République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ ABOU BEKR BELKAID DE TLEMCEN

FACULTÉ DE TECHNOLOGIE

DÉPARTEMENT D'informatique

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

Option : Réseaux et systèmes distribués (R.S.D.)

Mise au point d'une application de reconnaissance vocale

Présenté le 02/07/ 2018 devant le jury composé de :

Président:	M. Hadjila Fethallah	UABB Tlemcen
Examinateur:	Mme . Iles Nawel	UABB Tlemcen
Encadreur :	Mme. Lahfa Fedoua	UABB Tlemcen

Présenté par: Boukada Yassine

Bemmoussat Mohammed Hamza

Année académique: 2017-2018

Résumé

Passionné par les différentes techniques de reconnaissance d'une personne par l'exploitation des différentes sources d'information. Ce PFE a pour but de s'attaquer à l'utilisation de la synthèse vocale dans l'authentification et la reconnaissance d'une personne, suite au développement notable de la biométrie, à sa grande utilisation dans de nombreux domaines, et à ses nombreux aspects : empreintes digitales, empreinte rétinienne, empreinte vocale, empreinte de l'ECG, etc. Les caractéristiques de la source vocale propre à chaque individu sont généralement jugées moins discriminants mais difficiles à extraire, nous avons tous des timbres de voix différents. La voix de chaque personne dépend de caractéristiques à la fois anatomique et comportementales. Ces caractéristiques servent à créer une signature vocale qui permet d'authentifier la voix de chacun. Suite à l'évolution de la technologie les chercheurs et les développeurs sont poussés à reconsidérer certains préjugés et mieux utiliser ces informations complémentaires pour améliorer les performances du système de reconnaissance du locuteur. Notre système vocal fournit principalement les indices acoustiques et aussi la personnalité individuelle pour caractériser le locuteur et s'assurer de son identité. Le développement et la croissance planétaire des communications, implique le besoin de s'assurer de l'identité des individus. Motiver par l'importance des enjeux, les fraudeurs ne cessent d'essayer de mettre en échec les systèmes de sécurité existants.

Dans ce projet donc, on s'est posé comme objectif principal la mise au point d'un système de reconnaissance vocale.

abstract

Passionate about the different techniques of recognition of a person by the exploitation of the different sources of information. The purpose of this EFP is to address the use of voice synthesis in the authentication and recognition of a person, following the notable development of biometrics, its extensive use in many areas, and its many aspects: Fingerprints, retinal fingerprints, voice prints, ECG prints, etc. The characteristics of the voice source specific to each individual are generally considered less discriminatory but difficult to extract, we all have different voice stamps. The voice of each person depends on both anatomical and behavioral characteristics. These features are used to create a voice signature that allows you to authenticate everyone's voice. Following the evolution of technology researchers and developers are encouraged to reconsider some prejudices and use this additional information better to improve the performance of the speaker's recognition system. Our voice system mainly provides the acoustic cues and also the individual personality to characterize the speaker and make sure of his identity. The development and global growth of communications implies the need to ensure the identity of individuals. Motivating by the importance of the stakes, the fraudsters keep trying to defeat the existing security systems.

In this project, the main objective was to develop a voice recognition system.

ملخص

مطورو النظم المعلوماتية متحمسون لمختلف تقنيات التعرف على شخص عن طريق استغلال مصادر المعلومات المختلفة. ، بعد التطور الملحوظ لنظام البيومتري، واستخدامه علي نطاق واسع في كثير من المجالات ، وجوانبه العديدة : بصمات الأصابع ، وبصمات الشبكية ، والبصمات الصوتية ، ...

الصوت البشري في مجمل الحالات يميز الفرد ولكن من الصعب استخلاص مكوناته ، ولدينا جميعا طوابع صوتيه مختلفة تستخدم هذه الميزات لإنشاء توقيع صوتي يسمح مصادقه صوت بتعريف الافراد . صوت كل شخص يعتمد علي كل من الجميع. وبعد تطور البحوث لمطورين في مجال التكنولوجيا ، هناك تشجيع علي أعادة النظر في بعض التحيزات واستخدام هذه المعلومات الإضافية علي نحو أفضل لتحسين أداء نظام التعرف علي الاصوات.

نظامنا الصوتي يوفر أساسا البحة الصوتية وأيضا شخصيه الفرد لتوصيف المتكلم والتأكد من هويته. ويقتضي تطوير نظام تعرف على الأصوات دراسة معمقة لمختلف خصائصه و معرفة المخاطر و الهفوات الممكن مصادفتها اثناء تطوير نظامنا المقترح كمشروع تخرج .

وكان الهدف الرئيسي في هذا المشروع هو وضع نظام للتعرف على الأصوات بأكبر دقة ممكنة واقل استغلال للمصادر المتاحة.

Sommaire

Introduction Générale

Si l'être humain a le privilège de comprendre un message vocal d'une autre personne quelconque quel que soit l'environnement, la syntaxe, le vocabulaire utilisé, la machine sera-elle un jour capable de faire autant ? Une solution robuste et efficace sera-elle trouvée et proposée pour satisfaire ces contraintes ? Le moins qu'on puisse dire c'est que malgré son importance actuellement, seules des solutions partielles sont proposées dans le langage machine, qui peuvent faire des taches déjà prédéfinies et déjà préenregistrées, mais est incapable de faire ce dont l'homme est capable. Pour le moment on utilise l'ordinateur plus pour faire de la reconnaissance sous toutes ses formes : faciale, vocale, rétine, empreinte, ECG…afin de sécuriser et contrôler l'accès à des ressources, des bureaux, des ordinateurs ou des informations sensibles. Ou pour lutter contre la fraude, le vol d'informations personnelles, etc.

Dans ce PFE on va cibler la reconnaissance vocale qui à elle seule est un domaine très vaste en recherche, beaucoup de travaux y sont consacrés, et aucune solution satisfaisante n'a encore été proposée, d'où notre intérêt pour ce sujet, et donc on va éssaier de mettre au point une application de reconnaissance vocale ainsi une petite application de reconnaissance de la parole. Pour se faire notre PFE est composé en 3 chapitres, le premier présentera la biométrie vocale soit la reconnaissance automatique de la voix, son déploiement, son fonctionnement et sa nécessité.

Le deuxième chapitre introduira le traitement de la voix et sa conversion et proposera quelques techniques utilisées pour la conversion ainsi que le mécanisme de conversion.

Dans le troisième on détaillera notre application, les méthodes ainsi que les outils utilisés dans la réalisation du projet.

Chapitre I : Généralités

1. Introduction

La reconnaissance vocale appelée biométrie vocale est en passe de révolutionner l'identification des services à distance en remplaçant nos mots de passe, trop souvent vulnérables, oubliés, craqués, devinés ou trop long à taper. Elle consiste à employer des techniques d'appariement afin de comparer une onde acoustique à un ensemble d'échantillons, composés généralement de mots mais aussi plus récemment de phonèmes (unité sonore minimale). Cependant personne n'est encore arrivé à une solution satisfaisante à 100%.
La reconnaissance vocale est encore en chantier et attire beaucoup de développeurs amenés par la complexité de la technologie d'analyse de la voix (aussi appelée analyse du locuteur). Cette technologie s'applique avec succès là où les autres technologies sont difficiles à employer. Elle est utilisée dans des secteurs comme les centres d'appel, les opérations bancaires, l'accès à des comptes, sur PC domestiques, pour l'accès à un réseau ou encore pour des applications judiciaires. Le traitement vocal vise donc aussi un gain de productivité puisque c'est la machine qui s'adapte à l'homme pour communiquer, et non l'inverse et c'est pour ça que la reconnaissance Vocale est quasiment imparfaite dans son domaine. [1]

2. Qu'est-ce que la biométrie vocale ?

La biométrie vocale : est une technologie se basant sur plusieurs algorithmes à programmer pouvant distinguer les variations naturelles (silence, essoufflement…), et identifier plusieurs caractéristiques vocales pour authentifier une personne à travers sa voix, ainsi la biométrie vocale reconnaît votre voix, même altérée par un rhume ou par un changement de timbre, mais ne fonctionne pas si vous utilisez un enregistrement. Pour cela plusieurs applications ont été mis en point pour remplacer les méthodes d'authentification classiques beaucoup moins sécurisées comme le mot de passe ou la dates de naissance.

2.1. Reconnaissance vocale d ‘une personne et son fonctionnement

La reconnaissance vocale automatique d’une personne ou d’un locuteur se basant sur la voix fait encore parler d‘elle, suite aux nombreuses difficultés rencontrées par plusieurs équipes de programmeurs et développeurs formés pour trouver une solution optimale. Cette technologie s’est limitée à la vérification ou la détection de l‘identité d’une personne à partir de sa voix. En reconnaissance du locuteur, on fait la différence entre la vérification et l’identification du locuteur, bien que le problème reste le même. Est-ce que cette voix détectée correspond bien à cette personne sensée la produire parmi des centaines de voix d‘individus déjà préenregistrés ou non. Cette différence se détermine dans la reconnaissance d‘un locuteur, dépendante du texte, avec texte dicté, ou reconnaissance indépendante du texte. Dans le premier cas, la reconnaissance est limitée par la prononciation d’une phrase, déjà fixée dans la conception du système ; ou dictée en forme de mot de passe dans le deuxième cas, et non précisée dans le dernier.

La vérification vocale a pour but de filtrer en acceptant ou refusant une identité proclamée par un locuteur. En se basant sur le calcul d’un modèle stochastique sur la base dune expression vocale prononcé par ce dernier et comparer ce modèle à d’autre modèle de d’autres locuteurs déjà enregistrés.

Au fil des année les techniques de reconnaissance vocale automatique de locuteur, s’est considérablement élargi du au progrès des algorithmes utilisés, l’évolution remarquable des technologies utilisées, et la puissance de traitement disponible.

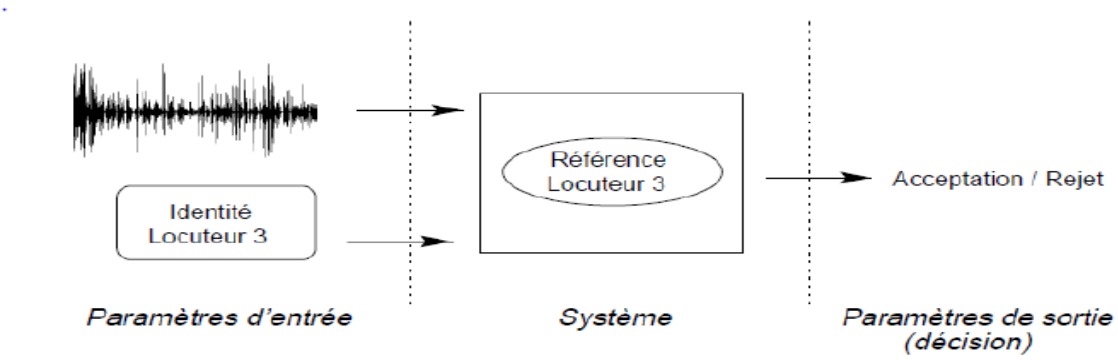


Figure 1.1 : Vérification vocale

a. Son fonctionnement

Il est important de ne pas confondre analyse du locuteur et dictée vocale. Dans un cas on cherche à déterminer l'identité d'un individu grâce à sa voix ; dans l'autre on cherche à déterminer ce que dit la personne sans se soucier de son identité.

La plupart des systèmes d'identification de la voix utilisent l'affichage d'un texte, des mots spécifiques doivent être lus puis parlés afin de vérifier que la personne à authentifier est bien présente et qu'il ne s'agit pas d'un enregistrement. Ils se concentrent sur les seules caractéristiques de voix qui sont uniques à la configuration de la parole d'un individu. Ces configurations de la parole sont constituées par une combinaison des facteurs comportementaux et physiologiques.

Les imitateurs essayent habituellement de reproduire les caractéristiques vocales qui sont les plus évidentes au système auditif humain et ne recréent pas les caractéristiques moins accessibles qu'un système automatisé d'identification de voix analyse. Il n'est donc pas possible d'imiter la voix d'une personne inscrite dans la base de données.

La variabilité d'une personne à une autre démontre les différences du signal de parole en fonction du locuteur. Cette variabilité, utile pour différencier les locuteurs, est également mélangée à d'autres types de variabilité : variabilité due au contenu linguistique, variabilité intra-locuteur (qui fait que la voix dépend aussi de l'état physique et émotionnel d'un individu), variabilité due aux conditions d'enregistrement du signal de parole (bruit ambiant, microphone utilisé, lignes de transmission) .qui ces variabilité peuvent rendre l'identification du locuteur plus difficile.

Malgré toutes ces difficultés apparentes et le problème qui consiste à extraire l'information contenue dans un signal de parole, typiquement par échantillonnage du signal électrique obtenu à la sortie d'un microphone, afin qu'il puisse être comparé à des modèles sous forme numérique, la voix reste un moyen biométrique intéressant à exploiter car pratique et disponible via le réseau téléphonique, contrairement à ses concurrents.

2.2. Le système est-il robuste ?

La reconnaissance vocale est la moins connue mais la plus sûre des méthodes d'authentification à distance À ce jour, c'est l'outil le plus sécurisé. L'empreinte de l'œil est unique elle aussi, mais les technologies sont moins matures. De même pour l'empreinte digitale qui n'atteint pas un niveau de reconnaissance suffisant. Les technologies de la reconnaissance vocale, affichent aujourd'hui un taux d'erreur très bas et un haut niveau de sécurité. Ce n'est bien sûr pas le cas des autres techniques d'authentification en ligne, qui exigent de se protéger efficacement . Mais, le système est-il capable de fonctionner proprement dans des conditions difficiles ?

En effet, de nombreuses variables pouvant affecter significativement les performances des systèmes de reconnaissance (Bruits d'environnement, Déformation de la voix par l'environnement, Qualité du matériel utilisé, Bande passante fréquentielle limitée). Certains systèmes peuvent contourner un ou plusieurs de ces perturbations, mais en règle générale, les systèmes de reconnaissance vocale sont encore sensibles et influençable par ces perturbations.

2.3. Les Domaines d'utilisation de la Reconnaissance vocale

L'identification de la voix est considérée par les utilisateurs comme une des formes les plus prisées de la technologie biométrique, car elle n'est pas invasive et n'exige aucun contact physique avec le lecteur du système.

La technologie d'analyse du locuteur s'applique avec succès là où les autres technologies sont difficiles à employer. Elle est utilisée dans des secteurs comme les opérations bancaires, les centres d'appel, sur PC domestiques, l'accès à des comptes, pour l'accès à un réseau ou encore pour des applications judiciaires

2.4. Problématique de la reconnaissance vocale

L'idée consiste à faire prononcer un mot par plusieurs personnes et puis enregistrer leurs voix, sous forme de vecteurs acoustiques (représentation numérique du signal sonore). Cette suite de vecteurs acoustiques caractérisera l'évolution de l'enveloppe spectrale du signal enregistré, on pourra dire qui correspond à un enregistrement d'un spectrogramme. L'étape de reconnaissance consiste alors à analyser le signal inconnu sous la forme d'une suite de vecteurs acoustiques similaires, et à comparer la suite inconnue à chacune des suites préalablement enregistrés. La personne « identifiée » sera alors celle dont la suite de vecteurs acoustiques s'apparente le plus à celle de la suite inconnue. Il s'agit en quelque sorte de voir dans quelle mesure les spectrogrammes se superposent, d'où alors l'apparition de certains problèmes. Un même mot peut en effet être prononcé d'une infinité de façons différentes, en changeant le rythme de l'élocution. Il en résulte des spectrogrammes plus ou moins distordus

dans le temps. La superposition du spectrogramme inconnu aux spectrogrammes de base doit dès lors se faire en acceptant une certaine «élasticité» sur les spectrogrammes candidats. Cette notion d'élasticité est formalisée mathématiquement par un algorithme nommé : l'algorithme DTW (Dynamic Time Warping).

