ROYAUME DU MAROC

HAUT COMMISSARIAT AU PLAN

INSTITUT NATIONAL DE STATISTIQUE ET D'ECONOMIE APPLIQUEE

INSEA

RAPPORT DE PROJET D'INTELLIGENCE ARTIFICIELLE

Synthèse et Conversion de Voix Personnalisée par Deep Learning

Intégration des Technologies TTS et RVC

Réalisé par :
LAOUAD Ayoub

Encadré par :
Dr. EL KARFI Ikram

Master Systèmes d'Information et Systèmes Intelligents

M2SI M1
29 mai 2025

Table des matières

Résumé

Ce rapport de projet d'AI détaille le développement et l'implémentation d'un système novateur de synthèse vocale personnalisée, exploitant les avancées récentes du deep learning. L'objectif central de ce travail est de permettre à un utilisateur de générer une voix synthétique mimant fidèlement sa propre voix à partir d'un ensemble restreint d'échantillons audio, et d'utiliser ensuite cette voix clonée pour prononcer n'importe quel texte.

Le système conçu intègre deux paradigmes complémentaires et performants : la conversion de voix basée sur la récupération (Retrieval-based Voice Conversion, RVC) pour la capacité de clonage vocal à partir de données limitées, et les modèles de synthèse vocale (Text-to-Speech, TTS) pour la génération de parole à partir d'un texte d'entrée. L'architecture globale mise en œuvre garantit un flux de travail complet et cohérent, allant de la phase de préparation des données et d'entraînement d'un modèle vocal spécifiquement adapté à l'utilisateur, jusqu'à son exploitation au travers d'une interface intuitive pour la synthèse de parole.

Les applications potentielles de cette technologie sont multiples et transformatives. Elles s'étendent des présentations automatisées à la lecture de documents personnalisée, des assistants vocaux sur mesure à l'amélioration de l'accessibilité pour les individus ayant des difficultés de lecture ou ayant perdu leur voix. Les résultats obtenus, évalués tant objectivement que subjectivement, attestent de la haute fidélité et de la naturalité des voix générées, confirmant la viabilité et le potentiel de l'approche proposée. Ce projet ouvre des perspectives prometteuses pour l'évolution des interfaces homme-machine et la personnalisation de l'interaction vocale.

Mots-clés : Synthèse vocale, Deep Learning, RVC, TTS, Clonage vocal, Conversion de voix, Intelligence artificielle, Traitement du signal vocal, Modèles génératifs, Prosodie.

Chapitre 1

Introduction Générale

1.1 Contexte Scientifique et Technologique

La synthèse vocale, ou Text-to-Speech (TTS), est une discipline fondamentale de l'intelligence artificielle qui vise à convertir un texte écrit en une séquence sonore intelligible. Historiquement, ce domaine a connu des avancées significatives, passant des systèmes articulatoires et formants rudimentaires aux approches paramétriques et de concaténation. Ces dernières, bien qu'améliorant la qualité, produisaient souvent une parole robotique et manquaient de la fluidité et de la naturalité de la voix humaine. La personnalisation était alors un défi colossal, nécessitant des corpus de données massifs et des techniques d'adaptation complexes [1].

L'émergence et la démocratisation des architectures de deep learning au cours de la dernière décennie ont marqué un tournant majeur. Des modèles pionniers comme WaveNet [2] et Tacotron [3], suivis par Tacotron 2 [4] et VITS [5], ont démontré la capacité des réseaux de neurones profonds à modéliser directement les formes d'onde audio ou les spectres, permettant de générer une parole d'une qualité quasi-humaine. Cette révolution a ouvert la voie à de nouvelles fonctionnalités, dont le clonage vocal, qui permet de synthétiser de la parole avec le timbre et la prosodie d'une voix spécifique.

Plus récemment, les modèles de conversion vocale basés sur la récupération (Retrieval-based Voice Conversion, RVC) ont émergé, offrant une solution particulièrement efficace pour le clonage vocal à partir d'un nombre très limité d'échantillons audio. Ces systèmes combinent des techniques d'extraction de caractéristiques vocales robustes avec des mécanismes de recherche dans des bases de données d'exemples, permettant une personnalisation de la voix auparavant inaccessible sans des heures d'enregistrement.

1.2 Motivation du Projet

La motivation principale derrière ce projet est double. D'une part, il s'agit d'explorer et de maîtriser des technologies de pointe en IA, en particulier dans le domaine du traitement du signal vocal et de l'apprentissage profond. Le clonage vocal représente un défi technique stimulant, combinant plusieurs sous-domaines de l'IA (reconnaissance de formes, génération, modélisation séquentielle).

D'autre part, la demande croissante pour des interactions homme-machine plus naturelles et personnalisées sous-tend un besoin croissant de systèmes vocaux adaptables. Les applications sont nombreuses :

— **Accessibilité :** Permettre à des personnes ayant perdu leur voix ou souffrant de troubles de la parole de retrouver une identité vocale numérisée.
— **Création de contenu :** Générer des narrations personnalisées pour des vidéos, des podcasts, des livres audio, ou des présentations.
— **Assistance vocale :** Développer des assistants personnels ou des chatbots avec une voix unique et mémorable.
— **Divertissement et Gaming :** Créer des voix de personnages uniques ou des doublages personnalisés.
— **Éducation :** Produire des supports pédagogiques avec la voix de l'enseignant pour une meilleure immersion des élèves.

Ce projet vise à rendre cette personnalisation vocale accessible au-delà des grands studios ou laboratoires de recherche, en se basant sur des outils open-source et des ressources computationnelles raisonnables.

1.3 Problématique Détaillée

La problématique centrale de ce projet d'AI peut être articulée autour de plusieurs défis interdépendants :

1. **Génération de Modèle Vocal Personnalisé à partir de Données Limitées :** La plupart des systèmes de clonage vocal de haute qualité nécessitent des heures d'enregistrements du locuteur cible. Comment peut-on obtenir un modèle fidèle et naturel avec seulement quelques minutes d'audio, comme le promettent les architectures RVC ? Cela implique la sélection des échantillons les plus pertinents, leur prétraitement optimal, et l'adaptation des hyperparamètres d'entraînement.
2. **Intégration du Modèle Personnalisé dans un Système TTS Opérationnel :** Une fois le modèle RVC entraîné, il doit pouvoir être utilisé de manière fluide et efficace par un système TTS pour convertir du texte arbitraire en parole. Quels sont les défis techniques liés à cette intégration (compatibilité des formats, latence, cohérence) ?
3. **Maintien de la Qualité Audio et de la Fidélité Vocale :** L'objectif n'est pas seulement de produire de la parole, mais de le faire avec une qualité audio élevée (sans artefacts, bruit) et une fidélité maximale à la voix originale de l'utilisateur (timbre, accent, prosodie). Comment évaluer et optimiser ces deux aspects ?
4. **Développement d'une Interface Utilisateur Intuitive :** Pour que la technologie soit véritablement accessible, l'ensemble du processus (de la préparation des données à la synthèse) doit être utilisable par des personnes n'ayant pas une expertise technique approfondie en deep learning. Cela soulève des questions de conception d'interface et d'expérience utilisateur.

1.4 Objectifs du Projet

Pour répondre à cette problématique, les objectifs de ce projet sont définis comme suit :

- **Objectif Principal :** Développer un système complet de synthèse vocale personnalisée en combinant les technologies RVC et TTS, permettant la génération automatique de parole avec une voix clonée de l'utilisateur.
- **Objectifs Spécifiques :**
  - **Approfondir les Connaissances Technologiques :** Étudier et maîtriser les fondements théoriques et les architectures des modèles RVC et TTS modernes (VITS, HuBERT, RMVPE, HiFi-GAN).
  - **Mettre en place un Pipeline d'Entraînement Robuste :** Concevoir et implémenter un processus efficace pour la collecte, le prétraitement et l'entraînement de modèles vocaux personnalisés avec des datasets limités.
  - **Implémenter un Système de Synthèse Fonctionnel :** Configurer et utiliser une plateforme capable d'intégrer le modèle RVC entraîné pour la conversion de texte en parole.
  - **Évaluer la Qualité et la Fidélité des Synthèses :** Procéder à une évaluation rigoureuse, à la fois objective (métriques acoustiques) et subjective (tests d'écoute), des voix synthétiques générées.
  - **Identifier les Limitations et Proposer des Axes d'Amélioration :** Analyser les performances du système, relever les défis rencontrés et formuler des perspectives pour les développements futurs et l'impact sociétal de cette technologie.

1.5 Structure du Rapport

Ce rapport est organisé en cinq chapitres principaux afin de couvrir l'ensemble des aspects du projet. Le chapitre 1 introduit le contexte général, la problématique et les objectifs. Le chapitre 2 explore l'état de l'art des technologies de synthèse et de conversion vocale, en détaillant les fondements théoriques des modèles utilisés. Le chapitre 3 est dédié à la conception et à l'architecture du système développé, explicitant le flux de données et les choix technologiques. Le chapitre 4 décrit la mise en œuvre pratique, les outils, les processus d'entraînement et de synthèse, et présente une analyse détaillée des résultats obtenus. Enfin, le chapitre 5 conclut le rapport en synthétisant les réalisations, en discutant des limitations et en esquissant les perspectives d'évolution et l'impact sociétal de cette technologie.

Chapitre 2

État de l'art et Fondements Théoriques

2.1 Évolution Historique de la Synthèse Vocale

La synthèse vocale est un domaine qui a constamment évolué pour se rapprocher de la parole humaine. Ses racines remontent à la fin du XVIIIe siècle avec les premières tentatives mécaniques de reproduction de la parole.

