UMONS
Université de Mons

Faculté
des Sciences

# Traitement de la parole: Notes de cours

**Notes écrites par:** Anthony Rouneau
**Section:** 1$^{er}$ Bloc Master en Sciences Informatiques

Faculté des Sciences • Université de Mons • Place du Parc 20 • B-7000 Mons

# Part I
# Compétences requises pour l'examen

## 1 Introduction to speech

### 1.1 An Introductory Course on Speech Processing

- Dissocier le traitement de la parole du traitement du signal.
- Justifier par une raison valable la complexité du signal de parole.
- Citer les principales sciences et techniques concernées par le traitement de la parole.
- Identifier les domaines de recherche du traitement de la parole.

### 1.2 Acoustics

- Nommer les 7 profils de spécialistes travaillant en traitement de la parole.
- Expliquer en quoi consiste le travail d'analyse d'un acousticien.
- Définir ce qu'est un audiogramme, comment on l'obtient et préciser quels traits acoustiques il permet de mettre en évidence.
- Différencier sur un audiogramme les signaux voisés des signaux non voisés.
- Décrire le résultat d'une analyse de Fourier et préciser quels traits acoustiques elle permet de mettre en évidence.
- Caractériser ce qu'est un formant.
- Différencier formant et fondamental.
- Décrire le principe de construction d'un spectrogramme, préciser son utilité et son utilisation.
- Dessiner une courbe typique d'intonation et en expliquer l'utilité.

### 1.3 Phonetics

- Donner un diagramme schématique du conduit vocal et des cordes vocales et en décrire le fonctionnement.
- Citer et détailler les principaux modes de classification articulatoires en soulignant leurs différences de principes.
- Présenter par un exemple la classification par lieux d'articulation.
- Donner un exemple des caractéristiques complémentaires de sons exploitées en phonétique.
- Expliquer le principe qui a conduit à l'élaboration de l'alphabet phonétique international.
- Établir le lien entre la description acoustique et phonétique de la parole.
- Définir ce que l'on entend par « niveau de description segmental », par opposition au niveau « suprasegmental ».

- Citer les trois caractéristiques globalisées par le terme « prosodie ».

## 1.4 Phonology

- Différencier les niveaux de description de la parole dépendant de la langue des niveaux qui n'en dépendent pas.
- Donner la définition de phonème et le nombre de phonèmes dans la langue française.
- Donner un exemple d'allophone.
- Citer et définir sur base d'un exemple (à donner) la principale distinction entre phonème et son (ou phone).
- Commenter l'affirmation suivante : « Lorsqu'on connaît le fossé qui sépare la représentation acoustique de la parole de son niveau phonologique, la communication parlée tient du miracle ».
- Justifier l'expression « C'est du chinois ! » du point de vue de la phonologie.
- Définir la coarticulation et en justifier son impact en traitement de la parole en l'illustrant par un schéma.

## 1.5 Morphology

- Donner un ordre de grandeur du nombre de mots d'une langue naturelle (comme la langue française) et du nombre de mots utilisé dans le langage quotidien.
- Définir ce qu'est un morphème et en donner un exemple.
- Citer et expliquer les trois procédés morphologiques fondamentaux d'une langue naturelle.
- Illustrer par deux exemples le fait que chaque langue utilise à sa façon les trois procédés de transformation morphologique.

## 1.6 Syntax

- Donner le but premier de l'analyse syntaxique.
- Distinguer syntaxe et grammaire.
- Nommer le type de grammaires compréhensibles par un ordinateur (en donner un exemple) et la discipline scientifique qui en découle.
- Justifier pourquoi il est intéressant de réaliser une analyse syntaxique, au delà de la vérification de la validité syntaxique de la phrase.

## 1.7 Semantics & Pragmatics

- Opposer syntaxe et sémantique.
- Spécifier la base de construction des règles sémantiques.
- Nommer et décrire par un exemple un problème classique de la sémantique. Désigner en une phrase les caractéristiques de la parole concernées par la pragmatique.

### 1.8 From Source To Receiver

- Décrire le fonctionnement de l'oreille et schématiser ses principales parties.
- Justifier l'analogie entre la cochlée et un analyseur spectral.
- Justifier que la psychophysique ait permis de normaliser la transmission de parole sur les canaux téléphoniques et traduire cette justification par un schéma.
- Donner la définition des courbes isosoniques et les comparer aux courbes de seuil de l'audition et de seuil de la douleur.
- Décrire le phénomène de masquage auditif et donner un exemple d'impact technologique en traitement de la parole.

## 2 Modeling of the speech signal

### 2.1 Autoregresive modeling of the speech

- Distinguer les deux étapes du processus de mise en œuvre d'un nouveau modèle de la parole.
- Différencier erreurs de modélisation et erreur d'estimation, et montrer que la création d'un modèle résulte d'un compromis entre ces deux types d'erreurs.
- Citer et caractériser les trois grandes familles de modèles de la parole.
- Présenter le modèle autorégressif de traitement de la parole en réalisant l'analogie avec le système naturel de production de la parole.
- Donner un schéma de principe du modèle LPC de traitement de la parole, citer ses paramètres et caractériser leur rôle.

### 2.2 Autoregressive estimation of speech

- Expliquer pourquoi la modélisation autorégressive est basée sur un principe de minimisation d'erreur.
- Décrire le système d'équations de Yule-Walker, en expliquer les composantes et justifier sa facilité de résolution.
- Préciser l'avantage de l'algorithme de Schur sur celui de Levinson.
- Critiquer un choix de fréquence d'échantillonnage dans l'analyse d'un signal de parole.
- Donner une analyse critique du choix de l'ordre de prédiction pour modéliser la parole.

### 2.3 Extension of the AR model

- Expliquer comment on synthétise un signal de parole avec le modèle LPC.
- Préciser les éléments du modèle sur lesquels agir pour améliorer la qualité du signal modélisé.
- Citer deux types de problèmes typiques de l'analyse LPC.

- Définir ce qu'est un anti-formant et préciser son origine dans le système naturel de production de la parole.
- Préciser l'origine des sons mixtes dans le système naturel de production de la parole.
- Rappeler l'idée de base du modèle MP-LPC.
- Préciser comment estimer un modèle MP-LPC et expliquer la difficulté de cette estimation.
- Rappeler l'idée de base du modèle CELP et citer un problème de ce modèle.

# 3 Speech Coding

## 3.1 Speech Coding

- Citer les deux buts fondamentaux du codage d'un signal de parole.
- Expliquer pourquoi les algorithmes traditionnels de compression ne fonctionnent pas pour un fichier de parole.
- Donner l'ordre de grandeur du débit binaire minimal d'un signal de parole (en ne considérant que son contenu phonétique) et le comparer au débit binaire admissible par une ligne téléphonique.
- Décrire en les illustrant par un schéma, la technique de quantification uniforme et l'erreur de quantification qui lui est liée.
- Détailler l'emploi de la technique de quantification uniforme pour un CD audio et calculer le débit binaire correspondant.
- Détailler l'emploi de la technique de quantification uniforme pour le codage de parole.
- Expliquer le principe d'une compression selon la loi A et citer une technologie qui l'exploite.

## 3.2 Coders

- Schématiser un codeur par prédiction linéaire et commenter son fonctionnement.
- Citer la caractéristique typique de la parole codée en respectant la norme LPC10 et donner un exemple d'emploi dans la vie courante.
- Schématiser un codeur MP-LPC et commenter son fonctionnement.
- Justifier la dégradation de la parole lors d'une communication par GSM.
- Schématiser un codeur CELP et commenter son fonctionnement.
- Expliquer l'avantage que procure l'UMTS sur le GSM en ce qui concerne la transmission de la voix.

## 3.3 Conclusion

- Positionner sur un schéma qualité/débit les techniques de codage de parole.
- Citer une restriction importante à opérer sur le signal de parole traité, pour espérer un jour coder la parole avec un très bas débit (moins de 200 bits/s).

- Prédire l'avenir de la recherche en codage de parole et justifier cette prédiction.

# 4 Automatic Speech Recognition

## 4.1 Automatic Speech Recognition (ASR)

- Justifier, sur le plan économique, le développement des systèmes de reconnaissance de parole.
- Donner trois domaines d'application de la reconnaissance de parole, ainsi qu'un exemple pratique pour chacun de ces domaines.
- Expliquer pourquoi la coarticulation représente un important défi de la reconnaissance de parole.
- Citer la plus importante faiblesse des systèmes de reconnaissance de parole actuels.
- Citer et définir les deux principaux types de bruit, donner deux exemples par type.
- Donner une idée de ce qu'est l'effet Lombard.

## 4.2 ASR Systems

- Énoncer les contraintes qui régissent le choix d'un système de reconnaissance de parole.
- Positionner la complexité d'un système de reconnaissance de parole par rapport à un autre en fonction de leurs contraintes.
- Schématiser et commenter les principes de fonctionnement des trois grandes familles de reconnaisseurs de parole étudiés depuis les années 60.
- Expliquer quels types de connaissances embarque un système de reconnaissance basée sur des modèles statistiques.
- Préciser les contextes d'utilisations respectifs (en fonction des contraintes qui régissent le choix d'un système de reconnaissance) de la de la reconnaissance basée sur des modèles et de celle basée sur des modèles statistiques.

