Université Abou Bekr Belkaid
Tlemcen Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

République algérienne Démocratique et Populaire

Université Abou Bakr Belkaid, Tlemcen

Faculté de Science
Département d'Informatique

Mémoire de fin d'étude

Pour l'obtention du diplôme de Master en Réseaux et Système distribué.

Thème

Système sécurisé à base vocale

Réalisé par :
Mlle.BELGHITRI KARIMA.

Encadré par :
Mme. DIDI FADOUA.

Présenté le 23 Juin 2015 devant la commission d'examinassions composée de :

Mme. Iles Nawal	(President)
Mr. Belhocine Amine	(Examinateur)
Mme. labraoui Nabila	(Examinatrice)
Mme. DIDI FADOUA	(Promoteur)

Année universitaire 2014-2015

Remerciement

En préambule à ce mémoire, je souhaite adresser mes remerciements les plus sincères aux personnes qui m'ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable année universitaire.

Je tiens à remercier sincèrement Didi Fedoua d'abord en tant qu'encadreur de ce, mémoire ensuite pour m'avoir donné plus de confiance en moi en voulant bien accepter un thème qui me tenait à cœur, pour sa générosité et la grande patience dont il a fait preuve tout le long de mon travail malgré ses nombreuses charges académiques et professionnelle ,qui m'a appris à être rigoureux dans mes travaux afin d'éviter les obstacles qui pouvaient se présenter, qui s'est toujours montré attentive et disponible tout au long de la réalisation de ce mémoire, ainsi que pour l'inspiration, l'aide et le temps qu'il a bien voulu me consacrer sans quoi ce mémoire n'aurait jamais eu autant de succès.

J'exprime toute ma gratitude à tous les membres de jury lors des recherches effectuées et qui ont accepté de répondre à mes questions avec gentillesse.

Je n'oublie pas mes parents pour leur contribution, leur soutien et leur patience.

A mes parents, mes sœurs Houda et Amina, et à mon frère Sidahmed à mes beau frere Houari et Lakhdar : le petit ange Abd raouf, Serine, Mohamed et Abd Rahmane. Vous vous êtes dépensés pour moi sans compter. En reconnaissances de tous les sacrifices consentis par tous et par chacun pour me permettre d'atteindre cette étape de ma vie. A mes, amis, tantes, cousins et cousines affectueuses reconnaissances.

Je vous remercie pour votre patience et pour m'avoir aidé à avancer. Vous êtes tous pour moi comme une seconde famille.

Merci d'être toujours près de moi dans mes joies et mes peines. A tous mes camarades de département d'informatique. A tout le personnel de la faculté des sciences et de l'université de Tlemcen.

Enfin, j'adresse mes plus sincères remerciements à tous mes proches et amis, qui m'ont toujours soutenue et encouragé au cours de la réalisation de ce mémoire.

Merci à tous.

Belghitri Karima.

Table des matières

Chapitre I : Généralités

1. Introduction

La reconnaissance vocale est un domaine scientifique ayant toujours eu un grand attrait aussi bien auprès des chercheurs qu'auprès du grand public. Elle consiste à employer des techniques d'appariement afin de comparer une onde sonore à un ensemble d'échantillons, composés généralement de mots mais aussi, plus récemment, de phonèmes (unité sonore minimale). Le secteur de la reconnaissance vocale est en pleine croissance et cette technologie bien que très avancée, n'est pas encore aboutie, pouvant commencer à répondre aux attentes de l'homme. Bien que des progrès soient encore à faire sur les systèmes complexes de traitement et reconnaissance, il est à noter que la reconnaissance Vocale est quasiment parfaite. 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. 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.

2. Définition

La reconnaissance automatique du locuteur est une des branches de l'authentification afin de sécuriser un système, qui se réfère à la reconnaissance automatique de l'identité des personnes en utilisant certaines de leurs caractéristiques intrinsèques. Outre la voix, il y a beaucoup d'autres modèles physiques et comportementaux pour l'authentification, par exemple : l'iris, les réseaux veineux de la rétine, les réseaux veineux de la paume de la main, l'empreinte digitale, …etc. Pratiquement, la sélection d'un modèle adéquat devrait prendre en compte au moins les considérations suivantes : la robustesse, la précision, l'accessibilité et l'acceptabilité. Par rapport à ces critères de sélection, parmi toutes les technologies d'authentification , la reconnaissance du locuteur est probablement la plus naturelle et économique pour les systèmes de communication homme-machine parce que d'une part la collecte de données parole est beaucoup plus pratique que les autres motifs, et d'autre part, la parole est le mode dominant d'échange d'information pour les êtres humains et tend à être le mode dominant pour l'échange d'information pour les systèmes de communication homme-machine. [1]

2.1. La vérification et Identification du locuteur

La reconnaissance automatique du locuteur (Reconnaissance par la voix) fait toujours l'objet de travaux de recherches entrepris par de nombreuses équipes de recherches dans le monde, elle s'est limitée longtemps à la détection ou la vérification de l'identité d'une personne à partir d'un échantillon de sa voix, la vérification consiste à accepter ou refuser l'identité proclamée par un locuteur, en se basant sur un modèle qui lui est associé. Aussi bien la vérification, du locuteur se fait en calculant un modèle stochastique sur la base de l'expression vocale du locuteur à reconnaitre. Une fois calculé, ce modèle est comparé à des modèles pré entraînés sur la base de différents enregistrements prononcés par les locuteurs. [2]

D'une année à une autre le champ d'application des techniques de reconnaissance automatique de locuteur, s'est considérablement élargi suite au progrès à la fois des algorithmes utilisés, la puissance de traitement disponible, et l'évolution remarquable des technologies utilisées.

En reconnaissance du locuteur, on fait la différence entre l'***identification*** et la ***vérification*** du locuteur, selon que le problème est de vérifier que la voix analysée correspond bien à la personne qui est sensée la produire, ou qu'il s'agit de déterminer qui, parmi un nombre fini et préétabli de locuteurs, a produit le signal analysé. On sépare reconnaissance du locuteur ***dépendante du texte***, reconnaissance ***avec texte dicté***, et reconnaissance ***indépendante du texte***. Dans le premier cas, la phrase à prononcer pour être reconnue est fixée dès la conception du système ; elle est fixée lors du test dans le deuxième cas, et n'est pas précisée dans le troisième.

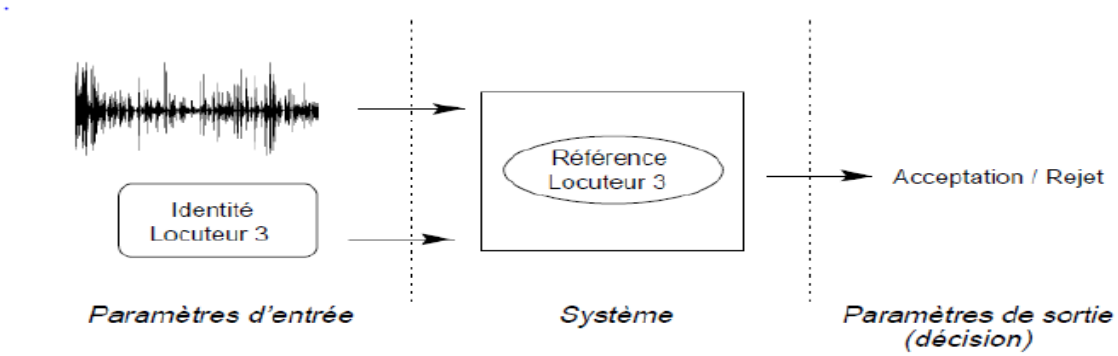


Figure 1-1 La vérification du locuteur

2.2. Le système est-il robuste

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

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

Certains systèmes peuvent être plus robustes que d'autres à l'une ou l'autre de ces perturbations, mais en règle générale, les systèmes de reconnaissance de locuteur sont encore sensibles à ces perturbations.

2.3. Fonctionnement

Le problème de la reconnaissance vocale 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.

2.4. Exemple de Reconnaissance

L'idée très simple dans son principe, consiste à faire prononcer un mot par plusieurs personnes et puis capturer ses voix, et les enregistrer sous forme de vecteurs acoustiques (représentation numérique du signal sonore). Puisque cette suite de vecteurs acoustiques caractérise complètement l'évolution de l'enveloppe spectrale du signal enregistré, on peut dire qui correspond à un enregistrement d'un spectrogramme. L'étape de reconnaissance proprement dite 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 «reconnue» sera alors celui 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.

Ce principe de base n'est cependant pas implémenté directement :

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. La voix dans un système de RAL (Reconnaissance Automatique du Locuteur)

La voix est porteuse d'informations variées. Emission de sons structurée, la parole humaine 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 parole peut cependant contenir de nombreuses informations telles que la langue parlée par le locuteur, son identité etc.

Les systèmes de reconnaissance de locuteur, cherchent à extraire du signal acoustique une information, à priori, indépendante de l'état présent du locuteur.

De par sa complexité, l'information portée par le signal de parole ne peut, quelle que soit la tache considérée, être utilisée dans sa totalité. C'est pourquoi l'utilisation d'un système automatique nécessite une sélection préalable d'informations à exploiter pour la tâche qui lui est confiée. En reconnaissance automatique du locuteur, seules les informations présentant une forte variabilité entre interlocuteurs permettent de discriminer les différents individus. A l'inverse, les informations dont la variabilité intra-locuteur est élevée rendent la tâche du RAL plus complexe.

Les informations les plus utilisées en RAL, du fait de leur fort potentiel discriminant, sont des informations acoustiques obtenues périodiquement par une analyse fréquentielle ou temporelle du signal. D'autres informations présentées dans le signal de parole peuvent s'avérer discriminantes dans le cadre de la reconnaissance du locuteur. Des paramètres tels que la prosodie ou la fréquence fondamentale, par exemple, contiennent une information spécifique au locuteur.

Les systèmes de reconnaissance du locuteur utilisent des représentations du signal de parole dans lesquelles le bruit et la redondance ont été réduits afin de ne conserve que les informations considérées comme utiles à la tache spécifiée.

3.1. Pourquoi l'authentification vocale ?

Le niveau de sécurité d'un système est toujours celui du maillon le plus faible. Ce maillon faible, c'est bien souvent l'être humain : mot de passe aisément déchiffrable ou note à cote de l'ordinateur. Dans la plupart des entreprises, on exige que les mots de passe soient modifiés régulièrement et comportent au moins 8 caractères, mélangeant lettres majuscules, minuscules et chiffres. L'objectif est d'échapper aux logiciels de décodage qui peuvent en peu de temps, balayer tous les mots du dictionnaire. Une protection qui peut s'avérer insuffisante pour l'accès à des applications sensibles.