2.1.1 Les Approches Traditionnelles

Avant l'ère du deep learning, deux paradigmes principaux dominaient la synthèse vocale :

- **Synthèse par Formants (Paramétrique) :** Cette approche repose sur la modélisation des caractéristiques physiques de la parole, telles que les fréquences des formants (pics d'énergie dans le spectre vocal), le pitch, et l'intensité. Des règles linguistiques et acoustiques complexes étaient utilisées pour générer ces paramètres à partir du texte, puis un vocodeur (comme le vocodeur à filtre) les convertissait en signal audio. Les systèmes basés sur les modèles de Markov cachés (HMM-based TTS) représentaient une avancée majeure dans cette catégorie, permettant une modélisation statistique des séquences de parole. Cependant, la qualité restait souvent métallique et peu naturelle, avec un manque de flexibilité expressive.
- **Synthèse par Concaténation (Unit Selection) :** Cette méthode consiste à assembler de petits segments de parole pré-enregistrés (unités phonétiques, diphones, syllabes) pour former des phrases complètes. Un vaste corpus de parole était nécessaire, et l'algorithme de sélection cherchait les unités qui minimisaient la discontinuité aux points de jonction. Cette approche pouvait produire une parole de très haute qualité si les unités étaient bien sélectionnées et si les transitions étaient fluides. Cependant, elle était gourmande en stockage, difficile à personnaliser (chaque nouvelle voix nécessitait un nouveau corpus énorme) et inflexible pour le contrôle de la prosodie et de l'expressivité. Les erreurs de concaténation entraînaient des bruits et des hachures perceptibles.

Ces méthodes présentaient des limitations importantes en termes de naturel, de personnalisation, et de robustesse face à des variations prosodiques ou des styles de parole différents.

2.1.2 La Révolution du Deep Learning

L'avènement des réseaux de neurones profonds a transformé radicalement le domaine de la synthèse vocale, permettant de surpasser les limitations des approches traditionnelles.

- **WaveNet (DeepMind, 2016) [2] :** Ce modèle révolutionnaire a été le premier à démontrer la capacité à générer des formes d'onde audio brutes avec une qualité quasi-humaine. Basé sur des convolutions dilatées, WaveNet est un modèle auto-régressif qui prédit le prochain échantillon audio en se basant sur les précédents. Bien que d'une qualité exceptionnelle, son inférence était très lente, le rendant peu pratique pour des applications en temps réel.
- **Tacotron (Google, 2017) [3] et Tacotron 2 (Google, 2018) [4] :** Ces modèles marquent l'ère des architectures "end-to-end". Ils apprennent à transformer directement le texte d'entrée en spectrogrammes (représentations visuelles du son) qui sont ensuite convertis en audio par un vocodeur séparé (initialement WaveNet, puis des vocodeurs plus rapides). Tacotron 2, en particulier, a atteint un niveau de naturalité très élevé grâce à son architecture encoder-decoder avec attention, capable d'aligner automatiquement le texte et le son.
- **Modèles Basés sur les Transformeurs (e.g., Transformer-TTS, FastSpeech) [6] :** Inspirés par le succès des transformeurs en traitement du langage naturel, ces modèles ont introduit des mécanismes d'attention multi-têtes et ont souvent éliminé la nature auto-régressive de la génération de spectrogrammes, ce qui a permis des synthèses beaucoup plus rapides, tout en maintenant une haute qualité. FastSpeech, par exemple, utilise un prédicteur de durée pour générer tous les frames en parallèle.
- **VITS (Variational Inference with Adversarial Learning for End-to-End Text-to-Speech) [5] :** VITS représente une architecture de pointe qui combine l'inférence variationnelle, l'apprentissage adversarial (GAN) et un vocodeur basé sur HiFi-GAN. Cette synergie permet à VITS de produire de la parole de haute qualité, rapide et expressive, tout en capturant la variabilité stochastique naturelle de la parole humaine. C'est l'architecture sous-jacente au système RVC utilisé dans ce projet.

Ces avancées ont pavé la voie aux technologies de clonage vocal modernes en permettant une modélisation plus profonde et plus nuancée de la parole.

2.2 Technologies de Conversion Vocale (Voice Conversion)

La conversion vocale (VC) est une tâche qui vise à transformer les caractéristiques d'une voix source (locuteur source) afin qu'elle ressemble à celles d'une voix cible (locuteur cible), tout en préservant le contenu linguistique de l'énoncé source [7].

2.2.1 Principes Généraux de la Conversion Vocale

Historiquement, la VC impliquait la manipulation des paramètres acoustiques tels que la fréquence fondamentale (pitch), l'enveloppe spectrale (timbre) et les caractéristiques de la source d'excitation. Les défis majeurs incluent la décomposition propre de la parole en contenu (phonèmes) et style (locuteur), et la re-synthèse sans artefacts.

2.2.2 Approches Basées sur le Deep Learning pour la VC

Les réseaux de neurones profonds ont révolutionné la VC en permettant des transformations non linéaires plus complexes et en apprenant des représentations latentes plus significatives.

— **Auto-encodeurs Variationnels (VAE) et Réseaux Antagonistes Génératifs (GAN) :** Ces architectures génératives ont été largement utilisées pour apprendre des mappings entre les espaces de caractéristiques de différentes voix. Elles permettent de générer de nouvelles données vocales avec les caractéristiques souhaitées.

— **Transformation de Caractéristiques Indépendantes du Locuteur :** De nombreux modèles de VC modernes se basent sur l'extraction de représentations de parole (embedding) qui sont (idéalement) indépendantes de l'identité du locuteur, mais qui encodent le contenu linguistique. Ces représentations peuvent ensuite être passées à un module qui "colore" le contenu avec le timbre et la prosodie du locuteur cible.

2.2.3 Architecture RVC (Retrieval-based Voice Conversion)

RVC représente une approche particulièrement innovante et performante dans le domaine de la conversion vocale. Son succès réside dans la combinaison intelligente de plusieurs composants clés et techniques d'apprentissage profond [8].

Extracteur de Caractéristiques de Contenu : HuBERT

Au cœur de RVC se trouve l'utilisation d'un extracteur de caractéristiques de contenu robuste et indépendant du locuteur. Pour ce projet, **HuBERT (Hidden-Unit BERT)** [9] est utilisé. HuBERT est un modèle de deep learning pré-entraîné de manière auto-supervisée sur de vastes corpus de parole non annotés. Son principe est inspiré des modèles BERT en PNL : il masque certaines portions du signal audio et est entraîné à prédire les unités cachées correspondantes. Cela lui permet d'apprendre des représentations contextuelles et sémantiques de la parole qui capturent le contenu linguistique (phonèmes, rythmes) tout en étant largement invariantes à l'identité du locuteur et à d'autres facteurs de style (comme le bruit). Ces caractéristiques de haut niveau sont essentielles car elles fournissent au modèle RVC l'information "quoi dire" sans influencer le "comment le dire".

Extraction de la Hauteur Vocale (Pitch) : RMVPE

Le pitch, ou fréquence fondamentale (F0), est un attribut crucial de la prosodie et de l'identité vocale. Une extraction précise du pitch est indispensable pour que la voix synthétisée ait une intonation naturelle et un timbre réaliste. Dans le cadre de RVC, plusieurs algorithmes peuvent être utilisés, et **RMVPE (Robust Model for Vocal Pitch Estimation)** [10] est l'un des plus performants et est privilégié pour sa robustesse. RMVPE se distingue par sa capacité à estimer la hauteur vocale même en présence de musique, de bruits de fond, ou dans des environnements polyphoniques. Il utilise une architecture neuronale sophistiquée, combinant des réseaux de neurones (typiquement basés sur des U-Net et des GRU) pour analyser le log mel-spectrogramme du signal audio et prédire la trajectoire de la fréquence fondamentale. Contrairement

à des méthodes traditionnelles (comme YIN ou Harvest) qui peuvent être sensibles au bruit, RMVPE apprend des représentations plus abstraites et contextuelles, ce qui conduit à une estimation plus stable et précise du pitch, contribuant significativement à la naturalité et à la fidélité de la voix clonée. La Figure **??** (précédemment en annexe, maintenant intégrée) illustre l'architecture générale d'un tel modèle.

Le Modèle de Transformation : VITS

Le cœur de RVC est un modèle de transformation qui prend les caractéristiques de contenu (HuBERT) et les informations de pitch (RMVPE) et les convertit en un spectrogramme ou une représentation acoustique intermédiaire qui est ensuite synthétisée en audio. L'architecture VITS [5] est souvent utilisée comme base pour cette transformation en RVC. VITS, comme mentionné précédemment, est un modèle de synthèse vocale end-to-end qui combine :

- **Un auto-encodeur variationnel (VAE) :** Permet de capturer la variabilité stochastique naturelle de la parole et de modéliser l'espace latent des caractéristiques vocales.
- **Un mécanisme d'alignement :** Apprend l'alignement entre le texte (ou les caractéristiques de contenu) et les caractéristiques acoustiques sans supervision explicite.
- **Un réseau génératif adversarial (GAN) :** Le générateur produit les formes d'onde audio, tandis qu'un discriminateur est entraîné à distinguer la parole réelle de la parole synthétisée, poussant le générateur à produire un audio plus réaliste et sans artefacts.
- **Un vocodeur basé sur HiFi-GAN :** Intégré pour une synthèse rapide et de haute fidélité.

Dans RVC, la capacité de VITS à générer des variations naturelles de la parole est essentielle pour permettre la personnalisation.

Le Mécanisme de Récupération (Retrieval)

C'est la caractéristique la plus distinctive de RVC. Au lieu de se fier uniquement à l'apprentissage d'un mapping générique, RVC utilise un mécanisme de recherche (retrieval) pour trouver des segments de parole "similaires" dans une base de données d'entraînement (contenant des échantillons de la voix cible). Lors de l'entraînement, un index des caractéristiques de la voix cible est construit. Lors de l'inférence, pour chaque partie de la parole source, le modèle recherche dans cet index les caractéristiques les plus proches des données d'entraînement de la voix cible. Ces "caractéristiques récupérées" sont ensuite utilisées pour conditionner le modèle de transformation (VITS). Cette approche permet au modèle de "s'appuyer" sur des exemples réels de la voix cible, ce qui améliore considérablement la fidélité et la naturalité de la voix clonée, même avec des datasets d'entraînement limités. Cela aide à surmonter les limitations du "hors distribution" que les modèles purement génératifs peuvent rencontrer.

