Université Sorbonne Paris Nord
Sup Galilée
Spécialité Informatique

Conduite et gestion de projet

Spécialité : INFO2

Interface de Génération sonore par IA

Étudiants :
Seyfeddine JOUINI
Yosra SASSI
Sirine TLILI
Hasna ELGARANI
Kaoutar BRAHIMI

Client :
Vittascience

7 avril 2025

Table des matières

Introduction

Ce document rend compte de l'exécution du projet de développement d'un module de génération sonore par intelligence artificielle, ayant pour but la mise en place d'une interface intuitive et interactive, pour la découverte et l'expérimentation par les élèves des technologies de synthèse vocale et de génération musicale, de façon pédagogique et immersive.

1. Contexte du projet

Le projet de génération sonore par intelligence artificielle vise à fournir une plateforme éducative innovante destinée aux enseignants et aux élèves, permettant de générer de la musique et des voix synthétiques via l'IA.

Les fonctionnalités clés du projet incluent les cas d'utilisation suivants :

- **Synthèse vocale** : Permettre à l'utilisateur de convertir un texte écrit en parole, en choisissant des paramètres tels que la voix, la vitesse et la hauteur.
- **Reconnaissance vocale** : Convertir l'entrée vocale de l'utilisateur en texte écrit, avec une prise en charge locale sans connexion Internet.
- **Conversion de la parole** : Assurer une interaction fluide entre l'utilisateur et le système, en permettant une communication naturelle via la reconnaissance et la synthèse vocale.
- **Synchronisation labiale** : Synchroniser les mouvements de la bouche d'un visage dans une image en entrée avec un audio.
- **Génération musicale** : Générer automatiquement de la musique (une mélodie) à partir d'un thème ou d'un prompt donné.

Le diagramme de cas d'utilisation ci-dessous modélise les interactions entre les acteurs principaux (enseignants, élèves) et les fonctionnalités du système. Il permet de visualiser de manière synthétique les objectifs opérationnels du projet et les relations entre les composants.

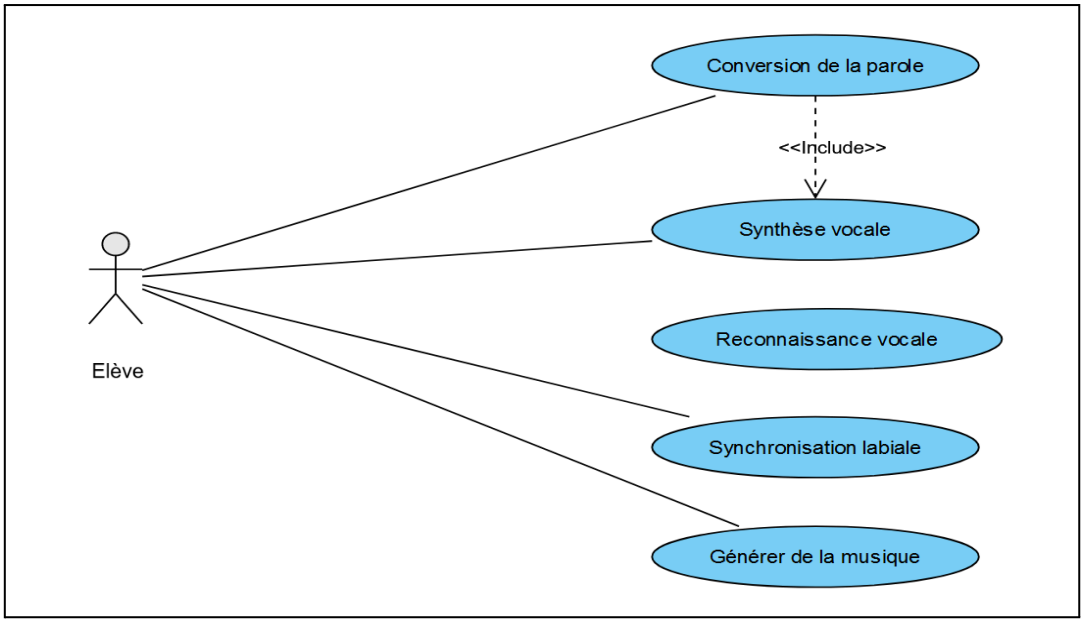

Figure 1 : Diagramme de cas d'utilisation

2. Méthodologie de travail

Conformément à la méthodologie Agile (SCRUM) définie dans le cahier des charges, le projet a été structuré en sprints itératifs de 4 semaines, permettant une adaptation continue aux défis techniques. Cette approche a favorisé :

- Une flexibilité maximale pour ajuster les priorités et les solutions techniques.
- Des livraisons fréquentes de fonctionnalités opérationnelles, validées en collaboration avec le client.
- Une transparence totale via des réunions régulières

Organisation de l'équipe :

L'équipe projet a été organisée de manière à couvrir toutes les compétences nécessaires pour le succès du projet.