3.1.1. Les applications de l'authentification vocale

Le champ d'application de l'authentification couvre potentiellement 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, le contrôle des frontières, l'accès aux réseaux, systèmes d'information, stations de travail et PC, 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, citons quelque exemples :

- Contrôle d'accès physiques aux locaux

« Salle informatique, Site sensible (service de recherche, site nucléaire) »

- Contrôle d'accès logiques aux systèmes d'informations

« Lancement du système d'exploitation Accès au réseau informatique Commerce électronique, paiement en ligne Transaction financière pour les banques»

- Equipements de communication

« Terminaux d'accès à Internet, Téléphones portables »

- Machines & Equipements divers « Coffre-fort avec serrure électronique, Distributeur automatique de billets, Cantine d'entreprise, cantine scolaire (pour éviter l'utilisation d'un badge par une personne extérieure et améliorer la gestion, Contrôle des temps de présence, Voiture (anti-démarrage) »
- Etat / Administration « Fichier judiciaire »

Nos voix ne sont pas seulement un moyen de communiquer. Elles offrent également un moyen fiable de nous reconnaître, et font partie intégrante de notre identité. C'est la raison pour laquelle les banques et d'autres grandes entreprises se tournent aujourd'hui vers l'authentification vocale.

La voix humaine est unique. Elle est avec nous tout le temps 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.

d) Aide aux handicapés

La reconnaissance de locuteur est très utile dans ce cas, elle offre la possibilité de saisir le données à la voix, commandes vocales (ouverture porte, contrôle des équipements au domicile).

Du coup, de nombreuses pistes sont explorées dans l'espoir de trouver la parade infaillible, susceptible de fournir la sécurité totale des échanges, inspirant à quelques observateurs l'idée que, bientôt, les mots de passe seront rangés au placard des souvenirs. Le recours à la reconnaissance vocale constitue une nouvelle trouvaille le fait de faire des achats sur Internet, valider des transactions bancaires, consulter son compte client... et tout cela avec le simple son de sa voix. C'est n'est pas de la science-fiction, mais une réalité à laquelle chacun d'entre nous sera confronté prochainement.

4. Les défis et les problèmes

Le principal défi dans la technologie de reconnaissance du locuteur a donc été d'améliorer la robustesse des systèmes dans des conditions incompatibles. La variation intra-locuteur du style de parler, les variations de l'environnement acoustique.

Notre système vocal fournit principalement les indices acoustiques pour la classification des phonèmes, et aussi la personnalité individuelle pour caractériser le locuteur. La variation interlocuteur pourrait être importante, même pour le même contenu parole. Un système de reconnaissance du locuteur essaye de comprendre ces variations interlocuteur sur lesquelles est basée la discrimination d'un locuteur par rapport aux autres. Dans le même temps, le système vocal produit un certain degré de variation intra-locuteur pour le même contenu parole réitère a différents moments. La plupart des erreurs de reconnaissance sont causées par ces types de variations intra locuteur.

D'autre part, la variation de l'environnement acoustique est causée par les diverses distorsions imprévisibles lors de la collecte des données et la transmission. Par exemple, dans les applications de reconnaissance du locuteur par téléphonie (par exemple, des transactions bancaires par téléphone), les données vocales pourraient être recueillies dans des environnements avec un bruit de fond différent, avec différents téléphones, et via différents canaux. Le bruit de fond et la combinaison combiné/distorsion du canal change la structure spectrale des données paroles et les paramètres acoustiques dérivant.

5. Panorama sur les plateformes existantes

En premier lieu on a posé la question comment peut-on faire un appel distant et que la machine réceptrice ou l'ordinateur récepteur enregistre ce qu'on a dit ?

En deuxième lieu, une fois la machine réceptrice, ou l'ordinateur récepteur a enregistré ce qu'on a dit comment peut-il le reconnaitre ?

Pour répondre à ces questions et pour réaliser un tel projet on a fait bien évidement une recherche bibliographique sur les technologies et les outils existant dans le cadre de la reconnaissance, et le résultat était étonnant, vu la profusion des logiciels et les travaux intégrant la ToIP. Nous en résumons le tout dans ce qui suit :

Les outils de la reconnaissance vocale les plus célèbres et les plus utilisés jusqu'à nos jours sont le HTK et le Sphinx.

Le HTK en anglais (The Hidden Markov Model Toolkit) est un outil de la construction des modèles cachés de Markov, dont les premières utilisations ont été pour la reconnaissance vocale, bien que maintenant cet outil soit utilisé dans d'autres domaines tel que la recherche sur la synthèse vocale, la reconnaissance des caractères et le séquençage de l'ADN. Le HTK est un ensemble de bibliothèques écrites dans le langage C, et qui contient plein de documentation et d'exemples.

Sphinx est un projet lancé par l'université Carnegie Mellon (CMU de Pittsburgh (Pennsylvanie)) dans le but est de concevoir un environnement pour la recherche dans le domaine de la reconnaissance vocale. CMU Sphinx 4 est une librairie de classes (en langage java). Sphinx est un système basé sur les Modèles Cachés de Markov (HMM) constitué d'un ensemble d'outils de reconnaissance vocale flexibles modulaires et extensibles formant un véritable banc d'essais et un puissant environnement de recherche pour les technologies de reconnaissance vocale. Comme notre environnement de développement choisi est le langage de programmation java et le système d'exploitation Windows, on a préféré réaliser notre propre application, en réutilisant d'autres bibliothèques qu'on a trouvées dans l'état de l'art. Pour se faire, on va introduire certaines notions essentielles à la compréhension de la suite du PFE.

6. La Perception du son

6.1. Qu'est-ce que le son ? [3]

C'est un phénomène moins évident à saisir que les images, il est nécessaire de rappeler quelques notions de base. Le son est une vibration de l'air. A l'origine de tout son, il y a mouvement (par exemple une corde qui vibre, une membrane de haut-parleur...). Il s'agit de phénomènes oscillatoires créés par une source sonore qui met en mouvement les molécules de l'air. Avant d'arriver jusqu'à notre oreille, ce mouvement se transmet entre les molécules à une vitesse de 331 m/s à travers l'air à une température de 20°C : c'est ce que l'on appelle la propagation.

Un son est d'abord défini par son volume sonore et sa hauteur tonale. Le volume dépend de la pression acoustique créée par la source sonore .Plus elle est importante plus le volume est élevé. La hauteur tonale est définie par les vibrations de l'objet créant le son. Plus la fréquence est

élevée, plus la longueur d'onde est petite et plus le son perçu est aigu. Chacun peut constater que le niveau sonore diminue à mesure que l'on s'éloigne de la source. Cette diminution est la même chaque fois que la distance est doublée. Cependant, les hautes fréquences ne se propagent pas aussi loin que les sons graves. Il faut plus d'énergie pour restituer les basses que les aigus.

6.2. Traitement du signal vocal [4]

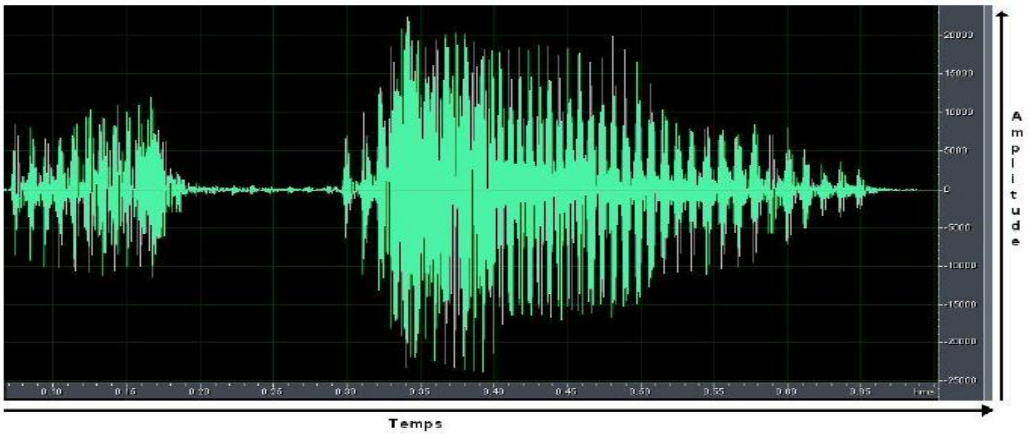


Figure 1-2 Représentation d'un signal sonore

a) Intensité d'un son *[5]*

L'intensité d'un son, appelée aussi volume, permet de distinguer un son fort d'un son faible. Elle correspond à l'amplitude de l'onde. L'amplitude est donnée par l'écart maximal de la grandeur qui caractérise l'onde. Pour le son, onde de compression, cette grandeur est la pression. L'amplitude sera donc donnée par l'écart entre la pression la plus forte et la plus faible exercée par l'onde acoustique. Lorsque l'amplitude de l'onde est grande, l'intensité est grande et donc le son est plus fort L'intensité du son se mesure en décibels (dB). On distingue différentes façons de mesurer l'amplitude d'un son :

- La puissance acoustique : La puissance acoustique est associée à une notion physique. Il s'agit de l'énergie transportée par l'onde sonore par unité de temps et de surface. Elle s'exprime en Watt par mètre carré ($W.m$-2).
- Addition de sons : L'échelle des décibels est une échelle dite logarithmique, ce qui signifie qu'un doublement de la pression sonore implique une augmentation de l'indice d'environ 3 : avec 3 dB de plus, l'intensité est en fait doublée.

b) 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.

c) Le timbre

Le timbre est l'ensemble des caractéristiques qui permettent de différencier une voix.

Il provient en particulier de la résonance dans la poitrine, la gorge la cavité buccale et le nez ; ceux sont les amplitudes relatives des harmoniques du fondamental qui déterminent le timbre du son. Les éléments physiques du timbre comprennent :

- la répartition des fréquences dans le spectre sonore,
- les relations entre les parties du spectre, harmoniques ou non,
- les bruits existant dans le son (qui n'ont pas de fréquence particulière, mais dont l'énergie est limitée à une ou plusieurs bandes de fréquence), l'évolution dynamique globale du son,
- l'évolution dynamique de chacun des éléments les uns par rapport aux autres.

d) Les composantes fondamentales du son

La durée : Représente l'étalement du son dans le temps. (long/bref). La durée est étroitement liée au rythme.

La hauteur : Représente une sensation auditive plus ou moins aiguë.

La densité : C'est la quantité d'éléments contenus dans un son.

Le contraste : créé par la juxtaposition d'intensités, de hauteurs, de timbres… différents.

Le mouvement mélodique : C'est la direction auditive que prend la mélodie : elle monte, elle descend, elle reste à la même hauteur, elle fait des vagues…

Le tempo (ou mouvement) : il peut être vif, rapide, médium ou lent. C'est la vitesse avec laquelle s'enchaînent les éléments sonores.

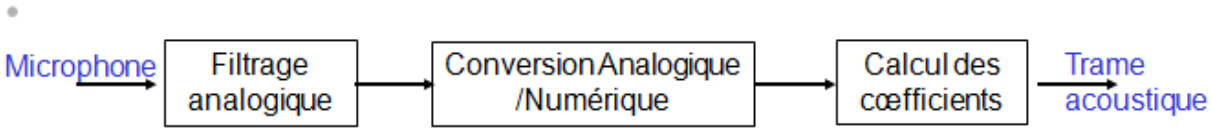


Figure 1-3 Numérisation d'un son analogique

7. La classification du son [6]

Il faut d'abord différencier les deux types de sons : le son analogique et le son numérique.

