كلية العلوم و التقنيات فاس
ⵜⴰⵖⵚⵍⵓⵍⵜ ⵏ ⵜⵇⵇⵓⵔⵉⵙ ⴷ ⵜⵓⵙⵙⵙⵜⵉ
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵓⴰⵍⵓⵙⵜ ⵓⵙⴷ ⵎⵓⵃⴰⵎⴷ ⵏ ⵄⴰⴱⴷⵍⵍⴰⵀ
Université Sidi Mohamed Ben Abdellah

Département d'Informatique

Projet de fin d'études

Master Sciences et Techniques
Systèmes Intelligents & Réseaux

Identification automatique du locuteur par les réseaux de neurones convolutifs

Lieu du stage : Laboratoire Systèmes Intelligents et Applications

Réalisé par :

- Rbiha Louaie

Encadré par :

- Mr. JAMAL KHARROUBI
- Mr. SOUFIANE HOURRI

Soutenu le 12.06.2018 devant le jury composé de :

- Pr. A. Talibi	Faculté des Sciences et Techniques de Fès	(Président)
- Pr. A. Benabbou	Faculté des Sciences et Techniques de Fès	(Examinateur)
- Pr. A. Majda	Faculté des Sciences et Techniques de Fès	(Examinatrice)
- Pr. J. Kharroubi	Faculté des Sciences et Techniques de Fès	(Encadrant)
- Mr. S. Hourri	Faculté des Sciences et Techniques de Fès	(Encadrant)

Année Universitaire 2017 – 2018

Remerciements

Avant toute chose je tiens à exprimer toute ma reconnaissance envers Monsieur **Jamal Kharroubi**, encadrant de mon stage, pour l'accord d'une attention particulière tout au long du stage ainsi que Monsieur **Soufiane Kourri** doctorants au sein de laboratoire des systèmes intelligents et applications pour ses conseils, orientations et son précieux aide.

Mes vifs remerciements vont également aux **membres du jury** pour l'intérêt qu'ils ont porté à ma recherche en acceptant d'examiner mon travail et de l'enrichir par leurs propositions.

Un grand merci à ma mère et mon père, pour leurs conseils ainsi que leur soutien inconditionnel, à la fois moral et économique.

Enfin, nous tenons également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Résumé

La reconnaissance automatique de locuteur est le processus d'authentification d'un locuteur à partir des caractéristiques de sa voix.

Dans ce rapport nous étudions la reconnaissance automatique de locuteur en mode indépendant du texte pour la tâche de vérification et principalement la tâche d'identification en se basant sur les images spectrogrammes de chaque locuteur et en utilisant les réseaux de neurones de convolutions pour l'étape de génération de modèles de références et aussi pour l'apprentissage de ces derniers.

Mots-clés : CNN, Spectrogramme, Vérification de locuteur, Identification du locuteur.

Abstract

Automatic speaker recognition is the process of authentification of a speaker based on the characteristics of his voice.

In this report we study automatic speaker recognition in text-independent mode for the verification task and mainly the identification task based on the spectrogram images of each speaker and using convolutional neural networks for the step of generation of reference models and also for the learning phase.

Keywords : CNN, Spectrogram, Speaker Verification, Speaker Identification.

Table des matières

Liste des figures

Listes des acronymes

RAL : Reconnaissance automatique du locuteur.

IAL : Identification automatique du locuteur.

VAL : Vérification automatique du locuteur.

UBM : Universal Background Model.

CNN : Convolution al Neural Network.

RBM : Restricted Boltzmann Machine.

MFCC : Mel-Frequency Cepstral Coefficients.

LPC : Linear Predictive Coefficients.

GMM : Gaussian Mixture Model.

HMM : Hidden Markov Model.

SVM : Support Vector Machine.

MAP : Maximum A Posteriori

DCT : Discrete Cosine Transform.

VQ : Vector Quantization.

TDNN : Time Delay Neural Network.

NTN : Neural Tree Network.

PLDA : Probabilistic Linear Discriminant Analysis.

Introduction générale

Depuis plusieurs dizaines d'années, la reconnaissance automatique du locuteur fait l'objet de travaux de recherche entrepris par de nombreuses équipes dans le monde.

La R.A.L présente deux branches principales : La vérification et l'identification automatique du locuteur, cette dernière permet de déterminer l'identité d'un individu inconnu sur la base de sa voix.

Cependant la majorité des systèmes actuels de reconnaissance automatique du locuteur sont basés sur l'utilisation des Modèles de Mélange de lois Gaussiennes (GMM) et/ou des modèles discriminants SVM.

En collaboration avec le laboratoire des systèmes intelligents et applications, le travail vise à présenter de nouvelles variantes basées sur les réseaux de neurones de convolutions ainsi qu'une approche différente qui est plus ou moins utilisée par rapport à ce qui est présent ces derniers jours.

La réalisation de ce travail suit une démarche commençant par une étude bibliographique sur la reconnaissance automatique du locuteur ainsi que les réseaux de neurones convolutifs suivi par une étape de développement en utilisant le langage Python et le Framework TensorFlow.

Ce présent rapport se compose de trois chapitres dont le premier donne un aperçu sur le laboratoire S.I.A ainsi que la problématique, les solutions proposées et les objectifs du projet. Le deuxième chapitre a pour but d'introduire les principes de reconnaissance automatique du locuteur et les réseaux de neurones de convolutions. Le dernier chapitre est consacré pour la présentation de la démarche suivie, les résultats obtenus et les perspectives proposées permettant de les améliorer.

Chapitre 1

Cadre général du projet

I. Introduction

Dans ce chapitre, nous allons présenter le laboratoire SIA et détailler l'objectif du travail avec la démarche adoptée.

II. Présentation du laboratoire S.I.A

Le laboratoire SIA, crée en 2011, est une unité de Recherche du Centre d'Etudes Doctorales en Sciences et Techniques de l'Ingénieur domicilié à la Faculté des Sciences et Techniques de Fès et regroupant des laboratoires de recherche tous accrédités par l'Université Sidi Mohamed Ben Abdellah de Fès, et domiciliés à la Faculté des Sciences et Techniques, l'Ecole Supérieure de Technologie, la Faculté Polydisciplinaire de Taza, l'Ecole Nationale des Sciences Appliquées de Fès et l'ENS de Fès.

Le LSIA est composé de 12 enseignants-chercheurs du département d'Informatique de la FST de Fès et de 23 doctorants. Cette imbrication étroite entre enseignement et recherche, est un élément essentiel de la dynamique du laboratoire.

Les thématiques de recherche se situent au cœur des Sciences et Technologies de l'Information et de la Communication et s'articulent essentiellement autour des thématiques de recherche des enseignants chercheurs du laboratoire et assure une large couverture thématique présentant un atout très important pour le laboratoire.

III. L'objectif du travail

Le présent travail s'inclue principalement dans le domaine de reconnaissance automatique du locuteur qui a pour objectif en premier lieu de savoir si on peut identifier des locuteurs en mode indépendant du texte à travers une approche différente qui est basée sur l'étude des spectrogrammes et les réseaux de neurones profonds, par la suite on vise a amélioré les résultats obtenus en employant différentes techniques et en réexaminant la démarche suivie pour aboutir à nos fins.

IV. Vue générale sur la démarche adoptée

Dans le cadre de la reconnaissance automatique du locuteur la démarche adoptée va suivre des étapes bien précises commençant par une étude des spectrogrammes qui vont être le principale outil dans notre tâche d'identification automatique du locuteur par la suite une étape de prétraitement de données va avoir lieu pour bien personnaliser nos données d'apprentissage, la prochaine étape consiste à extraire les caractéristiques les plus pertinentes à partir des images spectrogrammes et générer les modèles à partir de ces derniers en se basant sur les CNN, et à partir d'une nouvelle image spectrogramme d'un locuteur on va obtenir différents scores que l'un d'eux va représenter l'identité du locuteur.

Chapitre 2

La reconnaissance automatique du locuteur et les réseaux de neurones profonds

I. Introduction Générale

Ce chapitre est consacré pour l'état de l'art de la reconnaissance automatique de locuteur, sa structure générale et ces principales tâches ainsi qu'une présentation des réseaux de neurones profonds.

II. La reconnaissance automatique du locuteur

La reconnaissance automatique de locuteur est un terme générique utilisé pour toute procédure qui implique la connaissance de l'identité d'une personne sur la base de sa voix.

C'est un domaine très vaste où on recherche des méthodes pour extraire les caractéristiques vocales propres à chaque individu tel que le bilan socio-culturelle, l'état physique, émotionnelle et l'état de l'environnement, qui vont servir à créer une signature vocale qui permet de reconnaitre la voix de chacun de nous.

1. Branches de la reconnaissance automatique du locuteur

La discipline de reconnaissance des locuteurs comporte nombreuses branches qui sont directement ou indirectement liées, les branches simples de reconnaissance de locuteur sont celles qui sont autonomes (vérification, identification, classification), les branches composées sont celles qui utilisent une ou plusieurs manifestations simples avec éventuellement des techniques supplémentaires (segmentation, détection et suivi du locuteur).

a) Vérification automatique du locuteur

La vérification automatique du locuteur consiste à accepter ou refuser l'identité proclamée par un locuteur, en se basant sur un modèle qui lui est associé.

L'utilisateur s'identifie généralement par des méthodes non orthophoniques (un code), l'ID fourni est utilisé ensuite pour récupérer le modèle de ce dernier à partir d'une base de données.

Par la suite le signal vocal de l'utilisateur est comparé au modèle de référence afin de bien vérifier l'identité de l'utilisateur.

L'évaluation de similarité entre ces deux derniers se fait généralement par l'introduction d'un autre modèle concurrent (Cohort, UBM)[1].

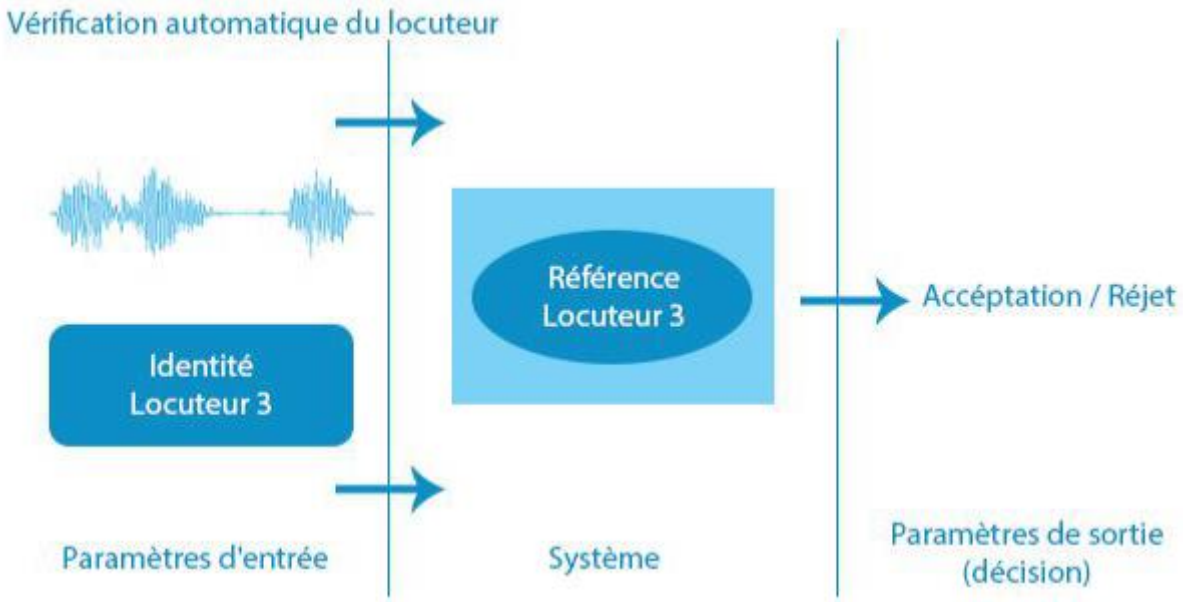


Figure 1 : La tache de vérification automatique du locuteur

b) Identification automatique du locuteur

L'identification automatique du locuteur consiste à reconnaitre un locuteur particulier parmi un ensemble de locuteurs possibles. Il existe deux types d'identifications dont la première est faite sur un ensemble fermé, où on suppose que l'utilisateur sera toujours connu par le système et par suite il n'y aura pas un schéma de rejet, la deuxième est faite sur un ensemble ouvert où l'utilisateur peut ne pas être connu.