3. Les défis et les motivations

Comme l'énonce la problématique, la biométrie est une science qui n'est pas encore au point. En effet, la reconnaissance vocale fait partie des domaines qu'il faut encore développer. Cette reconnaissance a des failles car elle dépend de plusieurs paramètres comme la qualité du micro, le fait qu'il ne doit pas y avoir de sons parasites lors de l'analyse ou des bruits de fond. Celle-ci est souvent utilisée pour les transactions par téléphone pour minimiser les fraudes mais elle n'est pas très efficace. L'un de ses avantages réside dans le fait que l'utilisateur n'est pas en contact direct avec l'appareil : son cerveau ne la perçoit pas comme intrusive. Selon certains sondages elle reste quand même une des méthodes les plus utilisées dans les systèmes de sécurité et d'authentification. Donc le principal défi dans la technologie de reconnaissance vocale du locuteur a donc été d'améliorer la robustesse et la fiabilité des systèmes dans des conditions incompatibles. La variation intra-locuteur du style du parlé, les variations de l'environnement acoustique, etc.

4. La voix dans un système de AAL (Authentification Automatique du Locuteur)

Il s'agit de reconnaître automatiquement l'identité d'une personne prononçant une ou plusieurs phrases, comme un auditeur humain identifie son interlocuteur au cours d'une conversation. Nous distinguerons les applications sur site (serrures vocales pour contrôle d'accès, cabines bancaires en libre service),les applications liées aux télécommunications (ces applications concernent l'identification du locuteur à travers le réseau téléphonique pour accéder à un service de transactions bancaires à distance ou pour interroger des bases de données en accès privé),et les applications judiciaires (recherche de suspects, orientations d'enquêtes, preuves lors d'un jugement).La difficulté de la tâche d'authentification n'est pas la même d'une application à une autre. Dans le cas des applications « sur site », l'environnement de prononciation de la phrase ou du mot de passe est plus facilement contrôlé que dans le cas des applications via le réseau téléphonique (distorsions dues au canal, différences entre les combinés téléphoniques, bande passante limitée). Les applications judiciaires présentent quant à elles des difficultés d'un autre ordre (locuteurs non-coopératifs, enregistrements de mauvaise qualité).

4.1. Capture et traitement de la voix

En fonction de l'application envisagée, la qualité demandée par la capture de la parole peut rapidement devenir très importante. En effet, cette qualité dépend de la variabilité de la voix du locuteur dans le temps comme dans le cas de maladie (un rhume), des états émotionnels (l'angoisse ou la joie) et de l'âge. De plus, les conditions d'acquisition de la voix telle que le bruit et la réverbération, ainsi que la fidélité des équipements tel que le microphone jouent très fortement sur la qualité de la capture, et donc sur la qualité des résultats. Pour pouvoir être traité numériquement, le signal sonore est numérisé sur 8 ou 16 bits à une fréquence d'échantillonnage qui varie entre 8 kHz et 48 kHz. Ainsi un système standard peut se décrire : sur l'analyse du signal acoustique afin d'en extraire des paramètres. Ces paramètres résultent, entre autres, d'une analyse spectrale du signal (coefficients de prédiction linéaires ou bancs de filtres). Les paramètres servent ensuite à l'élaboration éventuelle d'un modèle et sont introduites dans un classifieur qui permettra de déterminer l'identité du locuteur. De nombreuses techniques sont utilisées pour réaliser ce classifieur. On peut citer entre autre les réseaux de neurones, les chaines de Markov, les mélanges gaussiens, la quantification vectorielle, etc.

4.1.1. Dépendance et Indépendance au texte

La distinction est faite entre les systèmes dépendants et indépendants du texte. En mode dépendant du texte, le texte prononcé par le locuteur (pour être reconnu du système) est le même que celui qu'il a prononcé lors de l'apprentissage de sa voix. En mode indépendant du texte, le locuteur peut prononcer n'importe quelle phrase pour être reconnu. Néanmoins, il existe plusieurs niveaux de dépendance au texte suivant les applications (citée selon le degré croissant de dépendance au texte), systèmes à texte libre (ou free-text) ou le locuteur prononce ce qu'il veut ;aussi les systèmes à texte suggéré (ou text-prompted) içi un texte, différent à chaque session et pour chaque personne, est imposé au locuteur et affiché à l'écran par la machine ;plus à ça les systèmes dépendants de traits phonétiques (ou speech event dependent) , certains traits phonétiques spécifiques sont imposés dans le texte que le locuteur doit prononcer ;et les systèmes dépendants du vocabulaire (ou vocabulary dependent) içi le locuteur prononce une séquence de mots issus d'un vocabulaire limité (ex. : séquence de digits) ; Ajoutant à ça les systèmes personnalisés dépendants du texte (ou user-specific text dependent) de sorte que chaque locuteur a son propre mot de passe. Ces dernier donnent généralement de meilleures performances d'authentification que les systèmes indépendants du texte car la variabilité due au contenu linguistique de la phrase prononcée est alors neutralisée.

4.2. Pourquoi l'authentification vocale ?

Suite au attaque et aux tentatives de fraude, la sécurité informatique via des mots de passe ou les dates de naissance est devenu vulnérable face à ces derniers, la plupart des entreprises ont exigé que les mots de passe soient modifiés régulièrement avec un niveau de sécurité élevée (comportent au moins 8 caractères, avec variété entre majuscules, minuscules, chiffres et caractères spéciaux. L'objectif est d'éviter les logiciels de décodage qui peuvent balayer tous

les mots du dictionnaire en peu de temps. Une protection pas très fiable pour l'accès à des applications sensibles.

4.3. Evaluation des performances en AAL

L'évaluation des performances d'un système d'AAL est donné par le taux d'erreurs dans l'identification et le taux du rejet dans la vérification du locuteur, cependant cette évaluation n'est cependant pas un problème commun et on ne peut comparer deux systèmes à partir de ces seuls taux d'erreur qui dépendent de multiples facteurs. Ainsi, les éléments suivants doivent également être pris en compte :

- qualité de la parole : enregistrements en studio ou via le canal téléphonique ; qui dépend de l'environnement et du type de réseau téléphonique,
- quantité de parole : est la durée de parole prise pour l'apprentissage des références de chaque locuteur et des sessions de test ,
- variabilité intra-locuteur : tout dépend de l'état physique et émotionnel et le comportement du locuteur,
- population de la base de locuteurs : en identification du locuteur, la taille de la population, la qualité de la population ,la bonne répartition géographique des locuteurs parlant une même langue est également un facteur à intégrer et a une influence directe sur les performances,
- intention des locuteurs : la distinction est faite entre les locuteurs qui veulent être reconnus par le système et les locuteurs qui modifient leur voix pour ne pas être reconnus (cas de certaines applications judiciaires par exemple,

4.3.1. Domaine des applications de l'authentification vocale [2]

Les applications d'authentification couvrent en grande partie tous les domaines de la sécurité ou il est nécessaire de connaitre l'identité des personnes. Aujourd'hui, les principales applications sont la production de titres d'identité, le contrôle d'accès à des sites sensibles, l'accès aux réseaux, stations de travail et PC, le contrôle des frontières, systèmes d'information, le paiement électronique, la signature électronique et même le chiffrement de données. La liste des applications pouvant utiliser l'authentification pour contrôler un accès (physique ou logique), peut être très longue. La taille de cette liste n'est limitée que par l'imagination de chacun. Les banques et d'autres grandes entreprises se tournent aujourd'hui vers l'authentification vocale, distinguée comme unique. Notre voix fait partie de nous et elle est toujours avec nous contrairement à nos clés de voitures, et aux mots de passes ou codes PIN qu'on peut très souvent oublier. C'est à la fois cette sécurité et cette simplicité d'usage offerte par l'authentification vocale qui pousse les banques, les opérateurs de télécommunications et autres grandes organisations à choisir ce mode d'authentification, ci-dessous quelque exemple de ses applications :

a) Sécurisation des applications mobiles

Les grandes entreprises voient désormais leurs clients utiliser massivement les canaux mobiles pour prendre contact et effectuer les opérations courantes. C'est même devenu une attente forte des clients et des consommateurs. Mais la multiplication des applications et services en ligne fait qu'il devient difficile de gérer tous ces mots de passes, de forme et de tailles différentes.

L'authentification vocale devient dès lors le mode d'authentification mobile idéal. Il suffit simplement de donner une simple phrase clé à prononcer par un client pour vérifier son identité.

En plus d'éliminer la frustration née des mots de passe difficiles à mémoriser ou à saisir, le 'login vocal' réinvente véritablement l'authentification mobile. Le mobile devenant de plus en plus le point de contact principal entre un consommateur et un fournisseur de services, améliorer l'expérience utilisateur et la sécurité deviennent une priorité.

b) Sécurisation des transactions à risque par carte de crédit

La reconnaissance de locuteur constitue aussi une solution sûre et pratique pour vérifier les transactions par carte de crédit. Quand une opération à risque est détectée, une demande de vérification de la transaction peut être envoyée au titulaire de la carte de crédit, via un appel sortant automatique, sur son téléphone portable. Le détenteur est alors invité à prononcer une phrase clé : "J'autorise cette transaction par ma signature vocale".

A l'inverse, si la transaction est suspecte, il peut tout aussi facilement rejeter celle-ci, ce qui permet alors à l'institution financière d'investiguer sur les transactions marquées comme suspectes.

c) Paiement en ligne

La reconnaissance de la voix peut être utilisée pour sécuriser des paiements en ligne, typiquement des paiements à risque tels que le premier paiement en ligne sur un site d'e-commerce, par exemple le transfert de l'argent ou des opérations importantes. Lorsque ces opérations sont effectuées, un appel sortant automatique est émis vers le téléphone portable du titulaire du compte effectuant l'opération. Si cette opération est valide, l'utilisateur est invité à confirmer le paiement de la même façon qu'il peut confirmer l'achat par carte de crédit.

5. Reconnaissance vocale [3]

5.1. Quelles sont les applications directes de la reconnaissance vocale ?

La reconnaissance vocale permet de **dicter des mots** et de **contrôler-commander** un outil. Dans le domaine industriel, la reconnaissance vocale peut être utilisée dans le secteur de l'automobile (ex : entrer une destination dans le GPS), de l'aviation (ex : commandes diverses), de la domotique (ex : programmer la température de la maison), Dans le domaine médical, elle permet de numériser des comptes-rendus médicaux. Enfin, elle constitue également une aide pour la communication, que ce soit pour les personnes handicapées ou les apprenants d'une langue.

5.2. Où en est-on en matière de reconnaissance vocale?

Le principal domaine d'application concerne RAP (la reconnaissance automatique de la parole). Les premiers systèmes datent de la fin des années 1970. C'est aussi à cette époque que les recherches se sont multipliées. Une des premières applications concerne la dictée, toujours très utilisée dans certaines professions. Mais c'est l'application la plus récente qui est la plus célèbre : il s'agit du système Siri qui permet aux iPhone de répondre à des questions formulées à haute voix. Preuve que les techniques de reconnaissance automatique de la parole ont considérablement progressé en 40 ans, même si c'est loin de la perfection. Parallèlement, les recherches sur l'identification de la voix ont aussi bien avancé mais le taux d'erreur reste de l'ordre de quelques pourcents dans le meilleur des cas, inacceptable notamment en matière judiciaire.

5.3. Comment modélise-t-on la parole ?

Principalement par modélisation statistique à partir d'une base de données de parole, une technique postulée au début des années 1980. En pratique, on enregistre maintenant plusieurs milliers d'heures de parole de centaines voire de milliers de locuteurs dans une langue, à la radio ou au téléphone. Cette énorme base de données est annotée, c'est-à-dire transcrite dans la langue étudiée, par exemple en français. On utilise ensuite un système de reconnaissance de la parole pour la découper en sons afin de réaliser un nouvel apprentissage, de meilleure qualité. Chaque son est représenté sous la forme d'un automate caractérisé par la probabilité d'un état en fonction des états précédents et la probabilité d'émission d'un vecteur acoustique, sorte d'image acoustique de 20 à 30 millisecondes de signal sonore. Grâce aux capacités de stockage et à la puissance des ordinateurs qui ont démultiplié les possibilités de traitement numérique, ces approches statistiques sont de plus en plus correctes.

5.4. De telles modélisations sont-elles utilisables pour identifier une voix ?

Absolument pas. Comme on le voit la qualité de la base de données conditionne très fortement les résultats. Dès que les conditions de prise de son s'éloignent de celles utilisées pour enregistrer la base de données, les résultats se dégradent très fortement. Par exemple le système Siri a été paramétré sur des voix enregistrées par téléphone, proches des conditions d'utilisation de l'application. Un enregistrement d'aujourd'hui comparé à de vieux enregistrements serait donc inexploitable. Qui plus est, ces techniques sont inutilisables avec de courts enregistrements. D'ailleurs, de nombreuses campagnes d'évaluation ont été menées depuis les années 1990 pour utiliser la voix comme information biométrique, au même titre que l'empreinte digitale ou l'iris : elles se sont toutes soldées par un échec. Alors même que le locuteur était coopératif.

5.5. D'autres techniques permettront-elles un jour de définir une empreinte vocale, unique ?

Peut-être mais sans doute plutôt avec une autre technique - la modélisation physique de la parole - et probablement pas avant 10 ou 20 ans. Cette technique, basée sur la modélisation géométrique du conduit vocal en trois dimensions (mesuré par exemple par Imagerie par résonnance magnétique (IRM)) couplé à un modèle biomécanique de la langue, est de plus en plus étudiée. Elle permettrait peut-être d'identifier les particularités d'articulation et de prononciation de chacun, en tirant parti des puissances de calcul désormais disponibles. Mais beaucoup de difficultés restent à résoudre notamment quant à la pertinence des données du conduit vocal à retenir et à la complexité des algorithmes de biomécanique.

6. Etude d'un signal sonore

Moins évident à saisir que les images, le son est une matière complexe : sa vitesse dépend de nombreux facteurs, de même que sa transmission et sa réverbération. A l'origine de tout son, il y a un mouvement (par exemple une corde qui vibre, une membrane de haut-parleur...).

Le son est défini par une vibration mécanique d'un fluide, qui se propage sous forme d'ondes longitudinales grâce à la déformation élastique de ce fluide. cette vibration grâce est ressentis au sens de l'ouïe par les êtres humains, comme beaucoup d'animaux. la psycho acoustique étudie la manière dont les organes du corps humain ressentent et l'être humain perçoit et interprète les sons; on dira aussi que l'acoustique est la science qui étudie les sons ;.

6.1 Caractéristiques d'un signal sonore.

Un son est défini par 3 paramètres : son intensité, sa hauteur tonale et son timbre.

Son intensité ou volume dépend de la pression acoustique créée par la source sonore (nombre de particules déplacées) ; plus la pression est importante et plus le volume est élevé (fort).

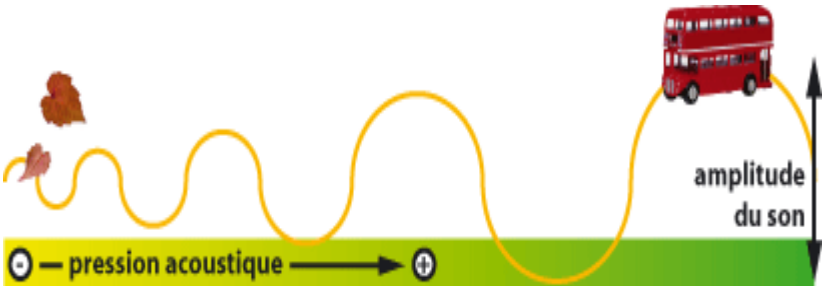


Figure 1.2 : Représentation de l'intensité du son

Sa hauteur tonale ou fréquence est définie par les vibrations de l'objet créant le son. Plus l'objet vibre rapidement, plus le son sera aigu. Le nombre de vibrations par seconde s'exprime en hertz. Le spectre audible de l'homme (de 16 Hz à 20.000 Hz) est divisé en octaves. Une octave représente l'intervalle séparant 2 notes dont la fréquence de l'une est le double de la fréquence de l'autre. La plupart des sources sonores produisent des sons complexes qui sont composés d'une fréquence fondamentale et d'harmoniques.
Les harmoniques sont des multiples entiers de la fréquence fondamentale f.
On distingue les harmoniques :
- paires : 2f, 4 f, 6 f, 8 f…
- impaires : 3 f, 5 f , 7 f, 9 f… que l'oreille n'apprécie guère (harmoniques anti-musicales)

Son timbre (ou couleur) est donné par le nombre et l'intensité des harmoniques qui le compose et permet de reconnaître la personne qui parle ou l'instrument qui est joué.

