UNIVERSITÉ LAVAL

Faculté des sciences et de génie

PROJET DE SYNTHESE VOCALE

ISABELLE EYSSERIC

PROJET SOLAIRE POUR LE COURS DE TRAITEMENT AUTOMATIQUE DU SIGNAL

Clonage de la voix

- État de l'art
- Collecte des données
- Prétraitement des données
- Formation du modèle
- Évaluation & Optimisation
- Tâches & Améliorations

État de l'art

Parole -> Texte

ASR & MS Tacotron2
(Speech-to-Text STT)

Texte -> Mel Spectrogrammes

Tacotron2 & MS Tacotron2
(Text-to-Mel)

Mel Spectrogrammes -> Parole

Vocoder: WaveGlow & HIFI-GAN
(Mel-to-Speech)

Modèle de
Reconnaissance Vocale(ASR)

Modèle de
Synthèse vocale (TTS)

État de l'art

Parole -> **Texte**

Google Speech Recognition

Encodeur + Décodeur
(Séquence à Séquence)

the little bunny hopped through the forest looking for carrots

Architecture des classiques ASR:

1. Analyse acoustique.
2. Association de fréquences à des mots.
3. Analyse de la parole avec 3 modèles (langue, prononciation & acoustique-phonetique)

Architecture des nouveaux ASR:

1. Encodeur: Signal audio -> Vecteurs
2. Décodeur: Vecteurs -> Texte

Lien de l'API Google Web Speech: https://www.google.com/intl/en/chrome/demos/speech.html

État de l'art

Texte -> **Mel Spectrogramme**

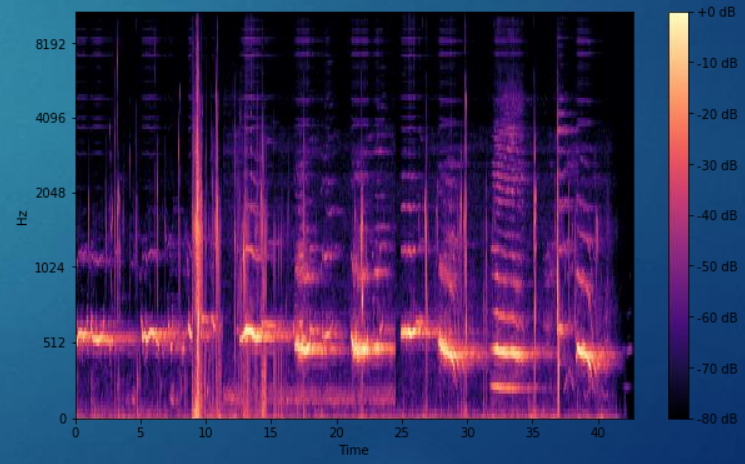
Tacotron2, FastSpeech2, Transformers TTS, …

Encodeur + Attention + Décodeur

Architecture du modèle Tacotron2:

1. **Encodeurr**: Texte -> Vecteurs
2. **Attention**: Calcule le contexte vectoriel
3. **Décodeur**: Génère le Mel Spectrogramme

the little bunny hopped through the forest looking for carrots



Lien de l'API Google Web Speech: https://www.google.com/intl/en/chrome/demos/speech.html

État de l'art

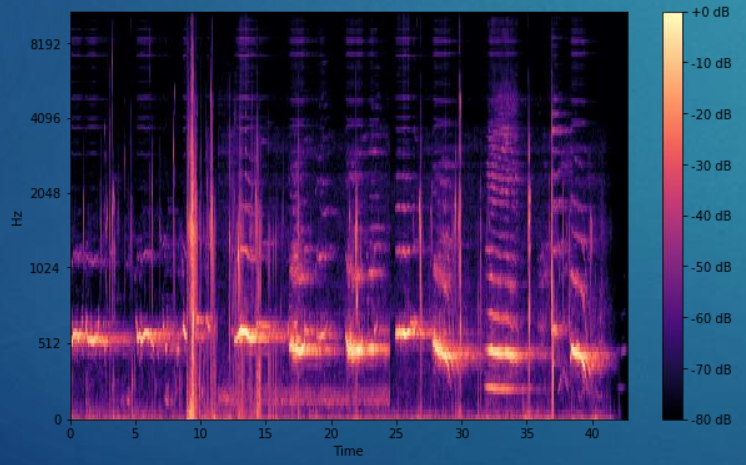
Mel Spectrogramme -> Parole

HIFI-GAN, WaveNet, WaveGlow, …

Architecture du vocodeur HIFI-GAN:

