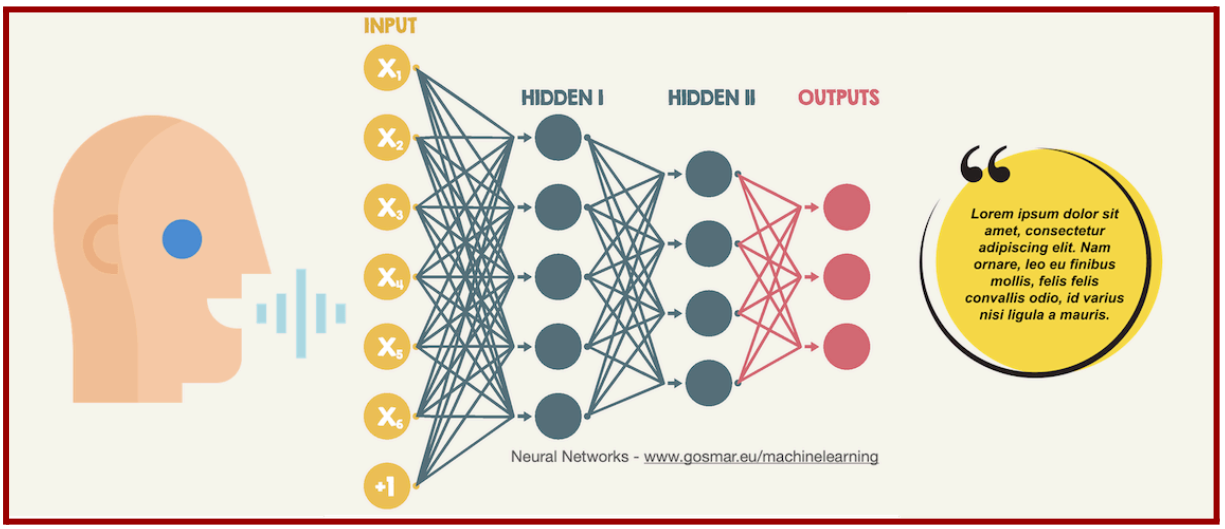
# Reconnaissance de Langue
# avec réseaux de neurones convolutif (CNN)




**Kenza AHMIA** (Inalco, numéro d'étudiant 21902213)
**Shami THIRION SEN** (Inalco, numéro d'étudiant 22200036 )

# Réseaux de neurones pour reconnaissance de l'oral et applications linguistiques

**Cédric GENDROT**
**Nicolas AUDIBERT**

## Introduction

Dans le cadre du cours sur **les réseaux de neurones dédiés à la reconnaissance vocale**, notre tâche consistait à élaborer un projet de reconnaissance de la parole en exploitant un système basé sur des réseaux de neurones.

Pour notre part, nous avons opté pour **l'étude de la reconnaissance automatique des langues** en utilisant la bibliothèque Keras sous l'environnement Python. L'objectif principal de ce projet était de mettre en œuvre l'entraînement des modèles à l'aide de spectrogrammes générés à partir d'enregistrements vocaux, suivis d'une évaluation de la performance sur une classification multi-label. Cette classification impliquait la reconnaissance des langues au sein d'un corpus comprenant l'arabe, l'anglais, l'espagnol et le français.

## Choix des données

Le choix des langues s'est fondé sur notre familiarité avec celles-ci ainsi que sur le degré de proximité et de divergence entre elles. Le français et l'espagnol, étant toutes deux des langues indo-européennes, sont donc étroitement liées, tandis que l'anglais est une langue de même origine, mais possède des particularités distinctes. De plus, l'arabe appartient à la famille chamito-sémitique, marquant ainsi une différence linguistique significative.

Le corpus oral se compose de 2000 fichiers audio extraits de Mozilla Common Voice. Pour chaque langue, nous avons utilisé des scripts Praat et Python afin d'extraire les phonèmes. La liste initiale des phonèmes pour toutes les langues était [a, u, v]. En procédant à des choix ultérieurs, nous avons ajusté cette liste en tenant compte des particularités linguistiques observées.

