République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté des Sciences
Département d'Informatique

Mémoire de fin d'études
Pour l'obtention du diplôme de Master en SIC

Thème

Utilisation de la Reconnaissance Vocale dans un jeu éducatif pour enfants

Réalisée par :

- **Melle. REZZOUG Wahiba**
- **Mme. MANKOUR Amina**

Encadré par :

- ***Mme. DIDI Fedoua***

Présenté le 18 Juin 2017 devant la commission d'examinassions composée de MM.

- ***Mme. DIDI Fedoua*** *(Promoteur)*
- ***Mme. ILES Nawel*** *(Président)*
- ***Mr. BELHOCINE Amin*** *(Examinateur)*

Année universitaire :2016 - 2017

Remerciements

A l'issue de ce travail, nous remercions, en premier lieu, le bon Dieu de nous avoir donné la force et le courage de le mener à terme.

Un grand remerciement très chaleureuxà notre encadreur

Madame DIDI FEDOUA

pour nous avoir donné l'opportunité de travailler dans un environnement idéal. Qui nous aguidées tout au long de ces mois. Pour sa disponibilité, ses conseils judicieux, sa réactivité et son aisance à faciliter les différentes étapes de ce mémoire.

Nous remerciements à tous les membres du jury pour l'honneur qu'ils nous ont fait en acceptant d'examiner ce travail.

Nous tenons, également, à exprimer notre sincère reconnaissance etnotre profonde gratitude à tous ceux qui ont contribué de près ou de

loin à la réalisation de ce mémoire, notamment tous les enseignants de l'informatique.

Merci

Dédicace

Je tiens à remercier mon dieu le tout puissant pour l'aide qu'il m'a donnée.

Je dédie ce travail à mon père (Allah yerhmah)

A la plus chère de ma vie qui a su m'apporter tendresse et amour, à ma mère.

A mes sœurs, ma belle sœur et mes frères à qui je souhaite tout le bonheur du monde ainsi que la réussite.

A mes nièces Rihame, Hadile, Rihabe, Meriem et mes neveux Yasser et Nadir

A toute ma famille.Petits et grands.

A ma directrice Melle. Hachemaoui Fatine.

Mes collègues de travail.

A l'ensemble des étudiants de la promotion Master 2 informatique 2016/2017, en particulier SIC.

A mon binôme Amina

A toute personne ayant contribué de prés ou de loin à la réalisation de ce mémoire

WAHIBA

Dédicace

Je dédie avec respect ce modeste travail à tous ceux qui m'ont inspirée, aimée et aidée.

Ma chère maman et ma belle mère qui ont partagés mes joies et pensées

Mes très chers petits : Mohamed, Mehdi, Yacine qui ont fait plein de sacrifices pour me voir réussir.

Mes chères sœurs, Fatima, Salima, Malika, Karima et leurs maris et leurs enfants.

A toute ma famille.

Ma directrice de travail Mme. Sakal Leila.

Mes collegues Asma et Khawla.

A mes amis(es) : Souad, Nouria, Bouchra, Imane Medahi, Imane Nor, Souad Sidhom, Houria, Abdellah, Abderrahime, Abderrezzak, Fouad, AlaaAdine qui m'a beaucoup aidée, Abdou.

Mon binôme Wahiba.

A mes collèges de promotion 2016/2017.

A touts mes enseignants de l'informatique.

Finalement une dédicace spéciale pour mon très cher mari Abdelhamid pour son aide pendant mes études.

AMINA

Table des Matières

Chapitre II : Développement

Liste des Figures

Introduction Générale

La parole est le mode de communication le plus naturel. Grâce à elle nous pouvons donner une voix à notre volonté et à nos pensées. Nous pouvons l'utiliser pour exprimer des opinions, des idées, des sentiments ou pour échanger, transmettre, demander des informations. Et aujourd'hui, nous ne l'utilisons pas uniquement pour communiquer avec d'autres humains, mais aussi avec des machines.

La reconnaissance vocale est un domaine de la science ayant toujours eu un grand attrait auprès des chercheurs comme auprès du grand public. En effet, qui n'a jamais rêvé de pouvoir parler avec une machine ou, du moins, piloter un appareil ou un ordinateur par la voix. Ne plus avoir à se lever pour allumer ou éteindre tel ou tel appareil électrique, ne plus avoir à taper pendant des heures sur un clavier pour rédiger un rapport (par exemple). L'homme étant par nature paresseux, une telle technologie a toujours suscité chez lui une part d'envie et d'intérêt, ce que peu d'autres technologies ont réussi à faire.

Le secteur de la reconnaissance vocale est en pleine croissance et nous verrons dans ce document que la technologie actuelle est très aboutie, pouvant commencer à répondre aux attentes de l'homme. Bien que des progrès soient encore à faire sur les systèmes complexes de reconnaissance, il est à noter que la reconnaissance de petits vocabulaires est quasiment parfaite, ce qui suffit largement pour des outils de traitements vocaux du quotidien. Sans compter le coût de ces systèmes qui a considérablement chuté ces dernières années mais aussi le gain qu'ils peuvent apporter à un particulier et surtout à une entreprise. D'où notre intérêt pour cette thématique en vue de développer nos compétences dans ce domaine. On a donc pensé après maintes essais et recherches d'utiliser la reconnaissance vocale dans un jeu éducatif pour enfants, facilitant son apprentissage de la langue en jouant.

Ce travail est composé de deux chapitres

- Dans le premier Chapitre, nous allons présenter d'une manière générale La reconnaissance vocale, ses techniques, ses utilisations..
- Le deuxième chapitre présente les différentes étapes de notre application, nous avons commencé par une modélisation de notre application puis la présentation des outils utilisés dans la réalisation de l'application qui se termine par une série de tests d'exécutions de cette dernière.
- Enfin, une conclusion synthèse notre travail et présente les perspectives envisagées

CHAPITRE I
Généralités sur la Reconnaissance Vocale

1. Introduction

La reconnaissance vocale fait intervenir un processus complexe, permettant à une machine d'identifier des mots ou des phrases en vue d'exécuter un ordre, d'identifier une personne ou bien encore de transcrire la parole humaine en texte écrit.

Elle peut être mise en œuvre par l'identification de spectrogrammes : des représentations graphiques des fréquences sonores émises en fonction du temps.

De plus en plus d'appareils d'utilisation courante utilisent aujourd'hui la reconnaissance vocale.

2. Historique [1]

Les débuts de la reconnaissance vocale remontent vers 1951. S.P. Smith, considéré comme étant le père de la reconnaissance vocale, présenta un détecteur de phonèmes. L'année suivante, un premier système pouvait reconnaître les chiffres de « zero » à « neuf » et était entièrement fait de câbles. Vers le début des années 60, l'amélioration des ordinateurs change entièrement l'orientation des recherches sur la reconnaissance vocale vers le numérique. C'est dans les années 70 que la recherche sur la reconnaissance vocale atteint son apogée avec le projet ARPA et que la première application commerciale de reconnaissance de mots voit le jour. Par la suite, l'amélioration des ordinateurs a facilité la réalisation d'applications de reconnaissance vocale.

Les principaux événements sont (voir Fig 1.1) :

1952 : Reconnaissance de 10 chiffres
1960 : Implantation des méthodes numériques
1965 : Reconnaissance de voyelles et consonnes en parole continue
1968 : Reconnaissance de mots isolés
1971 : Projet ARPA aux États-Unis pour tester la faisabilité de la compréhension automatique de la parole
1972 : Premier appareil commercial de reconnaissance de mots (500 mots)
1976 : Fin du projet ARPA avec 4 systèmes opérationnels

1983 : La reconnaissance vocale est implémentée dans un avion de chasse

1985 : Premier programme commercial pouvant reconnaître des milliers de mots

1986 : Projet japonais ATR pour développer un système de traduction téléphonique en temps réel

1988 : Apparition des premières applications à dicter par mots isolés

1997 : IBM lance le premier programme de dictée vocale en continu.

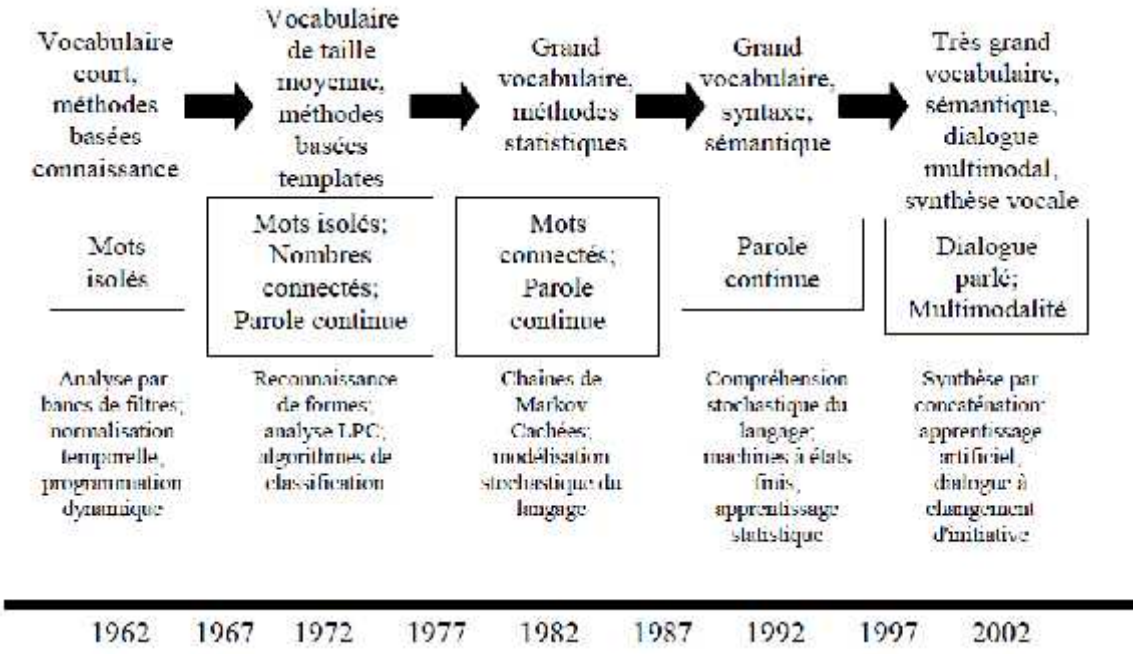


Figure1.1 : Historique du traitement de la parole [2]

3. Définitions

3.1. La reconnaissance vocale [3]

La reconnaissance vocale est la technique qui permet l'analyse des sons captés par un microphone pour les transcrire sous forme d'une suite de mots exploitables par les machines. Depuis son apparition dans les années 1950, la reconnaissance vocale a été constamment améliorée avec l'aide des phonéticiens, linguistes, mathématiciens et ingénieurs, qui ont défini les connaissances acoustiques et linguistiques nécessaires pour bien comprendre la parole d'un humain. Cependant, les performances atteintes ne sont pas parfaites et dépendent de nombreux critères. Les conditions favorables pour la reconnaissance de la parole impliquent une parole native, appartenant à un seul locuteur

ayant une diction propre (ne présentant pas une pathologie de voix), enregistrée dans un environnement calme et non bruité, basée sur un vocabulaire commun (mots connus par le système). La performance du système diminue lorsque l'on traite des accents non-natifs, différents dialectes, des locuteurs qui présentent une pathologie de voix, des mots inconnus par le système (généralement des noms propres), des signaux audio bruités (faible rapport signal-à-bruit), etc.

3.2 Le locuteur

Le locuteur est une personne qui parle, qui énonce quelque chose. Le locuteur désigne celui ou celle qui prend la parole au sein d'un discours oral ou écrit. Il est opposé au destinataire, qui lui reçoit la parole.

3.3 Le son [4]

3.3.1 Définition

Les sons qui parviennent à nos oreilles résultent de vibrations de l'air. Sous une excitation mécanique produite par un instrument de musique ou une personne qui parle, l'air se met à vibrer. Une molécule reçoit alors une impulsion qui la met en mouvement dans une direction donnée. Sur son parcours, elle rencontre d'autres molécules qu'elle pousse, formant ainsi une zone de compression. L'air possède une certaine élasticité, il ne tarde donc pas à se détendre. La matière traversée par l'onde acoustique est alors le siège de compressions et de dépressions successives et périodiques. Ce phénomène crée une onde progressive longitudinale.

Ces mouvements se propagent à une vitesse qui dépend du milieu (élément traversé) et des conditions (température, pression). Dans l'air calme, sous une pression atmosphérique normale et à 20° C, la vitesse de propagation du son est de 340 m/s. Dans un milieu homogène, les vibrations se propagent uniformément dans toutes les directions, mais elles finissent par s'amortir progressivement. L'amortissement est d'autant plus important que la fréquence est élevée. En conséquence, les sons aigus portent moins loin que les sons graves à intensité égale. On sait maintenant pourquoi nos voisins n'aiment pas nos chaînes HIFI. Pour que le son se propage, la présence d'un

milieu élastique est indispensable : le son ne se propage pas dans le vide. Ajoutons également que les molécules vibrent sur place. Un son sortant d'un haut-parleur n'a jamais créé de courant d'air !

