CLONAGE DE LA VOIX

Isabelle Eysseric

Université Laval
Département Informatique et génie logiciel,
1045 Av. de la Médecine, Québec, QC G1V 0A6

ABSTRACT

Dans le domaine de la radiodiffusion, la qualité des fichiers audio est très primordiale. Il arrive dans certaines situations que le contenu ou la qualité des enregistrements soient affectés au point d'être inutilisables. Le locuteur ne pouvant pas toujours réenregistrer l'émission, celle-ci se voit être annulée. Un système de synthèse vocale avec conditionnement de la voix pourrait pallier ce problème en recréant la voix de l'animateur à partir de simples transcriptions textuelles. Cette étude se concentre sur le développement d'un système de clonage de voix à partir d'un très petit jeu de données mais complet, couvrant l'intégralité des phonèmes de la langue cible. Le processus compte la préparation des données, son traitement, l'entraînement du modèle et son évaluation. Ce modèle a démontré une efficacité, atteignant une perte moyenne de 1.02 en un laps de temps très court. Cette performance est prometteuse et suggère la possibilité d'amélioration futures dans le domaine de la synthèse vocal sur de petits ensemble de données.

***Index Terms*—** Synthèse vocale, Conditionnement vocale, Spectrogrammes, Vocodeur, Phonèmes et Pangrammes.

1. INTRODUCTION

La synthèse vocale a beaucoup évoluée ces derniers temps avec les grandes avancées dans le domaine de l'intelligence artificielle et le traitement de la langue naturelle. On la retrouve dans plusieurs applications allant de l'assistance vocale, les livres audio ou encore la traduction à voix haute et en temps réelle.

Le conditionnement de la voix, qui permet d'imiter des voix spécifiques, vient augmenter les possibilités dans les secteurs de la communication et du divertissement. Cette technologie offre des solutions pratiques dans des situations où les enregistrements vocaux sont de très mauvaise qualité voir même indisponibles.

Le défi majeur avec la synthèse vocale est l'entraînement de modèles de haute qualité sur des jeux de données limités. Ces systèmes ont besoin de grande quantités de données afin de produire des résultats naturels. La capacité à conditionner efficacement la voix cible reste une tâche complexe.

Dans ce contexte, l'objectif de mon étude est de développer un système de synthèse vocale conditionné qui puisse être entraîné efficacement sur un petit jeu de donné. En utilisant des techniques d'optimisation et de traitement de la langue, je vise à créer un modèle capable de reproduire fidèlement la voix d'un interlocuteur.

Ce système pourrait avoir une utilité dans le secteur de la radiodiffusion par exemple, où la reproduction fidèle de la voix est essentielle.

2. REVUE DE LITTÉRATURE

Les modèles traditionnels de synthèse vocale nécessitent de vastes ensembles de données pour l'entraînement de leurs modèles. Dans mon cas, mon système se distingue par le fait d'utiliser de petits jeux de données.

Ce processus de synthèse se fait en deux étapes principales: la prédiction de spectrogrammes à partir de textes, suivie de la génération d'audio à partir de ces spectrogrammes. Sur la figure 1, nous pouvons observer l'architecture d'un tel système.

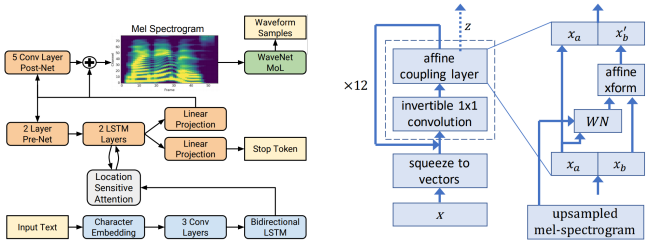
Parmi les modèles à l'état de l'art, Tacotron 2 [1], se distingue par le fait de synthétiser un discours avec un son naturel à partir seulement de transcription textuelles sans aucune information prosodique supplémentaire. À partir du texte d'entrée, il produit des spectrogrammes Mel en utilisant une architecture codeur-décodeur. Comme on peut le voir sur la figure 1, un vocodeur comme WaveNet, HIFI-GAN [2] ou WaveGlow [3] est ensuite appliqué pour générer la parole à partir de la prédiction de ces spectrogrammes.