La fréquence fondamentale est la même, mais le nombre et l'intensité de leurs harmoniques respectives sont différents et l'oreille distingue les deux instruments.

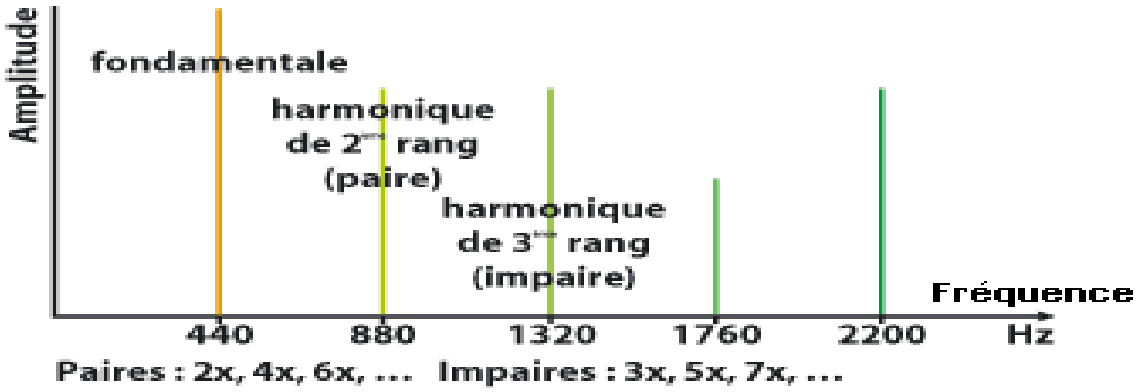


Figure 1.3 : Représentation de l'amplitude du son par rapport au fréquence.

Le rythme :

Le rythme est la durée des silences et des phones. Il est difficile de les en extraire car un mot prononcé d'une façon naturelle, sans aucun traitement, donne un mélange de phones chevauchés entre eux et un silence d'intensité non nulle.

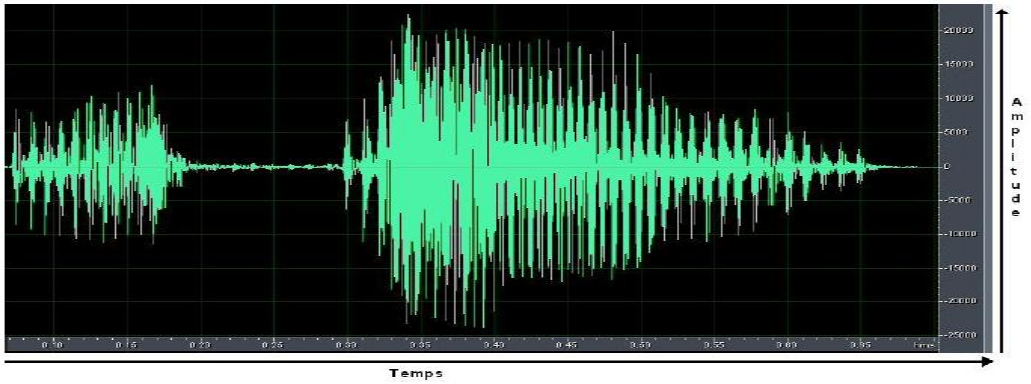


Figure 1.4 : Représentation d'un signal sonore

6.2 Les composantes du son :

Un son est souvent composé d'une durée qui représente l'étalement du son dans le temps, cette durée est strictement liée au rythme, une hauteur qui représente une sensation auditive plus au moine aiguë, une densité qui est la quantité d'éléments contenus dans un son. Le son contient un contraste qui est créer par la juxtaposition d'intensités, de hauteurs, de timbres, un son se compose aussi d'un mouvement mélodique qui est la direction auditive que prend la mélodie : elle monte, elle descend, elle reste à la même hauteur, et d'un tempo qui peut être rapide, vif, lent ou médium. Et représente la vitesse avec laquelle s'enchaînent les éléments sonores.

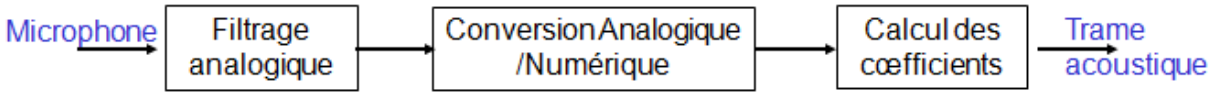


Figure 1.5 : Numérisation d'un son analogique

6.3. Traitement du son

6.3.1 La classification du son [5]

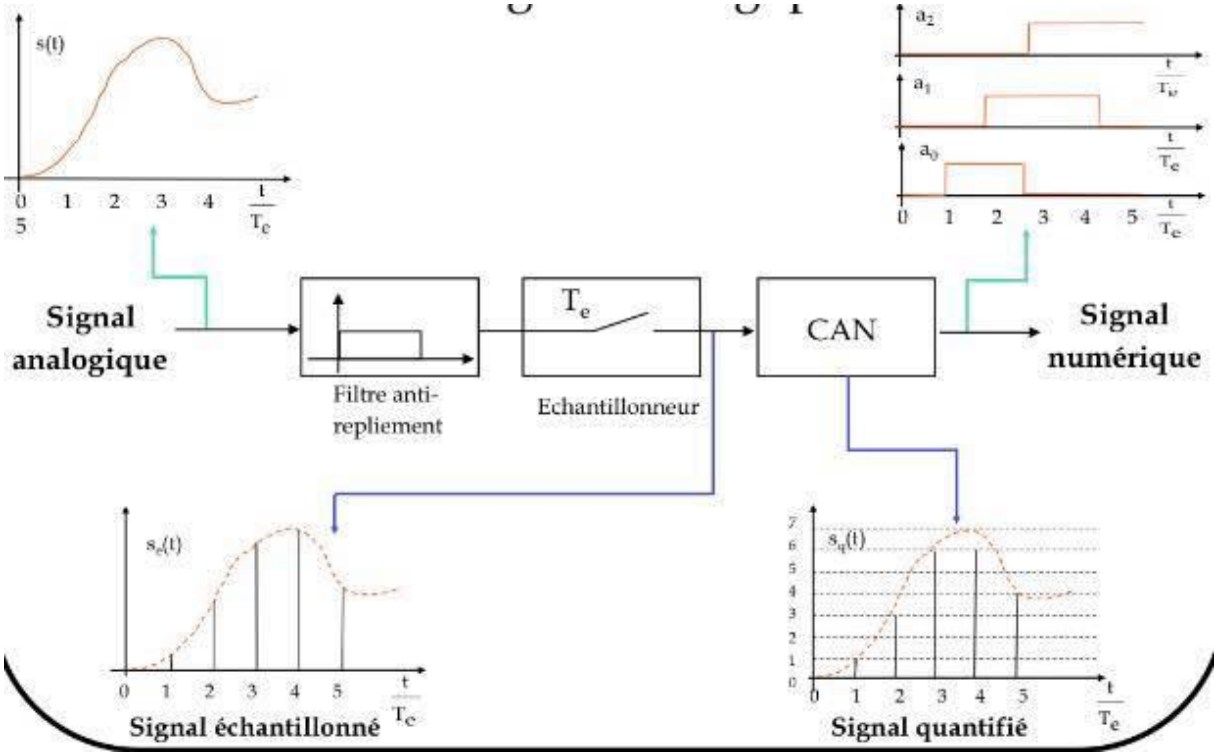


Figure 1.6 : Exemple d'un signal traité.

Il existe deux type de son: le son analogique et le son numérique.

Le premier est représenté sous la forme de signaux électriques d'intensité variable. Ces signaux sont issus d'un micro qui transforme le son acoustique d'une voix ou la vibration des cordes d'une guitare en impulsions électriques. Ces signaux sont enregistrables tels quels sur une bande magnétique (K7 audio par exemple) et peuvent être ensuite amplifiés, puis retransformés en son acoustique par des haut-parleurs. Le son analogique n'est pas manipulable tel quel par un ordinateur, qui ne connaît que les 0 et les 1.

Tandis que le son numérique est représenté par une suite binaire de 0 et 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. Une machine(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 la machine. C'est le rôle du convertisseur analogique/numérique.

Ainsi, la numérisation permet la transformation du signal sonore en fichier enregistré sur le disque dur de la machine, c'est le procédé qui permet de construire 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 qui permet la représentation de cet objet en informatique ou en électronique numérique. On utilise parfois le terme franglais digitalisation (digit signifiant chiffre en anglais).

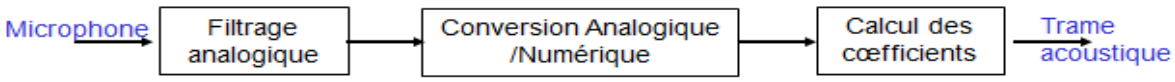


Figure 1.7 : Schéma général d'un traitement acoustique.

6.3.2. Numérisation du son [4]

C'est un procédé qui permet de construire 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).

Après l'enregistrement du son via 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. Il faut donc échantillonner le signal analogique pour convertir la tension en une suite de nombres qui seront traités par la machine. 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. Cette numérisation de son se réalise en deux étapes, l'échantillonnage et la quantification. Elle va permettre de transformer un signal continu en une suite de valeurs discrètes (distinctes) qui seront traduites dans le langage des machines, le langage binaire 0 et 1. Lorsqu'on capte un son à partir d'un microphone, ce dernier transforme l'énergie mécanique, en une variation de tension électrique continue. Ce signal électrique dit « analogique » pourra ensuite être amplifié, et envoyé vers un hautparleur dont la fonction est inverse: transformer à nouveau le signal électrique en une énergie mécanique.

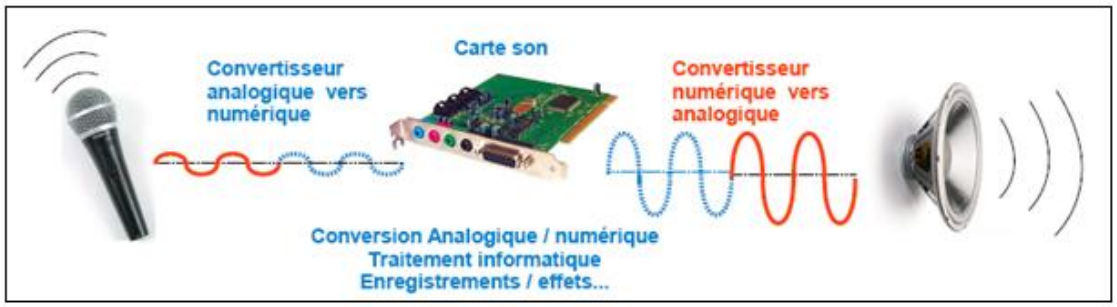


Figure 1.8 : Exemple d'une chaine numérique

6.3.3. L'échantillonnage

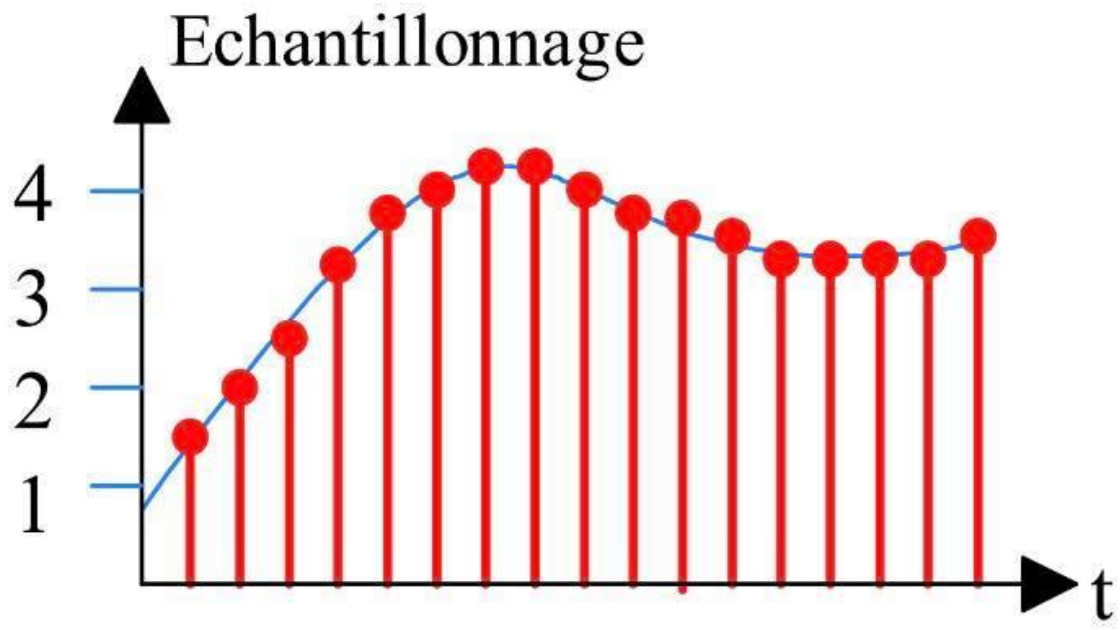


Figure 1.9 : Échantillonnage d'un signal audio.

Lorsqu'un son est numérisé, le signal analogique (continu) qui entre dans l'ordinateur est mesuré, un certain nombre de fois par seconde (d'où la discontinuité). Le son est donc découpé en "tranches", ou échantillons (en anglais « samples »). Le nombre d'échantillons disponibles dans une seconde d'audio s'appelle la fréquence d'échantillonnage exprimée en hertz. Pour traduire le plus fidèlement possible le signal analogique de notre micro, il faudra prendre le plus grand nombre de mesures possible par seconde. Autrement dit, plus la fréquence d'échantillonnage sera élevée, plus la traduction numérique du signal sera proche de l'original analogique. Attention tout de même à la taille des fichiers

