M2ISE – ou – M1Info

Logos des organismes d'accueil

Université Paris 8 - Vincennes à Saint-Denis

Master Informatique des Systèmes Embarqués

Reconnaissance automatique de la parole

Farès Belhadj

Date de soutenance : le JJ/MM/AAAA

Organisme d'accueil :	XXXXXXX (si stage)
Tuteur – Organisme d'accueil :	Prénom Nom (si stage)
Tuteur – Université :	Prénom Nom

Résumé

Ce document regroupe les informations nécessaires à la compréhension et à la mise en place de certaines méthodes dîtes de reconnaissance de la parole. En outre, il sera question des outils de traitement de la parole existants, dont certains que nous devrons apprendre à maîtriser afin de développer notre programme.

Le projet en question repose sur le développement d'un programme informatique qui devra permettre l'identification d'une personne à partir de sa voix. Avant d'en arriver là, le programme devra d'abord effectuer un traitement de la voix de cette personne, afin d'en obtenir un modèle qui sera utilisé pour comparer cette voix avec des voix diverses de test, puis générer un résultat nous permettant de savoir si les deux voix correspondent, et donc si il s'agit de la même personne ou non.

Nous entamerons ce rapport par une présentation des techniques permettant l'acquisition du son et son traitement. Puis, nous parlerons des méthodes permettant d'extraire les caractéristiques d'un son. Ensuite, nous aborderons les techniques permettant l'entraînement des modèles. Enfin, nous décrirons les méthodes utilisées pour tester l'identification d'une personne et nous analyserons les résultats obtenus.

Remerciements

La page des remerciements n'est pas obligatoire. Elle reste votre seul vrai espace de liberté complet. Il existe néanmoins une codification classique des remerciements consistant à remercier les personnes que vous citez de la relation la plus strictement professionnelle et hiérarchique à la relation la plus personnelle.

Table des matières

Table des figures

Liste des algorithmes

Introduction

Le domaine de la reconnaissance vocale se compose d'une multitude de méthodes et de techniques permettant d'effectuer des traitements de la parole.

Parmi ces techniques, nous retrouvons notamment la reconnaissance du locuteur (permet d'identifier une personne d'après sa voix), la reconnaissance de la parole (permet d'analyser une voix afin de la transcrire sous forme de texte), ou encore la synthèse de la parole (permet de créer une parole artificielle à partir d'un texte).

De nos jours, ces techniques sont très répandues dans les outils informatiques de tous les jours (ordinateurs, smartphones, objets connectés, etc), généralement pour la vérification d'identité ou encore l'exécution de commandes vocales. Ces plateformes offrent une utilisation simple et rapide de ces techniques, tout en restant performantes.

Afin de tester les capacités et les performances de certaines de ces techniques, nous allons travailler sur un projet de reconnaissance de locuteur, qui devra permettre d'identifier une personne d'après sa voix.

L'objectif final étant de pouvoir répondre à la question "Qui parle ?".

Nous allons nous appuyer sur les nombreux outils développés dans le but de faciliter le traitement de la parole, et ainsi réaliser notre propre programme d'identification de locuteur.

Chapitre 1

Etat de l'art

1.1 Acquisition du son

Une membrane vibre sous l'effet de la pression acoustique et un dispositif qui dépend de la technologie du microphone convertit ces oscillations en signaux électriques. La conception d'un microphone comporte une partie acoustique et une partie électrique, qui vont définir ses caractéristiques et le type d'utilisation. (https ://fr.wikipedia.org/wiki/Microphone)

1.1.1 Type de microphone

Le choix du microphone dépend des applications de notre modèle de reconnaissances , dans le cas ou on favorise d'une source émettrice loin ou proche avec beaucoup de bruit ou peu.

Microphones dynamiques

Les microphones dynamiques utilisent un ensemble diaphragme / bobine acoustique / aimant qui forme un générateur électrique miniature piloté par le son. Les ondes sonores frappent une fine membrane de plastique (diaphragme) qui vibre en réponse. Une petite bobine de fil (bobine mobile) est fixée à l'arrière du diaphragme et vibre avec ce dernier. La bobine acoustique elle-même est entourée d'un champ magnétique créé par un petit aimant permanent. C'est le mouvement de la bobine acoustique dans ce champ magnétique qui génère le signal électrique correspondant au son capté par un microphone dynamique. Les microphones dynamiques ont une construction relativement simple et sont donc économiques et robustes. Ils peuvent fournir une excellente qualité sonore et de bonnes spécifications dans tous les

domaines de la performance du microphone. En particulier, ils peuvent gérer des niveaux sonores extrêmement élevés : il est presque impossible de surcharger un microphone dynamique. De plus, les microphones dynamiques sont relativement peu affectés par les extrêmes de température et d'humidité. La dynamique est le type le plus utilisé dans le renforcement acoustique général.

