SPOKEN MALAGASY LANGUAGE DETECTION

DÉFINITION

La détection de la langue parlée (SLD) est le processus d'identification automatique de la langue parlée dans un enregistrement audio ou dans un texte.L'idée serait de faire les études sur la langue Malagasy à travers des données vocales puis d'identifier si la langue parlée est en malagasy.
Toutefois,cette étude comporte plusieurs contraintes et plusieurs challenges à prendre en compte:

- Complexité phonologique
- Variations des dialectes,(Merina,Betsilero,Sakalava,..)
- Utilisation d'autres langues étrangères

En raison de ces différentes difficultés, il est important de prendre en compte la diversité linguistique du malgache et de mener des études approfondies pour identifier de manière précise la langue parlée dans un contexte donné.

PROCESSUS

L'identification de la langue parlée repose sur l'analyse des caractéristiques acoustiques du signal vocal, telles que la fréquence, l'intensité, la durée des phonèmes et des mots, ainsi que les modèles de prosodie et d'intonation propres à chaque langue. Les avancées en matière de traitement du signal et d'apprentissage automatique ont permis le développement de modèles performants capables de reconnaître et de classifier les langues parlées avec une grande précision. Voici quelques étapes de l'étude :

1. Prétraitement du signal vocal : le signal vocal enregistré est prétraité pour éliminer le bruit de fond, normaliser le volume et extraire les caractéristiques acoustiques pertinentes.
2. Extraction des caractéristiques acoustiques : les caractéristiques acoustiques telles que la fréquence, l'intensité, la durée des phonèmes et des mots, ainsi que les modèles de prosodie propres à la langue malagasy sont extraites à partir du signal vocal.
3. Modélisation des caractéristiques : les caractéristiques acoustiques extraites sont utilisées pour entraîner des modèles d'apprentissage automatique capables de reconnaître les spécificités de la langue malagasy.
4. Classification de la langue : les modèles entraînés sont utilisés pour classifier le signal vocal enregistré et déterminer si la langue parlée est le malagasy ou une autre langue.
5. Évaluation et ajustement : le processus d'identification de la langue malagasy est évalué en utilisant des ensembles de données de test et ajusté si nécessaire pour améliorer la précision de la classification.

DIFFÉRENTES TECHNIQUES

La détection de la langue parlée de bas niveau fait référence au processus d'identification de la langue parlée par un locuteur en se basant sur des caractéristiques très basiques du signal vocal. Ces caractéristiques sont généralement dérivées de la forme d'onde audio brute elle-même et ne dépendent pas fortement de modèles d'apprentissage automatique sophistiqués. La détection de la langue de bas niveau peut être plus simple et plus rapide que les méthodes de détection de haut niveau, mais elle peut aussi être moins précise.

Les principaux challenges se trouvent dans la génération et dans le traitement de données et, après, dans l'architecture du modèle.
Voici quelques techniques pouvant servir lors du traitement de données vocales pour en tirer les caractéristiques significatives

- **FEATURE EXTRACTION**

Formant Analysis:

- Les formants sont les fréquences de résonance du tractus vocal. Ils peuvent varier d'une langue à l'autre et sont souvent utilisés pour les distinguer en analysant les formants dans le signal audio

Mel Frequency Cepstral Coefficients(MFCCs):

- Les MFCC sont une caractéristique largement utilisée en reconnaissance vocale qui représente le spectre de puissance à court terme d'un son. Ils sont dérivés de la transformée de Fourier et sont utilisés pour capturer la forme spectrale du signal vocal.Voici les étapes lors du MFCC :
    1. Prétraitement du signal : le signal audio est divisé en petites fenêtres de temps, généralement de 20 à 30 millisecondes, qui se chevauchent les unes par rapport aux autres.
    2. Transformation en domaine fréquentiel : chaque fenêtre de signal est transformée en domaine fréquentiel à l'aide de la transformée de Fourier discrète (DFT) ou de la transformée en cosinus discrète (DCT).
    3. Filtres de Mel : les fréquences sont ensuite passées à travers une série de filtres de Mel, qui sont conçus pour simuler la façon dont l'oreille humaine perçoit les différentes fréquences.
    4. Logarithme : les amplitudes des fréquences filtrées sont ensuite converties en échelle logarithmique pour mieux représenter la perception humaine de la fréquence.
    5. Transformation en cosinus discret (DCT) : enfin, une DCT est appliquée aux valeurs logarithmiques pour obtenir les coefficients cepstraux, qui sont les caractéristiques finales extraites par la MFCC.

Gammatone Cepstral Coefficient (GTCC):

- Les GTCC offrent une meilleure résolution aux basses fréquences que les MFCC. Cela est dû au fait qu'elle est basée sur l'échelle de largeur de bande rectangulaire équivalente, tandis que les MFCC sont basés sur l'échelle de Mel. Les GTCC utilisent la racine cubique, tandis que les MFCC utilisent le logarithme.

Perceptual Linear Prediction:

- utilise un modèle de prédiction linéaire pour estimer les coefficients de résonance d'un signal audio, en prenant en compte la sensibilité du système auditif humain.

Linear Predictive Coding (LPC):

- LPC est un outil utilisé en traitement de la parole pour modéliser l'enveloppe spectrale d'un signal vocal. Il peut être utilisé pour analyser les caractéristiques spectrales de la parole.

Pitch Tracking:

- Le ton du discours peut être suivi au fil du temps pour identifier la langue. Les différentes langues ont des plages de tonalité moyennes et des schémas de tonalité différents.