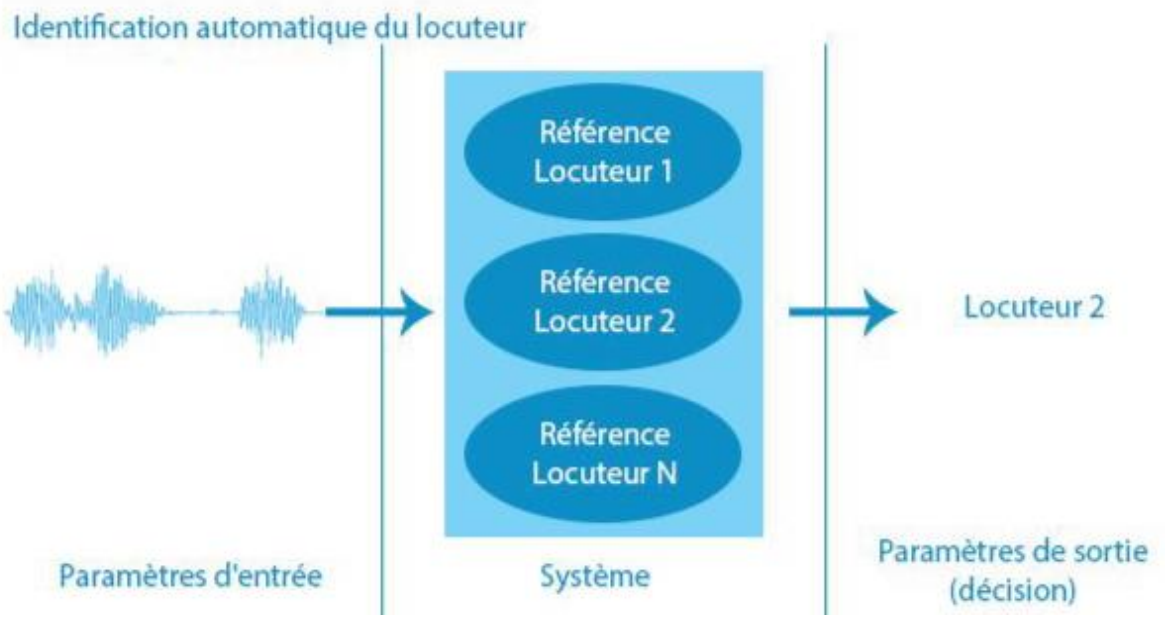


Figure 2: La tâche d'identification automatique du locuteur

c) Classification

Cette tâche peut être vue comme une étiquette générale pour n'importe quelle technique qui regroupe des signaux audios semblables dans des casiers individuels.

La classification peut utiliser des ensembles de caractéristiques légèrement différents de ceux utilisés pour la vérification et l'identification tel que l'âge et le sexe.

d) Segmentation

Le défi de la segmentation est de pouvoir séparer le discours produit par différents interlocuteurs les uns des autres.

Pour chaque deux segments adjacents des modèles de caractéristiques sont créés et une évaluation de similarité des deux segments est effectuée. Si les segments sont jugés suffisamment dissemblables, une marque de segmentation est réalisée à ce point de transition.

Si les identités des locuteurs sont connues, alors le sous-problème devient un problème d'identification.

e) Détection du locuteur

La détection du locuteur est l'acte consistant à détecter un ou plusieurs locuteurs spécifiques dans un flux audio. Par conséquent, la théorie derrière englobe la segmentation ainsi que l'identification et / ou la vérification des locuteurs, le choix dépend principalement de la formulation du problème.

f) Suivi du locuteur

Le suivi de locuteur est un peu similaire à la détection de locuteur avec la différence subtile qu'un ou plusieurs locuteurs sont suivis dans le flux.

Dans la plupart des cas, tant que les différents locuteurs de flux sont identifiés par des étiquettes générales telles que des étiquettes alphanumériques, le but est atteint.

2. Les approches scientifiques en reconnaissance automatique du locuteur

La majorité des systèmes actuels de reconnaissance automatique du locuteur sont basés sur l'utilisation de modèles de mélange de gaussiennes (GMM).

Ces modèles de nature générative sont généralement appris en utilisant les techniques de Maximum de Vraisemblance et de Maximum A Posteriori (MAP)[2]. Cependant, cet apprentissage génératif ne s'attaque pas directement au problème de classification étant donné qu'il fournit un modèle de la distribution jointe.

Ceci a conduit récemment à l'émergence d'approches discriminantes qui tentent de résoudre directement le problème de classification[3], et qui donnent généralement de bien meilleurs

résultats. Par exemple les machines à vecteurs de support (SVM), combinées avec les vecteurs GMM sont parmi les techniques les plus performantes en reconnaissance automatique du locuteur[4].

3. Modalités de reconnaissance automatique du locuteur

La reconnaissance du locuteur peut être mise en œuvre en utilisant différentes modalités qui sont liées à l'utilisation de la linguistique, du contexte et d'autres moyens.

Il existe différentes façons de reconnaître la voix d'un individu. On peut le faire, par exemple, à partir d'une phrase lue par une personne. Dans ce cas la reconnaissance sera dite text-dependant. Si le locuteur est libre de raconter ce qu'il veut, elle sera alors dite text-independant.

4. Points forts/faibles de la reconnaissance automatique du locuteur

Le signal est redondant et peut donc être incomplet : ceci est l'un des plus grands points forts de la reconnaissance automatique de locuteur vu que le signal de parole peut subir des altérations acoustiques provenant de diverses origines et malgré ces dégradations l'auditeur est capable de récupérer les informations manquantes mais cela n'empêche tout de même pas quelques erreurs de perception[5].

La reconnaissance automatique fait face à de nombreux problèmes liés au domaine applicatif comme l'utilisation des systèmes dans des conditions difficiles, les tentatives d'imposture la variabilité due au matériel, etc.

5. Applications de la reconnaissance automatique de la parole et du locuteur

La parole est certainement le moyen de communication directe entre humains qui est le plus sophistiqué. Les subtiles variations du langage sont capables de susciter chez l'auditeur non seulement une palette fort variée d'émotions et de sentiments, mais aussi une attention complète de son cerveau.

Les ordinateurs et les logiciels qui se construisent actuellement, bien que capables de traiter énormément d'informations en un temps très court, n'ont pas encore la capacité de générer ou de comprendre les finesses de la parole humaine.

Cependant, de nombreuses applications en reconnaissance de la parole sont déjà industrialisées, allant de la dictée vocale a la commande d'opérations diverses dans les navettes spatiales.

De plus en plus, les entreprises de télécommunications et de services (banques, assurances), désireuses d'améliorer leur service à la clientèle, tentent d'introduire des applications basées sur les technologies de la parole.

La palette de ces technologies est fort riche, partant de systèmes de reconnaissance de la parole entraines pour un seul locuteur a des systèmes capables de reconnaitre des centaines

de milliers de mots. Dans un autre registre, un grand nombre de services demandent une reconnaissance de l'identité du locuteur (accès aux boites vocales, à des services par abonnements, consultation de comptes en banques, etc…).

Finalement, pour le dialogue homme-machine soit complet, le domaine de la synthèse de la parole essaie de produire de la voix humaine (ou y ressemblant fort) automatiquement.

6. La mise en place d'un système de reconnaissance automatique du locuteur

Deux phases distinctes sont nécessaires pour la mise en place d'un système de reconnaissance automatique de locuteur, la première phase consiste à construire les modèles de références pour chaque locuteur à partir d'une ou plusieurs sessions d'enregistrements et la deuxième phase permet d'évaluer la robustesse du système de reconnaissance automatique de locuteur en présentant chaque locuteur devant ce dernier.

7. Structure des systèmes de reconnaissance automatique du locuteur

Les systèmes de reconnaissance automatique de locuteur peuvent être vus comme un enchaînement de processus qui sont : la paramétrisation, la modélisation et la décision.

a) La paramétrisation

Le processus de paramétrisation a pour rôle d'extraire les caractéristiques les plus pertinentes du signal de la parole. Ces dernières doivent avoir un degré de précision élevé avec un nombre raisonnable pour alléger les calculs, ils doivent être robustes des variations de niveau sonore et doivent donner une meilleure représentation des signaux vocaux et les rendre facilement séparables.

b) La modélisation

Le processus de modélisation vise principalement les taches d'identification et vérification automatique du locuteur pour une modélisation réalisable à travers plusieurs sessions d'enregistrements à partir des données d'apprentissage.
Par la suite une mesure de similarité va avoir lieu entre les modèles de références et les caractéristiques du signal vocal de l'utilisateur pour pouvoir continuer l'enchainement jusqu'à le processus de décision.

c) La décision

Généralement le processus de décision dépend fortement de la formulation du problème à résoudre avec les systèmes de reconnaissance automatique de locuteur ainsi que de la structure de l'environnement où on veut l'appliquer, pour cela un seuil de décision est choisi selon les besoins.

Conclusion

Les technologies basées sur le signal de la voix proposent plus de sureté et de confort d'utilisation que les techniques d'authentifications classiques qui sont vulnérables au vol et à la falsification.

III. Les réseaux de neurones profonds

Les réseaux de neurones artificiels sont des réseaux fortement inspirés du fonctionnement des neurones biologiques et connectés de processeurs élémentaires fonctionnant en parallèle où chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit.