## 4.3 Instance-based ASR

- Préciser le problème de base d'un système de reconnaissance basé sur des exemples et présenter la solution généralement employée.
- Préciser la notion de distance locale et de distance globale dans le cadre d'un système de reconnaissance basé sur des exemples.
- Schématiser la solution la plus adéquate au problème du calcul de la distance globale dans le cadre de la reconnaissance basé sur des exemples, lorsque les mots à reconnaître sont longs.
- Expliquer conceptuellement l'algorithme DTW.

## 4.4 Model-based ASR

- Présenter la classification bayesienne et définir les termes de la règle de Bayes dans le cadre de la reconnaissance basée sur des modèles.
- Développer la règle de Bayes afin d'obtenir une expression exploitable de la probabilité à posteriori de prononciation d'une phrase, justifier chaque étape de ce développement.
- Citer les probabilités estimées par les modèles acoustique, phonétique et de la langue.

## 4.5 Acoustic Model : Markoov Chain

- Décrire le principe des chaînes de Markov.
- Expliquer pourquoi une chaîne de Markov ne permet pas de calculer directement une probabilité dans un modèle acoustique en reconnaissance de parole basée sur des exemples.

## 4.6 Acoustic Model : Hidden Markov Model (HMM)

- Décrire le fonctionnement des modèles de Markov cachés sur base d'un exemple.
- Différencier les probabilités d'émission et de transition dans les modèles de Markov cachés.
- Définir en quoi les modèles de Markov cachés sont-ils doublement stochastiques.
- Citer et définir les trois types de problèmes à traiter pour utiliser les modèles de Markov cachés.
- Citer les deux algorithmes utilisés pour résoudre le problème d'estimation des modèles de Markov cachés.
- Illustrer par un exemple l'entraînement d'un modèle de Markov caché.
- Énoncer le principe général de l'algorithme EM permettant l'entraînement des modèles de Markov cachés.

## 4.7 Phonetic Model

- Citer la technique utilisée dans les modèles phonétique pour calculer la probabilité d'une transcription phonétique étant donné une suite de mots.
- Donner un exemple illustrant les limites des modèles phonétiques.
- Préciser comment les modèles phonétiques intègrent un traitement individualisé pour chaque phonème et signaler les avantages de cette technique.

## 4.8 Language Model

- Citer et définir les trois problèmes des modèles de la langue.
- Exprimer de trois façons différentes la probabilité de prononcer une phrase donnée en fonction des mots qui la composent et commenter l'utilisation de ces expressions dans les modèles de la langue.

- Donner une valeur réaliste au nombre n du modèle n-gramme et préciser les conséquences liées à ce choix.
- Citer le problème du modèle n-gramme et proposer deux techniques pour le résoudre.
- Commenter la qualité actuelle des systèmes exploitant les modèles de la langue.

### 4.9 ASR Conclusion

- Positionner le taux d'erreur d'un système de reconnaissance de parole en fonction du type de reconnaissance, de la tâche, de la dépendance du locuteur, de la taille du vocabulaire.
- Expliquer la différence d'erreur de reconnaissance entre la lecture d'un journal en laboratoire et la parole de la vie réelle.
- Souligner l'importance du modèle de la langue en reconnaissance de parole.
- Citer les champs d'investigation actuellement explorés en reconnaissance de parole.
- Justifier qu'un être humain reste plus performant qu'un ordinateur pour reconnaître la parole.

## 5 Speech Synthesis

### 5.1 Text-to-Speech Synthesis

- Citer trois applications de la synthèse de parole en télécommunication.
- Positionner le besoin en interactivité pour les applications actuelles en téléphonie.
- Donner un exemple d'application de la synthèse de parole dans le domaine du multimédia.
- Justifier l'intérêt de la synthèse de parole pour la communiquer homme-machine.
- Préciser en quoi la synthèse de parole apporte une aide aux handicapés.
- Expliquer pourquoi la synthèse de parole est d'une importance fondamentale pour les expériences scientifiques sur le langage naturel.

### 5.2 TTS Diagram / Phonetization

- Schématiser un synthétiseur de parole TTS.
- Définir le rôle des linguistes informaticiens.
- Identifier la première difficulté de l'étape de phonétisation et présenter quatre exemples supplémentaires de difficultés illustrant la complexité de cette étape.

### 5.3 Intonation / Coarticulation

- Préciser l'utilité des fréquents mouvements intonatifs dans le langage naturel.
- Préciser le problème rencontré lors de la modification artificielle de la courbe intonative.
- Citer les deux autres rôles majeurs de l'intonation.

- Caractériser les durées des phonèmes.
- Citer les connaissances nécessaires pour appliquer une intonation à une phrase.
- Justifier que la synthèse de parole tente de reproduire la coarticulation.
- Résumer les défis et contraintes auxquels doit répondre la synthèse de parole.

## 5.4 TTS Technique

- Résumer le principe de fonctionnement de la machine de Von Kempelen.
- Résumer le principe de fonctionnement du Voder d'Omer Dudley.
- Exposer l'idée de la synthèse par formants et justifier son nom.
- Caractériser la parole générée lors d'une synthèse par formants.

## 5.5 Diphone / Unit selection -based synthesis

- Énoncer le principe de fonctionnement de la synthèse par diphone.
- Préciser comment la synthèse par diphone respecte la coarticulation du langage naturel.
- Schématiser le fonctionnement d'un synthétiseur de parole par diphone et préciser sur ce schéma les deux principaux problèmes du modèle.
- Commenter le développement du projet MBROLA.
- Citer la différence entre la synthèse par sélection d'unité et la synthèse par diphone.
- Citer le problème qu'il restait à résoudre en synthèse par sélection d'unité.

## 5.6 Pre-processing / Morphological analysis / Contextual analysis

- Citer les sous-modules d'un module de traitement du langage naturel dans un système de synthèse de parole.
- Citer les rôles du prétraitement dans un système de synthèse de parole.
- Citer le type d'outils généralement utilisés pour résoudre une grande partie des problèmes de prétraitement en synthèse de parole.
- Justifier l'utilité de l'analyse morphologique en synthèse de parole.
- Préciser comment est réalisée l'analyse morphologique en synthèse de parole.
- Énoncer le but de l'analyse contextuelle et exposer la méthode pour y parvenir.

## 5.7 Syntactic-Prosodic Phrasing / Automatic Phonetization

- Offrir deux solutions pour identifier les groupes de mots d'une phrase en synthèse de parole.
- Spécifier la méthode généralement employée pour réaliser la phonétisation automatique des mots en synthèse de parole et illustrer cette méthode par un exemple.
- Présenter l'alternative à la méthode habituelle de phonétisation automatique et indiquer le point commun entre les deux méthodes.

### 5.8 Prosody generation

- Commenter les premiers efforts réalisés pour générer la prosodie en synthèse de parole.
- Définir ce qu'est un ton et spécifier son utilisation pour caractériser un corpus.
- Déterminer comment générer des tons sur base d'un texte.
- Déterminer comment générer par règles une courbe intonative sur base de tons.
- Déterminer comment générer par entraînement une courbe intonative sur base de tons.
- Établir l'analogie entre la génération de prosodie par entraînement et la synthèse par sélection d'unité et l'illustrer en schématisant un exemple.

### 5.9 TTS Conclusion

- Résumer la tendance suivie depuis 1995 en synthèse de parole et donner une justification technique à cette tendance.

## 6 Conclusion on Speech Processing

- Positionner l'évolution du codage de parole à l'heure actuelle.
- Positionner l'évolution de la reconnaissance de parole à l'heure actuelle.
- Positionner l'évolution de la synthèse de parole à l'heure actuelle et la comparer à celle de la reconnaissance de parole.
- Souligner l'importance qu'ont prise les grandes bases de données de parole dans le domaine du traitement de parole.
- Prédire l'avenir des scientifiques de la parole.
- Émettre un avis critique sur le traitement de la parole.

# Part II
# Matière du cours

## 1 Introduction to speech

### 1.1 Notes du présentiel

#### 1.1.1 Introduction

**Lancement de la recherche**
$60^s$ : FFT $\rightarrow$ $80^s$ : Télécom lancent la recherche par espoir d'aboutir vite à un résultat.
$90^s$ : Second souffle de la recherche avec des PME qui se lancent dans la recherche, avec par exemple Lernout & Hauspre en Belgique (la plus grosse entreprise en TDP) qui est dissoute en 2001 suite à des fraudes fiscales. Les employés se sont réfugiés dans Nuance (US).

**Avancement de la recherche**
Récemment, depuis 2013, les DNN (réseaux de neurones profonds) permettent d'envisager qu'un ordinateur ne fasse pas que comprendre les mots mais en comprenne le sens profond.

#### 1.1.2 Rappel : traitement du signal

**Échantillonngage**
Pour pouvoir traiter le signal de la parole, il faut tout d'abord échantillonner le signal électrique typiquement amené par un micro. Ceci dépend d'une fréquence d'échantillonnage.

$$F_e > 2 \cdot F_{max}$$

En effet, pour qu'une sinusoïde soit correctement numérisée, il faut prendre un point toutes les demi périodes. (On va échantillonner une fois à 1, une fois à -1 afin de reconstruire le signal)

**Quantification**
Transformer l'intensité du signal en valeur discrète $\rightarrow$ codage de l'intensité du signal par des paliers (voir 3.1).

**Signal de parole** Typiquement, $F_e = 10000Hz$ et Quantification sur 16bits.

**Transformée de Fourrier**
Permet de décomposer un signal complexe en une multitude de signaux simple de fréquences et d'amplitude différente. Fourrier nous dit que la somme de tous ces signaux simples reconstruit le signal de départ.