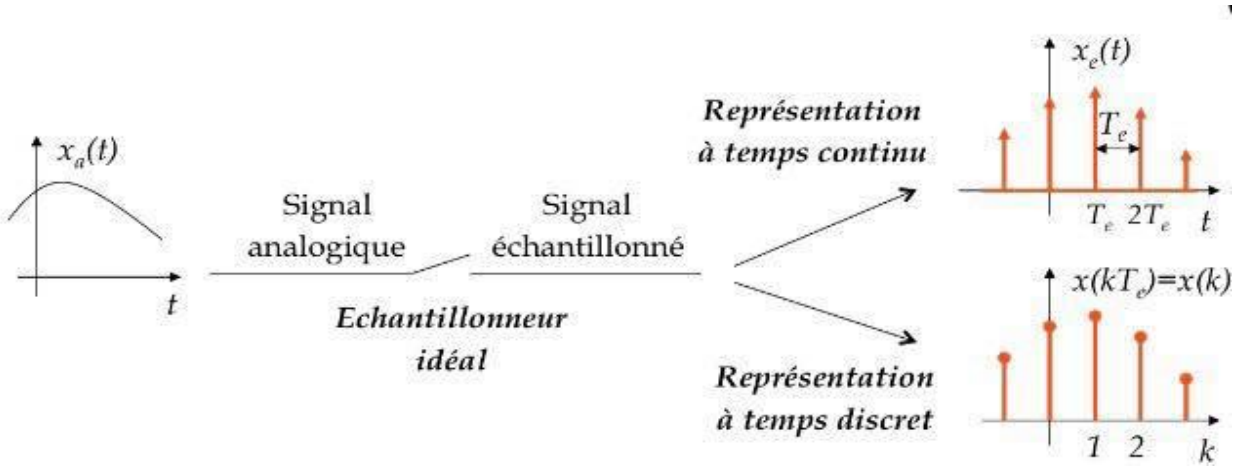


Figure 1.10 : Représentation d'un signal échantillonné

6.3.4 Résolution et quantification (bit)

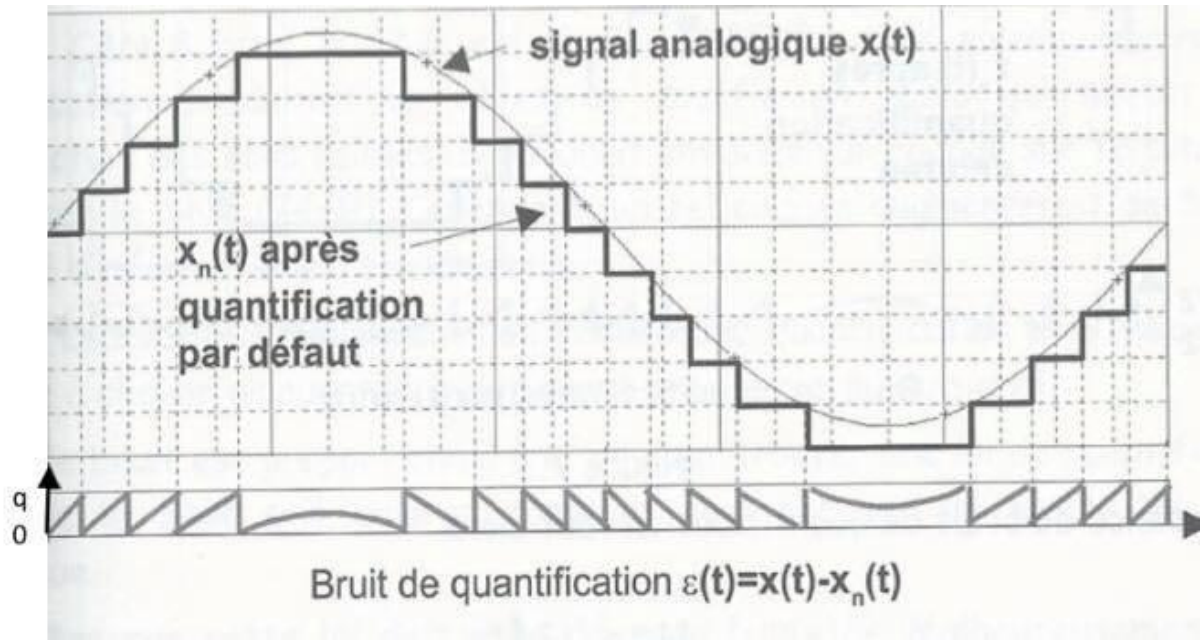


Figure 1.11 : Diagramme d'un signal quantifié

Une autre caractéristique importante est la résolution numérique du son, soit le nombre de « niveaux » ou de « paliers » qu'il est possible d'enregistrer pour reproduire l'amplitude du signal. Avec une résolution de 16bit, on dispose de 216, soit 65535 valeurs possibles pour traduire l'amplitude du son. Ainsi, plus la résolution est élevée, meilleure sera la dynamique (l'écart entre le son le plus faible et le plus fort qu'il est possible de reproduire).

Quelques exemples de résolutions fréquemment utilisées: Son qualité téléphone: 8000 Hz 8bit - Son qualité radio FM: 22050 Hz 16bit - Son qualité CD: 44100 Hz 16bit Son qualité DVD: 48000 Hz 24bit - Son audio professionnel: 96000 et 192000 Hz 24 et 32bit

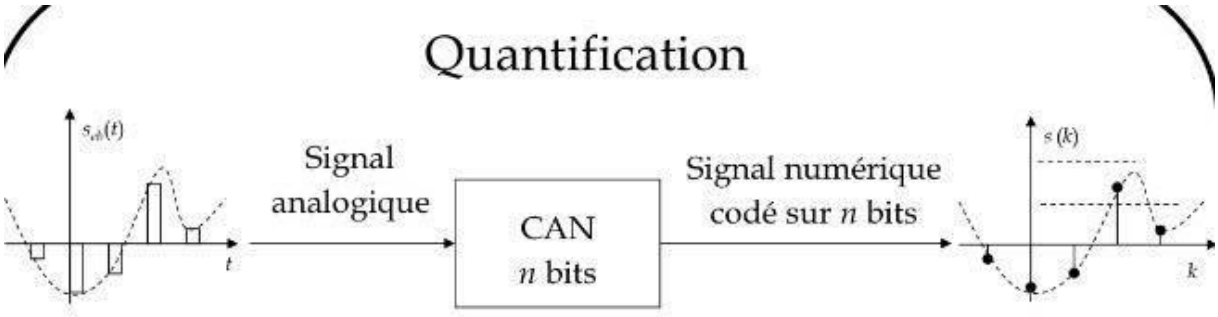


Figure 1.12 : Représentation d'un signal quantifié avec CAN.

6.4. Les fichiers audio numérique

6.4.1 LE FORMAT DE FICHIER WAVE

Un fichier audio non compressé est enregistré par défaut au format WAV(Waveform), qui est un dérivé de la spécification RIFF (Resource Interchange File Format) de Microsoft dédiée au stockage de données multimédias. Il s'agit d'un type de fichier mis au point par Microsoft. Un son d'une minute peut occuper entre 644Ko (kilo-octets) et 27Mo (mégaoctets). La taille de ce fichier dépend de la fréquence d'échantillonnage, du type de son (mono ou stéréo) et du nombre de bits utilisés pour l'échantillonnage (8 ou 16 bits). Le problème avec ce format est qu'il est évolutif et peut connaître de nombreuses formes (compressions audio, etc.

Les applications et logiciels d'édition sonore imposent que les sons soient dans ce format pour pouvoir les éditer. D'autres logiciels comme Audacity ont le privilège d'importer des fichiers mp3 qu'ils reconvertissent d'abord en Wave.

6.4.2. LE FORMAT DE FICHIER MP3

Très vite baptisé MP3,le MPEG 1 Audio Layer, a été créé en 1993 par l'institut Fraunhofer. A l'écoute, la différence entre le son d'un CD et le son de ce même CD compressé en MP3 est - selon la compression choisie - pratiquement imperceptible. En effet le MP3 filtre toutes les données non-audibles du fichier : tout ce qui se trouve en dessous de 20Hz ou au-dessus de 20000Hz est effacé mais aussi les sons qui sont couverts par d'autres. Le résultat est un fichier bien plus léger.

7. Conclusion

Dans ce chapitre on a parlé de la reconnaissance vocale, de l'authentification et l'identification d'une voix, on a aussi expliqué le traitement subi par un signal sonore nécessaire pour faire cette identification.

Les difficultés rencontrées dans cette partie commencent par l'acquisition du son afin de pouvoir l'utiliser pour les étapes suivantes.

Chapitre II : Etat de l'art

1. Introduction

Devenue comme instrument familier et tellement utilisée que l'on oublie sa complexité, la voix est un moyen pour nous de transmettre un message et nous permet de communiquer avec le monde extérieur, un lien entre deux personnes. La voix peut être qualifiée voire jugée sur certains critères nommés qualités vocales, ces qualités vocales se définissent comme la brillance, la raucité, la nasalité...etc. dans l'étude de la voix et du son on distinguera une discipline appelée : L'ethnomusicologie, c'est une discipline qui étudie les rapports entre musique et société. Elle emprunte aujourd'hui ses outils d'analyse et conceptuels à deux disciplines principales qui sont l'anthropologie et la musicologie. Elle s'ouvre aussi à d'autres domaines de recherche tels que les sciences cognitives et l'acoustique. Les répertoires étudiés émanent souvent de musiques de tradition orale souvent appelées ≪ les musiques du monde ≫.

Dans ce chapitre on donnera une description générale de la voix, son fonctionnement. La notion de qualité vocale sera décrite dans un premier temps, puis nous donnerons des explications sur les différentes qualités vocales utilisées, au niveau perceptif et acoustique. Ensuite, on citera la conversion du son pour donner une petite idée sur la comparaison entre deux sons différents.

2. Le timbre de la voix [7]

Pour explorer de manière perceptive et acoustique le timbre de la voix, nous devons expliciter le fonctionnement général de la voix. Celle-ci peut être caractérisée par son timbre mais aussi plus globalement par le terme de qualités vocales dont plusieurs d'entre elles seront décrites par deux points de vue, l'aspect perceptif et l'aspect acoustique.

2.1 La voix

2.1.1 Un instrument complexe

La voix est un instrument complexe permettant de produire des sons comme aussi de communiquer des informations et des émotions. La parole est vectrice de sens et d'interaction avec autrui. En tant qu'instrument de musique, la voix est assimilée à un instrument de la famille des vents. Elle est constituée d'une source et d'un corps sonore. L'air provenant des poumons met en vibration les cordes vocales situées dans le larynx, ce qui correspondrait `a la source. Le son est alors amplifié et filtré au niveau du conduit vocal qui joue le rôle de corps sonore et est constitué du pharynx, de la cavité buccale et des articulateurs (langues, lèvres, dents, ...). La voix est ainsi un modèle dit source/filtre. L'intensité, la hauteur ainsi que la signature vocale ou timbre d'un son sont caractérisés au niveau de la source. Le conduit vocal donne au son sa particularité, sa couleur et ses mouvements servent à l'articulation des sons émis.

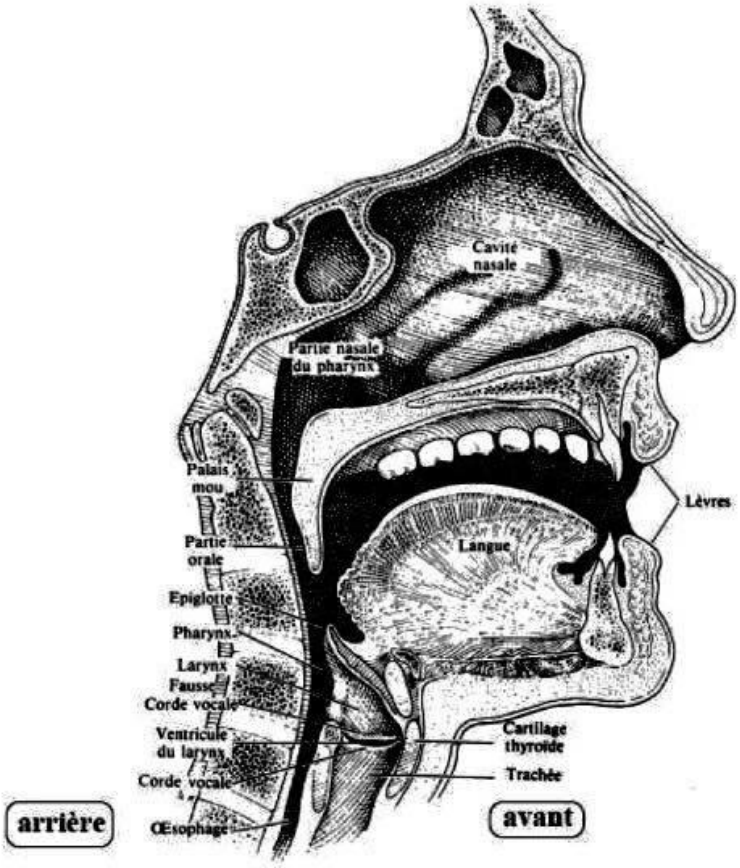


Figure 2.1 : Description de l'appareil phonatoire.

2.1.2 La registration

Un registre pourrait se définir de la façon suivante : ce serait une étendue de notes, par conséquent définie par des hauteurs. Le timbre serait homogène à l'écoute et le son serait produit de la même manière. Les registres correspondent ainsi à différents modes vibratoires des cordes vocales. De manière courante, on parle de voix de tête et de voix de poitrine. Dans le domaine de la technique vocale, sont référencés quatre types d´émissions du son ou quatre modes vibratoires dits mécanismes laryngés.

2.2 La qualité vocale

Dans le domaine du traitement de la parole et de la voix chantée, l'étude des qualités vocales est le sujet d'un certain nombre de recherches. Il s'agit, notamment, d'évaluer le timbre et la sonorité d'une voix donnée. Selon l'association américaine de normalisation (American Standards Association, 1960), le timbre est définie comme ≪ l'attribut de la sensation auditive qui permet à l'auditeur de différencié deux sons de même hauteur et de même intensité et présentés de façon similaire ≫. On peut considérer que l'oreille fonctionne en deux temps. La première écoute mise en place est l'écoute causale qui avais pour but d'identifier la source sonore. Cela correspondrait à reconnaître le timbre causal, c'est-à-dire dans le cas de la voix, l'empreinte vocale qui permet d'identifiée le locuteur. La deuxième écoute est l'écoute

qualitative d'un extrait vocal. Elle amène à une analyse des qualités du son, donc des qualités de la voix pour cette étude. Au niveau du timbre, c'est ce qu'on appelle la couleur ou la sonorité. Le terme de qualité vocale n'est pas encore bien définie, tout simplement parce qu'il repose sur la perception humaine, influencée par le contexte d'écoute et son environnement propre. L'évaluation se fait par comparaison entre diverses productions sonores. Il faut noter que la qualité vocale peut s'analyser de façon globale au niveau d'une phrase ou locale au niveau d'une voyelle comme l'illustre la figure 1.2.

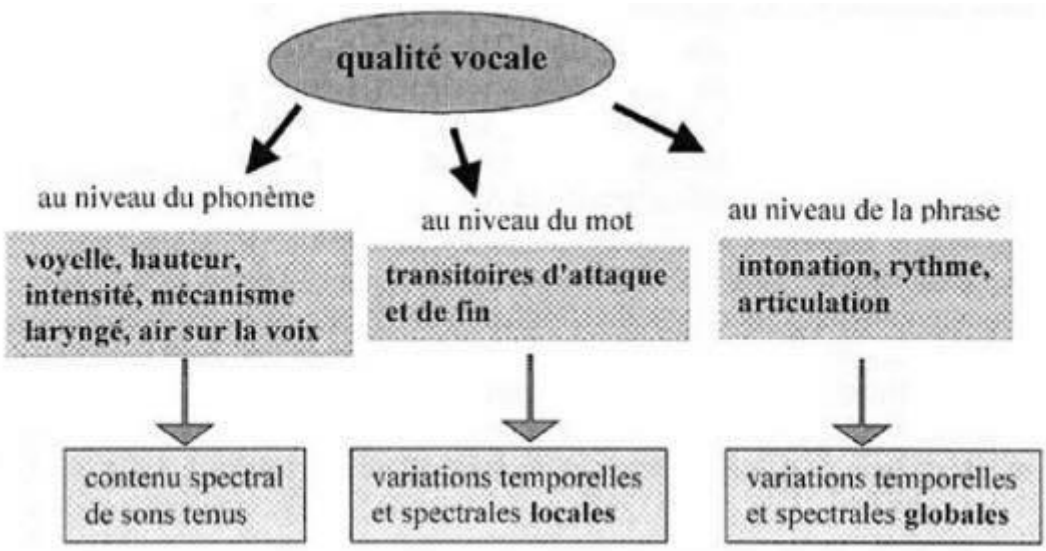


Figure 2.2 : Schéma sur la notion de qualité vocale.

Pour étudier la qualité vocale, on peut avoir recours au processus de verbalisation qui repose sur l'analyse syntaxique des propos recueillis lors de la qualification d'un stimulus sonore. Or le langage étant une notion complexe, les définitions des différents termes sont rarement universelles. La qualité vocale peut ainsi être définie de façon très générale, comme ce qui différencie deux productions vocales ayant le même contenu lexical. Plusieurs paramètres entrent en jeu, tels que la hauteur, la prosodie, le rythme, l'intonation, l'intensité et l'articulation. La qualité vocale peut aussi être considérée comme porteuse de sens et d'émotions. Pour l'analyse de la qualité vocale, on peut adopter plusieurs points de vue complémentaires : physiologique, perceptif, psycholinguistique, acoustique.

2.3 Dimensions perceptives et leurs corrélats acoustiques

Dans cette section il y a différentes caractéristiques et dimensions perceptives du timbre de la voix, en lien avec les paramètres du signal acoustique qui leur sont corrélés.

2.3.1 Le souffle

Le souffle est une notion plutôt vaste qui englobe notamment la gestion, le contrôle et la pression de l'air. Il se manifeste par la présence d'une composante bruitée dans le son. Contrairement à la tradition classique de la musique occidentale qui recherche plutôt la

“pureté de l’émission”, le souffle est un élément primordial dans certaines traditions orales ou il peut être considéré comme un symbole de vie. Le souffle peut correspondre à une voix blanche 1 (sans timbre), une voix chuchotée, une voix avec présence d’air. Le son peut, dans ce cas, être qualifié de sourd. Les analyses acoustiques décriraient, entre autres, un spectre pauvre en harmoniques et présentant un rapport signal à bruit (RSB) important.