3.3.2 Les sons simples et sons complexes

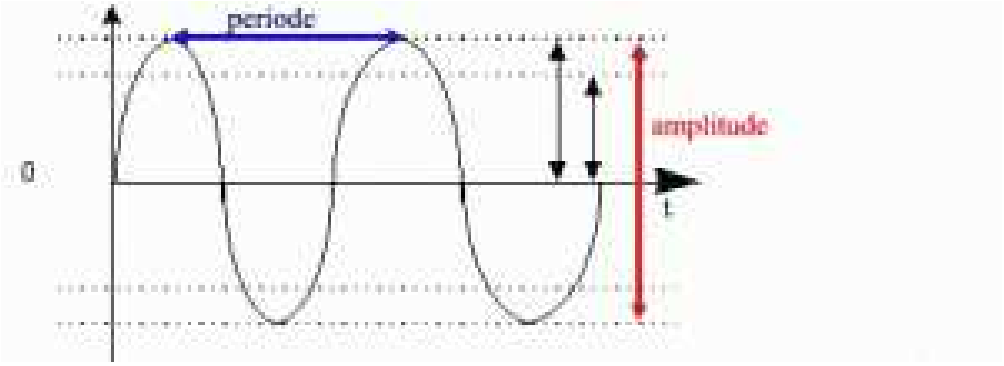


Figure 1.2Graphique d'un son simple ou pur

Ce graphique représente un son simple ou "pur", mais ce cas est très rare. En effet, la plupart des sources sonores produisent des sons complexes qui sont constitués par une fréquence fondamentale sur laquelle se superposent des harmoniques et des transitoires. Les harmoniques sont des multiples entiers de la fréquence fondamentale. On distingue les harmoniques pairs (2f, 4f, 6f, 8f…) et les harmoniques impairs (3f, 5f, 7f, 9f…). Les sons sont donc formés d'une superposition de vibrations ayant des amplitudes très variables dont la courbe résultante est très irrégulière. Le théorème de Fourier permet de démontrer qu'un signal complexe peut être décomposé en une série de signaux sinusoïdaux simples. Les transitoires sont constituées à l'établissement ou l'extinction d'un son d'une manière brusque comme par exemple les percussions, les cymbales, ou l'attaque d'une note de piano. Les transitoires ne sont pas décomposables en série de Fourier.

3.3.3 Les caractéristiques d'un son

Un son est défini par 3 paramètres : sa fréquence, son amplitude, et son timbre. Toutes les opérations que réalisent les logiciels de traitement du son tournent autour de ces 3 caractéristiques.

3.3.3.1 L'intensité

L'intensité d'un son correspond à l'amplitude de la vibration acoustique. En d'autres termes, elle caractérise le volume sonore qui nous permet de distinguer un son fort d'un son faible.

3.3.3.2 La hauteur

La hauteur d'un son est liée à la vitesse de vibration de l'air, c'est-à-dire la fréquence. Les variations de la fréquence fondamentale permettent de situer un son sur l'échelle des graves et des aigus. Le spectre audible (16 à 20000 Hz) est divisé en octaves qui couvrent un intervalle de fréquences dans un rapport de 1 à 2. La fréquence de référence est représentée par la note la3 dont la valeur a été fixée à 440 Hz.

3.3.3.3 Le timbre

Le timbre est donné par les harmoniques et les transitoires qui accompagnent la fréquence fondamentale. Il permet de différencier deux sons de même hauteur et de même amplitude. C'est ainsi que l'on reconnaîtra, à l'oreille, deux instruments de musique jouant une même note ou une personne qui parle. Le timbre est constitué d'un ensemble de fréquences appelé spectre. La richesse du spectre permettra de dire qu'un son est riche, brillant, profond …

3.3.4 La perception des sons

3.3.4.1 La pression acoustique

Nous savons maintenant que le son résulte d'une variation périodique de la pression de l'air. Les ondes produites transmettent une certaine quantité d'énergie. Dans le Système International (SI), l'unité de pression est le Newton par mètre carré (N/m^2) nommé en France le pascal (Pa). L'atmosphère est à environ 1020 hPa.

3.3.4.2 Le diagramme de Fletcher et Munson

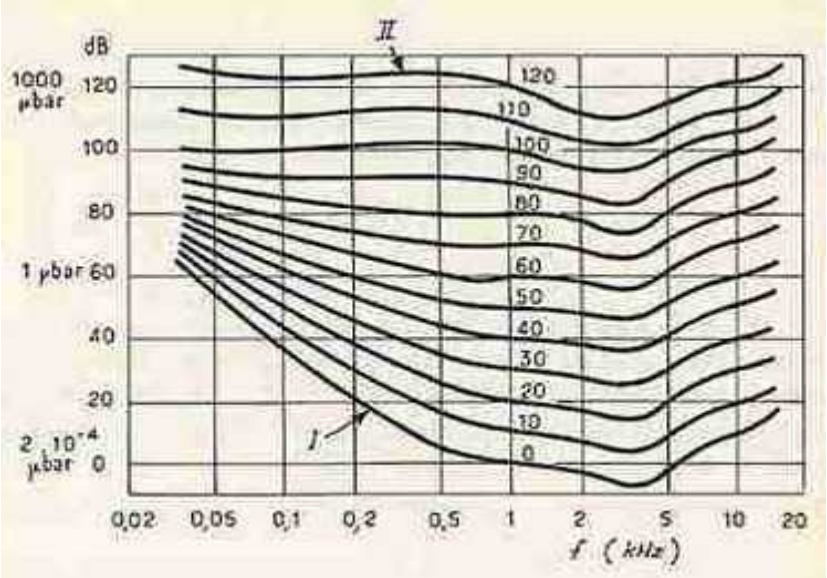


Figure 1.3 Le diagramme de Fletcher et Munson

Les courbes classiques de Fletcher et Munson mettent en relation la fréquence et la puissance sonore. Elles donnent ainsi la sensibilité moyenne de l'oreille en fonction de la fréquence. Lorsque l'on applique un signal sonore d'intensité croissante, l'oreille ne perçoit ce signal qu'à partir du seuil d'audibilité (courbe I). Puis l'oreille perçoit l'augmentation progressive d'énergie jusqu'à un niveau où l'audition devient douloureuse. L'oreille est saturée et aucune augmentation de sensation n'est plus perceptible. Le seuil de douleur est atteint (courbe II). Ce diagramme montre également que notre oreille est plus sensible aux fréquences médiums situées entre 500 et 5000 Hz (l'oreille est plus adaptée à la parole, notre principal moyen d'expression !). Les lignes intermédiaires relient les points du diagramme pour lesquels la sensation de volume est égale. Ces courbes résultent d'une étude statistique et correspondent donc à une oreille "moyenne". Dans la réalité, ces valeurs dépendent de nombreux facteurs comme par exemple l'âge ou l'état de santé d'un individu.

3.3.5 Le décibel

La sensibilité de l'oreille n'est pas linéaire car la sensation varie comme le logarithme de l'excitation. Pour doubler la sensation acoustique il faut multiplier par 10 la puissance sonore. C'est pourquoi on utilise le décibel pour définir le niveau

acoustique par rapport au seuil de sensibilité. Le décibel (dB) est la dixième partie du Bel. C'est l'unité de mesure du niveau sonore.

3.4 Le signal [5]

Le signal est le support de l'information émise par une source et destinée à un récepteur ; c'est le véhicule de l'intelligence dans les systèmes. Il transporte les ordres dans les équipements de contrôle et de télécommande, il achemine sur les réseaux l'information, la parole ou l'mage. Il est particulièrement fragile et doit être manipulé avec beaucoup de soins. Le traitement qu'il subit a pour but d'extraire des informations, de modifier le message qu'il transporte ou de l'adapter aux moyens de transmission ; c'est là qu'intervienne les techniques numériques. En effet si l'on imagine de substituer au signal un ensemble de nombres qui représentent sa grandeur ou amplitude à des instants convenablement choisis, le traitement, même dans sa forme la plus élaborée, se ramène à une séquence d'opérations logiques et arithmétiques sur cet ensemble de nombres, associées à des mises en mémoire.

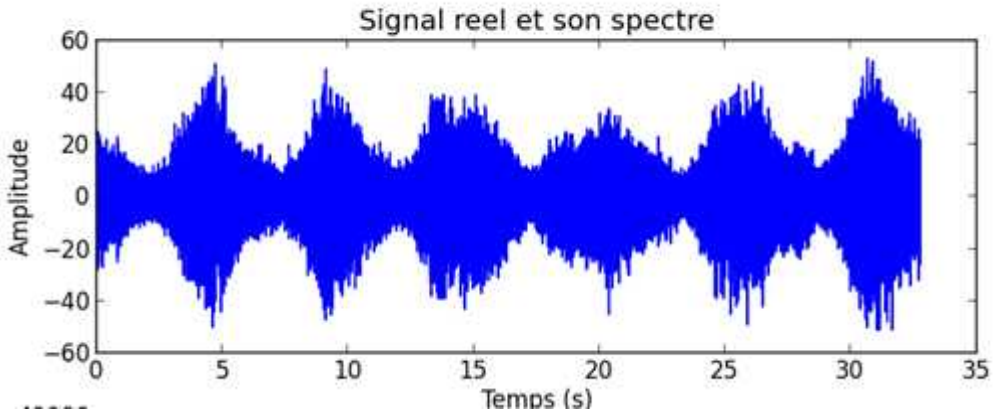


Figure 1.4 Représentation d'un signal sonore

3.5 Un fichier audio numérique : [6]

La reconnaissance vocale se base sur la comparaison de fichiers audio, ainsi nous devons tout d'abord maîtriser le format d'enregistrement utilisé avant d'effectuer des opérations de transformation du signal. On distingue deux types de format, les

formats compressés et les formats non compressés. Le format WAVE (Waveform) est un dérivé de la spécification RIFF (Resource Interchange File Format) de Microsoft dédiée au stockage de données multimédias. Ce format est libre d'utilisation et est sûrement le plus répandu parmi les nombreux formats de fichiers sons. Ce format est lisible sur la plupart des systèmes d'exploitation et par n'importe quel logiciel de traitement de son digne de ce nom. Le seul problème avec ce format est qu'il est évolutif et peut connaître de nombreuses formes (compressions audio, etc.). Nous allons donc nous limiter au format PCM (Pulse Code Modulation) dans lequel les échantillons sont codés de manière "brute" (aucune compression). Les logiciels d'édition sonore nécessitent également que les sons soient dans ce format pour pouvoir les éditer. Des logiciels comme Audacity permettent tout de même d'importer des fichiers mp3 qu'il reconvertit d'abord en Wave.

4. Les avantages les plus importants de la reconnaissance vocale:[7]

- Les tâches de travail sont réalisées de manière plus efficace car le temps de traitement des documents est réduit. Grâce à la reconnaissance vocale, les documents sont établis jusqu'à 3 x plus rapidement que par la rédaction.
- L'utilisation de la reconnaissance vocale diminue largement la charge de travail, notamment pour le secrétariat qui n'a plus normalement que quelques petites corrections à apporter aux documents.
- La reconnaissance vocale s'améliore à chaque correction d'erreur de reconnaissance. Le taux de reconnaissance peut ainsi encore progresser.
- Les dictées enregistrées par les dictaphones numériques sont transformées en texte sans difficulté à l'aide de la reconnaissance vocale.
- Pour ceux qui utilisent la dictée numérique, l'adaptation est facile et permet de bénéficier des avantages de la reconnaissance vocale.
- La reconnaissance vocale est disponible avec un lexique spécifique, par exemple pour les avocats. Une adaptation personnalisée selon les branches et les domaines spécifiques peut améliorer le taux de reconnaissance.
- La reconnaissance vocale permet de travailler sans aucune barrière.

- La reconnaissance vocale améliore l'efficacité, aboutit à un travail plus structuré et procure avant tout beaucoup de plaisir, car rien n'est plus fascinant que la transformation rapide des paroles en texte clair.

5. Les domaines d'application [8]

La reconnaissance vocale est une technologie relativement récente, qui séduit énormément les professionnels et le grand public par son aspect pratique, ludique et ergonomique.