Le son analogique 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.

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).

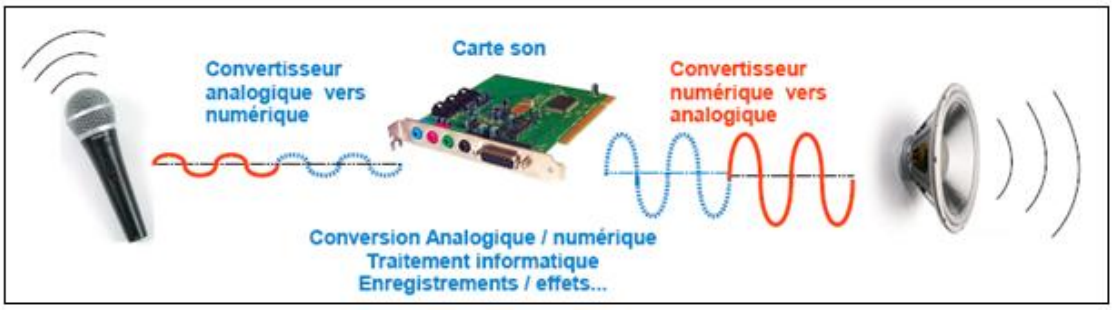


Figure 1-4 Exemple d'une chaine numérique

8. Numérisation du son

8.1. L'échantillonnage

L'échantillonnage consiste à transformer une fonction $a(t)$ à valeurs continues en une fonction $\hat{a}(t)$ discrète constituée par la suite des valeurs $a(t)$ aux instants d'échantillonnage **t = kT** avec k un entier naturel. Le choix de la fréquence d'échantillonnage n'est pas aléatoire car une petite fréquence nous donne une présentation pauvre du signal. Par contre une très grande fréquence nous donne des mêmes valeurs, redondance, de certains échantillons voisins donc il faut prélever suffisamment de valeurs pour ne pas perdre l'information contenue dans $a(t)$. Le théorème suivant traite cette problématique : *La fréquence d'échantillonnage assurant un non repliement du spectre doit être supérieure à 2 fois la fréquence haute du spectre du signal analogique*. $\boldsymbol{Fech = 2 \times Fmax}$ Par contre pour le signal audio (parole), on exige une bonne représentation du signal jusqu'à 20 kHz et l'on utilise des fréquences d'échantillonnage de 44.1 ou 48 kHz. Pour les applications multimédia, les fréquences sous-multiples de 44.1 kHz sont de plus en plus utilisées : 22.5 kHz, 11.25 kHz

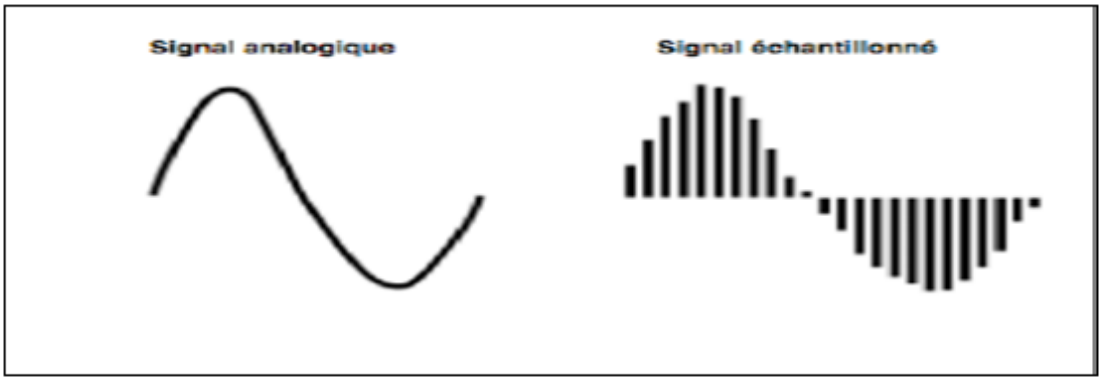


Figure 1-5 Échantillonnage d'un signal audio

8.2. La Quantification

Alors que l'échantillonnage opère un découpage temporel, l'opération de quantification crée une échelle de valeurs discrètes permettant d'attribuer à chaque échantillon une valeur d'amplitude. La quantification s'exprime en « bit » (un acronyme de binary digit). Les valeurs couramment utilisées en audio sont 16bit et 24bit.

L'amplitude de chaque échantillon doit impérativement prendre l'une des valeurs définies par l'échelle de quantification. Si la valeur d'amplitude de l'échantillon se situe entre deux paliers de l'échelle de quantification, elle est approximée au palier le plus proche. Cette approximation induit une erreur que l'on nomme « erreur de quantification ».

Par suite, plus le nombre de bits est élevé, plus le nombre de paliers est important et l'erreur de quantification faible. Autrement dit, les petites variations d'amplitude du signal échantillonné sont d'autant mieux approximées que la résolution de la quantification est élevée. La fidélité de la forme d'onde numérisée à la forme d'onde du signal analogique dépend donc de la résolution (exprimée en bit) et de la fréquence d'échantillonnage (exprimée en kHz).

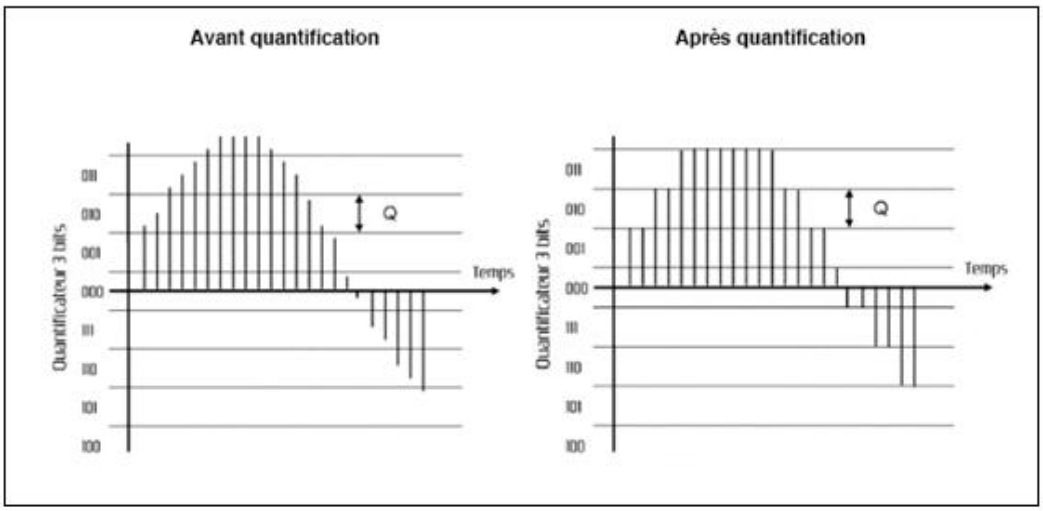


Figure 1-6 Signal échantillonné avant et après quantification

8.3. Codage

C’est la représentation binaire des valeurs quantifiées qui permet le traitement du signal sur machine.

9. Qu’est-ce qu’un fichier audio numérique

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.

10. Conclusion

Dans ce chapitre on a fait une étude, qui suit le son depuis sa production jusqu'à son analyse, dans le but de générer des vecteurs acceptant l'application de la méthode MFCC pour la reconnaissance du locuteur.

Les difficultés rencontrées dans cette partie commencent par l'acquisition du son, et sa transformation en représentation temporelle à la représentation fréquentielle, et comme ces deux représentations ne sont pas les plus adéquates pour la reconnaissance du locuteur il était nécessaire de passer au traitement du signal. Et sans oublier que la voix est passée de locuteur vers le pc via un microphone qui nous oblige de veiller à la crédibilité du fichier reçu en prenant en considération toutes formes de distorsions. Comme cette partie est la base de notre travail ses résultats seront entièrement utilisés dans la partie suivante, ou nous introduisons les techniques de comparaison entre fichiers audio dans le but de retrouver la bonne correspondance.

Chapitre II : Etat de l'art

1. Introduction

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, poussent les chercheurs à reconsidérer les préjuges connus jusque-là et essayer de tirer le maximum de ces informations complémentaires pour améliorer les performances du système de reconnaissance du locuteur. Pour atteindre cet objectif, nous essayons de répondre aux questions suivantes : Comment représenter efficacement les informations spécifiques au locuteur à partir du signal vocal ? Est-il vraiment utile de prendre en compte l'information de la source vocale pour la reconnaissance du locuteur ? Nous avons introduit dans ce chapitre aussi la technique que nous avons utilisée dans l'implémentation de notre solution.

2. Le signal

Le signal de parole est provoqué par des mécanismes complexes issus de plusieurs sources. Les sons de la parole se produisent normalement lors de la phase de l'expiration grâce à un flux d'air contrôlé, en provenance des poumons et passant par la trachée-artère (=conduit respiratoire). Ce flux d'air s'appelle « air pulmonaire (ou pulmonique) égressif ». Il va rencontrer sur son passage plusieurs obstacles potentiels qui vont le modifier de manière plus ou moins importante.

Le signal de la parole véhicule plusieurs types d'informations, tels que le fondamental, la prosodie, le timbre et les phonèmes. Par conséquent, ceci impose, aux systèmes de reconnaissance vocale, de n'extraire que l'information nécessaire à son application, les phonèmes pour les machines de dictée par exemple.

En général, on considère que la plage de fréquence d'un signal de parole se situe dans la bande de 100Hz-5KHz. Le signal de la parole est un phénomène de nature acoustique porteur d'un message. L'information du message parlé réside dans les fluctuations de l'air engendrées puis émises par l'appareil phonatoire. Ces fluctuations constituent le signal vocal, elles sont détectées par l'oreille, qui procède à une certaine analyse. Les résultats sont ensuite transmis au cerveau qui les interprète.

A l'image de ce processus naturel, le processus de reconnaissance de Locuteur inclut deux grandes phases : dans la première phase on s'attache à extraire du signal continu de la parole un certain nombre de paramètres qui le caractérisent. Ainsi, étant donné un signal en entrée du système, celui-ci va subir un prétraitement qui consiste généralement en un filtrage et un échantillonnage qui permet de passer d'un signal continu à des valeurs discrètes, de ces valeurs dont le nombre est important seront extraites des caractéristiques qui permettent de représenter de façon compacte et pertinente le signal originel.

3. L'information vocale

3.1. L'analyse du signal [7]

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.

3.2. 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.

3.3 La modélisation des paramètres acoustiques

- Dans un système de reconnaissance automatique du locuteur, les paramètres
- acoustiques sont utilisées 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.