Microphones à condensateur

Les microphones à condensateur sont basés sur un assemblage diaphragme / plaque arrière chargé électriquement qui forme un condensateur sensible au son. Ici, les ondes sonores font vibrer un diaphragme très fin en métal ou en plastique recouvert de métal. Le diaphragme est monté juste devant une plaque arrière en métal rigide ou en céramique revêtue de métal. En termes électriques, cet ensemble ou élément est appelé un condensateur (appelé historiquement un "condensateur"), qui a la capacité de stocker une charge ou une tension. Lorsque l'élément est chargé, un champ électrique est créé entre le diaphragme et la plaque arrière, proportionnel à leur espacement. C'est la variation de cet espacement, due au mouvement du diaphragme par rapport à la plaque arrière, qui produit le signal électrique correspondant au son capté par un microphone à condensateur. La construction d'un microphone à condensateur doit inclure une disposition permettant de maintenir la charge électrique ou la tension de polarisation. Un microphone à condensateur électret a une charge permanente, maintenue par un matériau spécial déposé sur la plaque arrière ou sur le diaphragme. Les types non-électret sont chargés (polarisés) au moyen d'une source d'alimentation externe. La majorité des microphones à condensateur pour l'amplification du son sont du type électret. Tous les condenseurs contiennent des circuits actifs supplémentaires permettant à la sortie électrique de l'élément d'être utilisée avec des entrées de microphone classiques. Cela nécessite que tous les microphones à condensateur soient alimentés : soit par piles, soit par alimentation fantôme (méthode consistant à alimenter un microphone par le câble du microphone lui-même). Les microphones à condensateur présentent deux limitations potentielles dues aux circuits supplémentaires : premièrement, les composants électroniques produisent une faible quantité de bruit ; deuxièmement, il existe une limite au niveau de signal maximal que l'électronique peut gérer. Pour cette raison, les spécifications du microphone à condensateur incluent toujours un facteur de bruit et un niveau sonore maximal. Les bonnes conceptions, cependant, ont des niveaux de bruit très bas et sont également capables de très grande plage dynamique.

Différence entre microphones à condensateur et microphone dynamique

Les microphones à condensateur sont plus complexes que les dynamiques et ont tendance à être un peu plus coûteux. De plus, les condenseurs peuvent être affectés par des températures et des taux d'humidité extrêmes, ce qui peut les rendre bruyants ou en panne de façon temporaire. Cependant, les condenseurs peuvent facilement être fabriqués avec une sensibilité plus élevée et peuvent fournir un son plus doux et plus naturel, en particulier à des fréquences élevées. La réponse en fréquence plate et la plage de fréquence étendue sont beaucoup plus faciles à obtenir dans un condenseur. De plus, les microphones à condensateur peuvent être très petits sans perte significative de performances. Image Microphone à condensateur La décision d'utiliser un microphone à condensateur ou dynamique dépend non seulement de la source sonore et du système de renforcement acoustique, mais également du réglage physique. D'un point de vue pratique, si le microphone doit être utilisé dans un environnement sévère tel qu'un club de rock'n'roll ou pour le son en extérieur, des types dynamiques constitueront un bon choix. Dans un environnement plus contrôlé, tel qu'une salle de concert ou une configuration théâtrale, un microphone à condensateur peut être préféré pour de nombreuses sources sonores, en particulier lorsque la qualité sonore optimale est désirée. (https ://www.shure.com/en-US/support/find-an-answer/difference-between-a-dynamic-andcondenser-microphone)

1.1.2 Etape de traitement du son

Le traitement numérique du signal par ordinateur exige que le signal soit converti en une suite de nombres (numérisation). Cette conversion se décompose, sur le plan théorique, en trois opérations

