Master 2 Langue et Informatique
NOM : XIE
Prénom : Shilin
M4SOL202
N °étudiant : 21102552

Rapport final du project Synthèse de la parole spontanée et analyse paralinguistique

Introduction

La synthèse vocale est une technologie d'interaction homme-machine importante qui vise à convertir du texte en sortie vocale naturelle et fluide. Avec le développement continu de l'intelligence artificielle et des technologies de traitement du langage naturel, la technologie de synthèse vocale est largement utilisée dans des domaines tels que les assistants vocaux, les services clients intelligents, la navigation vocale, etc. Le développement de la technologie de synthèse vocale améliore non seulement la naturalité et la commodité de l'interaction homme-machine, mais fournit également des outils d'assistance importants pour les personnes handicapées de la parole.

Le principe de la synthèse vocale consiste à simuler le mouvement des organes vocaux humains, en combinant les principes de la phonétique et les techniques de traitement du signal numérique pour générer un son vocal proche de la parole humaine naturelle. Dans le processus de synthèse vocale, des facteurs tels que la durée des phonèmes, les caractéristiques de prononciation, l'intonation, le volume, la vitesse de parole, etc., doivent être pris en compte pour générer une parole naturelle, fluide et expressive. Ces facteurs affectent directement la naturalité et la compréhensibilité de la parole synthétique, ce qui rend la recherche et l'amélioration de la technologie de synthèse vocale significatives. De plus, le choix du moteur de synthèse vocale et la configuration des paramètres ont également un impact sur la qualité de la synthèse.

Actuellement, les méthodes de synthèse vocale courantes comprennent les méthodes basées sur des règles, les méthodes de modélisation statistique et les méthodes d'apprentissage profond. Ces méthodes ont leurs propres avantages et inconvénients, et peuvent être choisies en fonction des besoins spécifiques et des scénarios d'application pour la synthèse vocale.

Ce projet vise à utiliser des outils de synthèse vocale tels que Praat, Mbrola, Espeak et Emofilt pour refaire la synthèse vocale à partir de matériaux existants, puis à évaluer la qualité des fichiers audio synthétisés à l'aide de méthodes d'évaluation objectives (audeep, un modèle basé sur le DNN) et subjectives. En approfondissant l'étude des effets de la synthèse vocale de différents outils et méthodes, ainsi que la faisabilité et l'efficacité des méthodes d'évaluation, ce projet vise à fournir des références et des enseignements pour le développement et l'application ultérieurs de la technologie de synthèse vocale.

Préparation du corpus

Le corpus de parole a été méthodiquement choisi afin de représenter une conversation authentique entre deux locuteurs. Un extrait de trois minutes a été délibérément sélectionné, mettant en scène un journaliste radio (L2) et un expert du football (L1).

Dans ce projet, mon jeu de données de base est composé de cinq segments audio extraits des trois minutes d'enregistrement. Ces segments ont une durée supérieure à 7 secondes et contiennent uniquement des enregistrements clairs d'un seul locuteur. Ces segments audio ont été annotés dans un projet précédent à l'aide du logiciel Praat, et chaque fichier audio est accompagné d'un fichier texte (txt) correspondant contenant la transcription textuelle du contenu audio.

Pour mieux comprendre la prononciation correcte de chaque mot, une transcription graphème-phonème est réalisée, ce qui constitue une étape clé du processus de synthèse vocale. La transcription graphème-phonème convertit le texte écrit en une séquence de phonèmes correspondante, permettant ainsi au système de synthèse de comprendre précisément les règles de prononciation de chaque mot. Dans ce projet, cette étape utilise les logiciels Praat et WebMAUS. Tout d'abord, un script Praat est utilisé pour extraire les caractéristiques de chaque phonème, telles que sa durée et sa fréquence fondamentale, à partir des fichiers audio. Ensuite, les fichiers audio et textuels sont transférés vers un serveur à l'aide de l'outil WebMAUS, où son fonctionnement automatique permet de convertir le texte en phonèmes SAMPA et de les aligner chronologiquement avec l'audio, générant ainsi de nouveaux fichiers TextGrid. Ces nouveaux fichiers TextGrid enregistrent chaque phonème ainsi que sa plage de temps dans le fichier audio.

De plus, les fichiers pho sont également essentiels dans ce projet, leur première colonne étant toujours le phonème conforme à la norme SAMPA, tandis que la deuxième colonne représente généralement la durée de ce phonème. Les différentes méthodes pour obtenir ces fichiers seront abordées ultérieurement.