5.1 L'État, l'armée et les renseignements :

Le domaine d'application des programmes de reconnaissance vocale est particulièrement vaste. Tout d'abord, il est impossible d'aborder ces technologies sans parler des problématiques de défense militaire. L'enjeu stratégique et militaire de la maîtrise de ce procédé s'impose comme une évidence à l'heure des guerres industrielles et des menaces terroristes. À ce sujet, les excellentes performances des logiciels à destination des particuliers laissent supposer que les systèmes utilisés par l'armée et les services de renseignement bénéficient d'une puissance de calcul et d'une capacité de traitement particulièrement importantes.

5.2 Le domaine professionnel :

En second lieu, on pourra bien sûr parler des applications professionnelles. Les programmes de reconnaissance vocale sont largement utilisés dans le domaine médical. Leur utilisation accélère le traitement des rapports et comptes rendus. Ce domaine n'est évidemment pas le seul à bénéficier de ces technologies, de nombreux autres corps de métiers tertiaires utilisent la reconnaissance vocale (justice, bureau d'études, etc.).

5.3 Le domaine personnel :

Enfin, les particuliers ne sont pas en reste. La reconnaissance vocale est présente partout autour de nous et de nombreuses personnes l'utilisent au quotidien sans même s'en rendre compte. On pourra citer l'exemple

- Des téléphones portables capables de composer un numéro lorsque l'on prononce le nom d'un contact donné. De nombreux standards téléphoniques utilisent aussi la reconnaissance pour naviguer sur un serveur.

- Bien sûr, les personnes qui sont amenées à taper de longs textes pour un usage personnel ou professionnel simplifieront leur travail en utilisant un logiciel de reconnaissance vocale.
- Dans un autre registre, il existe un certain nombre d'applications médicales capables de faciliter la vie des personnes attentes de lourds handicaps.
- Apprentissage d'une langue étrangère.
- Identification vocale dans les zones sécurisées grâce à la signature vocale
- La messagerie
- La possibilité de commande et de contrôle d'appareils à distance.

6. Fonctionnement des logiciels de reconnaissance vocale

Les logiciels de reconnaissance vocale, utilisent la voix humaine comme principale interface entre l'utilisateur et l'ordinateur. Bien qu'étant simple d'utilisation, un logiciel de reconnaissance vocale utilise une technologie hautement sophistiquée appliquant la « modélisation linguistique » pour reconnaître et distinguer les millions d'énoncés qui composent une langue. En utilisant des modèles statistiques, les programmes de reconnaissance vocale analysent un flux de son entrant et interprètent ces sons en tant que commandes ou dictées. Le procédé d'interprétation est appelé reconnaissance vocale. Son taux de réussite correspond au pourcentage d'interprétations correctes.

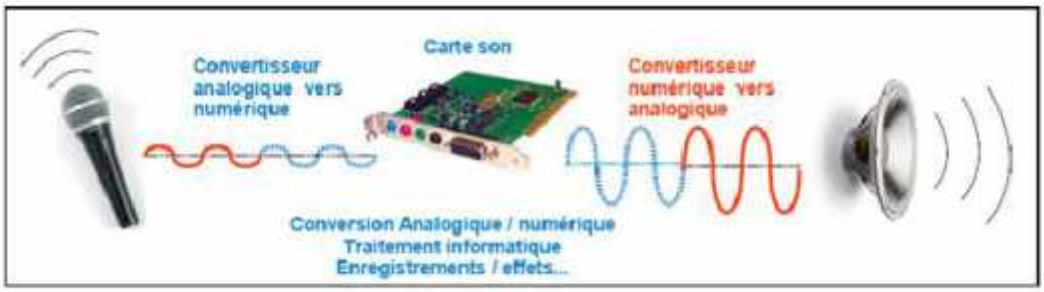


Figure 1.5 Exemple d'une chaine numérique

7. Complexité du problème

Pour appréhender le problème de reconnaissance automatique de la parole, il est bon d'en comprendre les différents niveaux de complexité et les différents facteurs qui en font un problème difficile.

Le signal parole est des plus complexes, et sujet à beaucoup de variabilités. Ayant été produit par un système phonatoire humain complexe, il n'est pas facile de le caractériser à partir d'une simple représentation bidimensionnelle de propagation de sons. On peut distinguer certaines de ses caractéristiques comme les sons élémentaires ou phonèmes, la hauteur, le timbre, l'intensité, la vitesse... Mais en réalité, la voix est beaucoup plus complexe qu'on ne peut percevoir par l'oreille. L'onde sonore varie non seulement avec les sons prononcés, mais également avec les locuteurs. La très grande qui peut chanter, créer, murmurer, être enroué ou enrhumé, et aussi selon le locuteur lui-même (homme, femme, enfant, voix nasillarde, différences de timbre), sans parler des accents régionaux, rend très délicate la définition d'invariants. Il faut pouvoir séparer ce qui caractérise les phonèmes, qui devrait être une constante quel que soit le locuteur et sa prononciation, de l'aspect particulier à chaque locuteur. Qu'est ce qui permet à notre cerveau de distinguer un mot d'un autre, indépendamment de celui qui nous parle ? De plus, la mesure du signal de parole est fortement influencée par la fonction de transfert du système de reconnaissance (les appareils d'acquisition et de transmission), ainsi que par le milieu ambiant. Ainsi l'obstacle majeur d'avoir une grande précision de la reconnaissance, est la grande variabilité des caractéristiques d'un signal vocal. Cette complexité du signal de parole provient de la combinaison de plusieurs facteurs, la redondance du signal acoustique, la grande variabilité inter et intra locuteur, les effets de la coarticulation en parole continue, et les conditions d'enregistrement.

7.1 Redondance du signal de parole

Quiconque a vu une représentation graphique de l'onde sonore a certainement été frappée par le caractère répétitif du signal de parole. Un grossissement à la loupe d'une brève émission de parole donne à voir une succession de figures sonores semblant se répéter à l'excès. Un peu de recul laisse apparaître des zones moins stables qu'il convient, de qualifier de transitoires. Ce qui semblerait de prime abord superflu, s'avère en réalité fort utile. Les répétitions confèrent à ce signal une robustesse. La redondance

le rend résistant au bruit. Dans une certaine mesure, elle fonctionne comme un code correcteur d'erreur, puisqu'un interlocuteur humain sait décrypter un message même s'il est entaché de bruits dus à de possibles interférences.

7.2 Une grande variabilité

À contenu phonétique égal, le signal vocal est très variable pour un même locuteur (variabilité intra locuteur) ou pour des locuteurs différents (variabilité interlocuteur).

7.2.1 Variabilité intra-locuteur

Une même personne ne prononce jamais un mot deux fois de façon identique. La vitesse d'élocution en détermine la durée. Toute affection de l'appareil phonatoire peut altérer la qualité de la production. Un rhume teinte les voyelles nasales; une simple fatigue et l'intensité de l'onde sonore fléchit, l'articulation perd de sa clarté. La diction évolue dans le temps: l'enfance, l'adolescence, l'âge mûr, puis la vieillesse, autant d'âges qui marquent la voix de leurs sceaux.

7.2.2 Variabilité interlocuteur

Est encore plus flagrante. Les différences physiologiques entre locuteurs, qu'il s'agisse de la longueur du conduit vocal ou du volume des cavités résonnantes, modifient la production acoustique. En plus, il y a la hauteur de la voix, l'intonation et l'accent différent selon le sexe, l'origine sociale, régionale ou nationale. Enfin toute parole s'inscrit dans un processus de communication où entrent en jeu de nombreux éléments comme le lieu, l'émotion, l'intention, la relation qui s'établit entre les interlocuteurs. Chacun de ces facteurs détermine la situation de communication, et influe à sa manière sur la forme et le contenu du message

7.3 La continuité

La production d'un son est fortement influencée par les sons qui le précédent et le suivent en raison de l'anticipation du geste articulatoire. L'identification correcte d'un segment de parole isolé de son contexte est parfois impossible. Évidement il est plus simple de reconnaître des mots isolés bien séparés par des périodes de silence que de

reconnaître la séquence de mots constituant une phrase. En effet, dans ce dernier cas, non seulement la frontière entre mots n'est plus connue mais, de plus, les mots deviennent fortement articulés.

7.4 Le système est-il robuste ?

Autrement dit, le système est-il capable de fonctionner proprement dans des conditions difficiles? En effet, de nombreuses variables pouvant affectés significativement les performances des systèmes de reconnaissance ont été identifiées :

- Bruits d'environnement (dans une rue, un bistrot etc.…) ;
- Déformation de la voix par l'environnement (réverbérations, échos, etc.…) ;
- Qualité du matériel utilisé (micro, carte son etc.…) ;
- Bande passante fréquentielle limitée (fréquence limitée d'une ligne téléphonique) ;
- Elocution inhabituelle ou altérée (stress, émotions, fatigue, etc.…).

Certains systèmes peuvent être plus robustes que d'autres par rapport à l'une de ces perturbations, mais en règle générale, les systèmes de reconnaissance de la parole sont encore sensibles à ces perturbations.

8. Approche de la reconnaissance [9]

Il existe deux approches permettant d'aborder la reconnaissance de la parole : l'approche globale et l'approche analytique [5]. Elles se distinguent essentiellement par la nature et par la taille des unités abstraites qu'elles s'efforcent de mettre en correspondance avec le signal de parole.

8.1 Approche globale

Dans l'approche globale, l'unité de base sera le plus souvent le mot considéré comme une entité globale, c'est à dire non décomposée. L'idée de cette méthode est de

donner au système une image acoustique de chacun des mots qu'il devra identifier par la suite.

8.2 Approche analytique

L'approche analytique, qui tire parti de la structure linguistique des mots, tente de détecter et d'identifier les composantes élémentaires. Celles-ci sont les unités de base à reconnaître. Cette approche a un caractère plus général que la précédente : pour reconnaître de grands vocabulaires, il suffit d'enregistrer dans la mémoire de la machine les principales caractéristiques des unités de base. Dans l'approche globale, l'unité de base est le mot : le mot est considéré comme une entité indivisible. Une petite phrase, de très courte durée, peut aussi être considérée comme un mot. Dans l'approche analytique, on tente de détecter et d'identifier les composantes élémentaires de la parole que sont les phonèmes.

9. La classification du son [10]

Il existe deux grandes familles de son : le son analogique et le son numérique.

Le son analogique est un signal électrique continu pour lequel il existe une valeur de tension en concordance avec la variation de la pression de l'air. Analogique vient du mot "analogue" ce qui signifie "ressemblance". En effet, un son analogique est enregistré de façon analogue à l'onde sonore qu'elle produit. Un son est une pression d'air qui vibre en fonction du temps. Cette pression est captée par un microphone et transforme cette pression en tension.

Le son numérique est représenté par une suite binaire de 0 et de 1. L'exemple le plus évident de son numérique est le CD audio Lorsqu'un son est enregistré à l'aide d'un microphone, les variations de pression acoustique sont transformées en une tension mesurable. Il s'agit d'une grandeur analogique continue représentée par une courbe variant en fonction du temps. Un ordinateur ne sait gérer que des valeurs numériques discrètes. Il faut donc échantillonner le signal analogique pour convertir la tension en une suite de nombres qui seront traités par l'ordinateur. C'est le rôle du convertisseur analogique/numérique. Ainsi, la numérisation permet de transformer un

signal sonore en fichier enregistré sur le disque dur de l'ordinateur, c'est le procédé permettant la construction d'une représentation discrète d'un objet du monde réel. Dans son sens le plus répandu, la numérisation est la conversion d'un signal audio en une suite de nombres permettant de représenter cet objet en informatique ou en électronique numérique. On utilise parfois le terme franglais digitalisation (digit signifiant chiffre en anglais).

10. Techniques de modulation [11]

En transmission, un des problèmes essentiels est d'adapter le signal transmis au support de communication. La transmission en bande de base (sans modulation) utilise le câble coaxial, la paire torsadée ou la fibre optique connue support de transmission pour acheminer les trains d'impulsions. Par contre, la transmission sur canal téléphonique et la transmission à large bande font appel à des techniques de modulation dont nous allons étudier les principaux types.

10.1 Modulation d'impulsions en amplitude (PAM)

Cette technique de modulation consiste à varier l'amplitude de chaque impulsion en fonction de l'amplitude du signal analogique. La figure 6 a) en est une illustration.

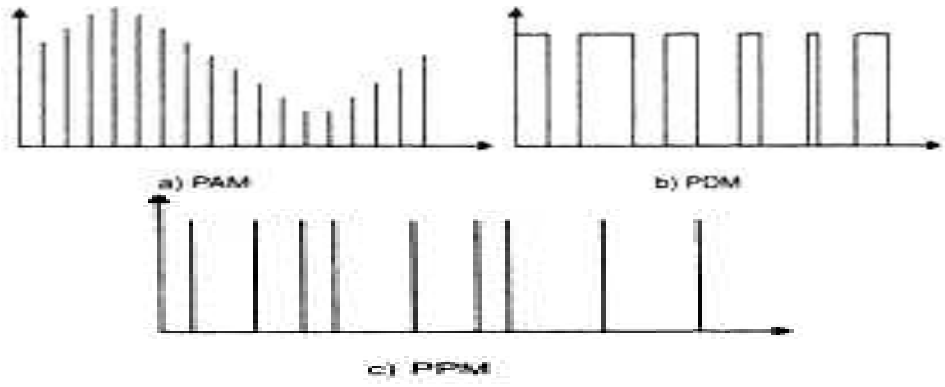


Figure 1.6 Les trois types de modulation d'impulsions.

