الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Master

Filière Electronique
Spécialité Instrumentation
Présenté par

TACHOUCHE Hayet

&

BENBELLIL Salim

Segmentation automatique de l'Arabe parlée en phonème à base de Deep Learning

Proposé par :

M. ABED Ahcéne MCB USD Blida

Année Universitaire 2021-2022

Remerciements

Au terme de cette étude, je tiens à présenter mes sincères remerciements au bon dieu de nous avoir accordé la connaissance de la science et de nous avoir aidé à réaliser ce travail.

Nous tenons à remercier en premier lieu le promoteur Mr ABED Ahcéne pour avoir répondu positivement à notre demande de direction, objet de la présente rédaction. Sa rigueur scientifique et ses remarques ont été utiles pour la qualité de ce travail.

Un grand remerciement aux membres du jury chacun par son propre nom pour l'honneur et l'intérêt qu'ils nous ont accordé en acceptant d'examiner et d'évaluer notre mémoire.

Nous présentons également notre gratitude à tous les professeurs, chefs de travaux et assistants département de génie électrique (électronique) en général.

Ainsi, nous remercions pour leur soutien tant moral, spirituel et matériel, nos parents.

À tous nos frères, sœurs, amis et compagnons qui nous ont aidé, conseillé et encouragé ; trouvent ici l'expression de notre profonde reconnaissance.

Nous saluons également de façon spéciale l'amitié mêlée de complicité dont nous avons bénéficié de (Rania Zerrouk), (Youness boularass).

Dédicaces

À mes très chers parents, source de vie, d'amour et d'affection

À mes chers frères et leurs enfants, source de joie et de bonheur

À mon petite sœur Sabrina ma joie et le bonheur de ma famille.

À toute ma famille, source d'espoir et de motivation (abde nour, brahim, Djamel).

À tous mes amis, (amin bouchiba), (Assia medjroub), (khire dine djoudi)

À mon binôme merci pour ta compréhension tout au long de ce projet

À vous cher lecteur.

Hayet

ملخص:

تتطلب معظم أنظمة المعالجة الآلية وتركيب الكلام عملية تقطيع الصوت .يعد التقطيع اليدوي لإشارات الكلام أمرًا صعبًا للغاية ويستغرق وقتًا طويلاً .لحل هذه المشكلة نقترح في هذا العمل طريقة آلية سريعة وفعالة للغاية . تعتمد هذه الطريقة على الشبكات العصبية العميقة .تم تنفيذ واختبار ثلاثة نماذج ، وهي CNN و Sinc Net و Constant Sinc Layer في .Matlab تظهر معدل تجزئة 97.98٪ للنماذج الثلاثة .يتم استخراج المعلمات باستخدام .MFCC

كلمات مفتاحية :

تقطيع الكلام ، الشبكات العصبية التلافيفية ، التعليم المعمق.

Résumé:

La majorité des systèmes de traitement automatique et la reconnaissance de la parole nécessite une opération de segmentation. La segmentation manuelle des signaux de parole est très difficile et prend un temps énorme. Pour résoudre ce problème nous proposons dans ce travail une méthode automatique très rapide et efficace. Cette méthode est basée sur les réseaux de neurones profonds. Trois modèles, le CNN, le Sinc Net et le Constant Sinc Layer, ont été implémentés et testés sous Matlab. Ils montrent un taux de segmentation de 97.98% pour les trois modèles. L'extraction des paramètres est effectué à l'aide des MFCC..

Mots clés :

CNN ; Sinc Net ; Constant Sinc Layer ; MFCC. Segmentation de parole ; deep learning

Abstract :

Most automatic processing systems and speech synthesis require a segmentation operation. Manual segmentation of speech signals is very difficult and time consuming. To solve this problem we propose in this work a very fast and efficient automatic method. This method is based on deep neural networks. Three models, the CNN, the Sinc Net and the Constant Sinc Layer, have been implemented and tested in Matlab. They show a segmentation rate of 97.98% for the three models. The extraction of the parameters is done using the MFCC...

Keywords:

CNN ; Sinc Net ; Constant Sinc Layer ; MFCC ; Speech segmentation ; deep learning

Table des matières

Liste des Abréviations

AD	:	Arabe Dialectal
AS	:	Arabe Standard
dB	:	décibel
FFT	:	Fast Fourier Transform
F_0	:	Fréquence Fondamentale
GMM	:	Gaussian Mixture Model
HMM	:	Hidden Markov Model
MFCC	:	Mel Frequency Cepstral Coefficients

Liste de Notations :

F_0 La fréquence fondamentale.

Δc_m Premières dérivées des coefficient cepstraux.

$\Delta\Delta c_m$ Deuxième dérivée des coefficient cepstraux.

M Le nombre de filtres.

f_h La fréquence la plus haute.

f_1 La fréquence la plus basse pour le traitement du signal.

$\sigma(z)$ Fonction d'activation sigmoïde.

E Energie.

Liste des Figures

Liste des Tableaux

Introduction Générale

L'objectif principal du projet est d'implémenter un système de segmentation automatique de la parole en phonème pour la langue Arabe. Cette segmentation se base sur la technique de deep Learning. L'implémentation des différents modèles de ce système est basée sur le langage de programmation Matlab. La segmentation de la parole est une opération nécessaire dans le traitement de la parole, qui consiste à découper le signal en segments assez extrêmement homogènes pouvant être transcrits en unités de base (Phonème) ou on trouve ces unités variées selon la nature du segment considéré. Il existe plusieurs types de segmentation selon la taille du segment traité. Généralement les coefficients MFCC sont les paramètres acoustique (caractéristiques) les plus utilisés dans la segmentation automatique de la parole en phonème.

Le traitement de la parole est un domaine de recherche scientifique renouvelé. Lorsque, un groupe de chercheurs ont effectué plusieurs études pour apprendre les meilleures méthodes qui donnent une meilleure classification. La classification est un processus d'apprentissage profond qui travaille pour que le comportement de la machine soit plus intelligent, Il y a deux phases principales dans le processus de la classification qui :

• Apprentissage : La phase d'apprentissage est la phase la plus importante dans la reconnaissance, elle est intéressée de la création du vocabulaire (dictionnaire linguistiques).

• classification : La décision est la deuxième étape de reconnaissance, Lorsque la machine fonctionne pour Déterminer le modèle le plus proche dans le dictionnaire à La nouvelle entrée. Cette estimation doit être en temps court que possible. Comme nous avons dit précédemment. Il y un facteur appelé le taux de classification qui permet de mesurer la performance de la classification.

Ce mémoire est constitué de trois chapitres : le premier chapitre représente une vision générale sur la langue arabe est les concepts de la production de la parole avec explication sur le traitement de la parole. Le deuxième chapitre nous avons présente les diffèrent

méthode utilise pour la segmentation de signal de parole (standard CNN, constant Sinc Layer, SincNet Layer). Après nous parlons sur Les extractions de paramètre reposent sur les coefficients MFCC. Le Troisième chapitre nous allons concevoir notre application. Nous montrerons Notre application est également implémentée en utilisant le langage Matlab. Commençons Le premier est une introduction au langage de programmation choisi. Ensuite nous Montrons une capture d'écran de notre application en cours d'exécution. Enfin, nous terminons le travail par une conclusion sur les résultats obtenus par la méthode proposée, et proposons quelques perspectives qui nous croyons utiles et nécessaires pour améliorer et rendre le processus de segmentation efficace pour la parole arabe.

Chapitre 1

Langue arabe et parole

1.1 Introduction

La segmentation automatique de l'Arabe parlée en phonème, consiste a réussir a séparer la parole en segment (qui sont des phonème dans notre cas) ceci afin d'être utilisé dans de nombreuse applications.

Ce chapitre présente des généralités sur la langue arabe pour mieux se familiariser avec la langue, ainsi que l'origine et les caractéristiques de la parole humaine et quelque exemple sur les différentes applications réaliser sur le signale de parole.

1.2 Langue arabe

La langue arabe fait partie des langues sémitiques (ougaritique, phénicien, araméen, Hébreu, et arabe)rassemblant un large groupe de langues, nommées "Afro-asiatique" et plus lointainement reliées aux familles des langues indigènes de l'Afrique du Nord, elle possède un héritage littéraire très riche remontant jusqu'à l'ère préislamique, parlée par plus de 530 millions de locuteurs dans le monde avec plus de 377 millions de personnes dont elle est la langue maternelle. Trois catégories sont distinguées : l'arabe classique, l'arabe standard MSA (Modern Standard Arabic) et l'arabe dialectal.

L'arabe classique est défini comme la langue formelle parlée pendant l'époque de premières rédactions du Coran, il est fixé par l'écrit et offre une situation linguistique très caractérisée c'est la langue officielle de plus de vingt-deux pays (377 millions de personnes) tandis queL'arabe dialectal est lié à l'origine du locuteur et varie selon les pays arabophones ou même selon les différentes régions dans un même pays. Cependant, du point de vue pratique, les arabophones utilisent, dans la vie quotidienne, l'arabe dialectal plus que l'arabe standard.

Dans notre travail, seul l'arabe standard a été considéré. Il représente une forme de langue commune à tout personne parlants la langue arabe, il est enseigné à l'école. L'arabe standard est essentiellement écrit et lu mais secondairement parlé. Elle s'écrit de droite à gauche, alors que les nombres sont écrits de gauche à droite [1].

1.2.1 Écriture et Phonologie

L'Arabe Standard est constitué par 40 phonèmes ; 28 consonnes, 6 voyelles (3 longues et 3 courtes voyelles) et 6 variantes vocaliques en contexte emphatique.

1.2.1.1 Voyelle

Le système vocalique de l'Arabe Standard a six voyelles, dont trois sont courtes et les autres sont longues [2].

Les voyelles long : sont représenter par des lettres [ا/ (a:) ;/و (u:) ; /ي (i:)] elles sont écrites dans le mot en tant qu'éléments de l'épellation mais aussi en tant que consonnes.

Les voyelles courtes : ne sont pas des lettres indépendantes mais des signes diacritiques qui complémente les lettre (consonnes), exemple :

- Le signe "fatha" au-dessus de la consonne :دَ /da/ c'est la voyelle /a/.
- Le signe "damma" au-dessus de la consonne :دُ /du/ c'est la voyelle /u/.
- Le signe "kasra" au-dessous de la consonne :دِ /di/ c'est la voyelle /i/.
- د étant la consonne dans ces exemples.

Il existe aussi d'autres signes diacritiques :

- Le tanwin : Le signe de tanwin est ajouté à la fin des mots, il correspond à la prononciation du son /n/ à la fin du mot. Ceci consiste à doubler un des signes diacritiques déjà mentionnés : بٌ /bun/ بً /ban/ بٍ /bin/
- Le "sokun" : Ce signe est utilisé pour indiquer l'omission d'une voyelle. Il s'agit d'un petit cercle au-dessus de la consonne : لَكِنْ /lakin/ 'mais'.
- La "shadda" : Ce signe en forme que la lettre "w" : بّ est utilisé pour distinguer les consonnes géminées des consonnes simples : نَزَّلَ /nazzala/ "faire descendre".

Généralement, dans les textes modernes, les signes diacritiques correspondant aux voyelles ourtes ne sont pas représentées [3].

1.2.1.2 Consonnes géminées

Les consonnes de la langue arabe peuvent prendre deux formes : simple et géminée. Pour la consonne géminée on a l'ajout d'un signe diacritique appelé "shadda" au-dessus de la consonne concernée. Acoustiquement, une consonne géminée possède une durée supérieure

à celle de son homologue simple. Le remplacement d'une consonne simple par son homologue géminée change le sens du mot.

Exemple :

دَرَسَ "darasa" qui veut dire "il a étudié", devient دَرَّسَ "darrasa" qui veut dire "il a enseigné" [2].

1.2.2 Consonnes emphatiques

Il s'agit du point de vue acoustique chez les consonnes pharyngalisées une concentration de l'énergie dans les plus basses régions du spectre et dirigent vers le bas le second formant de la voyelle suivante de la pharyngalisation de certaines consonnes dans certains contextes. Certaines voyelles peuvent présenter l'aspect emphatique si elles sont précédées ou suivies par une consonne emphatique. Les consonnes en arabe peuvent être regroupées en trois classes :

- Consonnes de nature emphatique : comme les consonnes خ ; غ ; ظ ; ض ; ق ; ط ; ص.
- Consonnes qui peuvent être emphatiques dans certains contextes : ل ; ر.
- Les autres consonnes qui ne peuvent pas être emphatiques [4].

1.2.3 Description phonétique et classifications des consonnes

Rappelons que le phonème est le plus petit élément des unités de la parole, Il indique la différence dans le sens, le mot et la phrase. Les phonèmes arabes contiennent deux classes distinctives appelées pharyngale et emphatique. Ces deux classes caractérisent les langues sémitiques comme l'arabe et l'hébreu

Tableau 1: Transcription Orthographique-Phonétique (TOP)

Graphème Arabe	Transcription phonétique	Graphème Arabe	Transcription phonétique
ء	[ʔ]	ض	[dˤ]
ب	[b]	ط	[tˤ]
ت	[t]	ظ	[ðˤ]
ث	[θ]	ع	[ʕ]
ج	[ʤ]	غ	[ʁ]
ح	[ħ]	ف	[f]
خ	[χ]	ق	[q]
د	[d]	ك	[k]

ذ	[ð]	**ل**	[l]
ر	[r]	**م**	[m]
ز	[z]	**ن**	[n]
س	[s]	**ه**	[h]
ش	[ʃ]	**و**	[w]
ص	[sˤ]	**ي**	[j]

Les consonnes peuvent être regroupées en différentes classes selon le mode et le lieu d'articulation :

- Nasales : Produites en abaissant le voile du palais. Selon l'emplacement elles sont soit nasales : dentale-alvéolaire ن ou labiale م.
- Plosives : elle est produite par la fermeture complète et momentanée du chenal expiratoire résultant du contact entre deux articulateurs.
  - Les plosives voisées sont non-aspirées, tandis que les plosives non-voisées sont aspirées, à l'exception de ق
  - Au voisinage d'une voyelle frontale haute, telle que /i/ ou /i:/ (courte ou longue), la plosive ك est palataliser.
  - La bilabiale voisée ب est souvent dévoisée au voisinage d'un son voisé.