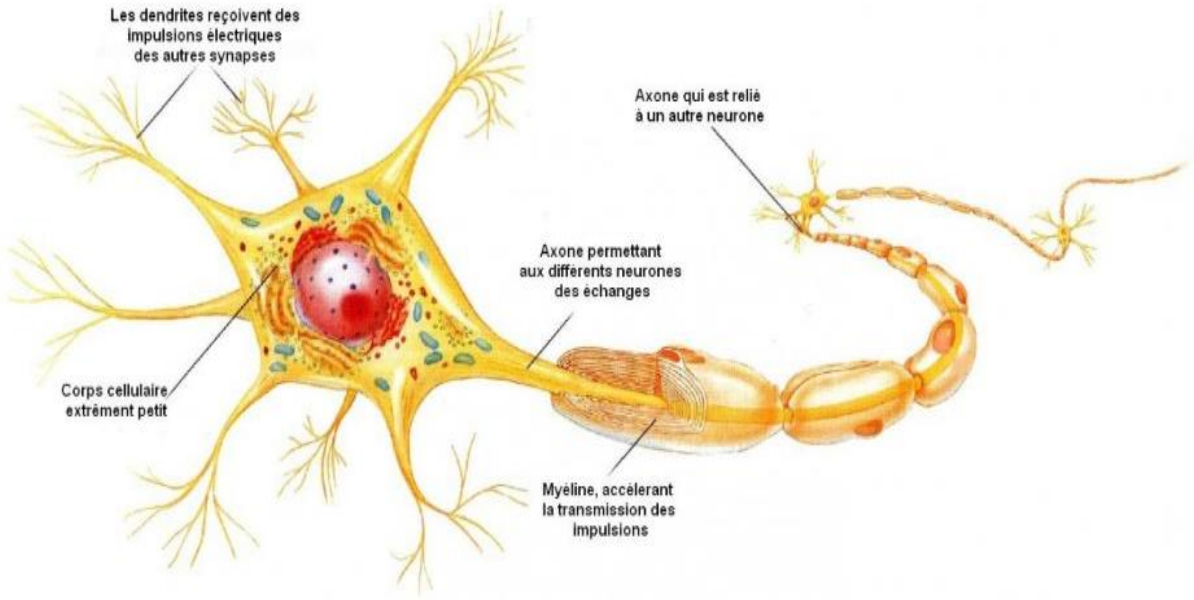


Figure 3 : Modèle du neurone biologique

1. Historique des neurones formels

L'histoire a commencé avec la construction d'un modèle simplifié de neurone biologique qui est appelé neurone formel. Les travaux menés sur ce dernier ont montré qu'avec un réseau de neurones formels on peut théoriquement réaliser des fonctions logiques, arithmétiques et symboliques complexes[6].
Ce neurone formel est doté d'une fonction de transfert qui lui permet, selon des règles, de transformer ces entrées en une sortie, il possède des paramètres importants tel que les coefficients synaptiques et le seuil de chaque neurone, et la façon de les ajuster, la chose qui détermine l'évolution du réseau en fonction de ses informations d'entrées.

Pour que le réseau de neurones formels propose une solution optimale, ces paramètres doivent converger vers des valeurs qui assurent la bonne classification lors de la phase d'apprentissage et par suite l'apprentissage dans un réseau de neurones formels revient à adapter les coefficients synaptiques pour classifier les exemples présentés en entrée.

2. Perceptron

Comme les derniers travaux n'ont pas proposé une méthode pour adapter ces coefficients synaptiques, on a proposé une règle simple qui permet de les modifier en fonction de l'activité des unités qui les relient.

C'est la règle de Hebb qui est presque partout présente dans les modèles actuels, même les plus sophistiqués[7].

A partir de ces travaux on a pu constituer un modèle de perceptron qui est le premier système artificiel capable d'apprendre par expérience, y compris lorsque son instructeur commet quelques erreurs, ce qui le diffère d'un système d'apprentissage logique formel[8].

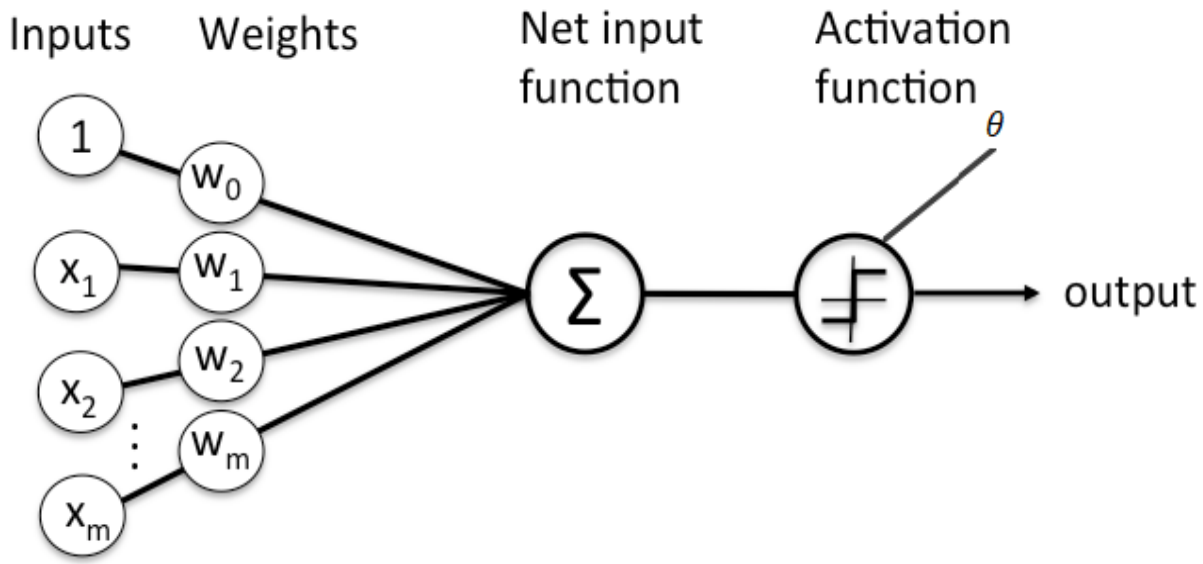


Figure 4 : Modélisation du perceptron

La sortie du perceptron est calculée selon la formule suivante :

$$O = \begin{cases} 1 \; si \sum_i Wi.Xi > \theta \\ 0 \qquad\qquad Sinon \end{cases}$$

Lorsque deux neurones sont excités conjointement, un lien se crée pour les unir, c'est ce que Donald Hebb a proposé comme règle d'apprentissage des réseaux de neurones artificiels.

Pour un seul neurone artificiel qui utilise la fonction signe comme fonction d'activation on a alors :

$$Wi' = Wi + \alpha \, (O.Xi)$$

Avec Wi' le nouveau poids i et α représente le pas d'apprentissage.

Frank Rosenblatt à créer aussi une règle qui ne diffère de la précédente seulement dans la vision et le calcul de l'erreur observée en sortie :

$$Wi' = Wi + \alpha(O_t - O).Xi$$

Où O_t représente la sortie désirée.

a) Apprentissage

L'apprentissage consiste à modifier les poids de connexions entre les neurones selon des règles de modification comme cité ci-dessus, généralement on a deux algorithmes qui sont les plus répandues dans ce domaine pour l'apprentissage d'un réseau de neurone monocouche :

- L'algorithme de Widrow-Hoff.
- L'apprentissage par descente de gradient.

La loi de Widrow-Hoff est la règle d'apprentissage du perceptron qui est locale, c'est à dire que chaque cellule de sortie apprend indépendamment des autres et qui s'écrit de la manière suivante :

$$W_{i,j}(t+1) = W_{i,j}(t) + \alpha(O_t - O_j).X_i$$

$W_{i,j}(t+1)$: poids de la liaison i -> j au temps t + 1

$W_{i,j}(t)$: poids de la liaison i -> j au temps t

X_i = sortie (0 ou 1) de la cellule i

O_j = réponse de la cellule j de sortie

O_t : réponse désirée

L'apprentissage se déroule de la manière suivante :

- On présente nos données au perceptron dans un ordre aléatoire
- S'il y'a des erreurs, les poids seront corrigés selon la règle de Widrow-Hoff et on retourne à l'étape perceptron.
- Si ce n'est pas le cas on considère que notre perceptron a bien appris.

Plutôt que d'obtenir un perceptron qui classifie correctement les exemples, l'apprentissage par descente de gradient s'intéresse aux calculs d'erreur et de la minimisation de cette dernière.
Pour introduire cette notion d'erreur, on utilise des poids réels et on élimine la notion de seuil, ce qui signifie que la sortie sera donc réelle.

L'algorithme de Descente de gradient est présenté donc comme suit :

Entrée : n poids reliant les n informations au neurone ayant des valeurs quelconques.
N exemples (I, O) où I est un vecteur à n composantes I_i, chacune représentant une information de cet exemple.

Sortie : les n poids modifiés
Pour 1 <= i <= n
dwi = 0
Fin pour

Pour tout exemple e = (I, O) Calculer la sortie Sk du neurone
Pour 1 <= i <= n
di = dwi + α.(O- Sk).Oi
Fin pour
Fin pour

Pour 1 <= i <= n
$dw_i = w_i + dw_i$
Fin pour

b) Limitation du perceptron

Le perceptron linéaire a seuil divise l'espace en deux qui sont délimités par un hyperplan, cependant les problèmes de classification ne sont pas forcément linéairement séparables et c'est de la d'où viens sa faiblesse, l'exemple le plus simple pour illustrer ce problème est l'opérateur logique XOR où on peut pas séparer les cercles rouges des losanges noirs avec une seule ligne.

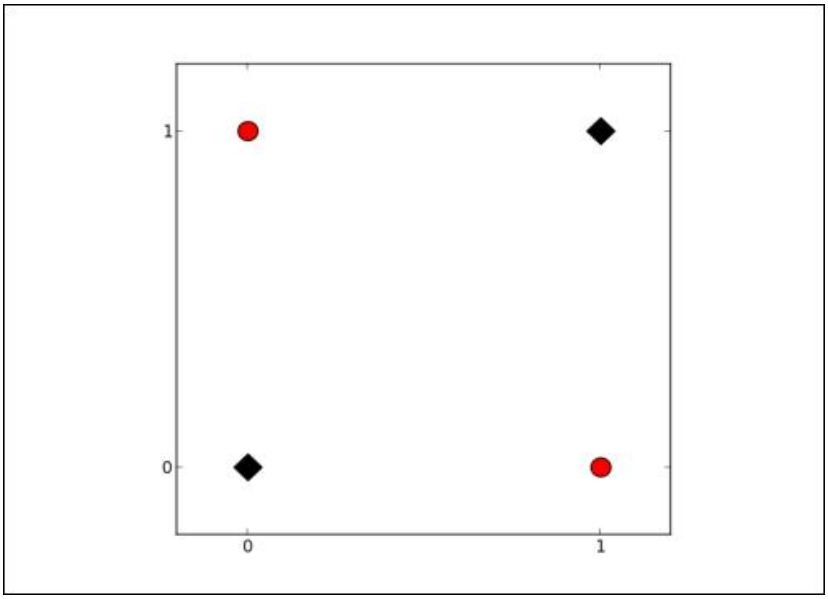


Figure 5 : Représentation de la fonction XOR

Une approche initiale consiste à utiliser plus d'un perceptron, chacun mis en place pour identifier de petites sections linéairement séparables des entrées, puis en combinant leurs sorties dans un autre perceptron, ce qui produirait une indication finale de la classe à laquelle appartient l'entrée[9].