Dans le cas de l'anglais, nous avons opté pour /i:/ en raison de la rareté du phonème /i/ par rapport aux autres langues. Pour l'espagnol, notre choix s'est porté sur /B/ en raison de la rareté, du phonème /b/ par rapport aux autres langues. En ce qui concerne l'arabe, notre sélection s'est orientée vers /?//, équivalent du /R/. La liste finale des phonèmes retenus pour toutes les langues est donc ['a', 'i', 'i:', 'u', 'b', 'B', 'R', 'v', '?/'].

Les décisions de choix de phonèmes ont été orientées par des considérations linguistiques spécifiques, incluant la présence ou l'absence de certains sons dans chaque langue, ainsi que le degré de proximité ou d'éloignement phonétique entre elles.

## Hypothèses

Avant d'entreprendre la prochaine phase, nous avons formulé certaines hypothèses concernant les résultats que nous anticipons. Ces hypothèses peuvent être illustrées de la manière suivante : il est possible que le réseau démontre des performances élevées en ce qui concerne l'anglais et l'arabe, réussissant à les différencier de l'espagnol et du français. Cependant, il pourrait rencontrer des difficultés particulières avec le français et l'espagnol. Une autre possibilité serait que le modèle ne soit performant pour aucune langue, en raison de la similitude entre plusieurs phonèmes, tels que les voyelles [‘a’, ‘u’] et la consonne [‘v’].

## Prétraitement des données

En complément de l'objectif initial, nous avons également saisi l'opportunité pour évaluer la praticité et l'efficacité des divers outils de prétraitement. Cette démarche avait pour but d'analyser leur facilité d'utilisation et de déterminer leur impact potentiel, en vue d'établir des paramètres favorables pour des travaux futurs.

1. Alignement et génération de TextGrid

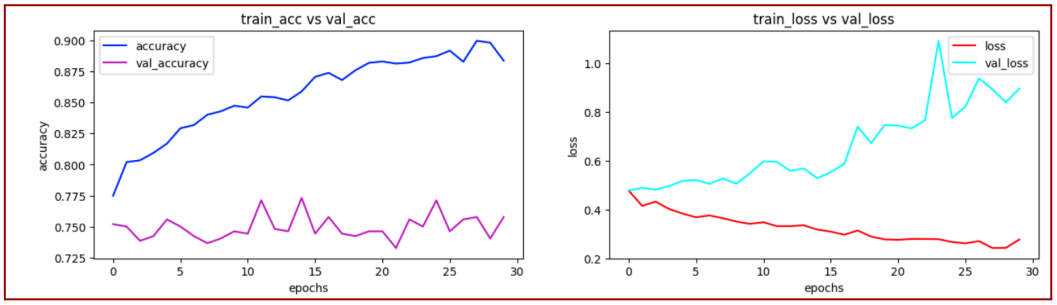
Après avoir téléchargé les fichiers audio et extrait les enregistrements validés avec leurs transcriptions, nous avons effectué l'alignement crucial en utilisant les outils WebMAUS et WebMINNI. Ces services, dédiés à la segmentation et à l'étiquetage phonétique, ont démontré leur efficacité, WebMAUS pour les langues latines et WebMINNI en tant qu'alternative performante pour l'arabe sans translittération. Cependant, leur performance diminue notablement avec des jeux de données volumineux, limitant ainsi leur capacité à gérer de grandes quantités de données.

2. Génération de spectrogrammes

Pour générer les spectrogrammes, nous avons utilisé les scripts fournis par notre enseignant encadrant. Praat s'est révélé moins efficace pour le français et l'arabe, car le processus de passage des fichiers audio aux spectrogrammes était long. Il aurait fallu d'abord couper au niveau des sons, ce qui a entraîné des surplus, nécessitant un tri pour éliminer les sons indésirables, parfois des mots entiers, surtout pour le français. En revanche, l'utilisation de Python pour l'espagnol et l'anglais s'est avérée plus pratique.
Nous devons préciser que les spectrogrammes pour le FR et AR ne sont pas sans padding comparé à ceux de l'ES et de l'EN.

## Résultats et observation du modèle de classification

Sur 30 epochs nous observons que l'accuracy de l'entraînement individuel continue à augmenter mais l'accuracy globale ou « value accuracy » oscille et continue à diminuer à à partir de epoch 13 et 14. Ainsi, l'entraînement n'a pas d'incidence sur la perte globale du modèle.