- Vibrante : C'est un son voisé dentoalvéolaire ر.
- Fricatives : Elles sont produites par resserrement du conduit vocal à un endroit variable. La friction est produite par différents organes et combinaisons (lèvres, langue, dents contre lèvres, dents contre langue, voile du palais. . .) : ; ع ; ذ ; ز ; ظ ; ط غ ; ح ; خ ; ف ; ه ; ش.
- Affriquées : Elle est composée d'une phase occlusive (où le flux d'air est bloqué) suivie par une étape fricative (où l'air retenu est relâché pour passer par une ouvertureplutôt étroite) : ج.
- Approximantes : Elles sont produites par un rapprochement modéré des organes phonateurs qui ne va pas jusqu'à produire le bruit caractéristique de friction des fricatives : ي ; و.
- Latérale : Elle est formée par l'affaissement de l'avant de la langue et le contact de sondosavec le palais : ل [5].

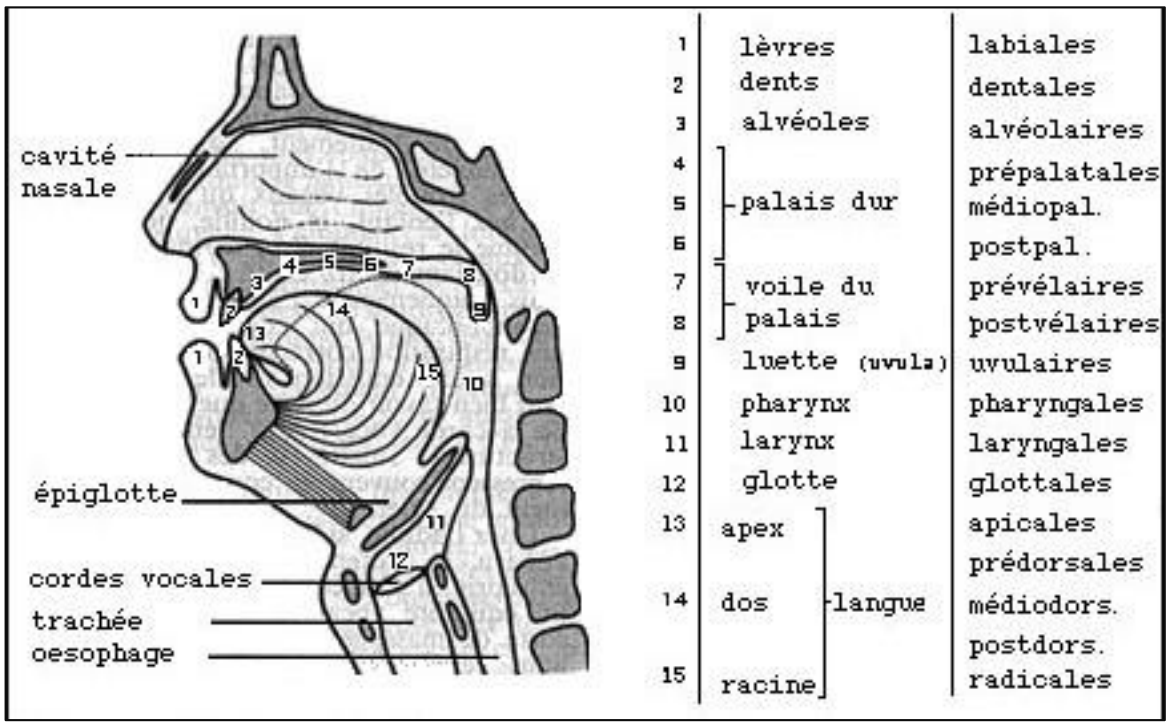


Figure 1. Lieux d'articulation des phonèmes arabes [21]

1.3 Parole

Dans cette parie on présentera les différentes parties responsables de la production de laparole.Le processus de la génération de la parole est un mécanisme complexeassuré par l'appareil phonatoirequi fait intervenir plusieursorganes tels que les poumons, le larynx, la langue etles lèvres.

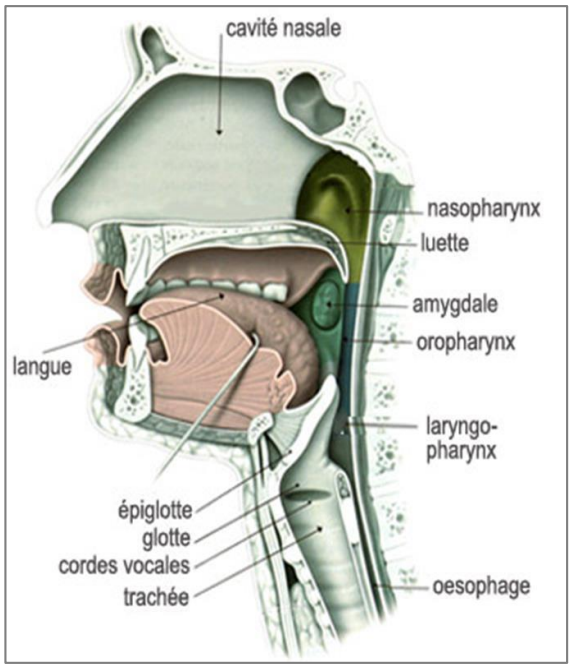


Figure 2. Appareil phonatoire [22]

Son fonctionnement repose sur une interaction à trois niveaux :

- Le système subglottique : composé par les poumons, le diaphragme et la trachée c'est la source du souffle.
- Le système phonatoire ou larynx : elle comprend les cordes vocales, des cartilages et des muscles c'est la source vocale ou sonore.
- Le système supra-glottique : qui contient les résonateurs qui sont principalement les cavités pharyngale, buccale, nasale et labiale [6].

1.3.1 Fonctionnement

La production de la parole résulte d'une variation de la pression de l'air générée par les poumons qui se gonflent et se dégonflent afin d'entretenir un courant d'air ensuite il est transmis par la trachée à travers les cordes vocales. Les cordes vocales sont définies par la superposition des muscles et de ligaments (figure 3). Si les cordes vocales sont écartées un bruit est produit grâce au passage de l'air librement dans le cas contraire quand les cordes vocales sont rapprochées, l'air sous pression peut engendrer leur vibration ce qui mène à avoir un son quasi-périodique dont la fréquence fondamentale correspond à la hauteur de la voix perçue.

On a vu que les cordes vocales peuvent laisser l'air circule librement quand elles sont écartées ou le contraire quand elles sont accolées. La troisième positions Cordes rapprochées et vibrant : l'air circule en faisant vibrer les cordes vocales, c'est le phénomène de voisement.

Dans les cas où les cordes vocales sont écartées ou rapprochées, le flux d'air continue son chemin à travers le conduit vocal. Ce dernier est considéré comme résonnateur de parole. Sa forme est déterminée par la position des articulateurs tels que la langue, la mâchoire, les lèvres ou le voile du palais ce qui définit le timbre des différents sons de la parole.

Le conduit vocal est généralement considéré comme un organe de production de la parole au-dessus des cordes vocales. Il inclut les organes d'excitation et les articulateurs. Les poumons, la trachée et les cordes vocales sont considérés comme des organes responsables de la production d'excitation. Les articulations incluses dans le tractus sont groupées dans le pharynx ; le larynx ; la cavité buccale ;la cavité nasale [7].

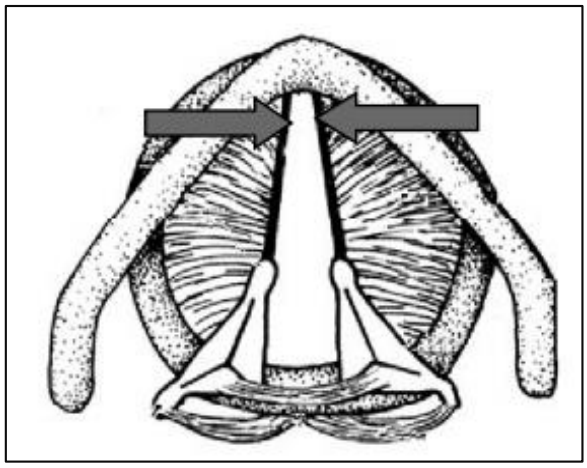


Figure 3. Coupe d'une corde vocale

1.3.2 Signal de la parole

Le signal de parole obtenu est un signal continu périodique ou aléatoire. Ainsi, le signal de parole est caractérisé par :

1.3.2.1 Voisement

Un signal de parole est formé par une succession des sons voisés et des sons non voisés dont l'amplitude diffère :

Les sons voisés :

Comme vu précédemment Ils sont considérés comme des signaux quasi-périodiques ayantune fréquence fondamentale et des harmoniques. Ce sont des sons prononcés avec une vibration des cordes vocales.

Les sons non-voisé :

Ils sont produits par une constriction des tractus vocaux étroits pour faire une turbulence du flux d'air, qui est produit par un bruit ou par une voix respiratoire. Le son non-voisé est souvent considéré comme un bruit blanc. Lors de la production des sons non-voisés, les cordes vocales ne vibrent pas.ces sons sont généralement considérés comme des bruits fricatifs [7].

1.3.2.2 Fréquence fondamentale :

La fréquence fondamentale d'un signal de parole est la fréquence d'oscillation des cordes vocales qui est quasi-périodique. Permettant la perception de la hauteur tonale de la voix d'un individu qualifié aussi de "pitch" et utilisée comme synonyme de la fréquence fondamentale, même si le pitchfait référence à la fréquence de la tonalité perçue, c'est ce que l'être humain peutentendre. Ainsi, la fréquence fondamentale est un paramètre acoustique mais le pitchreste un paramètre de perception. Généralement, la fréquence fondamentale varie selon le genre et l'âge : les hommes de 100 à $150Hz$; les femmes de 200 à $300Hz$; enfant de 300 à $450Hz$.

1.3.2.3 Energie

Elle caractérise l'intensité sonore d'un segment de parole.

1.3.2.4 Spectre fréquentiel

Le principe de Fourier permet dedécomposer le signal de parole en une somme d'ondes sinusoïdales. L'analyse de Fourier permet de déterminer quelles sinusodes sont àconsidérer pour reconstruirele signal original. Les amplitudes de ces ondes sinuso$1\tilde{3}053'$fdales révèlent le contenufréquentiel du signal d'origine.

1.3.3 Prétraitement du signal de parole

La parole est physiquement représentée comme unchangement de pression de l'air, et la phonétiqueacoustique étudie ce signal en le convertissant d'abord enun signal électrique à

l'aide d'un capteur approprié : tel unmicrophone. Aujourd'hui, les signaux électriques produitssont le plus souvent numérisés. Il peut ensuite être soumisà un ensemble de traitements statistiques destinés àmettre en évidence ces caractéristiquesacoustiques : fréquence fondamentale,son énergie et son spectrefréquentiel [8].

1.3.3.1 Filtre de garde

Afin de réduire le cout du traitement numérique d'une façon notable, on doit limiter le spectre en utilisant un filtre dont la fréquence de coupure (f_c) est choisie en fonction de la fréquence d'échantillonnage.

1.3.3.2 Échantillonnage

Le signal de la parole étant analogique, il s'avère nécessaire de le numériser avant touttraitement. Cette opérationconsiste en l'échantillonnage du signal qui est présenté dans la figure 4

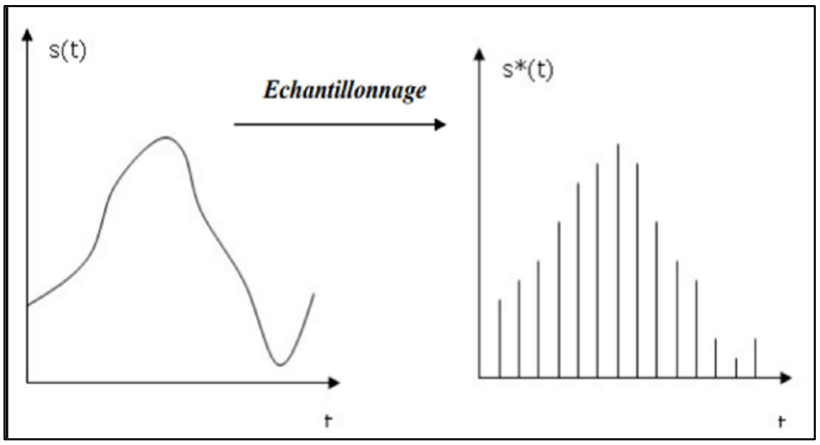


Figure 4. Echantillonnage d'un signal

D'après Shannon, la perte d'information entre le signalanalogique et le signal discretcorrespondant est nulle si etseulement si on a :

$$f_e \leq 2 \times f_max \quad (01)$$

- f_e : La fréquence d'échantillonnage.
- f_{max} : la fréquence maximale que contient le signal à traiter.

1.3.3.3 Quantification

La quantification définit le nombre de bits sur lesquels on veut réaliser la numérisation. Elle permet de mesurer l'amplitude de l'onde sonore à chaque pas d'échantillonnage. Le

choix de la fréquence d'échantillonnage est aussi déterminant pour la définition de la bande passante représentée dans le signal numérisé qui est présenté dans la figure (1.5)

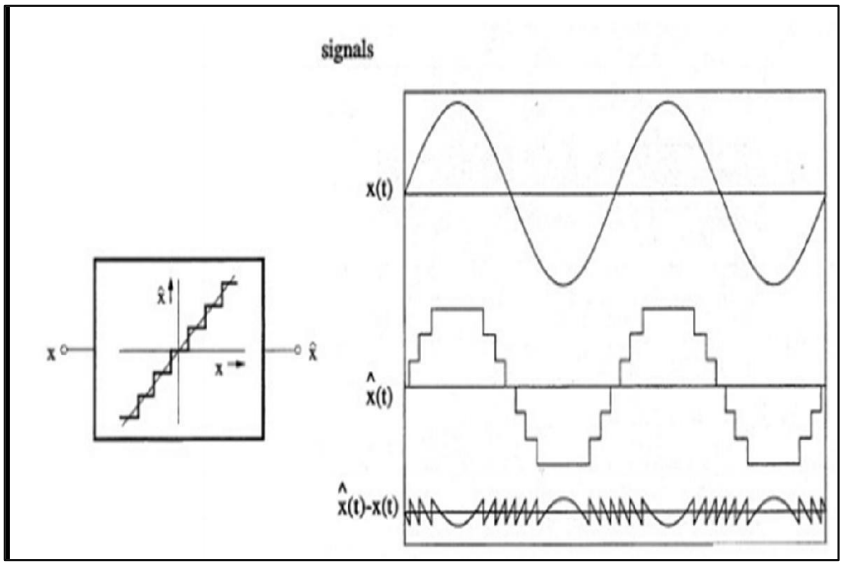


Figure 5. Quantification

1.3.3.4 Préaccentuation

En général, le signal vocal se caractérise par une perte de 6dB/octave, due à l'influence de la source d'excitation et aurayonnement des lèvres. Une perte de 6 dB/octave veut direqueles hautes fréquences ont une énergie plus faible quecelle des basses fréquences. Pour pallier cet inconvénientla préaccentuation permet d'égaliser les sons aigus avec lesSons graves.

