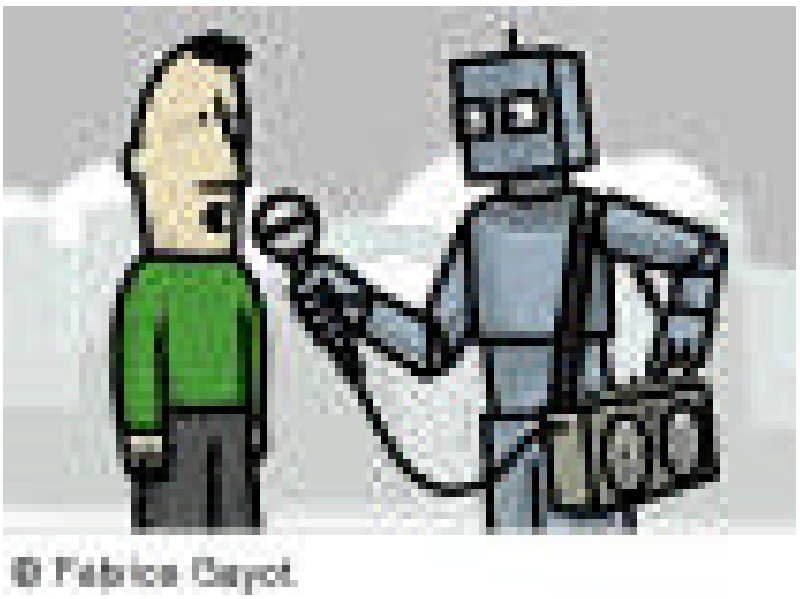
La

Reconnaissance

Vocale



Reconnaissance vocale : conversion de la voix en fichier numérique qui permet de décoder un signal acoustique de parole en une suite de mots effectivement prononcés

SOMMAIRE

INTRODUCTION

1 . L’approche d’un rêve

« Sésame, ouvre-toi ! »

Cette phrase mythique n’est pas sans signification, car en dépit du trésor caché derrière la porte de pierre, une autre découverte s’ouvre à nous :

La recherche en Reconnaissance Automatique de la Parole (RAP).

Celle-ci ne cesse de s’étendre dans nos foyer en dépit de l’étonnement qu’avaient nos « chères petites têtes blonde » en regardant bioman donner des ordres à son vaisseau.

Nous sommes cependant en dessous de la fiction étant donné la difficulté que nous avons encore a analyser un signal vocal complètement aléatoire. Si dans un téléphone, on écoute les sons qu’émettent un Minitel, un fax ou un micro-ordinateur pour échanger des données, ils se présentent à nous comme un sifflement suraigu bourré de parasites : le message semble parfaitement inintelligible. A l’inverse, alors que notre propre langage nous paraît simple et clair, la machine, elle, n’y détecte rien de cohérent.

SOMMAIRE

2 . Origine

Les USA sont encore une fois en première loges

C'est dans les années 40 au USA, que les premières tentatives de création d'une machine capable

de comprendre le discours humain eurent lieu. Leurs principaux objectifs étaient d'interpréter les messages russes interceptés.

SOMMAIRE

3 . Quelques applications

On utilise la reconnaissance vocale dans différents domaines.

- Une dictée vocale peut être associée à un traitement de texte : Un locuteur parle et le texte s?affiche ; ainsi, il n'a plus besoin de taper son texte au clavier.
- Les serveurs d?informations par téléphone
- La messagerie
- Elle permet l'autonomie : par exemple en médecine, lorsqu'un chirurgien a les deux mains occupées, il peut parler pour demander une information technique au lieu de taper sur un clavier (autonomie qui est aussi valable en industrie).
- La sécurité possible grâce à la signature vocale
- La possibilité de commande et de contrôle d'appareils à distance.

SOMMAIRE

4 . La reconnaissance vocale : oui mais à quel prix ?

Pendant ces premières années, il a fallu énormément de temps et de ressources informatiques pour enregistrer et emmagasiner la représentation de chaque mot dans chaque langue.

La représentation de symboles en discours n'est pas si simple, d'autant que différents symboles peuvent résulter de sons similaires. D'autres problèmes peuvent se poser : les sons individuels peuvent varier en fonction des sons qui suivent et qui précèdent.

La vitesse de traitement de la parole ne s'aligne pas encore avec celle d'un être humain : celle-ci est de 180 mots par minute, alors que des systèmes de reconnaissance vocale bien entraînés traitent au alentour de 130 mots par minute. Et là encore, «l' enrolling » (entraînement) qui consiste à lire un certain nombre de phrases de base en nombre suffisant pour créer un profil d'utilisateur de base, peut donner des performances médiocres s'il est mal établi.

Même le meilleur système de reconnaissance vocale ne pourra fonctionner correctement sans appui matériel. Le bruit de fond réduit considérablement le taux de précision ; par conséquent, des écouteurs conçus spécialement pour réduire le bruit sont recommandés.

SOMMAIRE

5 . Historique

Une évolution rapide

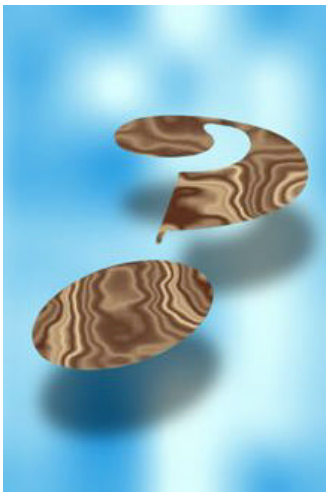
- 1952 : reconnaissance des 10 chiffres par un dispositif électronique câblé
- 1960 : utilisation des méthodes numériques
- 1965 : reconnaissance de phonèmes en parole continue
- 1968 : reconnaissance de mots isolés par des systèmes implantés sur gros ordinateurs (jusqu'à 500 mots)
- 1971 : lancement du projet ARPA aux USA (15 millions de dollars) pour tester la faisabilité de la compréhension automatique de la parole continue avec des contraintes raisonnables
- 1972 : premier appareil commercialisé de reconnaissance de mots
- 1978 : commercialisation d?un système de reconnaissance à microprocesseurs sur une carte de circuits imprimés
- 1983 : première mondiale de commande vocale à bord d?un avion de chasse en France
- 1985 : commercialisation des premiers systèmes de reconnaissance de plusieurs milliers de mots
- 1986 : lancement du projet japonais ATR de téléphone avec traduction automatique en temps réel

- 1988 : apparition des premières machines à dicter par mots isolés
- 1990 : premières véritables applications de dialogue oral homme-machine
- 1994 : IBM lance son premier système de reconnaissance vocale sur PC
- 1997 : lancement de la dictée vocale en continu par IBM

SOMMAIRE

I. LA RECONNAISSANCE VOCALE :

VISION D'ENSEMBLE



1 . Présentation

La parole est le principal vecteur d'information dans notre société humaine.