Membre	Rôle	Email
Seyfeddine Jouini	Chef de projet	jouini.seyf.js@gmail.com
Yosra Sassi	Responsable Back-End	sassiyosra66@gmail.com
Hasna Elgarani	Développeur Front-End	elgaranihasna@gmail.com
Sirine Tlili	Spécialiste IA	cyrinetlili07@gmail.com
Kaoutar Brahimi	Développeur Back-End	kaoutarbrahimi28@gmail.com

Tableau 1 : Répartition des rôles

3. Déroulement du travail et évolution des objectifs

Cette partie retrace l'évolution itérative du projet, depuis sa conception jusqu'à sa finalisation. Adoptant une méthodologie Agile (Scrum), le développement s'est articulé autour de 5 sprints de 4 semaines, chacun visant à livrer des fonctionnalités opérationnelles tout en intégrant les contraintes techniques découvertes en cours de route.

Nous décrirons ci-dessous, pour chaque sprint :

- Objectifs initiaux et livrables prévus
- Problèmes rencontrés et décisions prises
- Ajustements des objectifs

Le tableau ci-dessous résume les adaptations majeures au fil des sprints, illustrant comment les objectifs initiaux ont été affinés pour répondre aux défis techniques.

Sprint	Objectif initial	Problème majeur	Ajustement
Sprint 1	Prototype Synthèse vocale (API Web Speech Synthesis)	Problème de qualité de la voix.	- Utilisation des voix du navigateur adaptées à chaque langue
Sprint 2	Synchronisation labiale (transcription par modèle *Vosk* + génération d'images)	- Performance faible de la transcription par Vosk - Animation labiale peu réaliste	- Remplacement par *Whisper* - Intégration de *Wav2Lip*
	Amélioration Synthèse vocale	Besoin d'enregistrer l'audio après sa génération	Utilisation de *MediaRecorder*
Sprint 3	Synchronisation labiale (*Wav2Lip*)	Problème de licence (non utilisable dans les produits commerciaux)	Retour à une solution de synchronisation manuelle.
	Reconnaissance vocale (*API SpeechRecognition*)	Transcription imprécise	Migration vers *Whisper* (OpenAI)
Sprint 4	Conversion de la parole avec *YourTTS*	Licence restrictive	Migration vers *XTTS*
	Génération musicale : Génération de paroles avec *GPT-2*	Absence de structure musicale cohérente	Focus sur la génération exclusive de mélodies
Sprint 5	Synchronisation labiale (bouche animée)	Effet visuel indésirable	Application d'un léger flou sur les contours de la bouche pour adoucir les transitions
	Génération musicale : mélodie (bibliothèque Python *music21*)	- Qualité musicale artificielle et peu réaliste - Manque de naturel dans les arrangements instrumentaux	Remplacement par l'API *Stable Audio 2* de Stability AI

Tableau 2 : Evolution du projet

4. Fonctionnalités du projet

Synthèse vocale :

Cette fonctionnalité a pour objectif de permettre à un utilisateur de générer un son synthétique réaliste à partir d'un texte saisi, en utilisant l'API Web Speech (SpeechSynthesis). Cette solution fonctionne entièrement en local, sans dépendance à une connexion Internet, répondant ainsi à des besoins en accessibilité et en utilisation hors ligne.

La figure ci-dessous montre l'interface de synthèse vocale avec les différents paramètres ajustables :