On la représente donc en amplitude en fonction de la fréquence afin de comprendre l'amplitude des sinusoïdes qui se cachent dans le signal complexe.
On distingue deux types de signaux :

- Périodique : sa T.F. se rapproche de points discrets de toutes les sinusoïdes qui le compose.
- Non-périodique : sa T.F. est une fonction non-discrète.

### Système linéaire

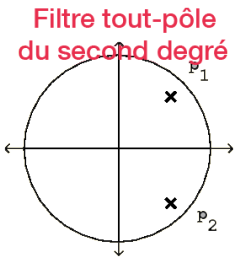
C'est un filtre du type : x(t) $\rightarrow$ système $\rightarrow$ y(t).
On parle de système linéaire lorsque :

$$\alpha \cdot x_1(t) + \beta \cdot x_2(t) \rightarrow Syst.Lin. \rightarrow \alpha \cdot y_1(t) + \beta \cdot y_2(t)$$

La caractéristique des systèmes linéaires est que ceux-ci peuvent être représentés par une réponse en fréquence, calculable sur base d'une entrée et sortie de ce filtre. Et quelque soit l'entrée et la sortie de ce filtre, la réponse en fréquence sera la même.

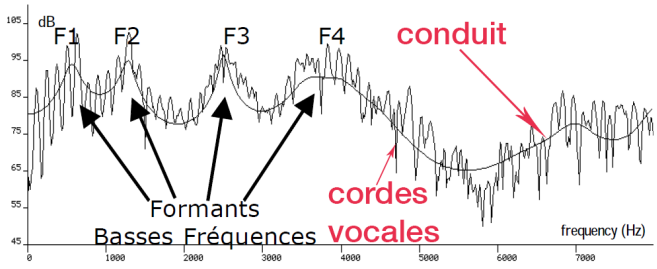
### Comprendre un filtre

Quand on se promène sur le cercle unitaire, la fréquence résonne quand on s'approche d'un pôle




### 1.1.3 Autre

### Description haut niveau d'un formant




On peut donc en dire que la partie conduit vocal ne change pas si on ne change pas de lettre. Un "a" aigu aura les même formants qu'un "a" grave. Ce qui change entre un "a" aigu et un "a" grave c'est le rapprochement des sinusoïdes formées par les cordes vocales (=> fréquence produite par conduit vocal).

## 1.2 An Introductory Course on Speech Processing

**Dissocier le traitement de la parole du traitement du signal**

- Les signaux telecom (étudiés par le traitement du signal) sont les signaux les plus simples; ils sont créés par des machines d'ingénieur et sont reçus par des machines construites par des ingénieurs également.
- La parole est un signal créé par un cerveau et ce signal est fait pour être compris par un cerveau également. Les signaux biologiques étaient là bien avant les ingénieurs, donc plutôt que de le créer/modifier, on va essayer de le comprendre.

**Justifier par une raison valable la complexité du signal de parole**

- Le cerveau humain est plus complexe que n'importe quelle machine crée par l'homme, et la complexité d'un signal est fonction de la complexité de l'émetteur et de la complexité du receveur

**Citer les principales sciences et techniques concernées par le traitement de la parole**

- Informatique : Les outils utilisés pour traiter la parole.
- Mathématique : Mathématisation très forte de la parole.
- Linguistique et linguistique informatique : Le signal reçu fait parti d'une langue.
- Ingénierie : Il faut construire des machines capable de comprendre/traiter ce signal.

**Identifier les domaines de recherche du traitement de la parole**

- Alphabet phonétique international : On travail toujours sur la précision de cet alphabet.
- Codage de la parole : Comment encoder un signal de voix sur l'ordinateur?
- Synthèse vocale : Créer de la voix par ordinateur.
- Reconnaissance vocale : Comprendre la voix humaine sur ordinateur.

## 1.3 Acoustics

Acoustique : comprendre la physique derrière le signal de la parole

**Nommer les 7 profils de spécialistes travaillant en traitement de la parole.**

- Indépendant de la langue :
  - Acousticiens
  - Phonéticien
- Dépendant de la langue :
  - Phonologie
  - Morphologie
  - Syntaxe
  - Sémantique
  - Pragmatique

**Expliquer en quoi consiste le travail d'analyse d'un acousticien.**

- Leur travail consiste à analyser un signal acoustique, et donc, ils en analysent :
  - La fréquence fondamentale du signal.
  - L'amplitude du signal.
  - Le spectre de fréquence du signal.

**Définir ce qu'est un audiogramme, comment on l'obtient et préciser quels traits acoustiques il permet de mettre en évidence**

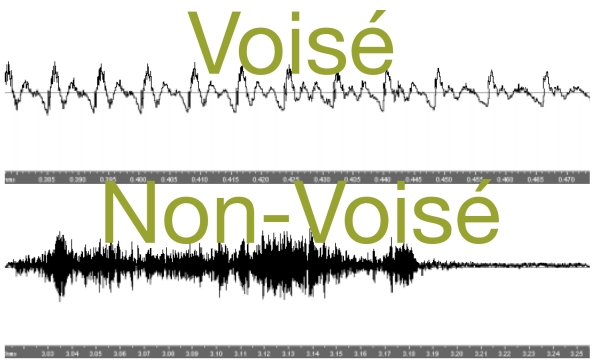
- Courbe qui représente en fonction du temps l'intensité du son en fonction du temps et il permet de calculer les valeurs citées précédemment (la fréquence fondamentale du signal, l'amplitude du signal, le spectre de fréquence du signal).
  On y représente donc la variation de la pression par rapport à la pression acoustique.
- Il permet de mettre en évidence les variations de pression causés par le conduit vocal.
  On peut noter que sans l'aide des phonèmes (les sons produits par le conduit vocal), aucun oeil humain ne peut déchiffrer tel quel un audiogramme.

**Différencier sur un audiogramme les signaux voisés des signaux non voisés**

- Signaux voisés : Signal à structure périodique faisant intervenir les cordes vocales.
- Signaux non-voisés : Signal ne faisant pas intervenir les cordes vocales.

On peut donc les distinguer sur un audiogramme car les signaux voisés montrent une structure plus ou moins périodique.

En tout cas, on a l'impression qu'il est périodique mais en réalité, les nombres représentant le signal ne permettent pas de retrouver une période à ce signal, c'est notre cerveau et son intelligence qui permet d'imaginer une périodicité (d'environ 10 ms) à ce signal.




**Décrire le résultat d'une analyse de Fourier et préciser quels traits acoustiques elle permet de mettre en évidence**

- L'analyse de Fourier prend en entrée un signal (30ms en pratique) et donne en sortie les signaux formants du signal.
- Une analyse de Fourier sur un signal voisé met en évidence la fréquence fondamentale et ses harmoniques (se trouvant à des fréquences multiples de la fondamentale) et ce pour chaque formant du signal.
- Une analyse de Fourier sur un signal non-voisé ne montre pas de fréquence fondamentale, (et donc pas d'harmoniques) mais seulement des pics d'amplitude qui sont les formants.

**Caractériser ce qu'est un formant**

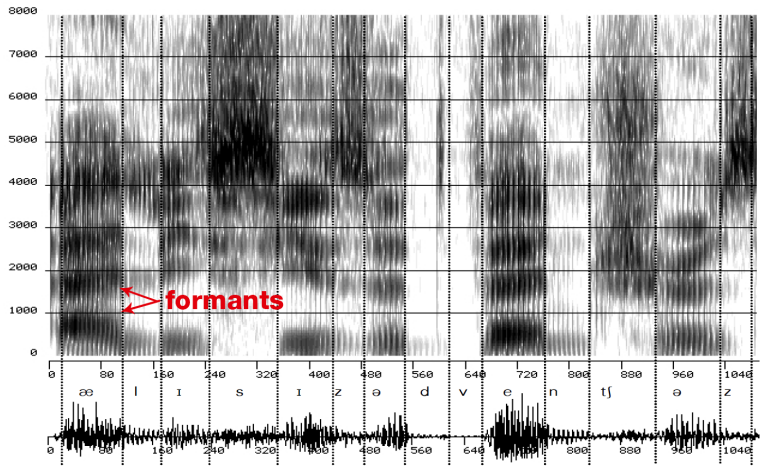
- Formant : Résonance de l'enveloppe spectrale du signal vocal. Il(s) change(nt) lorsque le conduit vocal change de forme.
- Lorsqu'on garde la même voyelle, le conduit vocal ne change pas de forme, et l'enveloppe spectrale reste donc identique : les formants ne bougent pas.
- La position des formants indiquent le son prononcé.

**Différencier formant et fondamental**

- Fondamentale : La fondamentale est un signal qui se répète en harmonique. Lorsque l'on change la note sur une même voyelle, les formants ne bougent pas mais la fondamentale et ses harmoniques oui.
- Formant : Un formant étant une fréquence de résonance de l'enveloppe spectrale d'un son, un signal sinusoïdal ou quasi-sinusoïdal ne définit pas de formant.

## Décrire le principe de construction d'un spectrogramme, préciser son utilité et son utilisation

- Un spectrogramme permet de combiner les informations de l'analyse de Fourier et de l'audiogramme. Il consiste en des analyses spectrales par fenêtres de temps.
- Il est constitué de 3 dimensions. On reprend la fréquence, il devient l'axe vertical, l'axe horizontal étant le temps (qui n'était pas présent dans une analyse spectrale, étant donné qu'on se concentrait sur une petite portion de temps). La troisième dimension est le niveau de gris représentant l'amplitude du signal.
- On multiplie le signal par des fenêtres de temps qui se recouvrent un peu (pas forcément des fenêtres linéaire, souvent des courbes genre Gaussienne aplatie).

