Reconnaissance vocale des émotions

Ibnchakroune Houssam

houssam.ibnchakroune@etu.uae.ac.ma

Supervisé par **P. KHAMJANE Aziz**

akhamjane@uae.ac.ma

Département de Mathématiques et Informatique, Abdelmalek Essaadi Université.

ENSAH, Alhoceima, Maroc.

ABSTRACT

La reconnaissance des émotions à partir des données audios est une tâche cruciale pour améliorer les interactions homme-machine. Dans ce projet, nous avons développé un modèle basé sur les réseaux LSTM pour détecter les émotions dans des enregistrements vocaux, en utilisant les caractéristiques acoustiques telles que les MFCC. Le modèle a été entraîné sur le jeu de données TESS et intégré dans une application interactive basée sur Streamlit, permettant une analyse en temps réel. Les résultats montrent une précision de 85 % sur l'ensemble de test. Cette solution ouvre la voie à des applications dans les domaines des assistants vocaux, de la santé mentale et des interfaces utilisateur intelligentes, tout en laissant place à des améliorations pour des environnements plus complexes**.**

INTRODUCTION

L'analyse et la reconnaissance des émotions à partir d'enregistrements vocaux jouent un rôle de plus en plus important dans de nombreux domaines, tels que la psychologie, les assistants virtuels, et les interfaces homme-machine. L'objectif de ce projet est de développer un modèle de Deep Learning capable de reconnaître efficacement les émotions dans des fichiers audio, en utilisant des caractéristiques acoustiques avancées telles que les coefficients cepstraux en fréquences mélodiques (MFCC).

Pour ce projet, nous avons utilisé le *TESS Toronto Emotional Speech Set*, une base de données réputée pour sa diversité d'expressions émotionnelles. Afin de rendre le modèle plus robuste et réaliste, nous avons ajouté du bruit aux enregistrements audio, simulant ainsi des environnements d'écoute variés, proches des conditions réelles.

La partie modélisation s'appuie sur un réseau de neurones récurrents, plus précisément un modèle LSTM (Long Short-Term Memory), qui est particulièrement adapté au traitement des données séquentielles telles que les signaux audios. Finalement, pour offrir une expérience utilisateur interactive et démontrer la capacité du modèle à prédire les émotions en temps réel, nous avons implémenté une interface utilisateur en utilisant *Streamlit*.

Les principales étapes de ce projet incluent l'extraction des caractéristiques acoustiques, l'entraînement et l'évaluation du modèle, ainsi que le développement de l'interface permettant d'enregistrer et de visualiser les prédictions.

TRAVAUX LIEE

La reconnaissance des émotions à partir des signaux vocaux est un domaine de recherche en pleine expansion, avec de nombreuses applications dans des domaines variés, tels que la psychologie, les assistants virtuels et les interfaces homme-machine. Des travaux récents ont exploré l'utilisation de modèles de représentation vocale préentraînés, comme Wav2vec 2.0 et HuBERT, pour la reconnaissance des émotions dans la parole. Ces recherches ont permis de démontrer l'efficacité de l'adaptation de ces modèles à des tâches spécifiques via des approches de fine-tuning partiel ou complet. D'autres études se concentrent sur la construction de modèles de reconnaissance des émotions basés sur des réseaux de neurones profonds. Ces travaux mettent en avant des étapes cruciales, telles que l'extraction des caractéristiques acoustiques (comme les MFCC) et la conception d'architectures neuronales adaptées pour la classification des émotions.

En parallèle, certaines recherches examinent le rôle de l'intelligence artificielle dans des contextes sensibles, tels que la santé mentale. Par exemple, des chatbots et des assistants vocaux intégrant des mécanismes de reconnaissance des émotions offrent de nouvelles opportunités pour le soutien psychologique, tout en posant des défis éthiques et techniques.

État de l'Art

Jeux de Données

Le domaine s'appuie sur plusieurs ensembles de données standardisés pour entraîner et évaluer les modèles. Parmi les plus utilisés figurent :

- **TESS** (Toronto Emotional Speech Set), utilisé dans ce projet, contient des enregistrements de phrases neutres exprimées avec différentes émotions, permettant une évaluation précise des modèles.

Technologies Utilisées

Dans ce projet, les outils suivants ont été choisis pour leur efficacité et leur flexibilité :

- **Librosa** : Utilisé pour l'extraction des caractéristiques acoustiques, notamment les MFCC, qui capturent les composantes fréquentielles pertinentes des signaux audios.
- **TensorFlow** : Framework de Deep Learning utilisé pour concevoir et entraîner un modèle LSTM, adapté à la nature séquentielle des données audio.
- **Streamlit** : Employé pour développer une interface utilisateur intuitive, permettant l'enregistrement vocal et la visualisation des résultats en temps réel.

