .	10
2.1.2 Perception auditive . . . . .	11
2.1.3 Analogies avec la reconnaissance vocale artificielle . . . . .	11
2.2 Modèles de reconnaissance vocale . . . . .	12
2.2.1 Caractéristiques acoustiques du son . . . . .	12
2.2.2 Réseaux de neurones convolutifs (CNN) . . . . .	15
2.2.3 Whisper d'OpenAI . . . . .	15
2.3 Outils et technologies utilisés . . . . .	17
<b>3 Conception et Réalisation du prototype</b>	<b>18</b>
3.1 Analyse du besoin . . . . .	18
3.2 Architecture proposée du prototype . . . . .	19
3.3 Développement du modèle CNN . . . . .	20
3.4 Développement du module Whisper . . . . .	20
3.5 Mise en place d'une interface utilisateur . . . . .	21

---

<b>4 Résultats et Évaluation</b>	<b>22</b>
4.1 Résultats et évaluation des performances . . . . .	22
4.2 Résultats et évaluation des performances . . . . .	22
4.3 Discussion et limites . . . . .	24
4.4 Perspectives futures . . . . .	25
 Conclusion générale	 <b>26</b>
 Liste des acronymes	 <b>28</b>
 Bibliographie	 <b>29</b>

# Introduction générale

La langue arabe, parlée par plus de 400 millions de personnes réparties dans une vingtaine de pays, est une langue riche et complexe sur les plans phonétique, lexical et grammatical. Son apprentissage peut s'avérer particulièrement difficile, surtout pour les apprenants non natifs, en raison de la prononciation spécifique de certaines lettres et sons, de l'accentuation, et des intonations propres à chaque dialecte. La maîtrise de la prononciation correcte est essentielle, car elle conditionne la compréhension orale et la communication efficace.

Depuis plusieurs décennies, les avancées en intelligence artificielle ont transformé de nombreux domaines, y compris l'éducation et l'apprentissage des langues. Les modèles d'IA modernes, capables de traiter et d'analyser de grandes quantités de données, ont ouvert la voie à des outils innovants pour l'apprentissage personnalisé. La reconnaissance vocale, le traitement automatique du langage et la correction automatique de la prononciation sont désormais des applications concrètes qui facilitent l'acquisition de compétences linguistiques, offrant aux apprenants un retour immédiat et précis sur leur performance.

Dans ce contexte, ce projet vise à exploiter les techniques d'apprentissage automatique pour la correction de la prononciation en langue arabe. Deux approches principales ont été étudiées : l'utilisation de réseaux de neurones convolutionnels (CNN) pour l'analyse et la correction des lettres arabes, et l'utilisation du modèle Whisper pour la transcription automatique et la correction des expressions courantes. L'objectif est de proposer un système pédagogique capable de détecter les erreurs de prononciation, de fournir des corrections adaptées et d'accompagner efficacement les apprenants dans leur progression.



# Chapitre 1

## Contexte général du stage

### 1.1 Présentation de l'entreprise et domaine d'activité

**Yafa Technologies** est une startup basée à Tétouan, au Maroc, spécialisée dans le développement de solutions numériques et technologiques adaptées à divers secteurs d'activité. Sa mission principale est de proposer des services innovants qui couvrent des domaines tels que le développement web et mobile, l'ingénierie SIG, la formation professionnelle, le marketing digital et les solutions métiers. L'entreprise intervient auprès de clients issus de secteurs variés, notamment la topographie, le bâtiment, l'agroalimentaire, l'industrie, l'e-commerce, ainsi que la recherche scientifique.

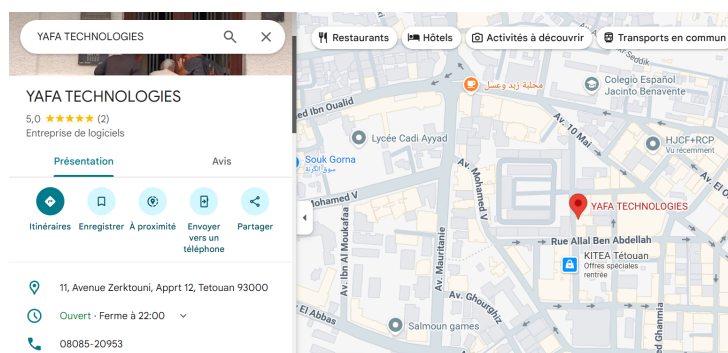


FIGURE 1.1 – Localisation de l'entreprise Yafa Technologies

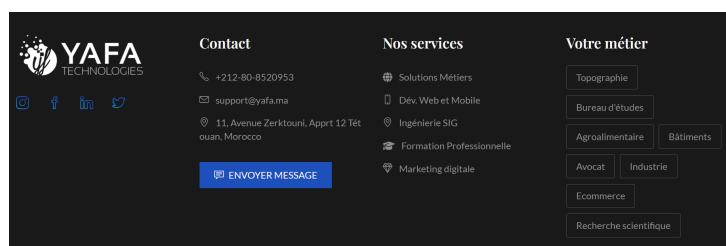


FIGURE 1.2 – services proposées et clients

---

## 1.2 Présentation du projet

Dans le cadre de son développement, **YAFA Technologies** a lancé Lissani, une plateforme innovante destinée à l'apprentissage de la langue arabe pour les non-arabophones, adaptée à différentes tranches d'âge. La plateforme propose un parcours pédagogique complet, débutant par les lettres de l'alphabet et allant jusqu'aux expressions courantes. Les utilisateurs peuvent s'entraîner à travers différents types d'exercices : **images, voix et écriture**. Elle supporte également la traduction en plusieurs langues, notamment le français, l'anglais et le néerlandais, afin de faciliter la compréhension et l'assimilation des contenus.

Loin de se limiter à l'enseignement classique, la startup a voulu tirer parti des possibilités offertes par l'intelligence artificielle pour rendre l'apprentissage plus vivant et personnalisé. L'IA est ainsi intégrée à plusieurs niveaux : correction automatique de la prononciation, adaptation des exercices au rythme et au niveau de chaque utilisateur, et accompagnement interactif tout au long du parcours.