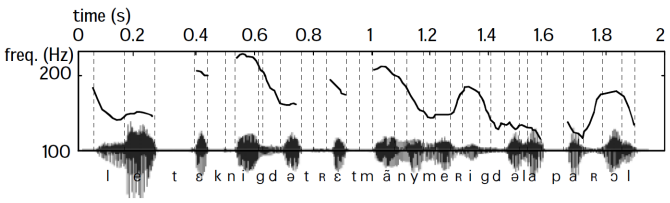



- Le spectrogramme est le premier outil qui permet une analyse à l'oeil nu. Avec beaucoup d'expérience on pourrait en déduire directement les sons émis. On ne peut par contre toujours pas connaitre l'intonation avec laquelle les sons ont été émis.

## Dessiner une courbe typique d'intonation et en expliquer l'utilité

- Courbe de pitch : Courbe représentant le déplacement de la fondamentale en fréquence des cordes vocales du conduit vocal. Cette courbe permet de regrouper les paquets de sons qui vont ensemble. Sans cette variation de fréquence, on aurait du mal à savoir quand un mot se finit et que le prochain commence.

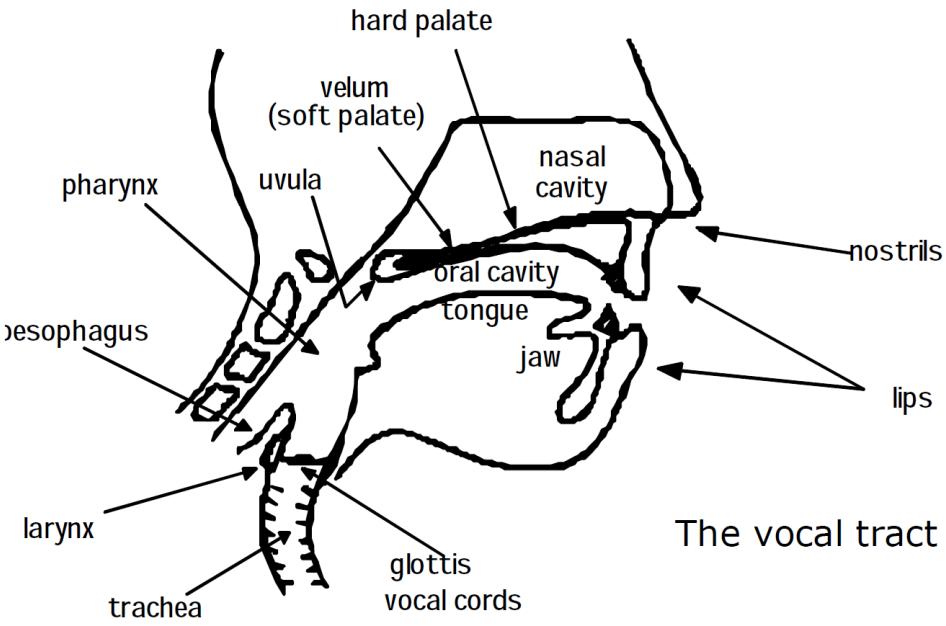
"les techniques de traitement numériques de la parole"



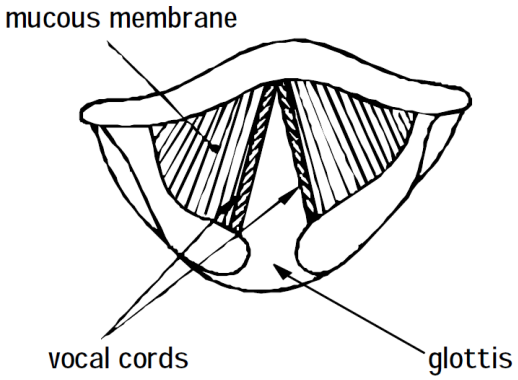
## 1.4 Phonectics

Phonétique : différencier les sons présents dans les langues

**Donner un diagramme schématique du conduit vocal et des cordes vocales et en décrire le fonctionnement.**

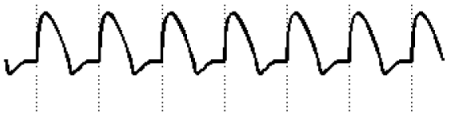



The vocal tract




The vocal folds

1. Membranes tendues par les muscles qui tendent les cartilages autours d'elles.
2. Comme les cordes vocales sont fermées, la pression augmente du à l'afflux d'air des poumons.
3. Ouverture due à la pression.
4. Perte de pression => son brut émis (modifié par la mâchoire, le larynx, ...) .
5. Fermeture et retour à l'étape 2. (=> périodique)



Plus les muscles sont tendus, plus la fréquence sera faible, et donc plus la période sera grande.

**Citer et détailler les principaux modes de classification articulatoires en soulignant leurs différences de principes.**

- Mode voyelle : L'air passe sans encombre, sans obstacle particulier.
  - Mode voyelle nasale : utilise les cavités nasales.
  - Mode voyelle orale : utilise la cavité orale uniquement.
- Mode consonne : L'air passe par une constriction (obstruction respiratoire en un point d'articulation donné) dans le canal vocal.
  - Mode consonne nasale : Obstruction totale.
  - Mode consonne fricative : Blocage partiel quelque part. $< fff > < sss >$
  - Mode consonne plosives : Son tenu suivi d'un blocage total suivi d'une "explosion" $< parenTHèse >$
  - Mode consonne liquide : Sons caractérisés par une vibration basse fréquence, soit des lèvres soit de la glotte.
- Mode semi-voyelle : Pas d'obstacle important mais pas tenu $< ye > < we >$

**Présenter par un exemple la classification par lieux d'articulation**

- On peut parler de voyelle avant, arrière ou centrale. (passage de $\acute{e} \rightarrow o$).
  On change de lieu selon l'espace qu'on laisse au son pour qu'il se développe.
- On peut changer le lieu d'obstruction des consonnes (h aspiré + bouger sa glotte...)

**Donner un exemple des caractéristiques complémentaires de sons exploitées en phonétique**

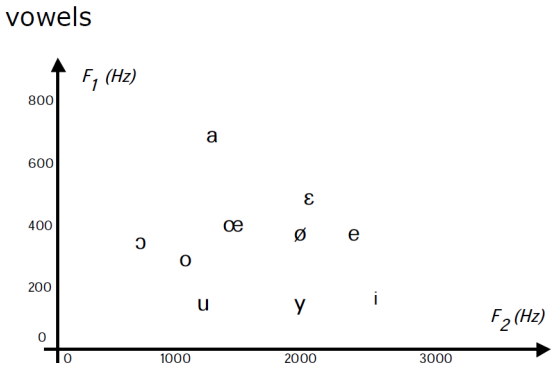
- Ouverture des voyelles : $\acute{e}$ = fermé, $\grave{e}$ = ouvert.
- Voyelles tendues/breathée = $a$ motivé / $a$ pas motivé
- Plosives apirées/non aspirées = $pfff$ ou $p$

**Expliquer le principe qui a conduit à l'élaboration de l'alphabet phonétique international**

- Chaque son peut être caractérisé par une suite binaire de caractéristiques phonétiques.
  (Ex: oral = 1, nasal = 0, ou encore labial = 1, non-labial = 0, ...).
  Chaque suite binaire donne une lettre de l'alphabet phonétique.

**Établir le lien entre la description acoustique et phonétique de la parole**

- Description acoustique : Analyse en fréquence, en formants.
- Description phonétique : Analyse en lettre de l'alphabet phonétique.




On peut remarquer qu'on peut déterminer la voyelle à laquelle nous avons affaire simplement en mettant en graphique la relation entre les fréquences des deux premiers formants. (On aurait pu établir une troisième dimension pour le troisième formant).

En fait, les cordes vocales, pour produire une voyelle, se configure dans une forme particulière, Ce qui donne des fréquences de résonance: les formants des acousticiens.
On en déduit donc que les phonéticiens ont donné une description phonétique du signal de parole.

**Définir ce que l'on entend par « niveau de description segmental », par opposition au niveau « suprasegmental »**

- Description Segmental : description en fonction du mode et du lieu d'articulation.
- Description Suprasegmentale : description de l'intonation.

**Citer les trois caractéristiques globalisées par le terme « prosodie »**

- Description Suprasegmental (prosody) : Description en fonction du pitch, durée et intensité du son. Il n'y a pas d'accord des phonéticiens à ce niveau.

## 1.5 Phonology

Phonologie : différencier les sons entendus et ce qu'on a voulu dire.

**Différencier les niveaux de description de la parole dépendant de la langue des niveaux qui n'en dépendent pas.**

- Acoustique et Phonétique sont indépendants de la langue, tandis que la Phonologie, la Morphologie, la Syntaxe, la Sémantique et les Pragmatiques en sont dépendants.
- On passe d'une analyse objective de la parole (niveaux indépendant de la langue) à une analyse presque subjective de la langue (niveaux dépendants de la langue).