L'opération consiste à faire passer le signal à travers unfiltre de transmittance :

$$H(Z) = 1 - aZ^{-1} \qquad (02)$$

Le facteur de préaccentuation est pris entre 0.9 et 1 (souvent 0.95). Comme conséquence, la préaccentuation introduit une légère distorsion spectrale.

1.4 Traitement automatique de la parole

Le traitement de la parole est une discipline qui fait le lien entre le traitement du signal ainsi que du traitement du langage. Ce qui donne lieu a différente famille de module de parole selon les besoins d'utilisation, qui elles-mêmes contiennent différentes applications sur la parole. Tel que la reconnaissance vocale, mais aussi la synthèse de la parole, l'identification du locuteur ou la vérification du locuteur.

Tous ces différentes techniques de traitement de la parole font intervenir des procédures que nous somme amener à utiliser dans notre travail. C'est pour cela qu'on va vous présenter une de ses techniques, qui est la reconnaissance vocale.

1.4.1 Reconnaissance vocale

Tout commence par une phrase qui sera enregistrée et numérisée, ensuite elle subit un découpage fonctionnel de la manière suivante :

Le traitement acoustique (front-end) permet essentiellement d'extraire du signal de parole une image acoustique la plus significative envisageable sur des tranches de temps d'environ 30ms. Cette image se présente sous la forme d'un vecteur de caractéristiques (features extraction) de 10 à 15 composantes principales, auxquelles sont ajoutées les différences de premier et second ordre pour obtenir une taille de 30-45 en final.

Cette opération vise à numériser le signal de parole sous forme de vecteurs acoustiques qui forment les données d'observation pour le dispositif dereconnaissance. On utilise pour cela les techniques de traitement du signal .

Tel que la technique de fenêtrage de Hemming qui consiste à la découpe du signal en tranches de 30 ms en procédant pour chaque tranche à un décalage de 10 ms afin d'obtenir 10 ms de données significatives pour chaque vecteur. Ensuite le signal est alors numérisé paramétré par une technique d'analyse fréquentielle utilisant les transformées de Fourier qui est MFCC, Mel-Frequency Cepstral Cœ fficients.

L'apprentissage automatique qui exécute une association entre les segments élémentaires de paroles et les éléments lexicaux. Cette association fait appel à une modélisation statistique entre autres par modèles de Markov cachés (HMM, Hidden Markov Models) et/ou par réseaux de neurones artificiels (ANN, Artificial Neural Networks).

Dans notre cas ça sera une autre forme de deep Learning qui est le CNN (Convolutional Neural Network), le SINC(Neural Networks Activation Function), SINC Layer, afin d'élire le système le mieux adapter pour notre tâche.

La reconnaissance (back-end) qui en concaténant les segments élémentaires de paroles auparavant appris reconstitue le discours le plus probable.

1.4.2 Reconnaissance de parole

La reconnaissance automatiquedu parole vise à déterminer automatiquement le conteni linguistique dans un enoncé de parole. La majorité des systèmes de reconnaissance automatique de parole repose sur les modèles de Markov cachés.

1.5 Conclusion

Dans ce chapitre on a donné une vision générale sur la langue arabe est les concepts de la production de la parole avec explication sur le traitement de la parole, dans le chapitre suivant nous essayons d'introduire la segmentation automatique de parole en phonèmes. Le système proposé est basé sur l'apprentissage profond (deep Learning).

Chapitre 2

Segmentation de parole et deep learning

2.1 Introduction

Dans ce chapitre nous avons présente les diffèrentes méthodes utilisées pour la segmentation du signal de parole en phonème à savoir : standard CNN, constant Sinc Layer, SincNet Layer). L'extraction des paramètres reposent sur les coefficients MFCC.

2.2 Segmentation de la parole

C'est la représentation du message (signal vocal) sous la forme d'une suite de segments de parole. On s'intéresse a détecter les boundaries de chaque unité de base (mot, syllabe, phonème, .etc.

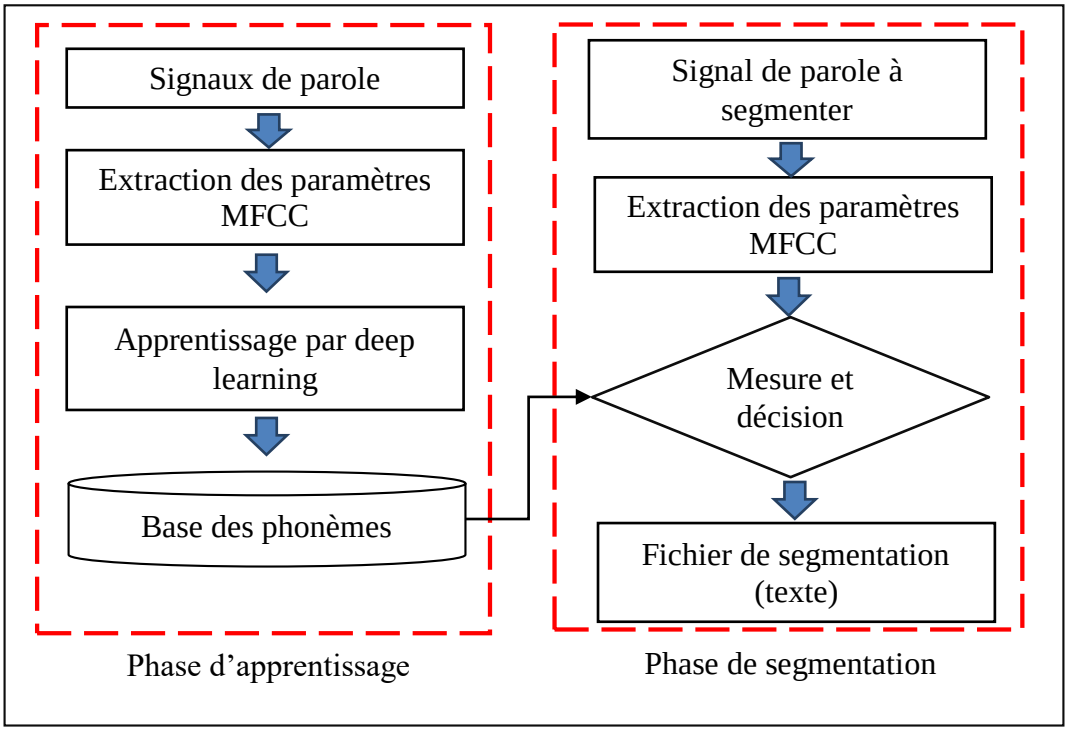


Figure 6. Segmentation automatique des signaux de parole

La majorité des systèmes du traitement automatique de la parole passe par une phase très importante. C'est la phase de segmentation, car elle prépare le signal de parole pour les traitements ultérieurs.

Cette phase possédé une grande influence sur la qualité des caractéristiques à obtenir et par conséquent, le taux de classification à obtenir. Le but de cette phase est l'extraction des segments de base à traiter selon l'unité de base à savoir : le mot, le syllabe ou le phonème ... etc. ce processus est très influencé par le bruit intégré dans le signal enregistré [10].

2.2.1 Type de segmentation

De point de vu pratique les systèmes de segmentation regroupent plusieurs types selon l'unité de base.

2.2.1.1 Segmentation en voisées/ non voisées

Les sons voisés sont produits par la vibration des cordes vocale. Les voyelles sont intrinsèquement voisées, tandis que les consonnes peuvent l'être ou non. On peut donc considérer qu'un mot est constitué d'une suite de segments voisés, de segments non voisés et de silences brefs. Cependant toute suite de ces trois segments de base ne correspond pas à un mot, du bruit peut être constitué par des sons voisés. Un des paramètres de voisement est le pitch [11].

2.2.1.2 Segmentation en phonèmes

La segmentation d'un signal de parole en phones consiste à délimiter sur le continuum acoustique de ce signal une séquence de segments caractérisés par des étiquettes appartenant à un ensemble discret et fini d'éléments, qui est l'alphabet phonétique de la langue. La segmentation phonétique de la parole est une tâche difficile car le signal de parole n'est pas clairement composé de segments discrets bien délimités [12].

D'un côté, nous constatons que l'élocution d'un énoncé se caractérise par un mouvement continu des organes de la parole et par l'absence d'un quelconque positionnement statique de ces organes. Le passage d'une cible articulatoire d'un phone, à une autre cible articulatoire d'un autre phone, se fait de manière continue, avec un chevauchement entre les deux configurations articulatoires, ce qui donne naissance au phénomène de coarticulation.

2.2.1.3 Segmentation en mots

La segmentation d'un message parlé en ses constituants élémentaires est un sujet difficile. Pour l'éviter, de nombreux projets de la RAP se sont intéressés à la reconnaissance de mots prononcés isolement.

La reconnaissance des mots isolés ou tous les mots prononcés sont supposés être séparés par des silences de durée supérieure à quelques dixièmes de secondes, se fait essentiellement par l'approche globale [13].

2.2.1.4 Segmentation en locuteurs

Et tour de parole La segmentation selon le locuteur est née relativement récemment pour répondre au besoin créé par le nombre toujours croissant de documents multimédia devant être archivés et accédés. Les tours de parole et l'identité des locuteurs constituent une intéressante clé d'accès à ces documents.

Le but de la segmentation selon le locuteur est donc de segmenter en tour de parole (un tour de parole est un segment contenant une intervention d'un locuteur) un document audio contenant N locuteurs et d'associer chaque tour de parole au locuteur l'ayantprononcé.

En général, aucune information apriori n'est disponible, sur le nombre de locuteur sous leurs identités [14].

2.2.2 Extraction de coefficients MFCC

Le principe de calcul des MFCC (Mel Fréquence Cepstral Coefficients) est issu des recherches psychoacoustique sur la perception des différentes bandes de fréquences par l'oreille humaine. Le principal intérêt de ces coefficients est extraire des informations pertinentes en nombre limité en s'appuyant à la fois sur la production (théorie Cepstrale) et sur la perception de la parole (échelle des Mels). Plusieurs étapes sont nécessaires pour transformer un fichier audio en Cepstre MFCC.

Les calculs de coefficients MFCC est réalisé de la maniéré suivante :

2.2.2.1 Fenêtrage

Le signal vocal est un signal non stationnaire ; il présente une évolution lente dans le temps. Le but du fenêtrage est de découper le signal de parole en petites tranches (chacune de durée $32ms$ environ) ou il peut être considéré localement comme quasi- stationnaire. En outre, et pour l'évolution lente du signal vocal, le fenêtrage permet le traitement en temps réel et il facilite aussi l'analyse des signaux sur la machine. Les ressources d'une machine étant limitées, le signal ne peut pas être traité dans sa globalité. Il existe plusieurs types de fenêtres d'analyse on représente quelques-unes :

Fenêtre rectangulaire

Elle est définie par :

$$f(nT) = \begin{cases} 1 & |nT| < T' \\ 0 & ailleurs \end{cases} \quad (03)$$

Fenêtre Hanning :

Elle est définie par :

$$f(nT) = \begin{cases} 0.5\left(1 - \cos\left(\frac{n\pi T}{T'}\right)\right) & |nT| < T' \\ 0 & ailleurs \end{cases} \quad (04)$$

Fenêtre Hamming :

Elle est définie par :

$$f(nT) = \begin{cases} 0.54 - 0.46\cos\left(\frac{n\pi T}{T'}\right) & |nT| < T' \\ 0 & ailleurs \end{cases} \quad (05)$$

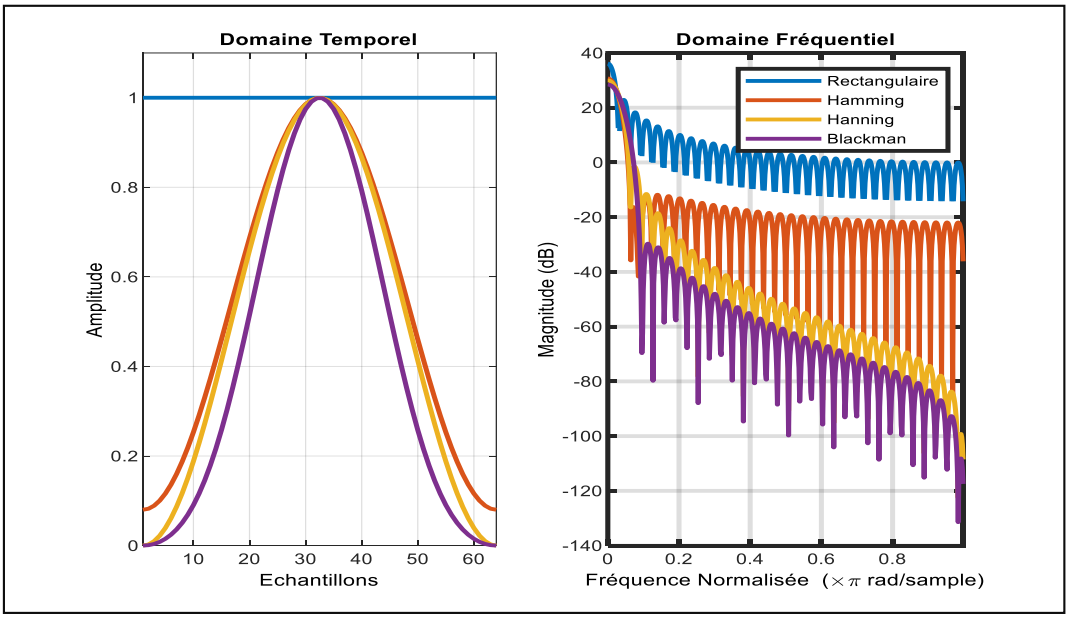


Figure 7. Fonctions de fenêtrage

Parmi ces fenêtres, la fenêtre de hamming est la plus convenable à la parole, car elle entraïne un minimum de distorsion spectrale du signal de parole, par rapport aux autres fenêtres.