Sur les données de test :

```
[Test loss: 0.8947702050209045 Test accuracy: 0.7576923370361328]
```

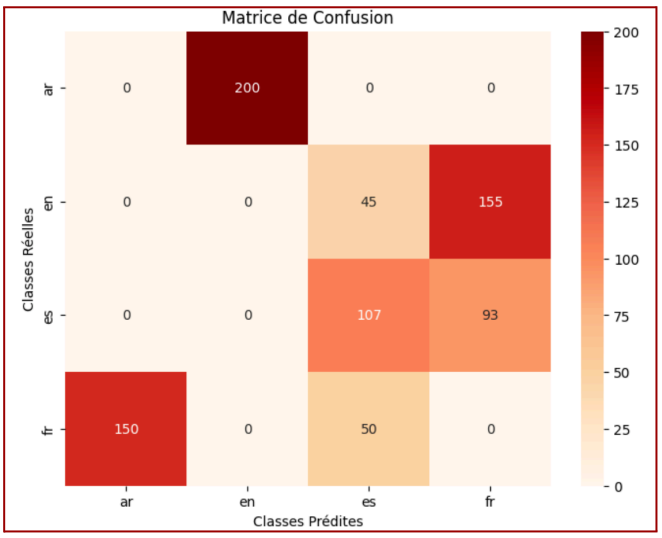
Les résultats montrent que notre modèle a une perte de 0.895 et une précision de 75.8% lorsqu'il est testé. Ces chiffres indiquent comment bien le modèle se comporte sur un ensemble de données différent de celui sur lequel il a été entraîné. En général, une perte plus basse et une précision plus élevée suggèrent que notre modèle est performant.

Sur les données inconnus :

- Pour la classe arabe (ar), aucune prédiction correcte n'a été faite, toutes les observations en ont été prédites comme anglaises (200).

- Pour la classe anglaise (en), aucune prédiction correcte n'a été faite, toutes les observations en ont été prédites comme espagnoles (45) ou françaises (155).

- Pour la classe espagnole (es), 107 prédictions correctes ont été faites, tandis que 93 observations espagnoles ont été incorrectement classées comme françaises.
- Pour la classe française (fr), toutes les prédictions sont incorrectes, toutes les observations françaises ont été incorrectement classées comme arabes (150) et espagnole (50).
→ Ces résultats pourraient s'expliquer par les similitudes phonétiques entre les langues, validant ainsi nos hypothèses initiales. De plus, la non-suppression du padding pour le français (FR) et l'arabe (AR), ainsi que la distribution déséquilibrée du nombre de sons dans les ensembles d'entraînement et de test, en plus de la sélection aléatoire des données inconnues, pourraient avoir constitué des facteurs significatifs.




Après un premier essai de 1000 spectrogrammes par langues , nous avons pu observer qu'avec 1500 fichiers langues les résultats s'améliorent considérablement. Ceci nous mène à formuler l'hypothèse qu'avec une plus grande quantité de données nous pouvons obtenir de meilleurs résultats.

## Conclusion

En conclusion, ce projet a été une expérience enrichissante qui nous a permis de développer une compréhension approfondie du fonctionnement et de l'utilisation des Réseaux de Neurones. Cependant, en raison de contraintes de temps, nous avons été limités dans l'exploration de divers types de données. Pour améliorer notre approche, il serait bénéfique d'expérimenter avec une variété de données, d'augmenter la quantité des données disponibles, et d'affiner davantage les processus de prétraitement.
En vue de perspectives futures, nous recommandons une amélioration du modèle en divisant les données en ensembles distincts de formation, de test et de validation. L'utilisation **d'ImageDataGenerator** pendant l'entraînement peut contribuer à atténuer le surapprentissage et à renforcer la généralisation du modèle.

Un point crucial à considérer serait le choix judicieux des phonèmes, en privilégiant la qualité des langues et des phonèmes plutôt que la quantité brute. Cela permettra une meilleure adaptation du modèle aux spécificités linguistiques des langues étudiées. En somme, bien que le projet ait été stimulant, il ouvre la voie à des perspectives futures passionnantes et plus approfondies dans le domaine des réseaux de neurones pour la reconnaissance vocale.