1. **Générateur**: Mel -> Signaux audio
2. **Discriminateurs**: Évaluation de la qualité de l'audio généré

Générateur + Discriminateurs



Article **Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions**: https://arxiv.org/pdf/1712.05884v2.pdf

 Collecte de données

À partir d'un micro et de fichiers audios

Attention: Tous les phonèmes de la langue

Attention: Bonne prononciation et fluidité

Collecte des données

Voyelles:
- **Courtes** : ɪ, ɛ, œ, ʌ, ɒ, ʊ, ɒ
- **Longues** : iː, ɑː, ɔː, uː
- **Diphtongues** : ɑɪ, ɑʊ, eɪ, əʊ, oʊ, ɔɪ, ɪər, eər, ʊər
- **Voyelles Centrales** : ə, ɜːr

Consonnes:
- **Plosives** : p, b, t, d, k, g
- **Fricatives** : f, v, θ, ð, s, z, ʃ, ʒ, h
- **Affriquées** : tʃ, dʒ
- **Nasales** : m, n, ŋ
- **Liquides** : r, l
- **Semi-voyelles** : j, w

"The little bunny hopped through the forest looking for carrots"

{DH AH0} {L IH1 T AH0 L} {B AH1 N IY0} {HH AA1 P T} {TH R UW1} {DH AH0} {F AO1 R AH0 S T} {L UH1 K IH0 NG} {F AO1 R} {K AE1 R AH0 T S}.

Liste des phonèmes anglais

Phrase en anglais

Phonèmes correspondants en anglais

Collecte des données

"In autumn, the leaves turn orange, red, and yellow."

"The princess and the prince lived happily ever after."

"On a sunny day, the flowers in the garden dance in the breeze."

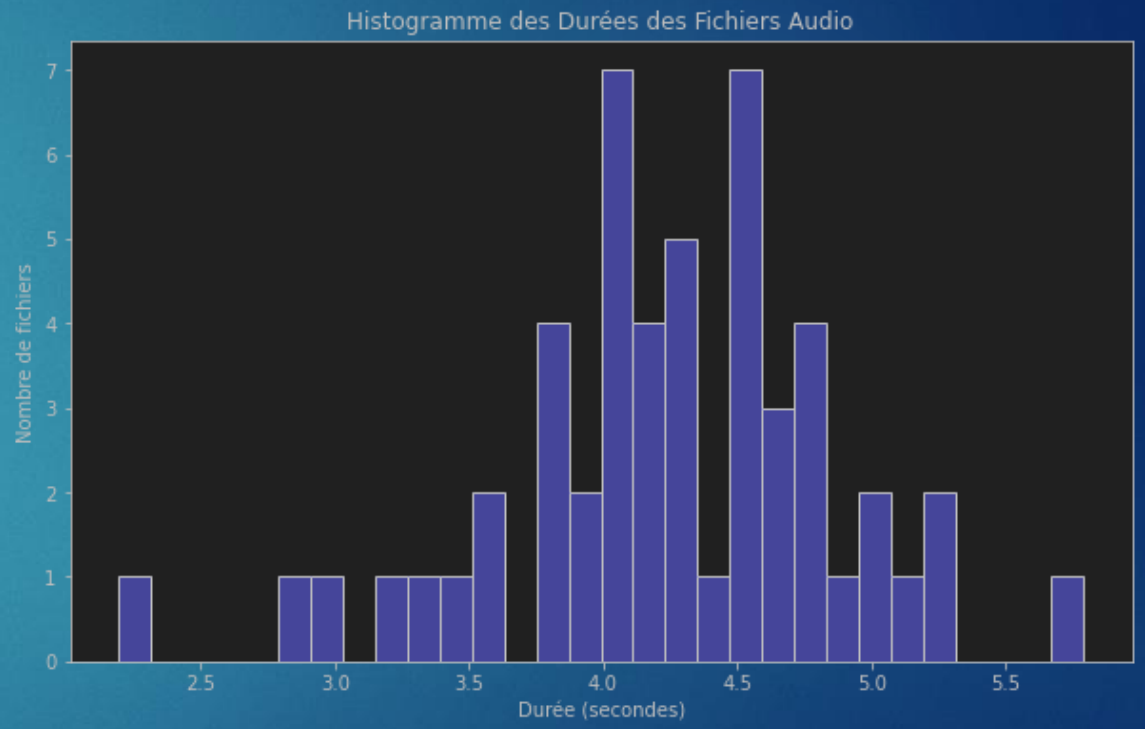
Script pour enregistrer les fichiers audios