10.2 Modulation d'impulsions en durée (PDM)

La modulation d'impulsions en durée consiste à varier la durée des impulsions en fonction de l'amplitude du signal analogique (figure 6 b).

10.3 Modulation d'impulsions en position (PPM)

La modulation d'impulsions en position (figure 6 c) consiste à varier les intervalles de temps entre des impulsions identiques en fonction de l'amplitude de l'information analogique. On peut facilement générer un signal modulé en position à partir d'un signal PDM à l'aide d'un simple monostable.

10.4 Modulation MIC différentielle (DPCM)

Nous avons vu que dans un système MIC, chaque échantillon du signal est quantifié et codé. Or, il arrive souvent que le changement de niveau d'un échantillon à l'autre soit petit et, de ce fait, il est avantageux de coder uniquement la différence entre deux échantillons successifs.

Les schémas bloc d'un codeur MIC différentiel et d'un décodeur MIC différentiel sont donnés aux figures 6.a. et 6.b. On voit, qu'au moment de l'échantillonnage, une tension proportionnelle à la différence entre le signal d'entrée filtré et un autre signal analogique généré à partir de la valeur numérique de l'échantillon précédent est présente à l'entrée de l'échantillonneur-bloqueur. Cette différence sera ensuite quantifiée et codée. En pratique la MIC différentielle offre la même performance avec un code à 4 bits que le PCM avec un code à 8 bits, nécessitant deux fois moins de bande passante pour transporter efficacement le signal codé.

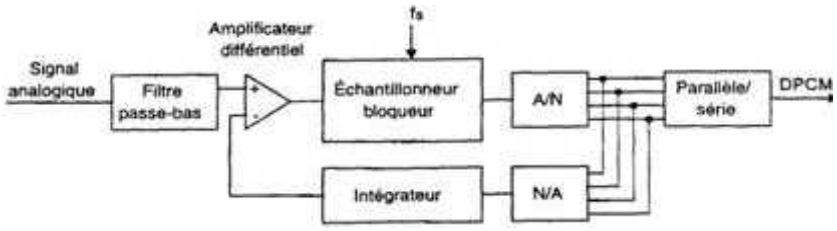


Figure 1.7 Schéma bloc d'un codeur MIC différentiel.

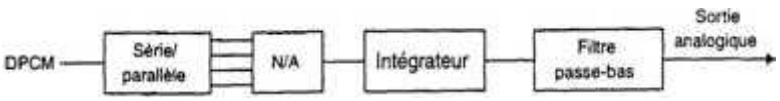


Figure 1.8 Schéma bloc d'un décodeur MIC différentiel.

10.5 Modulation Delta (DM)

Au lieu de coder la valeur de chaque échantillon ou la différence entre deux échantillons successifs, la modulation Delta code uniquement le sens d'évolution (ou la dérivée) du signal analogique et transmet un seul bit par échantillon. La synchronisation sera alors plus simple et la réalisation matérielle moins complexe. Un modulateur Delta linéaire est représenté à la figure 8. Le signal binaire obtenu à la sortie est semblable à celui obtenu par un modulateur d'impulsions en durée (PDM).

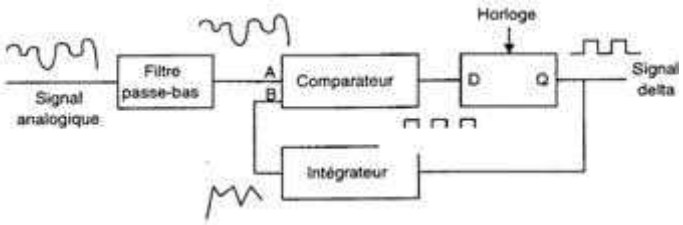


Figure 1.9 Schéma bloc d'un modulateur Delta linéaire.

10.6 Modulation Delta adaptative ou à pente continuellement variable (CVSDM)

Le principe de la CVSDM est assez simple: on mémorise dans un registre quelques bits (3 ou 4) de sortie du codeur pour prévoir la situation de surcharge et ainsi augmenter ou diminuer la pente de l'intégrateur. Donc, la pente s'ajuste constamment pour réduire l'erreur entre la sortie de l'intégrateur et le signal original.

Le schéma bloc d'un modulateur CVSDM apparaît à la figure 10

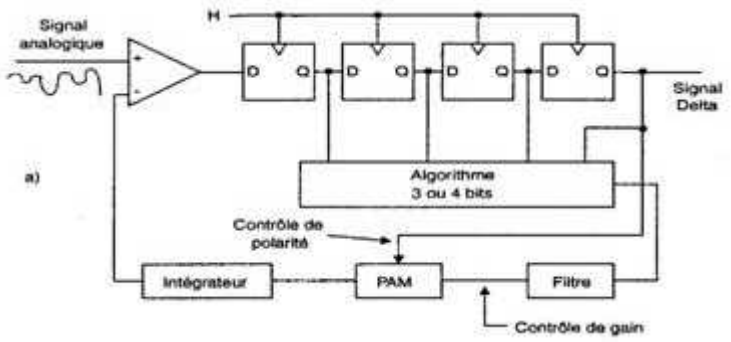


Figure 1.10 Schéma bloc d'un modulateur CVSDM

10.7 Modulation par impulsions codées (MIC)

Plusieurs travaux ont été publiés pour comparer les différentes techniques en modulation. Les meilleurs résultats ont été obtenus en utilisant la méthode MIC.

La modulation MIC (PCM pour Pulse Code Modulation) est une technique qui consiste à convertir un signal analogique en une série d'impulsions binaires codées.

Ainsi, pour réaliser un système de modulation MIC, trois opérations de base sont indispensables :

- échantillonnage
- quantification
- codage

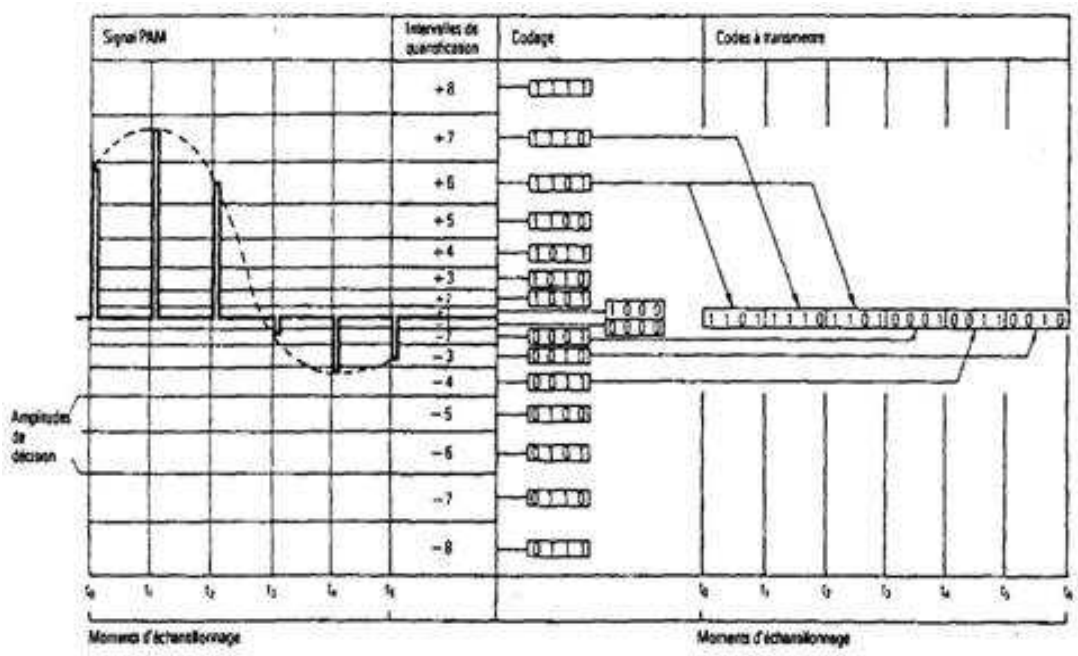


Figure 1.11 Principe de ce type de modulation.

N.B : On va détailler ce modèle dans le point suivant (Traitement du signal Numérique page 19)

11 Traitement du signal

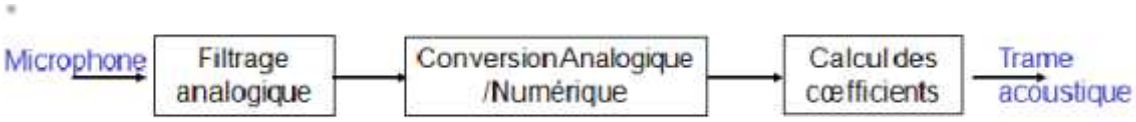


Figure 1.12 Numérisation d'un son analogique

11.1 Traitement du signal analogique [12]

Le traitement du signal analogique repose essentiellement sur l'utilisation d'opérateurs linéaires qui modifient les propriétés d'un signal de façon homogène dans le temps. La transformées de Fourier diagonalise ces opérateurs et apparait donc comme le principal outil d'analyse mathématique. Nous étudions la synthèse de filtres homogènes et une application à la transmission par modulation d'amplitude.

11.2 Traitement du signal Numérique

La numérisation est le procédé permettant la construction d'une représentation discrète d'un objet du monde réel. Dans son sens le plus répandu, la numérisation est la conversion d'un signal (vidéo, image, audio, caractère d'imprimerie, impulsion) en une suite de nombres permettant de représenter cet objet en informatique ou en électronique numérique. On utilise parfois le terme franglais digitalisation (digit signifiant chiffre en anglais).Lorsqu'un son est enregistré à l'aide d'un microphone, les variations de pression acoustique sont transformées en une tension mesurable. Il s'agit d'une grandeur analogique continue représentée par une courbe variant en fonction du temps. Un ordinateur ne sait gérer que des valeurs numériques discrètes. Il faut donc échantillonner le signal analogique pour convertir la tension en une suite de nombres qui seront traités par l'ordinateur. C'est le rôle du convertisseur analogique/numérique Ainsi, la numérisation permet de transformer un signal sonore en fichier enregistré sur le disque dur de l'ordinateur.

La numérisation se réalise en trois étapes, l'échantillonnage, la quantification et le codage. Elle va permettre de transformer un signal continu en une suite de valeurs discrètes (distinctes) qui seront traduites dans le langage des ordinateurs, en 0 et 1.

11.2.1 L'échantillonnage :

L'échantillonnage est le passage d'un signal continu en une suite de valeurs discrètes (discontinues).

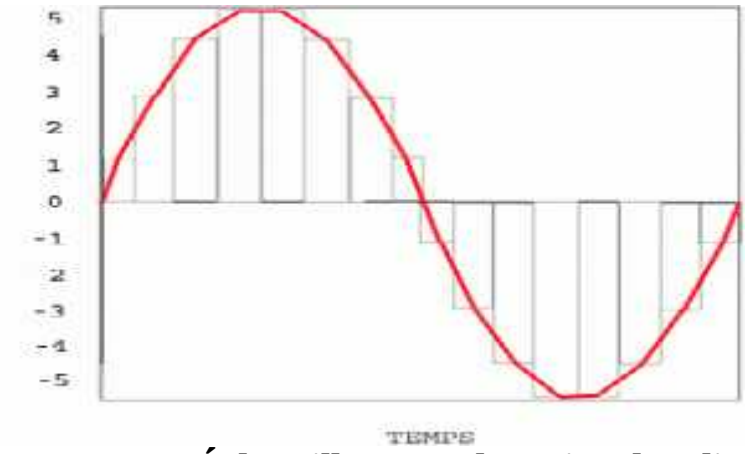


Figure 1.13 Échantillonnage d'un signal audio

C'est la première phase de la numérisation qui consiste à passer d'un signal à temps continu en une suite de valeurs mesurées à intervalles réguliers. Le signal analogique est ainsi découpé en "tranches" ou échantillons (samples). Le nombre d'échantillons par seconde d'audio représente la fréquence d'échantillonnage ou sampling rate. Celle-ci est exprimée en Hertz (Hz).