2.2.2.2 Application de la FFT (Fast Fourier Transform)

La transformée de Fourier rapide (acronyme anglais : FFT ou Fast Fourier Transform) est un algorithme de calcul de la transformée de Fourier discrète (TFD). Ainsi, pour le temps de calcul de l'algorithme rapide peut être 100 fois plus petit que le calcul utilisant la formule

de définition de la TFD. Cet algorithme est couramment utilisé en traitement numérique du signal pour transformer des données du domaine temporel au domaine fréquentiel.

2.2.2.3 Banc de filtres Mel au spectre de puissance

Banc de filtres Mels : L'échelle spectrale dans le calcul du cepstre est linéaire en fréquence. Cependant, les études physiques et perceptives de l'oreille indiquent qu'elle est sensible à une échelle presque logarithmique de la fréquence [15]. Pour s'approcher donc du modèle de l'audition, on fait appel à une échelle pseudo logarithmique appelée échelle MEL linéaire pour des fréquences allant de 0 à $1KHz$, et logarithmique au-delà.

Les paramètres MFCC utilisent une échelle fréquentielle non linéaire. La fréquence Mel-échelle est définie par :

$$B(f) = 2595\log\left(1 + \frac{f}{700}\right) \quad (06)$$

Où

- f : est la fréquence en Hz,
- $B(f)$: est la fréquence en échelle Mel de f.

Soit un signal discret $s[n]$ avec :

$$0 < s[n] < s[N-1] \quad (07)$$

- N : est le nombre d'échantillons de la fenêtre analysée.
- F_s : est la fréquence d'échantillonnage,

La transformée de Fourier discrète $S[k]$ est obtenue par :

$$S[k] = \sum_{n=0}^{N-1} s[n] e^{\frac{-j2\pi nk}{N}} \quad 0 \leq k < N \quad (08)$$

Le spectre du signal est multiplié avec des filtres triangulaires dont les bandes passantes sont équivalentes en domaine mèl-fréquence. Les points frontières $B[m]$ des filtres en Mel fréquence sont calculés ainsi :

$$B[m] = B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1} \quad (09)$$

Avec :

- M : le nombre de filtres.
- f_h : la fréquence la plus haute.
- f_l : la fréquence la plus basse pour le traitement du signal

Dans le domaine fréquentiel, les points $f[m]$ discrets correspondants sont calculés par l'équation :

$$f[m] = \left[\frac{N}{F_s}\right] B^{-1}\left[B(f_l) + m\frac{B(f_h)-B(f_1)}{M+1}\right] \quad (10)$$

Où B^{-1} est la transformée de Mel-fréquence en fréquence.

$$B^{-1}(b) = 700 * \left(10^{\frac{b}{2595}} - 1\right) \quad (11)$$

Le coefficient $H_m(k)$ de chaque filtre est déterminé par le système suivant :

$$H_m(k) = \begin{cases} 0 & k \leq f[m-1] \\ \frac{k-f[m-1]}{f[m]-f[m-1]} & f[m-1] \leq k \leq f[m] \\ \frac{f[m+1]-k}{f[m+1]-f[m]} & f[m] \leq k \leq f[m+1] \\ 0 & k \geq f[m+1] \end{cases} \quad (12)$$

Pour un spectre lissé et stable, à la sortie des filtres un logarithme d'énergie (ou un logarithme de spectre d'amplitude) est calculé [16] :

$$E[m] = \log[\textstyle\sum_{k=0}^{N-1} |S[k]|^2 H_m(k)] \quad 0 \leq m < M \quad (13)$$

2.2.2.4 Coefficient cepstraux

C'est l'étape finale, on transforme les données dans l'échelle des Mels, fréquentielle donc vers l'échelle des temps. Le résultat de cette étape sera les MFCC proprement dit. Il suffit d'effectuer l'inverse de la transformée de Fourier. Dans la pratique, on effectue une transformée en $FFT^{-1}(DCT^{-1})$ indiquées dans les équations ?? et ce qui revient au même puisque la transformée en Cosinus inverse donne la partie réelle de la transformée de Fourier ; or ici on a que des réels. Il faut noter que la transformée en sinus donnera la partie imaginaire de la transformée de Fourier.

$$c[n] = \textstyle\sum_{m=0}^{M-1} E[m]\cos\left(\frac{\pi n\left((m+\frac{1}{2}\right)}{M}\right) \quad 0 \leq n < M \quad (14)$$

2.2.2.5 Calcul des caractéristiques dynamiques des MFCC

Jusqu'ici aucune information sur l'évolution de temps n'est incluse dans les MFCC. L'information dynamique dans le signal de parole est égalementdifférente d'un locuteur à l'autre. Cette information est souvent donner par les dérives cepstrales. La première dérivée

des coefficients cepstraux s'appelle les coefficients Δ, et la dérivée deuxième des coefficients cepstraux s'appelle les coefficients ΔΔ. Les coefficients Δ nous donnent quelques informations sur la variation de ces vecteurs dans le temps, et les coefficients ΔΔ nous donnent des informations sur l'accélération de la parole. Ces coefficients sont donnés par :

$$\Delta c_m = \frac{\sum_{k=-1}^{l} K.c_{m+k}}{\sum_{k=-1}^{l} |k|} \quad (15)$$

$$\Delta\Delta c_m = \frac{\sum_{k=-1}^{l} k^2.c_{m+k}}{\sum_{k=-1}^{l} k^2} \quad (16)$$

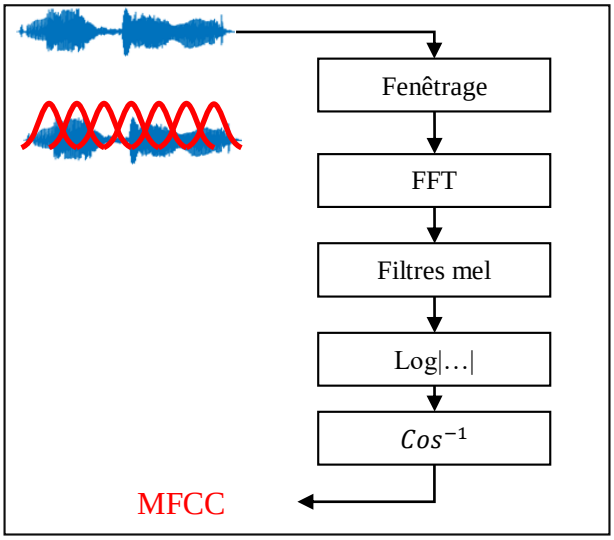


Figure 8. Calcul des MFCC

2.2.3 Apprentissage du système de segmentation

Dans la pratique on trouve plusieurs méthodes pour effectuer cette opération. Généralement la majourité des sytsème de segmentation de la parole repose sur les HMM et l'intélégence artificiel

2.2.3.1 Modèles de Markov Cachés (Hidden Markov Model)

Le modèle de Markov caché est une méthode statistique Les fonctionnalités puissantes sont des échantillons de données observés d'un processus à l'autre. Temps discret. Non seulement il offre une efficacité Modèle paramétrique, mais il intègre également des principes de programmation Segmentation et classification dynamiques des séquences de données unifiées Changer avec le temps. Les modèles de Markov cachés peuvent être utilisés

pour représenter des séquences Un son qui fait partie d'une unité de parole, comme un phonème. Phonème, un son Discours de base, qui peut être modélisé par des individus HMM de gauche à droite. Dans la modélisation HMM du processus, les échantillons peuvent être est caractérisé par un processus paramétrique stochastique dont les paramètres peuvent être Estimer dans un cadre bien défini

HMM est devenu la méthode la plus couramment utilisée pour la modélisation des signaux vocaux dans les applications suivantes : la reconnaissance et le suivi automatique de la parole, l'extration de la fréquence fondamentale et des formants, la synthèse vocale, la traduction automatique, l'étiquetage syntaxique et la compréhension de la langue parlée. Un processus aléatoire est un processus de Markov si son évolution est exactement déterminé par la probabilité initiale et la probabilité de transitions entre les états (son évolution ne dépend pas de son passé, mais seulement de son état présent. L'état actuel du système contient toutes les informations pour prédire son état futur [16].

2.2.3.2 Intelligence artificielle

L'intelligence artificielle (IA) est devenue un sujet très important au cours de la dernière décennie. Ce domaine regroupent les systèmes basés sur l'apprentissage automatique, l'apprentissage profond "Deep Learning" et Intelligence Artificielle (Figure 9) [17].

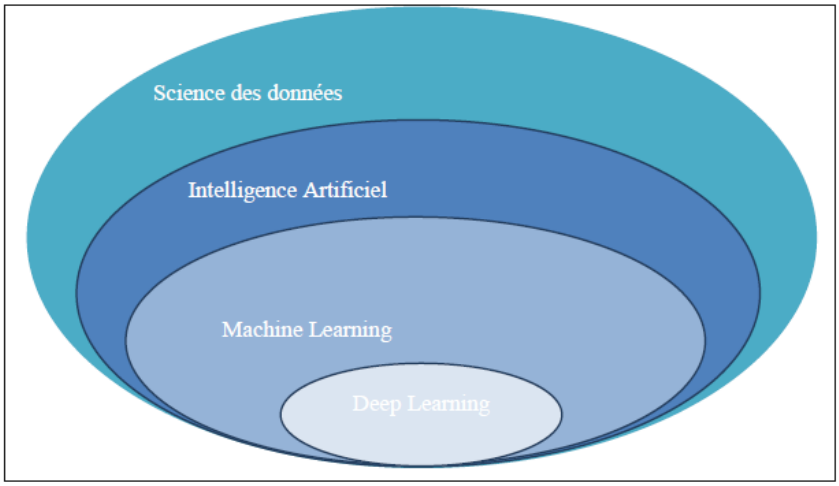


Figure 9. Représentationdes techniques de science des données

a) Apprentissage automatique (AA)

Les progrès récents de l'intelligence artificielle passent par l'apprentissage appliqué (apprentissage automatique) à de très grands ensembles de données. Cette Les algorithmes d'apprentissage automatique détectent les modèles et apprennent à faire Prédiction et

recommandation par le traitement des données et de l'expérience, plutôt que de recevoir des instructions de programmation explicites. Algorithme Peut également être ajusté pour s'améliorer en fonction de nouvelles données et expériences leur efficacité dans le temps [17].

b) Apprentissage profond (Deep Learning)

C'est un ensemble de techniques d'apprentissage automatique qui a permis des avancées importantes en intelligence artificielle dans les dernières années. Dans l'apprentissage automatique, un programme analyse un ensemble de données afin de tirer des règles qui permettront de tirer des conclusions sur de nouvelles données. L'apprentissage profond est basé sur ce qui a été appelé, par analogie, des « réseaux de neurones artificiels » , composés de milliers d'unités (les « neurones ») qui effectuent chacune de petites opérations simples. Les résultats d'une première couche de « neurones » servent d'entrée aux calculs d'une deuxième couche et ainsi de suite. Par exemple, pour la reconnaissance visuelle, des premières couches d'unités identifient des lignes, des courbes, des angles... des couches supérieures identifient des formes, des combinaisons de formes, des objets, des contextes... Les progrès de l'apprentissage profond ont été possibles notamment grâce à l'augmentation de la puissance des ordinateurs et au développement de grandes bases de données «big data».

2.3 Réseaux de neurones profonds (deep learning)

Les réseaux de neurones d'apprentissage en profondeur, ou réseaux de neurones artificiels, tentent d'imiter le cerveau humain grâce à une combinaison d'entrées de données, de pondérations et de biais. Ces éléments fonctionnent ensemble pour recon-naitre , classer et décrire avec précision les objets dans les données.

Les réseaux de neurones profonds se composent de plusieurs couches de nœ uds interconnectés, chacun s'appuyant sur la couche précédente pour affiner et optimiser la prédiction ou la catégorisation. Cette progression des calculs à travers le réseau est appelée propagation vers l'avant. Les couches d'entrée et de sortie d'un réseau neuronal profond sont appelées couches visibles. La couche d'entrée est l'endroit où le modèle d'apprentissage en profondeur ingère les données pour le traitement, et la couche de sortie est l'endroit où la prédiction ou la classification finale est effectuée

1. **Couche d'entrée :** reçoit des données d'entrée et transmet les entées à la première couche masquée.
2. **Couches cachées :** effectuent des calculs mathématiques sur nos entrées. L'un des défis de la création de réseaux neuronaux est de déterminer le nombre de couches cachées, ainsi que le nombre de neurones pour chaque couche.
3. **Couche de sortie :** renvoie les données de sortie. Chaque connexion entre neurones est associée à un poids. Ce poids détermine l'importance de la valeur d'entrée. Les poids initiaux sont définis aléatoirement.