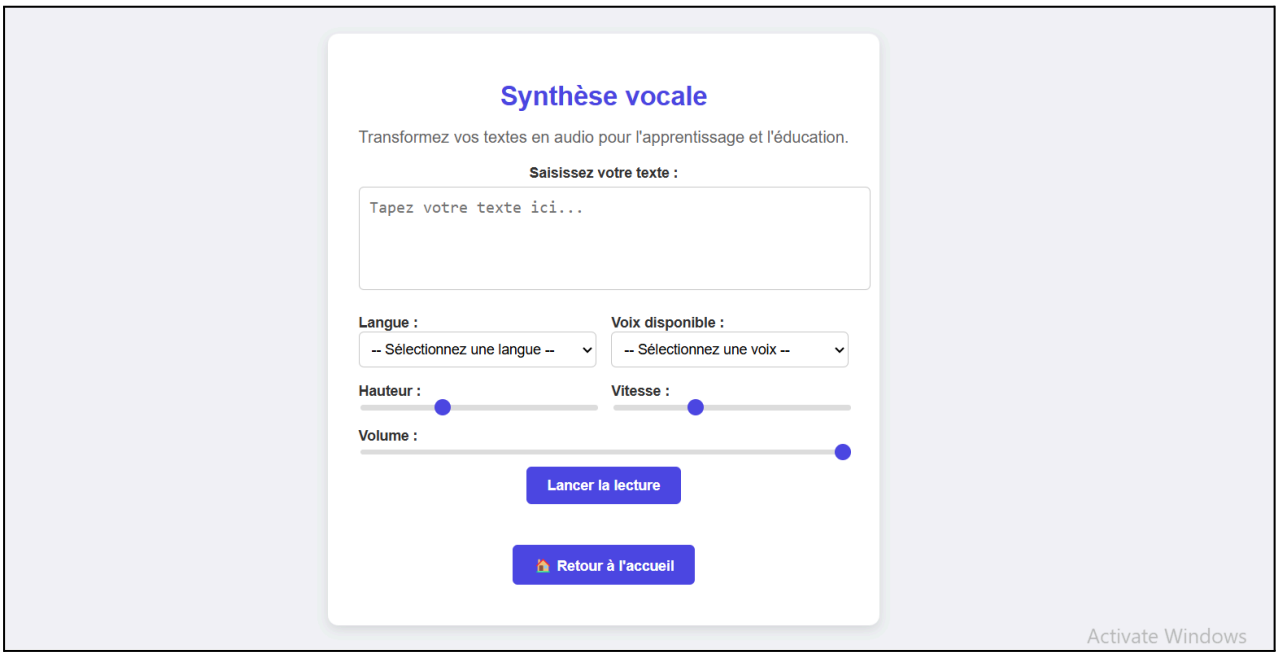

Figure 2 : Interface de la synthèse vocale

Fonctionnalités clés :

- Interface intuitive avec prévisualisation audio
- Réglage fin de la voix (paramètres personnalisables) :
  - La hauteur de la voix (pitch) : ajuster le timbre et apporter des variations émotionnelles
  - La vitesse de lecture (rate) : ralentir ou accélérer l'énonciation
  - Le volume sonore (volume) : ajustable selon les besoins
  - Le choix de la langue (support multilingue français/anglais)
- Export des enregistrements en MP3

Les figures suivantes présentent l'interface optimisée de synthèse vocale avec ses fonctionnalités clés :

Figure 3 : Panneau de contrôle de la synthèse vocal

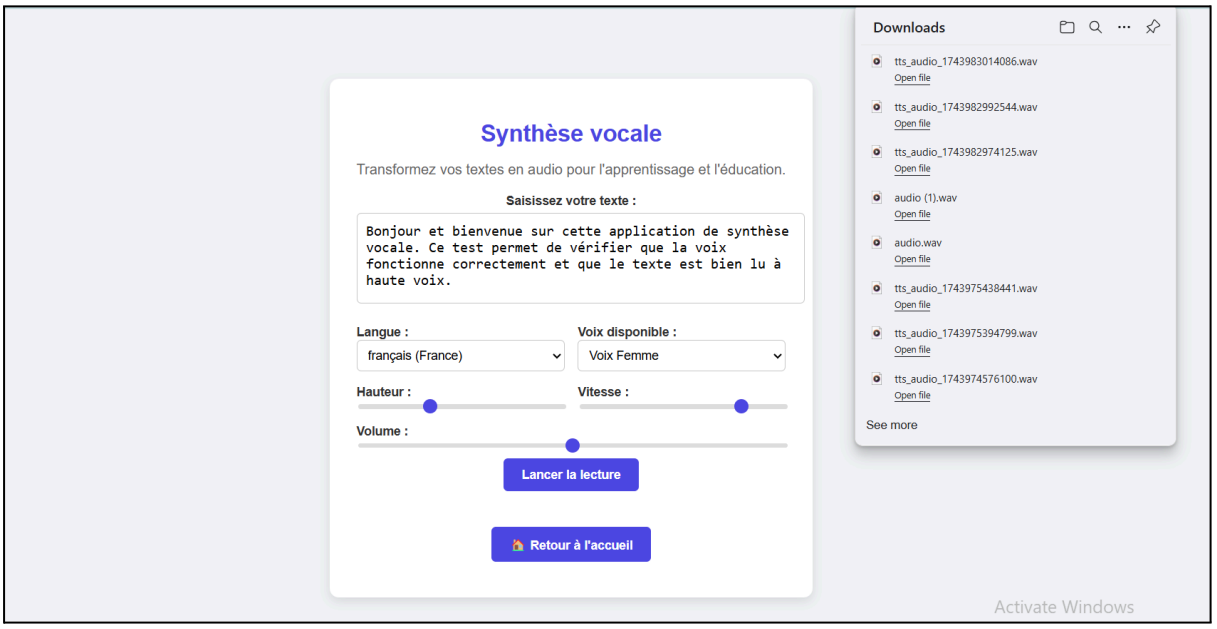

Figure 4 : Affichage du résultat de la synthèse vocale

Reconnaissance vocale :