Située entre celui du signal numérique et du langage, son traitement s'est fortement développé parallèlement au développement des moyens et des techniques de télécommunications. Sa particularité, tient du rôle que joue le cerveau humain dans la production et la compréhension de la parole, par l'emploi automatique de diverses fonctions.

L'étude des mécanismes de phonation isole la parole de ce qui n'en est pas, et l'étude des mécanismes d'audition et de perception dit ce qui est réellement perçu dans le signal de parole. Perception et C*ompréhension* influence la production de la parole : on ne parle que dans la mesure où l'on s'entend et se comprend soi-même;
la complexité du signal qui en découle s'en ressent forcément !

S’il n’est pas de parole sans cerveau humain pour l’entendre, et la comprendre, les techniques modernes de traitement de la parole tendent à produire des systèmes automatiques et plus précisément les ***reconnaisseurs,*** **qui ont pour mission de décoder l’information portée par le signal vocal à partir des données fournies par l’analyse.**

SOMMAIRE

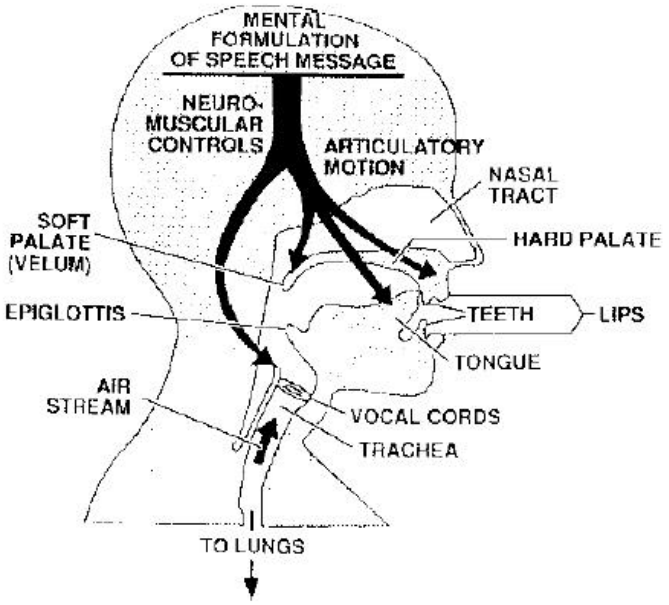
2 . La parole : c’est quoi déjà ?

La parole correspond à une variation de la pression de l’air causée par le système articulatoire. La phonétique acoustique étudie ce signal en le transformant dans un premier temps en signal électrique grâce au transducteur approprié qui de nos jours est le plus souvent numérisé.

phonétique acoustique : étude des propriétés physiques du son.

Il peut alors être soumis à un ensemble de traitements statistiques qui visent à mettre en évidence les ***traits acoustiques.***

La production de la parole



- Le son émis par le locuteur est capté par un microphone.

- Le signal vocal est numérisé à l'aide d'un convertisseur analogique-numerique

- Comme la voix humaine est constituée d'une multitude de sons, souvent répétitifs, le signal peut être compressé pour réduire le temps de traitement et l'encombrement en mémoire.

- L'analyse peut alors commencer ...

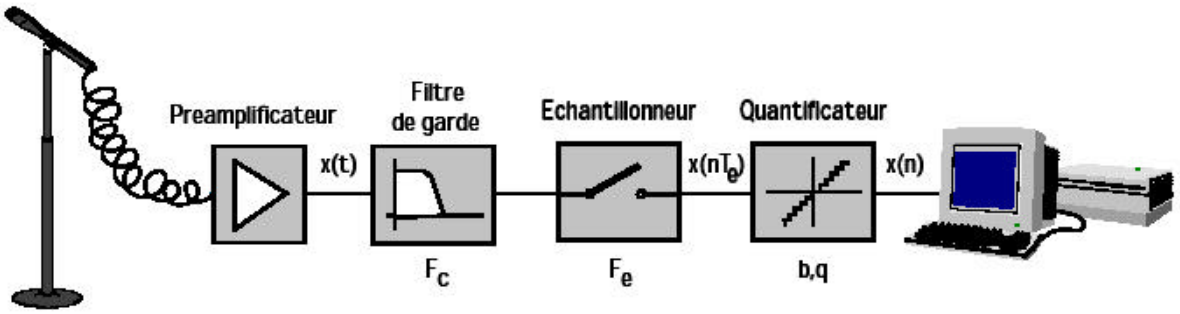
SOMMAIRE

3 . Paramétrisation

La paramétrisation du signal vocal s'effectue en deux temps et permet d'obtenir une « empreinte

caractéristique » du son, sur laquelle on pourra ensuite traiter la reconnaissance ...