Pour conditionner la voix d'une personne, mon système se base sur les modèles de Nvidia, Tacotron2 et WaveGlow [4], tous deux entraînées sur JLSpeech, un grand ensemble de données avec un seul locuteur. Contrairement au

Thanks to XYZ agency for funding.

modèle d'origine, l'implémentation de Tacotron 2 avec Pytorch utilise Dropout au lieu de zone pour régulariser les couches LSTM [5].

Outre le choix du modèle pré-entraîné, la collecte de données est cruciale afin d'optimiser l'apprentissage de mon modèle sur un petit ensemble de données contrairement aux systèmes traditionnels.



a) Architecture de Tacotron2 (b) Architecture de WaveGlow

Fig. 1. Architecture des modèles utilisés pour un système de synthèse vocale de bout en bout.

3. MÉTHODOLOGIE

Le processus de synthèse vocale a été réalisé en plusieurs étapes afin de produire un système efficace de synthèse vocale.

Première étape: Préparation et traitement des données. Pour simplifier la tâche, la langue anglaise a été choisi. Une sélection initiale de phrases, incluant des pangrammes phonétique, a été effectuée pour couvrir l'ensemble des phonèmes de l'anglais, essentiels à l'apprentissage du modèle. La collecte de données a permis de rassembler les 59 enregistrements vocaux et leurs transcriptions textuelles correspondantes pour une durée totale de seulement 4 minutes et 6 secondes. Une phase de nettoyage a été appliquée aux enregistrements pour éliminer le bruit de fond, les silences et normaliser le volume audio. Les textes ont été convertis en phonèmes en utilisant le dictionnaire de prononciation CMU et la méthode ARPAbet [6], suivi de la génération des spectrogrammes correspondants. Pour un fichier audio nous avons donc sa transcription textuelle et son spectrogramme correspondant.

Deuxième étape: Modélisation avec Tacotron2. Tacotron2, un modèle de synthèse vocale avancé basé sur un mécanisme d'attention, a été utilisé. L'entraînement du modèle s'est concentré sur la conversion de texte en spectrogrammes correspondants. Des validations périodiques ont été effectués afin de surveiller son apprentissage et d'ajuster les hyperparametres. Le calcul de la perte moyenne et la conversion du spectrogramme généré en signal audio avec le vocodeur ont été cruciaux dans cette tâche.

Troisième étape: Évaluation et optimisation. Cette dernière étape a impliqué une analyse de la qualité de la voix synthétisée en fonction du texte d'entrée et de sa similitude avec la voix cible. Un test final a été réalisé pour confirmer l'atteinte des objectifs du projet. Ce processus a permis d'obtenir une voix synthétique de haute qualité avec la possibilité d'améliorer le naturel de la voix générée par la suite.