File: 24.wav
Sample Rate: 44100
Duration: **2.19** seconds
Total frames: 96682

File: 23.wav
Sample Rate: 44100
Duration: **4.51** seconds
Total frames: 199103

File: 16.wav
Sample Rate: 44100
Duration: **5.79** seconds
Total frames: 255339

Informations sur les fichiers audios



Histogramme sur la longueur des fichiers audios

Prétraitement des données

Segmentation & Découpage:
Audio en Phrases

Renommer les fichiers
Audio en Phrases

Formater les données:
Taux échantillonnage approprié, Encodage et Canaux

Nettoyer & Normaliser:
Couper les blancs, normaliser le volume

Transcription textuelle:
Reconnaissance vocale

Prétraitement du texte:
Vérification de la transcription & Ajustements

Extraction des caractéristiques:
Phonèmes

Alignement:
Dictionnaire de paires Audio - Texte

Prétraitement des données

Formatage, Nettoyage & Normalisation:
Taux échantillonnage,
Encodage, Canaux, Blancs…

Validation de l'audio

Taux d'échantillonnage:
44 100 Hz -> 16 000 Hz

Encodage:
32 bits -> 16 bits

Canaux:
Stéréo -> mono

Audio match avec le script
baby birds **chirp** happily in their nest

"**The** dragon breathed fire but was friendly to everyone."
Dragon **breathe** fire **what** was friendly to everybody

Prétraitement des données

Transcription textuelle
Reconnaissance vocale

"The little bunny hopped through the forest looking for carrots."
the little bunny **hop** through the forest looking for carrots

"The friendly bear waved hello to all the animals in the wood."
the friendly bear waved hello to all the animals in the **world**

"In **the** autumn, the leaves turn orange, red, and yellow."
in the **items** the leaves turn orange red and yellow

"The busy bees buzzed from flower to flower."
The Busy Bees **buzz** from flower to flower

"The fluffy cloud looked like a bunny in the sky."
The Fluffy cloud **look** like a bunny in the sky

"The fluffy cloud looked like a bunny in the sky."
The Fluffy cloud **look** like a bunny in the sky

Validation de la transcription
Manuellement

"**The** baby birds chirped happily in their nest."
baby birds **chirp** happily in their nest

"The frog leaped high to catch the flying bug."
the Frog **leap hi** to catch the flying bug

"**The** dragon breathed fire but was friendly to everyone."
Dragon **breathe** fire **what** was friendly to everybody

"The rainbow fish had scales in all different colors."
the rainbow fish **at** Scales in all different colors

"The clock struck twelve, and the mice ran up the clock."
the Clock Struck **12** and the mice ran up the clock

"The **superhero** saved the day with courage and kindness."
the **superheroes save** the day with courage and kindness