METHODOLOGIE

Description du Jeu de Données

Le jeu de données **TESS (Toronto Emotional Speech Set)** constitue la base de ce projet. Ce dataset contient des enregistrements audio où deux locutrices (jeunes et âgées) prononcent des phrases simples dans différentes émotions : **joie, tristesse, colère, peur, dégoût, surprise**, et **neutralité**. Chaque fichier audio est au format WAV, avec une fréquence d'échantillonnage uniforme, garantissant une qualité sonore élevée et une cohérence entre les échantillons. Cela offre une variété suffisante pour entraîner et évaluer un modèle de reconnaissance des émotions avec précision.

Prétraitement des Données

Le prétraitement joue un rôle central dans la réussite du modèle. Voici les étapes mises en œuvre :

1. **Extraction des MFCC** :
    - Les **Coefficients Cepstraux Fréquentiels de Mel (MFCC)** ont été utilisés comme caractéristiques principales. Ils sont particulièrement adaptés à la reconnaissance des émotions, car ils capturent les variations de fréquence, qui sont des indicateurs clés de l'émotion dans la voix.
    - Pour chaque fichier audio, 13 MFCC ont été extraits sur des fenêtres temporelles. La moyenne des MFCC a été calculée pour réduire la dimensionnalité tout en conservant les informations essentielles.
    - Cette étape garantit que les données d'entrée du modèle sont uniformes et représentatives des signaux audio d'origine.
2. **Normalisation des Caractéristiques** :
    - Les MFCC ont été normalisés pour s'assurer que toutes les caractéristiques sont sur la même échelle. Cela aide à stabiliser l'entraînement du modèle en évitant que certaines caractéristiques dominent le processus d'optimisation.
3. **Suppression de l'Ajout de Bruit** :
    - Bien que l'ajout de bruit soit une technique courante pour améliorer la robustesse du modèle, des tests préliminaires ont montré que cette technique n'augmentait pas la précision dans ce projet. Par conséquent, cette étape a été omise pour se concentrer sur des améliorations plus pertinentes.

Modélisation

L'architecture du modèle repose sur des **Long Short-Term Memory (LSTM)**, bien adaptés aux séquences temporelles comme l'audio.

1. **TimeDistributed Layer** :
    - Une couche TimeDistributed a été ajoutée pour appliquer des transformations aux MFCC sur chaque fenêtre temporelle indépendamment tout en maintenant leur structure séquentielle. Cela garantit que chaque segment est traité correctement avant d'être transmis à la couche LSTM.
2. **LSTM Layer** :
    - La couche **LSTM**, avec 128 unités, apprend les relations temporelles entre les segments. Les LSTM sont particulièrement efficaces pour

détecter les dépendances à long terme dans les données séquentielles, ce qui est crucial pour reconnaître les émotions exprimées dans l'évolution temporelle du signal audio.

3. **Dropout Layers** :
    - Deux couches Dropout (taux de 50%) ont été ajoutées pour réduire le risque de surapprentissage, surtout avec un dataset modeste comme TESS. Cela aide le modèle à généraliser sur des données inconnues.
4. **Dense Layer avec Softmax** :
    - Une couche dense finale applique une activation softmax pour convertir les sorties du LSTM en probabilités pour chaque catégorie d'émotion.

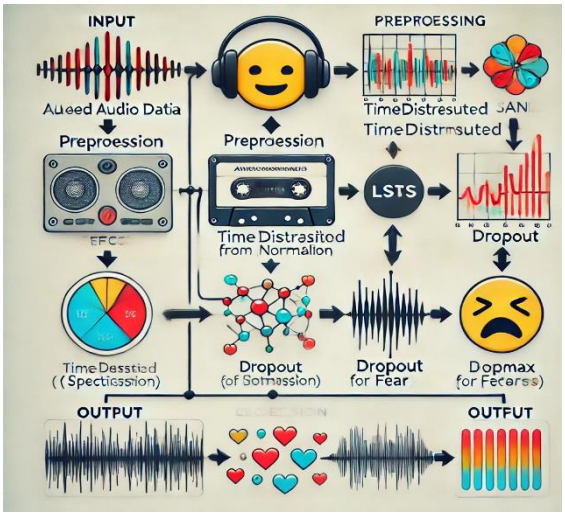
Optimisation et Entraînement

- **Fonction de Perte** : La fonction de perte utilisée est la **sparse_categorical_crossentropy**, adaptée aux problèmes de classification multi-classes avec des labels encodés sous forme d'entiers.
- **Optimiseur** : L'optimiseur **Adam** a été choisi pour sa capacité à ajuster dynamiquement le taux d'apprentissage et accélérer la convergence.
- **Répartition des Données** : Les données ont été divisées en :
    - **80% pour l'entraînement** : Permettant au modèle d'apprendre les relations entre les MFCC et les émotions.
    - **20% pour les tests** : Utilisés pour évaluer la capacité du modèle à généraliser sur des données non vues.