Grâce à ce processus, nous obtenons des annotations de phonèmes précises, jetant ainsi les bases des travaux ultérieurs de synthèse vocale et d'évaluation.

Méthodologie

Extraction la prosodie des tours de parole

Dans le traitement vocal, l'extraction du rythme et de l'intonation du discours est une tâche importante, et comprendre la hauteur du son est essentiel. La hauteur du son fait référence à la fréquence du son, et elle influence directement notre perception du son, comme les tons élevés et bas. Au cours de ce processus, nous devons faire la distinction entre deux concepts liés mais différents : la hauteur (Pitch) et la fréquence fondamentale (F0).

Le pitch est la perception subjective de la hauteur du son par les humains, indiquant si un son est aigu, moyen ou grave. Le pitch est déterminé par la fréquence du son, généralement exprimée en hertz (Hz). En revanche, la F0 est la fréquence fondamentale réelle du son, représentant les vibrations périodiques les plus basses dans la forme d'onde sonore. Elle est le composant de fréquence le plus fondamental du son, déterminant sa hauteur. La F0 est un paramètre important dans la synthèse vocale et l'analyse acoustique, généralement exprimée en hertz (Hz). En synthèse vocale, en contrôlant les variations de F0, on peut modifier la hauteur du son synthétisé, créant ainsi différents tonalités et styles de discours.

L'outil de génération vocale utilisé dans ce projet, mbrola, est un synthétiseur vocal basé sur une concaténation de deux phonèmes. Il nécessite une liste de phonèmes en entrée, accompagnée d'informations sur le rythme (durée des phonèmes) et l'intonation (description linéaire segmentée des tons) pour générer des échantillons vocaux sur 16 bits, à la fréquence d'échantillonnage de la base de données de phonèmes bimoraux. Ainsi, ce n'est pas un synthétiseur de texte à voix (TTS), car il n'accepte pas de texte brut en entrée. Pour obtenir un système TTS complet, il est nécessaire de combiner la génération vocale avec un système de traitement de texte qui produit des commandes de génération vocale et de rythme.

L'utilisation de cet outil permet de contrôler de manière flexible la durée des phonèmes et l'intonation, ce qui permet de générer des résultats de synthèse vocale plus naturels et précis.

Pour comparer l'impact de ces deux éléments sur les fichiers audio générés, j'utiliserai deux scripts Praat pour extraire, en plus de la durée des phonèmes, les valeurs de hauteur tonale ou les valeurs de F0 initiale (Hz) et centrale (Hz), puis générer deux types de fichiers pho: Premier type T1-T5.pho , seconde type T1_-T5_.pho.

Dans ces fichiers pho générés, il est parfois nécessaire de modifier les symboles des phonèmes et d'utiliser le caractère _ pour représenter les périodes de silence.

Ensuite, je passerai ces fichiers pho à travers mbrola en utilisant la commande suivante dans le terminal : "mbrola -t 1.2 -f 0.8 fr1/fr1 exemple.pho exemple.wav", pour obtenir les fichiers wav correspondants.

Les fichiers nommés T1 à T5 contiennent les valeurs de la fréquence fondamentale de chaque phonème, tandis que les fichiers nommés T1_ à T5_ contiennent les valeurs de la fréquence fondamentale initiale et centrale de chaque phonème.

Ces deux ensembles de fichiers ont des tailles identiques et je ne peux pas non plus entendre de différence entre les deux ensembles. Dans l'ensemble, lorsque les énoncés sont cohérents, les fichiers audio générés sonnent très semblables à un discours humain réel. Cependant, ils manquent d'une cohérence naturelle, ce qui est crucial en traitement vocal.

En ce qui concerne la cohérence vocale, nous devons d'abord nous concentrer sur la transition entre les phonèmes. Les transitions entre phonèmes consécutifs dans l'audio synthétisé devraient être fluides, sans rupture ni changement soudain évident. Des transitions phonémiques non fluides peuvent rendre l'audio incohérent et affecter la compréhension. De plus, la liaison entre les syllabes est également importante. Dans l'audio synthétisé, la liaison entre les syllabes consécutives devrait être naturelle, sans pause ni rupture inutile. Une liaison syllabique fluide aide à maintenir la cohérence vocale, rendant l'audio synthétisé plus facile à comprendre et à accepter. La vitesse et le rythme de l'audio synthétisé devraient correspondre à ceux de la parole réelle, sans paraître trop rapide ou trop lent. Une vitesse et un rythme naturels contribuent à rendre l'audio synthétisé plus réaliste et naturel.