1ère étape : Evolution temporelle du signal

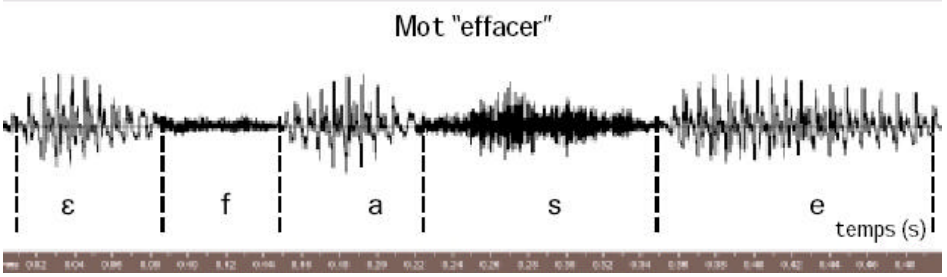


Enregistrement numérique d'un signal acoustique.
La fréquence de coupure du filtre de garde, la fréquence d'échantillonnage,
le nombre de bits et le pas de quantification sont respectivement notés fc , fe , b, et q.

L'échantillonnage transforme le signal à **temps continu** *x(t)* en signal à **temps discret** x(n) défini aux instants d'échantillonnage, multiples entiers de la période d'échantillonnage (inverse de la fréquence d'échantillonnage). Pour le signal vocal, il faut choisir une fréquence satisfaisant à peu près le **théorème de Shannon**.(24kHZ).

théorème de Shannon : L'information véhiculée par un signal dont le spectre est à support borné, n'est pas modifiée par l'opération d'échantillonnage, à condition que la fréquence d'échantillonnage soit au moins deux fois plus grande que la plus grande fréquence contenue dans le signal.

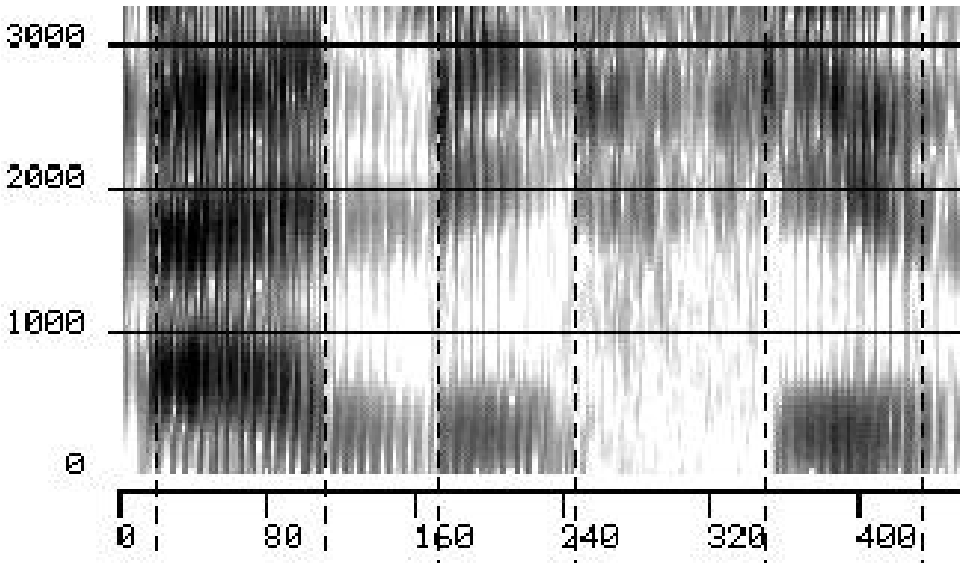
Parmi les valeurs possibles pour les échantillons *x(n)*, **la quantification ne retient qu'un nombre fini** *2b* de **valeurs** (*b* étant le **nombre de bits** de la quantification), espacées du **pas de quantification** *q*. Le signal numérique résultant est noté *x(n)*. Une quantification de bonne qualité requiert en général 16 bits.



Audiogramme de signaux de parole.

Il est souvent intéressant de représenter l'évolution temporelle du spectre d'un signal, sous la forme d'un **spectrogramme.** L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme en deux dimensions temps - fréquence. Ils mettent en évidence l'enveloppe spectrale du signal, et permettent donc de visualiser l'évolution temporelle des formants.

Les formants : ils constituent le facteur fondamental de la caractérisation du timbre.



La position et l'évolution des formants caractérise des sons produits. La seule lecture d'un spectrogramme (sans l'écoute du signal correspondant) permet d'ailleurs à l'œil expérimenté de certains phonéticiens de retrouver le contenu du message parlé : le spectrogramme présente sous une forme simple l'essentiel de l'information portée par le signal vocal.

L'évolution du signal vocal en fonction du temps n'est que la première étape de la paramétrisation Pour en déduire ses traits acoustiques, deux méthodes principales sont applicables...

2ème étape : Empreinte caractéristique du son

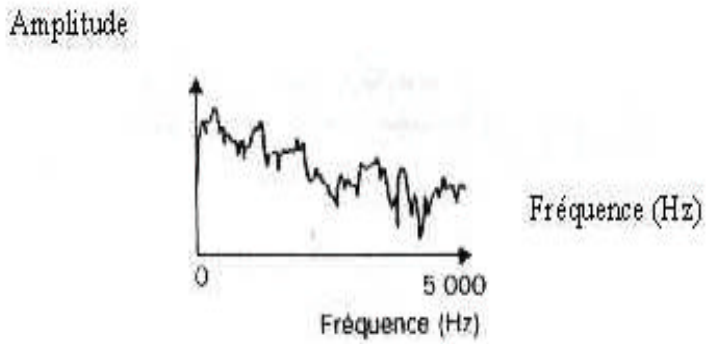
- Les méthodes spectrales :

Elles sont fondées sur la décomposition fréquentielle du signal sans tenir compte de sa structure fine.

La plus connue étant : **Fast Fourier Transform** (**FFT**). Tout son est la superposition de plusieurs ondes sinusoïdales. Grâce à la FFT , on peut isoler les différentes fréquences qui le composent.

La transformée de Fourier dite "à court terme", est obtenue en extrayant de l'audiogramme une trentaine de millisecondes de signal vocal et en effectuant un transformée de Fourier sur ces échantillons. Le résultat de cette transformation mathématique est souvent présenté dans un graphique qui donne, en fonction de la fréquence, l'amplitude des composantes présentes dans le

signal analysé.