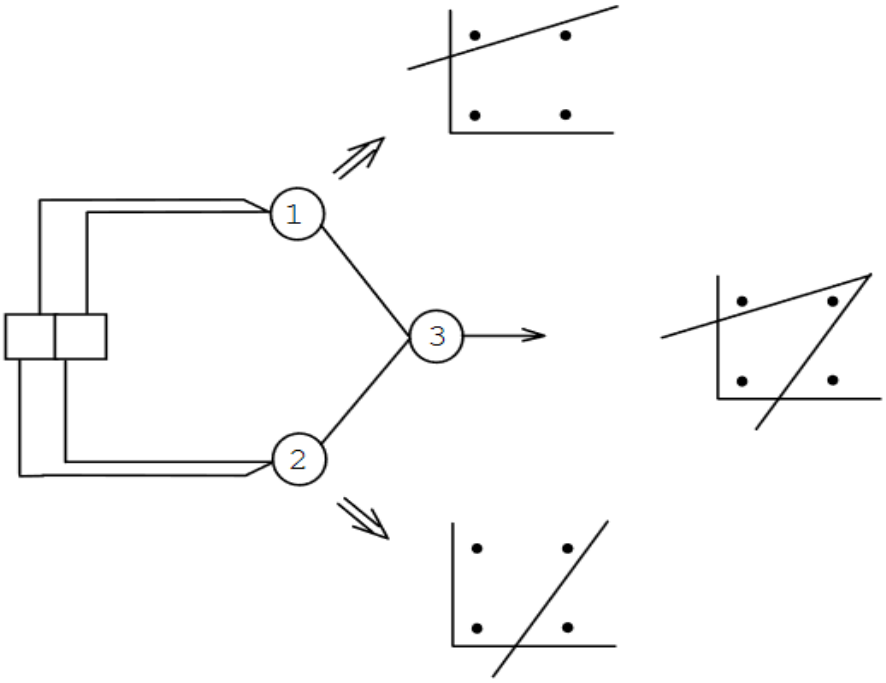


Figure 6 : Combinaison de sorties pour résoudre le problème XOR

3. Perceptrons multicouches

Le perceptron multicouche est un réseau de neurones formels organisé en plusieurs couches au sein desquelles une information circule de la couche d'entrée vers la couche de sortie uniquement.

Chaque couche est constituée d'un nombre variable de neurones, les neurones de la dernière couche sont des neurones de sorties du système global.

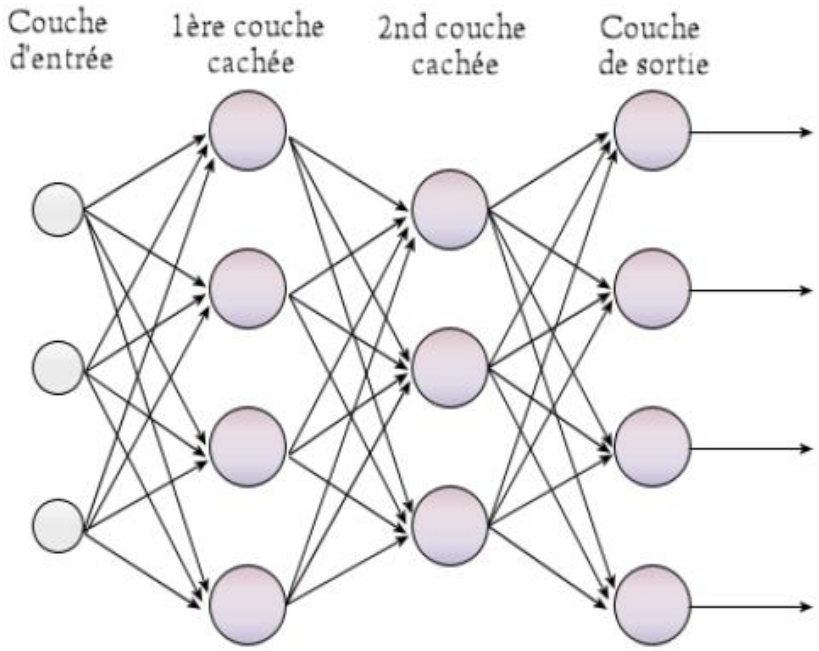


Figure 7 : Modélisation d'un perceptron multicouches avec deux couches cachées

Comme on a vu précédemment que les réseaux de neurones n'étaient pas capables de résoudre les problèmes non linéairement séparables, le perceptron multicouches répond à cette limitation en introduisant le principe de rétropropagation du gradient de l'erreur dans les systèmes multicouches[10].

a) Fonction de transfert

C'est une fonction mathématique qui doit renvoyer un réel dans l'intervalle [0,1] comme la fonction Heaviside qui renvoie tout le temps 1 si le signal en entrée est positif et 0 dans le cas contraire.

Les fonctions d'activations ou de transfert sont utilisés selon leurs caractéristiques :

Non-linéarité : qui permet de rendre un réseau neuronal multicouches équivalent à un réseau neuronal monocouche.

Différentiable : Cette caractéristique permet de créer des optimisations basées sur les gradients.

Monotone : Lorsque la fonction est monotone, la surface d'erreur associée avec un modèle monocouche est certifiée convexe.

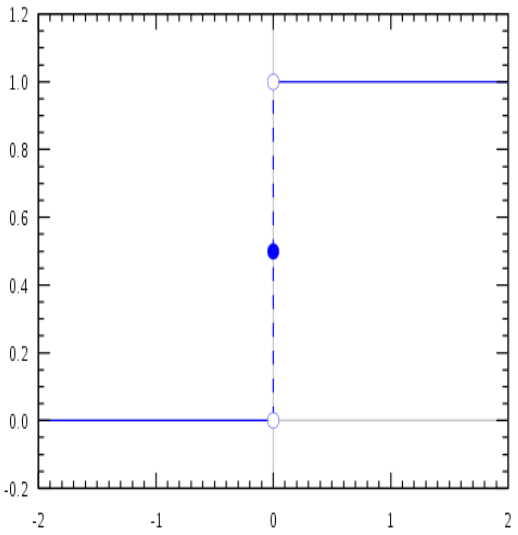


Figure 8 : Graphe de la fonction Heaviside

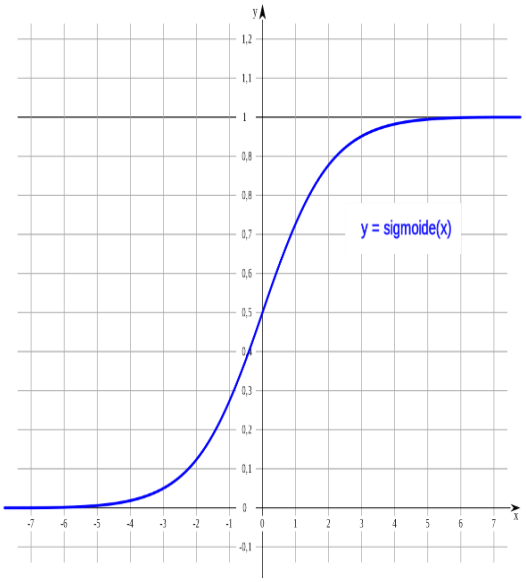


Figure 9 : Graphe de la fonction Sigmoïde

Contrairement à la Heaviside, la sigmoïde est une fonction qui est partout différentiable et qui renvoie non seulement 1 et 0 mais aussi des valeurs dans cet intervalle.

b) Principe de rétropropagation

Dans le perceptron multicouche à rétropropagation, les neurones d'une couche sont reliés à la totalité des neurones des couches adjacentes. Ces liaisons sont soumises à un coefficient altérant l'effet de l'information sur le neurone de destination.

Ainsi, le poids de chacune de ces liaisons est l'élément clef du fonctionnement du réseau : la mise en place d'un perceptron multicouches pour résoudre un problème passe donc par la détermination des meilleurs poids applicables à chacune des connexions inter-neuronales. Ici, cette détermination s'effectue à travers l'algorithme de rétropropagation.

Le taux d'apprentissage par rétropropagation du gradient doit être bien choisis, car pour un taux d'apprentissage rapide on risque d'avoir des instabilités dans le réseau, un taux trop lent peut ramener le réseau à être bloqué dans un minimum local et l'apprentissage par inertie qui permet de conserver les informations relatives au dernier apprentissage pour en tenir compte dans l'apprentissage courant.

4. L'apprentissage profond

L'apprentissage profond fait partie d'une famille de méthodes d'apprentissage automatique fondées sur l'apprentissage de modèles de données.

C'est un ensemble de méthodes tentant de modéliser avec un haut niveau d'abstraction des données grâce à des architectures articulées de différentes transformations non linéaires.

Ces techniques ont permis des progrès importants et rapides dans les domaines de l'analyse du signal sonore ou visuel et notamment de la reconnaissance faciale, de la reconnaissance vocale et du traitement automatique du langage.

Il fait référence à un type d'intelligence artificielle particulier utilisant le réseau de neurones et certains modèles d'algorithmes particuliers afin de générer des modèles intelligents grâce à l'apprentissage.

a) Les réseaux de neurones profonds

Ces réseaux de neurones multicouches peuvent comporter des millions de neurones, répartis en plusieurs dizaines de couches. Ils sont utilisés en apprentissage profond pour concevoir des mécanismes d'apprentissage supervisés et non supervisés.

Dans ces architectures mathématiques, chaque neurone effectue des calculs simples, mais les données d'entrées passent à travers plusieurs couches de calcul avant de produire une sortie. Les résultats de la première couche de neurones servent d'entrée au calcul de la couche suivante et ainsi de suite. Il est possible de jouer sur les différents paramètres de l'architecture du réseau : le nombre de couches, le type de chaque couche, le nombre de neurones qui composent chaque couche.

Le raisonnement c'est que plus on augmente le nombre de couches, plus les réseaux de neurones apprennent des choses compliquées, abstraites, correspondant de plus en plus à la manière dont un humain raisonne.

Il est néanmoins difficile de mettre au point des mécanismes d'apprentissage efficaces pour chacune des couches intermédiaires (couches profondes ou couches cachées). La solution consiste à allier l'apprentissage supervisé (comme cela est fait pour un réseau à deux couches) et l'apprentissage non-supervisé, utilisant des exemples dont la valeur de sortie est inconnue.