3.4 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 RAL, 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 RAL. 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***

Les principaux paramètres de l'analyse spectrale utilisés en RAL 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. [8]

4 Pourquoi l'échelle de Mel

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. [9]

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 2-1 Formule de fréquence en Mels

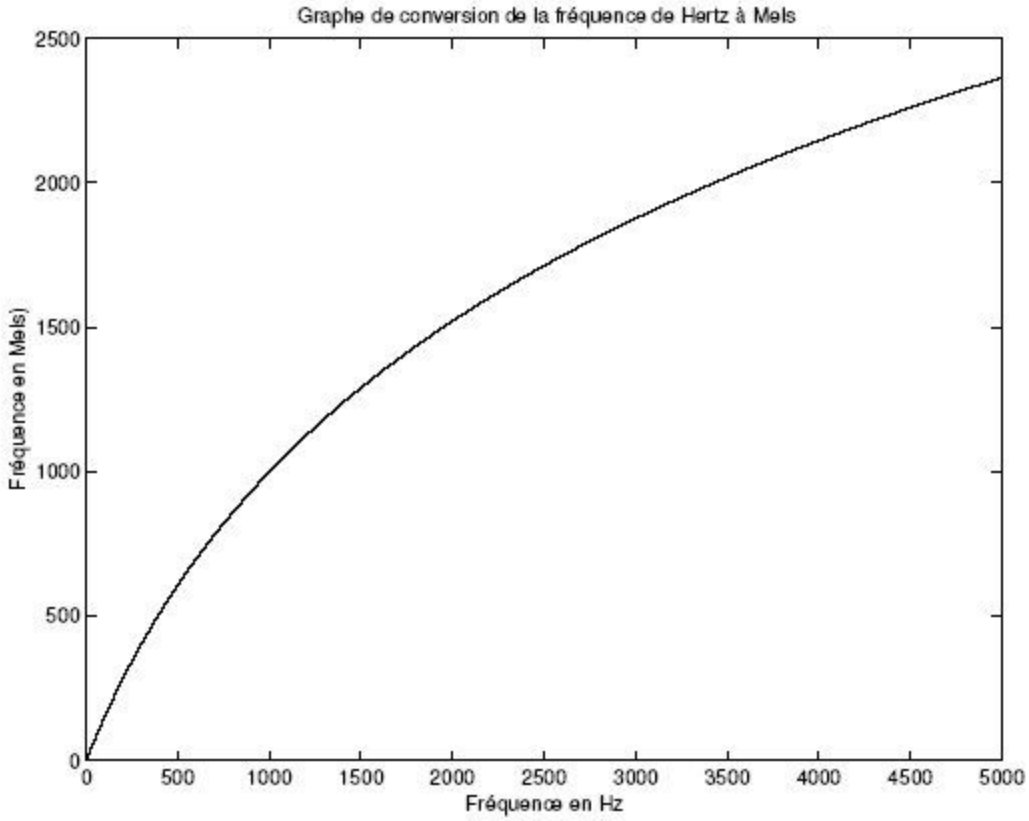


Figure 2-2 Exemple de conversion des hertz en mels.

L'objectif de cette phase est d'extraire des coefficients représentatifs du signale 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-Scale Frequency Ceptral 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 (Perceptual Linear Predective).

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.

5.Étapes de calcul du vecteur caractéristique de type MFCC

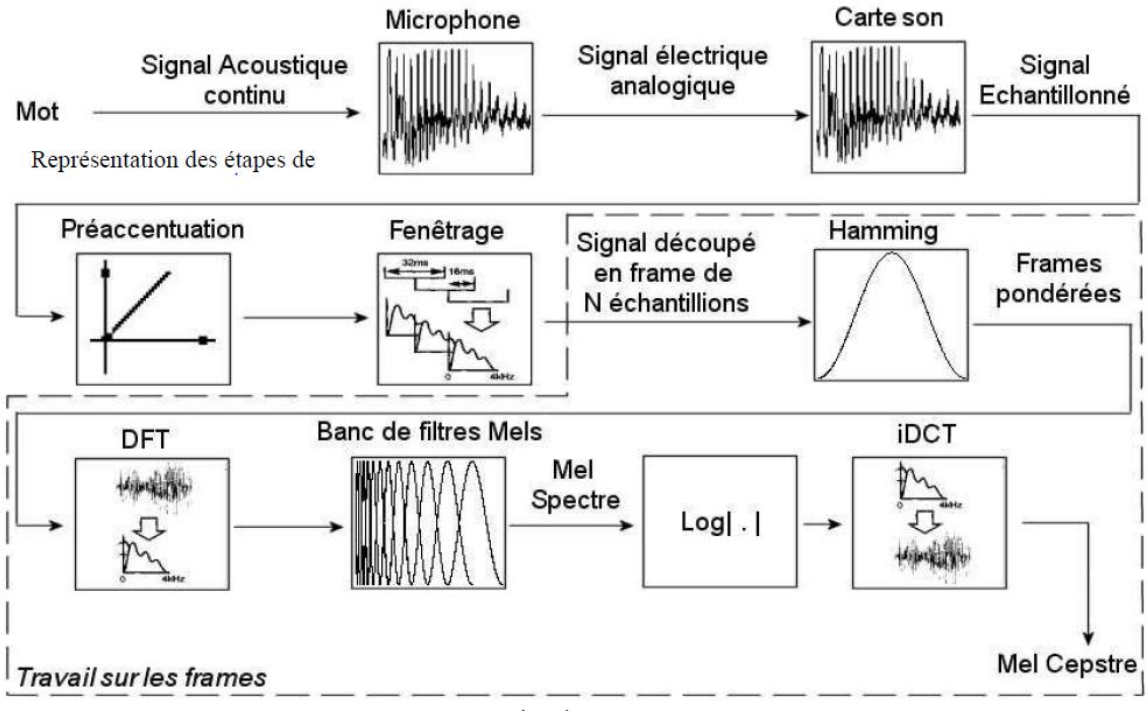


Figure 2-3 Étapes de calcul d'un vecteur caractéristique de type MFCC

1.1. Groupement en trames (Frame blocking)

Le signal acoustique continu est segmenté en trames de N échantillons, avec un pas d'avancement de M trames ($M < N$), c'est-à-dire que deux trames consécutives se chevauchent sur $N - M$ échantillons. Les valeurs couramment utilisées pour M et N sont respectivement 10 et 20. Comme prétraitement, il est d'usage de procéder à la préaccentuation du signal en appliquant l'équation de différence du premier ordre aux échantillons $x(n)$, avec l'équation $x0(n) = x(n) - kx\,(n-1)$, $0 < n < N - 1k$ représente un coefficient de préaccentuation qui peut prendre une valeur dans l'étendue $0 < k < 1$.

1.2. Fenêtrage

Si nous définissons $w(n)$ comme fenêtre où $0 < n < N - 1$ et N représente le nombre d'échantillons dans chacune des trames, alors le résultat du fenêtrage est le signal xa, donné par la formule $xa = x(n)\, w(n)$, $0 < n < N - 1$ .Les fenêtres les plus utilisées « Fenêtre de Hamming »

1.3. Calcul de la transformée de Fourier rapide (Fast Fourier Transform, FFT)

Au cours de cette étape chacune des trames, de N valeurs, est convertie du domaine temporel au domaine fréquentiel. La FFT est un algorithme rapide pour le calcul de la transformée de Fourier discret (DFT) et est définie comme suit, Les valeurs obtenues sont appelées le spectre.

$$x[k] = \sum_{n=0}^{N-1} x_a[n] e^{\frac{-2j\pi}{N} kn}, \qquad 0 \leq k \leq N-1$$

En général, les valeurs $X[k]$ sont des nombres complexes et nous nous utilisons que leurs valeurs absolues (énergie de la fréquence).

1.4. Filtrage sur l'échelle Mel

Le spectre d'amplitude est pondéré par un banc de M filtres triangulaires espacés selon l'échelle Mel. Dans l'échelle de mesure Mel, la correspondance est approximativement linéaire sur les fréquences au-dessous de 1kHz et logarithmique sur les fréquences supérieures à celle-ci. Cette relation est donnée par la formule :

$$m = 2595 \log_{10}(1 + \frac{f}{700})$$

Le logarithme de l'énergie de chaque filtre est calculé selon l'équation:

$$S[m] = \ln[\sum_{k=0}^{N-1} X_a[k] H_m[k]], \qquad 0 < m \leq M$$

1.5. Calcul du cepstre sur l'échelle Mel

Le cepstre sur l'échelle de fréquence Mel est obtenu par le calcul de la transformée en cosinus discrète du logarithme de la sortie des M filtres (reconversion du log-Mel-spectre vers le domaine temporel).

$$c[n] = \sum S[n] \cos \pi n(m - \frac{1}{2})/M, \qquad 0 \leq n < M$$

Le premier coefficient, c [0], représente l'énergie moyenne dans la trame de la parole c [1] reflète la balance d'énergie entre les basses et hautes fréquences ; pour i > 1, c[i] représente des détails spectraux de plus en plus fins.

1.6. Calcul des caractéristiques dynamiques des MFCC

Les changements temporels dans le cepstre (*c*) jouent un rôle important dans la perception humaine et c'est à travers les dérivées des coefficients des MFFC statiques que nous pouvons mesurer ces changements. En résumé, un système de parole typique de l'état de l'art effectue premièrement un échantillonnage à une fréquence de 16 kHz et extrait les traits suivants :

$$\begin{pmatrix} c_k \\ \triangle c_k \\ \triangle \triangle c_k \end{pmatrix}$$

O*u* :

– *ck* est le vecteur MFCC de la kièmetrame

– 4*ck* = *ck*+2 − 4*ck*−2, dérivée première des MFCCs calculée à partir des vecteurs

MFCC de la kième + 2 trames et kième - 2

– 44 $ck = 4ck-1 -4ck+1$, seconde dérivée des MFCCs.

Pour terminer, reste enfin la dernière étape qui consiste à comparer deux cepstres MFCC, pour cela on utilise l'algorithme DTW (Dynamic time Wrapping). Comme dit précédemment, il est impossible de comparer directement deux spectres (ou cepstres) entre eux, tout simplement parce qu'une même personne ne peut prononcer deux fois le même mot sur la même durée, le même rythme, la même intensité. Il est donc nécessaire de développer une méthode de comparaison cepstre à cepstre par l'algorithme de comparaison dynamique détaillé ci-après.

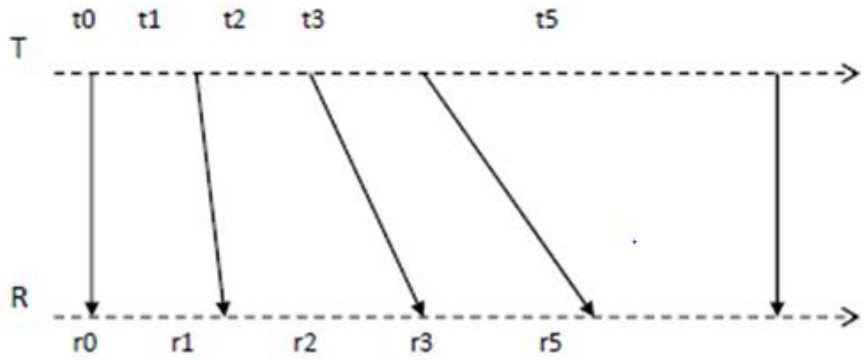
2. L'algorithme DTW

La **déformation temporelle dynamique** (algorithme **DTW** pour **Dynamic Time Warping** en anglais) est un algorithme permettant de mesurer la similarité entre deux suites qui peuvent varier au cours du temps. Par exemple des similarités entre des pas dans des vidéos peuvent être détectées même si dans l'une ou l'autre des vidéos le sujet a marché plus rapidement ou plus lentement, ou encore si au cours de l'une ou l'autre le sujet a accéléré ou ralenti.