Formation du modèle

Création du jeu de données:
Paires Audio – Texte - Spectrogramme

Entrainement du modèle pré-entrainé:
Modèle TTS Tacotron2 & HIFI-GAN

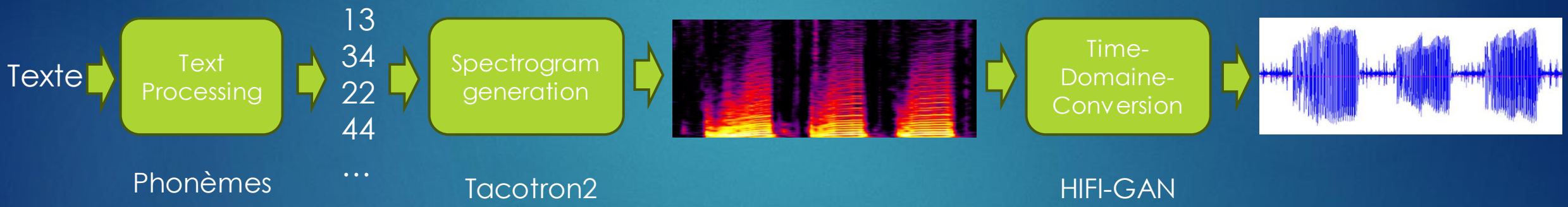
Transformation du jeu de données:
Dataset vers Dataloader

Adaptation du Modèle:
Conditionnement vocal, Embedding

Division du jeu de données:
Entrainement (90%), Test (10%)

Configuration du Modèle:
Hyperparamètres, Taux apprentissage

Formation du modèle



🎯 Évaluation & Optimisation

Ajustement du modèle et de ses paramètres

Taux d'apprentissage, Nombre d'époques, Taille des batchs

Mesure de performance

Mean Squared Error (MSE) & Mel Cepstral Distortion (MCD)

🎯 Évaluation & Optimisation

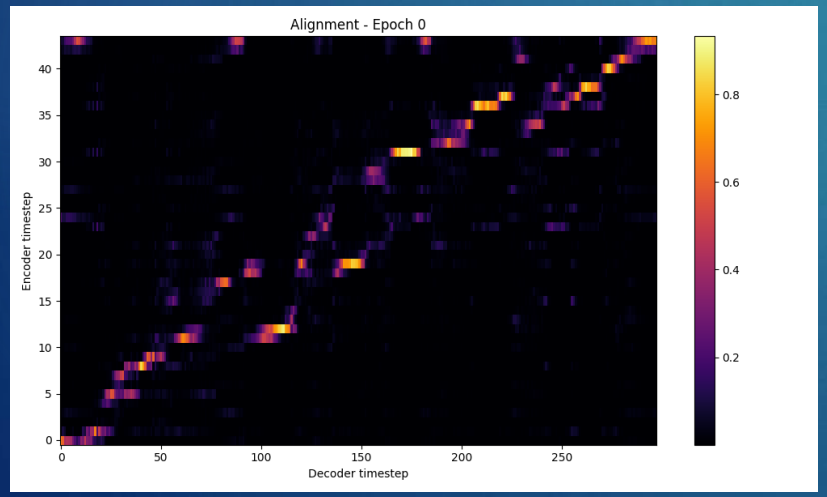
Ajustement du modèle et de ses paramètres
Taux d'apprentissage, Nombre d'époques, Taille des batchs