En appliquant la FFT à un son complexe et en la répétant de nombreuses fois, on dresse un graphique donnant l'évolution de l'amplitude et de la fréquence en fonction du temps. **On obtient ainsi une empreinte caractéristique du son.**

- Les méthodes d'identification :

Elles reposent sur un modèle. Celui-ci possède un ensemble de paramètres numériques, dont les niveaux de variation représentent l'ensemble des signaux couverts par le modèle. Pour un signal et un modèle donné, l'analyse estime les paramètres du modèle pour lui faire correspondre le signal analysé. Un algorithme d'analyse cherche à minimiser la différence, appelée erreur de modélisation, entre le signal original et celui qui serait produit par le modèle s'il était utilisé en tant que synthétiseur .

Le modèle **prédictif linéaire** (**LPC : Linear Predictive Coding**) est le plus connu.
De la même façon que la parole naît du passage à travers notre conduit vocal d'un signal d'excitation créé par les poumons et les cordes vocales, elle peut être modélisé par le passage d'un signal d'excitation numérique à travers un filtre numérique récursif.

filtre récursif : la sortie dépend de l'entrée et de la valeur précédente de la sortie.

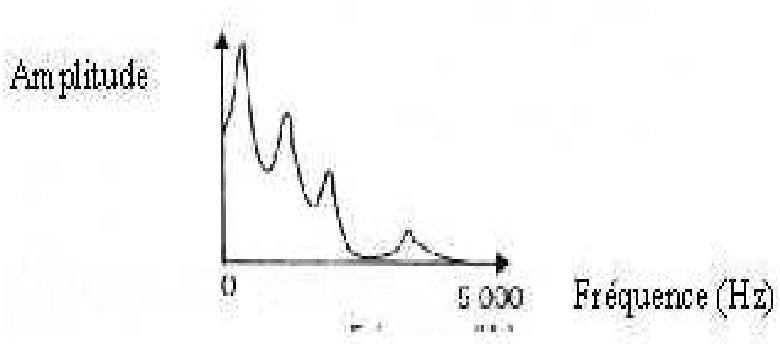
Le signal d'excitation sera soit :

- une suite d'impulsions numériques (qui serviront à simuler les impulsions de débit créées par les cordes vocales).
- du bruit numérique (qui reproduira le souffle poussé par les poumons).

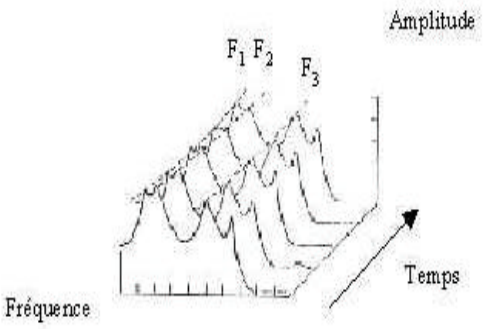
Ce modèle est appelé « prédictif linéaire » car il correspond à une régression linéaire très simple entre le signal d'excitation et le signal vocale produit. Les coefficients de cette régression linéaire sont les coefficients du filtre numérique récursif.

On repère alors facilement les **fréquences formantiques**.

fréquences formantiques : fréquences de résonance du conduit vocal.



En effet, elles correspondent au maximum d'énergie dans le spectre. En répétant cette méthode plusieurs fois, **on obtient l'empreinte du signal.**



D'autres méthodes existent, mais elles sont cependant moins employées. Les tendances actuelles visent à améliorer l'analyse fine des sons.

Le codage vectoriel permet de diminuer la quantité d'informations nécessaires pour coder un mot (et donc l'espace0. mémoire), en s'appuyant sur un dictionnaire de spectres instantanés .

Une fois que l'on a obtenu l'empreinte caractéristique du signal, on peut passer à l'étape suivante, qui est le décodage acoustico-phonétique ...

SOMMAIRE

4 . Décodage acoustico - phonétique

Il sert à décoder le signal acoustique en unités linguistiques (phonèmes, syllabes, les mots…).

phonème: élément sonore d'un langage donné, déterminé par les rapports qu'il entretient avec les autres sons de ce langage.

Par exemple, le mot " cou " est formé des phonèmes "keu " et " ou ". Il en existe une trentaine en français. Cette notion est assez importante en reconnaissance vocale.

1ère partie : Faire apparaître les segments du signal

1ère étape : segmenter le signal en segments élémentaires et étiqueter ces segments. Le principal problème est de choisir les unités sur lesquelles portera le décodage.

- Si des unités longues telles que les syllabes ou les mots sont choisies, la reconnaissance en elle-même sera facilitée mais leur identification est difficile.
- Si des unités courtes sont choisies, comme les phones (sons élémentaires), la localisation sera plus facile mais leur exploitation nécessitera de les assembler en unités plus larges.

Les phonèmes constituent un bon compromis, leur nombre est limité : ils sont donc souvent utilisés. Mais le choix dépend également du type de reconnaissance effectuée : mots isolés ou parole continue. Cela sera abordé plus loin.

2ème étape : identifier les différents segments en fonction de contraintes phonétiques, linguistiques… Il faut que le système ait intégré un certain nombre de connaissances : données articulatoires, sons du français, données phonétiques, prosodiques , syntaxiques , sémantiques …

Deux sortes d'outils sont utilisées :

- Les outils de reconnaissance de formes structurelle (ex : grammaires déterministes)
- Les outils provenant de systèmes experts (souvent associés pour de meilleures performances).Un système expert effectue les interprétations et déductions nécessaires grâce à la modélisation préalable du raisonnement de l'expert (domaine de l'intelligence artificielle).

Une fois que tout cela a été effectué, la reconnaissance en elle-même peut commencer, que ce soit pour des mots isolés ou pour de la parole continue…

2ème partie : Reconnaissance …

…des mots isolés

Retrouver les phonèmes et les mots dans un signal vocal est une réelle difficulté pour la reconnaissance vocale. De ce fait, séparer tous les mots prononcés par des silences permet de simplifier le problème.

a) Les techniques

Deux approches :

> Dans l'approche **globale**, l'unité de base est le **mot** (donc non décomposable). Cette méthode fournit une image **acoustique** de chaque mots à identifier et permet donc d'éviter l'influence mutuelle des sons à l'intérieur des mots. Elle se limite aux petits vocabulaires prononcés par un nombre restreint de locuteurs(les mots peuvent être prononcés de manière différente suivant le locuteur).