2.3.2 La raucité

La raucité désigne la sensation et la perception de bruit dans le signal vocal. On parle du caractère rauque de la voix. Cette qualité est beaucoup plus courante pour des voix masculines que féminines et proviendrait d’une irrégularité de l’onde glottique. Elle serait liée à deux paramètres de bruits nommés jitter et shimmer.

2.3.3 Le vibrato et autres ornements

Les techniques ornementales sont omniprésentes aussi bien dans les musiques dites populaires et savantes que dans les chants de tradition orale. Elles sont liées à la gestion du souffle. Le vibrato est un caractère intrinsèque de la voix. Toutefois, il peut être plus ou moins accentué par les chanteurs par des techniques vocales. Le vibrato est perçu comme une oscillation autour d’une hauteur émise ce qui rend la ligne mélodique instable (voir la figure 1.3).

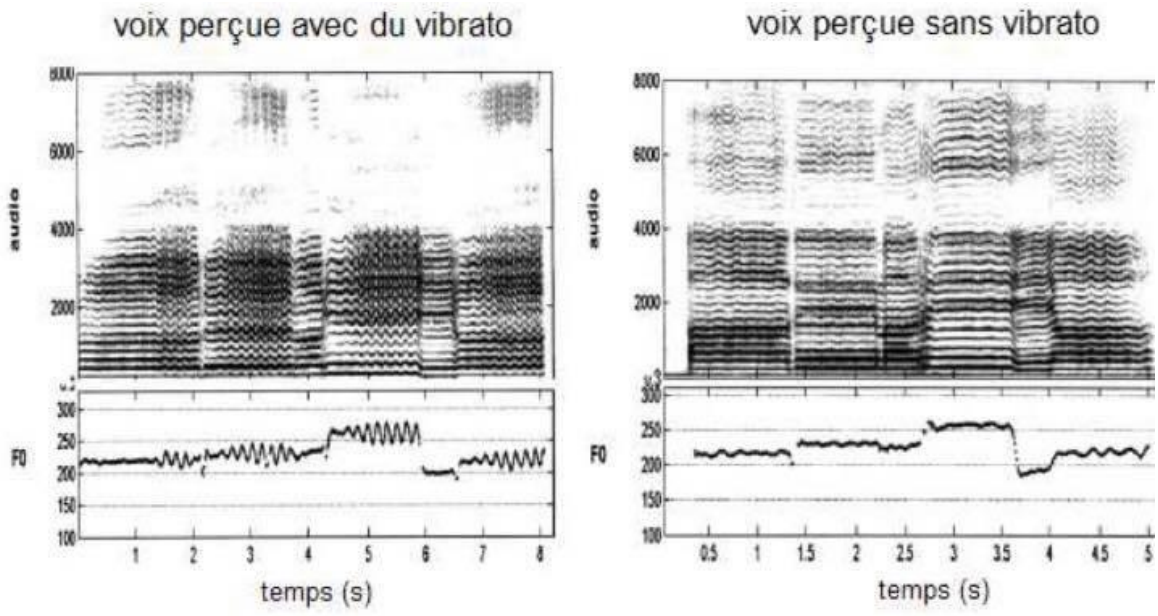


Figure 2.3 : Analyse de la fréquence fondamentale mettant en évidence la présence ou non de vibrato

On qualifie de son droit, un son produit avec l’absence de vibrato. Au niveau acoustique, le vibrato se définit par une modulation en fréquences et en amplitudes de la fréquence fondamentale. Quatre paramètres permettent de le caractériser : la vitesse, la profondeur, l’homogénéité, la dynamique interne. Dans le cas d’un vibrato, la norme est de 5 `a 7 oscillations par seconde. Si la vitesse du vibrato est élevée, c’est-à-dire supérieure à 7 Hz, la

perception est alors modifiée, il apparait alors un trémolo. Au-dessous de 5 Hz, la voix chevrote. Le portamento, est une technique qui permet à la voix de lier deux notes en parcourant l'étendue du spectre sonore séparant ces deux notes.

2.3.4 Voix timbrée/détimbrée

Le caractère timbré de la voix serait lié au renforcement du formant du chanteur (région de 2000Hz à 4000Hz) et aussi à la richesse spectrale dans les fréquences moyennes. Les représentations de la répartition d'énergie spectrale par bandes de fréquences confirment ces hypothèses. La voix détimbrée pourrait être assimilée à une atténuation du spectre dans la zone du formant du chanteur, la présence restreinte d'harmoniques aigues. Le son peut alors paraitre terne, aggravé ou avec la présence de souffle.

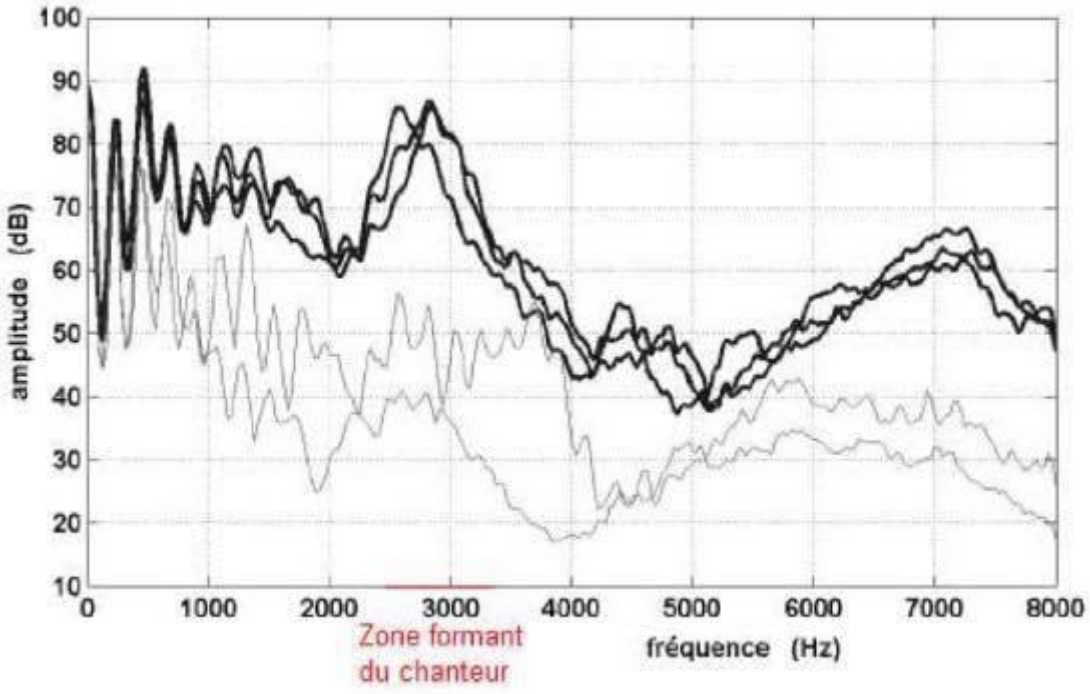


Figure 2.4 : Spectre d'amplitude d'une phrase musicale. Trait foncé : voix jugée timbrée; trait clair : voix jugée détimbrée

2.3.5 La brillance

La brillance est un attribut du timbre beaucoup étudié en psycho acoustique. Cette notion pourrait être liée à la différence d'amplitude entre les maximas des deux zones spectrales [0-2000Hz] et [2000-4000Hz]. Il faut aussi prendre en compte l'émergence du formant du chanteur certes moins caractéristique que pour la qualité de timbrage de la voix. La brillance pourrait se définir comme un timbrage dans l'aigu. Cela correspond à la présence d'harmoniques aigues qui se caractérise au niveau acoustique par le centre de gravité spectrale (CSG).

2.3.6 La nasalité

La nasalité est une qualité facilement identifiable à l'écoute. Elle est très présente dans certaines cultures. Du point de vue physiologique, elle est liée à l'abaissement du voile du palais et aux résonances du son dans les cavités nasales. Cependant la nasalité peut être induite par deux mécanismes différents. On parlera du twang 1 nasal et du twang pharyngé. Dans le cas du twang pharyngé, il n'y a pas de résonance dans les cavités nasales mais la constriction du pharynx produit un son très mince qui peut être perçu comme étant nasal.

3. Caractérisation de l'identité vocale [8]

Pour développer un système de conversion de voix, il faut connaitre les paramètres acoustiques caractérisant le locuteur. Cela nécessite une bonne compréhension du processus de production de la parole. Dans la première partie du présent chapitre, nous décrivons le processus de production de la parole ainsi que les mécanismes mis en œuvre lors de la phonation. Ensuite, nous décrivons les variabilités interlocuteurs, ainsi que les paramètres acoustiques servant à la discrimination des locuteurs par des humains.

3.1 Mécanismes de production de la parole

Le processus de production de la parole est un mécanisme à caractéristique très complexe qui repose sur une interaction entre les systèmes neurologique et physiologique. La parole commence par une activité neurologique. Après que soient survenues l'idée et la volonté de parler, le cerveau dirige les opérations relatives à la mise en action des organes phonatoires. Le fonctionnement de ces organes est bien, quant à lui, de nature physiologique.

Une grande quantité d'organes et de muscles entrent en jeu dans la production des sons des langues naturelles. Le fonctionnement de l'appareil phonatoire humain repose sur l'interaction entre trois entités : les poumons, le larynx, et le conduit vocal.

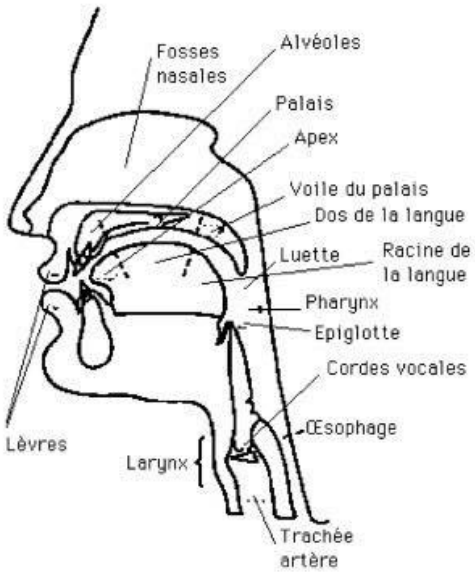


Figure 2.5 : Vue globale de l'appareil de production de la parole.

Le larynx est une structure cartilagineuse qui a pour fonction de réguler le débit d'air via le mouvement des cordes vocales. Le conduit vocal s'étend des cordes vocales jusqu'aux lèvres dans sa partie buccale et jusqu'aux narines dans sa partie nasale.

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. L'air des poumons est comprimé par l'action du diaphragme. Cet air sous pression arrive ensuite au niveau des cordes vocales. Si les cordes sont écartées, l'air passe librement et permet la production de bruit. Si elles sont fermées, la pression peut les mettre en vibration et l'on obtient un son quasi périodique dont la fréquence fondamentale correspond généralement à la hauteur de la voix perçue. L'air mis ou non en vibration poursuit son chemin à travers le conduit vocal et se propage ensuite dans l'atmosphère. La forme de ce conduit, déterminée par la position des articulateurs tels que la langue, la mâchoire, les lèvres ou le voile du palais, détermine le timbre des différents sons de la parole. Le conduit vocal est ainsi considéré comme un filtre pour les différentes sources de production de parole telles que les vibrations des cordes vocales ou les turbulences engendrées par le passage de l'air à travers les constrictions du conduit vocal.

Le son résultant peut être classé comme voisé ou non voisé selon que l'air émis a fait vibrer les cordes vocales ou non. Dans le cas des sons voisés, la fréquence de vibration des cordes vocales, dite fréquence fondamentale ou pitch, noté F0, s'étend généralement de 70 à 400 hertz. L'évolution de la fréquence fondamentale détermine la mélodie de la parole. Son étendue dépend des locuteurs, de leurs habitudes mais aussi de leurs états physique et mental.

Un exemple de signal de parole correspondant à la prononciation du mot (sa) est donné à la Figure 2.6. Le son (sa) est représenté dans le domaine temporel, la première partie (de 0 à 80 ms) est non voisée, c'est un signal non périodique de faible énergie.

La dernière partie représente un signal quasi-périodique avec une énergie plus grande, et est donc voisée.

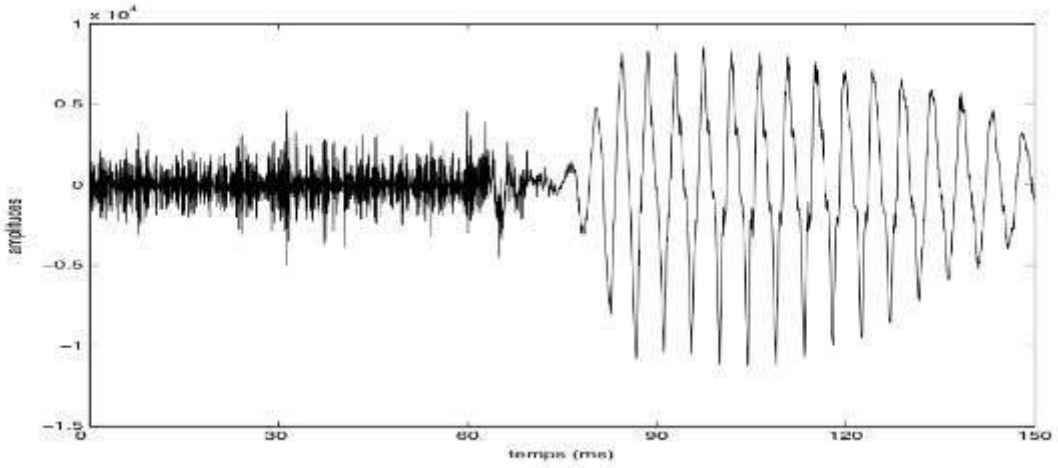
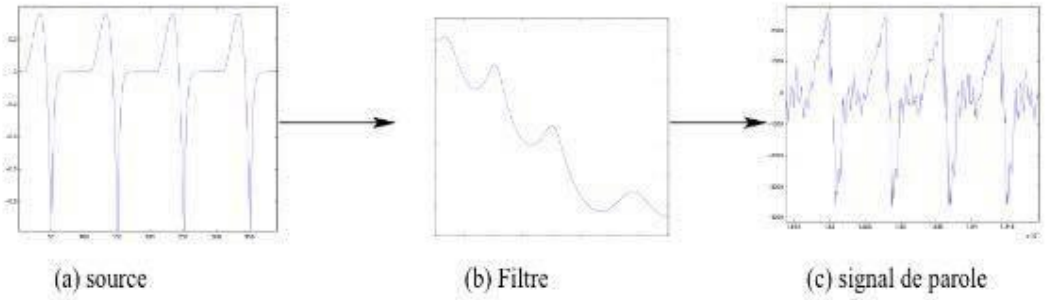


Figure 2.6 : Exemple d'un signal de parole

Le processus de production de la parole peut être représenté par le modèle source-filtre (Figure 2.7. (b)). Le signal de parole est modélisé comme la sortie d'un filtre linéaire variant dans le temps, qui simule les caractéristiques spectrales de la fonction de transfert du conduit vocal, excité par un signal source qui reflète l'activité des cordes vocales dans les zones voisées et le bruit de friction dans les zones non voisées. Quoique simpliste, cette représentation est capable de décrire la majorité de phénomènes de la parole et a été à la base de nombreux codeurs et synthétiseurs de parole.

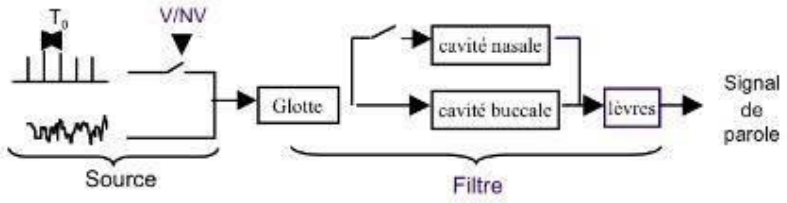
La décomposition source/filtre est une théorie particulièrement bien adaptée au problème de la conversion de voix. Transformer les paramètres de filtre revient à simuler la modification des caractéristiques du conduit vocal alors que la modification des paramètres du signal source simule les changements de la prosodie et des caractéristiques du signal d'excitation glottique. Des travaux de recherche ont permis d'apporter des informations a priori sur la forme du signal d'excitation glottique dans le cas des sons voisés. Ces études ont abouti à une modélisation théorique du signal glottique par un ensemble de paramètres pertinents : fréquence fondamentale, quotient d'ouverture, bruit de friction, etc... Cependant, l'extraction des paramètres pertinents du signal glottique reste un problème épineux. C'est d'ailleurs le manque de robustesse de ces techniques de dé convolution source-filtre qui fait que le signal glottique est encore peu utilisé tel quel en conversion de voix.