De plus, l'expression vocale dans l'audio synthétisé est également un facteur important pour évaluer sa naturalité. L'expression vocale comprend des aspects tels que la hauteur, le volume, l'intonation, etc. L'audio synthétisé devrait pouvoir reproduire avec précision l'expression vocale de la parole réelle, donnant ainsi l'impression à l'auditeur que l'audio synthétisé sonne aussi naturel qu'un être humain réel.

Dans ces fichiers audio synthétisés, il arrive parfois qu'il y ait une distorsion, peut-être en raison de la longueur excessive ou insuffisante des phonèmes, du décalage de la fréquence fondamentale, etc. De plus, dans les cas où le volume du son original est relativement faible, le son de l'audio généré peut avoir une tonalité ou une hauteur plus élevée que prévu, ce qui donne souvent une impression de stridence ou d'aigreur. Cela pourrait être dû à une détection de la fréquence fondamentale moins précise lorsque le volume du son original est réduit.

Étant donné que la prosodie de ces fichiers est extraite directement à partir de l'audio source, le rythme et l'intonation sont plus ou moins similaires à ceux de l'audio source. Cependant, si l'on génère des fichiers audio uniquement à partir du texte sans référence à des fichiers audio, cela crée une situation différente.

Génération des durées

Lors de la conversion directe du texte en audio, il est nécessaire tout d'abord de représenter le texte en termes de phonèmes SAMPA, puis d'ajouter à chacun de ces phonèmes la durée correspondante, ainsi que la fréquence fondamentale (F0) et d'autres paramètres. La durée des phonèmes peut être influencée par les phonèmes environnants, tels que la durée des phonèmes précédents et suivants ainsi que la position de l'accent tonique dans la syllabe. Par conséquent, il est possible de déterminer la durée des phonèmes en fonction de leur contexte. Les 11 règles de Klatt sont souvent utilisées à cet effet, permettant de calculer la durée appropriée des phonèmes en fonction de leur contexte et d'autres facteurs, afin d'obtenir une synthèse vocale plus naturelle et cohérente.

De plus, certains systèmes de synthèse vocale peuvent avoir des paramètres par défaut pour la longueur des phonèmes, ce qui permet de déterminer la durée des phonèmes en fonction de ces paramètres. J'ai choisi d'utiliser espeak. eSpeak est un moteur de synthèse vocale open-source qui convertit le texte en parole de manière naturelle. Il est écrit en langage C et peut fonctionner sur plusieurs plateformes, notamment Windows, Linux et macOS.

En entrant la commande suivante dans le terminal : "espeak -w example.wav -v fr -x “text”", vous pouvez obtenir un fichier audio correspondant au texte d'entrée généré automatiquement. Les durées des phonèmes dans ce cas sont uniformes, ce qui signifie que chaque phonème a la même durée et la même hauteur tonale, ce qui donne une impression de synthèse vocale très artificielle, semblable à la voix d'un robot.