**Donner la définition de phonème et le nombre de phonèmes dans la langue française.**

- Phonème : Son abstrait dans le cerveau de la personne qui parle, et chaque son s'oppose à tous les autres, non pas par ce qui est dit, mais par leur intention. Ils représentent donc les sons que le cerveau imagine avant de les envoyer au canal vocal. Il y a 36 phonèmes dans la langue française.

**Donner un exemple d'allophone.**

- Allophone : Plusieurs sons à la phonétique différente (e.g. $R$ normal ou $R$ roulé) pour un même phonème.
  - Les variantes géographiques : Si un auvergnat roule ses $R$, ce n'est pas voulu, ça ne change pas le sens du mot pour autant.
  - La coarticulation : On est constamment en train de gérer un transitoire dans les cordes vocales pour faire sortir un son voulu, mais qui ne sera jamais le bon en réalité (car son trop compliqué ou contraintes physiques). Exemple: entre ann$ue$l et act$ue$l, notre cerveau voudrait faire sortir le même phonème pour les deux mots, mais il en sort deux sons phonétiquement différents, on "siffle" le u de actuel, et pas celui de annuel.

**Citer et définir sur base d'un exemple (à donner) la principale distinction entre phonème et son (ou phone).**

- Un phonème est définit par une intention de faire parvenir un son par le canal vocal.
- Le son est défini par le son réel qui sort du canal vocal.
- La différence entre les deux s'explique par le principe de coarticulation. (e.g. ann$ue$l/act$ue$l)

**Commenter l'affirmation suivante : « Lorsqu'on connaît le fossé qui sépare la représentation acoustique de la parole de son niveau phonologique, la communication parlée tient du miracle ».**

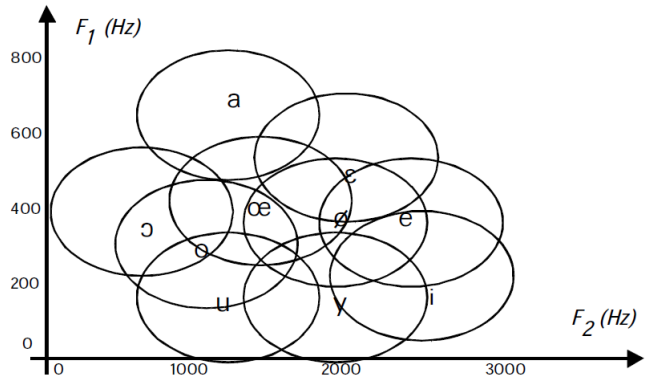
- Le miracle c'est que malgré que le mauvais son sorte, celui qui écoute est persuadé que le bon son est sorti, il a donc compris le phonème alors que le son n'y correspondait pas forcément.

**Justifier l'expression « C'est du chinois ! » du point de vue de la phonologie.**

- Le chinois est une langue à ton, c'est à dire que l'intonation a un sens dans la langue. chaque intonation désigne un mot différent, apporte un information particulière.
- Il est assez compliqué d'apprendre une langue à ton quand on en parle pas une car depuis l'enfance, le cerveau a appris à ne pas faire de distinction de phonème entre deux intonations différentes.

**Définir la coarticulation et en justifier son impact en traitement de la parole en l'illustrant par un schéma.**

- La coarticulation : Phénomène qui est du au fait que le conduit vocal est un système physique qui possède une inertie qui lui est propre, et donc, il est impossible à un être humain de passer infiniment rapidement d'un son à un autre.
- La coarticulation implique que chaque son émis est différent, même s'il est sensé désigner un même phonème.




- La synthèse vocale est affectée par la coarticulation parce qu'il faut être capable de l'imiter. En effet, si on ne respecte pas la coarticulation, les gens ont beaucoup de mal à comprendre ce qui est dit.
- La reconnaissance vocal est affectée par la coarticulation parce qu'il faut pouvoir la surmonter, et savoir quel phonème était voulu par rapport au son émis.

## 1.6 Morphology

Morphologie : étude de l'assemblage de sons pour faire un mot.

**Donner un ordre de grandeur du nombre de mots d'une langue naturelle (comme la langue française) et du nombre de mots utilisé dans le langage quotidien**

- L'ordre de grandeur est de 50 000 mots dans le dictionnaire, et en y ajoutant les noms masculin/féminin/pluriels, on tombe sur 500 000 voir 1 000 000 de mots.
- On utilise dans le langage quotidien environ 5000 mots.

**Définir ce qu'est un morphème et en donner un exemple**

- Morpheme : Unité abstraite qui compose les mots et qui apporte, chacune, une information de sens.
- Ces unités sont programmées dans le cerveau, encodées sous forme de mots qui respectent un certain nombre de règles (règles de morphologie de la langue).

  Exemple : went = $go$ + $past$, visible = $see$ + $able$.

**Citer et expliquer les trois procédés (règles) morphologiques fondamentaux d'une langue naturelle**

- Flexion : Entrée : un morphème lexicale + un morphème grammatical et donne un nouveau mot qui a la même nature que le morphème lexical de départ. (conjugaison des verbe, passage au pluriel, ...)
- Dérivation : Entrée : morphème lexical + morphème grammatical et donne un nouveau mot qui n'a pas la même nature que le morphème lexical de départ (image $\rightarrow$ imaginer, imagination, ...).
- Composition morphologique : Entrée : plusieurs morphèmes lexicaux et les composer en un nouveau mot (rouge-gorge, sous-marin, ...).

**Illustrer par deux exemples le fait que chaque langue utilise à sa façon les trois procédés de transformation morphologique**

- La flexion des verbes en français fait apparaitre 41 formes verbales, alors qu'en anglais on en trouve que 8. Donc la façon dont on utilise la morphologie de la langue dépend de la complexité de formation des mots de la langue.

## 1.7 Syntax

Syntaxe : étude de l'ordre dans lequel les mots peuvent se retrouver (sujet -> verbe -> complément)

**Donner le but premier de l'analyse syntaxique**

- Le but est de faire le tri parmi les suite de mots de la langue, les suites de mots possibles et les suites de mots impossibles.

**Distinguer syntaxe et grammaire**

- Syntaxe : Contraintes abstraites de la validité d'une suite de mot.
- Grammaire : Formalisation de la syntaxe à travers une grammaire (type de mots, ...). Elles sont crées par des grammairiens, spécialistes de la langue. Chaque spécialiste a sa façon de voir les choses (grammaire historique vs grammaire d'apprentissage).

**Nommer le type de grammaires compréhensibles par un ordinateur (en donner un exemple) et la discipline scientifique qui en découle**

- Les grammaires de langue que l'on apprend à l'école sont des grammaires spéciales : elles ont pour but de formaliser une langue qui est déjà parlée par l'étudiant.

- Les grammaires utilisables par l'ordinateur sont donc des grammaires qui ne supposent aucune connaissance préalable de la langue. Il faut donc lui donner des listes de mots que le langage va utiliser.
  Ex : les grammaires syntagmatiques (phrase = group_nom + verb_conj, group_nom = ...).
- Ça a donné naissance à la linguistique informatique dans les années 50.

**Justifier pourquoi il est intéressant de réaliser une analyse syntaxique, au delà de la vérification de la validité syntaxique de la phrase**

- Ça permet de former des groupes de mots (groupe nominal, groupe sujet, ...) et ça va beaucoup aider la synthèse et la reconnaissance vocale. La structuration syntaxique va donc permettre d'avoir une information sur le sens de la phrase (Exemple: [Time] [flies] [like an arrow] ou [Time flies] [like] [an arrow]).

## 1.8 Semantics & Pragmatics

Sémantique : étude liée à l'ordre des mots avec leur sens (lier l'objet qui se cache derrière le mot avec d'autres mots). Exemple: un concept abstrait n'a pas de couleur (la richesse jaune ?).
Pragmatique : étude des sens caché dans les phrases (concepts n'ayant pas besoin d'être précisés pour être compris par un autre humain).

**Opposer syntaxe et sémantique**

- Sémantique : Faire le tri entre les suites de mots syntaxiquement correctes qui veulent dire quelque chose et celles qui ne veulent rien dire.
  Cependant, comme la syntaxe, c'est une idée abstraites qu'il faudra formaliser en grammaire.

**Spécifier la base de construction des règles sémantiques**

- La base des règles sémantiques sont les traits sémantiques . Chaque mot a un ensemble de traits sémantiques et dans les règles de ces grammaires sémantiques, il y a des implication dans la concordance entre les traits sémantiques des mots successifs.
  Exemples de traits : mot abstrait, adjectif de couleur, ...
- On peut en conclure pas mal de choses. Ex : un mot abstrait ne peut pas être suivi d'un adjectif de couleur (politesse jaune ?).
- Un des problème qui se pose est : "comment définir un lexique informatique qui permet de choisir un trait parmi les traits sémantiques". Une solution peut être de lister de façon exhaustive tous les traits sémantiques liés à un vocabulaire (ce qui n'est pas envisageable dans le cas de la langue entière).