> L'approche **analytique**, qui tire parti de la structure des mots, identifie les composantes élémentaires (phonèmes, syllabes, …). Celles-ci sont les unités de base à reconnaître. Cette approche est plus générale que la précédente : pour reconnaître de grands vocabulaires, il suffit d'enregistrer dans la mémoire de la machine les principales caractéristiques des unités de base.

Pour la reconnaissance de mots isolés à **grand vocabulaire**, la méthode globale ne convient plus car la machine nécessiterait une mémoire et une puissance considérable pour respectivement stocker les images acoustiques de tous les mots du vocabulaire et comparer un mot inconnu à l'ensemble des mots du dictionnaire. Il est de plus impensable de faire dicter à l'utilisateur l'ensemble des mots que l'ordinateur a en mémoire.

C'est donc la méthode **analytique** qui est utilisée : les mots ne sont pas mémorisés dans leur intégralité, mais traités en tant que **suite de phonèmes** .

b) Principe général de la méthode globale et analytique

Le principe est le même que ce soit pour l'approche analytique ou l'approche global, ce qui différencie ces deux méthodes est l'entité à reconnaître : pour la première il s'agit du phonème, pour l'autre du mot.

On distingue deux phases:

- La **phase d'apprentissage** : un locuteur prononce l'ensemble du vocabulaire, souvent plusieurs fois, pour créer en machine le **dictionnaire de références** acoustiques. Pour l'approche analytique, l'ordinateur demande à l'utilisateur d'énoncer des phrases souvent dépourvues de toute signification, mais qui présentent l'intérêt de comporter des successions de phonèmes bien particuliers.

- La **phase de reconnaissance** : un locuteur prononce un mot du vocabulaire. Ensuite la reconnaissance du mot est un problème typique de reconnaissance de formes. Tout système de reconnaissance des formes comporte toujours les trois parties suivantes:

    - Un capteur permettant d'appréhender le phénomène physique considéré (dans notre cas un microphone),
    - Un étage de paramétrisation des formes (par exemple un analyseur spectral),
    - Un étage de décision chargé de classer une forme inconnue dans l'une des catégories possibles.

c) Choisir le mot

Le signal vocal paramétré est comparé aux mots du dictionnaire de référence. L'algorithme de reconnaissance permet de choisir le mot le plus ressemblant, en calculant le taux de similitude entre le mot prononcé et les diverses références.

Le programme va comparer le mot prononcé par le locuteur avec ceux qui sont en mémoire depuis l'apprentissage : la comparaison consiste à soustraire les nuances de gris des pixels du mot prononcé à ceux des mots en mémoire et de répéter cette opération pour chaque ligne et colonne. On pourra donc trouver, selon le résultat de la comparaison, le signal le plus ressemblant.

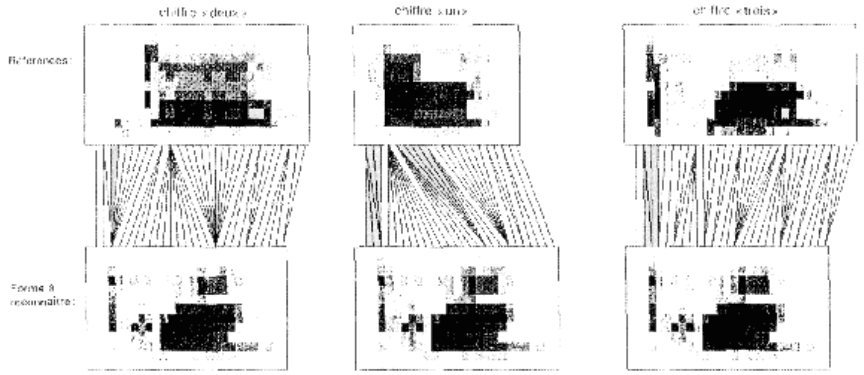
Inconvénients : Ce calcul n'est pas simple car les mots à comparer ont des durées et des rythmes

différents . En effet, un locuteur même entraîné ne peut prononcer plusieurs fois une même séquence vocale avec exactement le même rythme et la même durée. Les **échelles temporelles** de deux occurrences d'un même mot ne coïncident donc pas, et les formes acoustiques calculées lors de la paramétrisation ne peuvent pas être comparées point à point.

Solutions : il existe des solutions pour résoudre le problème de l'alignement temporel entre un mot inconnu et une référence. En voici trois principales :

- La modélisation sous forme de **modèles markoviens** (chapitre fondamental de la reconnaissance qui sera développée plus loin).

- Une très efficace est l' **algorithme de comparaison dynamique** qui va mettre en correspondance optimale les échelles temporelles des deux mots. On démontre que cette méthode fournit la solution optimale du problème. Elle nécessite cependant beaucoup de calculs. Pour fonctionner en temps réel, il faut donc soit réaliser des composants spécialisés de programmation dynamique (plusieurs firmes proposent des systèmes de reconnaissance intégrant un tel processeur), soit simplifier l'algorithme initial.



Comparaison de formes par programmation dynamique où l'action de l'algorithme est symbolisée par les traits entre chaque mot

La figure montre les correspondances effectuées par un algorithme de programmation dynamique entre une forme à reconnaître (le spectrogramme du chiffre " trois") et un vocabulaire de référence (ici les chiffres " un ", " deux ", "trois "). Le " trois " de référence est plus long (prononciation plus lente) que le " trois " à reconnaître ; l'algorithme assure une mise en correspondance optimale entre les vecteurs des spectrogrammes. En

revanche, la comparaison avec les formes de référence " un " et " deux ", très différentes de " trois ", est plus aléatoire. La représentation des mots est la suivante : horizontalement => le temps ; verticalement => les fréquences ; nuance de gris => l'intensité.

Les méthodes de comparaison par programmation dynamique ont été largement utilisées pour la reconnaissance de mots isolés. De plus, elles ont été étendues à la reconnaissance de séquences de mots enchaînés sans pause entre eux.