Une approximation classiquement employée consiste à considérer que le signal de source est constitué d'impulsions générées aux instants de fermeture de la glotte auxquelles s'ajoute un bruit blanc. Dans un tel modèle présenté en figure 2.7. (a), le spectre de la partie "filtre" appelée aussi enveloppe spectrale est composée du spectre du filtre décrivant le conduit vocal auquel s'ajoute la partie lisse du spectre glottique. Suivant le modèle du signal glottique utilisé, cette partie lisse du spectre du signal glottique peut être modélisée par un modèle AR d'ordre 2 ou 4. Certaines caractéristiques de ce modèle AR telles que la position du formant glottique et la pente spectrale sont d'ailleurs utilisées pour caractériser la qualité vocale du signal de parole. La partie "filtre" ainsi modélisée est porteuse des informations relatives à "l'empreinte" vocale d'un locuteur, c'est pourquoi elle est également dénommée timbre.



(a) source (b) Filtre (c) signal de parole

(a) Modèle-source filtre équivalent (Sons voisés)



(b) Modèle source-filtre

Figure 2.7 : Production de la parole.

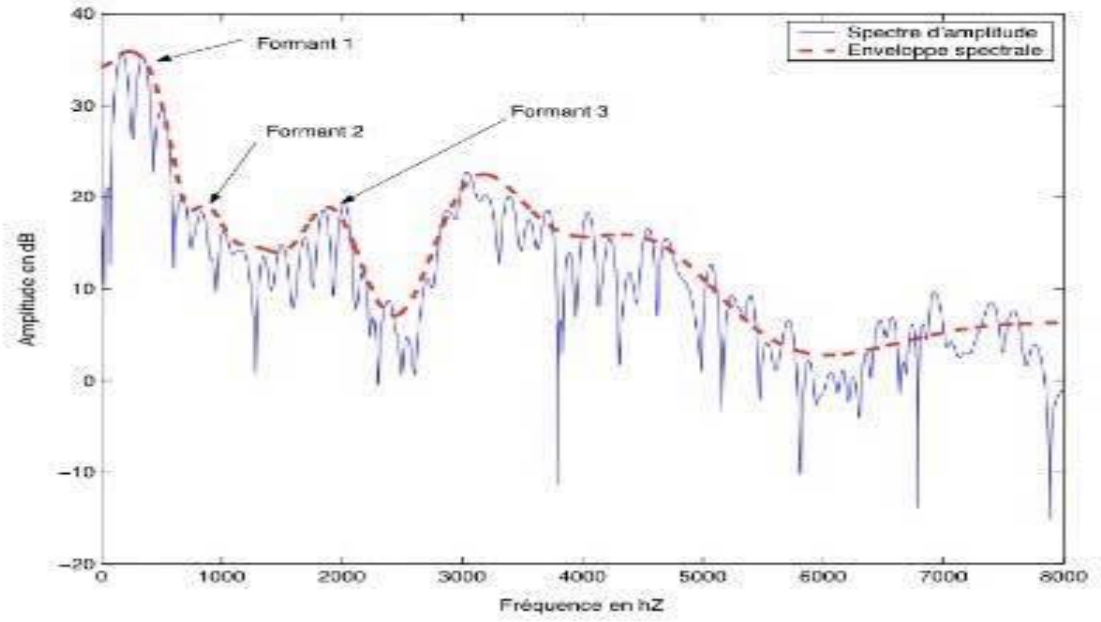


Figure 2.8 : Enveloppe spectrale.

3.2 Principe d'un système de conversion de voix

Dans la mise en œuvre d'un système de conversion de voix, on peut donc distinguer deux phases principales. La première est une phase d'apprentissage durant laquelle des signaux de parole source et cible sont utilisés pour estimer une fonction de transformation. La deuxième est une phase de transformation durant laquelle le système utilise la fonction de transformation précédemment apprise pour transformer des nouveaux signaux de parole source d'une façon qu'ils semblent, à l'écoute, avoir été prononcés par le locuteur cible.

3.2.1 Phase d'apprentissage

La figure 2.9(a) présente le principe général de la phase d'apprentissage d'un système de conversion de voix. La phase d'apprentissage nécessite trois composantes principales :

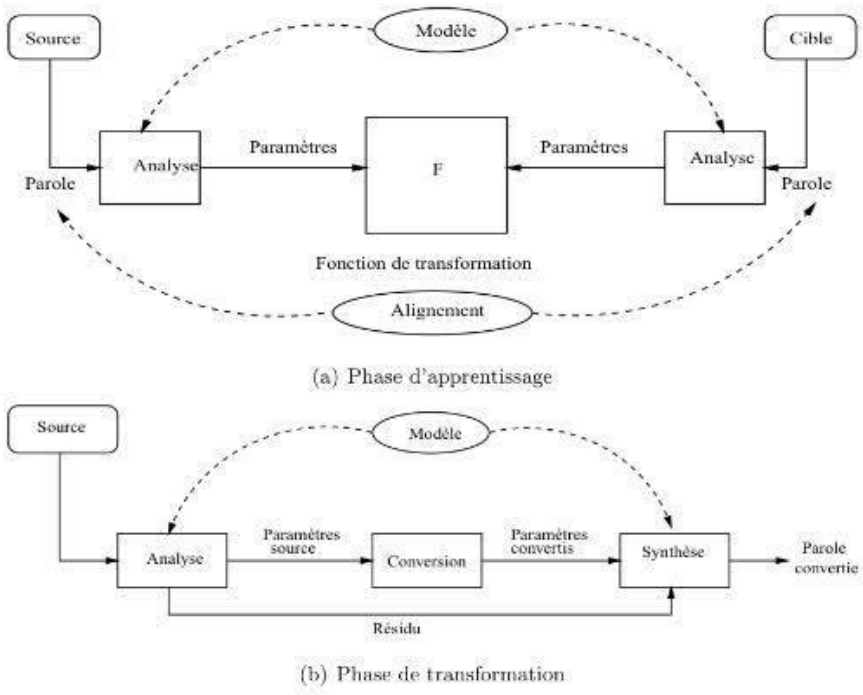


Figure 2.9 : Phase d'apprentissage et de transformation d'un système de conversion de voix.

Deux corpus de parole source et cible comprenant le même contenu phonétique;

– Un modèle mathématique du signal de parole pour déterminer quels paramètres vont être modifiés par le système;

– Une fonction de transformation décrivant la manière dont les paramètres sources seront modifiés.

La phase d'apprentissage commence par une étape d'analyse des corpus de parole source et cible suivant le modèle mathématique, afin d'extraire les paramètres acoustiques utiles à

l'estimation de la fonction de transformation. La nature des paramètres acoustiques utilisés a une dépendance vers le système de conversion. La plupart des travaux menés dans ce domaine traitent essentiellement de la transformation de l'enveloppe spectrale, modélisée généralement par une variante des coefficients de prédiction linéaire (LPC, LSF, LAR) par cepstre discret ou par des paramètres relatifs aux formants. A ces modifications d'enveloppe spectrale sont généralement associées des modifications de pitch allant d'une simple mise à l'échelle à une véritable prédiction de contours de pitch.

Le but de la fonction de transformation est d'établir le lien entre les paramètres de la source et ceux de la cible. Naturellement, le style d'élocution des locuteurs source et cible ne sont pas rigoureusement identiques, ce qui se traduit par une différence entre les unités linguistiques observées (durées de phonèmes par exemple). Or, pour l'apprentissage de la fonction de

transformation, les paramètres de la source et de la cible doivent être temporellement alignées de manière à décrire le même contenu phonétique. L'alignement temporel est réalisé à l'aide de l'algorithme DTW dans la plupart des systèmes de conversion de voix présentés. Cependant, il est possible de réaliser cet alignement avec d'autres méthodes comme les chaines de Markov cachées. Toutes ces techniques d'alignement opèrent essentiellement sur des paramètres d'enveloppe spectrale.

Cette base de données alignées sera, utilisée par la suite pour l'estimation de la fonction de transformation. Dans la littérature, la fonction de transformation a été implémentée par diverses méthodes, comme la quantification vectorielle, les réseaux de neurones, l'alignement fréquentiel dynamique, et les modèle de mélange de gaussiennes. Certaines de ces méthodes seront exposées dans la section suivante.

3.2.2 Phase de transformation

Après avoir défini la forme de la fonction de transformation et le modèle d'analyse du signal de parole, il reste à définir la stratégie de conversion de voix proprement dite. Cette stratégie est commune à tous les systèmes de conversion de voix : la conversion est simulée par l'application trame par trame de la fonction de transformation à des signaux de parole source.

La figure 2.9(b) présente l'architecture de la phase de transformation. D'un point de vue général, elle nécessite trois étape à la conversion de voix. Tout d'abord, une analyse est menée sur chaque trame de façon à extraire les paramètres de la source. Une partie des paramètres, par exemple ceux relatifs au timbre voire au pitch, sont modifié par le module de conversion. Ensuite, les paramètres modifiés ainsi qu'un résidu (les paramètres non modifiés) sont transmis à un module de synthèse qui réalise ainsi la génération du signal de parole converti.

3.3.Dynamic Time Warping[11]

La particularité de la méthode **Dynamic Time Warping (DTW)** est de savoir gérer les décalages temporels qui peuvent éventuellement exister entre deux séries (Berndt et Clifford 1994). Au lieu de comparer chaque point d'une série avec celui de l'autre série qui intervient au même instant t, on permet à la mesure de comparer chaque point d'une série avec un ou plusieurs points de l'autre série, ceux-ci pouvant être décalés dans le temps (Figuge.2.10).

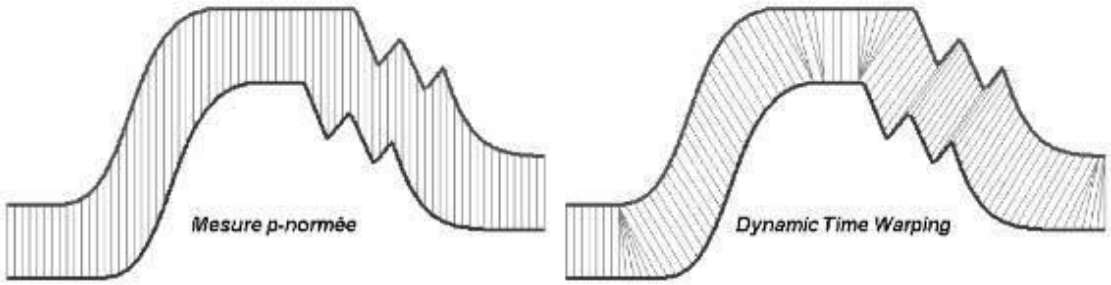


Figure.2.10 : Comparaison entre la mesure p-normée et le DTW

Le DTW possède une définition récursive qui calcule la similarité entre les séries

$Q = q_1,q_2,...,q_m$ et $C = c_1,c_2,...,c_n$ de la manière suivante :

Soit $D(i,j)$ la distance entre les sous-séquences $q_1,q_2,...,q_i$ et $c_1,c_2,...,c_j$ (avec $1 \leq i \leq m$ et $1 \leq j \leq n$) :

$$D(i,j) = \begin{cases} |q_1 - c_1|, & \text{si } i = j = 1 \\ |q_i - c_j| + min\{D(i-1,j), D(i-1,j-1), D(i,j-1)\}, & \text{sinon.} \end{cases}$$

Soit $Sim(Q,C)$ la mesure de similarité DTW entre les séries Q et C :

$$Sim(Q,C) = \frac{1}{D(m,n)}$$

On peut aussi paramétrer l'écart temporel maximum permis à la mesure pour comparer deux points. On définit ainsi une fenêtre temporelle (appelée delta) que l'on fera "glisser" sur chacune des séries à comparer. Par exemple, si on fixe la taille de la fenêtre delta = 3, chaque point d'une série qui intervient à un instant t ne pourra être comparé qu'avec les points de l'autre série qui interviennent aux instants t−3, t−2, t−1, t, t+1, t+2 et t+3. Le choix de la taille de cette fenêtre peut avoir une influence sur le résultat et il convient de la déterminer avec soin (Ratanamahatana et Keogh 2004 a).

4. Conclusion

La conception de la voix, son traitement sont les premières étapes pour la reconnaissance et l'identification vocale, les techniques sont nombreuses, certaines plus robustes et plus efficaces que d'autres. Le traitement automatique de la parole repose sur des données analogiques en fonction du temps. L'extraction des meilleurs paramètres aide, sans aucun doute, à ce traitement. L'intelligence artificielle peut intervenir pour trouver les paramètres pertinents, ou utiliser n'importe quels représentants de la parole pour faire la segmentation ou la classification. Dans le chapitre suivant, on va résumer notre solution et les résultats obtenus.

Dans notre application nous avons choisi de travailler avec le Maven et la bibliothèque recognition, on va montrer ça dans le chapitre qui va suivre et expliquer les utilités de notre application.

Chapitre III : La Réalisation

1. Introduction

L'étape la plus importante du projet est celle du développement afin d'atteindre notre but qui est de réaliser une application robuste et qui s'approche du concret avec la notion de reconnaissance vocale.

Dans ce chapitre on va citer les différentes étapes suivies pour la réalisation de notre application, en commençant par l'analyse et la conception jusqu'au développement final.

2. Les outils de réalisation

2.1 JAVA :[13]

Java est un langage de programmation orienté objet créé par James Gosling et Patrick Naughton, employés de Sun Microsystems, avec le soutien de Bill Joy (cofondateur de Sun Microsystems en 1982), présenté officiellement le 23 mai 1995 au SunWorld.

La société Sun a été ensuite rachetée en 2009 par la société Oracle qui détient et maintient désormais Java.

La particularité et l'objectif central de Java est que les logiciels écrits dans ce langage doivent être très facilement portables sur plusieurs systèmes d'exploitation tels que Unix, Windows, Mac OS ou GNU/Linux, avec peu ou pas de modifications. Pour cela, divers plateformes et Framework associés visent à guider, sinon garantir, cette portabilité des applications développées en Java.

2.2 JAVASCRIPT :

Utilisé dans le développement des applications web et mobiles actuelles, le JavaScript peut à la fois être utilisé côté client, c'est-à-dire interprété par le navigateur web de l'internaute, et côté serveur avec l'utilisation de Node.js. Le JavaScript est le langage de prédilection pour interagir avec le HTML permettant ainsi d'apporter du dynamisme à l'intérieur des pages web.

Ce dynamisme apporté aux pages web fut d'abord amené par Flash, technologie concurrente de JavaScript et apparue en même temps que ce langage dans les années 90. JavaScript ne s'est imposé que tardivement, à partir de 2010, jusqu'à devenir un standard du web. Cela étant notamment dû à l'évolution de ce langage qui permet d'obtenir des résultats équivalents à Flash sans nécessiter l'installation d'une extension au navigateur ainsi qu'au déclin de la présence de la technologie Flash sur mobile.

2.3 L'IDE Netbeans :

Netbeans est un IDE qui supporte une large variété de langages de programmation et d'outils de collaboration, pour la réalisation de notre projet on a utilisé la version8.0.

Pourquoi l'utiliser :

- Un contexte de déploiement runtime pour des fonctionnalités arbitraires qui simplifient le développement
- Une boîte à outils qui permet de gagner beaucoup de temps en développement et d'effort
- Un ensemble d'abstractions qui permet aux développeurs de se concentrer sur le business logique, et non de réécrire de la logique de routine et des composants requis par la plupart des applications
- Un ensemble de Standards pour rehausser et renforcer la consistance et l'interopérabilité entre les applications et les systèmes d'exploitation [9].