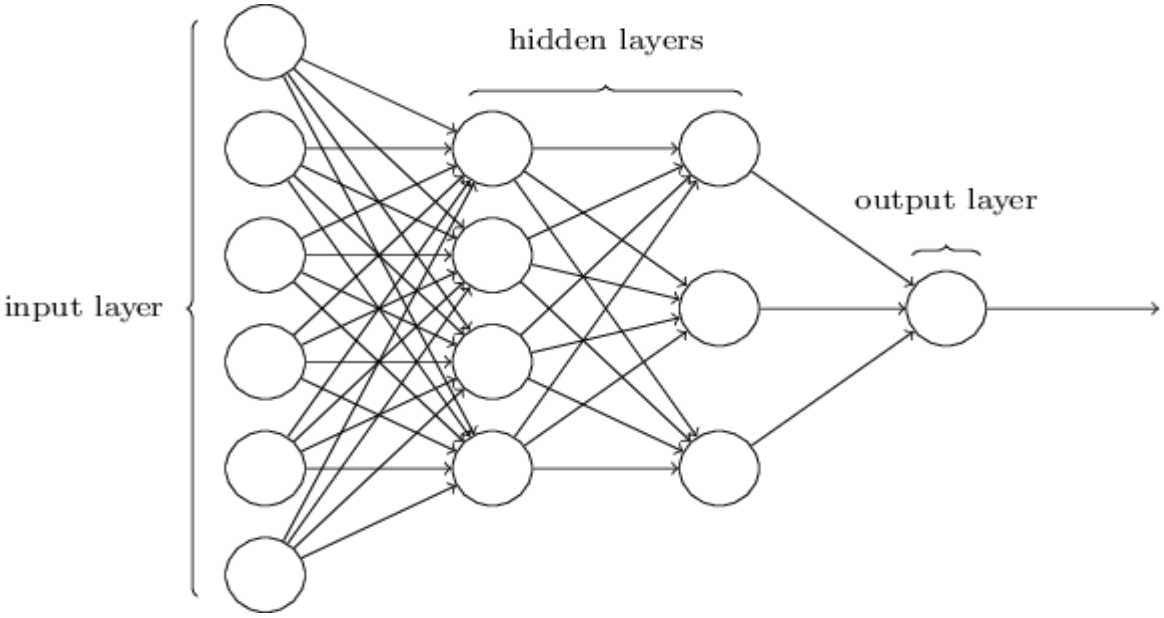


Figure 10 : Architecture générale d'un réseau de neurones profond

Le premier réseau de neurones profond a été publié par Alexey Girgoverich Ivankhnenko en 1965, il a été constitué avec 8 couches cachées en utilisant l'algorithme d'apprentissage appelé GMDH (Group Method of Data Handling)[11].

b) Les réseaux de neurones convolutifs

En apprentissage automatique, un réseau de neurones convolutifs est un type de réseau de neurones artificiels acycliques (feed-forward), dans lequel le motif de connexion entre les neurones est inspiré par le cortex visuel des animaux.

Les neurones de cette région du cerveau sont arrangés de sorte qu'ils correspondent à des régions qui se chevauchent lors du pavage du champ visuel.

Leur fonctionnement est inspiré par les processus biologiques, ils consistent en un empilage multicouche de perceptrons, dont le but est de prétraiter de petites quantités d'informations.

Les réseaux neuronaux convolutifs ont de larges applications dans la reconnaissance d'image et vidéo, les systèmes de recommandation et le traitement de langage naturel.

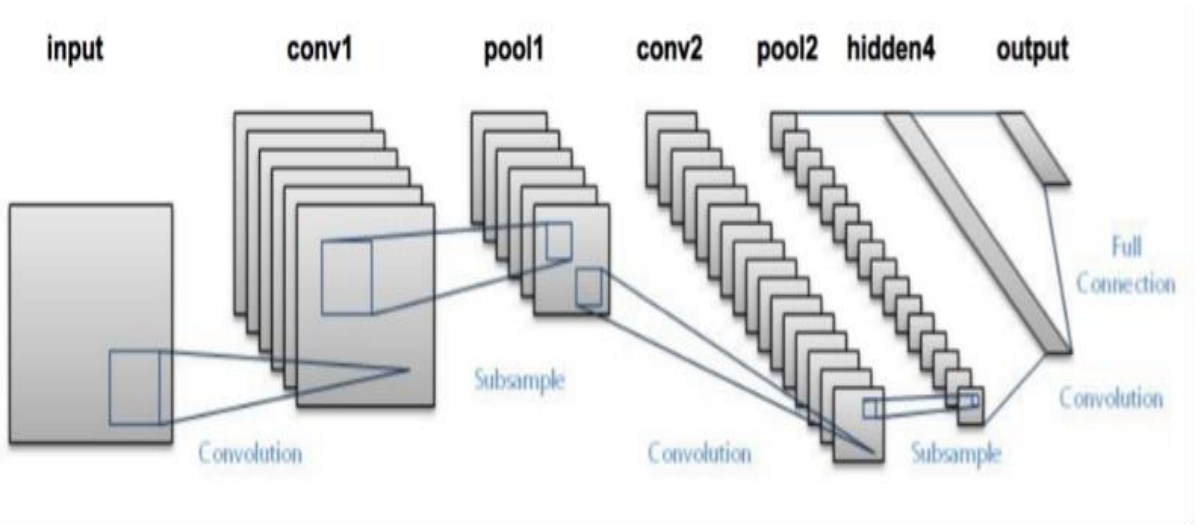


Figure 11 : Architecture d'un réseau de neurones à convolutions

Séquencement

a. Convolution

La couche de convolution permet de calculer la correspondance entre une caractéristique et une sous-partie de l'image en passant un filtre sur cette dernière et c'est comme ça comment le CNN cherche à trouver les caractéristiques dans une image (Spectrogramme dans notre cas).

L'étape suivante est de répéter le processus complet de convolution pour chacune des autres caractéristiques existantes. Le résultat est un ensemble d'images filtrées, dont chaque image correspond à un filtre particulier.

b. Pooling

L'étape de pooling permet de prendre une large image et d'en réduire la taille tout en préservant les informations les plus importantes qu'elle contient. Il suffit de faire glisser une petite fenêtre pas à pas sur toutes les parties de l'image et de prendre la valeur maximum de cette fenêtre à chaque pas.

Comme le CNN garde à chaque pas la valeur maximale contenue dans la fenêtre, il préserve les meilleures caractéristiques de cette fenêtre. Cela signifie qu'il ne se préoccupe pas vraiment d'où a été extraite la caractéristique dans la fenêtre.

La sortie aura le même nombre d'images mais chaque image aura un nombre inférieur de pixels ce qui permet ainsi de diminuer la charge de calculs.

c. Unités rectifiées linéaires

Chaque fois qu'il y a une valeur négative dans un pixel, on la remplace par un 0. Ainsi, on permet au CNN de rester en bonne santé (mathématiquement parlant) en empêchant les valeurs apprises de rester coincer autour de 0 ou d'exploser vers l'infinie.

Le résultat d'une couche d'unités rectifiés linéaires est de la même taille que ce qui lui est passé en entrée, avec simplement toutes les valeurs négatives éliminées.

d. Apprentissage profond

Ce qu'on donne en entrée à chaque couche ressemble à ce qu'on obtient en sortie ce qui permet d'ajouter les couches les unes après les autres ainsi les couches inferieures présentent des aspects simples de l'image alors que les couches supérieures présentent les aspects les plus complexes de l'image tels que des formes et des patterns qui ont tendance à être facilement reconnaissables.

En effet, chaque couche additionnelle laisse le réseau apprendre des combinaisons chaque fois plus complexes de caractéristiques qui aident à améliorer la prise de décision.

e. Couches entièrement connectées

Ces couches sont les principaux blocs de construction des réseaux de neurones traditionnels. Au lieu de traiter les entrées comme des tableaux de 2 dimensions, ils sont traités en tant que liste de votes appelée aussi poids ou force de connexion entre chaque valeur et chaque catégorie.

Lorsqu'une nouvelle image est présentée au CNN, elle passe à travers les couches inférieures jusqu'à la couche entièrement connectée. L'élection va ensuite avoir lieu et la solution avec le plus de vote gagne et est déclarée comme la catégorie de l'image.

f. Rétropropagation

Le nombre d'erreurs que l'on fait dans notre classification nous renseigne ensuite sur la qualité de nos caractéristiques et de nos poids. Les caractéristiques et poids peuvent ensuite être ajustés de manière à réduire l'erreur. La nouvelle valeur de l'erreur de notre réseau est recalculée, à chaque fois. Quel que soit l'ajustement fait, si l'erreur diminue, l'ajustement est conservé.

Comme on a parlé de filtres, de couches cachées et de nombre de neurones, une étape principale doit être prise en compte en vue de définir ces derniers qu'on appelle les hyperparamètres du CNN.

Dans la plupart des cas c'est la connaissance accumulée par la communauté scientifique, avec certains écarts occasionnels qui permettent des améliorations surprenantes en performances.

c) Réseaux de croyances profonds

De manière générale les réseaux de croyances profonds sont des réseaux de neurones générateurs qui empilent des machines de Boltzmann restreintes(RBM), ils sont des modèles probabilistes composés de plusieurs couches stochastiques de neurones cachés avec des connexions entre les couches du réseau[12].

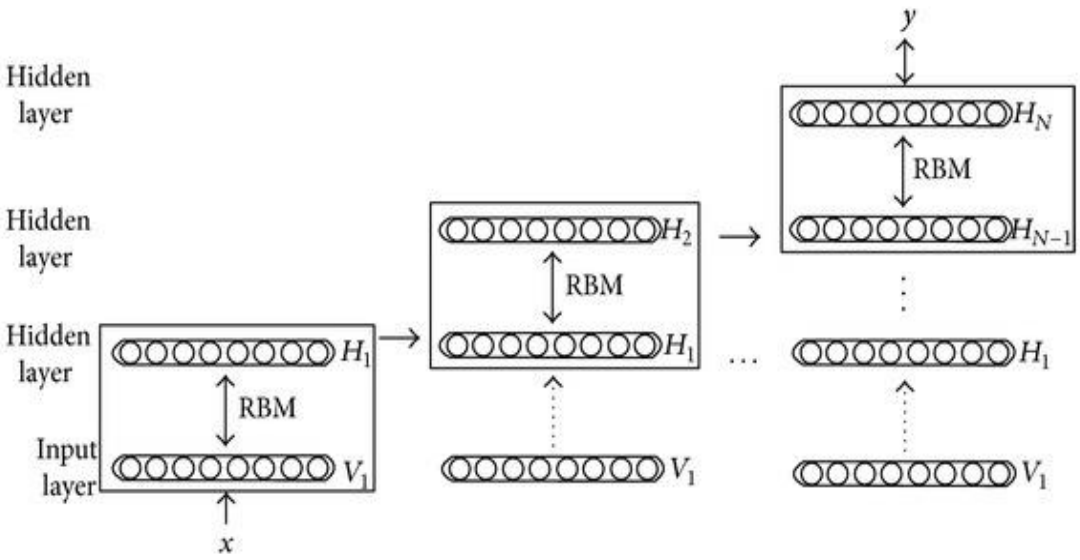


Figure 12 : Architecture d'un réseau de croyances profonds

Quand l'apprentissage d'un ensemble de données est en mode non-supervisé, le réseau de croyance profond peut apprendre la reconstruction des données de l'entrée. Donc, les couches actent comme des extracteurs de caractéristiques dans l'entrée. Après cette étape d'apprentissage, le réseau de croyance profond peut s'entrainer d'une manière supervisée pour faire la classification.

Les réseaux de neurones profonds ont le pouvoir de classifier des objets qui sont plus complexes et abstraits alors que la plupart des fonctions numériques peuvent être approchées par seulement des réseaux de neurones à une seule couche cachée et qui sont moins difficiles à apprendre.