Dans le cadre de mon stage, j'ai choisi de me focaliser sur la tâche de correction de la prononciation, car la prononciation des mots arabes constitue un défi majeur pour les non-arabophones et représente un aspect crucial pour un apprentissage efficace de la langue.

# Chapitre 2

## Étude théorique et état de l'art

### 2.1 Bases théoriques sur la voix et l'audition humaine

Parmi les systèmes les plus merveilleux et sans doute les plus complexes du corps humain, on trouve **le système vocal et le système auditif**. Ces systèmes permettent non seulement la production et la perception des sons, mais jouent également un rôle fondamental dans la communication, l'expression des émotions et l'apprentissage des langues.

#### 2.1.1 Production de la voix humaine

La voix occupe une place centrale dans la communication humaine et constitue une caractéristique aussi unique qu'une empreinte digitale. Mais avant tout, qu'entend-on réellement par « voix » ?

**La voix humaine** est l'ensemble des sons produits par le frottement de l'air des poumons sur les replis du larynx de l'être humain. La voix inclut la parole, le chuchotement, le gémissement, le cri, le rire et le chant [1].

La voix humaine est produite en trois étapes importantes :

1. **Respiration** : l'air est expulsé des poumons grâce au diaphragme et aux muscles thoraciques, fournissant l'énergie nécessaire à la production de la voix.
2. **Phonation et résonance** : l'air traverse les cordes vocales situées dans le larynx, qui vibrent pour produire un son brut, ensuite amplifié et modulé par la gorge, la bouche et le nez, donnant le timbre unique de chaque voix.
3. **Articulation et contrôle cérébral** : la langue, les lèvres, le palais et les dents façonnent le son en voyelles et consonnes, formant des mots et des phrases, tandis que le cerveau coordonne toutes ces actions pour assurer une parole fluide et naturelle.

Une fois produit, ce son se propage dans l'air sous forme d'ondes sonores, qui

---

transportent l'énergie acoustique jusqu'à l'oreille de l'auditeur, permettant ainsi la communication orale.

### 2.1.2 Perception auditive

L'audition humaine repose sur un mécanisme complexe qui permet de transformer les ondes sonores en signaux électriques interprétés par le cerveau. L'oreille, organe central de ce processus, se compose de trois parties principales : l'oreille externe, l'oreille moyenne et l'oreille interne, chacune jouant un rôle spécifique dans la transmission et l'analyse des sons.

**Oreille externe :** Elle comprend le pavillon et le conduit auditif. Le pavillon, de forme concave, capte les ondes sonores provenant de l'environnement et les dirige vers le conduit auditif. Ces ondes atteignent ensuite le tympan, qui commence à vibrer[2].

**Oreille moyenne :** C'est une cavité remplie d'air où se trouvent trois petits osselets (le marteau, l'enclume et l'étrier). Fixé au tympan, le marteau transmet les vibrations à l'enclume, qui les communique ensuite à l'étrier. Ce dernier est en contact avec la fenêtre ovale de l'oreille interne. Le rôle des osselets est d'amplifier les vibrations sonores pour faciliter leur transmission au liquide de la cochlée. La trompe d'Eustache, reliant l'oreille moyenne à la gorge, assure quant à elle l'équilibre de pression entre l'intérieur et l'extérieur de l'oreille, condition indispensable au bon fonctionnement du tympan [2].

**Oreille interne :** Elle renferme la cochlée, structure en forme de spirale remplie de liquide, qui joue un rôle central dans l'audition. Les vibrations transmises par l'étrier agitent ce liquide, mettant en mouvement les cellules ciliées tapissant la cochlée. Ces cellules sensorielle transforment les vibrations mécaniques en impulsions électriques, envoyées ensuite au cerveau via le nerf auditif. L'oreille interne comprend aussi les canaux semi-circulaires, spécialisés dans l'équilibre, qui détectent les mouvements de la tête grâce au déplacement des fluides et transmettent cette information par le nerf vestibulaire [2].

**Voyage du son vers le cerveau :** Les impulsions électriques issues de la cochlée circulent le long du nerf auditif pour atteindre différents centres nerveux, avant d'être traitées dans le cortex auditif, situé dans le lobe temporal. C'est à ce niveau que les sons sont triés, interprétés et mémorisés. Ce processus permet de reconnaître certaines informations sonores pertinentes tout en filtrant le bruit de fond.

### 2.1.3 Analogies avec la reconnaissance vocale artificielle

Le fonctionnement du système auditif humain a souvent été comparé aux approches de reconnaissance vocale artificielle. En effet, dans les deux cas, le processus commence par la capture des ondes sonores : l'oreille externe chez l'humain et le microphone dans

---

les systèmes artificiels. Ces ondes sont ensuite converties en un autre type de signal : les vibrations mécaniques deviennent des impulsions électriques dans le système nerveux, tandis que les signaux acoustiques sont transformés en représentations numériques telles que les coefficients cepstraux (MFCC) dans les systèmes automatiques, nous allons voir dans les sections qui suivent quelques types de ces représentations [3].

La cochlée, avec ses cellules ciliées spécialisées, réalise une analyse fréquentielle du signal auditif, jouant un rôle analogue à celui des transformations spectrales utilisées en traitement automatique de la parole [4]. Enfin, tout comme le cerveau humain interprète et donne du sens aux impulsions électriques, les réseaux de neurones convolutifs et modèles modernes comme Whisper d'OpenAI effectuent une classification et une interprétation des caractéristiques extraites afin de reconnaître et transcrire la parole [5, 6].