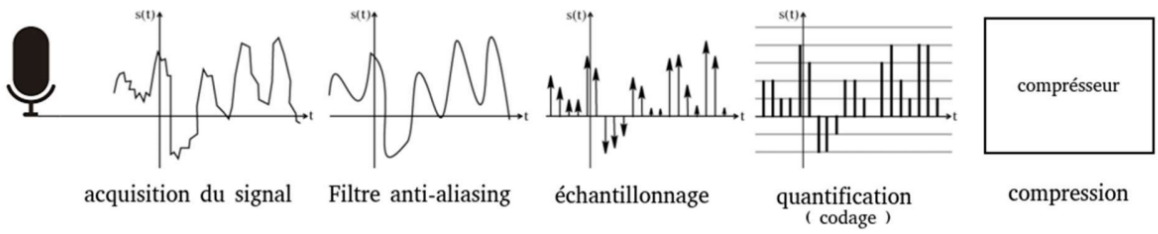

FIGURE 1.1 – étape de traitement du son

Échantillonnage

L'échantillonnage consiste à prélever les valeurs d'un signal à intervalles définis, généralement réguliers. Il produit une suite de valeurs discrètes nommées échantillons.(https ://fr.wikipedia.org/wiki/

Cadence d'échantillonnage : (théoreme de Shannon-Nyquist)

si toutes les fréquences du signal sont inférieures à la moitié de la fréquence d'échantillonnage, il peut être parfaitement reconstitué (frequence d'échantionnage = 2x frequenceMax du signal)

Quantification (conversion analogique numérique)

En traitement des signaux, la quantification est le procédé qui permet d'approcher un signal continu par les valeurs d'un ensemble discret d'assez petite taille. L'amplitude relevée à chaque étape d'échantillonnage va être codée en binaire sur un certain nombre de bits : 8, 16, 24, 32... C'est la quantification. Là encore, plus le nombre de bits va être élevé, plus la valeur numérique de l'amplitude sera proche de la valeur originale.

Compression

La compression audio est une forme de compression de données qui a pour but de réduire la taille d'un flux audio numérique en vue d'une transmission (contraintes de largeur de bande et de débit) ou d'un stockage (contrainte d'espace de stockage). On distingue la compression sans perte, qui permet de reconstituer exactement les données d'origine, de la compression en général, « avec pertes », qui abandonne des données jugées non nécessaires à l'écoute, au profit de la diminution du débit ou de la taille des fichiers.

1.2 Extraction de paramètres

Une fois l'enregistrement audio effectué, il sera traité afin d'obtenir des données utilisables par un programme informatique (conversion d'un signal analogique vers un signal numérique/digital).

1.2.1 Reconnaissance vocale – Paramétrisation / traitement

Le traitement de l'enregistrement obtenu passe par l'analyse de plusieurs paramètres qui le composent (volume sonore, bruits de fond, intonation, etc),

appelés traits prosodiques. Les différents traits prosodiques (paramètres prosodiques) :

— l'accent
— le ton
— l'intonation
— la jointure (ex : « coopérer »)
— la pause
— le rythme
— le tempo et le débit

Ces caractéristiques vont influer sur la manière dont certains sons vont être interprétés par le programme de reconnaissance vocale.

1.2.2 Algorithmes de traitement du signal audio

Afin de traiter le signal audio, il lui sera appliqué un algorithme spécifiquement crée pour ce type de signal. Il en existe plusieurs, certains plus efficaces selon la clarté de l'enregistrement, la présence de bruit, etc. Le signal ne sera pas traiter en un seul bloc, mais sera découpé en plusieurs segments (selon un intervalle de temps ou selon un intervalle de sons) de même longueur (environ 20 à 25 millisecondes) et qui se superposeront (la fin d'un segment – les 10 dernières millisecondes – et le début du segment suivant – les 10 premières millisecondes – seront à cheval sur les mêmes données). De cette manière, nous pouvons travailler sur des échantillons de sons plus petits et obtenir un résultat plus précis après traitement.

Transformation de Fourier

C'est l'une des opération les plus fréquemment effectuée pour le traitement des signaux. Elle permet de passer de la représentation temporelle d'un signal à sa représentation fréquentielle / spectrale.

Transformation de Fourier discrète

La transformation de Fourier discrète (TFD ou DFT en anglais) est un outil mathématique de traitement du signal numérique, qui est l'équivalent discret de la transformation de Fourier continue qui est utilisée pour le traitement du signal analogique. Elle est typiquement utilisée sur des sons.