L'algorithme DTW a été exploité en vidéo, audio, graphique par ordinateur, bio-informatique,... et peut être appliqué dans toute situation où les données peuvent être transformées en une représentation linéaire. Une application célèbre est l'application en reconnaissance de locuteur, où il est nécessaire de tenir compte de vitesses de locution très variables.

De façon générale, DTW est une méthode qui recherche un appariement optimal entre deux séries temporelles, sous certaines restrictions. Les séries temporelles sont déformées par transformation non-linéaire de la variable temporelle, pour déterminer une mesure de leur similarité, indépendamment de certaines transformations non-linéaires du temps.

2.1. Distance DTW [10] [11]



L'alignement temporel, plus connu sous l'acronyme de DTW, *Dynamic Time Warping*, est une méthode fondée sur un principe de comparaison d'un signal à analyser avec un ensemble de signaux stockés dans une base de référence. Le signal à analyser est comparé avec chacune des références et est classé en fonction de sa proximité avec une des références stockées. Le DTW est en fait une application au domaine de la reconnaissance de la parole de la méthode plus générale de la programmation dynamique. Elle peut ainsi être vue comme un problème de cheminement dans un graphe.

Ce type de méthode pose deux problèmes : la taille de la base de référence, qui doit être importante, et la fonction de calcul des distances, qui doit être choisie avec soin (Hamming dans notre cas). La taille de la base contenant les signaux de référence est directement liée aux capacités, variables, de reconnaissance du système d'alignement temporel

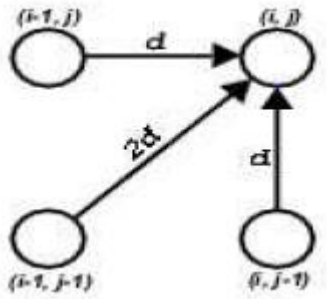
Soient *A* et *B* deux images acoustiques (vecteurs MFCC) de longueur *I* et *J* respectivement On crée un chemin $\{C(K) = (n(k), m(k)), "k \hat{I}[1,K]\}$ et il est nécessaire que les fonctions $n(k)$ et $m(k)$ soient croissantes et doivent correspondre à certaines contraintes : les seules chemins valides arrivants au point $(i\ j)$ sont ceux provenant des points $(i\text{-}1, j)$, $(i, j\text{-}1)$et $(i\text{-}1, j\text{-}1)$.

De plus on prend *K* tel-que $C(K) = (I, J)$. On pose $C(1) = (1,1)$.

La méthode consiste à choisir le chemin qui passe par les distances $d(i, j)$ les plus petites, de sorte que la distance cumulée le long de ce chemin soit la plus petite possible. On définit $g(i, j)$ la distance cumulée au point (i, j) comme :

$$g(i,j) = \min\begin{Bmatrix} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2.d(i,j) \\ g(i,j-1) + d(i,j) \end{Bmatrix}$$

On remplit ensuite la matrice *I* x *J* (le plan du chemin) avec en *i ème* et *j ème* colonnes le résultat de *g(i, j)*.



Enfin on définit la distance normalisée entre deux prononciations du mot :

$$G = \frac{g(I,J)}{I+J}$$

On obtient une distance entre deux spectres. On effectue ce travail entre le mot à reconnaître et tous les mots stockés auparavant. On prend ensuite le mot de la base de données qui a la plus petite distance spectrale avec le mot à reconnaître

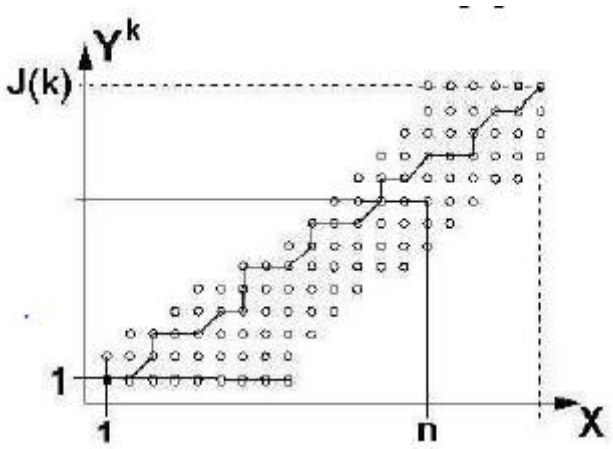


Figure 2-4 Représentation de la notion de chemin entre deux spectres.

Les différences entre les deux spectres « tordent ». Le chemin idéal est la diagonale.

3. Conclusion

L'algorithme DTW est un très bon outil capable de comparer deux spectres audio ayant des durées différentes, un débit, une intensité de la voix différente et cela de façon optimale en recherchant le meilleur chemin pour passer d'un spectre à l'autre. Néanmoins d'autres méthodes existent, comme les modèles de Markov cachés (HMM) par exemple bien plus puissant que l'algorithme DTW mais bien plus complexe. Nous avons donc vu ici une approche globale de la reconnaissance de la parole.

Néanmoins cette méthode peut être parfaitement utilisée dans des appareils d'utilisation courante comme les téléphone portables (utilisation type appel numérotation automatique), les consoles automobiles ou pourquoi pas dans le domaine de la domotique et même pour la commande vocale pour les handicapés.

Il existe des implantations très poussées des MFCC, comme par exemple le système Sphinx. Mais la méthode de comparaison des coefficients MFCC utilisée est statique (modèle de Markov Cachés et aussi approche Neuronale. 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.

Chapitre III : La Réalisation

1. Introduction

L'étape de développement est une étape essentielle dans notre projet, notre but était de réaliser une application ou un système robuste, protégé et sécurisé, avec la notion de l'authentification vocal, dans ce chapitre on va citer les différents 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 [12]

Java est un langage de programmation et une plate-forme informatique qui ont été créés par Sun Microsystems en 1995. Beaucoup d'applications et de sites Web ne fonctionnent pas si Java n'est pas installé et leur nombre ne cesse de croître chaque jour. Java est rapide, sécurisé et fiable. Des ordinateurs portables aux centres de données, des consoles de jeux aux superordinateurs scientifiques, des téléphones portables à Internet, la technologie Java est présente sur tous les fronts.

2.2. Le Framework JSF [13]

Java Server Faces (abrégé en JSF) est un Framework Java, pour le développement d'applications Web. À l'inverse des autres Framework MVC traditionnels à base d'actions, JSF est basé sur la notion de composants, comparable à celle de Swing ou SWT, où l'état d'un composant est enregistré lors du rendu de la page, pour être ensuite restauré au retour de la requête. JSF est agnostique à la technologie de présentation. Il utilise Facelets par défaut depuis la version 2.0, mais peut être utilisé avec d'autres technologies, comme JSP (qui était utilisé jusqu'à la version 1.2) ou XUL.

2.3. L'implémentation PRIMEFACES [14]

PrimeFaces pour JSF2 est une bibliothèque de composants, pour les applications de bureau et les mobiles, Pour notre projet on a utilisé la version 5.1de PrimeFaces.

Figure 3-1 Logo PrimeFaces

2.4. L'IDE Netbeans [15]

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.

2.5. Base de Données MySql [16]

Nous avons géré une boîte de location de matériel audiovisuel, et afin de toujours **savoir où nous en sommes dans notre stock**, nous voudrions un système informatique nous permettant de **gérer les entrées et sorties de matériel**, mais aussi éventuellement **les données de nos clients**. MySQL est une des solutions possibles pour gérer tout ça.

Généralement pour tout projet ou nous devons manipuler plusieurs clients **avec un espace membre, un forum, un système de news ou même un simple livre d'or**. Une base de données nous sera presque indispensable.

Dans notre cas, une base de données de voix préenregistrées pour éventuellement faire une reconnaissance vocale en vue d'authentifier les accès au réseau nous est nécessaire.

2.6. Serveur d'application (Tomcat) [17]

Apache Tomcat est un conteneur web libre de servlets et JSP Java EE. Issu du projet Jakarta, c'est un des nombreux projets de l'*Apache Software Foundation*. Il implémente les spécifications des servlets et des JSP du Java Community Process, est paramétrable par des fichiers XML et de propriétés, et inclut des outils pour la configuration et la gestion. Il comporte également un serveur HTTP.

3. Cycle de développement (Méthode en cascade)

Il existe différents types de **cycle de développement** entrant dans la réalisation d'un logiciel. Ces cycles prendront en compte toutes les étapes de la conception d'un logiciel.

Le modèle en cascade est hérité de l'industrie du BTP. Ce modèle repose sur les hypothèses suivantes :

- on ne peut pas construire la toiture avant les fondations ;
- les conséquences d'une modification en amont du cycle ont un impact majeur sur les coûts en aval (on peut imaginer la fabrication d'un moule dans l'industrie du plastique).

Les phases traditionnelles de développement sont effectuées simplement les unes après les autres, avec un retour sur les précédentes, voire au tout début du cycle. Le processus de développement utilisant un cycle en cascade exécute des phases qui ont pour caractéristiques :

- de produire des livrables définis au préalable ;
- de se terminer à une date précise ;
- de ne se terminer que lorsque les livrables sont jugés satisfaisants lors d'une étape de validation-vérification.

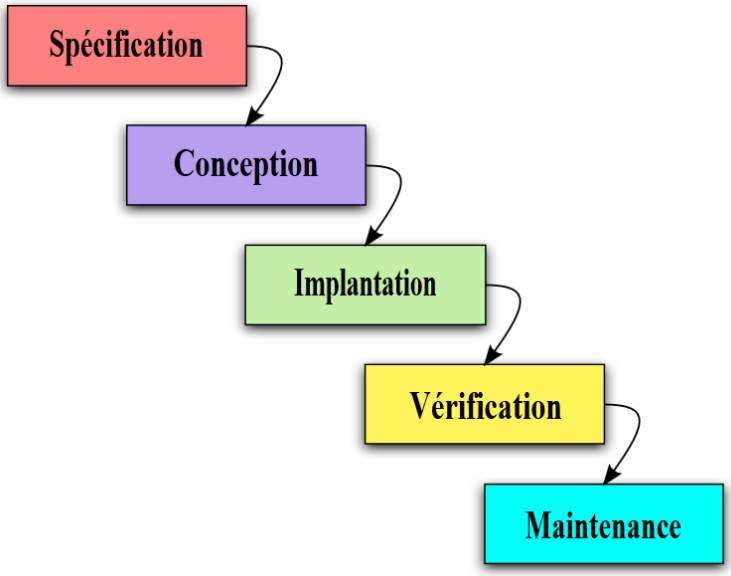


Figure 3-2 Méthode en cascade.