- Les **modèles neuromimétiques** qui sont constitués par l'interconnexion d'un très grand nombre de processeurs élémentaires fonctionnant comme le neurone.On parle de "neurone" car son fonctionnement est fondé sur celui d'un automate proposé comme une approximation du fonctionnement du neurone biologique

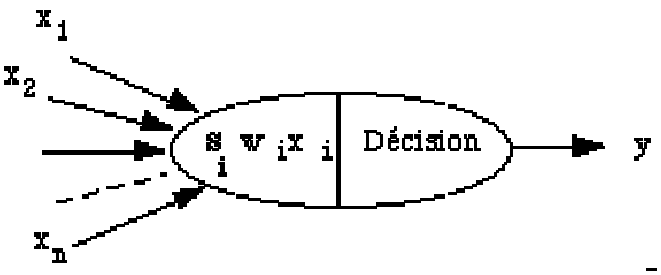
SOMMAIRE

5. Principe du neurone artificiel

Avec la méthode analytique, l'ordinateur procède identiquement pour décoder le message parlé (paramétrisation du signal, programmation dynamique, …) sauf que cette fois-ci il faut repérer une suite de phonèmes afin d'associer le mot au mot qui s'y rapporte dans le dictionnaire.

Les logiciels actuels utilises des dictionnaires de vocabulaire dépendant du type d'activité pour minimiser le taux d'erreur.

Dans une phrase, les mots s'enchaînent sans aucun moyen apparent de dissociation. Comment donc découper un signal afin de reconnaître les différents mots ou phonèmes qui le compose ? La notion de stratégie est lancée ...



… ou de la parole continue

a) Le premier niveau de stratégie : lire ou prédire ?

On distingue deux approches différentes. La première consiste à reconstituer la phrase à partir du signal en " lisant " tout simplement le signal (approche ascendante). On décrypte le résultat sans le comprendre ce qui nécessite de tester à chaque portion de phrase, l'ensemble des mots contenus dans le vocabulaire. Le vocabulaire peut très rapidement devenir gigantesque, et cela prend beaucoup de temps machine.

La deuxième approche consiste à prédire le mot à reconnaître (approche descendante) en implantant dans le système une certaine intelligence. Par exemple, si la machine reconnaît le mot " monsieur ", le vocabulaire testera est les noms de personnes contenus dans sa mémoire. Cette approche permet donc de ne pas tester tout le dictionnaire de la machine, et ainsi à gagner du temps.

On remarquera cependant qu'aucun système ne fonctionne en approche uniquement descendante, et rares sont ceux qui fonctionnent en approche uniquement ascendante. Seuls les systèmes à vocabulaire très restreint peuvent se permettre une approche uniquement ascendante.

b) Le deuxième niveau de stratégie : traitement de gauche à droite ou du milieu vers les cotés ?

L'analyse du signal peut s'effectuer dans différents sens. L'ordre chronologique reste le plus naturel (traitement gauche - droite). On peut aussi appliquer le traitement du milieu vers les cotés pour balayer le signal sans l'analyser complètement, afin de rechercher des mots - clés; on accentue la recherche de quelques mots du vocabulaire pour ainsi appliquer une stratégie descendante et combler les " trous ".

c) Le troisième niveau de stratégie : la recherche d'une solution optimale

On distingue deux grand types de stratégies.

-Les stratégies totales. Elles examinent toutes les solutions possibles. La machine teste tout son vocabulaire et attribue pour l'ensemble des phrases possibles un indice de probabilité de reconnaissance. Cette stratégie est applicable pour un vocabulaire très limité.

-Les stratégies heuristiques sont donc utilisées. Parmi les plus employées, on notera celle-ci :

- **Stratégie du meilleur d'abord** : A chaque analyse, le système ne retient que la solution offrant le meilleur score de probabilité. Elle est très simple à mettre en oeuvre, car elle n'effectue qu'une seule analyse à la fois. On gagne en temps de traitement, mais on perd en performance. Entre cette stratégie et une stratégie totale, il existe cependant un juste milieu.
- **Recherche en faisceau** ou **des " quelques meilleurs d'abord ":** Elle recherche en parallèle dans les différentes branches, les solutions plus probables et les conserve au fur et à mesure. Elle compare enfin les solutions partielles qui vont au même niveau de profondeur dans l'arbre de recherche. Cela est coûteux en temps machine, mais on se rapproche plus d'une solution optimale, car l'étendue des solutions explorées est plus vaste.
- **Recherche par îlots de confiance** : Dans les stratégies précédentes, une phrase est supposée analysée de la gauche vers la droite, en partant du début. Ici, on ne recherche que des mots - clés, dont la reconnaissance est quasi - certaine. On obtient donc une phrase à trous, avec ce qu'on appelle des îlots de confiance, dont on est sûr de la reconnaissance. On applique enfin une des stratégies précédentes pour découvrir ce qu'il y a entre.

SOMMAIRE

II. ZOOM SUR QUELQUES TECHNOLOGIES PERMETTANT LA RECONNAISSANCE VOCALE

1. Les automates stochastiques dans la reconnaissance

a. Définition du modèle

Les HMM sont définis par l'ensemble de données suivantes :

-Un automate de N états : 1, 2, …, N

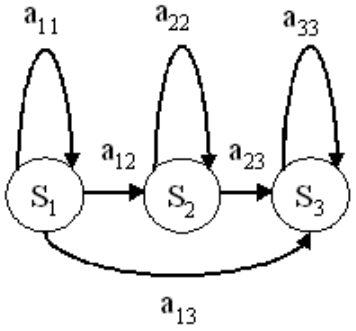
-un alphabet Y=(y_1, y_2, …, y_T)

-les probabilités a_{ij} associées à chacunes des transitions de l'état i vers l'état j.

-la probabilité b(m,i) pour l'automate d'émettre un symbole y_i lorsqu'il se trouve dans l'état m.