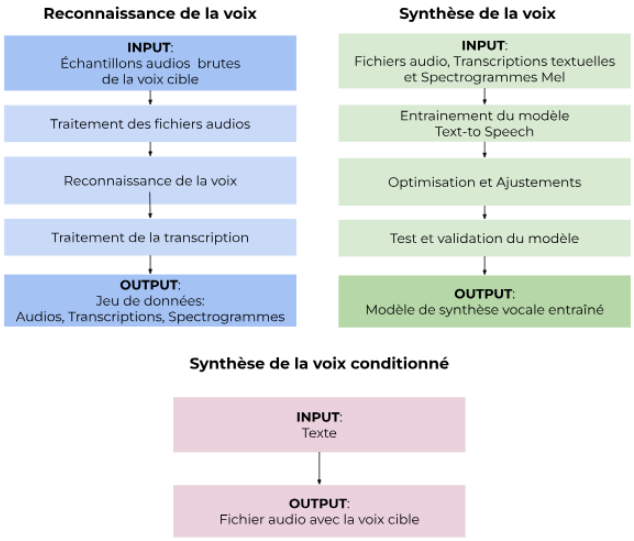


Fig. 2. Processus du projet de synthèse vocale

4. RÉSULTATS

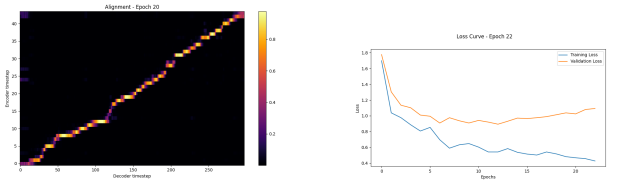
Les résultats du modèle ont permis de montrer l'efficacité de son apprentissage. Comme on peut le voir sur la figure 2, le modèle a démontré une grande capacité à retenir l'information essentielle en seulement 20 epochs. Il a réussi à synthétiser fidèlement la voix cible à partir d'un ensemble de données restreint de seulement 4 minutes d'enregistrements audio. De plus, on peut observer un très bon alignement entre les spectrogrammes et les textes correspondants, ce qui indique un bon apprentissage du modèle.

L'utilisation stratégique de pangrammes phonétiques ainsi que la sélection de modèles pré-entraînés sur un ensemble de données spécifique avec un seul interlocuteur a joué un rôle crucial dans la construction d'un système robuste. Cette approche a permis de donner au système de synthèse vocale, la capacité à reproduire une voix naturelle et précise en très peu de temps.

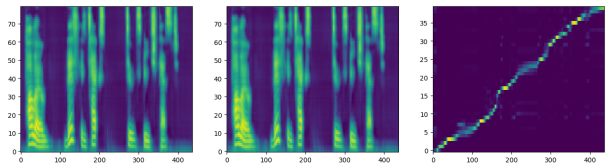
L'utilisation de pangrammes a permis de couvrir l'intégralité

des phonèmes existants de la langue et ainsi permettre un apprentissage complet du modèle. Tandis que l'exploitation des modèles pré-entraînés sur des données cillées à permis d'optimiser le processus d'entraînement du modèle et permettant ainsi de construire un système de synthèse vocale particulièrement efficace.

Ces résultats montrent la possibilité d'améliorations futures dans le domaine de la synthèse vocale. Ce projet apporte une contribution au domaine en explorant de nouvelles approches pour améliorer la fidélité de la voix synthétisée par rapport à la voix d'origine sur des données restreintes.



a) Alignement des spectrogrammes(b) Courbe d'apprentissage



(c) Test du modèle final

Fig. 3. Résultats de l'entraînement et de la validation du modèle

5. CONCLUSION

Les résultats obtenus pour ce projet montre la viabilité du système ainsi que les applications pratiques potentielles, en particulier dans le secteur de la radiodiffusion. La particularité du système de produire une voix synthétisée de qualité à partir d'un très petit jeu de données en un laps de temps très court est très prometteur.

Cependant, il reste encore quelques défis à relevé comme l'amélioration du naturel de la voix synthétisée ainsi que l'extension de cette technique avec d'autres langues moins communes comme le français ou encore l'intégration de ce système dans des applications en temps réelles.

En conclusion, ce projet à permis de jeter les bases pour de futures recherches dans le domaine de la synthèse vocale avec conditionnement et sur des ensembles de données limités.

6. REFERENCES

[1] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017.

[2] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *CoRR*, vol. abs/2010.05646, 2020.

[3] Rafael Valle Ryan Prenger and Bryan Catanzaro, "Waveglow - a flow-based generative network for speech synthesis," *CoRR*, vol. abs/1811.00002, 2018.

[4] "Nvidia/waveglow: A flow-based generative network for speech synthesis," .

[5] "Tacotron 2, the tacotron 2 model for generating mel spectrograms from text," .

[6] "The cmu pronouncing dictionary," .