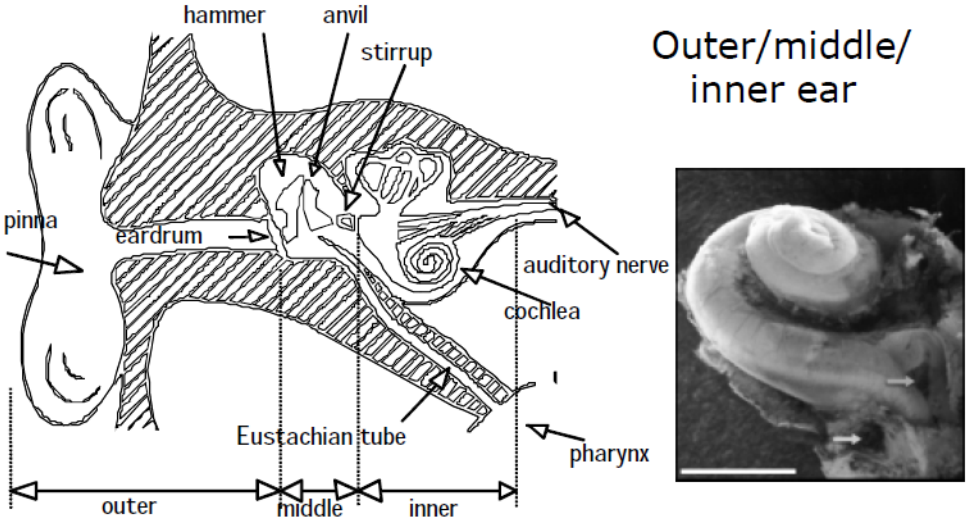
**Nommer et décrire par un exemple un problème classique de la sémantique. Désigner en une phrase les caractéristiques de la parole concernées par la pragmatique**

- Anaphore : Problème classique lié à des mots référant à d'autres mots dans une phrase.
  Exemple : "J'ai posé une tasse sur la table et puis je l'ai déplacée.". Rien ne permet de savoir de façon formelle si on a déplacé la table ou la tasse.
  D'ailleurs, aucun système informatique ne peut le déterminer à ce jour.

- Pragmatique : Concerne tout ce qui n'est pas traité par les points précédents.
  On retrouve dedans tout ce qui, pour la bonne interprétation du message, nécessite des informations qui ne se trouve pas de manière textuelle dans la phrase (entre les lignes).
  Ou encore des phrases qui requièrent des informations concernant le monde dans lequel la phrase est dite.
  Exemple: Si on parle de disques placés l'un après l'autre sur une tige pour faire une serrure, ils ne vont pas être soudé parallèlement à la tige, mais rien ne le dit dans le texte.

## 1.9 From Source To Receiver

**Décrire le fonctionnement de l'oreille et schématiser ses principales parties**




- L'oreille moyenne contient marteau, l'enclume et l'étrier. L'oreille communique avec le nez par la trompe d'Eustache.
- L'étrier met en vibration la fenêtre ovale qui est l'entrée de la cochlée (oreille interne) qui est un cone, enroulé sur lui même, rempli d'un liquide qui est mis en mouvement par le mouvement de l'étrier.

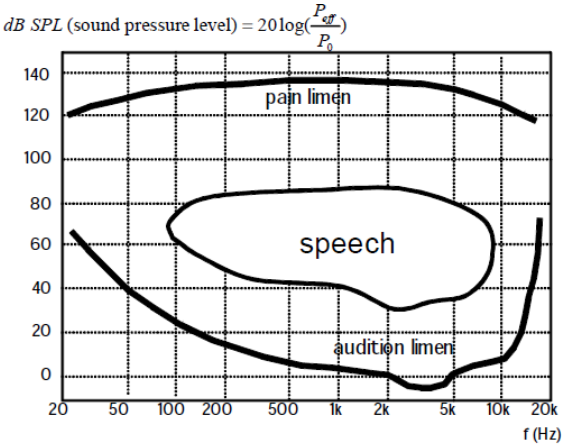
- Dans ce cone se trouvent des cellules ciliées qui sont raccordées chacune a un nerf auditif. Le mouvement du liquide a l'intérieur du cône va créer des impulsions électriques qui vont être comprises par le cerveau.

**Justifier l'analogie entre la cochlée et un analyseur spectral**

- En fonction de la fréquence du signal qui est poussé par l'étrier, les zones du cône qui sont mises en vibration varient. $\rightarrow$ ce ne sont pas les même cils qui seront stimulés si on change de fréquence du son entendu.
- À chaque zone ciliée correspond une gamme de fréquence entendue.

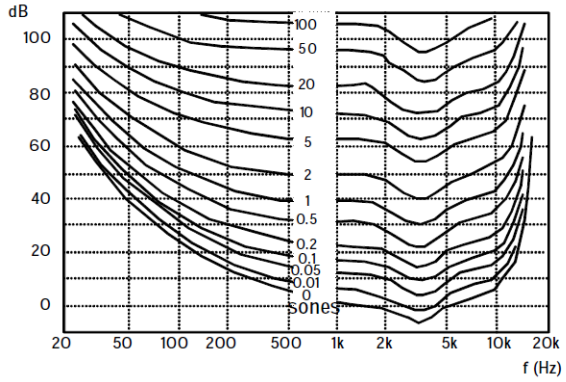
## Justifier que la psychophysique ait permis de normaliser la transmission de parole sur les canaux téléphoniques et traduire cette justification par un schéma

- Psychophysique : Branche de la psychologie qui étudie le cerveau en faisant appel à des ressentis de sujets de test.



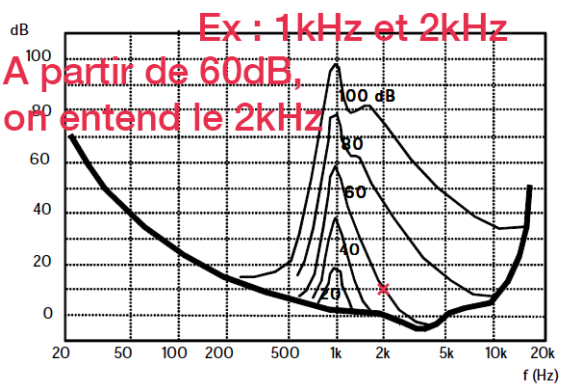

• En se basant sur le ressenti des gens, on a pu découvrir, en fonction de la fréquence, le seuil d'audition et le seuil de douleur de l'oreille humaine.
• En observant la courbe du seuil d'audition, on a pu formaliser en 1929 les fréquences qui allaient être utilisée pour la télécommunication: 300Hz → 3400Hz

## Donner la définition des courbes isosoniques et les comparer aux courbes de seuil de l'audition et de seuil de la douleur




• Courbes isosoniques : courbes d'égales sensations acoustiques. Elles servent à savoir quand est-ce qu'une fréquence peut être entendue avec la même intensité qu'une autre fréquence.
• Là ou le seuil d'audition et le seuil de douleur donnent des limites d'auditions, les courbes isosoniques donnent des informations sur l'équivalence d'intensité perçue pour deux fréquences différentes.

## Décrire le phénomène de masquage auditif et donner un exemple d'impact technologique en traitement de la parole




• Masquage auditif : Fait qu'une fréquence peut ne pas être entendue parce qu'elle est masquée par une autre fréquence. Lorsque quelqu'un écoute une fréquence, il faut dépasser un nouveau seuil pour que la nouvelle fréquence puisse être entendue.
• Ce masquage dépend à la fois de la fréquence, de l'amplitude du son de départ et de la fréquence du nouveau son à entendre. Ce phénomène peut devenir complexe car en réalité, il y a beaucoup de fréquences "masquantes" et de fréquences masquées.

• Dans les standardisation d'enregistrement sur média, on a décidé de ne stocker que les informations qui seront perçues. Problème : on a des courbes moyennes, mais en réalité, certaines personnes peuvent entendre des distorsions que d'autres n'entendront pas.
• Lorsqu'on calcule des taux d'erreur, il faut savoir que ces taux d'erreur n'auront pas toujours de signification forte, il faudra parfois les pondérés. On les pondérera grâce à des modèles de l'oreille humaine (modèles perceptuels) afin de prendre en compte non pas la distorsion physique mais la distorsion perçue.
• On peut également dire qu'on va se limiter aux fréquences audible par l'humain et ignorer les autres.

# 2 Modeling of the speech

## 2.1 Autoregressive modeling of the speech

**Distinguer les deux étapes du processus de mise en œuvre d'un nouveau modèle de la parole.**

- La création du modèle en lui-même : établir quels paramètres feront parti du modèle.
- La mise au point d'algorithmes qui permettent d'obtenir les paramètres de ce modèle : on peut voir ca comme une boucle d'estimation, on essaie des paramètres, on compare le résultat avec le résultat souhaité, on change certains paramètres et on recommence. (On peut avoir plusieurs algorithmes d'estimation différents pour un même modèle)

**Différencier erreurs de modélisation et erreur d'estimation, et montrer que la création d'un modèle résulte d'un compromis entre ces deux types d'erreurs.**

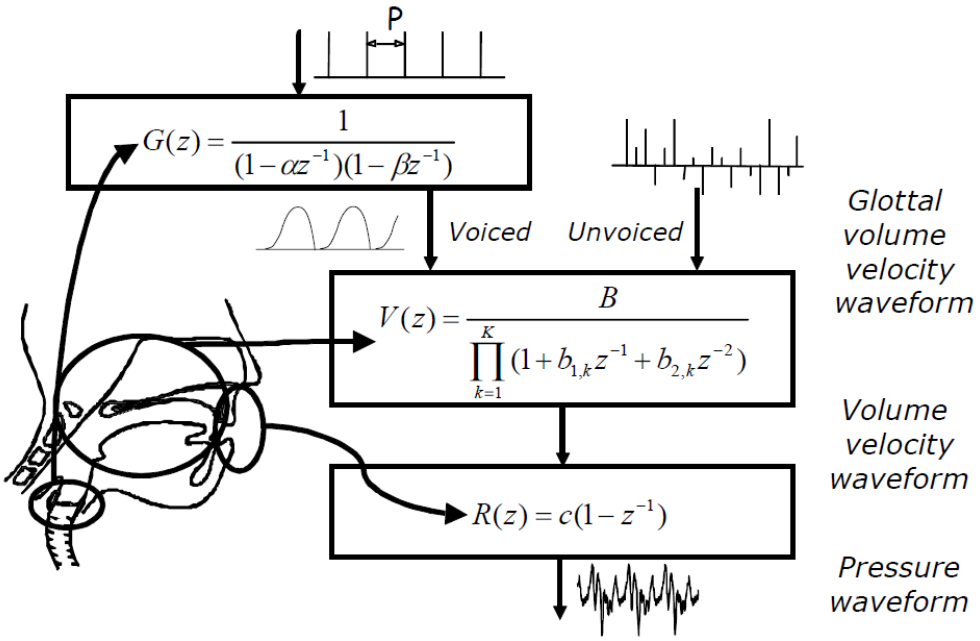
- Erreur de modélisation : dû à un modèle trop simple qui conduit à une impossibilité d'obtenir le résultat recherché.
- Erreur d'estimation : si le modèle est trop compliqué, on peut avoir du mal à développer un algorithme d'estimation.