-les probabilités d(m) de trouver l'automate à l'instant t=0 dans l'état m :

$d(m)=Pr(s_0=m)$



La modélisation gauche-droite ci-dessus tient compte du caractère changeant du rythme des mots prononcés.
Les boucles sur les états modélisent un ralentissement possible du rythme et la transition a13 modélise le fait que le phonème représenté par l'état S2 puisse être dit rapidement et ainsi avalé lors de la phase de reconnaissance.

b. Identification de la séquence d'états ayant engendrée l'observation d'une séquence d'observables

On cherche à identifier la séquence d'états S=(s_1, s_2, …,s_T) ayant observé la séquence Y= (y_1, y_2, …,y_T)) et connaissant le modèle [a(m,m'), b(m,n), d(m)].

Algorithme de Viterbi

On pose :

$r_t(m) = \max p(s_0, …, s_{t-1}, s_t = m; y_0, …, y_t)$: (9)

² Initialisation :

$r_0(m) = d(m)b(m, y_0)$: (10)

² Récurrence :

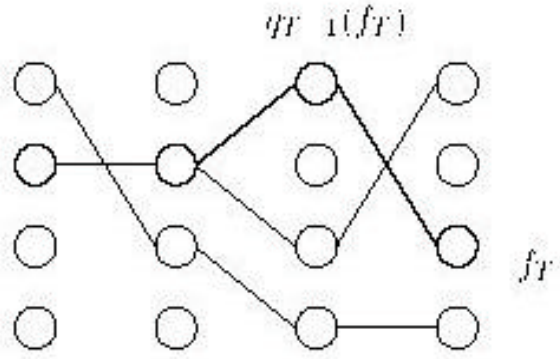
On suppose qu'à l'instant (t - 1) on a calculé $r_{t-1}(m)$ pour chacun des M états.
On a alors

$r_t(m') = \max r_{t-1}(m)a(m, m')b(m', y_y)$

L'état m le plus probable occupé à l'instant t -1 à partir duquel l'automate a évolué vers l'état M’ à l'instant t est l'état tel que rt¡1(m)a(m;m0)b(m0; yt) est maximum.

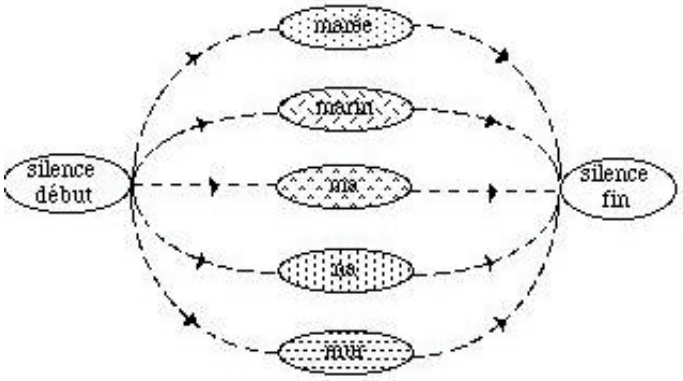
Si l’on mémorise le prédécesseur de chaque état m à l’instant t, alors il est enfantin d’en déduire la séquence d’état la plus susceptible d’avoir engendré la séquence d’observables Y

² Fin de l'algorithme :
L'état f_T retenu à l'instant T est celui pour lequel $r_T(m)$ est maximum. On effectue un chaînage arrière à partir de f_T en se servant du prédécesseur défini pour chacun des états à un instant t donné.



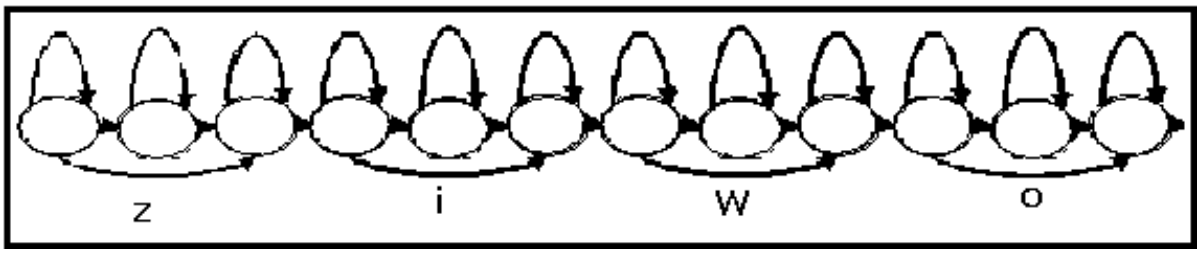
Dans le cas de la reconnaissance de mots isolés

Dans les applications de reconnaissance de mots isolés ne comportant qu'un vocabulaire modeste, une modélisation en mot est effectué. C'est à dire qu'un MMC est calculé pour chacun des mots du vocabulaire.
Le MMC global sur lequel va s'effectuer la recherche de la meilleure séquence d'état est réalisé en reliant l'entrée de chacun des MMC des mots à la sortie d'un MMC matérialisant un silence et reliant la sortie de chacun des MMC des mots à l'entrée d'un MMC modélisant le silence de fin.
Ceci est illustré par la figure ci-dessous :



Dans le cas de la parole continue

Dans ce type de reconnaissance, le vocabulaire est beaucoup trop large pour pouvoir calculer les MMC de l'ensemble des mots. Ceci nécessiterait une mémoire phénoménale ainsi qu'un corpus gigantesque pour la phase d'apprentissage
C'est pourquoi il est privilégié les modèles phonétiques qui permettent de modéliser des unités sonores beaucoup plus petite et à partir desquels on engendre tous les mots du dictionnaire par concaténation de ces modèles d'unité phonétique.



c. Apprentissage et obtention des paramètres caractérisant les MMC

Soit

$$\sigma_t(m', m) = p(s_{t-1} = m', s_t = m, [y_0, \ldots, y_T]),$$

on peut montrer que

$$\sigma_t(m', m) = \frac{\alpha_{t-1}(m')a(m', m)b(m, y_t)\beta_t(m)}{\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\alpha_{t-1}(m_1)a(m_1, m_2)b(m_2, y_t)\beta_t(m_2)}.$$

