Synthèse Vocale à partir de Coefficients MFCC à l'aide d'un Réseau de Neurones Récurrents

Résumé : Cette recherche a eu pour objectif de développer un modèle de synthèse vocale capable de générer des séquences audio à partir de coefficients de filtrage cepstraux en fréquence (MFCC). En utilisant un réseau de neurones récurrents (RNN), nous avons exploré la capacité du modèle à apprendre des représentations temporales complexes de la voix humaine et à générer de nouvelles séquences audio en se basant sur ces représentations.

Introduction : La synthèse vocale est un domaine de l'intelligence artificielle qui vise à imiter la voix humaine par des moyens computationnels. Les MFCCs sont des caractéristiques acoustiques couramment utilisées pour représenter le contenu spectral de la voix dans les systèmes de reconnaissance vocale. Cependant, leur utilisation pour la génération de voix est moins explorée. Cette recherche propose d'utiliser un modèle RNN pour convertir des séquences de MFCCs en séquences audio synthétiques.

Méthodologie :

1. Chargement et Prétraitement des Données : Nous avons utilisé un fichier audio d'exemple, Sample_Test_Voix1.wav, pour extraire les MFCCs à l'aide de la fonction extract_mfcc. Les MFCCs représentent les caractéristiques spectrales de l'audio, qui sont ensuite organisées en séquences pour l'entraînement du modèle. Les séquences ont une longueur fixe de 30 étapes, chaque étape contenant un vecteur de 13 dimensions correspondant aux coefficients MFCCs.
2. Préparation des Séquences : Les séquences de données ont été préparées pour l'entraînement en utilisant la fonction prepare_sequences. Chaque séquence d'entrée (X) est constituée de 30 étapes de MFCCs, tandis que la sortie correspond à la valeur de MFCCs à l'étape suivante. Cette préparation permet au modèle de prédire les valeurs futures en se basant sur les valeurs passées.
3. Création du Modèle : Un modèle RNN a été créé avec deux couches SimpleRNN, chacune contenant 64 unités, et une couche de sortie Dense dont la taille est égale au nombre de coefficients MFCCs (13). Ce modèle est entraîné pour minimiser l'erreur quadratique moyenne entre les séquences générées et les séquences cibles.
4. Entraînement et Validation : Le modèle a été entraîné sur 50 époques avec une taille de lot de 32, en utilisant 80 % des données pour l'entraînement et 20 % pour la validation. Les performances du modèle ont été évaluées sur la base de la perte de validation.
5. Génération de Voix : Après l'entraînement, le modèle a été utilisé pour générer une séquence de MFCCs à partir d'une séquence d'entrée. La fonction generate_voice a été employée pour produire une séquence synthétique en prédisant les valeurs de MFCCs à chaque étape et en mettant à jour la séquence courante avec ces prédictions.

6. Reconstruction de l'Audio : Les MFCCs générés ont été reconvertis en spectrogramme puis en signal audio à l'aide de la fonction mfcc_to_audio. Le fichier audio résultant a été sauvegardé sous le nom generated_voice.wav.

Résultats et Discussion : Le modèle a réussi à générer des séquences de MFCCs qui, après reconstruction, ont produit un fichier audio. Toutefois, des erreurs persistent, notamment dans la taille des prédictions par rapport à la taille des séquences d'entrée. Cette divergence souligne la nécessité d'ajuster les dimensions des vecteurs de sortie du modèle ou de revoir la préparation des séquences.

Conclusion : Cette étude démontre la faisabilité de l'utilisation des RNN pour la synthèse vocale à partir de MFCCs. Bien que des erreurs subsistent, ces résultats ouvrent la voie à des améliorations et à des explorations futures dans la génération de voix synthétiques. Les prochaines étapes incluront des ajustements du modèle et des méthodes de reconstruction audio pour améliorer la qualité des voix générées.

Références :

- Librosa: Documentation de Librosa
- TensorFlow: Documentation de TensorFlow