Cette fonctionnalité vise à permettre à un utilisateur de convertir une entrée vocale en texte écrit en utilisant Whisper, un modèle de reconnaissance vocale performant développé par OpenAI. Contrairement aux solutions basées sur des API de navigateur, Whisper offre une meilleure précision et peut être exécuté en local pour une utilisation hors ligne.

Comme montré ci-dessous, l'interface permet à l'utilisateur de déclencher la reconnaissance vocale d'un simple clic. Elle affiche en temps réel le texte reconnu, offrant ainsi une interaction instantanée. Cette conception garantit une utilisation intuitive et efficace, avec un retour immédiat du texte dicté.

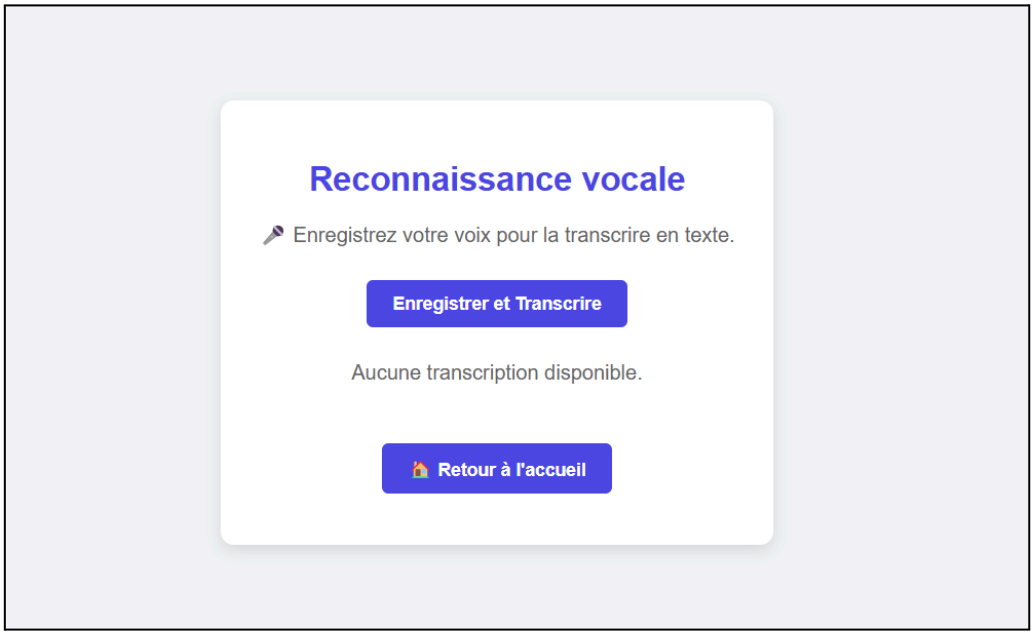

Figure 5 : Interface de la Reconnaissance vocale

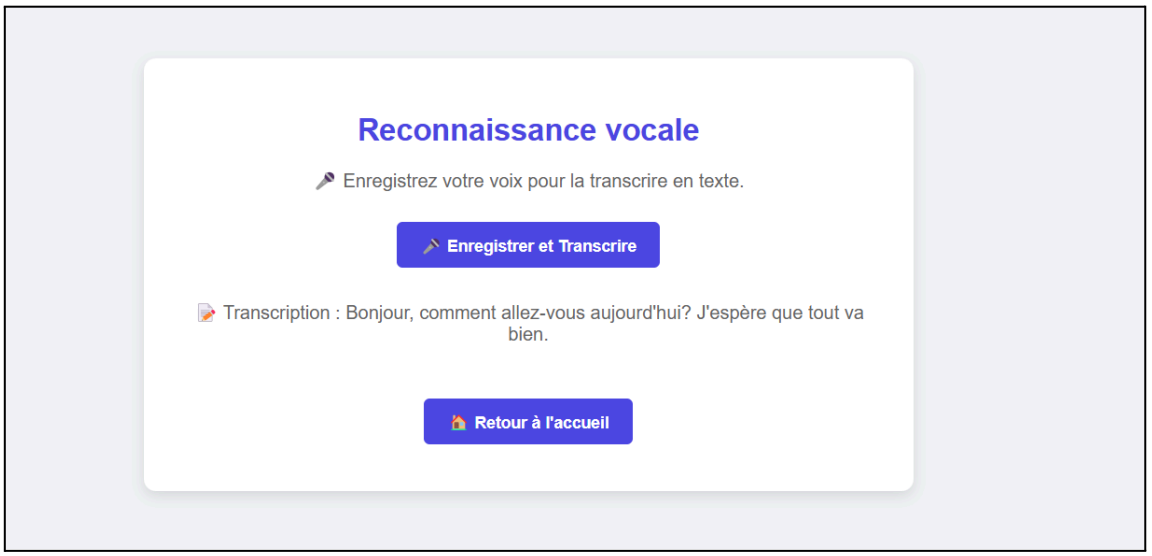

Figure 6 : Résultat de Transcription

Conversion de la parole :

La conversion de la parole consiste à transformer un signal vocal en un signal synthétisé, à travers plusieurs étapes techniques.

Tout d'abord, un signal audio est capté, généralement via un microphone. Ce signal, issu de la parole humaine, est numérisé pour être exploité par les algorithmes de traitement.

Ensuite, une étape de traitement centralisée intervient. Elle s'appuie sur des techniques avancées telles que la reconnaissance vocale, ou encore la transformation de la voix. Cette phase permet d'extraire les informations pertinentes et de préparer la synthèse vocale.

Enfin, le système produit une synthèse vocale basée sur un clonage de la voix et des émotions de l'utilisateur. Contrairement à une simple génération de texte lu, cette approche permet de reproduire fidèlement non seulement le timbre et les caractéristiques acoustiques de la voix originale, mais aussi l'expression émotionnelle transmise dans l'intonation et le rythme. Cette capacité à cloner la manière d'émettre une parole rend l'interaction plus naturelle, plus expressive.