2.4 WAMP Server : [10]

WampServer (anciennement WAMP5)est une plateforme de développement Web de type WAMP,permettant de faire fonctionner localement (sans avoir à se connecter à un serveur externe) des scripts PHP. WampServer n'est pas en soi un logiciel, mais un environnement comprenant trois serveurs (Apache, MySQL et MariaDB), un interpréteur de script (PHP), ainsi que phpMyAdmin pour l'administration Web des bases MySQL.

Il dispose d'une interface d'administration permettant de gérer et d'administrer ses serveurs au travers d'un tray icon (icône près de l'horloge de Windows).

La grande nouveauté de WampServer 3 réside dans la possibilité d'y installer et d'utiliser n'importe quelle version de PHP, Apache, MySQL ou MariaDB en un clic. Ainsi, chaque développeur peut reproduire fidèlement son serveur de production sur sa machine locale.

3 -Étape de développement

3.1 Analyse des besoins :

Pour l'analyse on doit discuter les différents cas pour notre application, donc cette dernière a des objectifs comme suit:

- Construction d'un système basé sur la reconnaissance vocale,
- Les utilisateurs peuvent tous utiliser l'application,
- Les utilisateurs peuvent récupérer ce qu'ils ont dit sous forme d'un fichier texte,

3.2. Technologies mises en œuvre

Le projet concerne la réalisation d'une application web, c'est-à-dire un logiciel utilisable via un navigateur internet standard. Ce type d'application repose principalement sur une architecture client-serveur : le client est le navigateur internet, le serveur est un programme qui fonctionne sur un ordinateur distant.

Et une application java qui permet la reconnaissance des bandes sonores déjà enregistrés dans le fichier de stockage.

3.3 Les taches des utilisateurs

L'utilisateur peut faire les tâches suivantes :

- Avoir accès au dictaphone afin de faciliter la rédaction des textes,
- Pouvoir reconnaitre les bandes sonores enregistrées pour des utilisations antérieures.

3.4. Maven et la bibliothèque recognito

Maven est un outil permettant d'automatiser la gestion de projets Java. Il offre entre autres les fonctionnalités suivantes :

- Compilation et déploiement des applications Java (JAR, WAR)
- Gestion des librairies requises par l'application
- Exécution des tests unitaires
- Génération des documentations du projet (site web, pdf, Latex)
- Intégration dans différents IDE (Eclipse, JBulder)

Dans le développement de notre application nous avons utilisé Maven sur laquelle on a utilisé une bibliothèque de reconnaissance vocale nommée recognito

Celle-ci regroupe les bibliothèques suivantes :

```
package com.bitsinharmony.recognito;

import java.io.Serializable;
import java.util.Arrays;
import java.util.concurrent.locks.Lock;
import java.util.concurrent.locks.ReentrantReadWriteLock;

import com.bitsinharmony.recognito.distances.DistanceCalculator;
```

-La **Sérialisation** nous permet de sauvegarder l'objet et l'ensemble de ses attributs.

-**Les tableaux** nous permet de stocker l'objet en cous de traitement.

-Les verrous de l'interface ***java.utils.concurrents.locks.locket concurrents.locks.reentrantReadWriteLock*** nous permet d'orchestrée les données qui serons utilisé.

```
import java.util.HashMap;
import java.util.Map;
```

Une Map est une classe qui permet d'associer une clé à un objet. Ainsi, il suffit de récupérer la clé (souvent un string ou un int) pour récupérer l'objet associé.

Une HashMap est une implémentation d'une Map. Elle consiste à calculer un Hash de la clé pour classer celles-ci et accélérer les temps de recherche.

```
import com.bitsinharmony.recognito.algorithms.LinearPredictiveCoding;
import com.bitsinharmony.recognito.algorithms.windowing.HammingWindowFunction;
import com.bitsinharmony.recognito.algorithms.windowing.WindowFunction;
```

Avec **LinearPredectiveCoding** la voix est reçue sous forme d'une séquence de segments.

HmmingWindowFunction utilisé en traitement du signal, le fenêtrage est utilisé dès que l'on s'intéresse à un signal de longueur volontairement limitée. En effet, un signal réel ne peut qu'avoir une durée limitée dans le temps ; de plus, un calcul ne peut se faire que sur un nombre fini de points.

Window.function effectue des calculs sur les enregistrements.

```
import java.io.IOException;

import javax.sound.sampled.AudioFormat;
import javax.sound.sampled.AudioInputStream;
import javax.sound.sampled.AudioSystem;
import javax.sound.sampled.UnsupportedAudioFileException;
```

-**AudioFormat** pour décoder le flux binaire

-**AudioInputStream** pour la lecture des fichiers audio.

-**AudioStream** pour que tous les sons soient supportés.

-**UnsupportedAudioFileException** gères les exceptions.

VoicePrint() : Recherchez l'empreinte vocale dans Wiktionary, le dictionnaire gratuit.

L'empreinte vocale peut se référer au spectrogramme d'une voix. Les utilisations plus spécifiques comprennent:

createVoicePrint() Crée le fichier audio dans le dossier spécifie après enregistrement.

initRecongito() Une méthode qui compare les signaux de la bande en cours avec celle déjà enregistre dans le fichier

getName() retourne le nom du fichier identifie.

3.5. Codage de la parole par prédiciton linéaire

Elle est la méthode utiliser par la bibliotheque récognito pour le codage et le traitement de la parole, Pour des considétations de capacités de canaux, il est nécessaire de réduire le débit de transmission de parole: sur un réseau numérique, le signal de parole est échantillonné au rythme de 8000 échantillons par seconde (pas d'échantillonnage de 125 µs) et codé sur 8 bits, soit un débit de 64 kbits/sec. Il faut ramener ce débit à 13 kbits/s. La technique retenue est celle de la prédiction linéaire dont nous tenterons d'expliquer le principe en quelques lignes. La génération du signal vocal peut être interprétée de la manière suivante. Les cordes vocales produisent un signal qui correspond à la mélodie; c'est une suite d'impulsions assez régulières donnant l'intonation. La forme du conduit vocal (ouverture de la bouche, position de la langue, ouverture ou fermeture du conduit vocal) caractérise le son qui est émis; en particulier les fréquences de résonance du conduit vocal sont caractéristiques des sons et sont reconnues par l'oreille . On peut modéliser le conduit vocal sous la forme d'un filtre linéaire à laquelle on applique le signal d'entrée généré par les cordes vocales ou bien dans le cas des sons fricatifs par un bruit. Comme les paramètres qui caractérisent le signal d'entrée et la forme du filtre varient relativment lentement, cette représentation permettra d'effectuer la réduction de débit recherchée.[12]

Le codage prédictif linéaire est reposer sur l'hypothèse de linéarité du estprocessus de production de la parole. Chaque échantillon peut être prédit à partir d'une pondération linéaire d'un nombre fini d'échantillons précédents, étant donné que la forme du conduit vocale n'évolue pas rapidement.

cette téchnique est largement utilisé en traitement de la parole notamment en transmissions. Le model est de plus facile à mettre en oeuvre dans les systèmes à temps réel. Ce model par définition ne prend pas en charge les phénomènes non linéaire, et de plus, il n'est pas optimisé pour la tâche de reconnaissance. En effet, des travaux ont montré qu'il est possible de mettre en oeuvre une meilleure représentation, et on fait naissance à des modèles plus optimisés comme la prédiction linéaire perceptive, la prédiction linéaire perceptive RASTH, le codage WLPC...

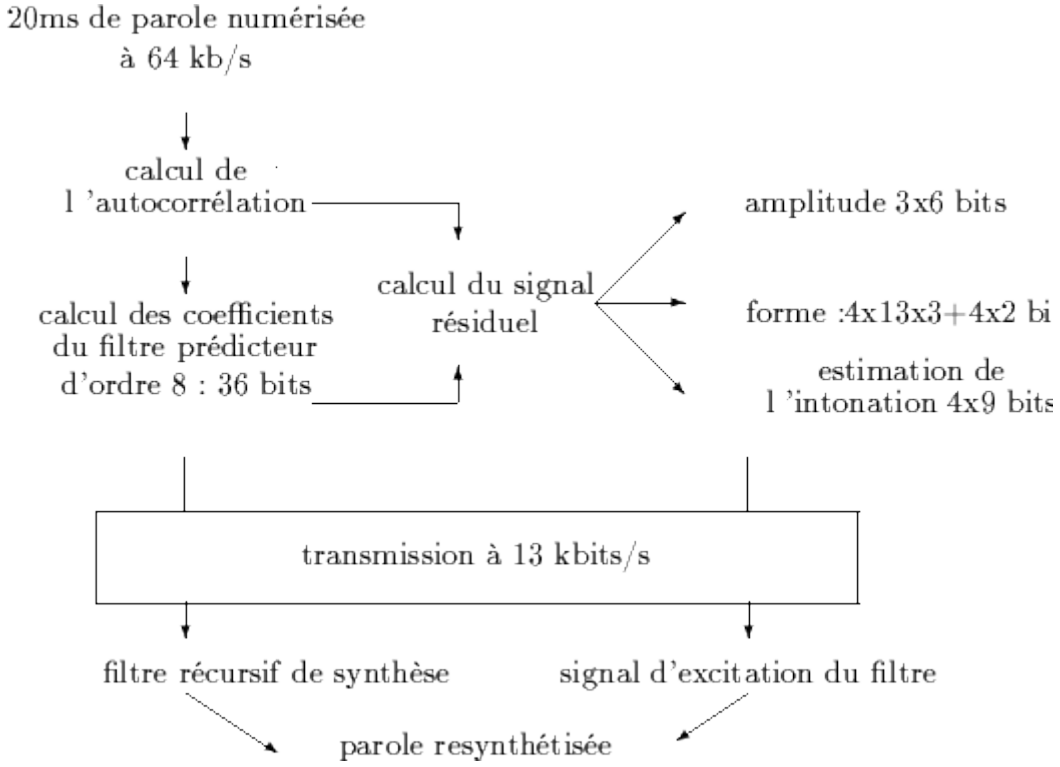


Figure 3.1 : Les différentes étapes du codage de parole par prédiction linéaire[12]

4. La Conception

La conception est une étape essentielle dans le développement d'une application, dans cette partie on va citer ces différentes étapes. L'étape 1 est la modélisation puis la réalisation.

On a un espace réservé aux utilisateurs qui peux soit rédiger des textes grâce à la reconnaissance vocale ou enregistrer des bandes sonores afin de les distinguer ensuite par l'application java

4.1. La modélisation

a. Les diagrammes

a.1.Diagramme de cas d'utilisation (application de reconnaissance vocale)

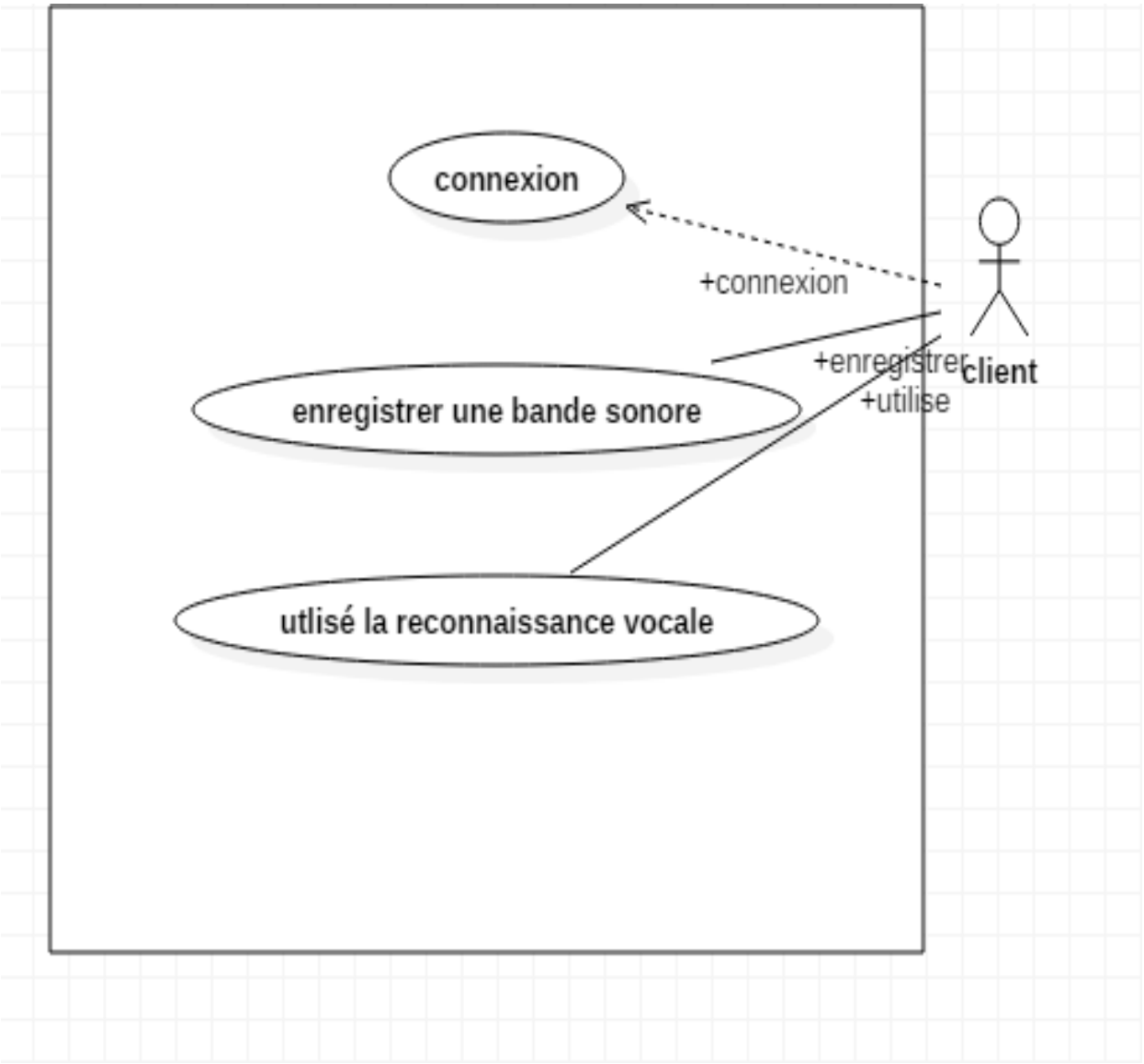


Figure 3.2 : Diagramme de cas d'utilisation.

a.2. Diagramme de séquence (application de reconnaissance vocale)

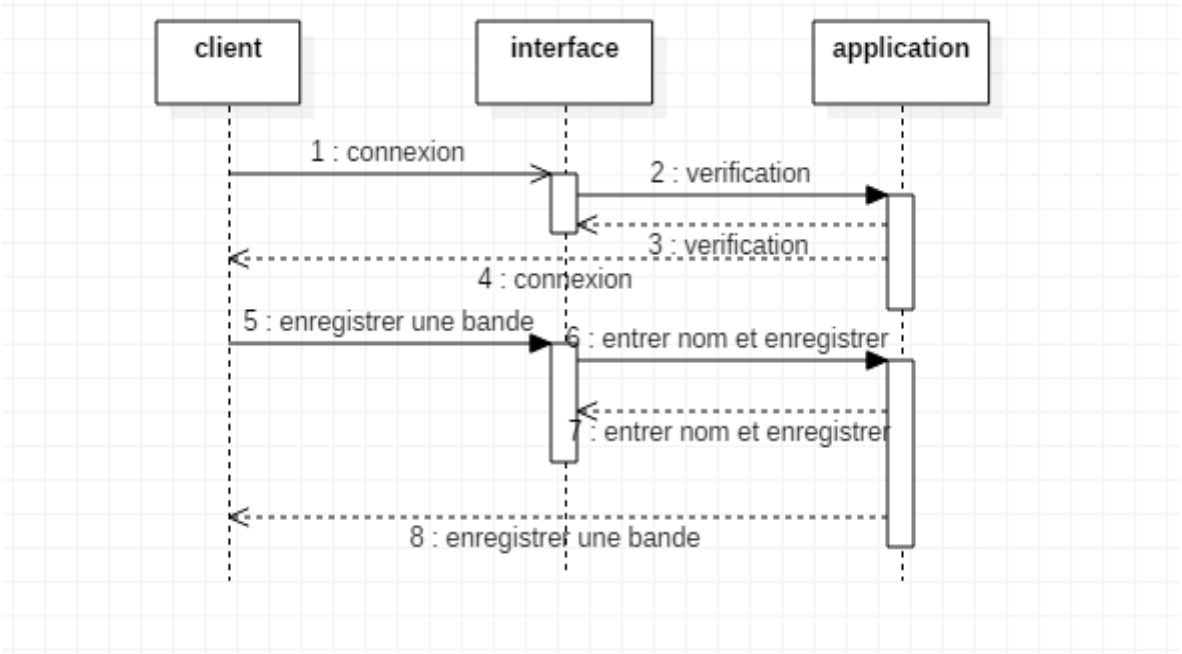


Figure 3.3 : Diagramme de séquence

a.3. Diagramme de cas d'utilisation (dictaphone)