Codage prédictif linéaire (LPC – Linear Predictive Coding)

Le codage prédictif linéaire est une méthode de codage et de représentation de la parole. Elle est appliquée sur un signal, afin d'en obtenir un

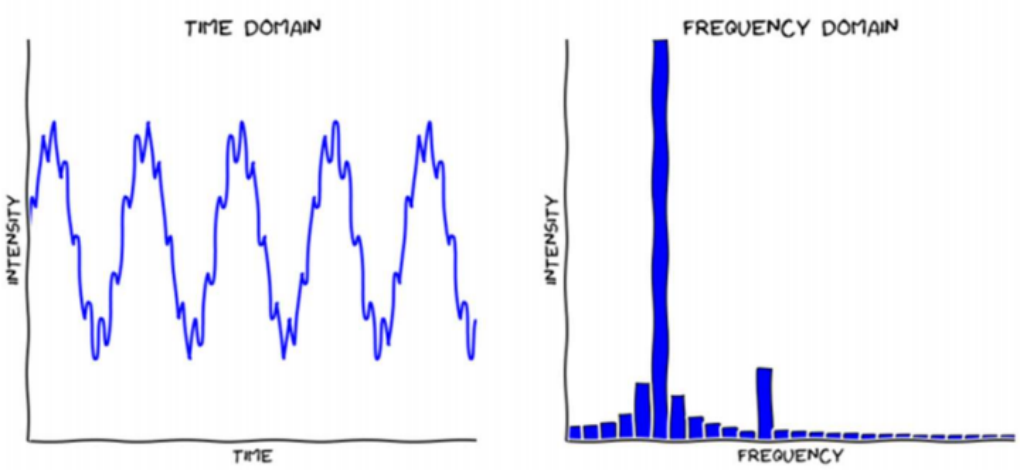

FIGURE 1.2 – Transformation d'un signal

modèle.

« Elle repose principalement sur l'hypothèse que la parole peut être modélisée par un processus linéaire. Il s'agit donc de prédire le signal à un instant n à partir des p échantillons précédents. La parole n'étant cependant pas un processus parfaitement linéaire, la moyenne que constitue la somme pondérée du signal sur p pas de temps introduit une erreur qu'il est nécessaire de corriger par l'introduction du terme e(n). » .

MFCC - Mel Frequency Cepstral Coefficients

Le MFCC permet d'appliquer des transformations à un signal (semblables à une transformation de Fourier), afin d'en obtenir une modélisation sous forme d'un spectre. C'est actuellement le plus utilisé pour les programmes de reconnaissance vocale.

Son avantage est qu'il utilise l'échelle de Mel pour mesurer la fréquence d'un signal, ainsi, son spectre sera plus précis et aura un aspect très proche de ce qui serait perçu par un humain.

Efficacité de MFCC

MFCC est considéré comme très efficace lorsqu'il est appliqué sur un enregistrement propre (pas de bruits de fond, bon volume vocal, etc), mais moins robuste lors de présence de bruit.

On notera néanmoins que l'analyse MFCC Aurora a été développée de manière à effectuer un dé-bruitage sur un tel signal.