## 2.2 Modèles de reconnaissance vocale

La reconnaissance vocale, ou reconnaissance automatique de la parole, consiste à analyser la voix humaine afin de la transformer en texte ou en requêtes interprétables par un système informatique. Ce processus commence par la capture des signaux sonores via un microphone, qui sont ensuite convertis en données numériques et traités à l'aide de techniques d'intelligence artificielle, notamment le deep learning. L'objectif principal des modèles de reconnaissance vocale est de fournir une traduction fiable de la parole en informations exploitables, permettant à la machine de répondre ou d'agir de manière appropriée. Dans de nombreux systèmes, cette étape est suivie par une analyse de la compréhension du langage naturel (Natural Language Understanding) pour que la machine puisse saisir le sens des mots et des phrases. Ces technologies sont largement utilisées dans les assistants vocaux tels que Siri, Alexa ou Google Assistant, dans les logiciels de dictée automatique, ainsi que dans les services clients interactifs et les commandes vocales sur ordinateurs et mobiles [7].

### 2.2.1 Caractéristiques acoustiques du son

Tout comme le cerveau humain interprète les sons grâce aux vibrations captées par les oreilles, les systèmes d'analyse vocale s'appuient sur une représentation numérique du signal sonore. Cette transformation du son en données permet de révéler les éléments qui le caractérisent — sa fréquence, son intensité, sa durée — et constitue la première étape vers sa compréhension par des méthodes d'apprentissage.

Parmi les caractéristiques acoustiques les plus utilisées, on retrouve :

- **Forme d'onde (Waveform)** : Représentation temporelle du signal audio, montrant les variations d'amplitude au fil du temps. Fournit des informations sur la

---

durée, le rythme et l'intensité, mais pas sur la composition fréquentielle [8].

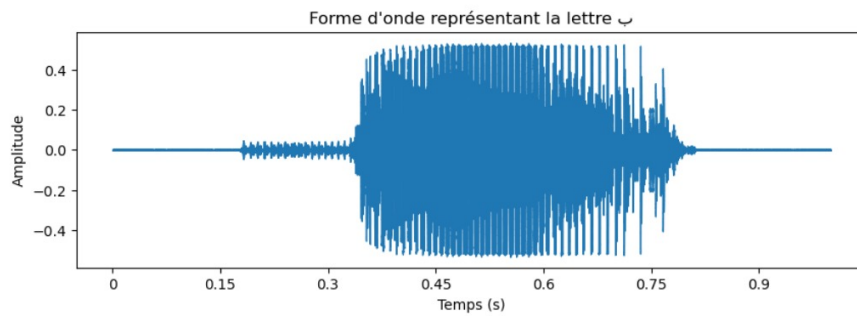


FIGURE 2.1 – Exemple de forme d'onde d'un signal audio pour la lettre BA

- **Spectrogramme** : Visualisation de l'évolution des fréquences du signal au cours du temps (transformée de Fourier à court terme). Permet de détecter les énergies présentes dans chaque bande fréquentielle [9].
- **Coefficients Cepstraux de Mel (MFCC)** : Reproduisent la perception humaine des fréquences, inspirée de la cochlée. Calculés à partir du spectrogramme via l'échelle de Mel, utilisés pour la reconnaissance automatique de la parole [10].
- **Coefficients Cepstraux de Mel (MFCC)** : Reproduisent la perception humaine des fréquences, inspirée de la cochlée. Calculés à partir du spectrogramme via l'échelle de Mel, utilisés pour la reconnaissance automatique de la parole [10]. MFCC (Melanfrequency Cepstral Coefficients) est une fonctionnalité utilisée en reconnaissance automatique de la parole et du locuteur. Il s'agit essentiellement d'un moyen de représenter le spectre de puissance à court terme d'un son, ce qui aide les machines à comprendre et à traiter la parole humaine plus efficacement. Imaginez votre voix comme une empreinte digitale unique. Les MFCC fonctionnent comme un code unique capturant les caractéristiques saillantes de votre parole et permettant aux ordinateurs de distinguer les mots et les sons. Dans les applications de reconnaissance vocale où les ordinateurs doivent traduire les mots prononcés en texte, ce code est particulièrement utile[11].

Les MFCC sont des représentations mathématiques du conduit vocal produit par l'humain lorsqu'il parle. Ce processus comprend plusieurs étapes pour capturer les caractéristiques essentielles de la parole humaine, les plus perceptibles à l'oreille humaine.

Voici comment les MFCC contribuent à la compréhension de la parole :

- **Analyse du signal** : La parole est un signal complexe caractérisé par des variations de fréquence et d'amplitude. Les MFCC permettent de décomposer ces signaux en composantes plus simples qui représentent la vitesse et les caractéristiques des variations des ondes sonores au fil du temps[11].

- Transformation de fréquence : L'être humain ne perçoit pas les fréquences sur une échelle linéaire. Par conséquent, les MFCC utilisent une échelle mel qui se rapproche étroitement de la réponse du système auditif humain, lequel est plus sensible aux variations des basses fréquences que des hautes[11].
- Représentation cepstrale : Après sa transformation à l'échelle mel, le signal est reconverti en une représentation temporelle appelée cepstre. Ce dernier sépare la variation périodique du signal (hauteur) de la variation lente (timbre), se concentrant sur cette dernière, qui véhicule la majeure partie de l'information nécessaire à la reconnaissance de la parole [11].