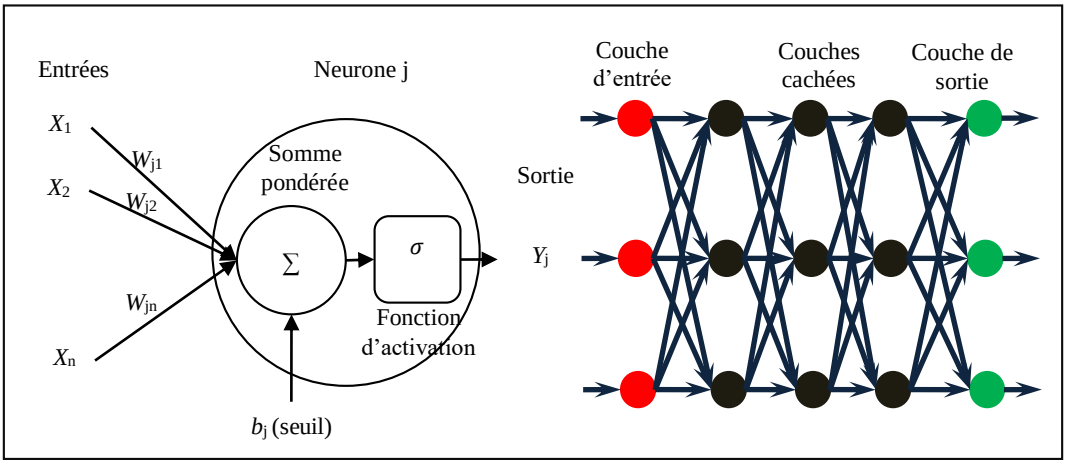


Figure 10. Schéma d'un réseau de neurone profond

La figure montre un schéma fonctionnel du reseau de neurones profond avec :

- **Neurone :** est une unité qui reçoit l'information, procède à des calculs simples, et la transmet à une autre unité.
- **Poids et biais** : application des poids à l'entrée de chaque neurone pour calculer les données de sortie
- **Fonctions d'activation** : est une transformation linéaire ou non linéaire. Elle calcule la valeur de l'état du neurone. C'est cette valeur qui sera transmise aux neurones suivants. Beaucoup de fonctions d'activation ont été introduites vu la variété des modèles RNA, on y trouve :
  - **Modèle à seuil** : ce modèle est très proche et conforme à la réalité biologique mais il pose des problèmes d'apprentissage.
  - **Modèles linéaires et sigmoïdaux** : ces modèles sont très adaptés aux algorithmes d'apprentissage comme celui de rétro propagation du gradient car leur fonction de transition est différentiable. Voici un exemple d'une fonction d'activation sigmoïde :

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (17)$$

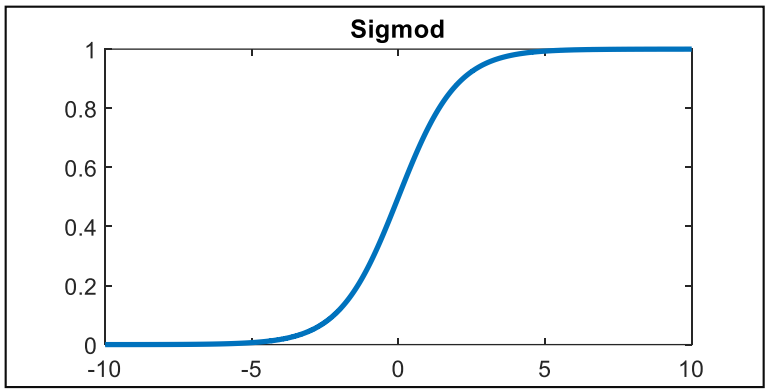


Figure 11. Fonction d'activation sigmoïde

2.3.1 Applications du Deep Learning

Dans la pratique on trouve plusieurs application reposent sur les réseaux de neurones profonds. On cite :

- Classification des images
- Reconnaissance des actions humaines
- Reconnaissance des gestes de la main
- Analyse des scènes

Il existe beaucoup de types de réseaux de neurones, chaque type étant développe pour un objectif particulier. Dans notre travail on s'interesse au reseaux de neurones convolutifs CNN, les réseaux de neurones SincNet et les réseaux de neurones ConstantSincLayer.

2.3.2 Réseaux de neurones convolutifs CNN

Les CNN désignent une sous-catégorie de réseaux de neurones artificiels qui désignent des modèles de classification d'images réputés être les plus performant, ils permettant notamment la reconnaissance d'images en attribuant automatiquement à chaque image fournie en entrée, une étiquette correspondant à sa classe d'appartenance. Contrairement à un modèle MLP (Multi Layers Perceptron) classique qui ne contient qu'une partie de classification, l'architecture du Convolutional Neural Network dispose en amont d'une partie convolutive et comporte par conséquent deux parties bien distinctes : partie convolutive et partie classification

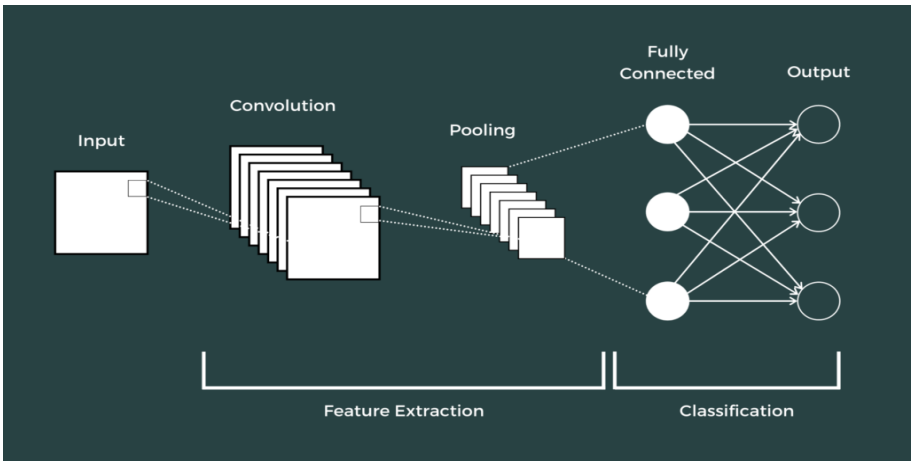


Figure 12. Architecteur de base du réseau CNN

2.3.2.1 Couche de convolution (Convolution Layer)

L'objectif de cette couche est de détecter les caractéristiques telles que les bords, les taches de couleur et d'autres éléments visuels, avec le filtre la convolution d'image crée des image appelées cartes de caractéristique de sortie, plus de filtres dans les couches de convolution plus des caractéristiques ont détecté.

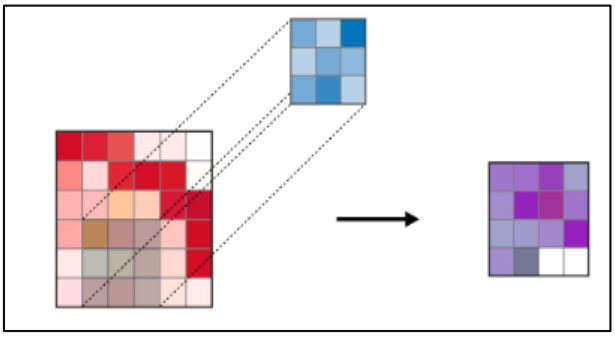
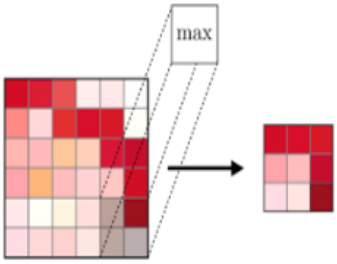
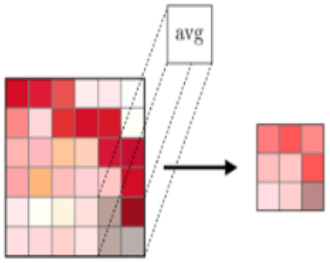


Figure 13. Couche de convolution

2.3.2.2 Couche de Pooling (POOL)

La couche de pooling (en anglais pooling layer) (POOL) est une opération de sous - échantillonnage typiquement appliquée après une couche convolutionnelle. En particulier, les types de pooling les plus populaires sont le max et l'average pooling, où les valeurs maximales et moyennes sont prises, respectivement.

Tableau 2: Différents entre max pooling/averagepooling

Type	**Max pooling**	**Average pooling**
But	Chaque opération de pooling sélectionne la valeur maximale de la surface	Chaque opération de pooling sélectionne la valeur moyenne de la surface
Illustration		
Commentaires	Garde les caractéristiques détectées Plus communément utilisé.	Sous échantillonne la feature map Utilisé dans le SincNet

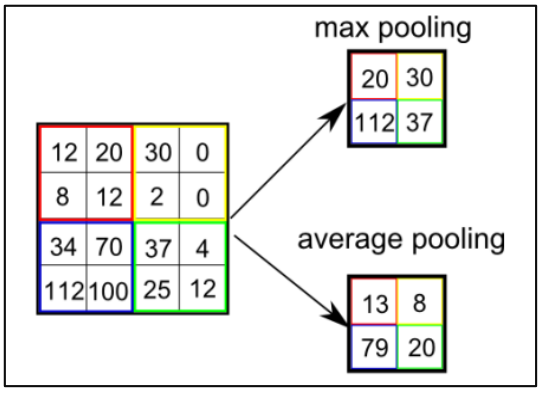


Figure 14. Types de pooling (max /average)

2.3.2.3 Couche de Fully-Connected FC

La couche de fully-connected (en anglaise fullyconnectedlayer) (FC) s'applique sur une entrée préalablement aplatie où chaque entrée est connectée à tous les neurones . Les couches de fullyconnected sont typiquement présentes à la fin des architectures de CNN et peuvent être utilisées pour optimiser des objectifs tels que les scores de class [18].

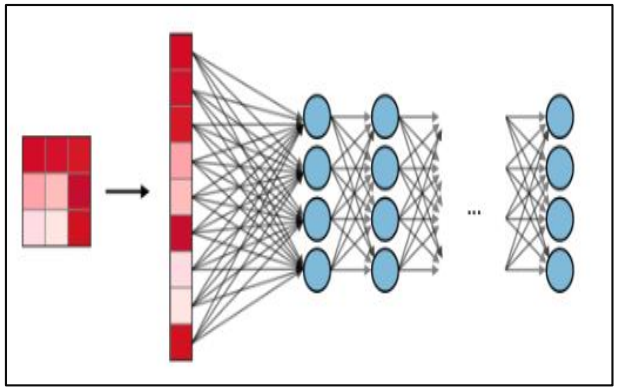


Figure 15. Couche de Fullyconnected (FC)

2.3.2.4 Couche de ReLU

ReLU (RectifiedLinear Unit) est une fonction non linéaire qui est utilisée après une couche de convolution dans l'architecture CNN. Il remplace toutes les valeurs négatives dans la matrice à zéro. Le but de ReLU est d'introduire la non-linéarité dans les CNN pour mieux performer

2.3.3 Réseau de neurones SincNET

SincNet est une architecture neuronale pour le traitement d'échantillons audio bruts. Il s'agit d'un nouveau réseau neuronal convolutif (CNN) qui encourage la première couche convolutive à découvrir des filtres plus significatifs. SincNet est basé sur des fonctions sinc paramétrées, qui implémentent des filtres passe-bande.

Contrairement aux CNN standard, qui apprennent tous les éléments de chaque filtre, seules les fréquences de coupure basses et hautes sont directement apprises à partir des données avec la méthode proposée. Cela offre un moyen très compact et efficace de dériver une banque de filtres personnalisée spécifiquement adaptée à l'application souhaitée.

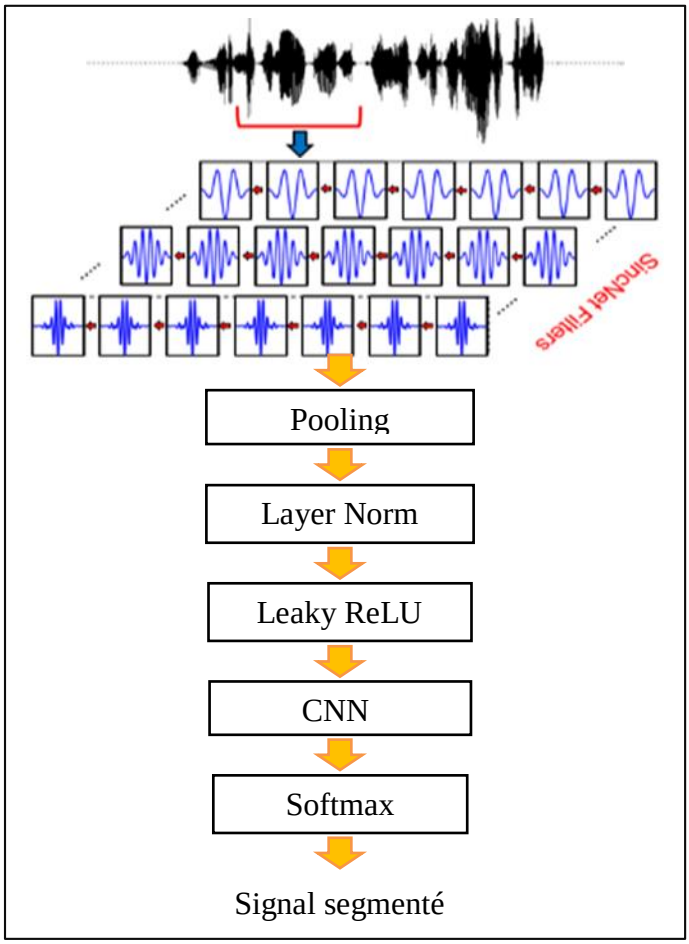


Figure 16. Architecture du réseau SincNet

Le premier étage de SincNet se présente sous la forme d'une convolution basée sur la fonction Sinc, qui utilise des opérations CNN communes (mise en commun, normalisation, activation, fuite et entrée dans des couches multi-convolutionnelles et un empilement de couche de connexion complète (ou couches circulantes) réseaux, et enfin, l'utilisation du classificateur SoftMAX [19].

2.3.3.1 Classifieur Softmax

La fonction d'activation Softmax, également connue sous le nom de SoftArgMax ou fonction exponentielle normalisée, est une fonction d'activation fascinante qui prend des vecteurs de nombres réels comme entrées et les normalise en une distribution de probabilité proportionnelle aux exponentielles des nombres d'entrée. Avant l'application, certaines données d'entrée peuvent être négatives ou supérieures à 1. De plus, elles peuvent ne pas totaliser 1. Après l'application de Softmax, chaque élément sera compris entre 0 et 1, et les éléments totaliseront 1. Cela Ainsi, ils peuvent être interprétés comme une distribution de

probabilité. Pour plus de précisions, plus le nombre d'entrée est grand, plus les probabilités seront grandes.

La fonction Softmax est souvent utilisé dans :

- Réseaux de neurones artificiels et convolutionnels : L'idée est de mapper la sortie non normalisée des données à la distribution de probabilité pour les classes de sortie. Il est utilisé dans les couches finales des classificateurs basés sur des réseaux de neurones. Ils sont formés sous le régime de perte logarithmique ou d'entropie croisée. De cette façon, le résultat est une variante non linéaire de la régression logistique multinomiale (Softmax Régression).

- Autres méthodes de classification multiclasse telles que l'analyse discriminante linéaire multiclasse, les classificateurs na1̃3053′ffs de Bayes, etc.
- Apprentissage par renforcement : La fonction Softmax peut être utilisée pour convertir des valeurs en probabilités d'action.