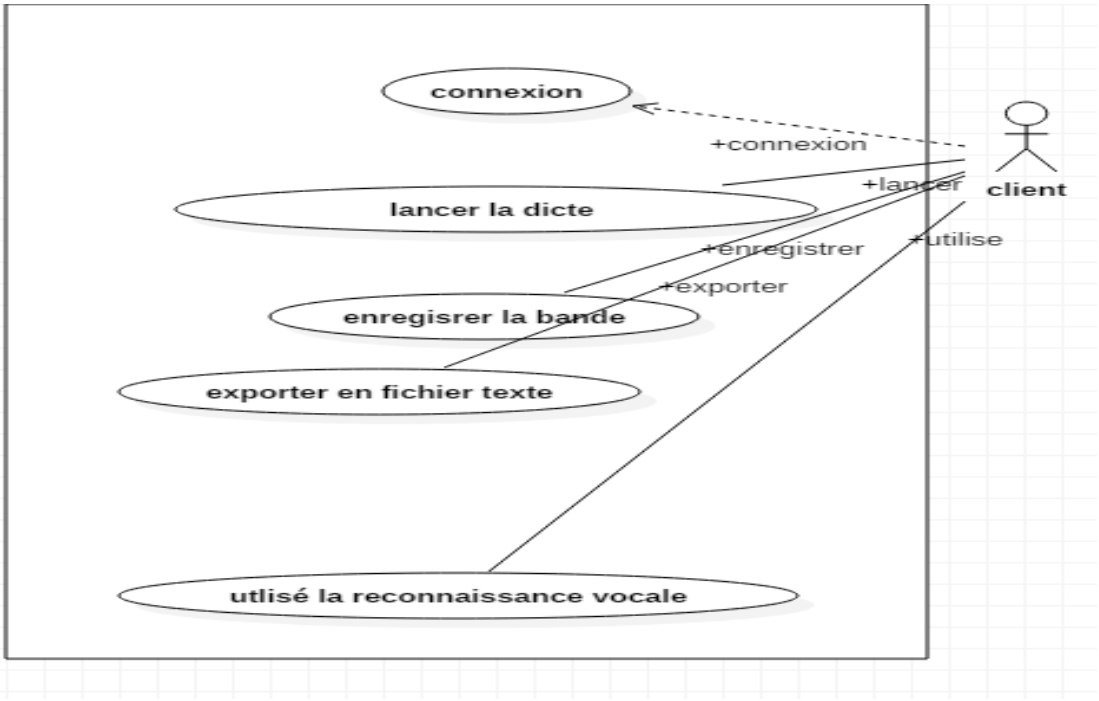


Figure 3.4 : Diagramme de cas d'utilisation (dictaphone)

a.4. Diagramme de séquence (dictaphone)

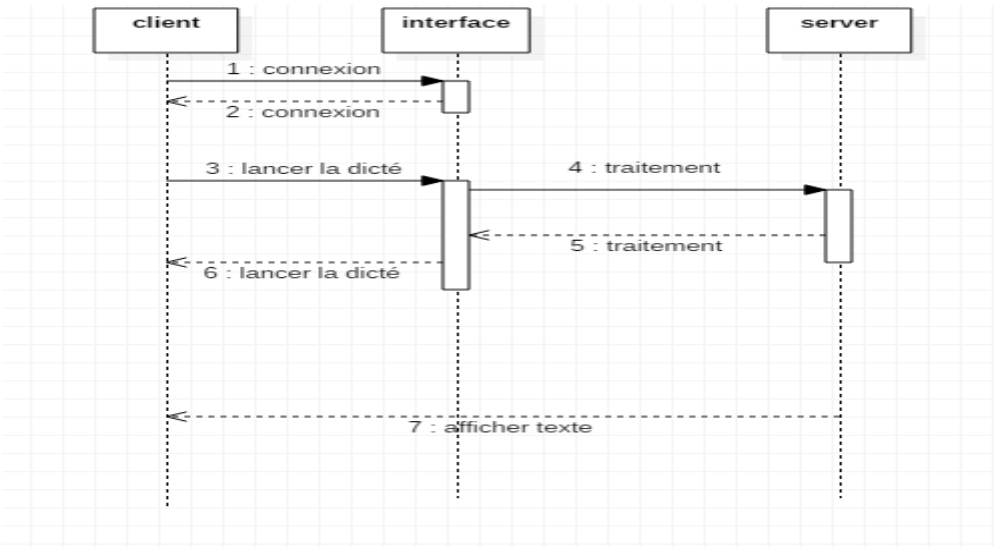


Figure 3.5 : Diagramme de séquence (dictaphone)

5. Réalisation

Ci-dessous des screenshots des exécutions de notre application.

5.1 Les étapes d'utilisation du dictaphone

Etape 1 :

L'utilisateur dois se connecte avec son mail et mot de passe pour accéder a l'application

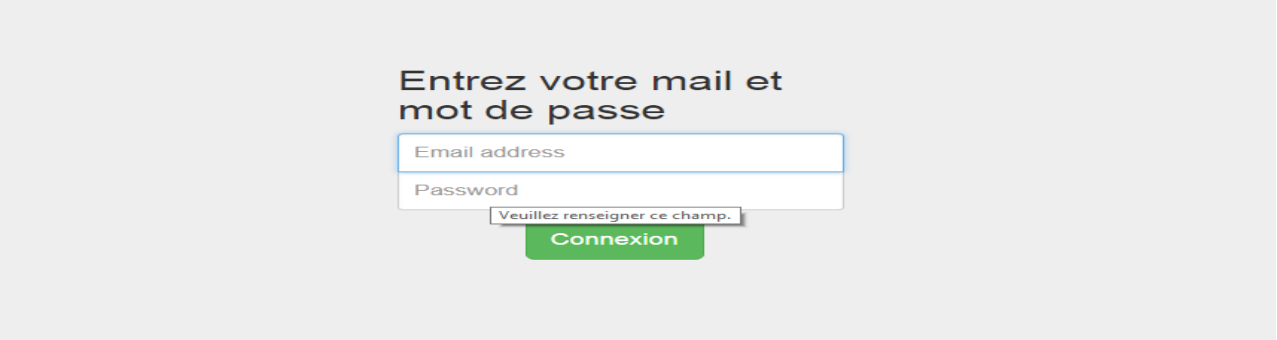


Figure 3.6 : Page login dictaphone.

Etape 2 :

Après avoir enter son mail et mot de passe correct il accède a l'interface suivante

Figure 3.7 : Page home dictaphone.

Etape 3 :

Après avoir cliqué sur le bouton « lancez la dictée » l'utilisateur doit parler et voit sa phrase s'afficher sur la zone texte, pour terminer il clique sur pause, si il souhaite il peut enregistrer sa voix afin de la récupérer par la suite ou bien l'exporter en fichier texte qui se trouvera dans le dossier ou l'application a été lancée.

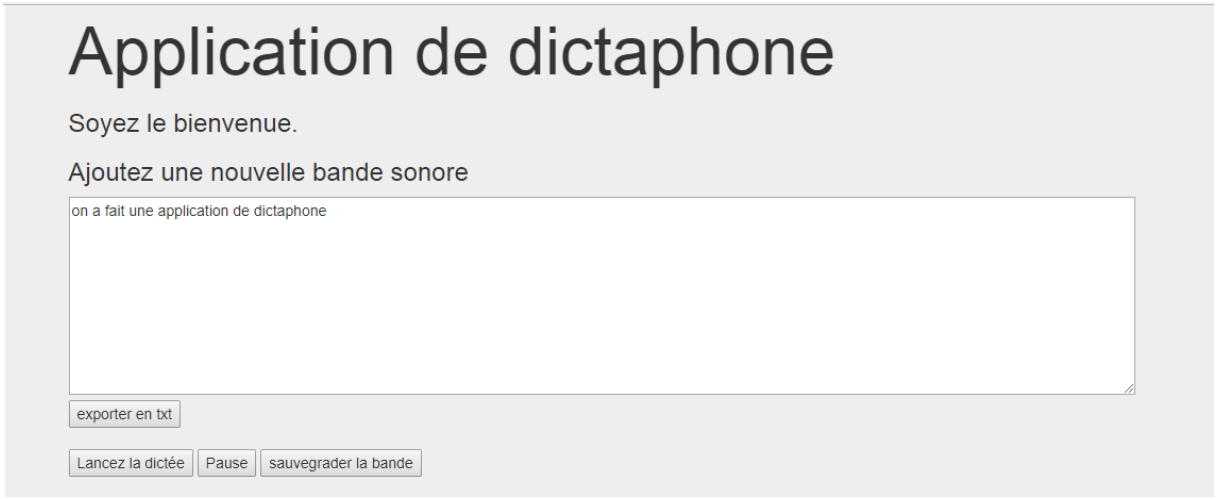


Figure 3.8 : Étape enregistrement de la voix.

5.2 Les étape d'utilisation de l'application de reconnaissance vocale

Etape 1 :

L'utilisateur dispose de cette interface afin de démarrer la reconnaissance des bandes sonores enregistrées ou bien d'enregistrer des nouvelles

Figure 3.9 : Menu home application reconnaissance vocale.

Etape2 :

Après avoir cliquer sur enregistrer une bande sonore l'utilisateur voit cette interface ou il rentre son nom ou donne un nom à la bande sonore, il peut lire la bande afin de savoir s'il la refait où s'il l'enregistre.

Figure 3.10 : Étape enregistrement de la parole

Après avoir enregistré la bande sonore, on peut ouvrir le dossier nommé voices ou nous retrouverons toutes les autres bandes sonores déjà enregistrées auparavant. On y remarque le fichier audioToReconize, qui contient la dernière bande enregistrée. Celle-ci sert à écraser les anciennes bandes qui portent le même nom que la nouvelle bande afin d'éviter la répétition des fichiers

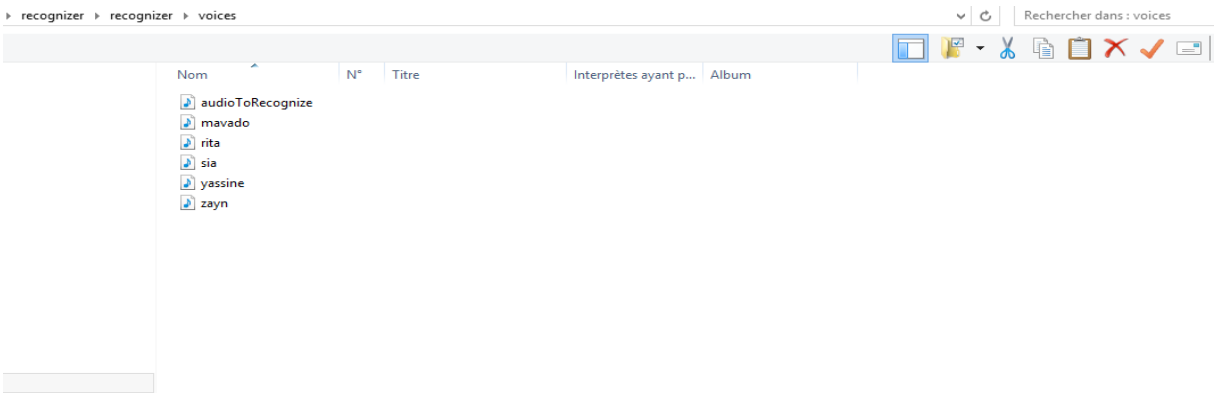


Figure 3.11 : Dossier contenant déjà les sons préenregistrés

Etape3 :

Pour lancer la reconnaissance vocale l'utilisateur clique sur démarrer puis expose au microphone la bande sonore après un certain temps il clique sur arrêter et l'interface retourne le résultat avec le nom de la bande sonore

Figure 3.12 : Résultat de la reconnaissance.

6. Conclusion

Dans ce chapitre on a cité les différentes étapes pour le développement et la réalisation de notre application, ce chapitre est le fruit de notre travail, durant plusieurs mois, on a fait des recherches bibliographiques, choisit les méthodes à suivre et enfin, c'était l'étape de l'implémentation et les résultats. Nous espérons avoir obtenu un produit satisfaisant, bien qu'encore imparfait, mais ce dont nous sommes surs, c'est que ce PFE nous a permis de mettre en pratique toutes nos connaissances informatiques et bien plus encore.

Conclusion Générale

Ce PFE traitant de la reconnaissance de musiques préenregistrées et de convertir la parole en fichier texte, a été mené à bien, et a répondu aux objectifs que nous nous sommes posés au préalable. Motivés par l'amélioration de la précision de la reconnaissance vocale par la fusion des différentes sources d'information, ce PFE se concentre sur l'exploitation de l'information de la source vocale. Les paramètres de la source vocale sont généralement jugés moins discriminants mais difficiles à extraire. Néanmoins, avec l'évolution de la technologie, des ressources de stockage et de calcul, des études dans le domaine de la compréhension du phénomène de production et de perception de la parole, ont poussé les chercheurs à reconsidérer ces préjugés et essayer de tirer le maximum de ces informations complémentaires pour améliorer les performances du système de reconnaissance vocale. Le principal défi dans la technologie de reconnaissance vocal est donc d'améliorer la robustesse des systèmes dans des conditions incompatibles. Notre système vocal fournit principalement les indices acoustiques pour la classification des phonèmes, et aussi la personnalité individuelle pour caractériser et identifier la parole. La voix est porteuse d'informations variées, la parole humaine considérée comme une émission de sons structurée, est essentiellement un vecteur de communication. A ce titre, un signal de parole est généralement porteur d'un message à destination d'une autre personne. 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 à du bruit. Nous avons fait une recherche bibliographique sur le sujet, ce qui nous a mené à notre choix de techniques et de bibliothèque utilisée, comme base à notre implémentation, puis nous avons procédé à la conception, réalisation, et conclu par des tests. Nous pensons encore améliorer le produit obtenu, et comme perspectives utiliser d'autres méthodes et comparer avec le taux de réponses positives et négatives. Ce PFE nous a permis d'apprendre à mener à bien un projet de A à Z, faisant face à toutes sortes de problèmes, tout en gérant le temps imparti.

Références

[1] M. F. Clemente Giorio, Kinect in Motion - Audio and Visual Tracking by Example, Packt Publishing, 2013.

[2] *Écrit par Joel Drakes - Responsable Avant-Vente - Nuance Communications Dans les Echos,2009*

[3]Isabelle Bellin -institut national de recherche en informatique et en automatique,2012

[4] SPECIALITE JANETTI- LA REPRESENTATION NUMERIQUE DU SON,2011

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

[6] Maxime Metzmacher- revue *Parcs & Réserves ,2008*

[7] Sarah Le Bagousse- Université Pierre et Marie Curie Paris,2008

[8] Taoufik En-Najjary. Conversion de voix pour la synthèse de la parole. Traitement du signal et de l'image. Université Rennes 1, 2005

[9] R. Dantas, NetBeans IDE 7 Cookbook, Packt Publishing, 2011.

[10] S. Pachev, Understanding MySQL Internals, O'Reilly Media, 2007.

[11] Rémi Gaudin, Nicolas Nicoloyannis, Apprentissage non supervisé de séries temporelles à l'aide des k-Means et d'une nouvelle méthode d'agrégation de series, 2012

[12] X. Lagrange, Ph. Godlewski et S. Tabbane, ``Réseaux GSM-DCS'', Hermès, 1995.

[13] C. D. Jeff Friesen, Beginning Java 7, apress, 2001.