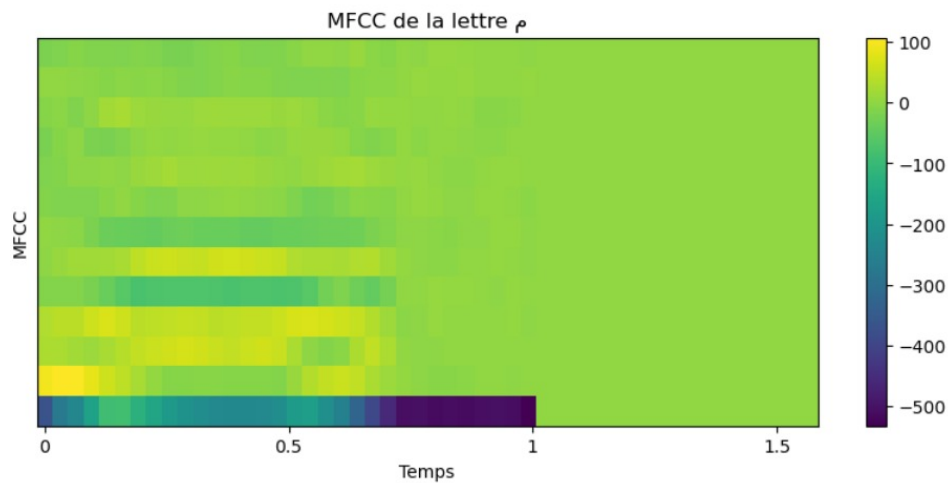


FIGURE 2.2 – Représentation MFCC pour la lettre MIM

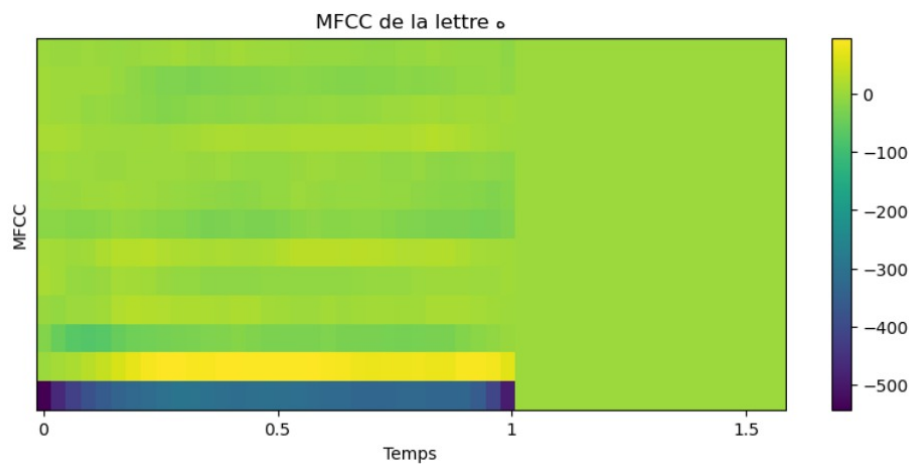


FIGURE 2.3 – Représentation MFCC pour la lettre HAA

---

### 2.2.2 Réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutifs (CNN) sont largement utilisés pour la reconnaissance vocale, en particulier pour traiter les représentations spectrales du signal audio, comme les MFCC ou les spectrogrammes.

L'architecture typique d'un CNN comprend :

- **Couches convolutives** : appliquent des filtres pour extraire automatiquement des motifs et caractéristiques locales du signal.
- **Couches de pooling** : réduisent la dimension des données tout en conservant les informations importantes.
- **Couches entièrement connectées (fully connected)** : effectuent la classification finale ou la prédiction à partir des caractéristiques extraites.

Cette architecture permet au modèle d'apprendre progressivement des caractéristiques de plus en plus complexes, allant des motifs simples (fréquences locales, formes d'onde) jusqu'à des structures plus globales du signal vocal [5].

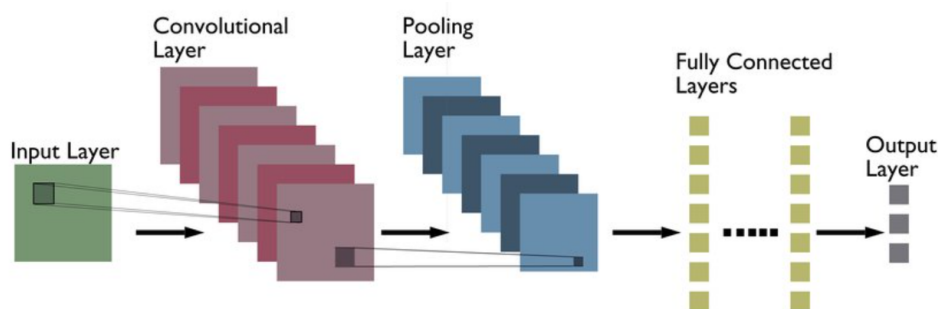


FIGURE 2.4 – L'architecture des réseaux CNN

### 2.2.3 Whisper d'OpenAI

**Whisper** est un modèle de reconnaissance automatique de la parole développé par OpenAI. Il repose sur une architecture **Transformer** qui permet de traiter des séquences temporelles complexes et de générer du texte à partir d'audio brut. Whisper a été entraîné sur des millions d'heures de données multilingues et multi-accentuées, ce qui lui confère une robustesse exceptionnelle face aux différents locuteurs, accents et bruits de fond [6].

L'architecture générale de Whisper comprend :

- **Prétraitement et encodage** : le signal audio est converti en spectrogramme, capturant les variations temporelles et fréquentielles.
- **Blocs Transformer** : ces blocs analysent les séquences de caractéristiques extraites, capturant les relations à court et long terme dans le signal vocal.



- 
- **Décodage autoregressif** : le décodeur génère la transcription textuelle de manière séquentielle, en tenant compte du contexte global et des dépendances entre les mots.

Whisper est particulièrement performant pour :

- La transcription multilingue et multiaspect (différents accents et locuteurs).
- La robustesse face au bruit et aux conditions audio variées.
- L'intégration directe dans des systèmes de traduction ou d'assistants vocaux.