1 KHz = 1000 Hz

La fréquence d'échantillonnage d'un signal audio n'est pas choisie arbitrairement. Elle doit être suffisamment grande, afin de préserver la forme du signal. Le Théorème de Nyquist – Shannon stipule que la fréquence d'échantillonnage d'un signal doit être égale ou supérieure au double de la fréquence maximale contenue dans ce signal. Si la fréquence choisie est trop faible, les variations rapides du signal analogique ne seront pas enregistrées. Ainsi pour un fichier de qualité téléphonique, on échantillonnera à 11,025 KHz – 8 bits – mono. Cela permettra de traiter des fréquences allant jusqu'à 5500 Hz, ce qui est largement suffisant pour rendre une voix parfaitement compréhensible.

Pour un enregistrement audio en qualité CD, la bande passante étant généralement de 20 KHz, on échantillonnera à 44,1 KHz – 16 bits – stéréo.

Dans un projet audio, pour la réalisation d'un CD de musique par exemple, on choisira dès le départ une résolution de 24 bit et une fréquence d'échantillonnage de

44.100 kHz (ou un multiple pair de cette fréquence 88.2 kHz , 176.4 kH). Les fréquences d'échantillonnage de 48 KHz, 96 KHz ou 192 Khz sont plutôt utilisées dans des projets vidéo et ne donnent pas le meilleur résultat lors de la conversion finale en 44.100 kHz. On verra plus loin dans ce dossier, la technique du dithering pour passer à une résolution numérique inférieure. Naturellement, la place occupée par notre fichier sur le disque dur sera fonction de la qualité choisie. Ces graphiques montrent l'influence de la fréquence d'échantillonnage :

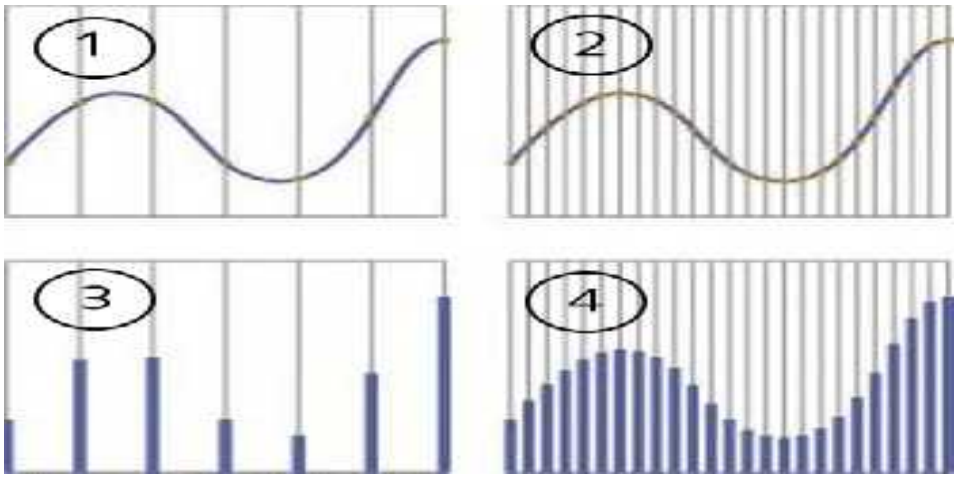


Figure 1.14 : Les figures 1 et 2 représentent les signaux analogiques. Les figures 3 et 4 montrent ces mêmes signaux après numérisation.

On notera que le signal est dans les 2 cas codé sur le même nombre de bits.

Dans la figure 4 la fréquence d'échantillonnage du signal analogique est le quadruple de celle utilisée en 3.On constate que plus la fréquence d'échantillonnage est élevée, plus le signal numérique se rapproche de la définition du signal analogique.

11.2.2 La quantification:

C'est la seconde phase de la numérisation. Après avoir découpé le signal continu en échantillons, il va falloir les mesurer et leur donner une valeur numérique en fonction de leur amplitude. Pour cela, on définit un intervalle de N valeurs destiné à couvrir l'ensemble des valeurs possibles. Ce nombre N est codé en binaire sur 8-16-20 ou 24 bits suivant la résolution du convertisseur A/N. L'amplitude de chaque échantillon est alors représentée par un nombre entier.

Codage sur 8 bits = 2 puissance 8 = 256 valeurs possibles

Codage sur 16 bits = 2 puissance 16 = 65536 valeurs possibles

Codage sur 20 bits = 2 puissance 20 = 1.048.576valeurs possibles
Codage sur 24 bits = 2 puissance 24 = 16.777.216 valeurs possibles.

Nous donnons ci-dessus le nombre de valeur possibles que peut prendre un échantillon. Cela signifie qu'en 16 bits cette valeur varie entre 0 et 65535 (en réalité en audio entre -32768 et +32767) et en 24 bits entre 0 et 16 777 215 (en réalité entre – 8 388 608 et + 8 388 607).

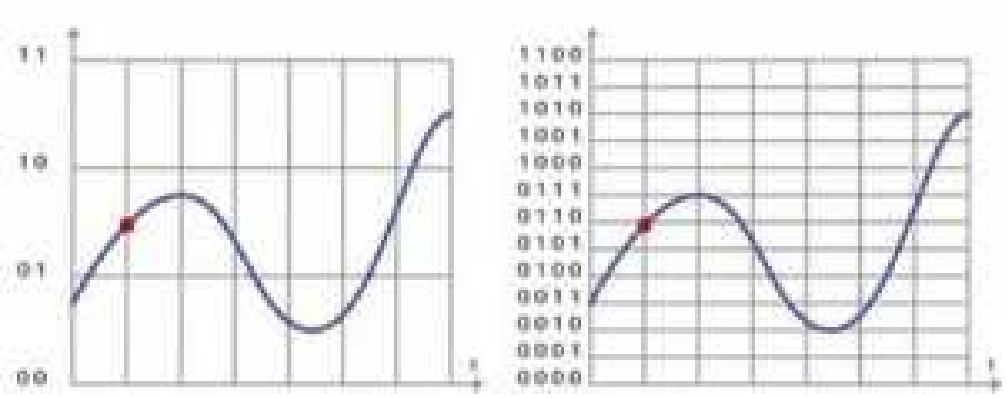


Figure 1.15 Signal échantillonné avant et après quantification

A gauche codage sur 2 bits: la valeur attribuée au point rouge sur la courbe est 01
A droite codage sur 4 bits: on peut attribuer à ce même point une valeur plus précise 0110

Le signal audio est maintenant numérisé. Notons que nombre de bits définit aussi l'amplitude dynamique du signal (6 dB/bit). Un équipement ayant une résolution de 16 bits offrira une dynamique maximale de 16*6=96 dB, pour une résolution de 8 bits, 8*6=48 dB. Plus l'encodage sera puissante, plus la dynamique sera élevée et le bruit de fond limité.

11.2.3 Le codage binaire:

On associe à chaque bit un poids. Ce poids est fonction de la position du bit dans l'octet.

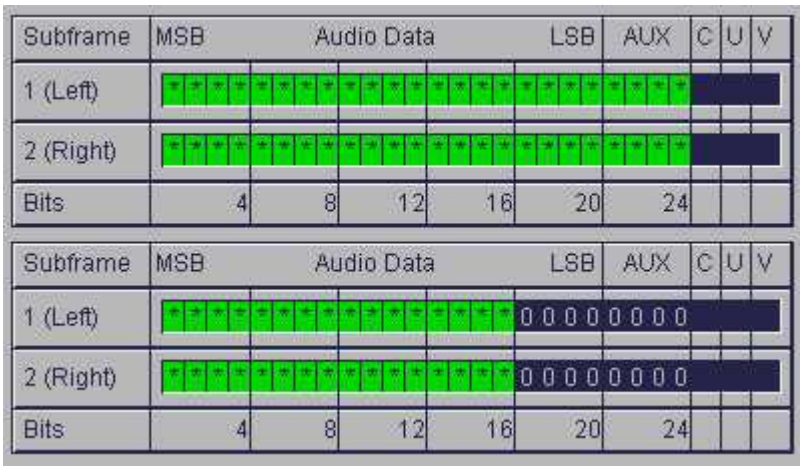


Figure 1.16 Le codage binaire

LSB : (Least significant Bit) bit de poids faible, c'est le bit qui a le moins de signification dans un octet. Par convention, c'est le bit le plus à droite dans l'écriture d'un mot. Exactement comme dans le nombre 105, c'est le chiffre le plus à droite qui à le moins de poids, la valeur la plus faible.

MSB : (Most significant bit) bit de poids fort, c'est le bit qui a le plus de signification dans un octet. Par convention, c'est le bit le plus à gauche dans l'écriture d'un mot. Exactement comme dans le nombre 105, c'est le chiffre le plus à gauche qui à le plus de poids, la valeur la plus forte.

12 Les modèles de la reconnaissance automatique [13]

12.1 Modèle de langage

Les modèles de langage statistiques sont des processus qui permettent d'estimer les probabilités des différentes séquences de mots P(W) = P(w1;w2; :::;wm□1;wm). Ces modèles servent à 'mémoriser' des séquences de mots à partir d'un corpus textuel d'apprentissage. Dans le contexte de la reconnaissance de la parole, les modèles de langage servent à guider et à contraindre la recherche parmi les hypothèses de mots alternatives.

12.2 Modèle de prononciation

Le lexique d'un système de reconnaissance vocale précise une ou plusieurs prononciations pour chaque mot. Pour le français, les prononciations multiples sont en partie dues aux événements de liaison ou de réduction, dans le cadre desquels un locuteur peut prononcer ou pas un certain phonème dans un certain contexte. Les accents et les dialectes peuvent aussi générer diverses variantes de prononciations.

La liaison implique la prononciation d'un phonème de liaison entre deux mots. Pour donner un exemple : les mots "les oiseaux" se prononcent séparément "l e" et "w a z o" (en notation API), alors qu'ensemble ils se prononcent "l e z w a z o".

La réduction implique l'omission d'un phonème a priori présent dans la prononciation standard d'un mot, comme dans le cas de "ce" qui se prononce normalement "c @", mais qui peut également être prononcé simplement "c" dans le cas d'une prononciation rapide. Les variantes de prononciation peuvent être obtenues manuellement, avec l'expertise des linguistes, ou automatiquement - avec des convertisseurs graphèmes-phonème basés sur des techniques comme les JMM (joint-multigrammodels) ou les CRF (conditional random fields). Le graphème est une lettre ou un ensemble de lettres représentant un phonème. Les règles de correspondance entre graphèmes et phonèmes sont complexes, irrégulières et spécifiques à chaque langue. En général, la conversion graphèmes-phonèmes peut être exprimée comme ~Q = ArgMaxQ P(Q;G), où G est l'orthographe du mot (séquence de graphèmes) et Q est une prononciation candidate. La méthode JMM applique un modèle de langage sur des couples {séquence de graphèmes, séquence de phonèmes}. L'algorithme d'apprentissage vise à déterminer l'ensemble optimal des séquences de graphèmes et de phonèmes ainsi que le modèle de langage associé, de façon incrémentale : un passage initial crée un modèle très simple, ensuite, chaque nouvelle passe d'apprentissage affine le modèle en agrandissant les séquences (si possible).

La méthode CRF modélise la distribution des probabilités conditionnelles d'une séquence d'étiquettes (la séquence de phonèmes) étant donnée une séquence d'observation (la séquence de graphèmes). En l'absence d'un corpus de données pré-étiquetées, les modèles HMM discrets peuvent être utilisés pour aligner les phonèmes avec les lettres.

12.3 Modèle acoustique

Le modèle acoustique est un modèle statistique qui estime la probabilité qu'un phonème ait généré une certaine séquence de paramètres acoustiques. Une grande variété de séquences de paramètres acoustiques sont observées pour chaque phonème en raison de toutes les variations liées à la diversité des locuteurs, à leur âge, leur sexe, leur dialecte, leur état de santé, leur état émotionnel, etc.

13 L'analyse du signal [14]

Le traitement numérique des signaux connaît depuis trois décennies un développement fulgurant. Une multitude de méthodes puissantes de traitement des signaux peuvent désormais être mise en œuvre grâce aux techniques numériques. L'étude de la parole a été un des domaines importants qui a bénéficié et qui continue de bénéficier du traitement numérique des signaux. L'étape d'analyse du signal est une opération essentielle, elle a pour but de fournir une représentation moins redondante du signal de la parole que celle obtenue par codage de l'onde temporelle tout en permettant une extraction précise des paramètres significatifs et pertinents. Le signal analogique est fourni en entrée et une suite discrète de vecteurs, appelée trame acoustique est obtenue en sortie. Mais avant tout traitement il faut discrétiser le signal continu sortant du microphone, puis le stocker en mémoire sous forme numérique.