3.1. Étape de développement

3.1.1. 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 sécurisé basé sur l'authentification vocale,
- Les utilisateurs doivent avoir chacun un compte,
- Les utilisateurs doivent utiliser ces comptes pour stocker des informations secrètes,
- Le système doit s'adapter à n'importe quelle application.

Comme on à plusieurs utilisateurs qui doivent utiliser cette application via internet, on doit développer une application web.

3.1.2. 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.

3.1.3. Les taches de l'administrateur

3.1.4. Les taches des utilisateurs

L'utilisateur peut faire les tâches suivantes :

- Avoir un compte sécurisé par une voix déjà enregistré au moment de sa création,

Le compte est considéré comme une base de données pour enregistrer des infos L'administrateur doit faire les tâches suivantes :

- l'accès à son compte,
- créer des comptes pour les utilisateurs de système,
- le contrôle d'accès aux comptes, c'est-à-dire la vérification des dates d'entrée et sortie,
- lecture des fichiers log
- confidentielles.

4. La Conception

La conception est une étape essentielle dans le développement d'une application, dans cette partie on va citer ces différentes étapes.

Comme nous l'avons déjà dit dans les parties précédentes, on a deux espaces, un espace pour l'administrateur et l'autre pour les utilisateurs, et l'utilisateur peut enregistre des messages, donc on peut distingue les table de base de données suivantes :

- Une table pour les utilisateurs,
- Une pour enregistrer les voix,
- Une pour enregistrer les messages des utilisateurs,
- Et l'autre pour enregistrer les dates d'entrée et de sortie des utilisateurs,

4.1. Les tables

Pour la **table de l'utilisateur**, elle doit contenir les informations suivant, un **ID** pour l'utilisateur, **le nom, le prénom, un email**, **un mot de passe**, et le mot de passe vocal.

Pour la **table des voix**, on va enregistrer jusqu'à **10 voix** pour chaque utilisateur pour les comparer après avec leur authentification ou leur voix au moment de l'authentification.

La table des messages doit contenir une **information pour l'utilisateur**, **un objectif** de message, **le contenu du message, la date de création** du message.

La table des entrée et sortie doit contenir une **information pour l'utilisateur**, la **date d'entrée** et une **date de sortie**.

Donc on peut distinguer le diagramme de classe suivant :

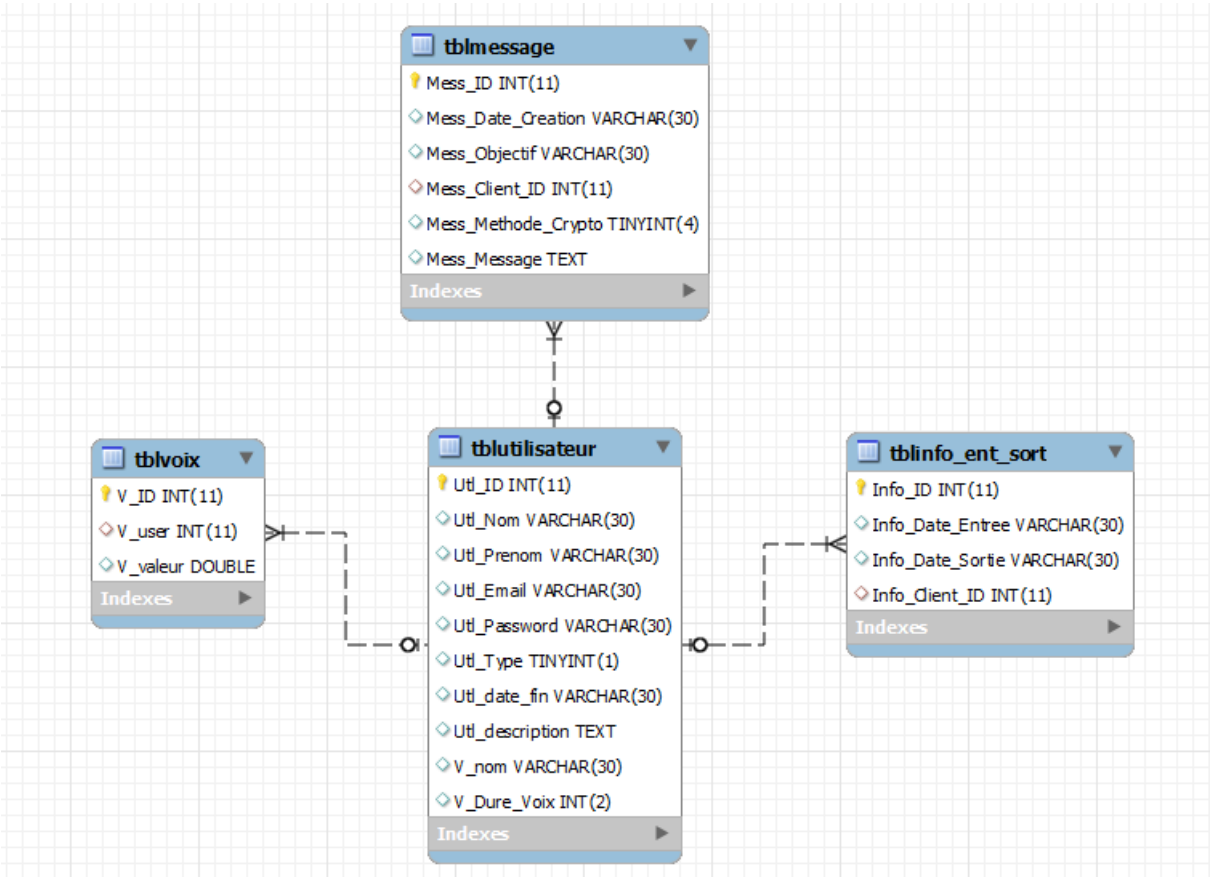


Figure 3-3 Diagramme de classe pour la base de données

5. Réalisation

Dans cette étape on a utilisé les algorithmes cités précédemment qui sont l'algorithme MFCC et DTW, et aussi on a utilisé la programmation JAVA, ou plutôt avec un Framework JAVA qui est PrimeFaces, et on a suivi les étapes ci-dessous dans la programmation :

Etape1 : Dessiner les pages et la forme de l'application avant la réalisation, pour mettre avoir une application souple et simple à utiliser.

Etape2 : créer la page principale c'est la page de LOGIN.

Etape3 : créer les Filtres.

Etape4 : créer les dossiers pour séparer les différents utilisateurs, par exemple un dossier pour l'administrateur, un autre pour les utilisateurs, …

Etape5 : créer la forme première des pages avec les composants de PrimeFaces.

Etape6 : créer des Beans pour chaque ensemble des données, par exemple un Bean pour les informations des utilisateurs, une autre pour les informations des voix, pour les messages, …

Etape7 : relier les pages avec les Beans.

Etape8 : vérifier les champs de chaque page.

Etape9 : relier les Beans avec la base de données.

5.1. Les étapes de comparaison entre deux voix

Les étapes de comparaison est comme suit :

1- Enregistrer les deux sons comme suit :

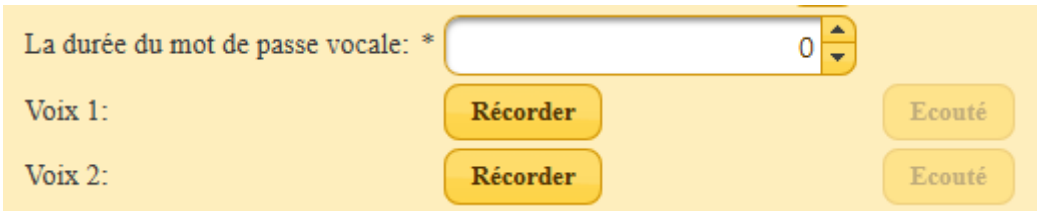


Figure 3-4 L'enregistrement des deux sons.

2- Récupérer une liste ou un fichier des fréquences à partir de ces deux voix,
3- Faire le para métrisation avec l'algorithme MFCC,
4- On va avoir deux listes de différents vecteurs,
5- Comparer ou plutôt calculer la distance entre ces deux listes,

Vue la difficulté de la comparaison entre les deux sons, nous avons choisi un repère pour valider le bon résultat, et à partir d'un ensemble de tests, si le résultat de cette distance est entre 0 et 2 donc on valide le test sinon, on donne la main à l'utilisateur pour recorder une autre fois.

5.2. Quelques pages de l'application

Figure 3-5 Page Login

Cette page de login est un espace unique pour l'administrateur et les autres utilisateurs, il suffit de changer d'utilisateur pour changer l'accès aux différents espaces.

Figure 3-6 Espace Administrateur pour le traitement des clients.

Cette page est pour l'ajout et la mise à jour des informations utilisateurs du système.

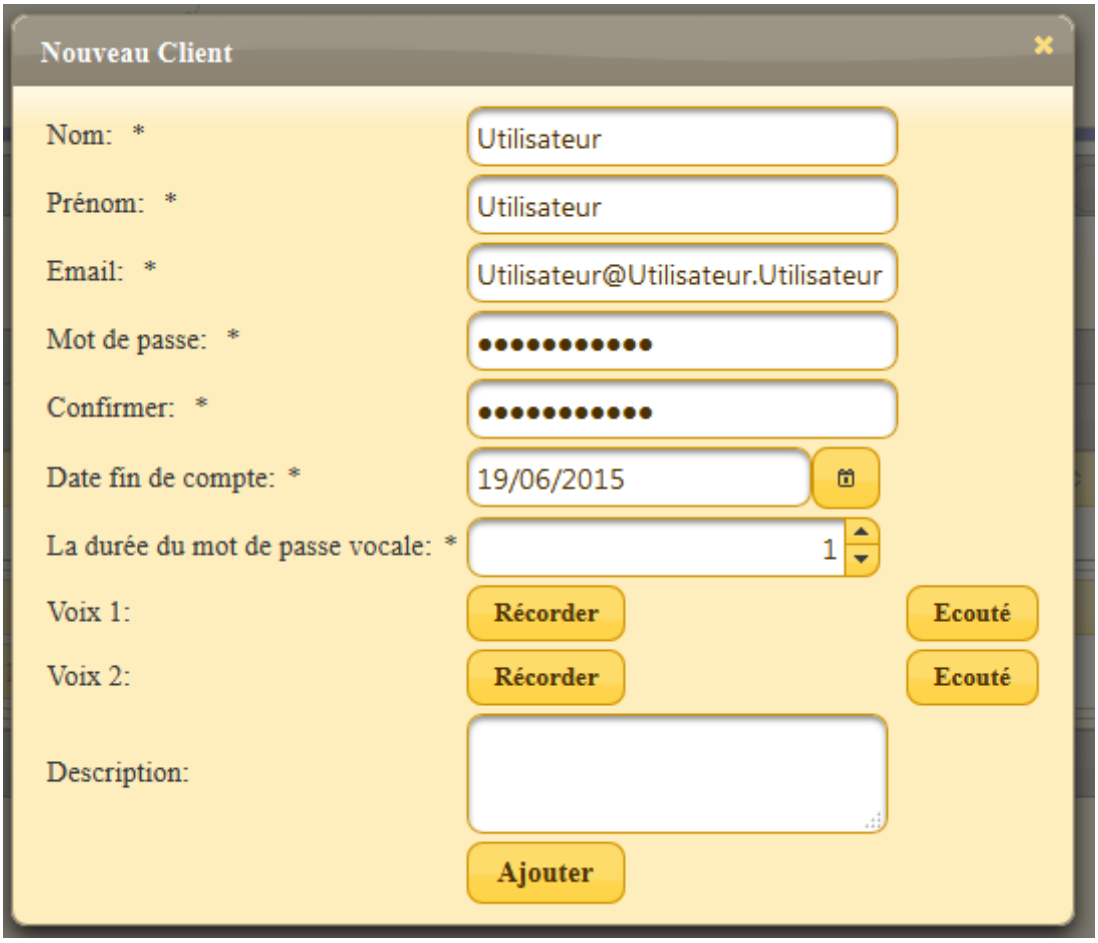


Figure 3-7 Traitement pour Ajouter un utilisateur.