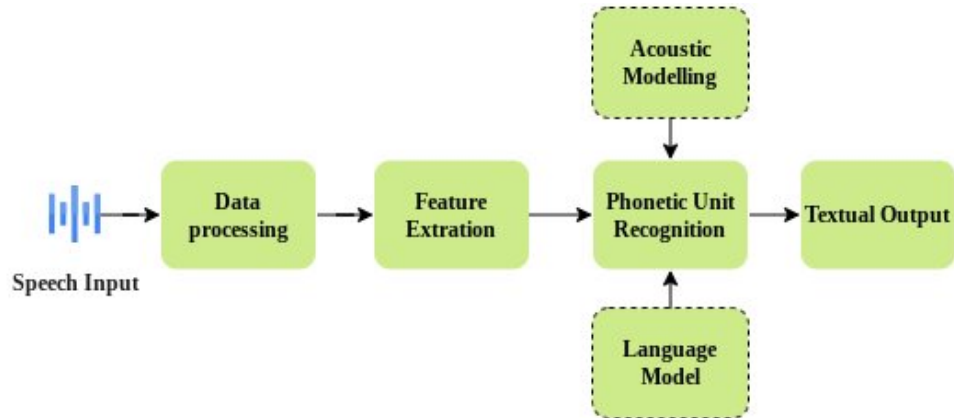









FIGURE 2.5 – L'architecture de Whisper

---

## 2.3 Outils et technologies utilisés

Pour le développement du prototype de correction de la prononciation, les outils suivants ont été utilisés :

-  **Python** : Langage de programmation principal pour le traitement audio, l'extraction de caractéristiques et l'entraînement des modèles.
-  **OS** : Module Python pour la gestion des fichiers, répertoires et opérations système nécessaires au traitement des données audio.
-  **Librosa** : Librairie Python spécialisée dans l'analyse et le traitement de signaux audio, utilisée pour extraire les spectrogrammes et MFCC.
-  **NumPy** : Librairie pour le calcul scientifique, manipulation des tableaux et opérations mathématiques sur les données audio.
-  **TensorFlow** : Framework open-source pour la construction et l'entraînement de réseaux de neurones, utilisé pour développer les modèles CNN et Transformer.
-  **Gradio** : Outil pour créer des interfaces interactives simples permettant aux utilisateurs de tester le modèle de correction de prononciation en ligne.
-  **Jupyter Notebook** : Environnement interactif pour l'expérimentation, le prototypage et le suivi des résultats des modèles d'apprentissage automatique.

# Chapitre 3

## Conception et Réalisation du prototype

### 3.1 Analyse du besoin

L'objectif principal du prototype est de fournir une solution automatisée pour l'amélioration de la prononciation des apprenants de la langue arabe. Pour cela, il est nécessaire de répondre à plusieurs besoins fonctionnels et techniques :

- **Correction de la prononciation** : Identifier les erreurs de prononciation et fournir un retour précis et compréhensible pour l'utilisateur.
- **Compatibilité avec différents types d'entrées audio** : Le système doit pouvoir traiter des fichiers audio variés (wav, mp3, etc.) et des enregistrements en direct via microphone.
- **Extraction fiable des caractéristiques acoustiques** : Les informations pertinentes du signal audio doivent être extraites pour permettre au modèle d'évaluer correctement la prononciation.
- **Interface utilisateur intuitive** : L'utilisateur doit pouvoir interagir facilement avec le système, tester sa prononciation et recevoir un retour clair.
- **Performance et rapidité** : Le traitement des fichiers audio et la génération des corrections doivent se faire rapidement pour offrir une expérience utilisateur fluide.
- **Adaptabilité et évolutivité** : Le prototype doit être conçu de manière modulaire pour permettre l'intégration de nouvelles fonctionnalités ou de modèles améliorés dans le futur.

Cette analyse constitue la base pour la conception du prototype et oriente les choix technologiques et méthodologiques qui seront présentés dans les sections suivantes.

## 3.2 Architecture proposée du prototype

Le prototype combine deux modèles d'intelligence artificielle pour corriger la prononciation et transcrire la parole des utilisateurs : **CNN** et **Whisper**.

- **CNN** : Conçu pour corriger la prononciation des lettres arabes. Le modèle analyse les caractéristiques acoustiques des lettres prononcées et fournit un feedback sur la justesse de la prononciation.
- **Whisper** : Conçu pour transcrire des expressions courantes enregistrées par les utilisateurs. Ces transcriptions sont ensuite comparées à une base de données contenant les expressions correctes, permettant de détecter les erreurs et de guider l'utilisateur pour améliorer sa prononciation.

Le pipeline global fonctionne de la manière suivante :

1. Acquisition de l'audio de l'utilisateur (fichier ou enregistrement direct).
2. Prétraitement du signal (normalisation, filtrage, segmentation).
3. Pour les lettres : extraction des caractéristiques et correction via CNN.
4. Pour les expressions : transcription via Whisper et comparaison avec la base de données.
5. Feedback présenté à l'utilisateur via l'interface (Gradio).



FIGURE 3.1 – Schéma de l'architecture du prototype combinant CNN et Whisper

Cette architecture permet de combiner les forces des deux modèles : la précision de CNN pour la correction des lettres et la puissance de Whisper pour la transcription

---

et l'évaluation des expressions complètes. Grâce à cette combinaison, le système fournit un apprentissage de la prononciation plus complet et adapté aux besoins des apprenants.

### 3.3 Développement du modèle CNN

Le modèle CNN conçu pour la correction de la prononciation des lettres arabes est composé de plusieurs blocs convolutionnels et de couches entièrement connectées. Il comprend :

- **Trois blocs convolutionnels** avec des filtres de tailles croissantes (32, 64 et 128), chacun suivi d'une normalisation par batch et d'un *max pooling* pour réduire les dimensions spatiales.
- **Couches fully connected** : deux couches denses de 256 et 128 neurones avec des *dropout* pour régulariser et éviter le surapprentissage.
- **Couche de sortie** adaptée au nombre de classes correspondant aux lettres arabes à corriger, avec activation *softmax*.

Le modèle est entraîné avec l'optimiseur **Adam**, la fonction de perte **categorical crossentropy** et évalué sur les métriques suivantes : **accuracy** et **val\_accuracy**. Un mécanisme d'*EarlyStopping* est utilisé pour arrêter l'entraînement en cas d'absence d'amélioration sur la validation, ce qui assure la robustesse et la généralisation du modèle.