13.1 Comment est vue un signal

Le signal acoustique d'une voix parlée contient différents types d'information : le message (ce qui est dit), des informations propres au locuteur (qui l'a dit) et à l'environnement (où, quand, comment cela a été dit, enregistré). Pour une transformation de voix, nous désirons modifier les attributs relatifs au locuteur Ces attributs spécifiques du locuteur peuvent être groupés en plusieurs niveaux :

- Phonématique (segmental) regroupe l'ensemble des facteurs définissant la qualité d'une voix : son timbre.
- Prosodique (suprasegmental) correspond aux composantes de l'expression et du style, c'est-à-dire l'intonation et l'accent. Au niveau

du signal, la prosodie correspond à la hauteur du son, à l'énergie et à la durée des phones et des silences.

13.2 La modélisation des paramètres acoustiques

Dans un système de reconnaissance vocale, les paramètres acoustiques sont utilisés pour estimer un modèle idéal qui doit satisfaire les contraintes suivantes **:**

- Il doit avoir une méthode d'estimation la moins complexe possible.
- Il doit permettre une décision rapide lors de la phase de test.
- Il doit être le plus robuste possible aux variations intra locuteur.
- Il doit permettre la meilleure séparation des locuteurs entres eux.
- Il doit avoir la représentation la plus complète possible des paramètres.

13.3 Para métrisation du signal vocal

La variation de la nature du signal acoustique rend le traitement des données brutes issues de ce dernier très difficile. En effet, ces données contiennent des informations complexes, souvent redondantes et mélangées.

La phase de para métrisation, qui traite le signal acoustique reçu, doit remplir plusieurs objectifs:

- Séparer le signal du bruit ;
- Extraire l'information utile à la reconnaissance ;
- Convertir les données brutes à un format directement exploitable par le système.

Afin de concevoir un bon système, il faut choisir des paramètres qui sont fréquents, (ne pas correspondre à des évènements ne survenant que très rarement dans le signal), facilement mesurables, robuste face aux imitateurs, ne pas être affecté par le bruit ambiant ou par les variations dues au canal de transmission. Pratiquement, il est très difficile de réunir tous ces éléments en même temps, la sélection des paramètres pose un problème très complexe, et influe fortement sur les résultats des systèmes. D'après plusieurs recherches effectuées sur cette étape, les types de paramètres efficaces et utilisables sont les paramètres de l'analyse spectrale.

- ***Les Paramètres de l'analyse spectrale [15]***

Les principaux paramètres de l'analyse spectrale utilisés en reconnaissance vocale sont les coefficients de prédiction linéaire et leurs différentes transformations (LPC, LPCC,..) ; ainsi que les coefficients issus de l'analyse en banc de filtres et leurs différentes transformations (coefficients banc de filtres, MFCC...).

Plusieurs travaux ont été publiés pour comparer les différentes techniques en para métrisation, l'enjeu de ces travaux était de cibler les meilleurs paramètres représentant de façon efficace les propriétés caractéristiques propres à chaque locuteur. Les meilleurs résultats ont été obtenus en utilisant la méthode MFCC.

13.4 L'analyse de Fourier : [16]

L'analyse de Fourier est un moyen de décomposer un signal en une somme de signaux élémentaires particuliers, qui ont la propriété d'être faciles à mettre en œuvre et à observer. L'intérêt de cette décomposition réside dans le fait que la réponse au signal d'un système obéissant au principe de superposition peut être déduite de la réponse aux signaux élémentaires. Ces signaux élémentaires sont périodiques et complexes, afin de permettre une étude en amplitude et en phase des systèmes ; ils s'expriment par la fonction $s_e(t)$telle que :

$$s_e(t) = e^{j\pi 2ft} = cos\ (2\pi ft) + j\ sin\ (2\pi ft)$$

Où f représente l'inverse de la période, c'est la fréquence du signal élémentaire.

13.5 Pourquoi l'échelle de Mel [17]

L'échelle des Mels est une échelle biologique. C'est une modélisation de l'oreille humaine. A noter que le cerveau effectue en quelque sorte une reconnaissance vocale complexe avec filtrage des sons… Prenons l'exemple suivant où vous êtes à table en compagnie de nombreuses personnes, l'ensemble de ses personnes parle en même temps et vous discutez avec votre voisin. Malgré le bruit, vous arrivez à discerner clairement ce que vous dit votre voisin, vous ignorez de façon naturelle le bruit de fond et vous amplifier le son qui vous paraît le plus important. Vous pouvez répéter cette expérience avec chacun des convives. Le cerveau ne se contente non pas seulement de

filtrer les sons et de les amplifier mais aussi de prédire. Prenons l'exemple suivant où une personne discute avec vous avec un volume sonore très bas, vous n'avez pas entendue une certaine partie de la phrase mais vous arrivez à la reconstituer et à la comprendre. A partir de l'étude du cerveau nous pouvons nous faire une idée de la complexité de la reconnaissance vocale et nous pouvons nous rapprocher d'un modèle de plus en plus puissant et parfait. On considère que l'oreille humaine perçoit linéairement le son jusqu'à 1000 Hz, mais après, elle perçoit moins d'une octave par doublement de fréquence. L'échelle de Mels modélise assez fidèlement la perception de l'oreille : linéairement jusqu'à 1000 Hz, puis logarithmiquement au-dessus.

La formule donnant la fréquence en Mels *m* à partir de celle en Hz *f* est :

$$m = \frac{1000 \cdot \ln\left(1 + \frac{f}{700}\right)}{\ln\left(1 + \frac{1000}{700}\right)} \approx 1127 \cdot \ln\left(1 + \frac{f}{700}\right) \approx 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$$

Figure 1.17 Formule de fréquence en Mels

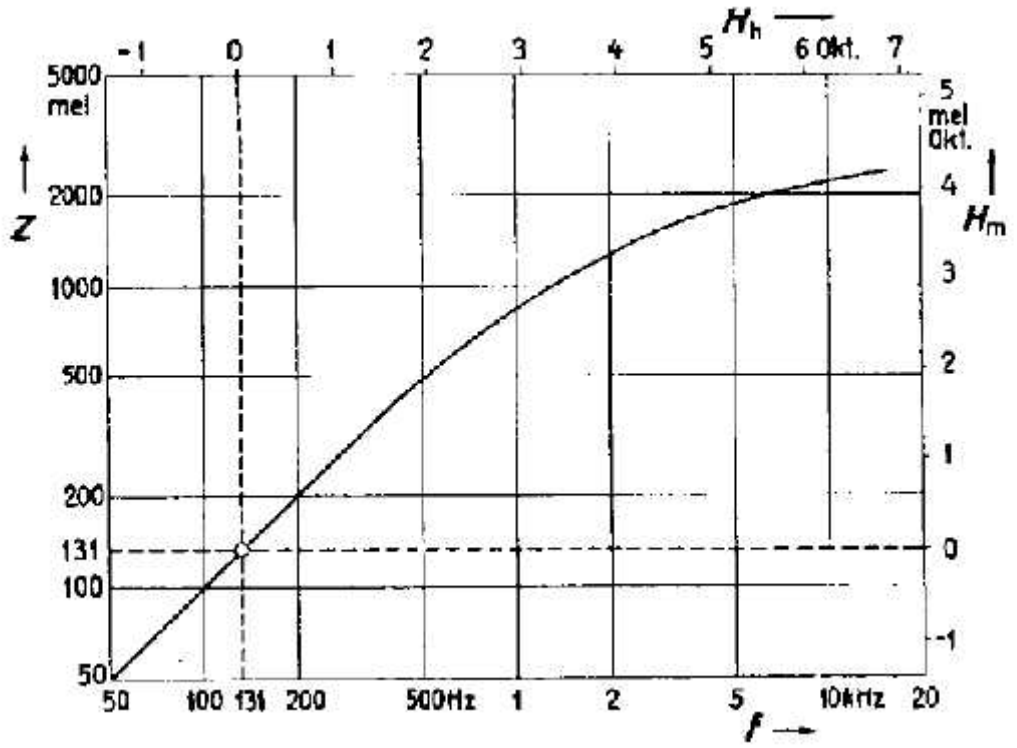


Figure 1.18 Exemple de conversion des HERTZ en MELS.

L'objectif de cette phase est d'extraire des coefficients représentatifs du signal de la parole. Ces coefficients sont calculés à intervalles réguliers. Le signal de la parole est transformé en une série de vecteurs de coefficients, ces coefficients doivent représenter au mieux ce qu'ils sont censés modéliser et doivent extraire le maximum d'informations utiles pour la reconnaissance. Parmi les coefficients les plus utilisés et qui représentent au mieux le signal de la parole, nous trouvons les coefficients ceptraux, appelés également ceptres. Les deux méthodes les plus connues pour l'extraction des ceptres sont : l'analyse spectrale et l'analyse paramétrique. Pour l'analyse spectrale (Mel-ScaleFrequencyCeptral Coefficients (MFCC)) comme pour l'analyse paramétrique (le codage prédictif linéaire (LPC)), le signal de parole est transformé en une série de vecteurs calculés pour chaque trame. Il existe d'autres types de coefficients qui sont surtout utilisés dans des milieux bruités, par exemple les coefficients PLP (PerceptualLinearPredective).

Il existe plusieurs techniques permettant l'amélioration de la qualité des coefficients, nous trouvons par exemple ; l'analyse discriminante linéaire (LDA), l'analyse discriminante non linéaire (NLDA), etc. Ces coefficients jouent un rôle capital dans les approches utilisées pour la reconnaissance vocale.

14 Outils existants

14.1 HTK

Hidden Markov Model Toolkit (HTK) est un ensemble d'outils portable permettant lacréation et la manipulation de modèles de Markov cachés. HTK est principalement utilisédans le domaine de la recherche de la reconnaissance vocale bien qu'il soit tout à faitutilisable dans de nombreuses autres applications telles que la synthèse vocale, lareconnaissance de l'écriture ou la reconnaissance de séquences d'ADN.

Il est composé d'un ensemble de modules et outils écrits en langage C. Ces différents outilsfacilitent l'analyse vocale, l'apprentissage des HMM, la réalisation de tests et l'analyse desrésultats. Il est à noter, que ce qui a contribué au succès de HTK, est qu'il est accompagnéd'une assez bonne documentation.

14.2 Sphinx 4

Sphinx 4 est un logiciel de reconnaissance vocale écrit entièrement en Java. Les butsde Sphinx sont d'avoir une reconnaissance vocale hautement flexible, [Philippe Galley et al.2006] d'égaler les autres produits commerciaux et de mettre en collaboration les centres derecherche de diverses universités, des laboratoires de Sun et de HP mais aussi du MIT.

Tout en étant hautement configurable, la reconnaissance de Sphinx 4 supportenotamment les mots isolés et les phrases (utilisation de grammaires). Son architecture estmodulable pour permettre de nouvelles recherches et pour tester de nouveaux algorithmes.La qualité de la reconnaissance dépend directement de la qualité des données vocales.Ces dernières étant les informations relatives à une voix propre. Ce sont par exemple lesdifférents phonèmes, les différents mots (lexique), les différentes façons de prononciation.Plus ces informations ne seront importantes et connues par le système, meilleure sera saréaction et ses choix à faire.

15 Conclusion

Dans ce chapitre, nous avons décrit les premiers modules du traitement de la parole en vue de sa reconnaissance. Nous avons abordé les outils les plus répandus de nos jours. Nous avons présenté ces données pour caractériser notre système d'expérimentation, décrit en détail plus loin dans ce manuscrit. Nous avons jusqu'ici volontairement éludé les modèles de langage. Ceux-ci étant la base de notre travail, le chapitre suivant leur sera entièrement consacré.

CHAPITRE II
Présentation de l'application

1. Introduction

Cette étude permet, à partir de besoins bien spécifiés de donner une solution concernant la structure du logiciel à savoir l'ensemble des programmes, et la structure des données utilisées. Dans ce chapitre nous présentons la modélisation de notre application, les outils utilisés et plusieurs tests sont présentés en vue de montrer la viabilité de notre application

2. Définition du thème

2.1 Un jeu éducatif

Le jeu a une potentialité éducative c'est donc être capable de montrer que le jeu est susceptible d'intervenir positivement dans l'un des quatre domaines au moins quedonne la définition :

- Le développement intellectuel du joueur
- La formation physique du joueur
- La formation morale du joueur
- L'adaptation sociale du joueur.

Un raccourci consiste à dire que si le jeu a une valeur éducative, c'est qu'il apprend quelque chose au joueur.

2.2 Les avantages du jeu éducatif

L'apport de la reconnaissance vocale permettra aux enfants :

- De démontrer une aptitude à utiliser les outils de reconnaissance et de saisie vocale (aujourd'hui un atout dans un cursus scolaire et demain dans la vie professionnelle).
- De prononcer et de lire clairement à l'aide des technologies de reconnaissance vocale.

- De pouvoir exprimer ses idées très rapidement par écrit sans en perdre le fil.
- De gagner un temps considérable dans la rédaction de tous leurs travaux scolaires et universitaires.