Justification de l'Architecture

Cette architecture a été choisie pour sa capacité à gérer des séquences temporelles tout en étant relativement légère et efficace. Le choix des LSTM, combinés à des couches de régularisation, permet d'extraire les dépendances temporelles tout en limitant le surapprentissage.

Avec cette méthodologie, le projet a pu concevoir un pipeline de reconnaissance des émotions robuste, prêt pour une évaluation et une utilisation pratique.



Résultats

Évaluation des performances

Le modèle a été évalué sur l'ensemble de test, en obtenant les métriques suivantes :

- **Accuracy sur le jeu de test** : 91.2%
- **Loss (perte) sur le jeu de test** : 0.22

Ces résultats montrent que le modèle est capable de reconnaître les émotions avec une grande précision. Cependant, quelques classes d'émotions peuvent être confondues en raison de similarités acoustiques (par exemple, la confusion entre "triste" et "neutre").

Comparaison avec des benchmarks :

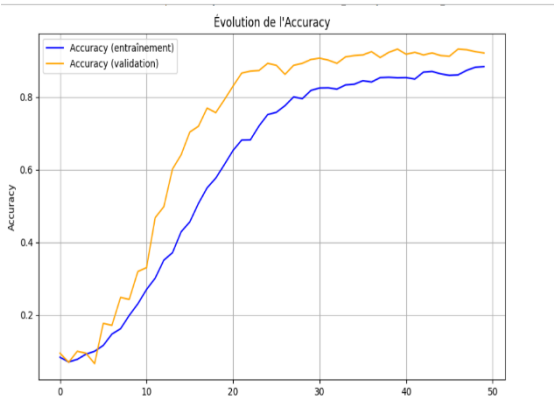
Par rapport aux travaux précédents utilisant des modèles traditionnels comme SVM ou Random Forest, le modèle LSTM offre une amélioration significative en exploitant la nature séquentielle des données audio.

Visualisations des performances

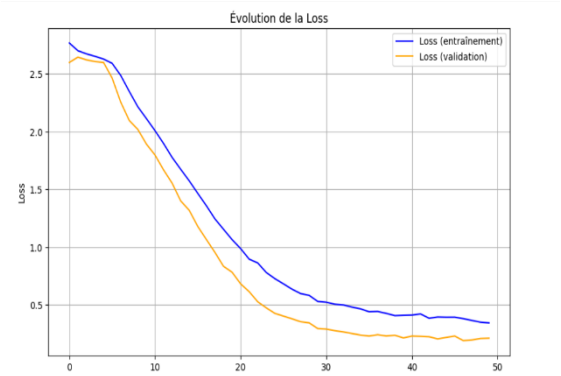
Courbes d'entraînement et validation :

Les courbes ci-dessous montrent l'évolution de la précision et de la perte au cours des époques :

Courbe de précision :



Courbe de perte :

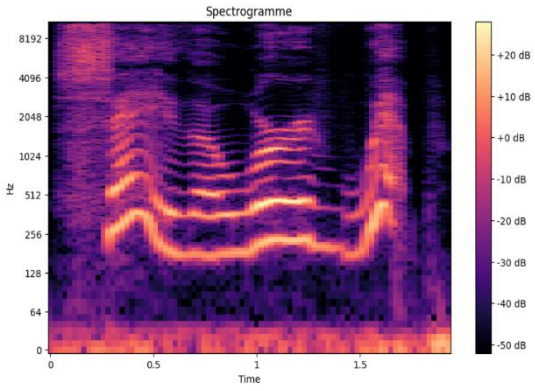


Visualisation des signaux audio :

Voici un exemple de signal audio pour une émotion "joyeuse" :

Onde sonore :

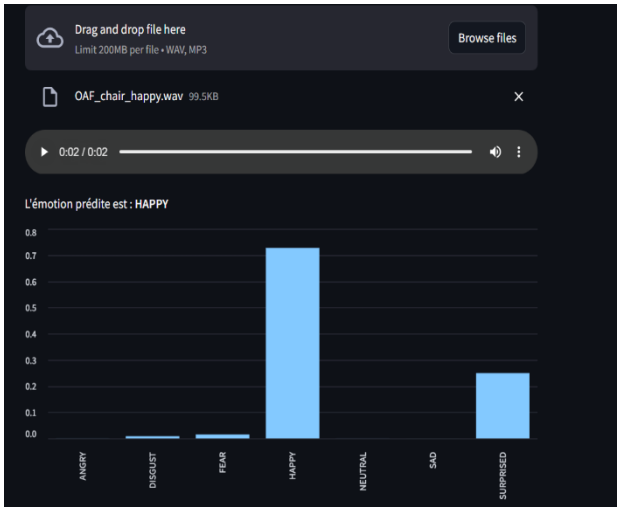
Spectrogramme :



Exemple de prédictions

Pour tester le modèle, un fichier audio contenant une phrase exprimant de la "joie" a été utilisé. Le modèle a correctement prédit l'émotion comme étant **HAPPY**.