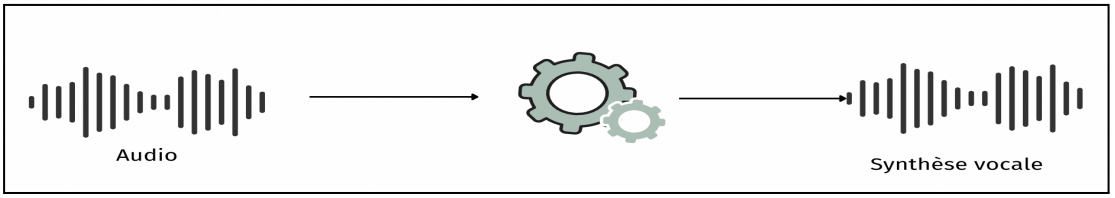

Figure 7 : Principe de la conversion de la parole

Technologies utilisées

Composant	Technologie / Outil	Rôle principal dans le projet
Langage de programmation	Python	Langage principal pour le développement du serveur et des traitements vocaux
Framework web	Flask	Serveur web léger permettant de gérer les requêtes HTTP
Synthèse vocale (TTS)	Coqui TTS (XTTS)	Génération de la voix clonée à partir du texte transcrit
Reconnaissance vocale	Whisper (OpenAI)	Transcription du signal audio en texte
Interface utilisateur	HTML / CSS + JavaScript (via Flask templates)	Interface web interactive pour dialoguer avec l'assistant vocal

Tableau 3 : Technologies de la conversion de la parole

Fonctionnement de l’application :

L’application permet de parler à un assistant vocal et d’entendre une réponse synthétique, générée automatiquement. Voici les étapes du fonctionnement :

1. Enregistrement :
   L’utilisateur clique sur "Enregistrer" et parle dans le micro. L’application capte sa voix pendant quelques secondes.

2. Transcription :
   Le signal audio est envoyé à un modèle de reconnaissance vocale (Whisper), qui le transforme en texte.

3. Synthèse vocale :
   Le texte est ensuite transmis à un modèle de clonage vocal (XTTS), qui génère une réponse audio en imitant une vraie voix.

4. Restitution :
   L’utilisateur voit le texte affiché à l’écran et peut écouter le fichier audio généré directement depuis l’interface et le télécharger.