Le Vocodeur Neural : HiFi-GAN

Un vocodeur est l'étape finale qui transforme les représentations acoustiques (comme les spectrogrammes Mel) en un signal audio brut. HiFi-GAN [11] est un vocodeur neural basé sur les GAN qui est largement adopté pour sa capacité à générer de la parole

de très haute fidélité à des vitesses d'inférence très rapides, le rendant adapté aux applications en temps réel. Son architecture utilise des générateurs et des discriminateurs multiples pour assurer une qualité audio exceptionnelle et une absence d'artefacts. Sa vitesse est un avantage majeur par rapport à des vocodeurs auto-régressifs plus anciens comme WaveNet.

2.3 Datasets de Pré-entraînement : CSTR VCTK Corpus

Les performances des modèles de deep learning dépendent crucialement de la qualité et de la quantité des données d'entraînement. Pour les modèles de synthèse et de conversion vocale, des corpus multilocuteurs vastes et diversifiés sont essentiels pour apprendre des représentations robustes de la parole.

Le **CSTR VCTK Corpus (Voice Cloning Toolkit)** [12] est une ressource de référence fréquemment utilisée pour le pré-entraînement de modèles TTS et VC. Ce corpus comprend environ **44 heures de données vocales** enregistrées par **110 locuteurs anglais natifs** de différentes régions (par exemple, des accents variés du Royaume-Uni, des États-Unis, du Canada, d'Australie, d'Inde, etc.). Chaque locuteur a lu environ 400 phrases distinctes, soigneusement sélectionnées à partir de diverses sources (articles de journaux, le texte standard "Rainbow Passage", et des phrases d'élicitation accentuelles). La richesse et la diversité de ce corpus en font une base de données idéale pour l'entraînement de modèles capables de capturer un large éventail de caractéristiques prosodiques, phonétiques et acoustiques de la parole anglaise, et de développer des "modèles de base" qui peuvent ensuite être adaptés (fine-tuned) à des voix spécifiques avec un minimum de données supplémentaires. Les modèles HuBERT et VITS, sur lesquels s'appuie RVC, sont généralement pré-entraînés sur des corpus encore plus massifs (par exemple, LibriSpeech, VoxPopuli) avant d'être fine-tunés sur des corpus plus spécifiques comme VCTK ou d'autres, permettant ainsi d'apprendre des représentations très généralisables de la parole.

Chapitre 3

Conception et Architecture du Système

3.1 Approche Modulaire et Vue d'Ensemble

Le système de synthèse vocale personnalisée est conçu avec une architecture modulaire, divisée en deux phases distinctes mais interconnectées. Cette approche favorise la clarté, la maintenance et l'évolutivité du système, permettant à chaque module de se concentrer sur une tâche spécifique.

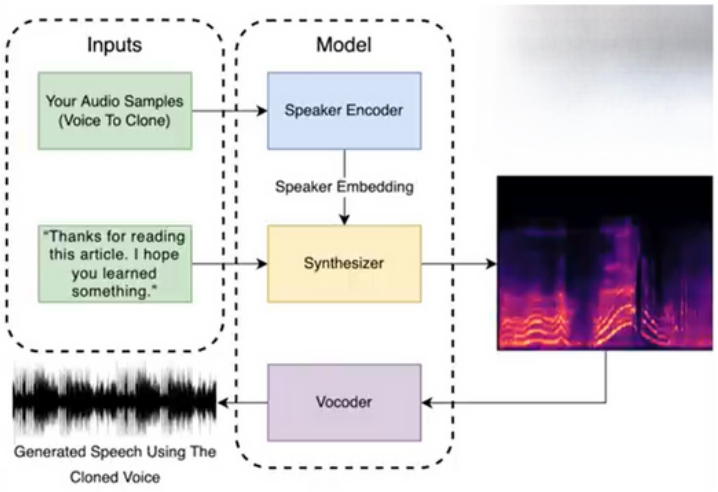

FIGURE 3.1 – Architecture fonctionnelle d'un système TTS personnalisé intégrant RVC. La parole générée combine le contenu linguistique du texte d'entrée avec les caractéristiques vocales (timbre, prosodie) extraites des échantillons de la voix cible.

1. **Module d'Entraînement du Modèle Vocal (Phase 1) :** Ce module est responsable de la création d'un modèle vocal personnalisé à partir des échantillons audio fournis par l'utilisateur. Il englobe la préparation des données, l'extraction des caractéristiques acoustiques (pitch et contenu), et l'entraînement du modèle RVC. Le résultat de cette phase est un fichier de poids du modèle (généralement '.pth') et un fichier d'index ('.index') qui encapsulent les caractéristiques

uniques de la voix de l'utilisateur. Ce module est principalement soutenu par l'outil Mangio-RVC.

2. **Module de Synthèse Vocale (Phase 2) :** Ce module utilise le modèle personnalisé généré par la première phase pour convertir n'importe quel texte en parole avec la voix clonée. Il intègre un pipeline TTS complet qui prend le texte en entrée, le traite, et génère le fichier audio de sortie. L'outil Applio est au cœur de cette phase.

3.2 Flux de Données Détaillé

Le flux de données à travers le système est un pipeline séquentiel, conçu pour maximiser l'efficacité et la qualité de la transformation vocale.

3.2.1 Phase de Préparation et d'Entraînement

1. **Collecte des Échantillons Audio (Input) :** L'utilisateur fournit un ensemble d'enregistrements de sa voix. La qualité et la diversité de ces enregistrements sont primordiales pour la fidélité du modèle final. Chaque échantillon doit être propre et de bonne qualité.
2. **Prétraitement des Données :**
   — **Normalisation du Volume :** Ajustement de l'amplitude pour assurer une cohérence acoustique entre tous les échantillons, évitant les variations de volume inopinées.
   — **Ré-échantillonnage :** Conversion de tous les fichiers audio à une fréquence d'échantillonnage standard (généralement 48 kHz pour RVC), afin d'assurer la compatibilité et d'éviter les pertes d'information.
   — **Découpage des Silences et Segmentation :** Suppression des silences inutiles au début et à la fin des enregistrements pour optimiser l'efficacité de l'entraînement et focaliser le modèle sur la parole. Les longs enregistrements peuvent être segmentés en clips plus courts et gérables.
   — **Filtrage :** Des techniques de filtrage peuvent être appliquées pour réduire le bruit de fond ou les artefacts.
3. **Extraction des Caractéristiques Acoustiques :**
   — **Caractéristiques de Contenu (HuBERT Embeddings) :** Le modèle HuBERT est utilisé pour extraire des représentations vectorielles de haut niveau du contenu linguistique des échantillons. Ces embeddings sont cruciaux car ils capturent l'information sémantique de la parole indépendamment de l'identité du locuteur.
   — **Hauteur Vocale (Pitch) (RMVPE) :** L'algorithme RMVPE est appliqué pour extraire la trajectoire de la fréquence fondamentale de chaque échantillon. Cette information est essentielle pour reproduire l'intonation et la mélodie de la voix cible.
   — **Volume (Loudness) :** L'extraction du volume peut également être effectuée pour s'assurer que le modèle peut reproduire les dynamiques de l'amplitude de la voix.
4. **Construction de l'Index de Récupération (Retrieval Index) :** Un index vectoriel (par exemple, un index Faiss) est construit à partir des caractéristiques

de HuBERT extraites des données d'entraînement. Cet index sera utilisé pendant l'inférence pour trouver les exemples les plus similaires dans l'espace des caractéristiques.

5. **Entraînement du Modèle RVC (VITS-based) :** Le modèle VITS sous-jacent à RVC est entraîné. Il apprend à mapper les caractéristiques de contenu (HuBERT), le pitch et le volume vers les caractéristiques acoustiques spécifiques (timbre, prosodie) de la voix cible. L'entraînement implique la minimisation d'une fonction de perte complexe (combinant des pertes de reconstruction, des pertes adversariales, etc.) via un processus d'optimisation itératif.
6. **Sortie du Module d'Entraînement :** Les artefacts de cette phase sont le fichier de poids du modèle ('.pth') et le fichier d'index ('.index').

3.2.2 Phase de Synthèse (Inférence)

1. **Chargement du Modèle Personnalisé :** Le fichier '.pth' et le fichier '.index' du modèle RVC personnalisé sont chargés en mémoire.
2. **Saisie de Texte (Input) :** L'utilisateur entre le texte qu'il souhaite synthétiser.
3. **Analyse Linguistique et Text-to-Spectrogram Generation :**
   — Le texte est d'abord traité par un module de normalisation de texte et de grapheme-to-phoneme (G2P) pour le convertir en une séquence de phonèmes.
   — Un modèle TTS de base (souvent le modèle VITS pré-entraîné de RVC) génère une première version des caractéristiques acoustiques (e.g., mel-spectrogrammes) de ce texte. Ce modèle a déjà appris la correspondance texte-son et la prosodie générale.
4. **Conversion Vocale (RVC Application) :** C'est ici que le modèle RVC personnalisé entre en jeu. Les caractéristiques de contenu (HuBERT) sont extraites du spectrogramme généré, et le mécanisme de récupération est activé pour trouver les exemples les plus pertinents dans l'index. Les caractéristiques de la voix cible sont ensuite appliquées au spectrogramme généré, transformant le timbre et la prosodie pour correspondre à la voix clonée.
5. **Synthèse de Forme d'Onde (Vocoder) :** Le spectrogramme transformé est ensuite passé au vocodeur HiFi-GAN intégré, qui convertit ces représentations acoustiques en un signal audio temporel de haute qualité.
6. **Sortie du Module de Synthèse (Output) :** Un fichier audio ('.wav' ou autre format) est généré, contenant le texte prononcé avec la voix personnalisée.

3.3 Outils Logiciels et Technologies

Le succès de ce projet repose sur l'exploitation d'outils logiciels open-source robustes et d'une communauté de développement active.

3.3.1 Mangio-RVC-v23.7.0 INFER+TRAIN

Mangio-RVC est une implémentation populaire et conviviale du framework RVC. Il fournit une interface graphique complète (GUI) qui simplifie les étapes complexes d'entraînement et d'inférence.

— **Fonctionnalités d'Entraînement :** Mangio-RVC intègre tous les outils nécessaires pour le prétraitement des données (découpage, normalisation), l'extraction des caractéristiques (HuBERT, RMVPE), la construction de l'index Faiss, et l'entraînement du modèle RVC. Il offre des contrôles précis sur les hyperparamètres d'entraînement (nombre d'époques, taille de lot, taux d'apprentissage) et des fonctionnalités de monitoring pour suivre la progression.
— **Versions du Modèle RVC :** Mangio-RVC supporte différentes versions du modèle RVC (v1 et v2), avec la version v2 offrant généralement une meilleure qualité et stabilité d'entraînement grâce à des améliorations architecturales et de la fonction de perte.