2.3.3.2 Couche de normalisation (Layer Norm)

La couche de normalisation permet de normaliser l'entrée dans les entités de dimension du lot . Un mini-lot se compose de plusieurs exemples avec le même nombre de fonctionnalités. Les mini-lots sont des matrices (ou tenseurs) où un axe correspond au lot et l'autre axe (ou axes) correspond aux dimensions de l'entité

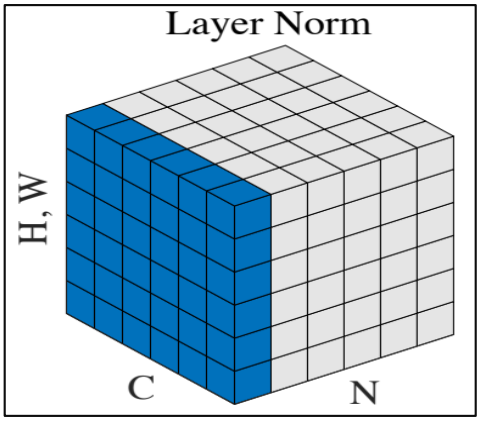


Figure 17. Couche de normalisation (Layer Norm)

2.4 Conclusion

Dans ce chapitre nous avons présenté la segmentation de la parole et le calcule des coefficients MFCC. En outre, nous avons introduit les Deep Learning (définition, architecture...etc.) en donnant une vision globale de Deep Learning, pour comprendre le mechanisme des méthodes choisies dans notre travail. Le chapitre suivant couvrira les différentes simulations et l'implémentation de notre application pour la segementation des signaux de parole en phonèmes.

Chapitre 3

Simulation et Evaluation Expérimentale

3.1 Introduction

Dans ce chapitre, nous allons montrer les différentes étapes pour l'impléméntation de notre système de segmentation automatiuqe de parole. Les simulations sont effectuées à l'aide du logiciel Matlab en utilisant un corpus de parole Arabe monolocuteur. Trois architecteurs sont impléméntées le CNN , le SincNet et le ConstantSinc Layer. La paramétrisation des signaux de parole repose sur les coefficients MFCC.

3.2 Contexte expérimentale

Pour arriver à implémenter notre système nous allons tout d'abord commencer par simulations des trois modèles. Nous citons le matériel necessaires et le logiciel de programmation. Nous présentons en suite le corpus de parole utilisé.

3.2.1 Matériel utilisé

Pour aboutir à nos objectifs, certains équipements sont indispensables :

- Ordinateur (Laptop I5)
- Logiciel de programmation Matlab (Avec toolbox deeplearning et audiobox)
- Corpus de parole Arabe

3.2.2 Logiciel utilisé

MATLAB (Matrix LABoratory) est un logiciel interactif basé sur le calcul matriciel. Il est utilisé dans les calculs scientifiques et les problèmes d'ingénierie parce qu'il permet de résoudre des problèmes numériques complexes en moins de temps requis par les langages de programmation courant, et ce grâce à une multitude de fonctions intégrées et à plusieurs programmes outils testés et regroupés selon usage dans des dossiers appelés boites à outils ou "toolbox". La figure 18 illustre l'interface principal du logiciel Matlab.

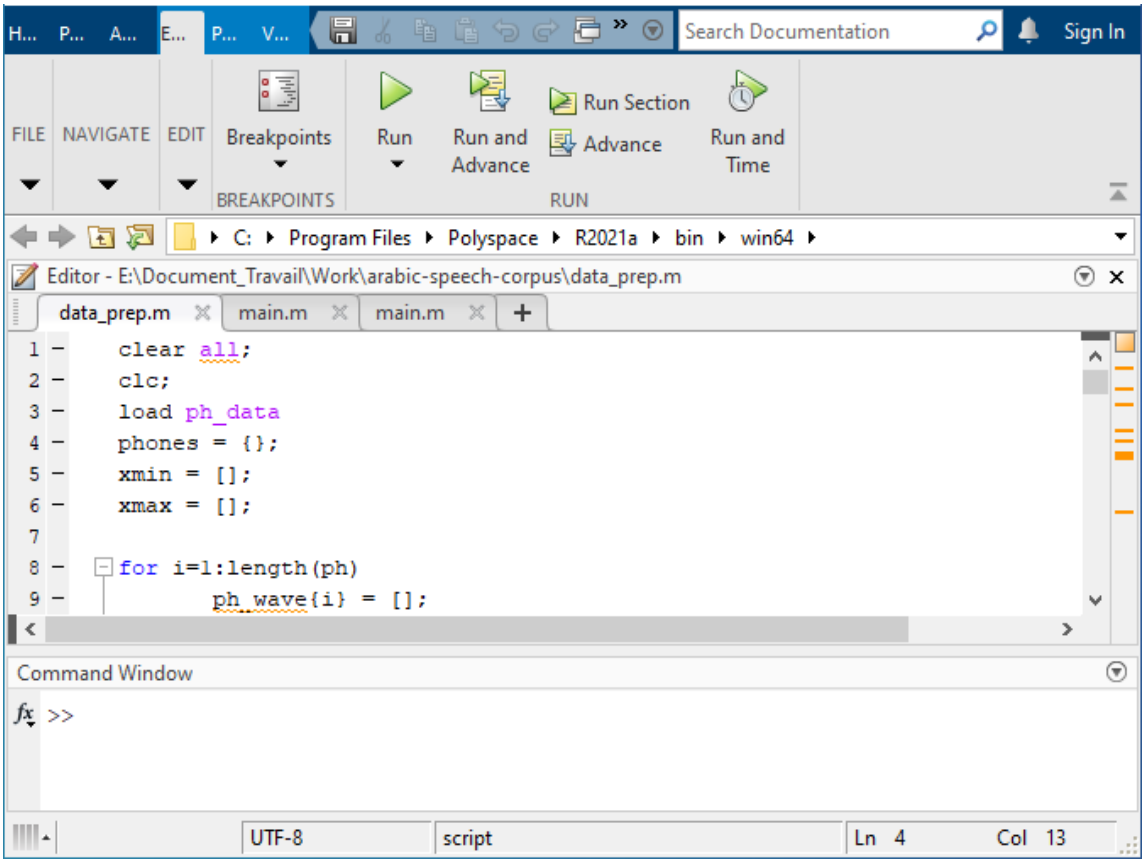


Figure 18. Interface principale du Matlab

a) Deep Learning Toolbox

Il fournit un outil très important pour la conception et la mise en œ uvre de réseaux de neurones profonds avec des algorithmes, des modèles pré-entraïnés et des applications. Vous pouvez utiliser des réseaux de neurones convolutifs (ConvNets, CNN) et des réseaux de mémoire longue à court terme (LSTM) pour effectuer une classification et une régression sur des données d'image, de séries chronologiques et de texte. Vous pouvez créer des architectures de réseau telles que des réseaux antagonistes génératifs (GAN) et des réseaux siamois à l'aide de la différenciation automatique, de boucles d'entraïnement personnalisées et de pondérations partagées. Avec l'application Deep Network Designer, vous pouvez concevoir, analyser et former des réseaux graphiquement.

L'application Expérimente Manager vous aide à gérer plusieurs expériences d'apprentissage en profondeur, à suivre les paramètres d'entraïnement, à analyser les résultats et à comparer le code de différentes expériences. Vous pouvez visualiser les activations de couches et surveiller graphiquement la progression de la formation.

b) Audio Toolbox

Ce toolbox fournit des outils pour le traitement audio, l'analyse de la parole et la mesure acoustique. Il comprend des algorithmes pour le traitement des signaux audio tels que l'égalisation et l'étirement temporel, l'estimation des métriques du signal acoustique telles que le volume et la netteté, et l'extraction des caractéristiques audio telles que le MFCC et la hauteur.

3.2.3 Corpus de parole

Les signaux de parole urilisés sont extraits à partir du corpus de parole monolocuteur "Arabic speech corpus" [4]. Ce corpus a été développé dans le cadre d'une thèse de doctaorat de l'Université de Southampton. Le corpus était enregistré dans un studio professionnel en arabe levantin du sud (Accent damascien). Les signaux sont échantillonnés à une fréquence de $16KHz$ et sont quantifiés à $16bits$.

Ce corpus contient 1813 fichiers audios sous format "wav" et 1813 fichiers des énoncés textuels. Il conteint egalement les fichiers de segmentation en phonème effectué grâce à une segmentation manuelle. Le tableau 3 illustre la transcription orthographique et les symboles phonétiques utilisés dans le corpus, qui contient 35 phonèmes de la langue arabe.

Tableau 3: Phonèmes et leurs notations choisies

Numéro	Phonème arabe	Symbole utilisé	Numéro	Phonème arabe	Symbole utilisé
1	"أ"	<	19	"غ"	g
2	"ب"	b	20	"ف"	f
3	"ت"	t	21	"ق"	q
4	"ث"	^	22	"ك"	k
5	"ج"	J	23	"ل"	l
6	"ح"	H	24	"م"	m
7	"خ"	x	25	"ن"	n
8	"د"	d	26	"ه"	h
9	"ذ"	*	27	"و"	w
10	"ر"	r	28	"ي"	y
11	"ز"	z	29	"ـَ"	a
12	"س"	s	30	"ـُ"	u0
13	"ش"	$	31	"ـِ"	i0
14	"ص"	S	32	"ـَا"	aa
15	"ض"	D	33	"ـُو"	uu
16	"ط"	T	34	"ـِي"	ii
17	"ظ"	Z	35	"sil"	sil
18	"ع"	E			

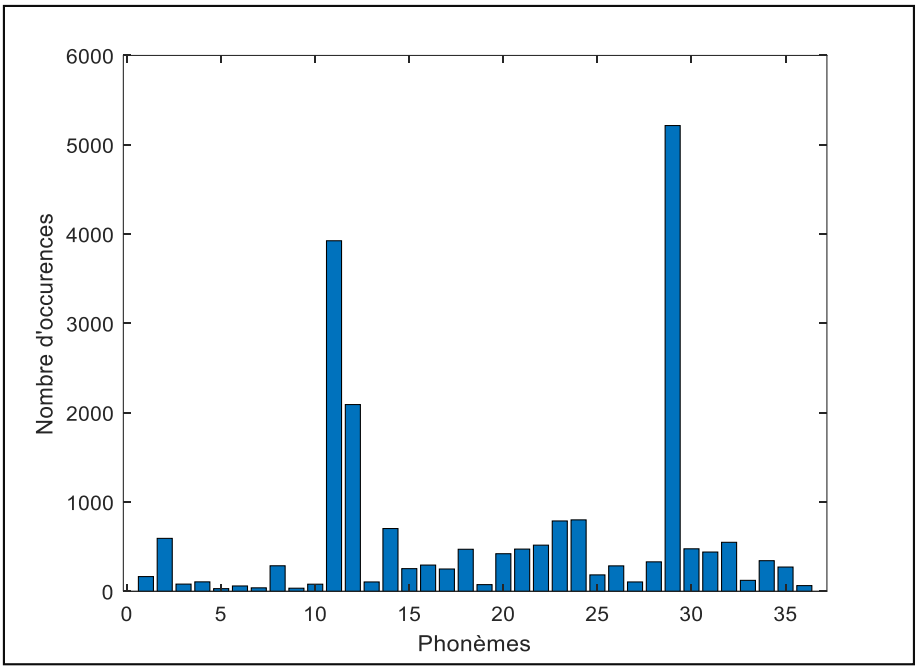


Figure 19. Occurrences de phonèmes dans le corpus

La figure 19 représente les 35 occurrences de phonèmes dans le corpus de parole, qui sont divisés en quatre catégories : moins de 100 Occurrences, de 102 à 400 occurrences, de 401 à 700 occurrences, et plus de 701 occurrences. Le nombre de phonèmes indiqué à la figure 19 a été extrait du tableau 3.

3.2.4 Extraction des paramètres

L'extraction des caractéristiques du signal de parole est une étape fondamentale dans le traitement du signal vocal. Dans notre travail nous nous sommes intéressés à l'étude des paramètres MFCC. Et on voit l'étude des paramètres MFCC se réalise par plusieurs étapes comme suit :

- **Etape 1 :** Découper le signal en plusieurs fenêtres qui se recoupent entre elles. Nous appliquerons la MFCC à chaque fenêtre.
- **Etape 2 :** Afin de diminuer la distorsion spectrale nous appliquons une fenêtre de Hamming. Par la suite nous multiplions cette fonction par le signal à transformer, nous minimisons ainsi la distorsion spectrale créée par le recoupement.
- **Etape 3 :** Appliquer ensuite la FFT à la fenêtre pour en ressortir l'amplitude, on obtient donc le spectre.

- **Etape 4 :** On passe à l'échelle de Mel. En effet, après des études sur l'ou1̃3053′fe humaine, il a été montré que l'homme se base sur une échelle fréquentielle spécifique, pour simuler l'oreille humaine, il faut passer par un banc de filtre, un filtre pour chaque fréquence que l'on cherche. ; Ces filtres ont une réponse de bande passante triangulaire.
- **Etape 5 :** Pour finir, nous travaillons avec le Cepstre, nous convertissons le spectre logarithmique de Mel en temps au moyen de la DCT (Discret Cosinus Transforme), FFT (Fast Fourier Transforme).

Dans notre cas et pour obtenir de meilleures performances en termes de taux de segmentation correcte. Nous avons configuré notre système comme suite :

- Le signal de parole est fenêtré par une fenêtre de Hamming de durée 32ms, avec un chevauchement de 50%.
- Pour chaque trame on extraire 14 coefficients (13 MFCC et 1 Energie)
- Nous avons utilisé en plus les dérivés premiers et seconds ce qui donne un vecteur de 42 coefficients $(13MFC + 1E + 14\Delta + 14\Delta\Delta)$. Il permet de voir l'effet de l'information dynamique contenue dans les paramètres acoustiques.
- Après nous avons construit un vecteur de 336 paramètres en concaténant 8 trames successives. Ce choix est interprité que la durée minimale d'un phonème est à l'ordre de $0.25s$. Ce vecteur est utilisé comme entrée dans les réseaux de neurones implémentés

$$42 \times 8\ trames = 336\ coefficients \tag{18}$$

- Chaque vecteur de paramètres st normalisé en utilisant la méthode CMS (Cepstral Mean Sustraction)

$$c_N = \frac{c_m - M}{S} \tag{19}$$

Avec :

- c_N : vecteur de paramètres normalisé.
- c_m : vecteur de paramètres MFCC.
- M : moyenne des vecteurs de paramètres.
- S : variance des vecteurs de paramètres.

La figure suivante (figure 20) représente les coefficients MFCC extraits d'un signal de parole (corpus de test).