Les figures ci-dessous illustrent un exemple de manipulation de l’interface de conversion de la parole :

Figure 8 : Interface de la conversion de la parole

Figure 9 : Début d'enregistrement

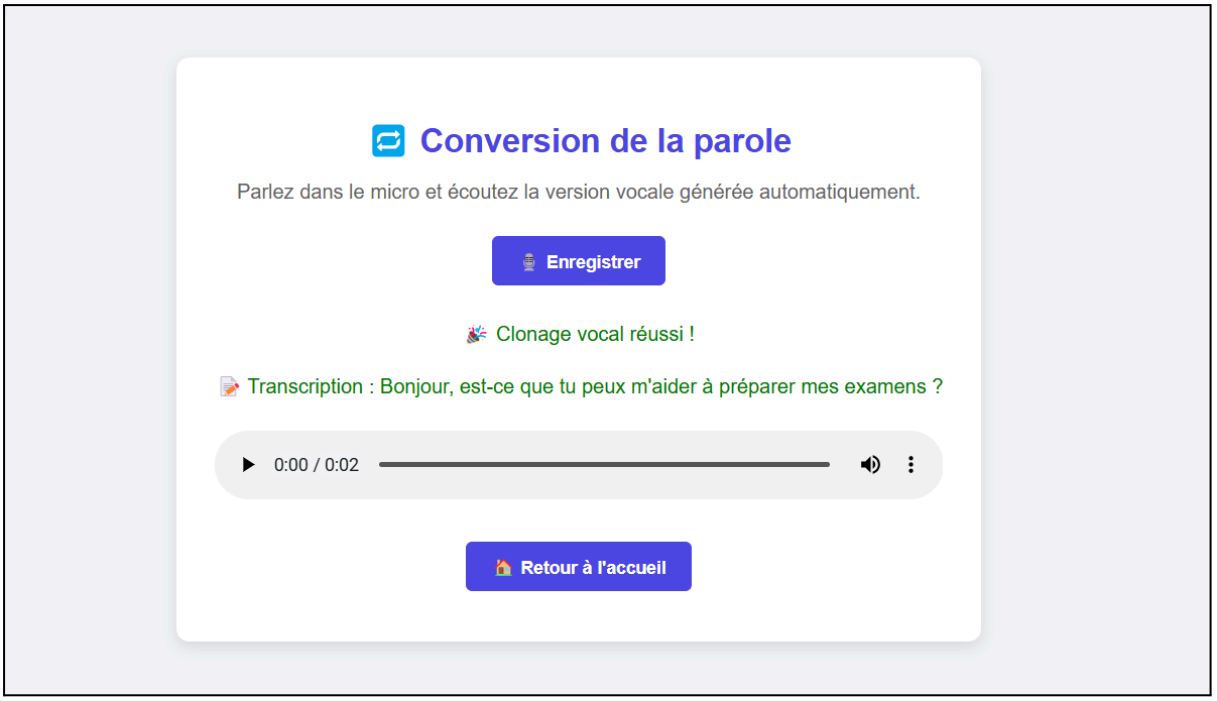

Figure 10 : Fin d'enregistrement, transcription de l'audio capté par l'utilisateur et synthèse vocale

Figure 11 : Résultat final de la conversion de la parole

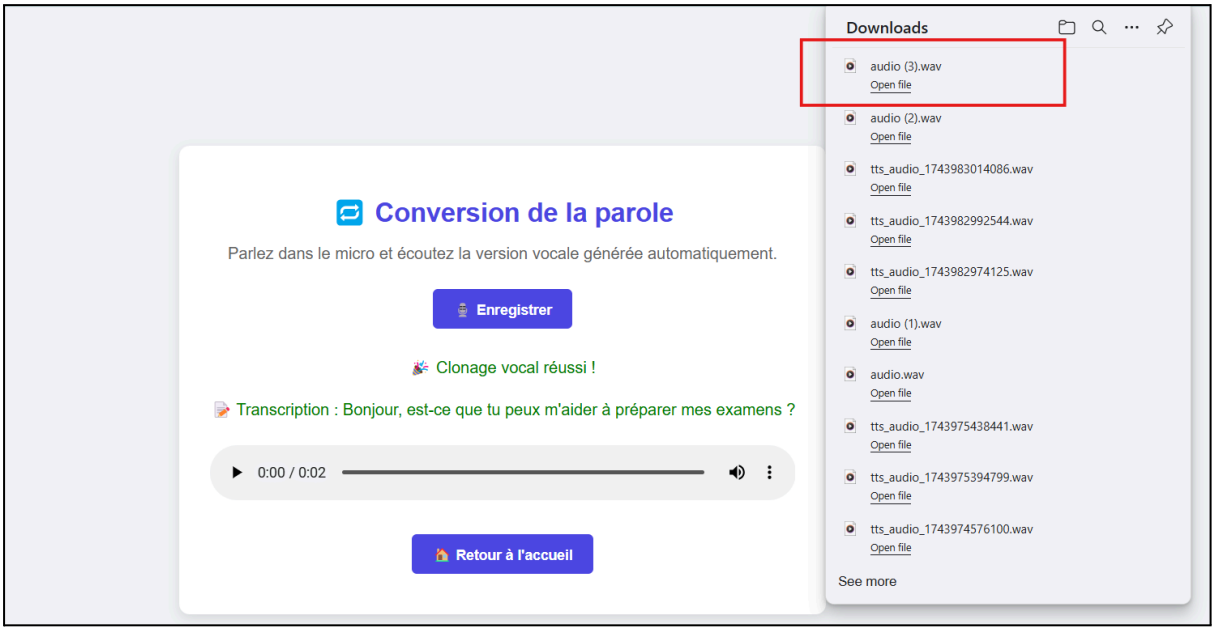

Figure 12 : Téléchargement du fichier audio généré

Synchronisation labiale :