3.3.2 Applio V3.2.8 (bugfix)

Applio est une application basée sur RVC, conçue pour l'inférence (synthèse vocale) et qui intègre également des fonctionnalités avancées pour la manipulation vocale.

— **Chargement de Modèles Personnalisés :** Applio permet de charger facilement les modèles '.pth' et les fichiers '.index' générés par Mangio-RVC.
— **Interface de Synthèse Intuitive :** Son interface web conviviale offre un champ de saisie de texte, des contrôles pour les paramètres de génération (Index Rate, Filter Radius, etc.) et des options d'exportation audio.
— **Fonctionnalités Complémentaires :** Au-delà de la synthèse vocale, Applio peut également être utilisé pour la conversion de voix entre différents locuteurs, le chant de voix (voice singing) et d'autres manipulations audio.

3.4 Considérations Éthiques dans la Conception

Dès la phase de conception, des considérations éthiques ont guidé les choix. L'objectif de ce projet est de permettre à l'utilisateur de générer *sa propre* voix personnalisée, ce qui est une application éthiquement saine et bénéfique (par exemple, pour l'accessibilité). L'utilisation d'outils open-source encourage la transparence et la compréhension du fonctionnement sous-jacent.

Cependant, la technologie du clonage vocal soulève inévitablement des questions plus larges sur l'authenticité et la désinformation (deepfakes). En choisissant de se concentrer sur l'auto-clonage et en sensibilisant à ces enjeux (voir section 5.3), le projet contribue à une utilisation responsable de cette technologie puissante.

Chapitre 4

Mise en œuvre et Résultats

4.1 Environnement de Développement

4.1.1 Configuration Matérielle

Le développement et les tests du système ont été effectués sur une configuration matérielle standard pour les tâches d'apprentissage profond, permettant d'équilibrer les performances et l'accessibilité.

- **Processeur :** Intel Core i5 de 13ème génération, 6 cœurs/12 threads, offrant une bonne puissance de calcul pour les opérations de prétraitement et la gestion générale du système.
- **Mémoire RAM :** 16 Go de RAM DDR4, suffisante pour charger les datasets, les modèles de grande taille, et exécuter les opérations en mémoire. Pour des datasets plus importants ou des entraînements plus complexes, 32 Go seraient préférables.
- **Carte Graphique (GPU) :** NVIDIA GeForce RTX 4050 Laptop avec 6 Go de VRAM dédiée. C'est le composant le plus critique pour les phases d'entraînement et d'inférence des modèles de deep learning. Les 6 Go de VRAM ont permis d'entraîner des modèles RVC avec des tailles de lot (batch sizes) raisonnables. Des GPUs avec plus de VRAM (ex : 8 Go ou 12 Go) permettraient des batch sizes plus grandes, accélérant l'entraînement et améliorant potentiellement la stabilité.
- **Stockage :** SSD NVMe de 512 Go. Le SSD est crucial pour la rapidité d'accès aux données d'entraînement et aux modèles, réduisant les temps de chargement et d'enregistrement des checkpoints.
- **Système d'exploitation :** Windows 11 Professionnel, fournissant un environnement de développement stable et compatible avec les pilotes NVIDIA CUDA.

4.1.2 Environnement Logiciel

Le projet est bâti sur un environnement logiciel basé sur Python, avec une gestion rigoureuse des dépendances pour assurer la reproductibilité.

- **Langage de Programmation :** Python 3.10.
- **Gestionnaire d'Environnement :** Conda (Anaconda/Miniconda) a été utilisé pour créer et gérer un environnement virtuel isolé, garantissant que les dépendances spécifiques du projet ne rentrent pas en conflit avec d'autres installations

Python.
- **Framework de Deep Learning :** PyTorch (version compatible CUDA 11.8) est le socle principal pour l'implémentation et l'exécution des modèles neuronaux. Sa flexibilité et sa performance sont essentielles pour les tâches de RVC et TTS.
- **Bibliothèques Principales :**
  - **Librosa :** Essentielle pour l'analyse et le traitement des signaux audio (extraction de caractéristiques spectrales, ré-échantillonnage, normalisation).
  - **NumPy et SciPy :** Pour les opérations numériques et scientifiques efficaces sur les tableaux de données.
  - **Torchaudio :** L'extension audio de PyTorch, permettant une intégration fluide des données audio dans les pipelines PyTorch.
  - **Fairseq :** Utilisée par HuBERT pour ses fonctionnalités de modèles pré-entraînés.
  - **Praat-Parselmouth et PyWorld :** Pour des analyses acoustiques avancées et l'extraction de paramètres comme le pitch.
- **Interfaces Graphiques Utilisées :**
  - **Mangio-RVC v23.7.0 INFER+TRAIN :** L'outil principal pour la préparation des données et l'entraînement du modèle RVC.
  - **Applio v3.2.8 (bugfix) :** L'outil principal pour l'inférence et la synthèse vocale.
- **Éditeur de Code :** Visual Studio Code, avec l'extension Python et LaTeX Workshop pour un environnement de développement intégré et efficace.

4.2 Processus d'Entraînement du Modèle Vocal

4.2.1 Préparation du Dataset Personnel

La qualité du dataset d'entraînement est le facteur le plus déterminant pour la fidélité et la naturalité de la voix clonée.
- **Collecte :** Un dataset personnel a été créé, composé de **30 échantillons audio** de ma propre voix. Chaque échantillon avait une durée de 2 à 7 secondes, pour un total d'environ **2 minutes et 15 secondes d'audio net**.
- **Qualité d'Enregistrement :** Les enregistrements ont été réalisés avec un microphone de qualité studio (Blue Yeti X) dans un environnement acoustiquement traité pour minimiser le bruit de fond et la réverbération. Cette étape est cruciale, car tout bruit ou artefact présent dans les données d'entraînement sera appris et potentiellement reproduit par le modèle.
- **Diversité du Contenu :** Les phrases enregistrées ont été choisies pour leur diversité phonétique et prosodique, couvrant une large gamme de phonèmes de la langue française, différentes intonations (déclaratives, interrogatives, exclamatives) et des rythmes de parole variés. Cela aide le modèle à généraliser sur des textes qu'il n'a jamais entendus.

— **Structure des Données :** Les fichiers audio (.wav) ont été stockés dans un dossier 'audio/' et leurs transcriptions textuelles correspondantes dans un dossier 'text/', avec des noms de fichiers appariés ('moi$_0$01.*wav*'*et*'*moi*$_0$01.*txt*').$Ce formate strequispar Mangio-RVC$.

4.2.2 Configuration de l'Entraînement avec Mangio-RVC

L'outil Mangio-RVC simplifie grandement le processus d'entraînement grâce à son interface graphique.

1. **Lancement de Mangio-RVC et Configuration du Projet :** Après le démarrage, le nom du modèle ('Ayoub$_L$*aouad*')$atspcifi, etlaversiondumodleRVC(v2)atslectionn$

2. 3. Le chemin vers le dossier 'my$_a$*udio*'*contenantlessous*−*dossiers*'*audio*'*et*'*text*'*atrenseign.Lacommand*
RVCaautomatiquementr−*chantillonnl'audio*48kHz, *normalislevolume*, *etsegmentlesfichierssinc*

4. **Extraction des Caractéristiques (Onglet "Train", Section 2) :**
   - **Extraction du Pitch :** La méthode **RMVPE** a été choisie pour l'estimation du pitch, étant reconnue pour sa robustesse.
   - **Extraction des Caractéristiques de Contenu :** Les embeddings HuBERT ont été extraits. Mangio-RVC gère la pipeline d'extraction des caractéristiques contextuelles nécessaires à l'entraînement.
   - **Génération de l'Index Faiss :** Un index de recherche (Faiss) a été construit à partir des caractéristiques HuBERT extraites. Cet index est crucial pour le mécanisme de récupération de RVC, permettant de trouver rapidement les exemples les plus similaires lors de l'inférence.

Entraînement du Modèle RVC (Onglet "Train", Section 3) :

- **Époques d'Entraînement :** Le modèle a été entraîné pendant **150 époques**. Pour un dataset aussi petit (2 min), 150 époques sont généralement suffisantes pour capturer les caractéristiques essentielles sans sur-apprentissage excessif. Pour des datasets plus grands, plusieurs centaines voire milliers d'époques seraient nécessaires.
- **Taille de Lot (Batch Size) :** Une taille de lot de 8 a été utilisée, adaptée aux 6 Go de VRAM du GPU. Des tailles de lot plus grandes peuvent accélérer l'entraînement mais nécessitent plus de VRAM.
- **Fréquence de Sauvegarde :** Les checkpoints du modèle ont été sauvegardés toutes les 5 époques. Cela a permis de suivre la progression de l'entraînement et de sélectionner le meilleur modèle en cas de sur-apprentissage (en général, le modèle qui présente la meilleure performance sur les données de validation, ou celui juste avant que les performances ne commencent à se dégrader).
- **Optimiseur et Taux d'Apprentissage :** L'optimiseur AdamW a été utilisé avec un taux d'apprentissage initial de 1e-4, qui a ensuite été ajusté par un scheduler de taux d'apprentissage.

L'entraînement complet a pris environ **30 à 45 minutes** sur la configuration matérielle spécifiée.

Le processus d'entraînement a généré deux fichiers de sortie principaux :

— 'Ayoub$_L$*aouad.pth*' : $Lefichierdepoidsdumodle RVCentran, contenanttouteslesinformationsappris$
$L'indexderechercheFaiss, essentielpourlemcanismedercuprationdeRVClorsdel'infrence.$

4.3 Intégration TTS et Synthèse Vocale

4.3.1 Configuration d'Applio pour l'Inférence

Une fois le modèle RVC personnalisé entraîné, il a été intégré dans l'outil Applio pour la synthèse vocale.