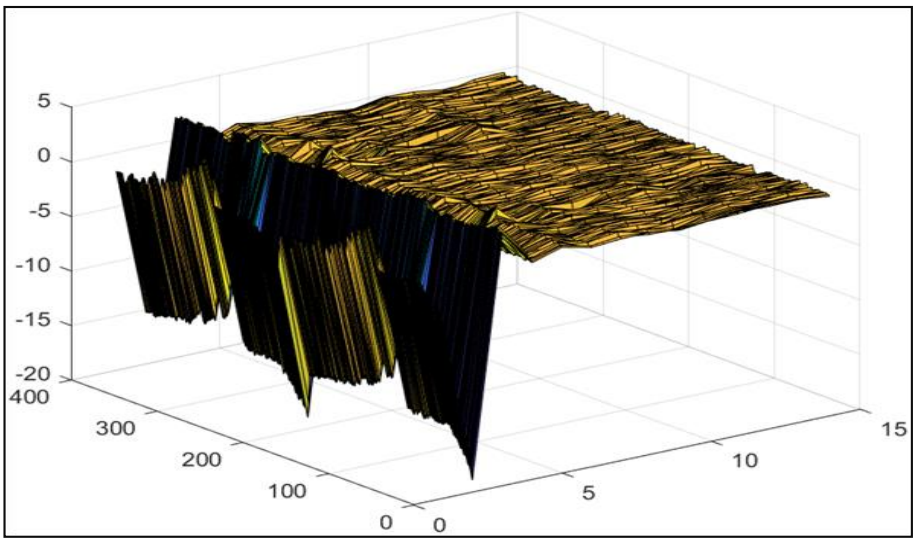
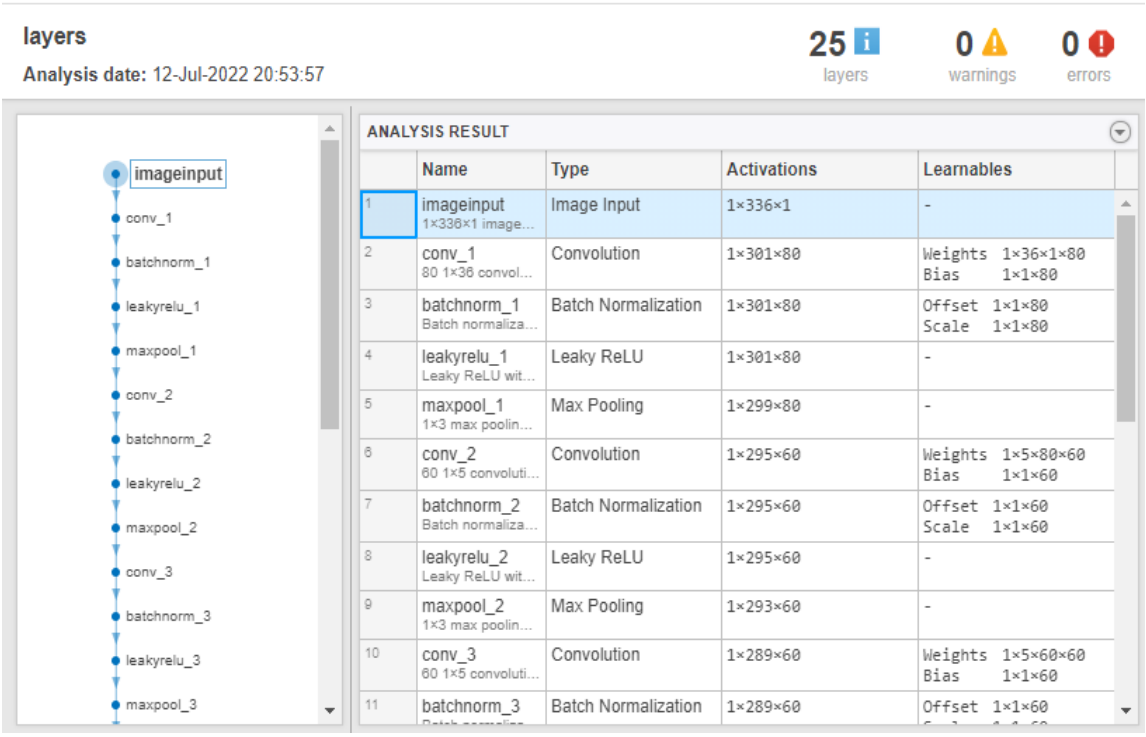


Figure 20. Coefficients MFCC

3.3 Segmentation de parole avec deep Learning

Nous avons mis en place trois architectes d'apprentissage en profondeur on a appliqué ces modèles sur 15 époques pour chaque modèle avec une précision de segmentation différente.



	Name	Type	Activations	Learnables
1	imageinput 1×336×1 image...	Image Input	1×336×1	-
2	conv_1 80 1×36 convol...	Convolution	1×301×80	Weights 1×36×1×80 Bias 1×1×80
3	batchnorm_1 Batch normaliza...	Batch Normalization	1×301×80	Offset 1×1×80 Scale 1×1×80
4	leakyrelu_1 Leaky ReLU wit...	Leaky ReLU	1×301×80	-
5	maxpool_1 1×3 max poolin...	Max Pooling	1×299×80	-
6	conv_2 60 1×5 convoluti...	Convolution	1×295×60	Weights 1×5×80×60 Bias 1×1×60
7	batchnorm_2 Batch normaliza...	Batch Normalization	1×295×60	Offset 1×1×60 Scale 1×1×60
8	leakyrelu_2 Leaky ReLU wit...	Leaky ReLU	1×295×60	-
9	maxpool_2 1×3 max poolin...	Max Pooling	1×293×60	-
10	conv_3 60 1×5 convoluti...	Convolution	1×289×60	Weights 1×5×60×60 Bias 1×1×60
11	batchnorm_3	Batch Normalization	1×289×60	Offset 1×1×60

Figure 21. Implémentation et Apprentissage du système de segmentation

- Dans la première architecture, le modèle est basé sur les réseau de neurones convolutifs CNN. Dans la première couche est une couche convolutive "standard", implémentée à l'aide de convolution2dLayer.
- Dans la deuxième architecture, la première couche convolutive est un banc de filtres sinus constants, mis en œ uvre à l'aide d'une couche personnalisée.
- Dans la troisième architecture, la première couche convolutive est un banc de filtres Sinc entraïnables, mis en œ uvre à l'aide d'une couche personnalisée. Cette architecture est appelée SincNet [20].

La figure 21 représente architecture du système basé sur CNN est composé de trois couches (Conv 1 ,Conv 2 ,Conv3) , batch Norm, Leaky ReLU, Max Pooling. Une couche de sortie composé Fully connected, Soft max.

Le tableau 4 montre les différentes paramètres utilisés pour l'implémentation de notre système.

Tableau 4: Résumé des paramétré utilise

Paramètre	Valeurs
Nombre des filtres	80
Taille des filtres	36
Nombre d'époques	15
Taille minimum du batch	128
Nombre des canaux	1
Itération par époques	105
Nombre totale d'itérations	1575

3.3.1 Apprentissage du modèle CNN

La figure 22 représente l'apprentissage du système de segmentation en phonème par CNN (Convolutional Neural Network) sur 15 époques. Les résultats sont montrés sous forme d'un accuracy. Le système montre des résultats très importants en terme d'accuracy de 97.98% ou bien de temps d'execution (le système montre un accuracy de 97.8274% seulment après 2 époques).

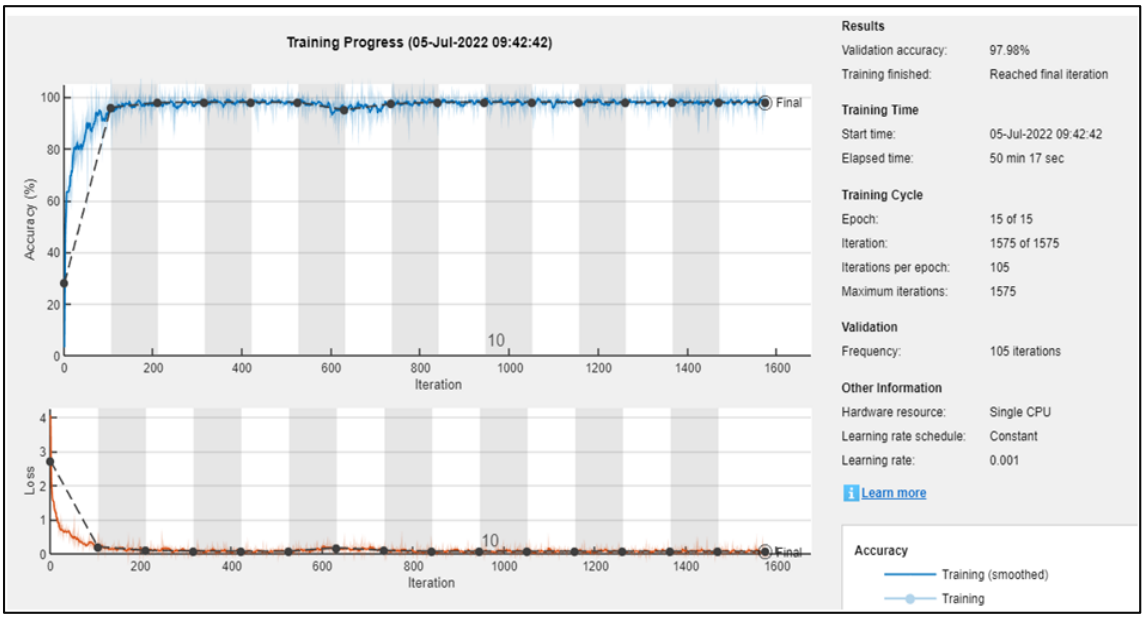


Figure 22. Apprentissage du système de segmentation de parole par CNN

La figure 23 illustre la réponse fréquencielle des différentes filtres de convolutions. La forme de ces filtres n'est pas intuitive et ne correspond pas à une connaissance perceptive.

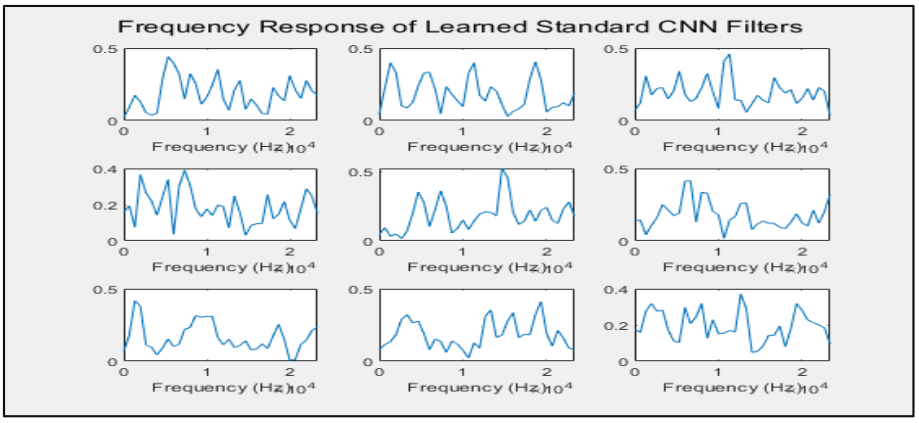


Figure 23. Réponse fréquentielle des filtres CNN

3.3.2 Implémentation de l'architecteur SincNet

Dans cette section, nous remplaçons la première couche convolutive du CNN standard par une couche de banc de filtres Sinc. La couche de banc de filtres à Sinc Net convolue les trames d'entrée avec un banc de filtres passe-bande. Les filtres passe-bande sont une combinaison linéaire de deux filtres Sinc dans le domaine temporel. Les fréquences des

filtres passe-bande sont espacées linéairement sur l'échelle mel. Ce modèle aussi montre des performaces très importantes. Il donne un accuracy final de 97.98%.

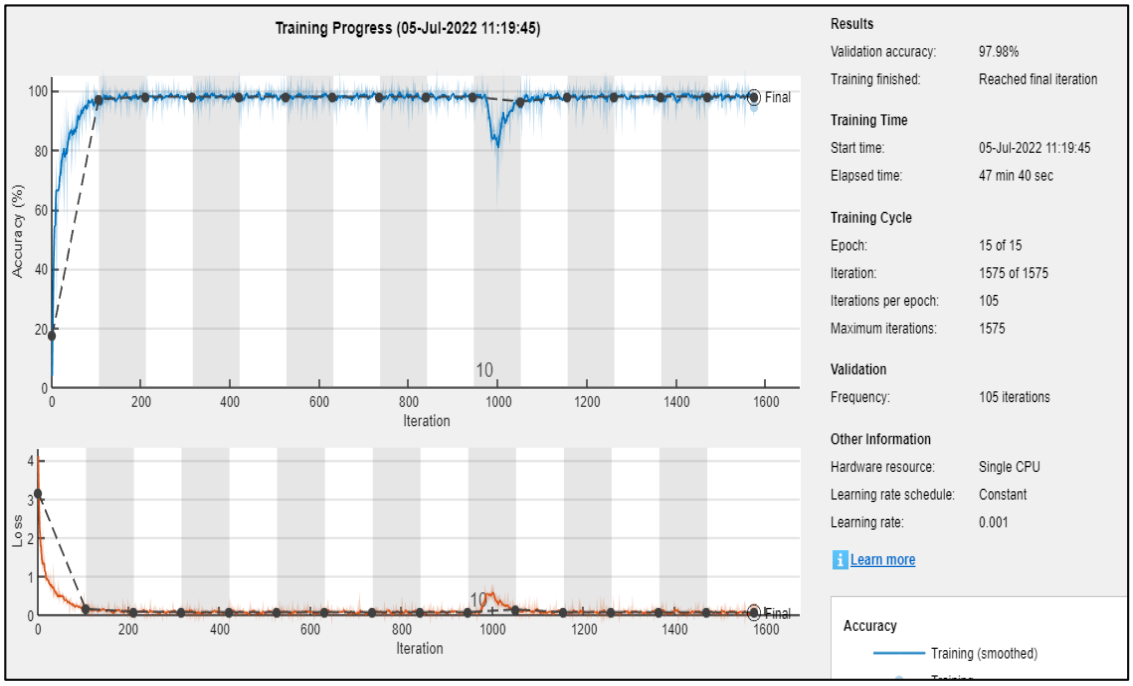


Figure 24. Apprentissage du système de segmentation de parole par SincNet

La figure 25 montre les réponses fréquentielles des filtres d'apprentissage utilisés dans le réseau SincNet.