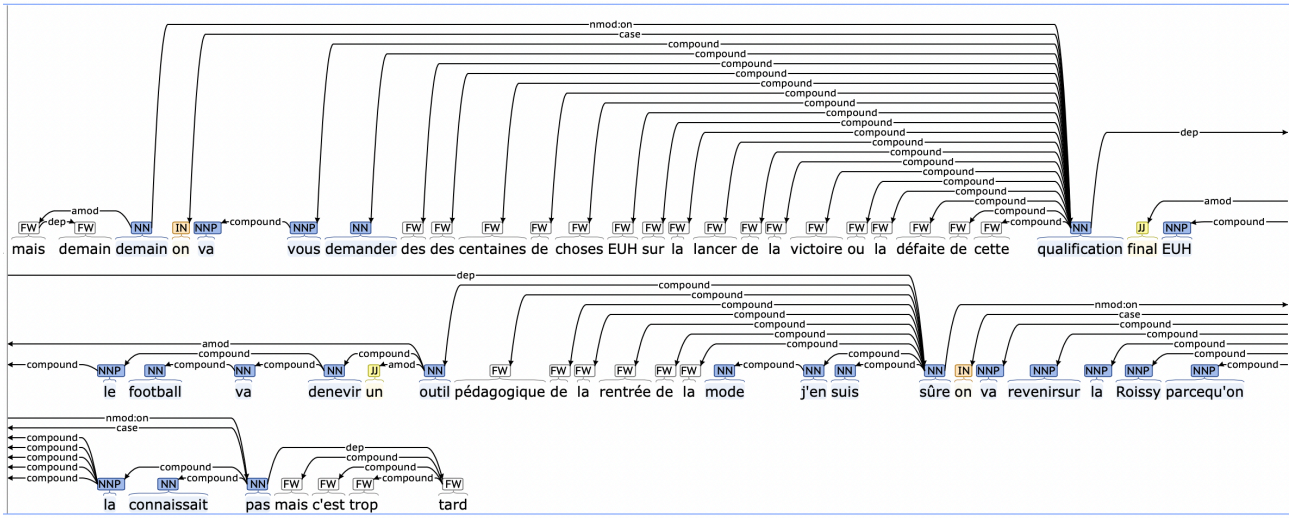
En revanche, lorsque l'on ajoute des durées aux phonèmes selon les règles de Klatt, le résultat est plus naturel et claire, avec des variations de rythme et de prosodie.

Génération des pauses et la courbe mélodique

Dans les fichiers audio synthétisés, les pauses jouent un rôle très important. Elles contribuent à segmenter le texte en unités sémantiques plus petites, telles que des phrases ou des syntagmes, ce qui renforce la compréhensibilité de la parole. Lors du processus de synthèse, déterminer les positions et les durées appropriées des pauses est crucial, car elles peuvent refléter le rythme et les variations d'intonation du locuteur, rendant ainsi la parole plus naturelle et fluide. Des pauses bien placées permettent aux auditeurs de mieux comprendre le sens du langage et ajoutent également à sa naturalité.

De plus, la longueur et la position des pauses peuvent également véhiculer l'attitude et l'état émotionnel du locuteur. Par exemple, la durée des pauses peut exprimer différents degrés de réflexion, d'hésitation ou d'accentuation, enrichissant ainsi les possibilités d'expression de la parole.
En outre, l'accent est également très important dans la synthèse vocale. Il est utilisé pour mettre en évidence des syllabes ou des mots spécifiques dans la parole, afin de souligner leur importance ou d'exprimer une intention particulière. Dans la synthèse vocale, l'accent est généralement réalisé en ajustant l'intensité et la hauteur des syllabes. L'accentuation dans les fichiers audio synthétisés rend la parole plus vivante, expressive et renforce son effet d’expression.

Pour ajuster les pauses et l'accentuation dans les fichiers audio synthétisés, il est possible de se baser sur l'analyse syntaxique. En utilisant corenlp pour analyser la syntaxe du contenu vocal, il est possible de générer des arbres syntaxiques, ce qui permet de mieux comprendre la structure linguistique et les règles grammaticales de la langue, et d'identifier les positions appropriées pour les pauses et l'accentuation, ce qui améliore encore la naturalité et l'expressivité des fichiers audio synthétisés.



Dans le traitement émotionnel, nous pouvons non seulement ajuster le pitch (F0) et les pauses en fonction du poids spécifique des mots, mais également ajuster l'ensemble de la voix en fonction de l'intensité et du type de l'émotion. Par exemple, lors de l'expression d'une émotion intense, il est possible d'augmenter le volume et l'amplitude du pitch pour renforcer l'expression émotionnelle. En revanche, lorsqu'une émotion est plus calme ou neutre, il est possible de réduire l'amplitude de ces paramètres pour maintenir la stabilité et l'équilibre de la voix.

Lors de l'ajustement du pitch et du ton, nous devons prendre en compte les caractéristiques vocales associées à différents états émotionnels. Par exemple, lors de l'expression d'une émotion excitée ou joyeuse, la voix peut être plus brillante et plus aiguë, avec une amplitude de pitch plus prononcée ; tandis que lors de l'expression de sentiments de tristesse ou de désespoir, la voix peut être plus sombre et plus grave, avec une amplitude de pitch réduite. De plus, l'ajustement de la vitesse et du rythme est également un moyen important d'exprimer les émotions. Dans des états émotionnels excités ou anxieux, la vitesse peut être accrue, avec un rythme plus serré ; tandis que dans des états émotionnels de réflexion ou de calme, la vitesse peut être ralentie, avec un rythme plus stable.