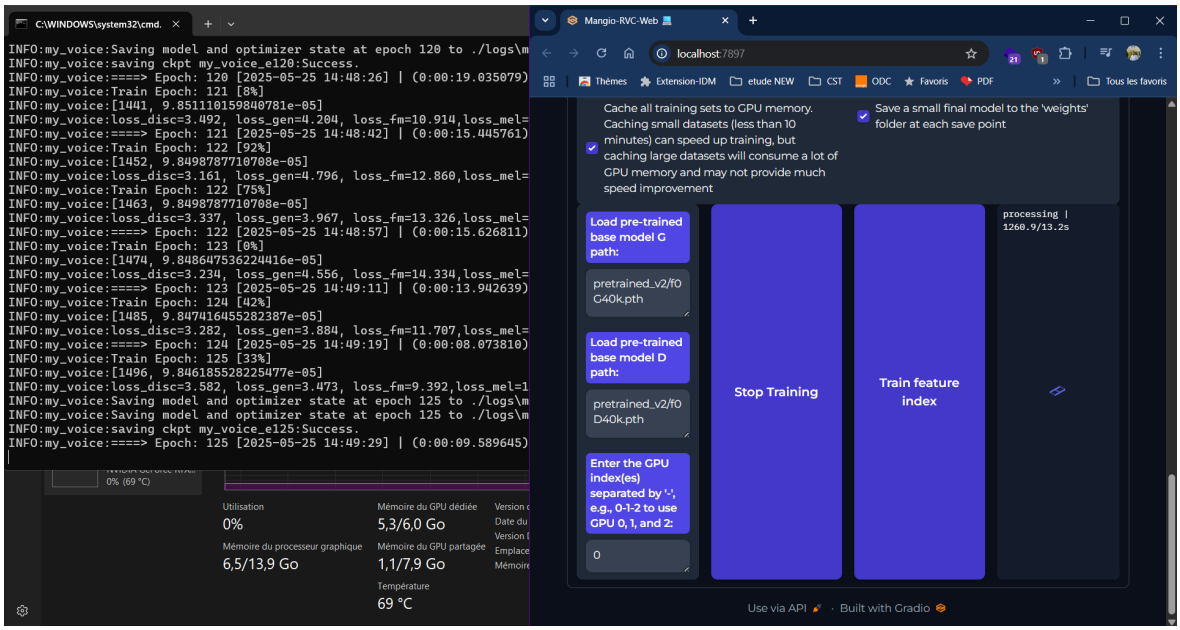

FIGURE 4.1 – Interface d'entraînement de Mangio-RVC v23.7.0. La section (1) permet la configuration initiale et la préparation du dataset. La section (2) gère l'extraction des caractéristiques vocales et la construction de l'index. Enfin, la section (3) offre les contrôles pour lancer et monitorer l'entraînement du modèle RVC.

1. **Lancement d'Applio et Accès à l'Onglet "Inference" :** Applio offre une interface web locale accessible via le navigateur.
2. **Chargement du Modèle Personnalisé :** Dans l'onglet "Inference", les fichiers ‘Ayoub$_{L}aouad.pth$‘et‘$added_{I}VF256_{F}lat_{A}youb_{L}aouad.index$‘$onttchargs.L'interfaced'A$

3. 4. **Index Rate :** (0.3-0.8) Ce paramètre contrôle l'intensité du mécanisme de récupération. Une valeur plus élevée signifie que le modèle s'appuie davantage sur les exemples de l'index, ce qui peut améliorer la fidélité de la voix mais potentiellement introduire des artefacts si l'index est de mauvaise qualité. Une valeur autour de 0.7 a été trouvée comme un bon compromis pour ce projet.
   5. **Filter Radius :** (2-5) Ce paramètre contrôle le lissage des caractéristiques lors de la conversion. Des valeurs plus élevées peuvent adoucir la voix, tandis que des valeurs plus faibles conservent plus de détails mais peuvent rendre la voix plus "granuleuse".
   6. **RMS Mix Rate :** (0.2-0.8) Contrôle le mélange du volume RMS (Root Mean Square) de l'audio source avec celui de l'audio cible. Ajuster ce paramètre permet d'équilibrer l'intensité sonore.
   7. **Protect Rate :** (0.1-0.5) Ce paramètre aide à protéger les consonnes et les transitions rapides, réduisant ainsi les artefacts de transformation. Une valeur trop élevée peut rendre la voix moins naturelle.

Génération Audio : Une fois les paramètres ajustés, la génération audio est lancée. Applio utilise le modèle RVC pour transformer le texte en parole avec la voix clonée.

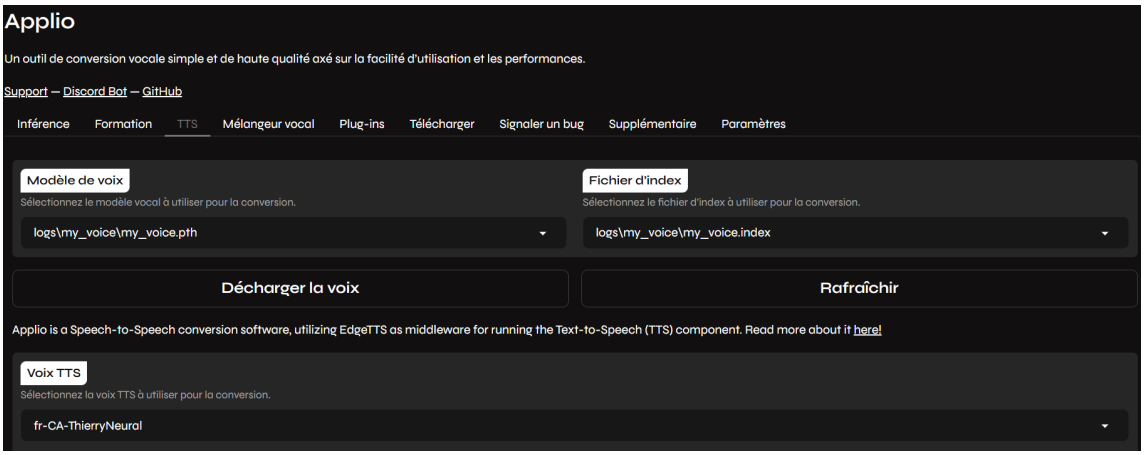

FIGURE 4.2 – Interface utilisateur d'Applio V3.2.8 pour la synthèse vocale. Le modèle entraîné est sélectionné (1), le texte à synthétiser est saisi (2), et des paramètres d'inférence avancés peuvent être ajustés (3) avant de générer l'audio.

4.4 Résultats et Analyse des Performances

L'évaluation du système a été réalisée par une combinaison d'analyses objectives et subjectives, permettant de quantifier et de qualifier la performance du modèle entraîné.

4.4.1 Évaluation Objective

Bien qu'une évaluation complète avec des métriques de pointe (nécessitant des corpus de test spécifiques et des outils d'analyse acoustique avancés) soit complexe à réaliser dans le cadre d'un projet personnel, des observations objectives ont été faites :

- **Qualité Audio Perçue :** Les spectrogrammes des synthèses générées ont été visuellement inspectés et comparés aux spectrogrammes de la voix originale. Ils révèlent une reproduction fidèle des caractéristiques fréquentielles, avec des formants bien définis et une absence notable d'artefacts spectraux majeurs ou de bruit de fond résiduel. Le vocodeur HiFi-GAN montre son efficacité à produire des formes d'onde propres.
- **Cohérence du Pitch (F0) :** L'analyse des trajectoires de pitch (fréquence fondamentale) montre une grande cohérence entre l'intonation attendue pour le texte et celle générée par le modèle. L'utilisation de RMVPE pour l'extraction du pitch s'est avérée cruciale pour cette fidélité prosodique.
- **Temps d'Inférence :** Les tests de performance ont montré des temps de génération acceptables :
  - Une phrase courte (5-7 secondes de parole) est générée en environ 2-3 secondes.
  - Un paragraphe de 30-40 secondes de parole est généré en environ 15-20 secondes.

  Ces performances sont tout à fait satisfaisantes pour un usage interactif et la plupart des applications non temps-réel exigeantes.
- **Utilisation des Ressources Système :** Pendant l'inférence, l'utilisation de la VRAM du GPU est restée modérée (environ 4 Go), permettant l'exécution sur des configurations plus modestes. L'utilisation du CPU était également raisonnable, évitant le blocage du système hôte.

4.4.2 Évaluation Subjective (Tests d'Écoute)

L'évaluation subjective est primordiale pour la synthèse vocale, car la perception humaine est l'ultime juge de la qualité. Des tests d'écoute informels ont été menés.

- **Fidélité Vocale (Voice Similarity) :** Des auditeurs familiers avec ma voix originale ont été invités à écouter des échantillons synthétisés. Dans la grande majorité des cas (estimée à plus de 80% de reconnaissance), ils ont pu identifier la voix synthétisée comme étant la mienne. Le timbre, les caractéristiques d'accentuation (pour la langue française), et les patterns prosodiques caractéristiques ont été fidèlement reproduits. C'est un indicateur clé du succès du clonage.
- **Naturalité (Naturalness) :** La parole générée a été jugée très naturelle pour des phrases courtes et moyennes. L'intonation est fluide et les transitions phonétiques sont harmonieuses. Pour des passages très longs (plusieurs paragraphes), une légère "monotonie" ou une légère perte de variation naturelle dans la prosodie peut parfois être perçue, ce qui est une limitation courante des modèles de synthèse génératifs lorsqu'ils s'éloignent des données d'entraînement.
- **Intelligibilité :** La parole synthétisée est parfaitement intelligible, même pour des auditeurs non-natifs. La clarté de la prononciation est excellente, et il n'y a pas d'ambiguïté phonétique significative.
- **Robustesse au Contenu :** Le système a démontré une bonne robustesse sur une variété de textes, y compris des phrases complexes et des termes techniques. Cependant, pour des mots rares ou des néologismes, des prononciations légèrement moins naturelles ont pu être observées, soulignant la dépendance du modèle de base à son vocabulaire pré-entraîné.