La probabilité

$$\lambda_t(m) = p(s_t = m, [y_0, \ldots, y_T]),$$

SOMMAIRE

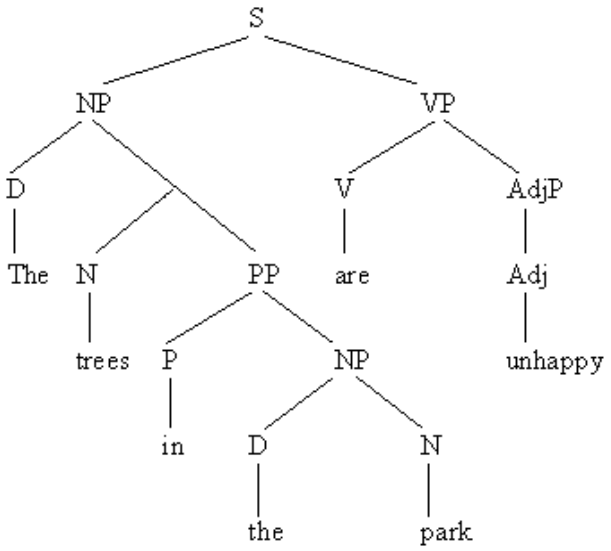
2. Grammaire décrivant la langue parlée

Dans une architecture d'un système à reconnaissance vocale, la partie implémentant une stratégie de recherche heuristique de la meilleure séquence de mot s'appelle un décodeur.

Il s'agit d'introduire des « règles contextuelles » afin de construire une grammaire propre au langage parlé.

La première étape consiste à associer des classes aux mots du lexique (le lexique étant l'ensemble des mots enregistrés dans la mémoire du système.

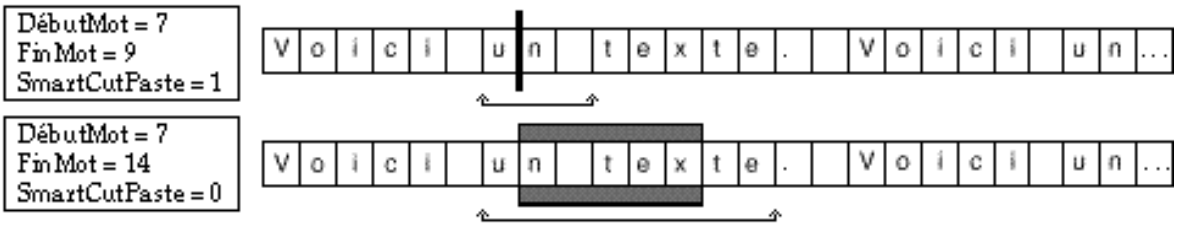
Par exemple, [enfant = pluriel] ou encore [noir = adjectif] ...



Exemple d'attribution de classes

Pour chaque variante de prononciation donnée, on lui applique un contexte, c'est à dire qu'elle n' existe dans une phrase que dans une situation bien particulière.

Pour cela, on utile la classe du mot précédent(celui qui a été reconnu dans l'analyse) et celle du mot suivant (le prochain mot qui sera reconnu).



Par exemple , la prononciation [lez] du mot [les] s'insère uniquement dans le contexte où le mot suivant est de classe « pluriel » où « voyelle initiale » .

Il es donc nécessaire de s'assurer que les contextes droits et gauches de la prononciation sont vérifiés. Cependant, il n'est évidemment pas possible de vérifier le contexte droit de l'hypothèse de prononciation puisque le mot suivant est encore inconnu.

En revanche il est possible de vérifier la compatibilité du contexte droit pour le mot précédent avec le contexte gauche de l'hypothèse courante, et, de manière symétrique, le contexte gauche de la règle courant avec le mot précédent.

Ainsi, toute absence de compatibilité rejetterait l'hypothèse courante.

Cet algorithme mémorise la règle qui a été utilisée pour le mot précédent afin d'en garder une trace. En effet, les hypothèses ayant pour prédécesseurs le même mot acoustique mais ayant utilisés des règles différentes sont considérées comme des mots différents puisqu'ils n'ont pas le même contexte droit.

Contextes dans le Système Sirocco

Pour tester l'apport des contraintes contextuelles, nous avons utilise des ressources MHATLex développées à l'IRIT.

MHATLex contient deux niveaux de représentation des transcriptions phonétiques:une représentation abstraite, dite phonotypique, qui condense un ensemble de représentation des transcriptions phonétiques valides dans un contexte linguistique défini.

Le passage d'un niveau a l'autre est opéré par application de règles

de réécriture, les transcriptions dérivées héritant héritant des contraintes contextuelles de leur ancêtre.

MHATLex inclut également diverses informations morpho_syntaxiques (lemme associe à une forme graphique, partie du discours, genre, nombre ...), desquelles ont été dérivés divers jeux de conception contextuelle.

Environ la moitié de ces classes se fonde sur des propriétés morpho-synthaxiques l'autre motif des classes encodant des propriétés phonologiques.

Pour contrôler plus finement l'effet de l'introduction de contraintes contextuelles, les règles dérivant les représentations phonétiques en fonction de phénomene phonologique sous-jacent.

Ceci conduit à marquer chaque transcription phonétique par l'ensemble des phénomènes linguistiques impliqués dans sa dérivation. une fois ce marquage construit, trois phénomènes agissant sur la frontière du mot sont étudiés:la liaison, les collisions ou réalisations de e-muets, et la chute des consonnes liquides finales, qui tous conduisent ˆ des transcriptions phonétiques dépendant de l'environnement linguistique dans lequel elles s'insèrent.

SOMMAIRE

CONCLUSION

Si les systèmes continuent d'évoluer comme ils l'ont fait au cours des dernières années, nul doute que plus personne ne pourra se passer de la reconnaissance vocale, car elle représente, en plus de tous les autres avantages, un confort de travail extraordinaire. La relève du clavier est assurée !

Nous avons cependant souvent du mal à nous faire comprendre et l'on peut se demander comment une machine, dont le principe de reconnaissance est fondé sur celui de l'homme, serait capable de faire mieux.

Les limites de la reconnaissance vocale risquent d'être rapidement atteintes.

SOMMAIRE