Exemple de mise en place

(page 16) : https ://hal.inria.fr/tel-01251128/document

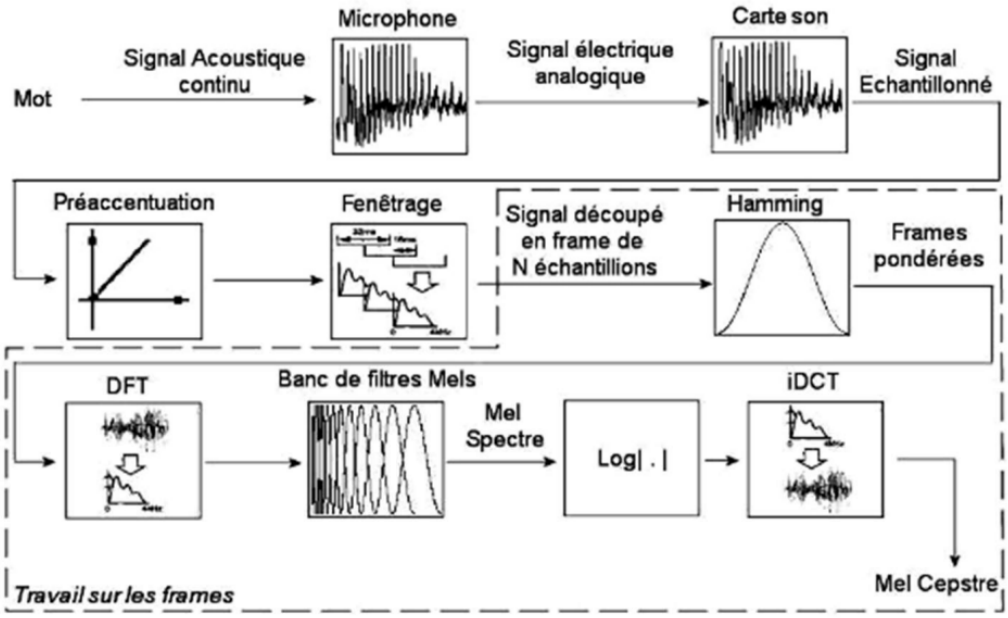

FIGURE 1.3 – Etape du calcul des coefficients MFCC

Résultat

Le spectre finalement obtenu représente ainsi les informations phonétiques citées précédemment (traits prosodiques). Nous pouvons alors utiliser les valeurs de ce spectre afin d'effectuer les opérations voulues (modification du signal, récupération d'un d'un son, etc).

1.2.3 Outils d'extraction des features

SPro

Htk

1.3 Apprentissage du modéle

Dans cette étape on réalise une association entre les segments élémentaires de la parole et les éléments lexicaux. Cette association fait appel à une modélisation statistique ou par réseaux de neurones artificiels ou par algorithme de déformation temporelle dynamique

1.3.1 Approche statistique

L'RAS vise à convertir le signal vocal en texte et ce processus peut être formulé statistiquement comme suit. Soit un ensemble d'observations acoustiques O = (o1, o2, o3,. . . ,on) (séquence de vecteurs de parole, où oi est le vecteur de parole observé à l'instant i), qui est la séquence de mots W = (w1, w2, . . . , Wn) qui a la probabilité maximale :

<equation1>

L'équation (1) spécifie la séquence de mots la plus probable à l'aide de la règle de Bayes et P(O) - la probabilité d'énonciation de la parole - peut être ignorée, car elle est indépendante de la séquence W. Ainsi, (1) devient :

<equation2>

L'équation (2) contient deux facteurs qui peuvent être directement estimés : la probabilité a priori de la séquence de mots P (W) et la probabilité des données acoustiques, étant donné la séquence de mots P (O | W). Le premier facteur P (W) peut être estimé en utilisant uniquement un modèle de langage et le second facteur peut être calculé à partir du modèle acoustique. La numérisation à deux modèles doit être construite indépendamment, mais ils seront utilisés ensemble pour reconnaître un message parlé.

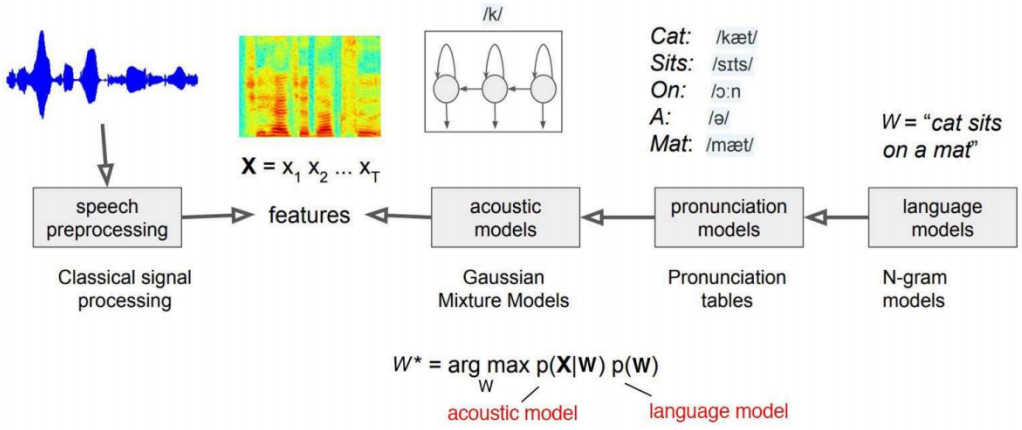

FIGURE 1.4 – Etape modéle statistique

Modèle de langue

Le modèle de langue décide si un mot (ou une phrase) est valide dans une langue donnée Un modèle de langage statistique est une distribution de probabilité sur des séquences de mots. Étant donné une telle séquence, disons de longueur m, il attribue une probabilité à la séquence entière.