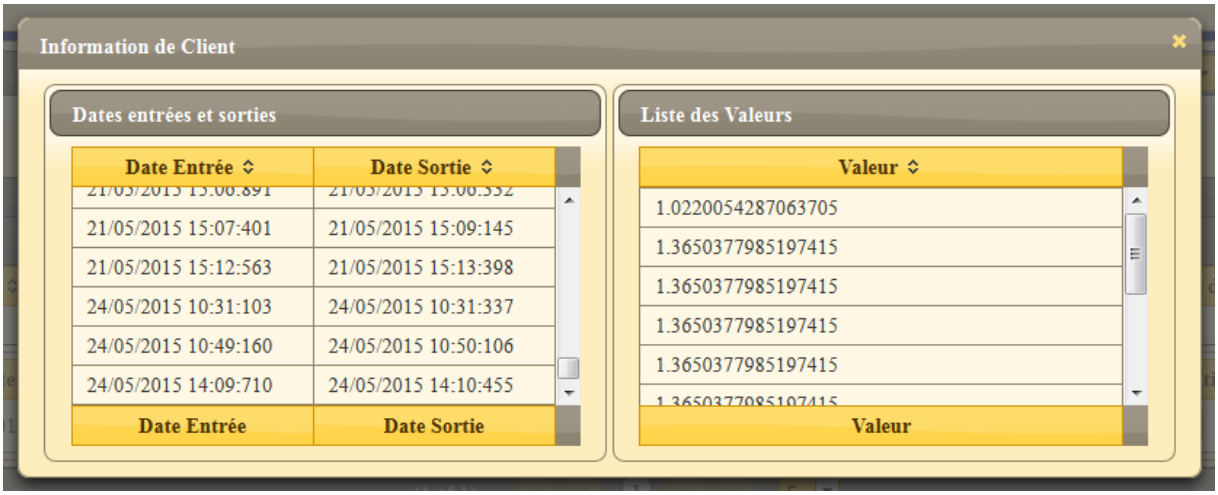


Figure 3-8 Les Informations de l'utilisateur.

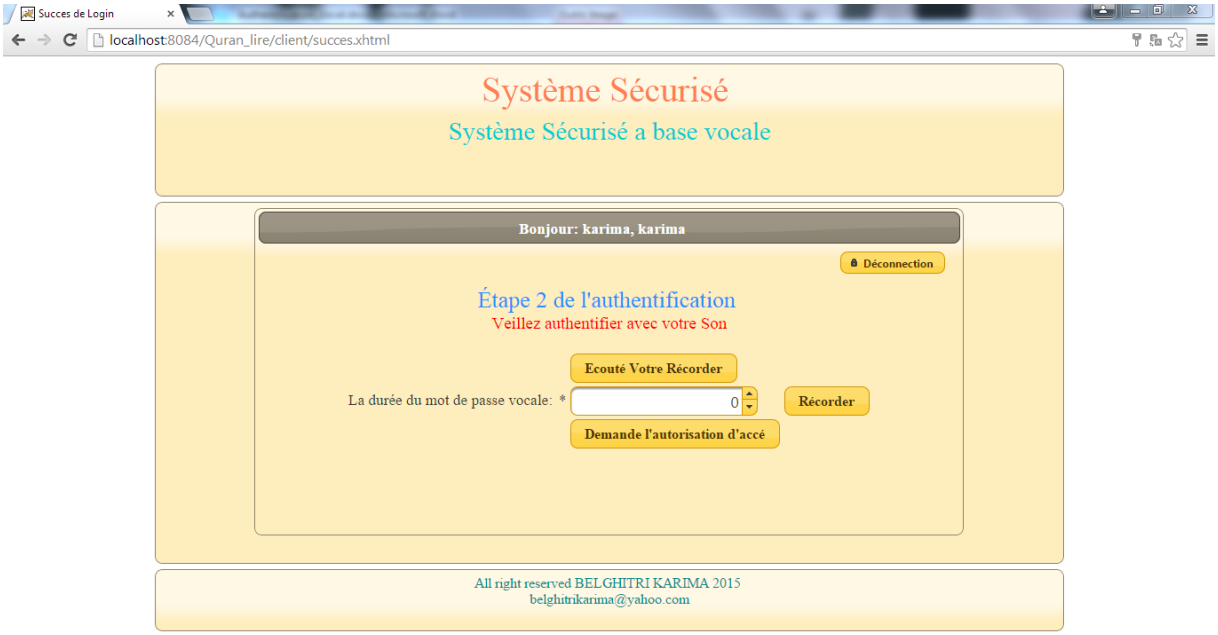


Figure 3-9 Etape deux de l'authentification des utilisateurs.

Cette page est pour la deuxième étape de l'authentification, cette étape est basée sur la comparaison entre les deux voix, c'est-à-dire la nouvelle voix et la voix qui existe dans la base de données.

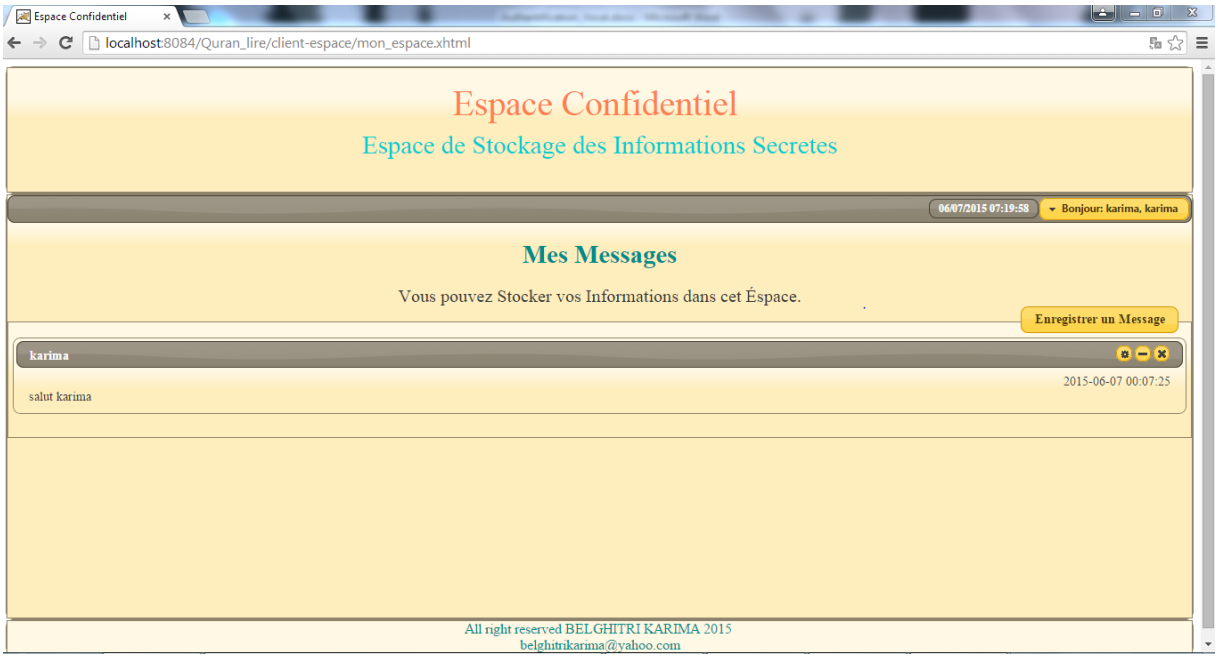


Figure 3-10 Espace personnel des utilisateurs.

Cette page est pour stocker les messages des utilisateurs.

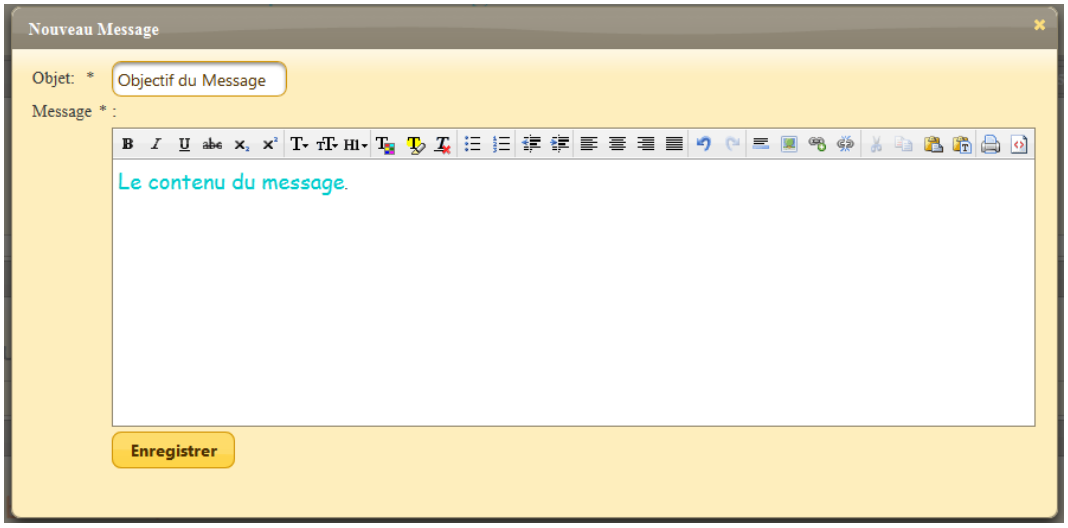


Figure 3-11 Création d'un message pour Stocker.

6. Les Tests

Voici quelques résultats de notre application.