IV. Les réseaux de neurones profonds en R.A.L

1. Introduction

Dans cette partie on va présenter la structure moderne et générale d'un système de reconnaissance automatique de locuteur avec les différentes techniques d'apprentissage profonds.

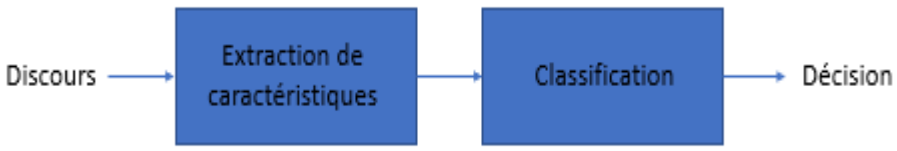
2. Les réseaux de neurones profonds en R.A.L

Dans la reconnaissance automatique du locuteur, un algorithme génère une hypothèse concernant l'identité ou l'authenticité du locuteur. Lorsque la tâche consiste à identifier la personne qui parle plutôt que ce que la personne dit, le signal vocal doit être traité pour extraire les mesures de la variabilité du locuteur au lieu des caractéristiques segmentales.

Les deux tâches principales dans la reconnaissance du locuteur sont l'identification du locuteur (ID) et la vérification du locuteur. L'identification des locuteurs concerne une situation dans laquelle la personne doit être identifiée comme faisant partie d'un ensemble de personnes en utilisant ses échantillons de voix.

L'objectif de la vérification du locuteur est de vérifier l'identité revendiquée de ce locuteur sur la base des échantillons de voix de ce locuteur seul.
Pour la reconnaissance du locuteur, les aspects acoustiques de ce qui caractérise les différences entre les voix sont obscurs et difficiles à distinguer des aspects du signal qui affectent la reconnaissance des segments.

La reconnaissance des locuteurs comprend deux étapes, à savoir l'extraction des caractéristiques et la classification.



a) Extraction de caractéristiques

L'extraction de caractéristiques est associée à l'obtention des motifs caractéristiques du signal qui sont représentatifs pour le locuteur en question.

Les paramètres ou les caractéristiques utilisés dans la reconnaissance du locuteur sont une transformation du signal vocal en une représentation acoustique compacte qui contient des informations utiles pour l'identification du locuteur. Cela est souvent fait en utilisant l'analyse prédictive linéaire (LP) à court terme[13].

Le classificateur utilise ses fonctions pour fournir une décision qui concerne l'identité du locuteur ou pour vérifier cette dernière revendiquée du locuteur.

Les coefficients LP sont convertis en vecteurs de caractéristiques, idéalement, il est souhaitable que les vecteurs caractéristiques présentent une grande similitude pour un locuteur particulier et montrent simultanément beaucoup de variabilité pour différents locuteurs. Cela permet une distinction facile entre les différents intervenants.

De plus, les vecteurs caractéristiques devraient être robustes à une grande variété de conditions (telles que les différents milieux de bruit et de transmission) en montrant peu de variabilité lorsque les conditions environnementales sont modifiées.

Une étude comparative des caractéristiques a révélé que le LP(prédiction linéaire) est le meilleur pour la reconnaissance du locuteur[14].

Cependant, le cepstre LP n'est pas robuste aux effets de canal et de bruit. Récemment, des caractéristiques telles que le cepstre à composante adaptative pondérée (ACW)[15] et le cepstre post-filtre (PFL)[16] se sont révélées plus robustes au canal et au bruit.

On parle aussi des MFCC [17] (Mel Frequency Cepstral Coefficients) et qui sont largement utilisés en reconnaissance automatique de la parole et en reconnaissance automatique du locuteur. Ils sont introduits par Davis et Mermelstrein en 1980[18], et depuis cette date ils restent l'état de l'art de l'extraction des caractéristiques.

L'algorithme des MFCC est simple :

- En premier lieu on découpe le signal en petite trames.
- On calcule le périodogramme du spectre de puissance pour chaque trame.
- On applique le Mel Filterbank aux spectres de puissance et on somme l'énergie dans chaque filtre.
- On applique le logarithme de toutes les énergies filterbank.
- On applique la transformée en cosinus discrète des énergies de log filterbank.
- On obtient plusieurs coefficients de DCT dont on peut garder de 2 à 13 coefficients.
-

b) Classification : Structure générale

Comme mentionné précédemment, l'analyse LP de la parole et l'extraction subséquente des caractéristiques conduit à un ensemble de vecteurs de caractéristiques. Le classifieur se compose des différents modèles de locuteurs et de la logique de décision.

Son fonctionnement est constitué de deux étapes importantes. Dans la phase d'apprentissage, des vecteurs de caractéristiques sont utilisés pour obtenir les modèles de locuteurs M. Pour chaque locuteur, un modèle différent est obtenu à partir de son discours.

Dans la phase de test, les vecteurs de caractéristiques d'un locuteur inconnu sont d'abord calculés. Pour l'identification des locuteurs, les vecteurs caractéristiques sont comparés avec chacun des modèles de locuteurs M pour obtenir les scores de 1 jusqu'à M, ces derniers sont utilisés pour la prise de décision.

Il existe deux catégories principales de modèles, à savoir ceux qui sont formés avec des algorithmes d'entraînement non supervisés et ceux qui sont formés avec des algorithmes d'entraînement supervisés.
Le terme « supervision » signifie que les données fournies à l'algorithme d'apprentissage du modèle contiennent des informations de classe via une étiquette. Les algorithmes non supervisés ne nécessitent aucune information concernant l'appartenance à une classe pour les données d'apprentissage.

En termes de reconnaissance du locuteur, les algorithmes de formation non supervisés n'utilisent que les données d'un locuteur spécifique pour créer un modèle.

Les exemples d'approches de modélisation non supervisées comprennent l'algorithme du plus proche voisin (NN), la quantification vectorielle (VQ), les modèles de mélange gaussien (GMM) et les modèles de Markov cachés (HMM).

Les algorithmes d'apprentissage supervisés attribuent des étiquettes différentes aux différents locuteurs d'une population. Des exemples d'algorithmes d'entraînement supervisé comprennent les perceptrons multicouches (MLP), les fonctions de base radiales (RBF), les arbres de décision et les réseaux d'arbres neuronaux (NTN).

Les algorithmes d'entraînement supervisé ont un avantage sur les algorithmes d'entraînement non supervisés dans le sens qu'ils permettent de mieux saisir les différences entre un locuteur particulier et d'autres locuteurs de la population. Cependant, la quantité de données d'apprentissage et, par conséquent, l'effort de calcul dans la dérivation d'un modèle de locuteur peut être supérieur à celui des algorithmes d'entraînement non supervisés.

Les réseaux neuronaux à temporisation (TDNN) [19]sont des réseaux neuronaux feed-forward à couches qui incorporent des versions retardées des entrées, de manière à apprendre les corrélations temporelles des données d'entrée. Les TDNN sont la contrepartie supervisée du HMM en ce sens qu'ils tentent de tirer parti des informations temporelles.

Les TDNN ont été considérés pour une identification de locuteur indépendante du texte [20]. Ici, les TDNN ont été évalués pour une population de 20 locuteurs (10 mâles et 10 femelles) et ont démontré un taux d'identification de 98%.

3. Conclusion

Dans cette partie nous avons mis en avant l'importance des réseaux de neurones profonds pour la reconnaissance automatique de locuteur qui se résume dans l'amélioration significative de la précision de la R.A.L et que la majorité des résultats montrent que ces derniers peuvent très bien détecter les locuteurs, sauf dans les cas où il y a un chevauchement important dans l'accent et le ton du locuteur[21].

Chapitre 3

Identification automatique du locuteur utilisant les spectrogrammes et CNN

I. Les spectrogrammes

Le spectrogramme est un diagramme représentant le spectre d'un phénomène periodique, associant à chaque fréquence une intensité ou une puissance.

L'idée derrière l'utilisation de ces spectrogrammes est de pouvoir identifier les différents locuteurs en utilisant l'image comme modèle au lieu de travailler avec les caractéristiques classiques qui donnent de bons résultats.

Le spectrogramme sonore ou sonagramme est une représentation visuelle d'un signal acoustique qui nous permet de voir comment le spectre d'un signal de parole évolue avec le temps.

En général, Le sonagramme permet de représenter dans un seul diagramme de deux dimensions trois paramètres qui sont : le temps et qui est présenté souvent sur l'axe horizontal, la fréquence qui est présentée sur le deuxième axe et la puissance sonore qui est représentée par la couleur des points de diagramme qui correspondent à un instant ou une durée suffisante permettant d'analyser la fréquence du signal.

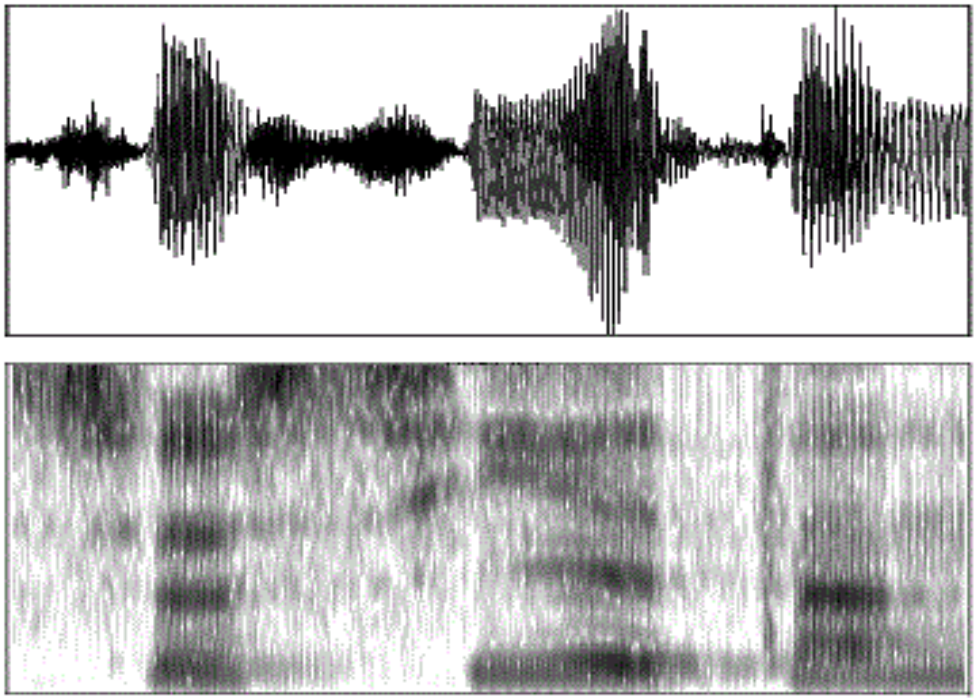


Figure 13 : Spectrogramme d'un signal vocal

Le concept de l'utilisation de spectrogrammes pour l'identification des locuteurs existe depuis des décennies. L'une des premières tentatives de reconnaissance automatique des locuteurs a été faite dans les années 1960 [22] en utilisant des bancs de filtres et en corrélant deux spectrogrammes numériques pour une mesure de similarité[23].