Prosodic Features:

- La prosodie fait référence au rythme, au stress et à l'intonation de la parole. Analyser ces caractéristiques peut aider à identifier la langue, car elles peuvent varier d'une langue à l'autre.

- **MODÈLES:**

<u>Acoustic Modeling:</u>

Hidden Markov Models(HMMs) :
Les HMM sont des modèles statistiques qui peuvent être utilisés pour modéliser la séquence des états cachés dans un signal vocal. Ils peuvent être utilisés pour identifier la langue en modélisant la séquence des phonèmes ou d'autres unités de parole.

Gaussian Mixture Models(GMMs):
Les GMM sont utilisés en reconnaissance vocale pour modéliser la distribution des caractéristiques de la parole. Ils peuvent être utilisés pour identifier la langue en comparant la distribution des caractéristiques dans le signal audio aux distributions connues pour différentes langues.

<u>Pronounciation Modeling</u>

DL :
Les modèles d'apprentissage profond, tels que les réseaux neuronaux convolutifs (CNN) ou les réseaux neuronaux récurrents (RNN), peuvent être entraînés sur de grands ensembles de données d'échantillons audio pour identifier la langue. Ces modèles peuvent saisir des schémas complexes dans le signal vocal qui ne seraient pas facilement capturés par des modèles plus simples.

ML :

Les modèles de classification classiques en ML et spécialement avec les times series peuvent aussi marcher parfois dans ce genre d'étude (SVM,Arbres , ..)

Language Modeling

N-gram model:
Un modèle N-gramme est un type de modèle de langue statistique qui prédit la probabilité d'une séquence donnée de mots dans un texte ou un discours.

DIFFÉRENTES APPROCHES :

1. **Spoken Language Identification Using Deep Learning (CNN)**
   Les réseaux de neurones convolutifs (CNN) peuvent être utilisés pour identifier plusieurs langues, et elle est efficace et rapide. L'apprentissage en profondeur et la technique du spectre ont été utilisés pour compléter l'analyse. Les caractéristiques des photos sont extraites à l'aide d'un réseau neuronal convolutionnel (CNN) afin de les classer [4] . Enfin, la catégorisation multilingue est réalisée à l'aide de la fonction d'activation softmax.

2. **CLIASR: A Combined Automatic Speech Recognition and Language Identification System**
   Cette étude décrit une approche appelée Système combiné de reconnaissance automatique de la parole et d'identification de la langue, qui associe les technologies de reconnaissance automatique de la parole et d'identification de la langue pour reconnaître les chiffres prononcés après avoir identifié la langue dans laquelle ils sont prononcés [5].

3. **A Language Identification System using Hybrid Features and Back-Propagation Neural Network [2020]**
   L'identification de la langue [7] est le processus d'identification précise d'une langue inconnue en comparant les données biométriques d'un échantillon de parole testé avec les modèles de langue élaborés au préalable. Pour les systèmes d'identification de la langue parlée (LID), cette étude propose et encourage l'utilisation d'algorithmes hybrides d'extraction de caractéristiques robustes.

4. **Spoken Language Identification System Using Convolutional Recurrent Neural Network**
   Cette approche d'identification est basée sur l'arrangement des features vectors[9] .le système proposé utilise un réseau neuronal récurrent convolutif hybride (CRNN), qui combine un réseau neuronal convolutif (CNN) et un réseau neuronal récurrent (RNN). L'architecture CRNN est créée par le système proposé en combinant les meilleures caractéristiques des architectures CNN et RNN

EXPÉRIENCES

On a fait une expérience sur l'identification et classification de 3 langues : Espagnol , Allemand et l'Anglais . On a pu essayer les techniques (1) de CNN et (4) de RNN .
Le Dataset est un ensemble de données vocales constituées des vocales des 3 langues ayant le même nombre pour chacune vu sur Kaggle .Les données ont été transformées en dataframes ayant la source du vocal dans une colonne ainsi que la langue parlée dans le vocal dans une autre colonne .
Après , le processus cité ci-dessus avait été appliqué , et puisque la **mfcc** est la technique de feature extraction la plus populaire ; en procédant ainsi , les données vocales ont été transformées en données numériques idéales pour les modèles de prédiction . La MFCC extrait les caractéristiques les plus importantes du signal audio en se concentrant sur les propriétés fréquentielles du son.
Ensuite , on a divisé les données en train ,test et validation et c'est après qu'on a testé des modèles de CNN et de RNN (LSTM) . On n'a pas pu faire de la fine-tuning mais on s'est basé sur des architectures existantes pour entraîner les données et c'est ainsi qu'on a constaté les résultats de CNN bien meilleurs que ceux des RNN .Certes , on pouvait bien améliorer la performance sur RNN mais le but était de se familiariser avec certaines méthodes et techniques concernant l'identification de la langue parlée .
Voici le lien du code github du code :
https://github.com/Harena21/Malagasy-language-identification

RÉFÉRENCES

- "National library of medicine." https://www.ncbi.nlm.nih.gov/ .
- "Researchgate." www.researchgate.net , 2008.
- A. A. Alashban, M. A. Qamhan, A. H. Meftah, and Y. A. Alotaibi, "Spoken language identification system using convolutional recurrent neural network," Applied Sciences, vol. 12, no. 18, 2022.
- K. Lounnas, H. Satori, M. Hamidi, H. Teffahi, M. Abbas, and M. Li chouri, "Cliasr: A combined automatic speech recognition and language identification system," in 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), pp. 1–5, 2020.
-
-