Il faut donc trouver le juste équilibre entre modèle pas trop compliqué et pas trop simple.

**Citer et caractériser les trois grandes familles de modèles de la parole.**

- Modèle articulatoire : les équations sont des équations de la mécanique des fluides. On y représente par exemple les cordes vocales par un ensemble de masses vibrantes. Les paramètres sont par exemple : la position de la langue, la forme du conduit vocal, ...
- Modèle de production : modèle pour lesquels on symbolise la production de la parole à l'aide d'outils simples qui sont : des générateurs et des filtres. Ce modèle considère que la parole humaine est le résultat du passage d'un signal de base, simple à travers une série d'amplificateurs, de filtres, ...
- Modèle phénoménologique : modèle le plus mathématique, car se basent uniquement sur des théories de traitement de signal pures (analyse de fourrier, ...)

**Présenter le modèle autorégressif de traitement de la parole en réalisant l'analogie avec le système naturel de production de la parole.**

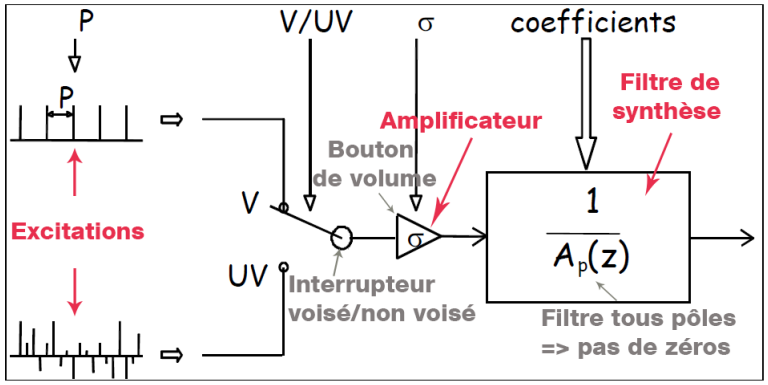



Modèle autoregressif = modèle prédictif linéaire = LPC

- Le signal glottique va être modélisé par un signal numérique (suite d'impulsion de Dirac) passant par un filtre numérique à deux pôles réels : $\alpha$ et $\beta$.
- On considère que le conduit vocal joue le rôle d'une suite de résonateurs, suite de système du second degrés à deux pôles. Ceci vient du fait qu'on considère donc que le conduit vocal produit des formants (résonances dans le spectre), et que donc on a besoin d'un nombre fini de systèmes résonants dont on va régler les coefficients $b_1$ et $b_2$ et chaque paire de coefficient servira a régler un formant en terme de fréquence et d'amplitude.
- Le débit en pression est réalisé par un simple système numérique qui simule la dérivation. La pression est proportionnelle à la dérivée du débit. Cette opération de dérivation peut être simulée par l'équation $R(z)$ en traitement de signal. Le zéro de l'équation est en z = 1. La mise en cascade des 3 filtres va faire en sorte de camoufler le zéro et ne faire apparaitre que les pôles.

**Donner un schéma de principe du modèle LPC de traitement de la parole, citer ses paramètres et caractériser leur rôle.**

G(z)V(z)R(z) ≈σ /Ap(z) : « All pole » model




On considère que si le son est non voisé, il peut "tout de même" (même si on ne met pas en cascade $G(z)$) être représenté par un filtre tous pôles.

Les pôles du modèle doivent rester dans le cercle unité pour que son filtre reste stable.

## 2.2 Autoregressive estimation of speech

**Expliquer pourquoi la modélisation autorégressive est basée sur un principe de minimisation d'erreur.**

- Le signal vocal n'est pas stationnaire. En effet, si le même son était entendu constamment, on ne pourrait pas transmettre de l'information. Mais étant donné les contraintes physiques des cordes vocales, on peut considérer le signal vocal comme stationnaire sur 30ms.
- Chaque fenêtre d'analyse aura donc une largeur de 30ms et le décalage entre chaque fenêtre est de 10ms. Ce qui permet de réduire l'erreur loors du passage au numérique sans pour autant demander trop de calculs.

**Décrire le système d'équations de Yule-Walker, en expliquer les composantes et justifier sa facilité de résolution.**

- Système de Yule-Walker : Système de p équations à p inconnues. La fonction $\phi_X$ est appelée fonction d'auto-corrélation. On va la calculer sur 30ms depuis l'ordre 0 jusqu'à l'ordre p. Ces valeurs sont les potentiomètres du modèle LPC (coefficients).
- La matrice qui contient les valeurs de $\phi_X$ est une matrice de Toeplitz(symétrie autour de la diagonale) ce qui permet de résoudre le système en $O(p^2)$ à la place de $O(p^3)$ pour un système standard (grâce à l'algorithme de Schur ou de Levinson).

$$\sum_{j=1}^{p} \phi_x(i-j)a_j = -\phi_x(i) \quad (i = 1...p)$$

$$\begin{bmatrix} \phi_x(0) & \phi_x(1) & \dots & \phi_x(p-1) \\ \phi_x(1) & \phi_x(0) & \dots & \phi_x(p-2) \\ \dots & \dots & \dots & \dots \\ \phi_x(p-1) & \phi_x(p-2) & \dots & \phi_x(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix} = - \begin{bmatrix} \phi_x(1) \\ \phi_x(2) \\ \dots \\ \phi_x(p) \end{bmatrix}$$

$$\mathbf{\Phi}_{\mathbf{x}}^{\mathbf{p-1}}\mathbf{a} = -\boldsymbol{\varphi}_{\mathbf{x}}^{\mathbf{p}}$$

**Préciser l'avantage de l'algorithme de Schur sur celui de Levinson.**

- L'algorithme de Schur permet de très bonnes performances dans le cas de limitation de la précision (*e.g.* précision limitée à 16 bits en virgule flottante => nombre de chiffres qui sont précis après la virgule est limité).

**Critiquer un choix de fréquence d'échantillonnage dans l'analyse d'un signal de parole.**

Ca dépend de l'application que l'on recherche.

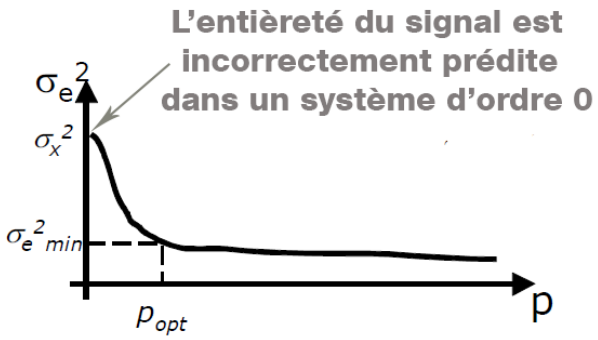
- Téléphonie : 8kHz : fréquence liée à la bande passante du signal téléphonique (on prend 2 fois la fréquence max).
- Reconnaissance vocale : 10kHz : Car ceci correspond en moyenne à 5 formants. Au delà, la qualité de la représentation vocale était meilleure mais pas la reconnaissance. Il ne sert donc à rien d'aller au delà.
- Synthèse de parole : 16kHz : on essaie d'avoir une voix la plus naturelle possible. Avec 16kHz, on arrive à reproduire plutôt bien les sons humains, pas besoin d'aller au delà.
- Multimédia : 11.025kHz, 22.05kHz, 44.1kHz : dépend du média et de l'utilisation.

**Donner une analyse critique du choix de l'ordre de prédiction pour modéliser la parole.**

L'ordre "p" de prédiction correspond à la taille du système à résoudre dans le modèle LPC. On peut faire une expérience simple qui consiste à mesurer la variance (l'énergie) du signal d'erreur, c'est-à-dire la partie du signal de parole qui n'est pas correctement prédite par un modèle d'ordre p. On choisit comme ordre de prédiction le "coude" du graphe.

En général $\sim p_{opt} = 2 + F_{sampling}$

La valeur à donner à l'amplificateur du modèle LPC est la racine carrée de la variance du signal d'erreur.