Unigram

Un modèle unigramme peut être traité comme la combinaison de plusieurs automates finis à un état. [1] Il divise les probabilités de différents termes dans un contexte, par exemple. de

<equation3>

Dans ce modèle, la probabilité de chaque mot dépend uniquement de la probabilité de ce mot dans le document, de sorte que nous avons uniquement des automates finis à un état en tant qu'unités. L'automate lui-même a une distribution de probabilité sur tout le vocabulaire du modèle, en faisant un total de 1. Ce qui suit est une illustration du modèle unigramme d'un document.

<tableau1>

n-gram

Dans un modèle à n-grammes, la probabilité P (w1, w2,. . . , wm) d'observer la phrase w1, w2,. . . , wm est approximée comme suit :

<eq4>

On suppose que la probabilité d'observer le i-ème mot wi dans l'historique de contexte des mots i-1 précédents peut être approximée par la probabilité de l'observer dans l'historique de contexte raccourci des n-1 mots précédents (propriété de Markov d'ordre n) .

La probabilité conditionnelle peut être calculée à partir des comptes de fréquence du modèle ngramme :

<eq5>

Les termes modèles de langage bigram et trigram désignent les modèles à n-grammes avec n = 2 et n = 3, respectivement.

Exponential

Les modèles de langage d'entropie maximum codent la relation entre un mot et l'historique ngram à l'aide de fonctions. L'équation est :

<eq6>

où Z (w1, w2,. . . , wm-1) est la fonction de partition, a est le vecteur de paramètre et f (w1, w2,. . . wm) est la fonction de fonction. Dans le cas le plus simple, la fonction caractéristique n'est qu'un indicateur de la présence d'un certain n-gramme. Il est utile d'utiliser un préalable sur un ou une forme de régularisation. Le modèle log-bilinéaire est un autre exemple de modèle de langage exponentiel.

Neural network

La probabilité d'une séquence de mots peut être obtenue à partir de la probabilité de chaque mot étant donné le contexte des mots qui le précèdent, en utilisant la règle de probabilité en chaîne (une conséquence du théorème de Bayes) :

<eq7>

La plupart des modèles de langage probabilistes (y compris les modèles de langage réseau neuronal publiés) approchent <eq8> en utlisant un contexte fixe de taille n - 1, c'est-à-dire en utilisant <eq9>, comme en n-grammes.

Dans le modèle introduit dans (Bengio et al 2001, Bengio et al 2003), la prédiction probabiliste <eq10> est obtenue comme suit. Tout d'abord, chaque mot wt - i (représenté par un entier dans [1, N]) dans le contexte de n-1 mot est mappé sur un vecteur de caractéristique ddimensionnel associé Cwt - i, qui est la colonne wt - i de la matrice de paramètres C. Le vecteur Ck contient les fonctions apprises pour le mot k. Soit le vecteur x la concaténation de ces n-1 vecteurs de caractéristiques :

<eq11>

La prédiction probabiliste du mot suivant, à partir de x, est ensuite obtenue à l'aide d'une architecture de réseau de neurones artificielle standard pour la classification probabiliste, à l'aide de la fonction d'activation softmax au niveau des unités de sortie (Bishop, 1995) :

<eq12>

où

<eq13>

où les vecteurs b, c et les matrices W, V sont également des paramètres (en plus de la matrice C). Notons <eq14> pour la concaténation de tous les paramètres. La capacité du modèle est contrôlée par le nombre d'unités cachées h et par le nombre de fonctions de mots apprises d.

Modéle acoustique

Le modèle acoustique doit estimer la probabilité de prononcer un message, à partir d'une séquence de mots.

Pour tout w donné, le modèle acoustique correspondant est synthétisé en concaténant des modèles de téléphone pour créer des mots tels que définis par un dictionnaire de prononciation.