Cette fonctionnalité permet à l'utilisateur de donner vie à une image en animant le visage et en synchronisant les mouvements des lèvres avec un dialogue ou une chanson provenant d'un fichier audio, créant ainsi une illusion réaliste de parole ou de chant.

La synchronisation labiale suit le principe du schéma ci-dessous :

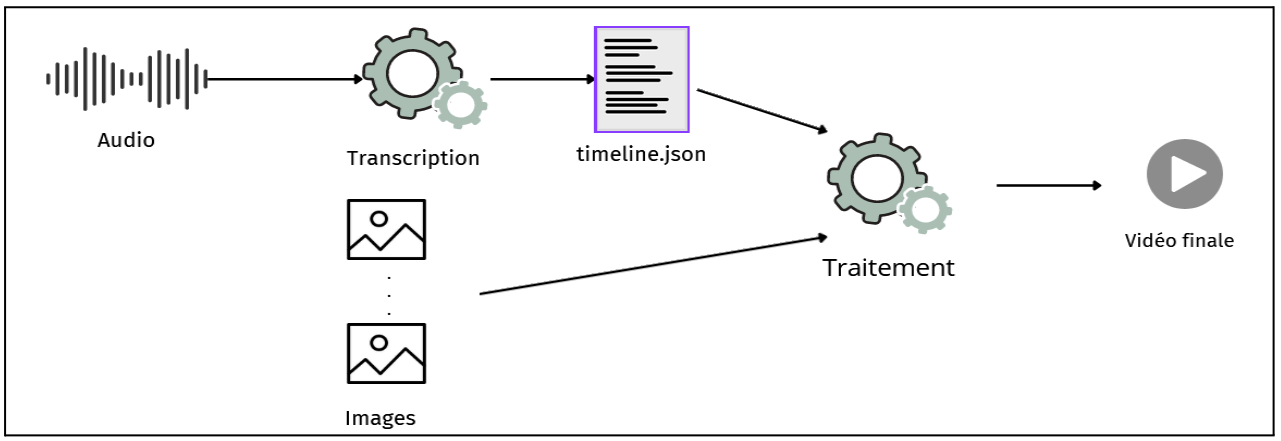

Figure 13 : Principe de synchronisation labiale

La synchronisation labiale se fait en deux phases :

1. Transcription de l'audio :
   Ceci se fait via le fichier timeliner.py avec le modèle **Whisper** qui va prendre en entrée l'audio (mp3, mp4, mpeg, mpga, m4a, wav et webm) et générer un fichier json : **"timeline.json"** contenant la timeline de l'audio.
   La figure suivante illustre un exemple du fichier timeline :

Figure 14 : Fichier **Timeline.json**

Comme affiché dans la figure ci-dessus, chaque caractère transcrit de l'audio a un début **"start"** et une fin **"end"**. Ceci va permettre la bonne synchronisation de l'audio et des lèvres.

2. Synchronisation des lèvres :
   Cette étape consiste à dessiner une bouche animée dans la zone de la bouche de la personne présente dans la photo. Nous avons généré un set d'images de bouche réalistes à l'aide de **ChatGpt Pro** qui vont correspondre aux différentes lettres et phonèmes.

   La figure ci-dessous montre des exemples d'images :

Figure 15 : Images de bouche

A partir du fichier **timeline.json** et des différentes images de lèvres, l'application va pouvoir faire correspondre les différentes lettres de l'audio avec les images correspondantes avec un durée d'apparition de la forme gérée via les temps disponibles dans le fichier timeline.json.