De plus, les variations de volume sont également un moyen important d'exprimer les émotions. Lorsque l'émotion est intense, la voix peut être plus forte et plus puissante, attirant ainsi l'attention et exprimant l'intensité émotionnelle ; alors que lorsqu'elle est plus douce ou plus retenue, la voix peut être plus douce et plus douce, maintenant ainsi l'équilibre émotionnel interne. De plus, l'ajustement de la qualité vocale peut également exprimer les émotions. Par exemple, lorsqu'on exprime la colère ou la nervosité, la voix peut être plus aiguë et plus rugueuse, exprimant ainsi l'impulsion émotionnelle et l'agitation ; alors que dans des états émotionnels calmes ou détendus, la voix peut être plus douce et plus claire, exprimant ainsi la tranquillité et le confort émotionnels.

Emofilt, en tant que système de synthèse vocale basé sur les émotions, peut exprimer de manière plus précise des états émotionnels spécifiques en utilisant un modèle émotionnel, un modèle acoustique, et des techniques de traitement vocal basées sur les émotions. En ajustant automatiquement les caractéristiques vocales de la synthèse vocale en fonction des étiquettes émotionnelles fournies par l'utilisateur, le système Emofilt peut générer des voix avec des nuances émotionnelles, rendant ainsi la synthèse vocale plus expressive et contagieuse. De cette manière, la synthèse vocale peut transmettre avec précision des états émotionnels spécifiques, renforçant ainsi son efficacité de communication et sa capacité d'expression émotionnelle.

Dans le terminal, allez dans le dossier où se trouve Emofilt, puis saisissez la commande suivante : *java -jar emofilt.jar -useGui -cf emofiltConfig.ini -if example.pho -voc fr1*. Une interface s'ouvrira alors où vous pourrez choisir l'émotion que vous souhaitez générer.

J'ai ajouté de la joie au troisième tour de parole, de la colère au quatrième et de la tristesse au cinquième.

Dans les fichiers audio où les émotions ont été ajoutées, les distorsions et les déformations sonores deviennent très prononcées, ce qui peut être dû à un ajustement insuffisamment précis des paramètres sonores lors du traitement émotionnel. Bien qu'on puisse distinguer les émotions correspondantes, les distorsions et les déformations altèrent l'efficacité globale de la synthèse, rendant le résultat non naturel. Les variations de vitesse, de tonalité, de qualité de la voix, etc., influent sur la cohérence et la naturalité de la parole. Par exemple, dans les fichiers audio associés à la joie et à la colère, la vitesse de la parole augmente clairement, tandis que dans ceux associés à la tristesse, elle diminue. Cette situation peut entraîner une déformation et une incohérence du son, réduisant ainsi la qualité des fichiers audio synthétisés.

Pour résoudre ce problème, il est possible d'essayer d'ajuster l'algorithme de traitement émotionnel pour qu'il identifie et ajuste les paramètres sonores de manière plus précise. De plus, l'utilisation de techniques de traitement émotionnel plus avancées, telles que les modèles d'apprentissage profond, peut améliorer la précision et l'efficacité du traitement émotionnel. Enfin, il est également possible d'effectuer des ajustements plus fins des paramètres sonores pour garantir la qualité et la naturalité des fichiers audio synthétisés. En optimisant continuellement l'algorithme de traitement émotionnel et les ajustements des paramètres sonores, il est possible d'améliorer la qualité et l'expression émotionnelle des fichiers audio synthétisés, répondant ainsi mieux aux besoins et aux attentes des utilisateurs.

Evaluation des fichiers synthèses

Après ces opérations, j’obtiens actuellement 25 fichiers audio, et je souhaite les classifier à l'aide d'un modèle pour évaluer si ces fichiers audio peuvent être jugés par une machine.

L'outil utilisé dans cette étape est Audeep. Audeep est un modèle de traitement de la parole basé sur les techniques de deep learning, conçu pour des tâches telles que la reconnaissance vocale et l'analyse des émotions. Ce modèle comprend généralement plusieurs couches de réseaux neuronaux, notamment des réseaux de neurones convolutionnels (CNN), des réseaux de neurones récurrents (RNN) et des réseaux de neurones à mémoire à court et long terme (LSTM). Ces réseaux sont capables d'apprendre automatiquement et d'extraire des caractéristiques des données vocales, puis d'utiliser ces caractéristiques pour la reconnaissance vocale et l'analyse des émotions.