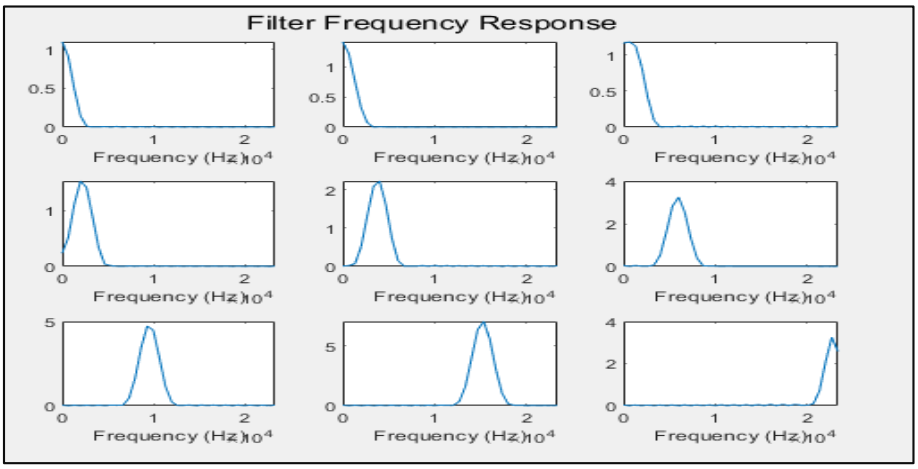


Figure 25. Réponse fréquentielle des filtres SincNet

3.3.3 Implémentation de l'architecteur constant Sinc Layer

Ce modèle d'apprentissage est pratiquement le même que le modèle précédent (Sinc) ou il utilise des fonctions de Sinc lors de la convolution mais avec des largeurs fixes.

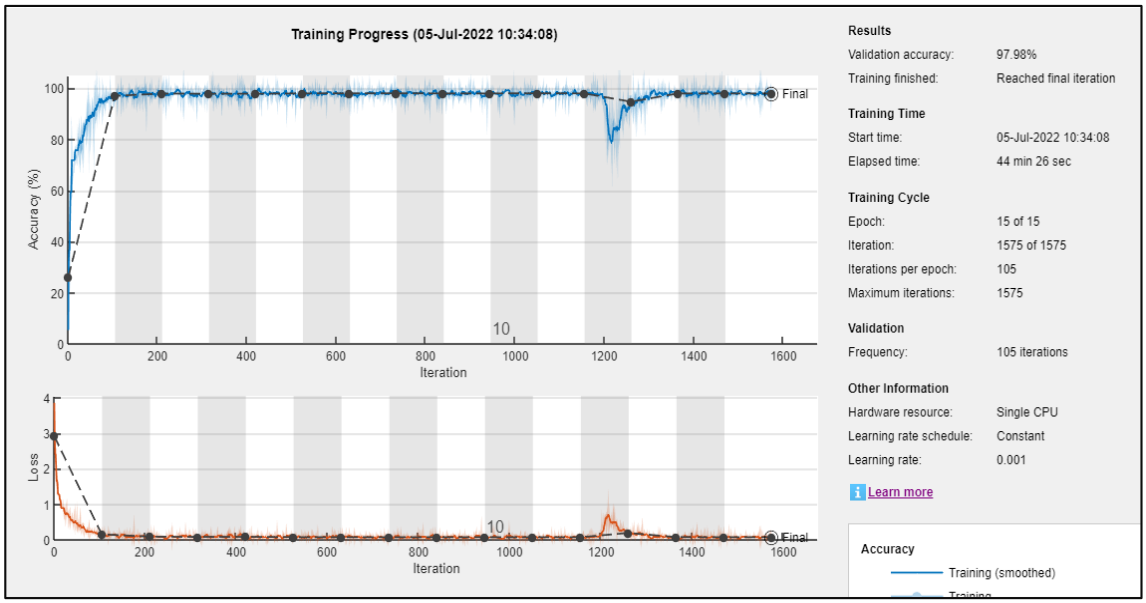


Figure 26. Apprentissage des systèmes de segmentation de parole par ConstantSinc Layer

De même ce système montre des résultats semblables avec ceux obtenus par les deux architectes prcédentes. Le système montre un accuracy de 97.98% (figure 26).

On obtient aussi la réponse de fréquence des filtres (figure 27).

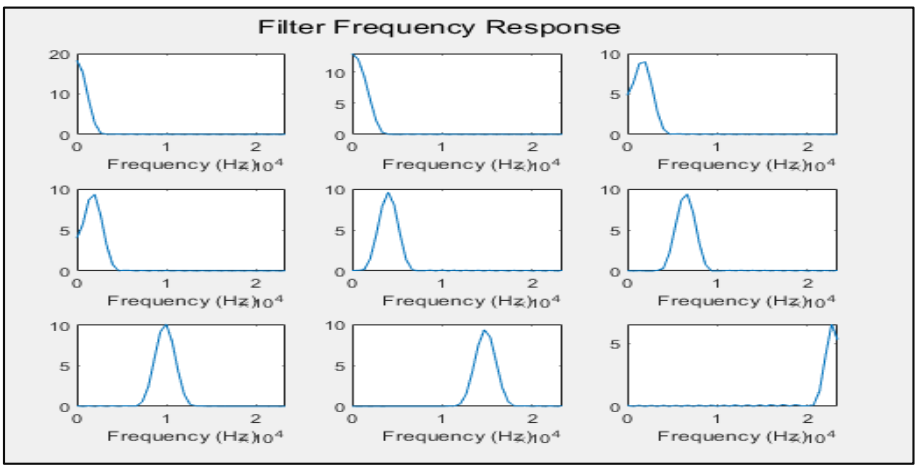


Figure 27. Réponse fréquentielle des filtres ConstantSinc Layer

3.3.4 Résultats de classification

Pour connaitre le meilleur modèle d'apprentissage de la segmentation de la parole dans notre projet on a fait la comparaison entres les trois modèles selon le taux de classification correct en fonction de nombre d'époques. Le résultat montre que les trois modèles est plus élevée

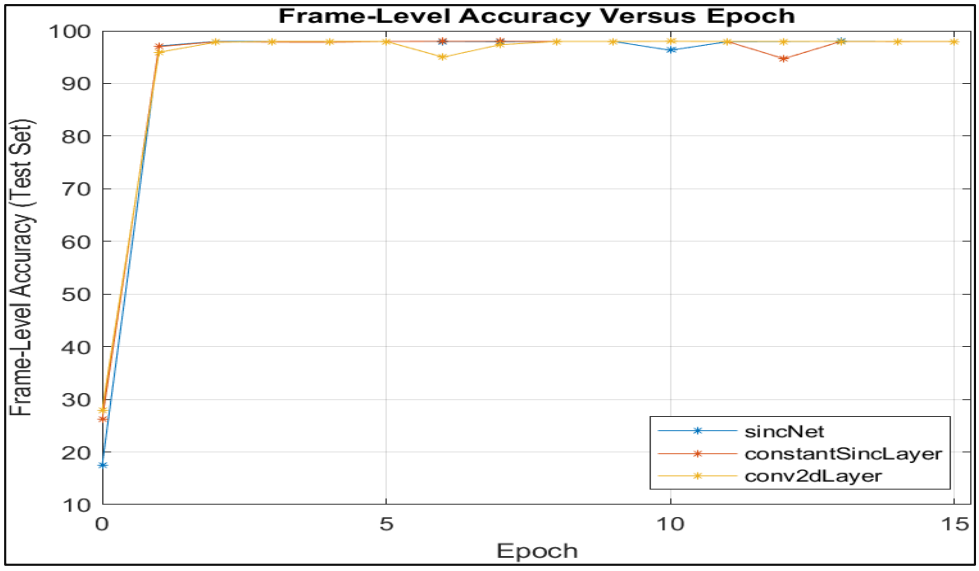


Figure 28. Comparaison entre les trois systèmes

Le tableau suivant montre les nombre de époque (1 à 15 époque) sur pourcentage (%) de taux de classification pour chaque architecteur (CNN, SincNet, ConstantSinc Layer).

Tableau 5: Comparaison entre les trois architecteurs de deep Learning

	Accuracy en %		
Némuro d'epoque	CNN	Sinc Net	ConstSinc
1	95.9226	95.9226	96.994
2	97.8274	97.8274	97.8274
3	97.9762	97.9762	97.9762
4	97.9762	97.9762	97.9762
5	97.9762	97.9762	97.9762
6	95	98.006	98.006
7	97.3512	98.006	98.006
8	97.9762	97.9762	97.9762
9	97.9762	97.9762	97.9762
10	98.006	96.3095	98.006
11	97.9762	97.9762	97.9762
12	97.9762	97.9762	94.7024
13	97.9762	97.9762	97.9762
14	97.9762	97.9762	97.9762
15	97.9762	97.9762	97.9762

3.4 Conclusion

Dans ce chapitre on a présenté la segmentation avec Deep Learning sur Matlab ou on a utilisé trois modèles d'apprentissage. Le système proposé repose sur troix modèles de deep

learning à savoir le CNN, le SincLayer et le ConstantSincLayer. Ces trois modèles montrent des performances très importantes pour la segmentation des signaux de parole Arabe.

Conclusion Générale

Dans ce projet de fin d'étude, nous avons traité le sujet intitulé « Deep Learning pour la segmentation automatique de l'Arabe parlée en phonème », et pu arriver à de très bons résultats. La segmentation de la parole est une opération nécessaire dans le traitement de la parole, qui consisteà découper le signal en segments extrêmement homogènes, pouvant être par al suite transcrits en unités de basephonème. Segment en phonèmes : Cette technique consiste à délimiter la continuité acoustique d'unsignal à une séquence de segments d'un ensemble discret et fini d'éléments, qui estl'alphabetphonétique de la langue (exemple : le mot "Arabe" en le divise en " A, r,a, b et e"). Dans cette étude on a utilisé l'apprentissage profond (deep Learning) selon trois modèles lesquels ont donné des résultats ayant la même précision :

- Réseau neuronal convolutif (CNN.
- Sinc Net Layer.
- Constant Sinc Layer

Ce mémoire a été organisé comme suit : Le premier chapitre, on fait un état de l'art qui donne une vision générale sur la langue Arabe et le traitement de parole. Le deuxième chapitre est consacré à la segmentation de la parole par deep learning en extraction des paramètres MFCC. Ainsi qu'une description plus détaillée sur les calculs de coefficient MFCC et les trois architecteurs (CNN, SincNetLayer, constantSincLayer) choisies dans notre projet. Le troisième chapitre on représente la conception de notre application ainsi que la méthode d'implémentation de notre travail en expliquant l'ensemble des choix techniques (logiciel Matlab) utilisés pour la réalisation de cette application. Les résultats de notre projet ont montré que les trois méthodes d'apprentissage sur 15 époques est de 1575 itérations, et a donné les meilleures performances pour la segmentation automatique de l'Arabe parlée en phonème.

Références Bibliographiques

[1] http://www.universalis.fr/encyclopedie/grammaires-histoire-des-la-tradition-arabe/ mars 2022.

[2] Newman, D. (2002). The phonetic status of Arabic within the world's languages: the uniqueness of the lughat al-daad. Antwerp papers in linguistics., 100:65–75.

[3] (Kouloughli, 2007) Kouloughli, D. (2007). Sur la valeur du tanwin. Nouvelle contribution à l'étude du Système déterminatif de l'arabe. Arabica, 54(1) :94.

[4] Halabi, N. Modern Standard Arabic Speech Corpus. Thèse de doctorat, University Of Southampton 2015.

[5] Watson, J. C . The phonology and morphology of Arabic. Oxford University Press On Demand 2002.

[6] Divenyi, P., Greenberg, S. et Meyer, G . Dynamics of speech production and perception, volume 374. Ios Press 2006.

[7] Vilain, C. E . Contribution à la synthèse de parole par modèle physique. Application À l'étude des voix pathologiques. Thèse de doctorat, Institut National Polytechnique De Grenoble-INPG 2002.

[8] En-Najjary, T. (2005). Conversion de voix pour la synthèse de la parole. Thèse de doctorat, Université Rennes 1(2005).

[9] Jean-François Bonastre pour l'AFCP, 17 novembre 2003.

[10] MAAMRA, OumElhana et SETTOU, Trablesse."Proposition d'un modèle de déscripteur structurel pour la voix arabe, Application saisie des notes ». 73P.Thése de master Academique, UNIVERSITÉ HAMA LAKHDAR D'EL-OUED,2015.

[11] Calliope (ouvrage collectif)."La parole et son traitement automatique". Collection technique et scientifique des télécommunications, CNET - ENST, Masson, 718 pp, 1989.

[12] Samir.N." Segmentation automatique de parole en phones. Correction d'étiquetage par l'introduction de mesures de confiance". Thèse de Doctorat de l'Université de Rennes 1en Informatique, Année 2004.

[13] Oualid.D."Reconnaissance Automatique De La Parole Arabe Par Cmu Sphinx 4 ». Mémoire pour L'obtention Du Diplôme De Magister en électronique, Université Ferhat Abbas -Sétif 1-, Année 2013.

[14] Daniel.M, Sylvain.M, Corinne.F, Laurent.B, Jean-François.B."Segmentation selon le locuteur : les activités du Consortium ELISA dans le cadre de Nist RT03". Avignon Cedex 9-France, Année 2004.

[15] L. MENSOR, A. SLIMANI « Reconnaissance des chiffres manuscrits par les réseaux de neurones artificiels », mémoire d'ingénieur, institut d'électronique, USTHB, Alger (2001).

[16] B. Ryadh, Traitement Automatique De La Parole Arabe par les HMM : Calculatrice Vocale, 2012, pp. 18-23.

[17] [En ligne]. Available : https://www.mckinsey.com/business-functions/mckinsey-analytics/ourinsights/an-executives-guide-to-ai. [Accès le 0ctobre 2022].)

[18] https://blog.csdn.net/gzj_1101/article/details/79376798

[19] (Mirco Ravanelli, Y. Bengio. "Interpretable Convolutional Filters with SincNet". arXiv:1811.09725. 2018)

[20] M. Ravanelli and Y. Bengio, "Speaker Recognition from Raw Waveform with SincNet," 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 1021-1028, 2018.

[21] http://www.claudegabriel.be/ mars 2022.

[22] (Kamina, 2014)