Au final, la vidéo obtenue est comme le montre la figure ci-dessous :

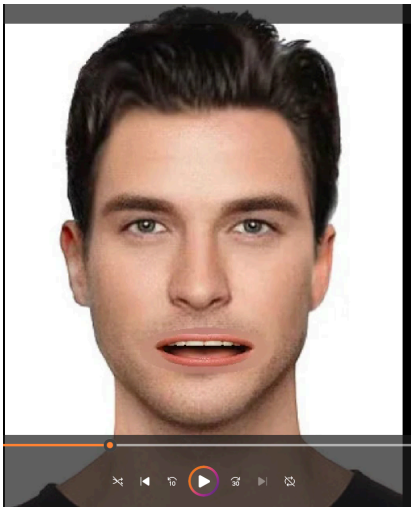

Figure 16 : Résultat du synchronisation labiale

Génération musicale par IA :

Cette fonctionnalité permet aux utilisateurs de créer des compositions musicales originales automatiquement à partir d'une simple description textuelle. Conçue spécialement pour un usage pédagogique et créatif, elle transforme en quelques secondes des idées en mélodies complètes de 30 secondes.

Les technologies clés sont :

Composant	Technologie	Rôle
Interface utilisateur	HTML5/CSS3/JavaScript	Interface conviviale pour saisir les requêtes et écouter les résultats
Serveur d'application	Node.js avec Express	Pont sécurisé entre l'interface et l'API d'IA
Moteur de génération	API Stable Audio 2 (Stability AI)	Transformation avancée du texte en fichier audio au format MP3

Tableau 4 : Technologies clés de la génération musicale

Le fonctionnement de la génération musicale est comme suit :

1. **Saisie intuitive** :
   L'utilisateur décrit le style ou l'ambiance musicale souhaitée via une interface web simple. Les figures ci-dessous montrent cette étape :

Figure 17 : Interface de génération musicale

Figure 18 : Saisie d'une description

2. **Traitement par l'IA** :
   La description est analysée par l'IA Stable Audio 2, pour produire un fichier MP3 en quelques secondes.

Figure 19 : Envoi de la demande de génération musicale

3. **Génération immédiate** :
   Un fichier audio au format MP3 est généré. L'utilisateur peut écouter la musique directement dans son navigateur (avec options de téléchargement de l'audio et changement de vitesse).

Figure 20 : Résultat de la génération musicale

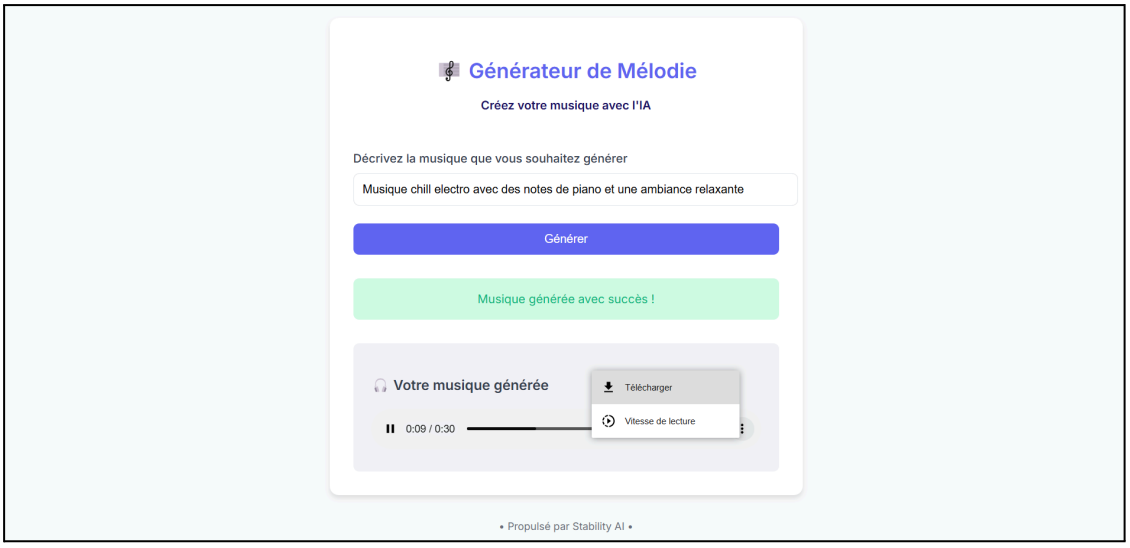

Figure 21 : Options pour l'audio généré

Conclusion

La réalisation du module de production sonore par intelligence artificielle a donné forme aux objectifs projetés au départ. Avec une interface intuitive et immersive, les élèves sont enfin capables d'explorer de façon autonome les principes de la synthèse vocale et de la composition musicale à l'aide de l'IA. Ce module est devenu ainsi une ressource pédagogique à la fois moderne et motivante, en rejoignant l'apprentissage par l'expérimentation.

Les résultats finaux offrent aussi des perspectives d'évolution, tant sur l'enrichissement fonctionnel, que sur la personnalisation des contenus, ou enfin l'intégration à d'autres environnements numériques éducatifs.