2.3 Définition de l'Application

Notre application (Mon Enfant Parle)est un programme d'apprentissage de jeux éducatifs conçus pour les enfants pour différentes tranches d'âge,dans lequel le joueur (l'enfant) apprend a prononcé correctement des mots de la langue française. Il propose des activités variées aux enfants, ces activités sont quelquefois ludiques, mais toujours pédagogiques. Les parents ont la possibilité de rajouter quotidiennement de nouveaux apprentissages adaptés à leur enfant et leur niveau scolaire puis de les tester.

Nous y trouverons des activités dans les domaines suivants :

- Découverte de l'ordinateur : Souris, Les mouvements de la souris, ...
  Mathématiques : Apprendre les chiffres, ...
- Lecture : Exercice d'entraînement à la lecture, lettre, chiffre, mot.
- Autres, ...

2.4 Le but de l'application

- « Mon enfant parle » améliore la lecture et l'écriture des enfants. Pour la dictée, les enfants doivent lire à voix haute. La lecture à haute voix étant peu pratiquée dans les établissements scolaires, cet exercice améliore leur lecture et peut également améliorer la reconnaissance des mots et leur prononciation, la fluidité de la lecture et la compréhension.
- Les enfants présentant des troubles d'apprentissage, notamment ceux ayant des difficultés liées au langage et des problèmes de mémoire, utilisent l'application Mon enfant parle. Ils n'ont plus besoin de se préoccuper du mécanisme de la composition (orthographe, structure des phrases, etc.), ce qui facilite le passage d'idées en mots.

- Grâce aux logiciels de reconnaissance vocale, les enfants apprennent une langue, améliorent leur expression orale et leur écriture tout en s'amusant.

3. Reconnaissance vocale de Windows

La reconnaissance vocale de Windows permet aux utilisateurs d'interagir avec leurs ordinateurs par la voix. Il a été conçu pour les personnes quiveulent limiter considérablement leur utilisation de la souris et le clavier tout en maintenant ou en augmentant leur productivité globale. Vous pouvez dicter des documents et des courriers dans les applications grand public, utiliser les commandes vocales pour lancer et basculer entre les applications, contrôler le système d'exploitation, et même remplir des formulaires sur le Web.

- **Cortana, l'assistante vocale au cœur de Windows 10 :**

Windows 10 est le premier système d'exploitation de Microsoft à succomber à la mode de l'assistante vocale en introduisant Cortana. Une assistante à la fois très proche et très différente de ce que l'on peut trouver chez Google ou Apple.

Les outils qu'on a choisi d'utiliser sont comme suit :

4. Choix du langage de programmation

4.1 Qu'est-ce que le C# :

Microsoft C# (prononcez C sharp) est un nouveau langage de programmation qui a été conçu pour permettre la création d'une large gamme d'applications d'entreprise s'exécutant sur le .NET Framework. Évolution du Microsoft C et du Microsoft C++, C# est simple, moderne, à sécurité de type et orienté objet. Le code C# est compilé en tant que code managé, c'est-à-dire qu'il bénéficie des services du Common

LanguageRuntime(CLR). Ces services incluent l'interopérabilité entre les langages, un garbage collection, une sécurité améliorée et une meilleure prise en charge du versioning.

C# est présenté en tant que Visual C# dans la suite Visual Studio .NET. La prise en charge de Visual C# comprend les modèles de projet, les concepteurs, les pages de propriétés, les Assistants Code, un modèle objet et d'autres fonctionnalités de l'environnement de développement. La bibliothèque de programmation de Visual C# n'est autre que le .NET Framework.

4.2 Qu'est-ce que SQLITE? [18]

SQLite est un système de base de données qui a la particularité de fonctionner sans serveur, on dit aussi "standalone" ou "base de données embarquée". On peut l'utiliser avec beaucoup de langages : PHP, Python, C# (.NET), Java, C/C++, Delphi, Ruby...

L'intérêt c'est que c'est très léger et rapide à mettre en place, on peut s'en servir aussi bien pour stocker des données dans une vraie base de données sur une application pour smartphone (iPhone ou Android), pour une application Windows, ou sur un serveur web.

Une base de données SQLite est bien plus performante et facile à utiliser que de stocker les données dans des fichiers XML ou binaires, d'ailleurs ces performances sont même comparables aux autres SGBD fonctionnant avec un serveur comme MySQL, Microsoft SQL Server ou PostgreSQL.

Pourquoi utiliser SQLite ?

Ce qu'il faut bien garder à l'esprit, c'est que SQLite n'est pas vraiment un concurrent des serveurs de base de données relationnelles, c'est plus une extension de leur

champ d'application à des applications offline (là ou avant elles n'étaient utilisé que sur des serveurs sur un modèle client/serveur).

SQLite peut aussi s'avérer très utile sur un site web pour créer des fonctionnalités qui ne sont pas directement liées au site (comme par exemple organiser un jeu concours avec un formulaire où les gens peuvent d'inscrire), évitant ainsi de polluer la base de données principale (fonctionnant sur MySQL par exemple) avec des données temporaires.

L'autre avantage est la simplicité : il n'y a aucune manipulation à faire, le fichier sqlite est créé automatiquement à la volée, toute la base est stockée dans un fichier unique qu'il est facile d'échanger en FTP...

En fait, SQLite peut être utilisé en remplacement de Microsoft Access, car il est bien plus performant, propre, portable et multiplateforme.

4.3 Qu'est-ce que UNITY ? [19]

Unity3D est un logiciel, ou plus précisément, un moteur de jeux vidéo en 2D et 3D. Il permet de créer des jeux multiplateformes. En effet, Unity3D nous permet de créer nos propres jeux vidéo assez facilement (en ayant des connaissances en développement) avec un choix de deux langages possibles : le JavaScript orienté jeux vidéo ainsi que le C#.

Le JavaScript, quand à lui, est beaucoup moins utilisé que le C#. Utiliser le C# permet de développer avec l'IDE Visual Studio qui comprend également l'auto complétion de toutes les classes d'Unity3D (à chaque création d'un jeu vidéo avec ce moteur de jeu, il crée lui-même une solution Visual Studio avec un .SLN).

L'intérêt de ce logiciel est que, tout comme Flash, celui-ci dispose d'une interface d'intégration d'objets et de scripts vraiment très intuitive. L'éditeur d'Unity intègre des composants préconfigurés évitant le développement de code assez fastidieux. Ces composants sont :

- Un moteur physique basé sur PhysX de la société Nvidia.
- Un système de collision.
- Un système de ragdoll.
- Un moteur d'ombre et de lumières PBR.
- Une interface en drag and drop.

Et plein d'autres choses… Le gros avantage d'Unity est qu'il permet **des exports vers de multiples plateformes** comme vous pouvez le voir ci-dessous

5. Modélisation

Modéliser un système avant sa réalisation permet de mieux comprendre le fonctionnement du système. C'est également un bon moyen de maîtriser sa complexité et d'assurer sa cohérence. Il existe plusieurs méthodes de modélisation : Merise, UML…

On a choisi UML, plus précisément le Diagramme de classes car ilest considéré comme le plus important de la modélisation orientée objet,

Le diagramme de classe est une représentation statique des éléments qui composent un système, il permet de modéliser les classes intervenant dans le système ainsi que les interventions entre elles.

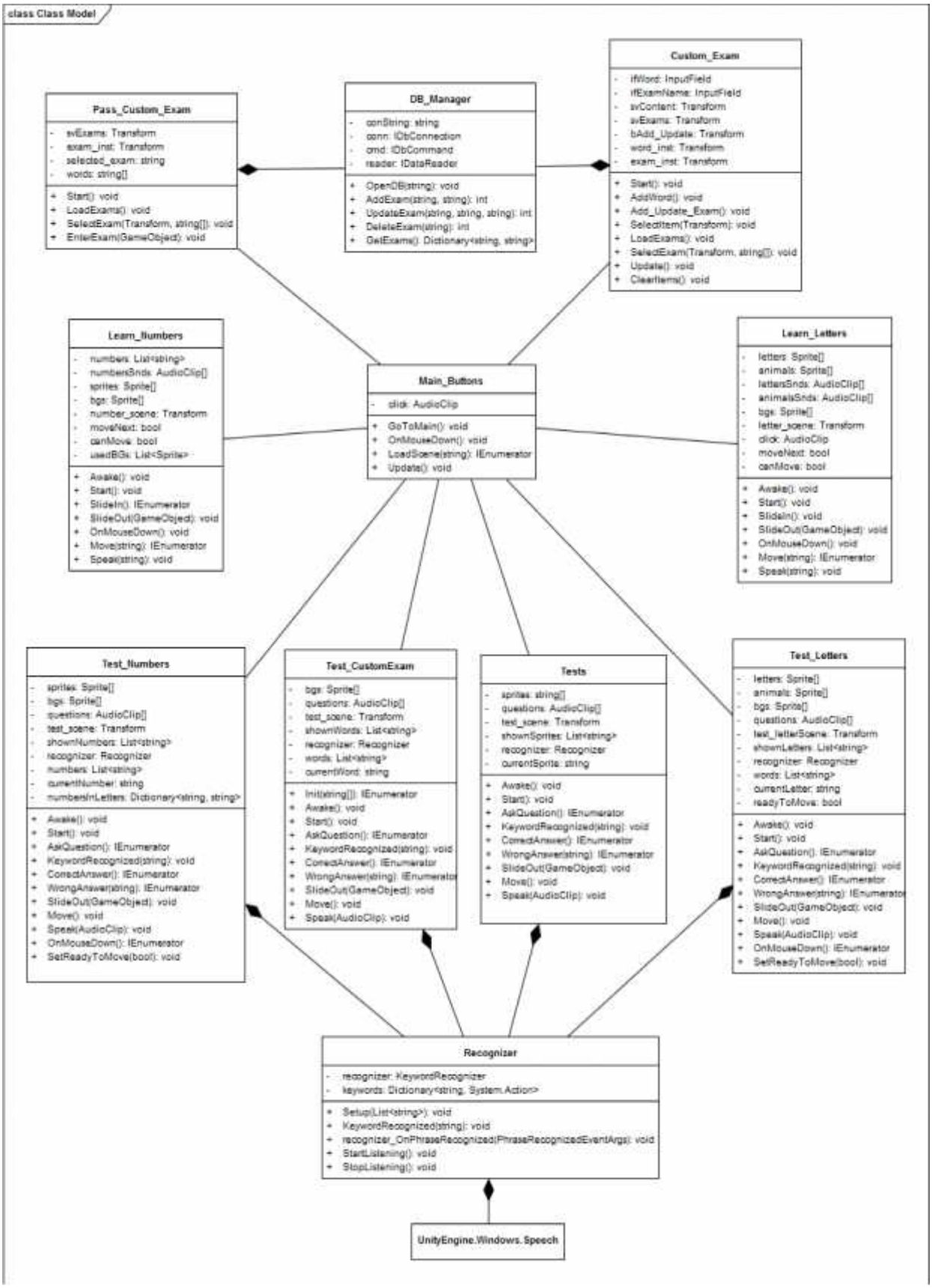
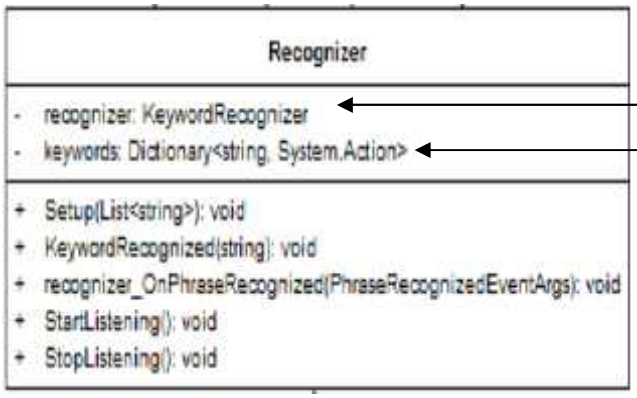


Figure 2.1 : Diagramme de classe

- Ce diagramme contient 11 classes.