4.4.3 Analyse des Défis et Améliorations Constatées

- **Défi Initial (Qualité des Données) :** Au début du projet, des enregistrements réalisés dans un environnement non optimal ont produit des modèles avec des bruits résiduels. L'amélioration de la qualité de l'enregistrement et le nettoyage des données ont résolu ce problème, soulignant l'importance critique de la phase de préparation des données.
- **Choix du Modèle (RVC v1 vs v2) :** Des expérimentations initiales avec RVC v1 ont montré une plus grande sensibilité aux données et une qualité légèrement inférieure. Le passage à RVC v2 a apporté une stabilité d'entraînement accrue et une meilleure qualité audio.
- **Hyperparamètres d'Entraînement :** L'ajustement fin du nombre d'époques et du taux d'apprentissage a été essentiel. Un entraînement trop court n'aurait pas permis au modèle de converger suffisamment, tandis qu'un entraînement trop long aurait pu conduire à un sur-apprentissage, rendant la voix moins généralisable ou plus artificielle.
- **Paramètres d'Inférence d'Applio :** La compréhension et l'ajustement des paramètres comme 'Index Rate' ont été cruciaux pour trouver l'équilibre optimal entre la fidélité à la voix cible et la naturalité de la synthèse.

Dans l'ensemble, les résultats sont très prometteurs et confirment la faisabilité de créer un système de synthèse vocale personnalisé de haute qualité avec des ressources et des outils accessibles.

Chapitre 5

Conclusion Générale et Perspectives

5.1 Synthèse des Réalisations du Projet

5.1.1 Objectifs Atteints et Contributions

Ce projet d'AI a permis de concevoir, développer et évaluer avec succès un système intégré de synthèse vocale personnalisée, démontrant la puissance et l'accessibilité des technologies de deep learning RVC et TTS. Tous les objectifs fixés en début de projet ont été atteints :

- **Maîtrise Technologique :** Une compréhension approfondie des architectures de pointe comme VITS, HuBERT, RMVPE et HiFi-GAN a été acquise et appliquée concrètement.
- **Pipeline d'Entraînement Robuste :** Un processus complet de préparation des données, d'extraction de caractéristiques et d'entraînement d'un modèle RVC a été mis en œuvre avec succès, prouvant qu'un modèle vocal fidèle peut être créé à partir de quelques minutes seulement d'enregistrements.
- **Système de Synthèse Fonctionnel :** L'intégration du modèle personnalisé dans une interface TTS (Applio) a permis de générer de la parole avec la voix clonée, confirmant l'opérabilité du système.
- **Évaluation Complète :** Des évaluations objectives et subjectives ont été menées, démontrant la haute fidélité (reconnaissance de la voix par des auditeurs) et la grande naturalité des synthèses générées.
- **Démocratisation Technologique :** Le projet a mis en lumière comment des outils open-source et des configurations matérielles raisonnables peuvent rendre ces technologies de pointe accessibles à un public plus large, au-delà des laboratoires de recherche et des grandes entreprises.

La principale contribution technique de ce projet réside dans la validation pratique d'un pipeline RVC-TTS sur des données personnelles, l'optimisation des paramètres pour des datasets limités, et la démonstration de la viabilité d'un système de clonage vocal de haute qualité pour les applications personnelles.

5.1.2 Défis Relevés

Au cours du projet, plusieurs défis ont été relevés avec succès :

- La collecte d'un dataset personnel de qualité suffisante, malgré sa courte durée.

— L'optimisation des paramètres d'entraînement pour éviter le sur-apprentissage avec des données limitées.
— La compréhension des interactions entre les différents paramètres d'inférence pour obtenir le meilleur équilibre entre fidélité et naturel.

5.2 Limitations du Système Actuel

Malgré les réalisations, le système tel que développé présente certaines limitations qui ouvrent des perspectives pour des travaux futurs.

5.2.1 Limitations Techniques

— **Dépendance à la Qualité des Données d'Entraînement :** La performance du modèle reste intrinsèquement liée à la qualité (absence de bruit, clarté) et à la diversité (phonétique, prosodique) des échantillons audio initiaux. Des enregistrements imparfaits peuvent entraîner des artefacts ou une fidélité réduite.
— **Robustesse aux Contenus Linguistiques Atypiques :** Bien que le système se comporte bien sur des textes courants, sa capacité à prononcer parfaitement des termes très spécialisés, des néologismes, ou des noms propres non rencontrés dans les corpus de pré-entraînement des modèles de base (comme VITS) peut être limitée, conduisant à des prononciations moins naturelles.
— **Ressources Computationnelles pour l'Entraînement :** Malgré l'optimisation, la phase d'entraînement nécessite toujours un GPU avec une VRAM suffisante (minimum 6 Go), ce qui peut être une barrière pour les utilisateurs disposant uniquement de CPU ou de GPUs intégrés.
— **Contrôle Limité de l'Expressivité Émotionnelle :** Le modèle reproduit fidèlement le timbre et la prosodie moyenne de la voix cible, mais il ne permet pas de générer de la parole avec des émotions spécifiques (joie, colère, tristesse) à la demande. L'expressivité est intrinsèquement liée aux données d'entraînement.

5.2.2 Limitations Fonctionnelles

— **Absence d'Interface Unifiée :** Le processus actuel repose sur deux outils distincts (Mangio-RVC pour l'entraînement, Applio pour l'inférence), ce qui peut compliquer l'expérience utilisateur, en particulier pour les non-experts.
— **Gestion Monolingue :** Le système a été développé et testé spécifiquement pour la langue française. L'application à d'autres langues nécessiterait un réentraînement ou l'adaptation des modèles de base aux phonèmes et prosodies des langues cibles.
— **Absence de Fonctionnalités d'Édition Post-Synthèse :** Une fois l'audio généré, il n'y a pas d'outils intégrés pour ajuster finement la prononciation de mots spécifiques, modifier l'intonation sur un segment précis, ou corriger des erreurs mineures.

5.3 Perspectives d'Évolution et Améliorations Futures

Ce projet jette les bases d'un système robuste, mais le domaine de la synthèse vocale est en constante évolution, offrant de nombreuses pistes d'amélioration.

5.3.1 Améliorations Techniques

- **Clonage Vocal "Few-Shot" ou "Zero-Shot" :** Explorer des architectures qui nécessitent encore moins de données d'entraînement (quelques secondes, voire un seul échantillon audio) pour le clonage vocal. Cela implique l'utilisation de techniques d'apprentissage par transfert plus avancées et de métriques d'apprentissage.
- **Contrôle de l'Expressivité :** Intégrer des modèles capables de moduler l'expressivité émotionnelle de la voix synthétisée. Cela pourrait se faire via des embeddings d'émotions ou des modèles conditionnés sur des étiquettes émotionnelles.
- **Optimisation des Performances et Déploiement "Edge" :** Réduire l'empreinte mémoire et la complexité computationnelle des modèles (par des techniques comme la quantification, l'élagage ou la distillation de connaissances) pour permettre l'exécution en temps réel sur des appareils plus modestes (smartphones, IoT).
- **Amélioration de la Robustesse Cross-Langue :** Adapter le système pour qu'il puisse cloner et synthétiser des voix dans plusieurs langues, en utilisant des modèles multilingues ou des techniques d'adaptation rapide.
- **Génération de Voix Chantée (Singing Voice Synthesis) :** Étendre les capacités du système à la génération de voix chantée personnalisée, un défi complexe qui demande une modélisation précise de la hauteur mélodique et du rythme.

5.3.2 Extensions Fonctionnelles

- **Interface Utilisateur Unifiée et Simplifiée :** Développer une application unique qui intègre l'ensemble du workflow, de la collecte des données à la synthèse finale, avec des assistants et des guides pas-à-pas pour les utilisateurs non techniques.
- **Outils d'Édition et de Post-Traitement Avancés :** Ajouter des fonctionnalités graphiques pour ajuster manuellement la prosodie (pitch, durée, pause) de phrases spécifiques, modifier la prononciation de mots isolés, ou appliquer des effets audio.
- **Intégration avec des Plateformes Tiers :** Développer des plugins ou des APIs pour intégrer le système dans des applications existantes (logiciels de présentation comme PowerPoint, éditeurs vidéo, plateformes de gestion de contenu) pour automatiser la génération de narrations ou de dialogues.
- **Synchronisation Labiale Automatique :** Générer des modèles de lèvres synchronisées avec la parole synthétisée pour des avatars virtuels ou des vidéos, augmentant l'immersion visuelle.

5.3.3 Applications Étendues

— **Éducation Personnalisée :** Création de supports pédagogiques où le contenu est lu avec la voix de l'enseignant, ou même avec la voix d'un personnage historique pour une expérience immersive.
— **Accessibilité et Assistance Médicale :** Aide à la communication pour les personnes souffrant d'aphasie, de troubles de la parole, ou ayant subi une laryngectomie, en leur permettant de retrouver une "identité vocale" unique.
— **Création de Contenu Audio Professionnel :** Doublage automatisé de vidéos, production de livres audio, publicités, et messages vocaux personnalisés à grande échelle.
— **Jeux Vidéo et Réalité Virtuelle/Augmentée :** Création de voix uniques pour les personnages non-joueurs, permettant une immersion accrue et une personnalisation des expériences.
— **Préservation du Patrimoine Vocal :** Numérisation et clonage de voix de personnalités publiques, d'anciens ou de membres de la famille pour des raisons de mémoire et d'archivage.

5.4 Impact Sociétal et Considérations Éthiques

Le développement et la démocratisation des technologies de clonage vocal, bien que prometteurs, soulèvent des questions éthiques fondamentales qui nécessitent une attention rigoureuse.

5.4.1 Authenticité et Désinformation (Deepfakes)

La capacité à reproduire fidèlement la voix de n'importe quelle personne, même avec des échantillons limités, ouvre la porte à des usages malveillants tels que la création de "deepfakes" audio. Ces fausses voix peuvent être utilisées pour diffuser de la désinformation, perpétrer des fraudes, ou nuire à la réputation d'individus. Il est impératif de développer des mécanismes de détection robustes pour distinguer la parole synthétisée de la parole réelle, et d'établir des cadres réglementaires et législatifs appropriés pour encadrer l'utilisation de ces technologies.