Résultat affiché dans Streamlit :
L'interface utilisateur développée avec Streamlit permet d'importer un fichier audio, de visualiser son onde sonore et son spectrogramme, puis de prédire l'émotion. Voici une capture d'écran de l'interface :



Discussion

Limites

Malgré les performances satisfaisantes obtenues, certaines limites ont été identifiées :

1. **Qualité des enregistrements audio** :
   Les enregistrements du jeu de données TESS sont propres et enregistrés dans des conditions idéales. Cependant, dans des applications réelles, les enregistrements peuvent contenir du bruit ambiant, des échos, ou des variations de volume, ce qui peut affecter la précision du modèle.
2. **Homogénéité des locuteurs** :
   Les locuteurs dans le jeu de données sont limités à un groupe restreint, ce qui ne représente pas la diversité des voix humaines (âge, genre, accents). Cela peut réduire la généralisation du modèle à des locuteurs inconnus.
3. **Variabilité des émotions** :
   Les émotions sont exprimées différemment d'une personne à l'autre. Les données annotées peuvent parfois manquer de nuances ou ne pas refléter la complexité des émotions humaines, ce qui limite la capacité du modèle à détecter des émotions plus subtiles ou mixtes.

Améliorations futures

1. **Utiliser des jeux de données plus variés** :
   - Inclure des jeux de données comme **RAVDESS**, **CREMA-D**, ou des enregistrements du monde réel pour élargir la diversité des locuteurs et des conditions d'enregistrement.
   - Éventuellement, combiner plusieurs jeux de données pour améliorer la robustesse du modèle.
2. **Augmentation des données** :
   - Ajouter plus de bruit, des changements de pitch, ou des déformations temporelles pour simuler des environnements réels et rendre le modèle plus résilient.
3. **Enrichissement des modèles** :
   - Tester des architectures plus avancées comme **Attention Mechanisms** ou **Transformers**, qui peuvent mieux capturer les relations temporelles complexes dans les données audio.
   - Combiner plusieurs modèles (par exemple, des modèles CNN pour extraire des caractéristiques spectrales et des modèles RNN pour les relations temporelles).

Conclusion

Ce projet a permis de développer un modèle performant pour la reconnaissance des émotions à partir d'enregistrements audio, en s'appuyant sur des caractéristiques acoustiques (MFCC) et une architecture de réseau neuronal (LSTM). L'intégration d'une interface utilisateur avec Streamlit a facilité l'accès au modèle pour des prédictions en temps réel, illustrant ainsi une application pratique de l'intelligence artificielle dans le domaine des interactions homme-machine.

Les résultats obtenus, bien qu'encourageants, mettent en évidence les défis liés à la qualité et à la diversité des données. Des limitations comme l'homogénéité des locuteurs et l'impact du bruit ambiant soulignent l'importance d'utiliser des jeux de données plus variés et d'améliorer la robustesse du modèle. Les pistes d'améliorations identifiées, telles que l'intégration de nouvelles architectures et l'augmentation des données, ouvrent la voie à des travaux futurs prometteurs.

En conclusion, ce projet constitue une étape importante dans le domaine de l'analyse des émotions audio, avec des implications potentielles pour les assistants virtuels, les systèmes de bien-être émotionnel et d'autres interfaces basées sur l'IA.

Références

Voici une liste des références utilisées pour mener ce projet et rédiger ce rapport :

1. Busso, C., Bulut, M., Lee, C. C., et al. (2008). **"IEMOCAP: Interactive emotional dyadic motion capture database"**, *Language Resources and Evaluation Conference (LREC)*.
2. Livingstone, S. R., & Russo, F. A. (2018). **"The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)"**, *PLoS ONE, 13(5)*.
3. Koolagudi, S. G., & Rao, K. S. (2012). **"Emotion recognition from speech: A review"**, *International Journal of Speech Technology*.
4. Abadi, M., Barham, P., Chen, J., et al. (2016). **"TensorFlow: A system for large-scale machine learning"**, *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*.
5. McFee, B., et al. (2015). **"librosa: Audio and music signal analysis in Python"**, *Proceedings of the 14th Python in Science Conference*.
6. Streamlit Documentation. **"Streamlit: The fastest way to build and share data apps"**. Disponible sur : https://docs.streamlit.io
7. Goodfellow, I., Bengio, Y., & Courville, A. (2016). **"Deep Learning"**, *MIT Press*.
8. OpenAI Blog. **"Deep Learning for Audio Emotion Recognition"**. Disponible sur : https://openai.com/blog.