1. La classe Recognizer responsable sur la reconnaissance vocale relie la bibliothèque **Unity.Engine.Windows.Speech** avec les autres classes
2. La classe Test_Letters représente la partie test des lettres
3. La classe Tests représente la partie des tests(test animaux, test école, test cuisine…)
4. La classe Tests_CustomExam représente la partie test des examens créé par les parents
5. La classe Test_Numbers représente la partie test des nombres
6. La classe Learn_Letters représente la partie apprentissage des lettres
7. La classe Learn_Numbers représente la partie apprentissage des nombres
8. La classe Custom_Exam permet de crée un examen
9. La classe DB_Manager pour connecter la base de données
10. La classe pass_Custom_Exam permet de passé un examen existe déjà
11. La classe Main_Buttonsreprésente la fenêtre principale de l'application, elle permet de relier toute les scènes du jeu

- **Unity.Engine.Windows.Speech :** qui représente la bibliothèquede la reconnaissance vocale sur laquelle on s'est basé pour créer notre application.
- **Exemple :**



La classe Recognizer contient :

L'attribut recognizer

L'attribut Keywords

Et les méthodes :

Setup

KeywordRecognizer

Recognzer_OnphraseRecognized

StarListening

StopListening

6. Quelque fenêtre de l'application

Après le lancement de l'application (Mon enfant parle) la fenêtre suivante s'affiche :

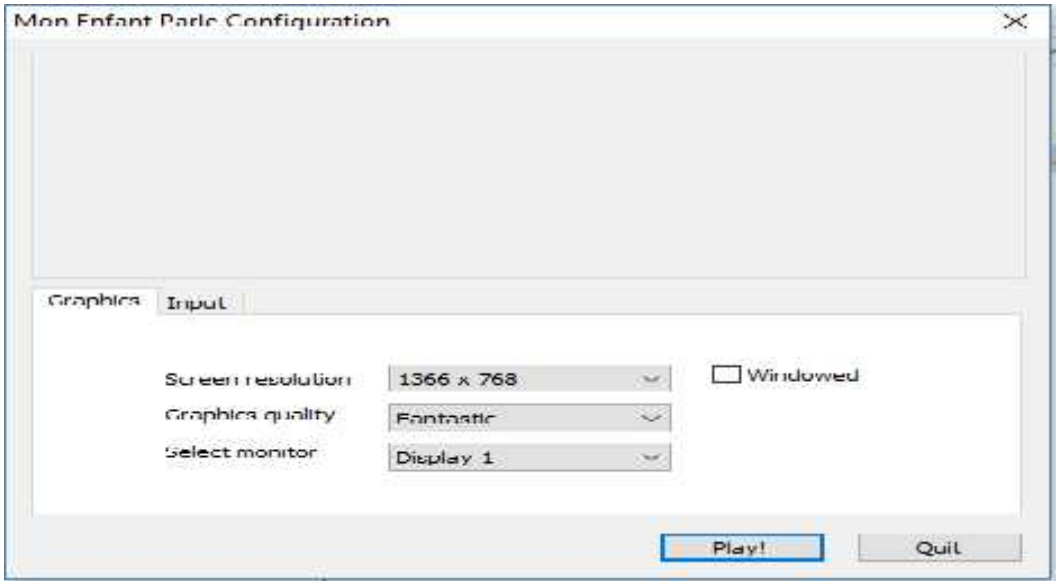


Figure 2.2 : Configuration de la résolution écran

Elle permet de règles la résolution de l'écran (par défaut elle prend la résolution de l'écran de PC ou l'application est installée)

6.1 La fenêtre principale du logiciel

La fenêtre principale contient trois scène Apprendre, Tests et Parents Coin.

Figure 2.3: Fenêtre Principale du logiciel

6.2 La partie de l'apprentissage

Permet de rentrer dans la lecture en douceur, et vont amener les enfants à développer leur conscience phonologique, à jouer avec les sons

Figure 2.4: Fenêtre d'apprentissage

- Partie Lettre pour apprendre les lettres avec une photo d'un animal qui commence par la lettre courante

Figure 2.5: Fenêtre d'apprentissage de l'alphabet

- Cette partie permet apprendre à lire les chiffres par observation

Figure2.6: Fenêtre d'apprentissage des chiffres

6.3 La phase des tests

Cette phase contient plusieurs tests. Elle vient juste après la phase d'apprentissage, pour tester si l'enfant apprend bien ou non.

Figure 2.7: Fenêtredes tests

- Dans ce test, l'enfant apprendra à compter

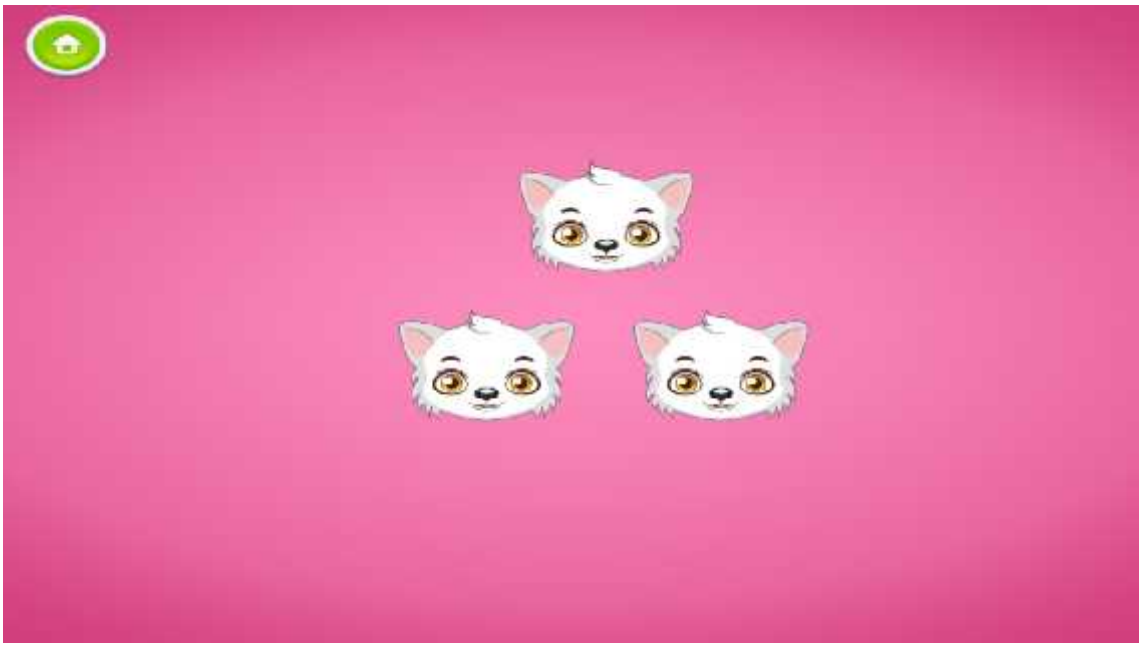


Figure 2.8: Fenêtretest nombres

- L'enfant apprendra à prononcer correctement le nom de l'animal affiché

Figure 2.9: Fenêtretest des animaux

6.4 La partie parents

Cette partie est la plus importante dans ce jeu, car elle rend le jeu dynamique

Figure 2.10: FenêtreParents Coin

6.5 Fenêtre nouveau examen

Permet aux parents de créer un nouveau test qui s'adapte au niveau de leurs enfants.

Figure 2.11: FenêtreNouveau examen

Ecrire le nom d'examen

Ecrire les mots au les phrases avec les quelles les parents veulent tester leurs enfants

Cliquer ici pour ajouter le nouveau test à la liste des examens

Figure 2.12: Comment créer unNouveau examen

6.6 Fenêtre tester votre enfant

Cette fenêtre contient les examens créés par les parents et c'est à l'enfant de choisir un examen à passer.

Figure 2.13: Fenêtretester votre enfant

6.7 Fenêtre résultat

Voici les résultats pour la réponse juste Voici les résultats pour la réponse fausse

Figure 2.14: Fenêtrerésultat

7. Conclusion

Dans ce chapitre on a cité les différentes étapes pour la réalisation et le développement de notre application, ce chapitre est la consécration de plusieurs mois de travail, incluant une étape recherche bibliographique, puis le choix des méthodes à implémenter et enfin l'étape de l'implémentation et les résultats. Lors de ce travail, nous avons dû faire face à plusieurs problèmes, car la tâche n'est pas aisée et les difficultés nombreuses. Nous espérons avoir obtenu un produit efficace, bien qu'encore imparfait, mais ce dont nous sommes sures, c'est que ce PFE nous a permis de mettre en pratique toutes nos connaissances informatiques et bien plus encore.

CONCLUSION GENERALE

La recherche en reconnaissance vocale a été influencée par les progrès technologiques. Ca a débuté avec les systèmes analogiques, et le développement rapide de l'informatique et de la microélectronique ont permis l'ouverture de nouveaux horizons pour ce domaine tant au niveau des techniques qu'au niveau des secteurs d'application.

La reconnaissance vocale recouvre tous les aspects liés à l'interprétation, par la machine, du langage humain. Les applications de cette technologie sont nombreuses : Navigation sur un serveur vocal au téléphone, Apprentissage d'une langue étrangère, commandes vocales dans les voitures, les téléphones ou bien encore dans les salles d'opérations chirurgicales, dictée vocale, identification vocale dans les zones sécurisées ou bien dans le cadre d'une enquête judiciaire, etc.

Dans notre mémoire, nous avons choisi d'appliquer la reconnaissance vocale dans le domaine éducatif sous forme d'un jeu vue l'importance de ce dernier pour nous enfants car, le jeu est indispensable à l'enfant surtout chez les plus jeunes. Le fait d'associer apprentissage et divertissement est un concept intéressant faisant intervenir le jeu. Ce dernier ne nécessitant aucun effort spécifique, il autorise l'enfant à s'instruire tout en s'amusant.

Dans la réalisation de l'application, nous avons utilisé la plateforme des bibliothèques de la reconnaissance vocale de Windows, C# et Unity (Unity.Engine.Windows.Speech) avec certains problèmes dus à la bibliothèque elle-même vue que cette dernière est spécifique pour la reconnaissance des mots. Ces problèmes sont résumés dans la difficulté de reconnaitre quelques lettres. Cependant, il reste quelque point à améliorer :

- Gestion du joueur (voir plusieurs joueurs)
- Meilleurs gestion du temps
- Meilleur gestion des formes
- Ajout d’animations
- Gestion des scores

Ce projet était intéressant car il nous a permis de découvrir de nouvelle façon d’interagir avec un programme(Unity). Par contre nous trouvons dommage que le temps mis à disposition ne soit pas plus grand. Nous pensons qu’avec plus de temps le projet pourrait être mieux développé et donc plus intéressant.

Bibliographie

Bibliographie

[1] Jean-François Compagnat ; François Carreau. Le principe de la reconnaissance vocale (Rapport de recherche)

[2] B. H. Juang, L. R. Rabiner. Automatic Speech Recognition—A Brief History of the Technology. In Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005.

[3] http://spcts.e-monsite.com/medias/files/15-t2-reconnaissance-vocale.pdf

[4] http://www.audio-maniac.com/?page_id=2

[5] Traitement numérique du signal. Théorie et pratique de M.Bellanger professeur à l'Université de Paris-Sud. Directeur du laboratoire des signaux et systèmes

[6] https://www.google.dz/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0ahUKEwiqwbzarbTUAhVLvRQKHduCCFYQFgglMAE&url=http%3A%2F%2Fwww.tigen.org%2Fperso%2Fgeogeo%2FTPE_Vocale%2FDossier%2520(Reconnaissance%2520vocale).doc&usg=AFQjCNHOTh2ZRp3dlxtFzY9Xsy3_1yJArg
20/05/2017 à 23 :15

[7] http://www.grundig-gbs.com/fr/loeschen/solutions/dicter/article//vorteile-der/
29/04/2017 à 18 :00

[8] http://www.clubic.com/article-161030-4-clubic-test-solutions-reconnaissance-vocale.html 15/05/2017 à 20 :13
Par Paul-E Graff

[9] A. Mokeddem, "Reconnaissance multilocuteur de mots isolés pour les systèmes miniaturisés", thèse de doctorat, Université de Neuchâtel institut de microtechnique, 1985.

[10] Y. Deville, Traitement du signal - Signaux temporels et spatiotemporels, Ellipses, 2011.

[11] http://www.technologuepro.com/transmission/chapitre4.htm
03/05/2017 à 00 :17

[12] Articles de Cristiano Mendès, Micael Henriques Lopes, Paulo-André Fontes mardi 27 mai 2008

[13] Thèse Doctorat de l'Université de Lorraine (mention informatique)
Par Luiza Orosanu (Reconnaissance de la parole pour l'aide à la communication pour les sourds et malentendants)
11 Décembre 2015

[14] R. L. Paul Gaillard, Analyse et traitement du signal : Siganux déterminsites et aléatoires, filtrage, estimation avec exercices et problèmes corrigés, Traitement du signal, Broché, 2006.

[15] G. Mahmoud, La Parametrisation Mfcc En Vue D'Une Reconnaissance Robuste de Parole, 2015

[16] Traitement numérique du signal. Théorie et pratique de M.Bellanger professeur à l'Université de Paris-Sud. Directeur du laboratoire des signaux et systèmes

[17] K. Dash, A Novel Bpnn Approach for Speaker Identification Using Mfcc, 2012.

[18] http://www.finalclap.com/faq/180-sqlite-definition

[19] Par Anthony DI STEFANO Publié le 13/07/2015 à 18:27:06 Noter cet article