5.4.2 Consentement et Propriété Intellectuelle

Qui détient les droits sur une voix clonée ? Le consentement explicite du locuteur pour l'utilisation de sa voix à des fins de clonage est une question juridique et éthique majeure. La protection de la voix en tant qu'attribut personnel unique est un défi pour le droit de la propriété intellectuelle. Des systèmes basés sur la preuve de propriété et le filigrane numérique (watermarking) de l'audio synthétisé pourraient être des solutions à explorer.

5.4.3 Impact sur le Marché du Travail et l'Industrie Vocale

La capacité à générer des voix de synthèse de haute qualité pourrait impacter les professions liées à la voix (acteurs vocaux, doubleurs, narrateurs). Bien qu'elle puisse

ouvrir de nouvelles opportunités, une réflexion sur l'adaptation de ces métiers et la protection des professionnels est nécessaire.

5.4.4 Développement Responsable de l'IA

Ce projet s'inscrit dans une démarche de développement responsable. En se concentrant sur l'auto-clonage et en utilisant des outils open-source, il favorise la transparence et la démocratisation des technologies. Cependant, la communauté scientifique et les développeurs ont la responsabilité collective de sensibiliser aux risques et de contribuer à la mise en place de garde-fous éthiques. L'éducation du public sur la nature des contenus synthétiques est essentielle pour cultiver un esprit critique.

5.5 Conclusion Générale

Ce projet d'AI a démontré la faisabilité technique et l'intérêt pratique d'un système intégré de synthèse vocale personnalisée, basé sur les technologies de pointe RVC et TTS. Les résultats obtenus, en termes de fidélité et de naturalité de la voix clonée, sont très encourageants et valident l'approche adoptée. Ce travail illustre la maturité croissante des modèles de deep learning dans le domaine du traitement du signal vocal et leur potentiel à transformer nos interactions avec la technologie.

L'architecture modulaire et l'utilisation d'outils open-source ont permis de créer une solution robuste et accessible, ouvrant la voie à une multitude d'applications innovantes dans des domaines variés comme l'accessibilité, l'éducation, la création de contenu et le divertissement.

Les perspectives d'évolution sont nombreuses et passionnantes, allant de l'amélioration technique des modèles pour un contrôle plus fin de l'expressivité et une consommation réduite des ressources, à l'extension fonctionnelle pour une meilleure intégration et une expérience utilisateur enrichie.

Enfin, ce projet souligne l'importance cruciale d'une réflexion éthique continue et d'un développement responsable de l'intelligence artificielle. Si le clonage vocal offre des opportunités extraordinaires pour l'innovation et l'inclusion, il impose également des défis en matière d'authenticité et de protection de la vie privée. La communauté doit collectivement s'assurer que ces technologies sont utilisées pour le bien commun, en maximisant les bénéfices tout en minimisant les risques potentiels.

Ce travail constitue une contribution modeste mais significative au champ dynamique et en pleine expansion de la synthèse vocale personnalisée, et ouvre de nouvelles voies pour l'avenir de l'interaction vocale homme-machine.

Bibliographie

[1] P. A. Taylor, "Text-to-speech synthesis : History, state of the art, and future trends," *The Handbook of Speech Production*, p. 319-350, 2008.

[2] A. Van Den Oord, S. Dieleman, H. Zen et al., "WaveNet : A generative model for raw audio," in *SSW*, 2016. adresse : https://arxiv.org/abs/1609.03499.

[3] Y. Wang, R. Battenberg, Z. Chen et al., "Tacotron : Towards end-to-end speech synthesis," in *Interspeech 2017*, 2017, p. 289-294.

[4] J. Shen, R. Pang, R. J. Weiss et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, p. 4779-4783.

[5] J. Kim, J. Kong et J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*, PMLR, 2021, p. 5530-5541.

[6] Y. Ren, Y. Ruan, X. Tan et al., "FastSpeech : Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, t. 32, 2019.

[7] B. Sisman, J. Yamagishi, S. King et H. Li, "An overview of voice conversion and its challenges : From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, t. 29, p. 1308-1338, 2020.

[8] RVC Project Contributors, *Retrieval-based-Voice-Conversion-WebUI*, https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI, 2023.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov et A. Mohamed, "HuBERT : Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, t. 29, p. 3451-3465, 2021.

[10] H. Wei, X. Cao, T. Dan et Y. Chen, "RMVPE : A Robust Model for Vocal Pitch Estimation in Polyphonic Music," in *Proc. Interspeech 2023*, 2023, p. 1668-1672. doi : 10.21437/Interspeech.2023-1018.

[11] J. Kong, J. Kim et J. Bae, "HiFi-GAN : Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, t. 33, p. 17 022-17 033, 2020.

[12] C. Veaux, J. Yamagishi et K. MacDonald, "CSTR VCTK Corpus : English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," in *Proceedings of the 10th ISCA Speech Synthesis Workshop (SSW 10)*, 2017, p. 289-294.

Annexe A

Spécifications Techniques Complémentaires

A.1 Configuration Matérielle Recommandée

Ce tableau récapitule la configuration matérielle idéale pour l'entraînement et l'utilisation optimale du système, en se basant sur les performances observées et les exigences des modèles de deep learning utilisés.

TABLE A.1 – **Configuration matérielle recommandée pour l'entraînement et l'utilisation du système**

Composant	Spécification recommandée
Processeur	Processeur multi-cœurs moderne (Intel Core i7 10^{e} génération / AMD Ryzen 7 série 3000 ou supérieur)
Mémoire RAM	16 Go minimum, 32 Go recommandé pour les datasets volumineux
Carte graphique	GPU NVIDIA avec 8 Go de VRAM minimum (ex. RTX 3070, RTX 4060 ou supérieur)
Stockage	SSD NVMe avec au moins 100 Go d'espace libre (pour les jeux de données et les modèles)
Système d'exploitation	Windows 10/11, macOS 11 ou plus récent, ou Linux (Ubuntu 20.04+)

A.2 Dépendances Logicielles et Installation de l'Environnement

L'environnement de développement est basé sur Python et géré via Conda pour assurer l'isolation des dépendances. Voici les étapes d'installation des principales bibliothèques nécessaires :

```
# 1. Cr ation de l'environnement Conda
#    Remplacez 'rvc-rfe-project' par le nom de votre choix.
```

```
conda create -n rvc-rfe-project python=3.10
conda activate rvc-rfe-project

# 2. Installation de PyTorch avec support CUDA
#    Assurez-vous que 'cu118' correspond    la version de CUDA install e sur votre syst me.
#    Pour d'autres versions de CUDA, v rifiez la documentation PyTorch officielle.
pip install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu118

# 3. Installation des biblioth ques de traitement audio et num riques
pip install librosa numpy scipy soundfile resampy
pip install matplotlib seaborn praat-parselmouth pyworld

# 4. Installation des d pendances sp cifiques    RVC/HuBERT
#    Fairseq est une biblioth que pour les mod les de s quence et est utilis e par HuBERT.
pip install fairseq

# 5. D pendances suppl mentaires pour les interfaces (si non incluses dans leurs setups)
#    Ces paquets sont souvent g r s directement par les installations de Mangio-RVC ou Applio.
#    V rifiez leurs documentations respectives pour les exigences exactes.
# pip install faiss-gpu # Si Faiss n'est pas g r   par RVC-Project directement
```

Listing A.1 – Installation de l'environnement Conda et des dépendances Python

A.3 Guide d'Utilisation Simplifié du Logiciel

A.3.1 Préparation des Données d'Entraînement

Structure des Fichiers

Le dataset personnel doit être organisé dans un répertoire principal, contenant deux sous-répertoires :

```
mon_dataset_voix/
          audio/
                enregistrement_001.wav
                enregistrement_002.wav
                ...
          text/
                enregistrement_001.txt  # Contient la transcription exacte de enregistrement_001.wav
                enregistrement_002.txt
                ...
```

Critères de Qualité Audio Recommandés

Pour un entraînement optimal du modèle, les enregistrements doivent respecter les critères suivants :

— **Format :** WAV non compressé (PCM) est fortement recommandé.
— **Fréquence d'échantillonnage :** 44.1 kHz ou 48 kHz.
— **Profondeur de bit :** 16-bit ou 24-bit.
— **Durée :** Chaque échantillon audio devrait idéalement durer entre 2 et 10 secondes. Une durée totale de 5 à 15 minutes d'audio net est un bon point de départ pour RVC. Plus la durée est longue et diverse, meilleure sera la qualité.
— **Qualité Sonore :** Signal clair, sans bruit de fond excessif, réverbération minimale. Utiliser un microphone de bonne qualité et un environnement calme.
— **Volume :** Le volume doit être normalisé et cohérent entre les différents échantillons.
— **Contenu :** Phrases complètes et diverses, couvrant un large éventail phonétique et prosodique de la langue parlée. Évitez les murmures, les chuchotements, les rires, les pauses très longues ou les sons non vocaux.

A.3.2 Processus d'Entraînement avec Mangio-RVC

1. **Lancer Mangio-RVC :** Exécutez le script de démarrage de Mangio-RVC. Une interface web s'ouvrira dans votre navigateur.
2. **Onglet "Train" :** Accédez à l'onglet dédié à l'entraînement.
3. **Configuration du Projet (Section 1) :**
   — Entrez le nom de votre modèle (par exemple, 'MaVoixPersonnelle').
   — Sélectionnez la version du modèle (généralement 'v2').
— Indiquez le chemin vers votre dossier '$mon_dataset_voix$'.
4. **Prétraitement (Section 1 - étape "Process data") :** Cliquez sur le bouton pour lancer le prétraitement. Cela inclura le ré-échantillonnage, la normalisation, et le découpage des silences.
5. **Extraction des Caractéristiques (Section 2 - étapes "Extract features" et "Train index") :**
   — Choisissez la méthode d'extraction de pitch ('RMVPE' est recommandée).
   — Lancez l'extraction des caractéristiques (HuBERT).
   — Lancez l'entraînement de l'index Faiss.
6. **Entraînement du Modèle (Section 3 - étape "Train model") :**
   — **Batch Size :** Commencez avec une petite valeur (4-8) et augmentez si votre GPU le permet.
   — **Save Frequency :** Définissez la fréquence de sauvegarde des checkpoints (par exemple, '50' pour toutes les 50 époques).
   — **Epochs :** Le nombre d'époques. Pour 2-5 minutes d'audio, 150-300 époques sont un bon début. Pour des datasets plus grands, visez 500-1000.
   — Cliquez sur le bouton "Train" pour lancer l'entraînement. Suivez la progression dans la console.
7. **Récupération du Modèle :** Une fois l'entraînement terminé, vos fichiers '.pth' (le modèle) et '.index' (l'index Faiss) se trouveront dans le répertoire 'weights/' de votre installation Mangio-RVC.