Un modèle acoustique est utilisé dans la reconnaissance automatique de la parole pour représenter la relation entre un signal audio et les phonèmes ou autres unités linguistiques qui composent la parole. Le modèle est appris à partir d'un ensemble d'enregistrements audio et de leurs transcriptions correspondantes. Il est créé en prenant des enregistrements audio de la parole et leurs transcriptions de texte, et en utilisant un logiciel pour créer des représentations statistiques des sons qui composent chaque mot.

HMM (hidden markov model)

Une chaîne de Markov contient tous les états possibles d'un système et la probabilité de passer d'un état à un autre.

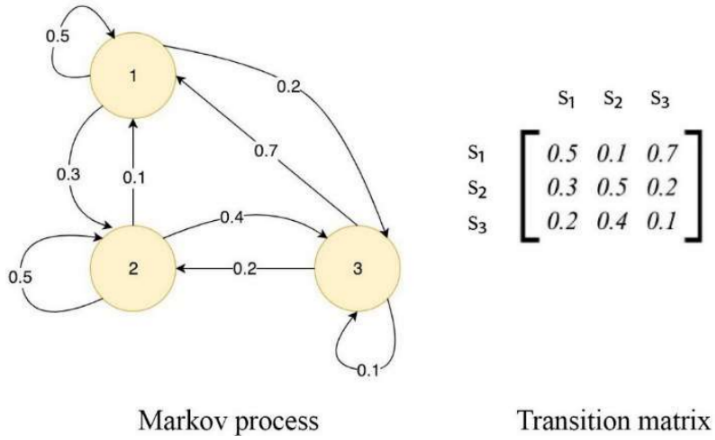

FIGURE 1.5 – Chaîne de Markov

A first-order Markov chain assumes that the next state depends on the current state only. For simplicity, we often call it a Markov chain

<eq14>

Ce modèle sera beaucoup plus facile à manipuler. Cependant, dans de nombreux systèmes ML, tous les états ne sont pas observables et nous appelons ces états états cachés ou états internes. Certains peuvent les traiter comme des facteurs latents pour les intrants. Par exemple, il peut être difficile de savoir si je suis heureuse ou triste. Mon état interne sera H ou S. Mais nous pouvons obtenir des indications de ce que nous observons. Par exemple, lorsque je suis heureux, j'ai 0,2 chance de regarder un film, mais quand je suis triste, cette chance monte à 0,4. La probabilité d'observer un observable étant donné un état interne s'appelle la probabilité d'émission. La probabilité de passer d'un état interne à un autre s'appelle la probabilité de transition.

Modéle acoustique HMM/GMM

Modéle acoustique HMM/DNN

Modèle Phonétique

1.3.2 Approche par réseau de neurones

1.3.3 Approche par Dynamic time wrapping approch

1.3.4 Outils d'apprentissage de modéle

Chapitre 2

Conception et réalisation

Chapitre 3

Tests et résultats

Chapitre 4

Conclusion et Perspectives

Vous arrivez à la presque-fin de votre périple (oui il restera le résumé à faire, rappelez-vous), la conclusion. Ici, il est attendu d'avoir un bilan du travail réalisé [1]. Ce dernier doit être consolidé par les réalisations et les résultats obtenus. Il est utile de rappeler les améliorations apportées en les replaçant brievement dans leur contexte. Aussi, il est conseillé d'avoir un point de vue critique vis-à-vis de votre travail et souligner les points pouvant être améliorés. Ceci s'enchainera parfaitement avec les perspectives qui ouvrent la voie vers les nouvelles réalisations possibles sur la base de vos travaux. Les perspectives peuvent être données à court, moyen et long terme.

Par exemple, une conclusion à ce document peut être : « Dans ce document, nous avons présenté un ensemble de règles permettant d'écrire un mémoire de stage ou de projet tuteuré. Ce document utilise un langage de formatage de texte nommé LaTeX. En perspectives, nous souhaitons que l'ensemble des étudiants lisent attentivement et utilisent ce document. Enfin, nous pensons que ce type d'exercie deviendra un standard pour chacun d'entre-eux ».

1. Ne pas utiliser de formules du type « Ce stage a été très enrichissant » ou « Ce projet m'a beaucoup apporté sur le plan professionnel ou personnel » car si le travail en question est important ou intéressant le mémoire doit naturellement le refléter.