### 3.4 Développement du module Whisper

Le module Whisper a été intégré pour la transcription des expressions courantes prononcées par l'utilisateur. Son fonctionnement repose sur le modèle **Whisper d'OpenAI** (version *medium* pour une meilleure précision) et comprend les étapes suivantes :

- **Prétraitement audio** : Les enregistrements des utilisateurs sont normalisés, éventuellement filtrés pour réduire le bruit, puis sauvegardés dans un chemin fixe.
- **Transcription** : Le modèle Whisper transcrit le signal audio en texte en langue arabe. Une normalisation du texte est effectuée pour retirer les caractères non pertinents.
- **Comparaison avec la base de données** : Les expressions transcrites sont comparées à une base de données contenant des mots et expressions attendus.

Cette étape permet de vérifier et corriger la prononciation ou l'expression de l'utilisateur.

L'évaluation du module repose sur la **précision de la transcription** et la correspondance avec la base de données, fournissant ainsi un retour utile pour l'apprentissage et la correction de la prononciation.

### 3.5 Mise en place d'une interface utilisateur

Pour rendre le prototype accessible et interactif, une interface conviviale a été développée à l'aide de **Gradio**. Cette interface combine les deux modèles (CNN et Whisper) et offre deux fonctionnalités principales :

- **Correction des lettres arabes** : Le module CNN évalue la prononciation des lettres isolées et fournit un retour immédiat à l'utilisateur.
- **Correction des expressions arabes courantes** : Le module Whisper transcrit les expressions prononcées, puis les compare à une base de données d'expressions attendues afin de détecter et corriger d'éventuelles erreurs de prononciation.

L'interface fonctionne en **temps réel**, ce qui permet de traiter aussi bien les fichiers audio préenregistrés que les enregistrements effectués directement depuis le micro. Cette approche interactive offre un retour instantané et intuitif, facilitant l'apprentissage et la pratique de la langue arabe.

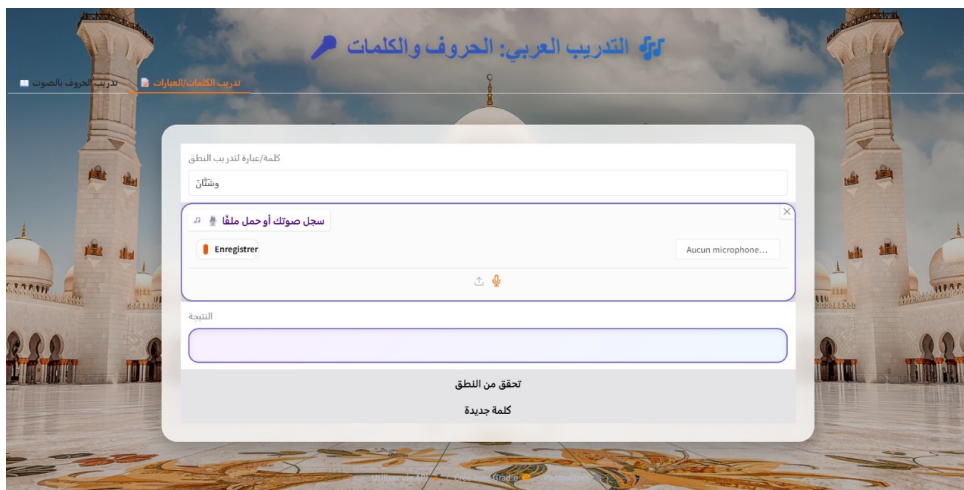


FIGURE 3.2 – Interface utilisateur.

# Chapitre 4

## Résultats et Évaluation

### 4.1 Résultats et évaluation des performances

### 4.2 Résultats et évaluation des performances

Les résultats obtenus à partir du modèle CNN et du modèle Whisper démontrent l'efficacité de l'approche proposée pour la correction de la prononciation des lettres et des expressions arabes.

#### Résultats du modèle CNN

Le modèle CNN, conçu pour la reconnaissance et la correction de la prononciation des lettres arabes, a montré d'excellentes performances. Après entraînement sur les coefficients MFCC extraits des enregistrements, le modèle a atteint une **précision de test de 98.25%** et une **perte de test de 0.0667**.

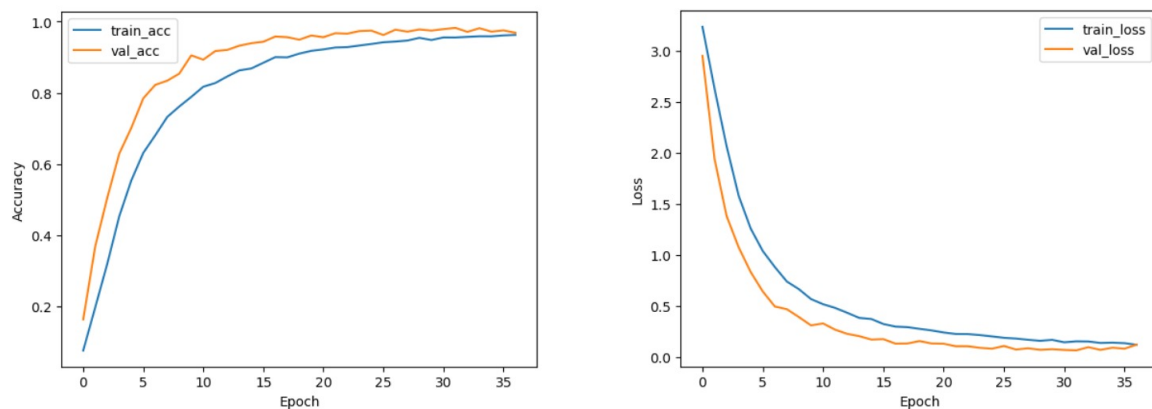


FIGURE 4.1 – Courbes d'entraînement du modèle CNN : précision (à gauche) et perte (à droite).

Les courbes de précision et de perte montrent une convergence rapide du modèle

dès les premières époques, sans signe de surapprentissage notable. Cela indique une bonne généralisation sur les données de validation.

## Résultats du modèle Whisper

Le modèle Whisper a été utilisé pour la transcription automatique des expressions arabes courantes enregistrées par les utilisateurs. Les transcriptions générées ont ensuite été comparées à une base de données de référence contenant les expressions correctes, permettant ainsi de détecter et de corriger les erreurs de prononciation.