Périodes de Plateau

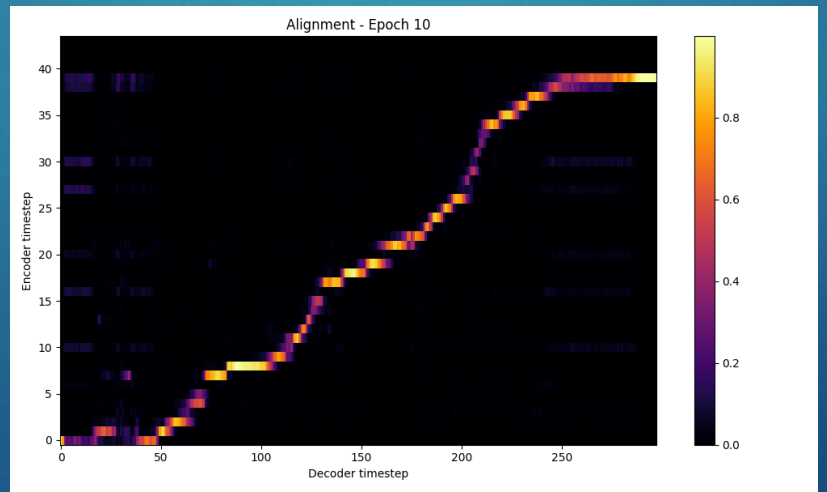
Sous ou Sur ajustement

Learning Rates: 0.00003 **Batch Size:** 6

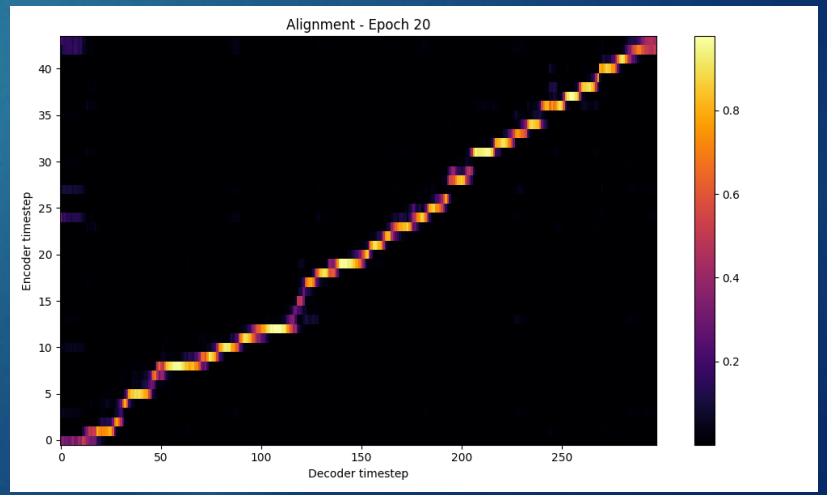
Epoch: 0 loss: 1.77676



Epoch: 10 loss: 0.941476



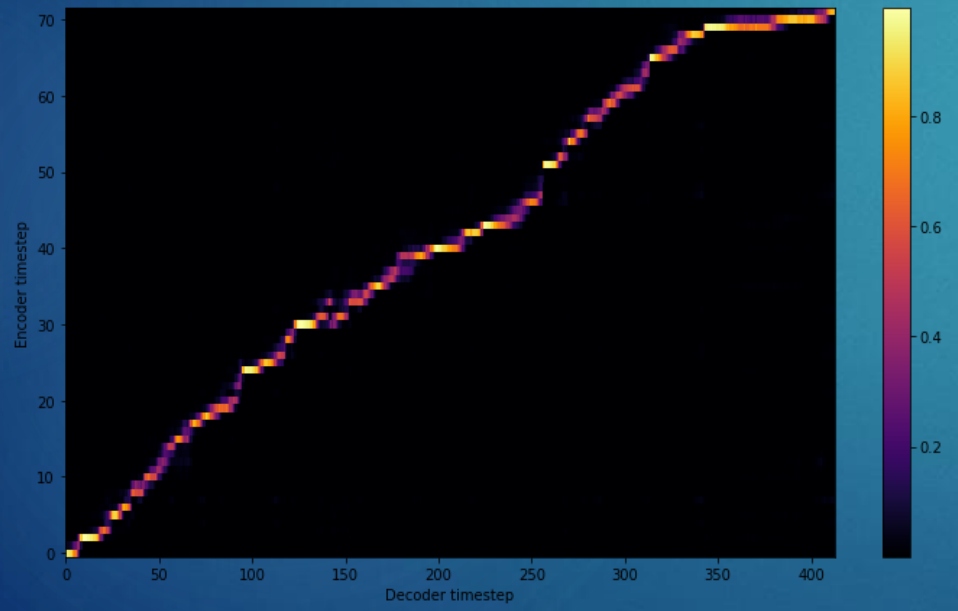
Epoch: 20 loss: 1.02405



🎯 Évaluation & Optimisation

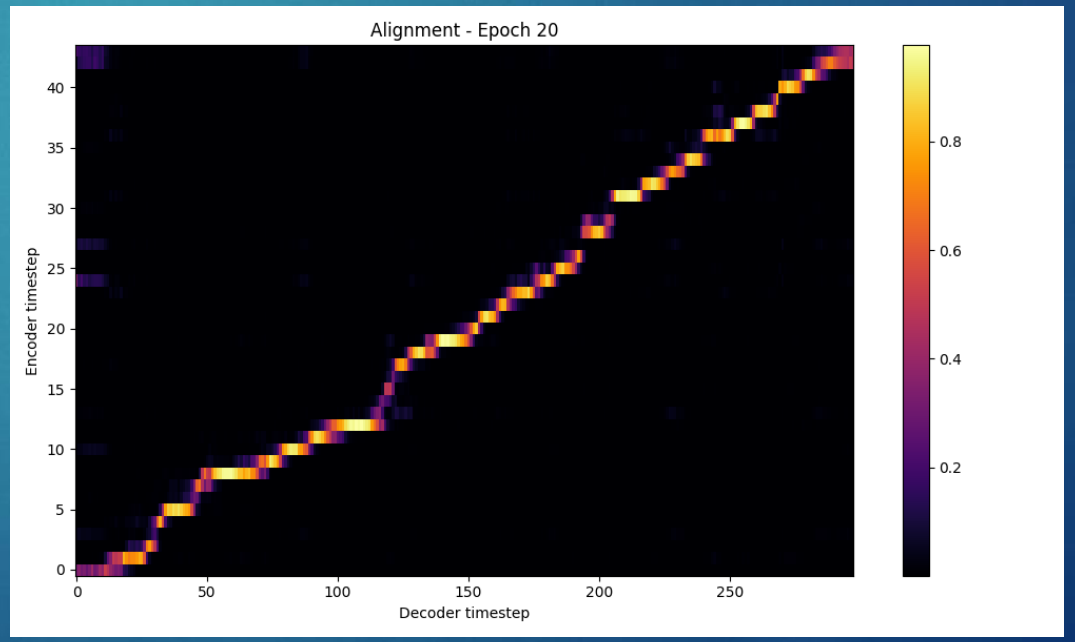
Mesure de performance

Mean Squared Error (MSE)



Courbe d'un bon entrainement

```
Epoch: 20 Iteration: 189 Validation Loss: 1.02405
```



Résultat de mon entrainement

🎯 Tâches & Améliorations

Amélioration de la voix synthétisée

Entrainer le vocodeur HIFI-GAN sur le dataset LJSpeech
et le tester sur mes données

Extension du modèle sur d'autres langues

Intégration du modèle en temps réel

Clonage de la voix

- **Librairies**:
  - SpeechBrain: https://speechbrain.github.io/
  - Pytorch: https://pytorch.org/
  - Modèles pré-entrainés: https://huggingface.co/models et https://catalog.ngc.nvidia.com/models
- **Applications**:
  - Google Web Speech: https://www.google.com/intl/en/chrome/demos/speech.html
  - TTS Maker: https://ttsmaker.com/
- **Articles**:
  - Article **Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions**: https://arxiv.org/pdf/1712.05884v2.pdf
  - HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesi: https://arxiv.org/abs/2010.05646
  - Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages https://ar5iv.labs.arxiv.org/html/2303.01037

Merci pour votre attention

Isabelle Eysseric
Isabelle.eysseric.1@ulaval.ca