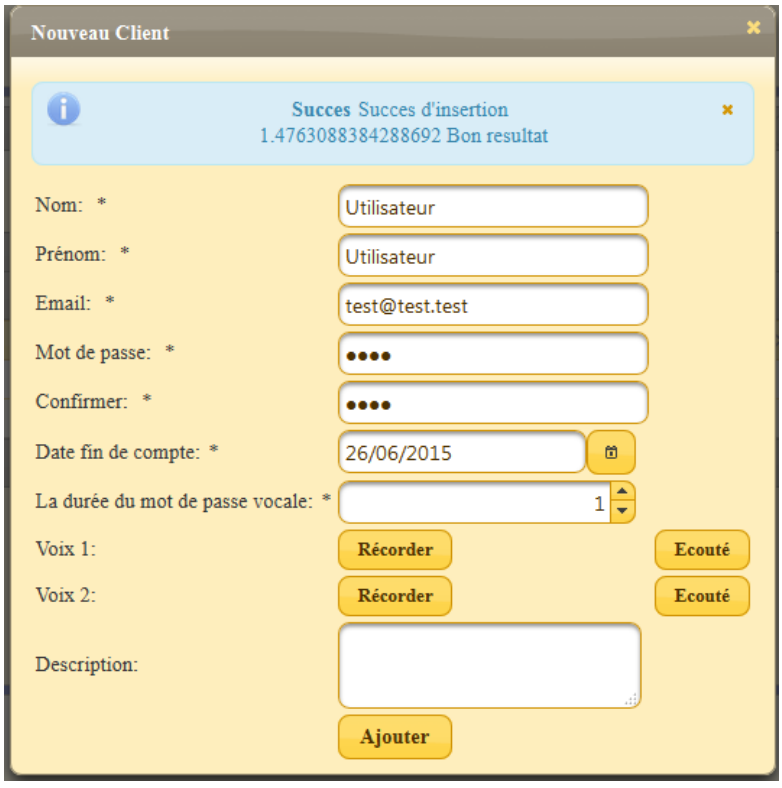


Figure 3-12 Résultat positif.

Voici un exemple positif pour la comparaison entre les deux sons, donc le résultat est entre 0 et 2, donc c'est considéré comme étant un bon résultat.

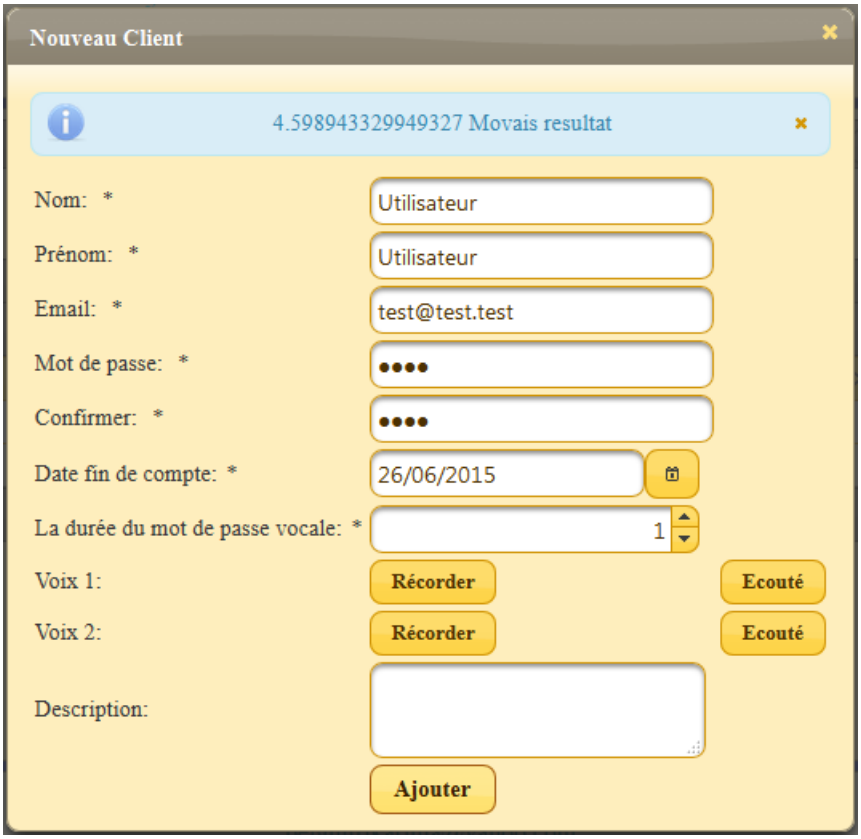


Figure 3-13 Résultat négatif

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, surtout dues aux algorithmes très complexes à programmer, cumulée avec la difficulté du traitement du son. 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 Générale

Ce PFE traitant de la reconnaissance vocale de personnes préenregistrées dans le but de faire une authentification lors d'un contrôle d'accès à un réseau, a été mené à bien, et a répondu aux objectifs que nous nous sommes posés au préalable. Motivée par l'amélioration de la précision de la reconnaissance d'une personne par la fusion des différentes sources d'information, ce PFE se concentre sur l'exploitation de l'information de la source vocale spécifique au locuteur. 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 du locuteur. Le principal défi dans la technologie de reconnaissance du locuteur a donc été 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 le locuteur. Le développement et la croissance exponentielle planétaire des communications, tant en volume qu'en diversité (déplacement physique, transaction financière, accès aux services...), implique le besoin de s'assurer de l'identité des individus. L'importance des enjeux, motive les fraudeurs à mettre en échec les systèmes de sécurité existants. 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 a 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 un choix de deux techniques les plus répandues, comme base à notre implémentation, puis nous avons procédé à la conception, réalisation, et conclue 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] memoirepfe, «La reconnaissance automatique du locuteur par la voix IP,» [En ligne]. Available: www.memoirepfe.fst-usmba.ac.ma/get/pdf/462. [Accès le janvier 2015].

[3] K. A. Chris Adamson, Learning Core Audio, Addison-Wesley, 2012.

[4] N. Moreau, Tools for Signal Compression, Wiley, 2011.

[5] G. Binet, Traitement numérique du signal, Ellipses, 2013.

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

[7] 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.

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

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

[10] R. McMunn, Speed, Distance and Time Questions (Testing Series), 2015.

[11] V. Nasser, Speed Distance Time and Numerical Reasoning Tests: Useful for the numerical part of AOSB tests, 2014.

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

[13] A. Leonard, JSF 2.0 Cookbook, Packt Publishing, 2010.

[14] M. C. Oleg Varaksin, PrimeFaces Cookbook, Packt Publishing, 2003.

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

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

[17] I. F. D. Jason Brittain, Tomcat: The Definitive Guide, 2007.

[18] G. Z. e. M. Sekler, Beginning JSP™, JSF™, and Tomcat Web Development, New York: Steve Anglin, 2007.

[19] C. DELANNOY, Programmer en Java (Java 5.0)., 75240 Paris Cedex 05: EYROLLES, 2006.

Résumé

Motivée par l'amélioration de la précision de la reconnaissance d'une personne par la fusion des différentes sources d'information, ce PFE se concentre sur l'exploitation de l'information de la source vocale spécifique au locuteur. 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 du locuteur. Le principal défi dans la technologie de reconnaissance du locuteur a donc été 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 le locuteur. Le développement et la croissance exponentielle planétaire des communications, tant en volume qu'en diversité (déplacement physique, transaction financière, accès aux services...), implique le besoin de s'assurer de l'identité des individus. L'importance des enjeux, motive les fraudeurs à mettre en échec les systèmes de sécurité existants. 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 a 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.Dans ce PFE donc, on s'est posé comme objectif principal la mise au point d'un système de reconnaissance vocale fonctionnant à partir de comparaison entre cepstres MFCC. Plusieurs étapes sont nécessaires pour transformer un fichier audio en cepstre MFCC et puis ces derniers vont subir des traitements par l'algorithme DTW qui sert à calculer des distances pour enfin arriver à résoudre le problème de l'authentification vocale.

ملخص

بدافع تحسين دقة التعرف على الشخص من قبل مختلف مصادر المعلومات، تركز هده المذكرة على استغلال معلومات عن مصدر خطاب محدد للمتكلم. تعتبر المعلومات الموجودة في الصوت عموما أقل تمييزا ولكن من الصعب استخراجها ولكن، مع تطور التكنولوجيا، والتخزين الموارد والدراسات في مجال فهم ظاهرة إنتاج وإدراك الكلام، أدت الباحثين إلى إعادة النظر في هذه الأحكام المسبقة ومحاولة الحصول على معظم هذه المعلومات إضافية لتحسين أداء نظام التعرف على مكبر الصوت. وكان التحدي الرئيسي في تكنولوجيا التعرف على المتكلم إلى تحسين متانة النظم في ظروف الغير المتوافقة. نظامنا يوفرا المؤشرات الصوتية أساسا لتصنيف شخصية الفرد، وكذلك لتميز المتكلم. التنمية والنمو المتسارع العالمي للاتصالات، سواء من حيث الحجم والتنوع (التهجير الجسدي، والمعاملات المالية، والحصول على الخدمات ...) تعني الحاجة إلى التحقق من هوية الأفراد. أهمية هذه القضايا، يحفز المحتالين لهزيمة الأنظمة الأمنية القائمة. صوت يحمل المعلومات المختلفة، والكلام البشري ينظر إليها على أنها قضية الأصوات منظم، هو في الأساس وسيلة للاتصال. على هذا النحو، وهو إشارة الكلام عادة ما يكون حاملها من جهة الرسالة إلى شخص آخر. التغيير في طبيعة الإشارة الصوتية يجعل معالجة البيانات الخام من الأخير صعبة للغاية. في الواقع، تحتوي هذه البيانات معلومات معقدة، وغالبا ما زائدة ومختلطة مع الضوضاء.

Abstract

Motivated by improving the accuracy of recognition of a person by the fusion of different sources of information, PFE focuses on the exploitation of information on the specific speech source to the speaker. Voice source parameters are generally considered less discriminating but difficult to extract. However, with the evolution of technology, storage and computing resources, studies in the area of understanding the phenomenon of production and perception of speech, have led researchers to reconsider these prejudices and try to get the most of this additional information to improve the performance of speaker recognition system. The main challenge in speaker recognition technology has been to improve the robustness of the systems in conditions incompatible. Our voice system provides mainly acoustic indices for classification of phonemes, the individual personality and, also to characterize the speaker. Development and global exponential growth of communications, both in volume and diversity (physical displacement, financial transactions, access to services ...) implies the need to verify the identity of individuals. The importance of the issues, motivates fraudsters to defeat existing security systems. The voice carries various information, human speech seen as an issue of structured sounds, is basically a means of communication. As such, a speech signal is usually the bearer of a message destination to another person. The change in the nature of the acoustic signal makes the processing of the raw data from the latter very difficult. Indeed, these data contain complex information, often redundant and mixed with noise. In this PFE so we landed as main objective the development of a speech recognition system operating from comparison cepstra MFCC.Several steps are required to transform an audio file cepstrum and MFCC then the latter will undergo treatment by the DTW algorithm used to calculate distances to finally get to solve the problem of voice authentication.