Comparaison et évaluation de vocodeurs pour la synthèse neuronale en Français

Thomas Granjon, Marie Tahon

Comparaison et évaluation de vocodeurs pour la synthèse neuronale en Français

Thomas Granjon ; Marie Tahon

LIUM, Le Mans, France

Contexte & Objectifs

- TTS : entrée graphème ou phonème, sortie signal de parole audio
- On veut construire un synthétiseur neuronal pour du Français
- On veut pouvoir évaluer ce système
- On veut savoir quelles conditions sont requises pour obtenir de bonnes performances

État de l'art

Les systèmes :

- **DeepVoice 3** : un synthétiseur neuronal convolutionnel [Ping et al., 2017, arXiv]
- **Tacotron** : un synthétiseur neuronal récurrent [Wang et al., 2017, arXiv]
- **Merlin** : un synthétiseur neuronal récurrent complètement public [Wu et al., 2016, SSW]

Les vocodeurs :

- **Griffin-Lim** : un algorithme effectuant une transformée de Fourrier inverse [Griffin et Lim, 1984, IEEE]
- **WORLD** : trois algorithmes utilisant la F0, l'enveloppe spectrale et le paramètre apériodique [Morise et al., 2016, IEICE]
- **WaveNet** : un réseau de neurones convolutionel qui transforme un spectrograme en forme d'onde [Van Den Oord et al., 2016, SSW]

Problématiques :

- DeepVoice 3 est-il un bon système pour synthétiser du Français ?
- Quelle quantité de données est requise pour obtenir de bonnes performances ?
- Quelle entrée pour la TTS, graphèmes ou phonèmes ?

Protocole

Données : corpus SynPaFlex [Sini et al., 2018, LREC]

- 70 h de livres audio lus par une seule locutrice
- Alignement entre le texte, les phonèmes et l'audio
- Extraits < 10 secondes → 28 000 phrases

Système TTS : DeepVoice 3

- Adaptation à une voix en Français
- Entrée graphèmes et/ou phonèmes (taux variable)
- Apprentissage du réseau : 10 jours avec un GPU et 80G de mémoire
- Hyperparamètres : 4 000 époques, 16 batches, $F_s = 22\,050$ Hz

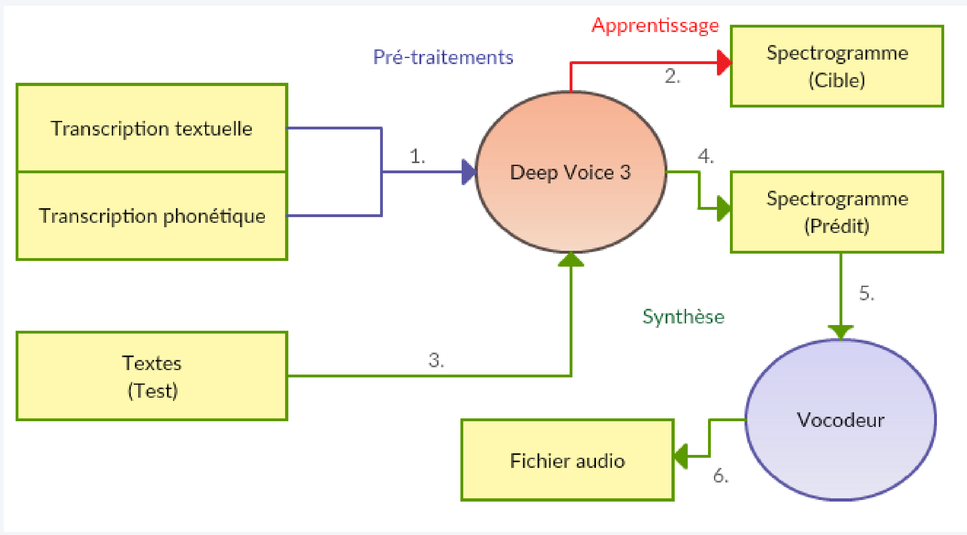


Schéma du fonctionnement de notre système

Différentes expériences

- Évaluer l'importance de la quantité de données apprises sur les performances : 100%, 75% ou 50% des 28 000 phrases
- Évaluer l'influence du taux de graphèmes/phonèmes appris p =0 (graphèmes seuls) ; 0.5 (mélange) ; 1 (phonèmes seuls)
- Evaluer l'influence du vocodeur Griffin-Lim ou WaveNet → seul Griffin-Lim est présenté

Nom	% phrases	p
GL100	100	0.5
GL75	75	0.5
GL50	50	0.5
PhoneGL0	36	0
PhoneGL1	36	1

Les modèles évalués

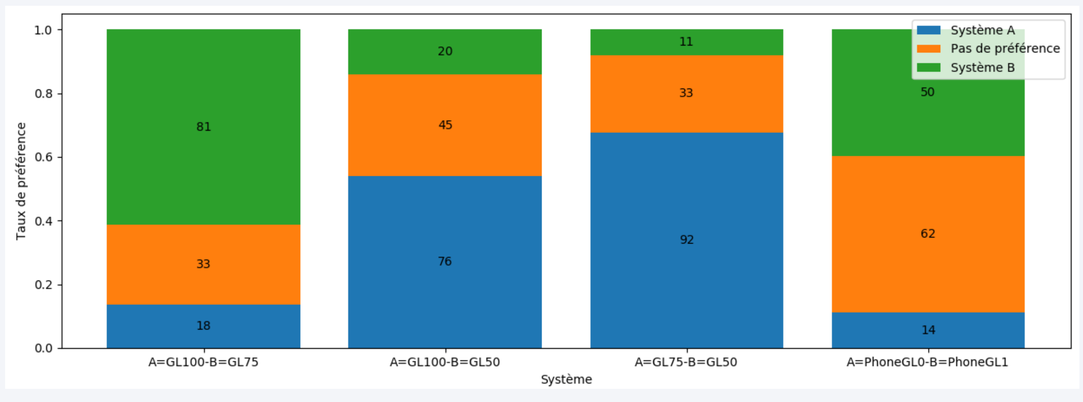
Test perceptif

Évaluation :

- 13 volontaires ont participé
- 40 phrases sélectionnées sur le nombre de caractères
- 2 questions : préférence A/B et transcription

Résultats :

- GL75 > GL100 > GL50
- PhoneGL1 > PhoneGL0



Résultats du test A/B

Retour et perspectives

Conclusion :

- Résultats de mauvaise qualité, difficile à évaluer
- Contrairement à ce qui était prévu, plus de données ne signifie pas plus de qualité
- Comme prévu, la version uniquement phonétique est préférée à la version uniquement textuelle

Pour plus tard :

- Reproduire l'expérience avec un autre système comme Tacotron par exemple
- Obtenir des performances convenables avec WaveNet
- Établir un corpus d'apprentissage plus adapté à notre tâche