Au départ, une version **small** de Whisper a été utilisée. Cependant, cette dernière s'est révélée **insuffisante en termes de précision**, car elle ne parvenait pas toujours à transcrire correctement certaines expressions, notamment en présence de variations de ton ou de légers bruits ambiants. Pour améliorer les performances, la version **medium** du modèle a ensuite été adoptée. Cette mise à niveau a permis d'obtenir des **résultats nettement plus fiables et cohérents**, même dans des conditions d'enregistrement plus naturelles, renforçant ainsi la robustesse du système global.

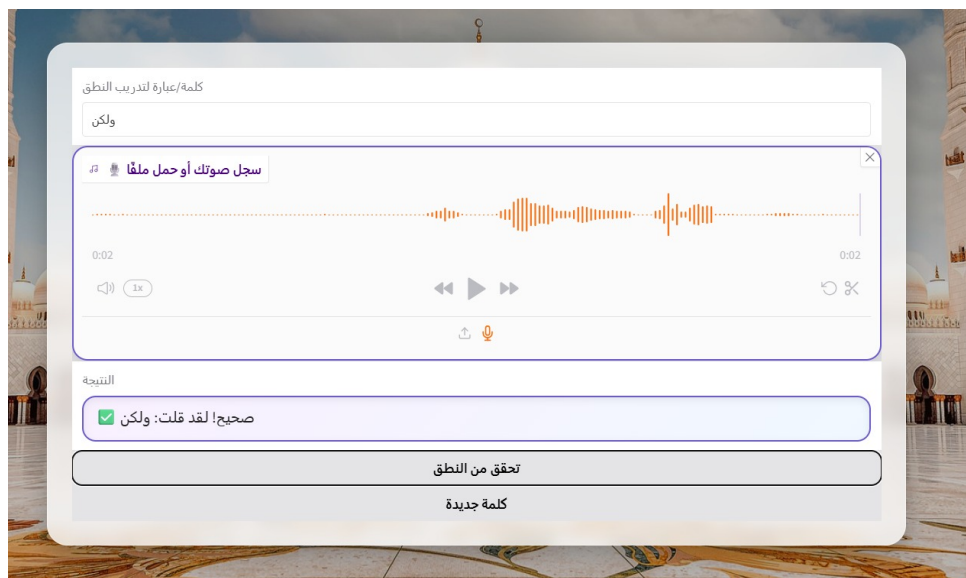


FIGURE 4.2 – Correction avec Whisper

## Évaluation globale

L'intégration des deux modèles dans une seule architecture a permis d'obtenir un système performant et convivial :

— **Correction de la prononciation des lettres arabes** : assurée efficacement



par le CNN avec une précision supérieure à 98%.

- **Correction des expressions arabes** : réalisée grâce à Whisper et au module de comparaison avec la base de données.
- **Interface utilisateur** : développée avec Gradio, permettant une utilisation intuitive et un traitement en temps réel des enregistrements vocaux.

Ainsi, les performances atteintes valident la pertinence du prototype proposé et confirment la complémentarité entre le traitement de la parole et les techniques de Deep Learning pour la correction de prononciation.

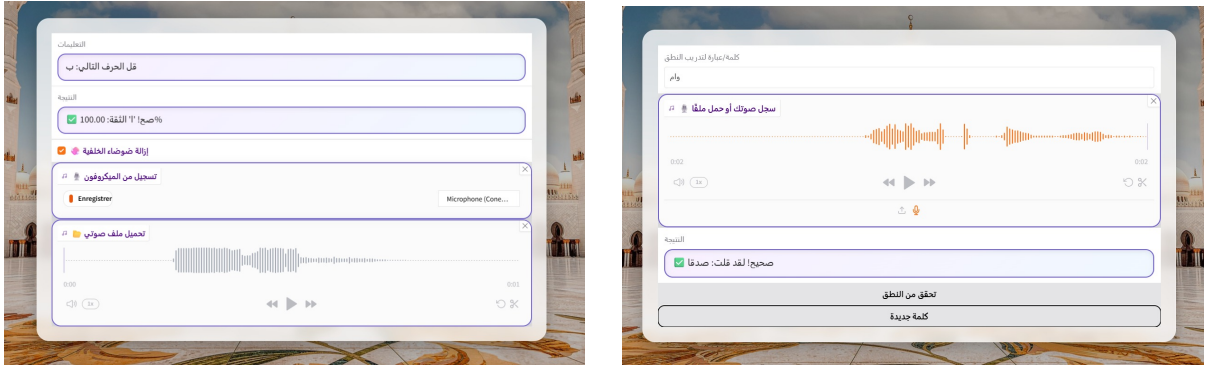


FIGURE 4.3 – Evaluation globale des deux modèles

### 4.3 Discussion et limites

Les résultats obtenus à partir du modèle CNN montrent une performance remarquable avec une précision de plus de **98%**, confirmant la capacité du modèle à reconnaître efficacement les lettres arabes à partir des caractéristiques MFCC. Toutefois, ces performances restent étroitement liées à la qualité du signal audio et à l'environnement d'enregistrement. En présence de bruit de fond ou lorsque le matériel d'enregistrement est de faible qualité, le taux de reconnaissance tend à diminuer.

Un autre point important concerne la **variation des prononciations selon les dialectes arabes**. Le modèle CNN, entraîné principalement sur un ensemble de données standardisées, n'a pas pu se généraliser efficacement à toutes les variantes de la langue arabe. Les différences d'accent, de ton et de rythme vocal influencent considérablement la reconnaissance des sons, ce qui limite la robustesse du modèle face à la diversité linguistique des utilisateurs.

Concernant le modèle Whisper, bien qu'il ait offert une **très bonne qualité de transcription** après le passage à la version **medium**, certaines limites persistent. Le modèle peut encore confondre des mots proches phoniquement ou introduire de

---

légères erreurs dans les accents ou la segmentation des phrases. De plus, l'utilisation d'un modèle de taille moyenne implique une **consommation de ressources plus importante** (RAM et GPU), ce qui peut ralentir le traitement sur des machines à capacité limitée.