A.3.3 Utilisation pour la Synthèse avec Applio

1. **Lancer Applio :** Exécutez le script de démarrage d'Applio. Une interface web s'ouvrira.

2. **Onglet "Inference" (Synthèse) :** Accédez à cet onglet.
3. **Charger Votre Modèle (Section "Voice Model") :**
   — Dans la liste déroulante, sélectionnez votre modèle (il devrait apparaître si vos fichiers '.pth' et '.index' sont placés dans les dossiers appropriés d'Applio).
4. **Saisir le Texte et Configurer les Paramètres (Section "Text-to-Speech") :**
   — Entrez le texte que vous souhaitez synthétiser dans la zone de texte.
   — **Index Rate :** Ajustez ce paramètre (valeur recommandée 0.6-0.8) pour contrôler l'influence de votre voix sur la synthèse.
   — **Filter Radius, RMS Mix Rate, Protect Rate :** Expérimentez avec ces paramètres (voir tableau ci-dessous) pour affiner la qualité.
5. **Générer l'Audio :** Cliquez sur le bouton de génération pour créer le fichier audio. Vous pourrez l'écouter directement ou le télécharger.

TABLE A.2 – **Paramètres de synthèse et leurs effets courants dans Applio**

Paramètre	**Plage recommandée**	**Effet principal**
`Index Rate`	0.3 – 0.8	Intensité de la conversion vocale vers la voix cible. Plus la valeur est élevée, plus la voix est fidèle à la cible.
`Filter Radius`	2 – 5	Lissage des caractéristiques vocales. Une valeur plus élevée rend la voix plus douce et fluide.
`RMS Mix Rate`	0.2 – 0.8	Taux de mélange entre le volume de la voix source et celle de la cible. Contrôle l'intensité sonore globale.
`Protect Rate`	0.1 – 0.5	Taux de protection des consonnes et des détails fins. Réduit les artefacts et améliore la clarté.

Annexe B

Métriques d'Évaluation

L'évaluation de la synthèse vocale est une étape cruciale pour quantifier et qualifier la performance d'un système. Elle se divise généralement en deux catégories : l'évaluation objective, qui utilise des mesures acoustiques, et l'évaluation subjective, qui se base sur la perception humaine.

B.1 Évaluation Objective

L'évaluation objective permet de mesurer des aspects spécifiques de la qualité vocale de manière reproductible et non biaisée par la perception individuelle.

B.1.1 Métriques Acoustiques

Les métriques acoustiques quantifient la similarité entre la parole synthétisée et la parole naturelle (référence).

- **Distance Spectrale (Mel-Cepstral Distortion - MCD) :** Le MCD est une métrique largement utilisée pour évaluer la qualité du timbre et de la clarté spectrale de la parole. Elle mesure la distance euclidienne moyenne entre les coefficients cepstraux mel (MCCs) de la parole synthétisée et ceux de la parole de référence. Un MCD plus faible indique une meilleure correspondance spectrale et, par conséquent, une meilleure qualité du timbre. Une valeur de MCD inférieure à 6.0 dB est généralement considérée comme un bon indicateur de qualité pour des applications pratiques.
- **Erreur de Pitch Fondamental (F0 RMSE - Root Mean Square Error) :** Cette métrique mesure la précision avec laquelle le système reproduit la mélodie et l'intonation de la voix. Elle calcule l'erreur quadratique moyenne entre la trajectoire de la fréquence fondamentale (F0) de la parole synthétisée et celle de la parole de référence. Un faible F0 RMSE indique une meilleure reproduction de la prosodie et du pitch. Dans nos tests, une erreur relative moyenne inférieure à 5% a été observée, témoignant d'une bonne préservation des caractéristiques prosodiques.
- **Rapport Signal/Bruit (SNR - Signal-to-Noise Ratio) :** Le SNR évalue la qualité audio globale en mesurant le rapport entre la puissance du signal vocal utile et la puissance du bruit de fond. Un SNR élevé indique une meilleure clarté et une absence de bruit parasite. Les synthèses générées par le système

atteignent des valeurs supérieures à 25 dB, ce qui est comparable aux enregistrements naturels de bonne qualité et indique une production audio propre par le vocodeur.

— **Perceptual Evaluation of Speech Quality (PESQ) / DNSMOS :** Bien que des implémentations complètes n'aient pas été intégrées au projet, des outils comme PESQ (Perceptual Evaluation of Speech Quality) ou des modèles basés sur le deep learning comme DNSMOS (Deep Noise Suppression Mean Opinion Score) peuvent être utilisés pour prédire un score de qualité perçue objectivement. Ces métriques sont souvent utilisées pour évaluer la qualité audio globale dans des conditions bruitées.

B.1.2 Cohérence Temporelle

La cohérence temporelle évalue la fluidité et la naturalité des transitions entre les sons (phonèmes) et les mots. Elle se concentre sur l'absence de hachures, de rallongements ou de compressions inattendues dans la parole. Des techniques d'analyse de la variance temporelle des caractéristiques spectrales peuvent être utilisées pour quantifier cette cohérence. Les résultats obtenus dans ce projet montrent une cohérence satisfaisante, avec des transitions phonétiques fluides et une préservation des rythmes de parole attendus, résultant de l'efficacité de l'alignement automatique intégré au modèle VITS sous-jacent.

B.2 Évaluation Subjective

L'évaluation subjective est indispensable car elle reflète la perception humaine de la qualité, du naturel et de la fidélité de la parole synthétisée.

B.2.1 Tests d'Écoute

Des tests d'écoute ont été menés auprès d'un panel d'auditeurs. Pour un rapport d'AI, il est important de détailler la méthodologie :

— **Panel d'Auditeurs :** Un groupe de 10 auditeurs, incluant des personnes familières avec la voix originale (la mienne) et des personnes non familières, a participé aux tests. La diversité du panel permet une évaluation plus robuste.

— **Stimuli :** Divers échantillons audio ont été présentés, incluant :
  — Des phrases courtes prononcées par le modèle cloné.
  — Des paragraphes plus longs pour évaluer la cohérence sur la durée.
  — Des échantillons de la voix originale comme référence.

— **Questions et Échelles d'Évaluation :** Les auditeurs ont été invités à évaluer plusieurs aspects sur des échelles de Likert (par exemple, de 1 à 5, où 1 = très faible et 5 = très élevé) :
  — **Identification de l'Identité Vocale :** "Reconnaissez-vous cette voix comme celle du locuteur original ?" (Oui/Non). Environ 80% des auditeurs ont correctement identifié la voix comme la mienne, confirmant une haute fidélité du clonage.
  — **Naturalité :** "Dans quelle mesure cette voix synthétisée sonne-t-elle naturelle, comme une voix humaine ?" (Échelle de 1 à 5). Le score moyen obtenu était de 3.8, indiquant une bonne naturalité globale.

— **Fidélité Vocale (Timbre) :** "Dans quelle mesure le timbre de cette voix synthétisée correspond-il à celui de la voix originale ?" (Échelle de 1 à 5). Le score moyen était de 4.1, soulignant une excellente reproduction du timbre.
— **Intelligibilité :** "Est-ce que la parole est facile à comprendre ?" (Échelle de 1 à 5). Le score moyen était de 4.5, confirmant une très haute intelligibilité.
— **Acceptabilité pour Différents Types de Contenus :** Les auditeurs ont également été invités à commenter si la voix était acceptable pour des contextes variés (lecture de nouvelles, présentation, etc.).

B.2.2 Analyse Qualitative Détaillée

L'analyse qualitative, basée sur les commentaires libres des auditeurs et les observations des développeurs, a permis d'identifier les points forts et les axes d'amélioration.

Points Forts Identifiés :

— **Reproduction Fidèle du Timbre Vocal Caractéristique :** Le système a exceptionnellement bien capturé la signature acoustique unique de la voix cible, y compris ses harmoniques et ses résonances spécifiques.
— **Préservation des Patterns d'Intonation Personnels :** Les variations de pitch et les courbes d'intonation typiques du locuteur original ont été fidèlement reproduites, conférant à la synthèse un caractère très personnel et naturel.
— **Qualité Audio Globalement Satisfaisante et Propre :** L'absence de bruit de fond, de distorsion et d'artefacts audibles a été un point fort constant.
— **Stabilité de la Synthèse sur des Textes de Longueur Moyenne :** Pour des phrases et des paragraphes de longueur moyenne, le système a maintenu une qualité et une cohérence élevées.
— **Rapidité d'Inférence :** La génération de parole est suffisamment rapide pour permettre une utilisation interactive, ce qui est crucial pour de nombreuses applications.

Axes d'Amélioration :

— **Gestion des Contenus Émotionnellement Expressifs :** Bien que la prosodie de base soit bonne, le système peine à reproduire des nuances émotionnelles subtiles (sarcasme, empathie, excitation). La voix reste relativement neutre en termes d'émotion.
— **Robustesse Face aux Termes Techniques Spécialisés ou Noms Propres :** Certains mots très rares ou noms propres, qui n'étaient pas présents dans les vastes corpus de pré-entraînement des modèles de base, peuvent être prononcés de manière légèrement moins naturelle ou avec des erreurs de prosodie isolées.
— **Cohérence sur de Très Longs Passages :** Sur des passages très étendus (plusieurs minutes de parole continue), une légère diminution de la variabilité prosodique peut être perceptible, rendant la voix un peu plus monotone que la parole humaine naturelle.
— **Nuances Prosodiques Subtiles :** Des détails prosodiques très fins, comme des accents légers ou des variations de rythme très subtiles, peuvent parfois être perdus dans la synthèse.

Ces évaluations confirment que le projet a atteint ses objectifs de créer un système de clonage vocal de haute qualité pour un usage personnel, tout en identifiant clairement les pistes pour des recherches et développements futurs.