Plusieurs techniques ont été développées pour le « Pattern Matching » dans des signaux de parole en utilisant les spectrogrammes dont on cite parmi eux celle utilisé dans [24] qui décrit la bande de spectrogramme comme un groupe et sa valeur de pixel moyenne comme le centroïde de ce dernier. Par conséquent, étant donné l'énoncé d'un mot connu par un locuteur inconnu on recherche l'échantillon de base de données de ce mot particulier avec des centroïdes de groupes ordonnés ayant la distance euclidienne la plus proche de celle du locuteur inconnu.

La génération des spectrogrammes consiste en général a appliqué une transformation de fourrier avec une fenêtre que l'on fait glisser tout au long de la durée du son enregistré électroniquement[25].

Les spectrogrammes sont utilisés en acoustique pour identifier des sons, comme des cris d'animaux [26] et des sons d'instruments musicaux ou même les sons d'environnement[27]. Ils servent en phonétique à l'étude et à la comparaison des phonèmes[28].

La structure temporelle du signal de la parole est une direction non seulement simple avec un ensemble d'étapes qui capturent les aspects fondamentaux de la représentation du signal vocal, mais aussi prometteuse pour construire un système de discrimination de locuteurs robustes avec un pourcentage de précision élevé.

II. Approche proposée

Nous allons présenter maintenant la base de données de locuteurs sur laquelle on a travaillé ainsi que la démarche suivie en vue de présenter et analyser les résultats.

1. Base de données

La base de donnees « Thuyg-20 SRE » se constitue de plus de 20 heures de signaux vocaux en totale qui ont été générés auprès de 371 locuteurs[29].

Les locuteurs étaient tous des étudiants âgés de 19 à 28 ans et originaires de 30 pays différents, les phrases choisies dans les signaux vocaux sont extraites de domaines généraux, notamment des romans, des journaux et divers types de livres.

« Thuyg-20 SRE » est basé sur la base de donnees utilisé « Thuyg-20 » qui sont but original était pour la reconnaissance de la parole mais qui peut être aussi utilisé pour la reconnaissance du locuteur, l'ensemble des enregistrements et de tests ont été générés par le même ensemble de personne qui se constitue de 153 locuteurs où chaque énoncé d'enregistrement est composé de 30 secondes et chaque énoncé de test est composé de 10 secondes.

Les signaux ont été enregistrés dans un bureau en plein silence par le même microphone, La fréquence d'échantillonnage est de 16 000 Hz et la taille de l'échantillon est de 16 bits et la base de données a été enregistrée de janvier à septembre en 2012.

2. Processus suivi

a) Prétraitement de données

La première étape consiste à détecter et extraire les segments de silence de tous les signaux audios, et pour faire plusieurs techniques peuvent être appliqué[30] où on fait généralement appel à des modèles de probabilité gaussiens pour représenter les paramètres relatives aux segments silence et non-silence.

Ces modèles permettent de détecter les segments de silence en assignant les observations de paramètres correspondant à ces segments à la classe silence, après on calcule la vraisemblance des segments avec cette classe et on fix le seuil.

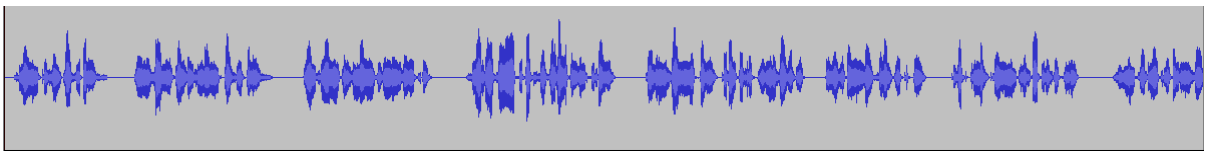


Figure 14 : Signal vocal du locuteur n°1 avant l'extraction du silence

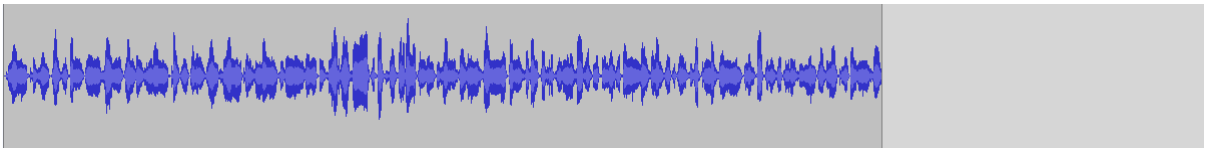


Figure 15 : Signal vocal du locuteur n°1 après l'extraction du silence

L'extraction du silence est une étape primordiale où on peut même dire que le silence représente du bruit car sans cette dernière on aura une autre caractéristique qui sera générée dans les images spectrogrammes à partir des signaux audios, la chose qui va diminuer les performances et surtout la précision du système de reconnaissance automatique du locuteur.

Ensuite, ce qu'on a obtenu comme signal après l'extraction de silence sera divisé en plusieurs signaux vocaux d'une seconde et c'est là que l'étape de génération de spectrogrammes fait son apparition.

b) Génération de spectrogramme

Cette étape consiste a généré les spectrogrammes de tous les signaux vocaux d'une seconde pour chaque locuteur de la base de données.

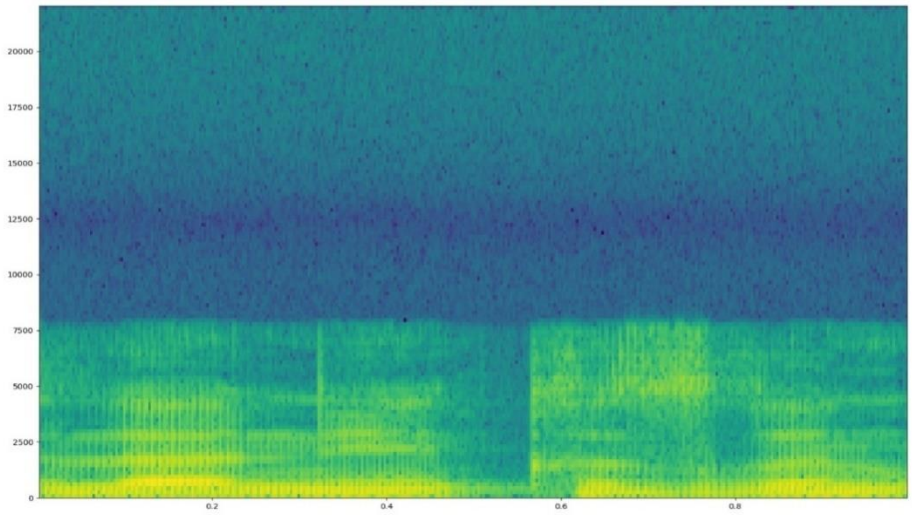


Figure 16 : Spectrogramme d'une seconde du locuteur n°1

c) Architecture du CNN

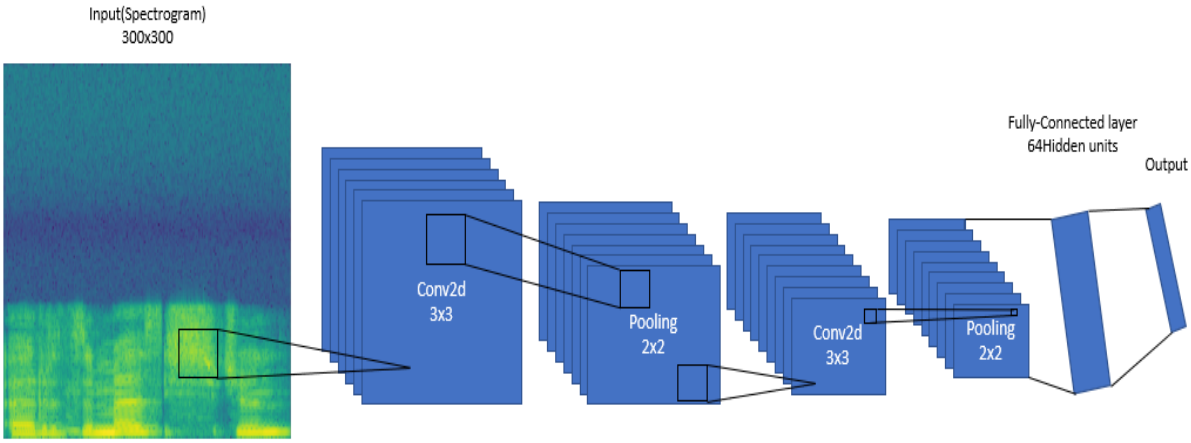


Figure 17 : Architecture du CNN employé

On passe une image en entrée à la première couche convolutionelle où les filtres dans cette dernière permettent d'extraire les caractéristiques les plus pertinentes de l'image spectrogramme.

Chaque filtre utilisé doit donner une caractéristique différente pour aider à la bonne prédiction de classes.

Les couches Pooling sont ensuite ajoutées pour réduire davantage le nombre de paramètres.

Une autre couche de convolution et de pooling sont ajoutées avant la prédiction. Au fur et à mesure que nous pénétrons plus profondément dans le réseau, des caractéristiques plus spécifiques sont extraites.

La couche de sortie dans un CNN comme mentionné précédemment est une couche entièrement connectée, où l'entrée des autres couches est aplatie et envoyée de manière à transformer la sortie en nombre de classes souhaité par le réseau.

Une fonction de perte est définie dans la couche de sortie entièrement connectée pour calculer l'erreur commise.

L'erreur est ensuite rétro-propagée pour mettre à jour les poids et les valeurs de biais.

d) L'identification automatique du locuteur

Dans cette partie on suppose que l'identification est faite sur un ensemble fermé c-à-d que le locuteur recherché sera toujours proche des modèles de références et par suite ce dernier sera toujours reconnu parmi un ensemble fini de 153 locuteurs.

Après qu'on a extrait toutes les caractéristiques présentes dans les images spectrogrammes et qu'on a généré un modèle pour chaque locuteur et fait l'apprentissage de ce dernier, l'étape suivante sera celle où on test les performances de notre réseau de convolutions.

L'ensemble de donnees de test de chaque locuteur passera donc par les mêmes étapes de pré-traitements et c'est ainsi qu'on va obtenir nos donnees tests qui auront la même forme que ceux d'apprentissage.

La sortie de notre CNN est une dernière couche comportant un neurone par locuteur ainsi on obtient des valeurs numériques qui sont généralement normalisées entre 0 et 1, de somme 1, pour produire une distribution de probabilité sur les locuteurs.

Par suite la classe qui proposera la probabilité la plus élevé sera considéré comme classe du locuteur présenté par ses données de test.

L'étape de l'identification est considérée comme délicate parce qu'il faut une grande quantité d'enregistrements dans des conditions acoustiques à peu près similaires afin d'avoir une bonne identification et d'être plus fiable, plus il y'a une différence d'un enregistrement a un autre et plus les performances sont dégradées.

e) Analyse des résultats

Les résultats obtenus à partir de la tâche d'identification sont intéressant pour une première étude basée sur les spectrogrammes qui est une idée originale,76,15% comme taux d'identification.

Certes les résultats sont moins performants à ce qu'on peut avoir avec les GMM/SVM mais ce taux obtenu explique fortement que les spectrogrammes ne peuvent pas être utilisées uniquement pour la parole mais aussi pour la reconnaissance du locuteur, la chose qui nous mène à se poser des questions sur la totalité de l'approche proposée.