Voici les instructions pour installer Audeep :
dnf install python37 virtualenv
virtualenv -p python3.7 audeep_virtualenv
source audeep_virtualenv/bin/activate
pip install --upgrade pip
pip install auDeep-master/
pip install 'protobuf<=3.20.1' --force-reinstall

Après installation, renommer ces 25 fichiers au format test_111.wav et les placer dans le dossier "wav" :
test_220.wav=espeak2
.
.
test_223.wav=espeak5
test_224.wav=OT1
.
.
test_228.wav=OT5
test_229.wav=T1_
.
.
test_233.wav=T5_
test_234.wav=T1
.
.
test_239.wav=T5
test_239.wav=T3Joie
.
test_241.wav=T5Sad

Ensuite, ajoutez les noms de ces fichiers ainsi que leur étiquette "?" dans lab/labels.csv.

Supprimer les caractéristiques du dossier features dans le grand dossier, et supprimer l'espace de travail du dossier workspace dans le dossier de base. Notez qu'il est nécessaire d'utiliser la commande ctrl+H sous Linux pour supprimer les fichiers cachés supplémentaires dans le dossier wav, sinon des erreurs pourraient survenir. Accédez au dossier de base dans le terminal, puis exécutez le fichier sh en utilisant la commande ./audeep-generate.sh. Une fois le fichier sh terminé, exécutez baseline.py (python3 baseline.py).

Une fois le programme terminé, nous obtiendrons un fichier csv contenant les résultats. Pour l'ensemble de mes tests audio, tous les résultats sont classés comme A, c'est-à-dire agréables.

Conclusion

Dans ce projet, l'objectif est d'améliorer la technologie de synthèse vocale pour produire des sorties vocales plus naturelles et émotionnellement expressives. J'ai utilisé une variété d'outils et de techniques, y compris Praat, Mbrola, Espeak et Emofilt, pour explorer comment générer des sorties vocales adaptées en fonction du contenu textuel et des besoins émotionnels.

Tout d'abord, j'ai analysé et traité les données vocales brutes en utilisant les outils Praat et Mbrola. J'ai extrait des caractéristiques acoustiques telles que la durée des phonèmes et la fréquence fondamentale à l'aide de Praat, puis j'ai utilisé Mbrola pour la synthèse vocale basée sur des règles. Ces outils nous ont fourni des fonctionnalités de synthèse vocale de base, mais présentaient des limitations en termes d'expression émotionnelle.

Pour améliorer la capacité d'expression émotionnelle de la synthèse vocale, j'ai introduit la technologie de traitement émotionnel Emofilt. Emofilt est un système de synthèse vocale basé sur l'émotion qui peut ajuster automatiquement les caractéristiques vocales de la synthèse en fonction du texte d'entrée et des étiquettes émotionnelles. En ajustant des paramètres tels que le ton, la vitesse, le volume et la qualité vocale, le système Emofilt peut transmettre plus précisément des états émotionnels spécifiques, rendant la synthèse vocale plus expressive et immersive.

Au cours de l'expérimentation, j'ai constaté que l'impact du traitement émotionnel sur la synthèse vocale était significatif, mais il y avait aussi des défis. Les audios avec traitement émotionnel pouvaient présenter des distorsions et des déformations sonores, ce qui affectait la qualité globale de la synthèse. De plus, les ajustements des paramètres vocaux dans le processus de traitement émotionnel nécessitaient une finesse accrue pour améliorer la qualité et la naturalité de l'audio synthétisé.

En résumé, ce projet m'a permis de mieux comprendre la technologie de synthèse vocale et d'explorer l'application du traitement émotionnel dans ce domaine. En combinant judicieusement différents outils et techniques, nous pouvons générer des sorties vocales plus expressives et naturelles, ce qui constitue une référence importante pour le développement et l'application ultérieure de la technologie de synthèse vocale. De plus, l'installation et l'utilisation de ces différents outils m'ont confronté à divers défis, me permettant de mieux comprendre les processus et les limitations opérationnelles sur macOS, Windows et Linux.

Références

https://www.fon.hum.uva.nl/praat/

https://github.com/numediart/MBROLA

https://espeak.sourceforge.net

https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic

http://emofilt.syntheticspeech.de

https://github.com/auDeep/auDeep