En ce qui concerne l'**interface développée**, elle a démontré une grande simplicité d'utilisation et une interaction fluide avec les utilisateurs. Cependant, quelques améliorations peuvent être envisagées, notamment dans la gestion des enregistrements multiples, l'optimisation du temps de réponse en temps réel, et l'ajout d'un retour audio ou visuel pour rendre l'expérience plus interactive.

Enfin, une autre limite du projet réside dans la **taille et la diversité du jeu de données**. L'extension du corpus vocal pour inclure un plus grand nombre de locuteurs, d'âges, d'accents et de conditions d'enregistrement permettrait d'améliorer la capacité de généralisation et la robustesse des modèles proposés.

## 4.4 Perspectives futures

- **Extension multilingue** : Adapter le prototype à d'autres langues afin de permettre l'apprentissage et la correction de la prononciation pour un public plus large.
- **Enrichissement des bases de données vocales** : Constituer des corpus audio plus variés pour chaque langue, incluant différents accents et niveaux de maîtrise, afin d'améliorer la précision des modèles.
- **Personnalisation de l'apprentissage** : Intégrer des profils d'apprenant permettant de suivre la progression individuelle et de proposer des exercices adaptés à chaque utilisateur.
- **Intégration d'un feedback multimodal** : Combiner retour audio, visuel et textuel pour rendre l'apprentissage plus interactif et efficace.
- **Optimisation des modèles** : Améliorer les performances du CNN et de Whisper pour réduire le temps de traitement et augmenter la précision de la correction.
- **Déploiement sur différentes plateformes** : Étendre l'accès via mobile, web et applications éducatives pour une utilisation plus large et flexible.

# Conclusion générale

Ce projet, qui s'inscrit dans le domaine de l'apprentissage automatique appliqué à la reconnaissance vocale, a constitué une véritable opportunité de mettre en pratique des concepts théoriques et de travailler sur un problème concret : la correction de la prononciation en langue arabe. Deux approches principales ont été explorées : les réseaux de neurones convolutionnels (CNN) pour l'analyse et la correction des lettres arabes, et le modèle Whisper pour la transcription et la correction des expressions courantes. L'entraînement et l'évaluation de ces modèles sur des données vocales arabophones ont permis d'obtenir des résultats fiables et pertinents.

À travers ce projet, j'ai pu approfondir mes connaissances en apprentissage automatique, traitement du signal et reconnaissance vocale, tout en développant des compétences pratiques essentielles pour la mise en œuvre de solutions basées sur l'IA. Cette expérience a également permis de comprendre les limites et les défis liés à la variabilité des voix, des accents et des intonations, et d'apprécier l'importance d'une approche méthodique dans la conception et l'évaluation des modèles.

Enfin, ce projet ouvre des perspectives prometteuses pour l'amélioration des outils d'apprentissage des langues. L'intégration de jeux de données plus variés, le développement de modèles capables de gérer différents accents et dialectes, ainsi que la création d'interfaces interactives pour un apprentissage plus personnalisé pourraient rendre ces technologies encore plus efficaces et accessibles. Ce travail témoigne du potentiel des outils d'IA pour enrichir l'expérience des apprenants et faciliter l'acquisition de compétences linguistiques.

# Table des figures

1.1	Localisation de l'entreprise YAFA Technologies . . . . .	8
1.2	services proposées et clients . . . . .	8
2.1	Exemple de forme d'onde d'un signal audio pour la lettre BA . . . .	13
2.2	Représentation MFCC pour la lettre MIM . . . . .	14
2.3	Représentation MFCC pour la lettre HAA . . . . .	14
2.4	L'architecture des réseaux CNN . . . . .	15
2.5	L'architecture de Whisper . . . . .	16
3.1	Schéma de l'architecture du prototype combinant CNN et Whisper	19
3.2	Interface utilisateur. . . . .	21
4.1	Courbes d'entraînement du modèle CNN : précision (à gauche) et perte (à droite). . . . .	22
4.2	Correction avec Whisper . . . . .	23
4.3	Evaluation globale des deux modèles . . . . .	24

# Liste des acronymes

**CNN** Convolutional Neural Network

**MFCC** Mel Frequency Cepstral Coefficients

**IA** Intelligence Artificielle

**API** Application Programming Interface

**ASR** Automatic Speech Recognition

**DL** Deep Learning

**ML** Machine Learning

**SIG** Systèmes d'Information Géographique

**GPU** Graphics Processing Unit (Unité de Traitement Graphique)

**RAM** Random Access Memory (Mémoire Vive)

**STFT** Short-Time Fourier Transform (Transformée de Fourier à Court Terme)

**NLU** Natural Language Understanding (Compréhension du Langage Naturel)

# Bibliographie

- [1] Scotto Di Carlo, « Voix humaine », dans *Quid*, 2008.
- [2] Mayo Clinic, « Comment entends-tu ? », juin 2025.
- [3] L. R. Rabiner et B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
- [4] S. Rosen, « Temporal information in speech : Acoustic, auditory and linguistic aspects », *Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences*, vol. 336, no. 1278, pp. 367–373, 1992.
- [5] G. Hinton et al., « Deep Neural Networks for Acoustic Modeling in Speech Recognition : The Shared Views of Four Research Groups », *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [6] A. Radford et al., « Robust Speech Recognition via Large-Scale Weak Supervision », *Proceedings of NeurIPS*, 2023.
- [7] A. Crochet-Damais, « Reconnaissance vocale : définition, algorithmes et fonctionnement », mai 2022.
- [8] L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
- [9] D. O’Shaughnessy, *Speech Communications : Human and Machine*, IEEE Press, 2000.
- [10] S. Davis, P. Mermelstein, « Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences », *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1980.
- [11] Geeks for Geeks, « Coefficients cepstraux de fréquence Mel (MFCC) pour la reconnaissance vocale », 23 juillet 2025.