Un autre travail [31] fait sur une base de données construite ,pour pallier aux problèmes des base de données existantes qui sont de petites taille, basée sur les MFCC pour l'extraction de caractéristiques, GMM-UBM et i-Vector/PLDA pour la modélisation du locuteur où les signaux vocaux sont numérisés à 16KHz avec une résolution de 16bits.

On a remarqué que les résultats obtenus pour le système basé sur GMM-UBM atteint de meilleurs performances que celui basé sur i-Vector/PLDA pour les deux taches et principalement l'identification automatique du locuteur (99% comme taux d'identification pour GMM-UBM).

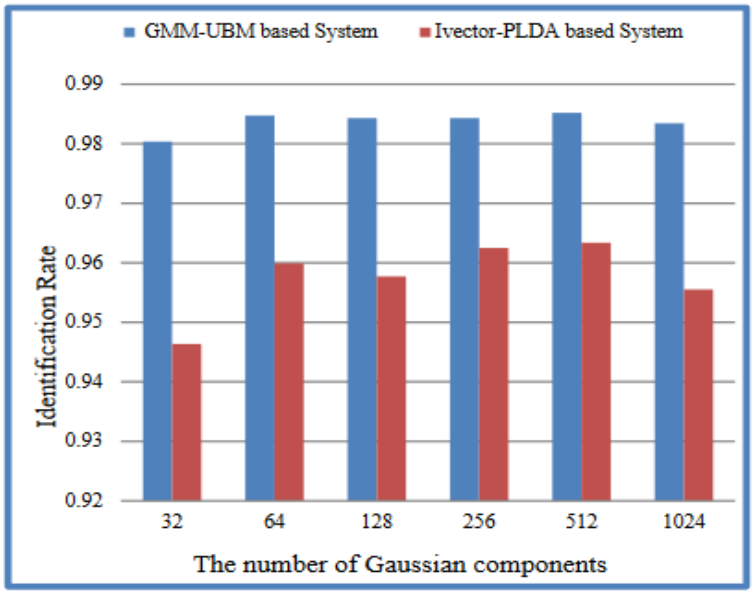


Figure 18 : Taux d'identifications pour différentes taille d'UBM

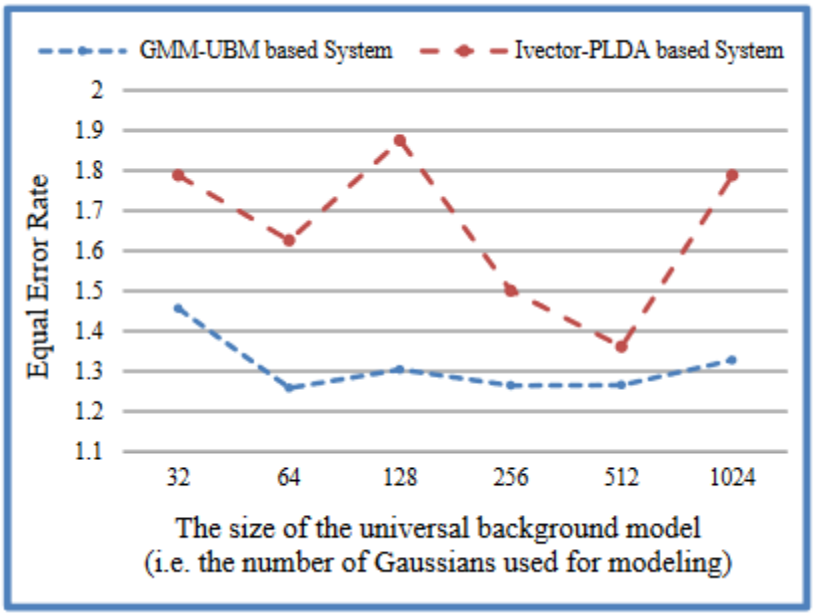


Figure 19 : Courbe d'erreur pour la tache de vérification automatique du locuteur

La courbe d'erreur confirme les résultats obtenus avec une erreur minime pour GMM-UBM par rapport à i-Vector/PLDA, on a remarqué aussi que le système basé sur GMM-UBM est moins sensible au choix de la taille d'UBM comparé au système i-Vector/PLDA.

Donc on peut se poser directement la question sur la taille de données, est ce que les données de notre système sont suffisant pour faire une meilleure identification ?

Est-ce que nos données d'apprentissage présentent des informations inutiles et des caractéristiques non pertinentes qui sont la principale cause de la dégradation des performances ?

Est-ce qu'un changement des hyperparamètres va pousser notre CNN à donner de meilleurs résultats que ceux obtenus avant ?

Est-ce que l'architecture du CNN basique qui fait baisser les performances de notre réseau ?

Conclusion

Ce stage de fin d'études m'a bien introduit dans le domaine de reconnaissance automatique de locuteur et m'a permis d'avoir une très belle expérience dans le domaine du Deep Learning auprès des gens ayant un grand savoir et une très bonne connaissance dans ce domaine.

Dans ce présent rapport nous avons présenté notre approche proposée basée sur les spectrogrammes pour faire une reconnaissance du locuteur et principalement la tache IAL, c'est une idée très originale qui peut devenir promettante dans le futur avec quelques modifications.

Les travaux de recherches présentés dans ce rapport de fin de stage peuvent être amélioré et poursuite de plusieurs façons, le premier travail qu'il faut faire c'est de pallier à l'insuffisance de donnees d'apprentissage en proposant une étape de préapprentissage permettant d'augmenter la taille des données d'apprentissage, ensuite plusieurs perspectives permettant de les améliorer peuvent être employées en faisant une étude sur la numérisation des signaux vocaux pour 10KHZ, 12KHZ, 16KHZ et 20KHZ pour but de garder que les informations pertinentes contenues dans les signaux vocaux ,en changeant l'architecture du CNN basique qu'on a utilisée par une autre plus personnalisée et en modifiant par exemple les hyperparamètres du réseau de neurones (nombres de nœuds, nombres de couches cachées, taille des filtres,…).

Dans ce cadre, nous pouvons penser à employer une méthode hybride entre les modèles générés par les spectrogrammes et les MFCCs où utiliser l'un d'eux comme un système de support pour le deuxième ce qui peut donner des résultats très intéressants.

Références

[1] H. Beigi, *Fundamentals of Speaker Recognition*. 2011.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, Jan. 2000.

[3] J. Keshet, “Automatic Speech and Speaker Recognition : Large Margin and Kernel Methods edited by.”

[4] W. M. Campbell, D. E. Sturim, and D. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, 2006.

[5] Jaquier, “The robustness of the speech signal,” 2008. [Online]. Available: http://theses.univ-lyon2.fr/documents/getpart.php?id=lyon2.2008.jacquier_c&part=148480. [Accessed: 27-Jun-2018].

[6] J. Y. Lettvin, H. R. Maturana, H. R. Maturana, W. S. Mcculloch, and W. H. Pitts, “What the Frog’s Eye Tells the Frog’s Brain,” *Proc. IRE*, vol. 47, no. 11, pp. 1940–1951, 1959.

[7] F. Attneave, M. B., and D. O. Hebb, “The Organization of Behavior; A Neuropsychological Theory,” *Am. J. Psychol.*, vol. 63, no. 4, p. 633, 1950.

[8] F. Rosenblatt, “The Perceptron - A Perceiving and Recognizing Automaton,” *Report 85, Cornell Aeronautical Laboratory*. pp. 460–1, 1957.

[9] Cornelia Denz, “Optical neural networks”.

[10] Paul J. Werbos, “Backpropagation through time : What it does and how to do it”.

[11] “Connectionists First Deep Learning Networks in 1965.” .

[12] G. Hinton, “2007 NIPS Tutorial on: Deep Belief Nets,” *Nips*, pp. 1–100, 2007.

[13] F. Transform, “Digital Processing of Speech and Image Signals,” no. April 2003, 2006.

[14] B.S. ATAL “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification”.

[15] “New LP-derived features for speaker identification - IEEE Journals & Magazine.” .

[16] C. I. Byrnes, P. Enqvist, and A. Lindquist, “Cepstral coefficients, covariance lags, and pole-zero models for finite data strings,” *IEEE Trans. Signal Process.*, vol. 49, no. 4, pp. 677–693, 2001.

[17] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing : a guide to theory, algorithm, and system development*. Upper Saddle River (N.J.) : Prentice-Hall PTR, 2001.

[18] S. Davis and P. Mermelstein, “Comparison of Parametric Representations for,” *IEEE Trans. Acoust.*, vol. 28, no. 4, pp. 357–366, 1980.

[19] a. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, “Phoneme recognition using time-delay neural networks,” *IEEE Trans. Acoust.*, vol. 37, no. 3, pp. 328–339,

1989.

[20] Y. Bennani and P. Gallinari, “On the use of TDNN-extracted features information in talker identification,” in *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, 1991, pp. 385–388 vol.1.

[21] S. Prabhakar, P. Selvaraj, and S. Konam, “Deep Learning for Speaker Recognition,” pp. 1–7, 1996.

[22] S. Furui, “50 Years of Progress in Speech and Speaker Recognition,” *Specom*, no. 1, pp. 1–9, 2005.

[23] “reference.” Pruzansky, S. (1963). Pattern-Matching Procedure for Automatic Talker Recognition. The Journal of the Acoustical Society of America, 35(3), 354-358.

[24] T. Dutta, “Text Dependent Speaker Identification Based on Spectrograms,” *Proc. image Vis. Comput.*, no. December, pp. 238–243, 2007.

[25] H. B. Kekre, V. Kulkanii, P. Gaikar, and N. Gupta, “Speaker Identification using Spectrograms of Varying Frame Sizes,” *Int. J. Comput. Appl.*, vol. 50, no. 20, pp. 27–33, 2012.

[26] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, “Bird detection in audio: A survey and a challenge,” *IEEE Int. Work. Mach. Learn. Signal Process. MLSP*, vol. 2016–Novem, 2016.

[27] A. Hiremath, “Environmental Sound Recognition Using Spectrogram Image Features,” pp. 91–95, 2017.

[28] “Reconnaissance de phonèmes par analyse formantique dans le cas de transitions.” p. 19, 1997.

[29] A. Rozi and Z. Wang, “AN OPEN / FREE DATABASE AND BENCHMARK FOR UYGHUR SPEAKER RECOGNITION Askar Rozi , Dong Wang , Zhiyong Zhang , Thomas Fang Zhen g ∗ Center for Speech and Language Technologies , Division of Technical Innovation and Development , Tsinghua National Laborator,” pp. 81–85, 2015.

[30] R. J. P. de Figueiredo, “Pattern recognition approach to exploration,” *Concepts Tech. oil gas Explor.*, no. 3, pp. 267–286, 1982.

[31] A. Bouziane, H. Kadi, S. Hourri, and J. Kharroubi, “An open and free speech corpus for speaker recognition: The FSCSR speech corpus,” *SITA 2016 - 11th Int. Conf. Intell. Syst. Theor. Appl.*, pp. 1–5, 2016.