## 2.3 Extensions of the AR model

**Expliquer comment on synthétise un signal de parole avec le modèle LPC.**

- On vient placer dans le filtre les coefficients qu'on a mesuré, et on remplace le résidu de prédiction par deux excitations type :
  - Une suite d'impulsions de Dirac parfaitement périodique (voisé)
  - Un bruit blanc (non voisé)

qui sont ensuite passés dans l'amplificateur $\sigma$.

**Préciser les éléments du modèle sur lesquels agir pour améliorer la qualité du signal modélisé.**

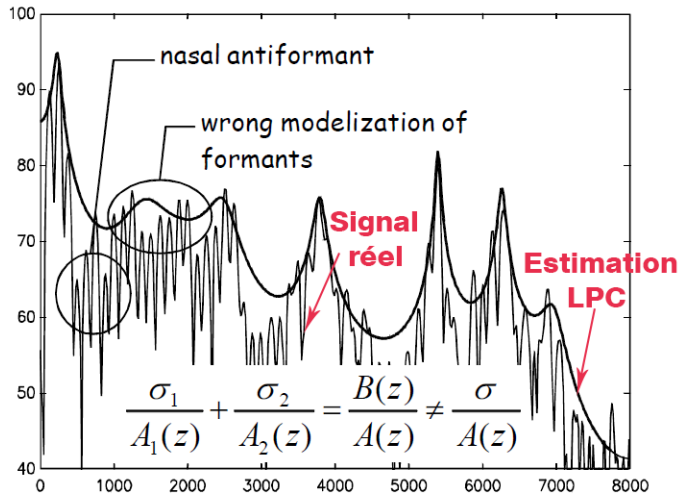
- Les excitations d'entrée sont parfaitement arbitraires, et c'est elles qu'il faut tenter d'améliorer pour améliorer la qualité du signal produit.

**Citer deux types de problèmes typiques de l'analyse LPC.**

- Problème des anti-formants (nasaux par exemple)
- Problème des sons mixtes (exemple : son $vv$)

**Définir ce qu'est un anti-formant et préciser son origine dans le système naturel de production de la parole.**

- Anti-formant : Signaux émis en opposition de phase qui fait que le contenu fréquentiel est très faible (2 signaux proches qui s'annulent pour une fréquence donnée).




La distance a parcourir et la différence du conduit entre le nez et la bouche peuvent causer un déphasage qui annule le signal total (somme des deux).

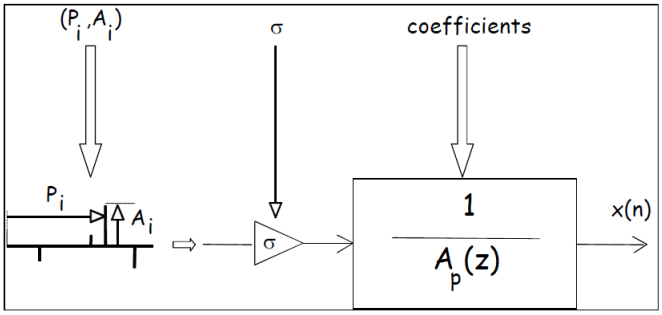
Étant donné que le modèle LPC est un modèle tous pôles (capable de reproduire des formants), il est difficile de représenter un anti-formant par ce modèle.

**Préciser l'origine des sons mixtes dans le système naturel de production de la parole.**

- Sons mixtes : sons pour lesquels les cordes vocales entrent en vibration, mais pour lesquels les cordes vocales ne se ferment pas complètement. Ce qui donne un mélange entre un signal voisé et un signal non-voisé.
- On pourrait être tenté de remplacer l'interrupteur par une somme entre signal voisé et non-voisé (ce qui existe comme modèle), mais cela rend le modèle excessivement complexe.

**Rappeler l'idée de base du modèle MP-LPC.**

- MultiPulse Linear Prediction (MP-LPC) : Créé pour palier aux problèmes de modélisation du modèle LPC. Il remplace l'excitation à deux types du modèle LPC par quelque chose de beaucoup plus flexible : un petit nombre d'impulsions réglables en position et amplitude.



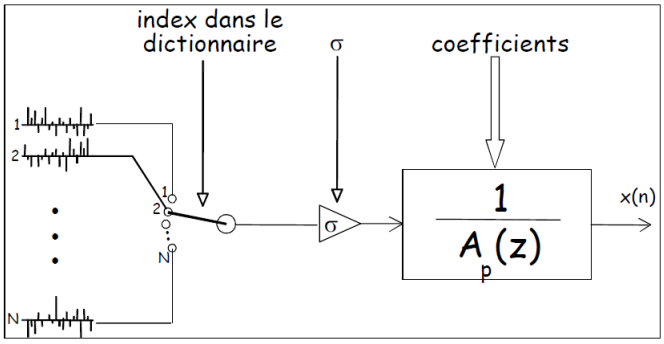

L'excitation est représentée par un petit nombre d'impulsions de Dirac dont on peut régler la position et l'amplitude de manière individuelle.

**Préciser comment estimer un modèle MP-LPC et expliquer la difficulté de cette estimation.**

- On devrait imaginer toutes les excitations possibles. Ce qui veut dire que pour chaque impulsion, il faudrait tester toutes les positions et les amplitudes possibles pour choisir les paramètres idéaux. Ce qui fait exploser le nombre de calculs.

**Rappeler l'idée de base du modèle CELP et citer un problème de ce modèle.**

- Code Excited Linear Prediction (CELP) : Créé pour palier aux problèmes de modélisation du modèle LPC. Il remplace l'excitation à deux types du modèle LPC par quelque chose de plus flexible : on remplace l'interrupteur à deux positions du modèle LPC par un interrupteur à 1024 ou 2048 positions.




L'excitation est choisie parmi un assez grand nombre de signaux différents. On considère qu'il y en aura au moins une qui conviendra. La difficulté revient alors à choisir la bonne position de l'interrupteur.

# 3 Speech Coding

## 3.1 Speech Coding

**Citer les deux buts fondamentaux du codage d'un signal de parole.**

- Stocker du signal de la parole sur un espace de stockage minimal
- Transmission du signal de parole sur un canal à bande passante minimale

**Expliquer pourquoi les algorithmes traditionnels de compression ne fonctionnent pas pour un fichier de parole.**

- Les systèmes de compression standards ne fonctionnent pas bien avec les fichiers sons. En effet, les algorithmes classiques (zip, rar, ...) recherche des suites de bytes qui se répètent. Dés lors, on ne stocke cette suite qu'une seule fois. Ceci n'arrive jamais dans un fichier audio. WAV $\rightarrow$ On stocke des suites de pression acoustique à des moments très rapprochés. Et ces suites ne se répètent presque jamais.

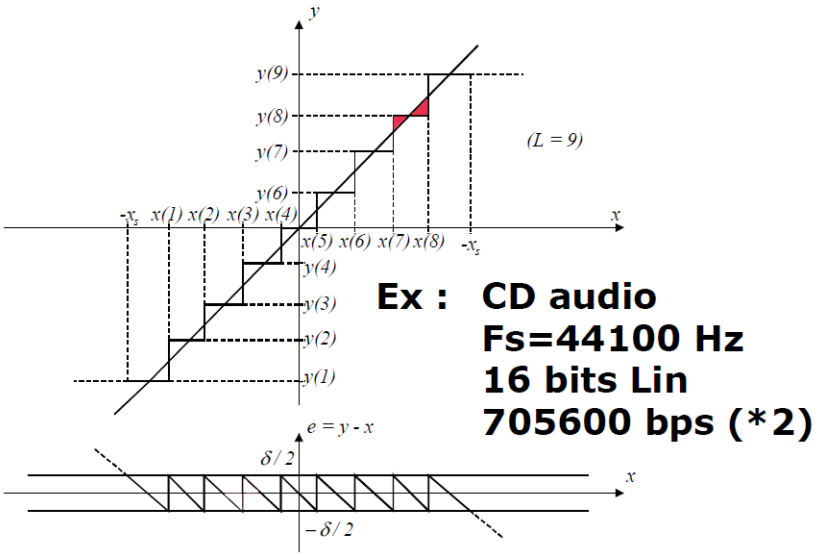
**Donner l'ordre de grandeur du débit binaire minimal d'un signal de parole (en ne considérant que son contenu phonétique) et le comparer au débit binaire admissible par une ligne téléphonique.**

- Débit minimal : Le nombre minimal de bits par seconde qui permettent de conserver l'information contenu dans le signal de parole :
  - Contenu : suite de phonèmes qui constituent le message à faire passer.
  - La voix de l'orateur : ce qui permet de reconnaître la voix de l'orateur.
  - Le para-linguistique : l'intonation
- On va se concentrer sur le stockage de contenu. On a vu qu'on a environ un phonème par $0,1$s, donc 10 phonèmes par secondes. Le nombre de phonèmes de la langue va définir le nombre de bits sur lesquels sont stocké le contenu (32 phonèmes = 5 bits). Pour l'exemple on a donc 5 bits * 10 phonèmes/s = 50 bits/s.
- Les lignes téléphoniques permettent l'échange de parole à environ 50kbits/s. On a donc 1000 fois le débit minimal.

**Décrire en les illustrant par un schéma, la technique de quantification uniforme et l'erreur de quantification qui lui est liée.**

- Quantification : opération consistant à remplacer les valeurs continues mesurées par l'ordinateur par des valeurs discrètes. Elle est caractérisée par un nombre de bits $b$, qui défini le nombre de pas de la fonction escalier à $2^b$.

- Quantification uniforme : quantification dans laquelle la taille des escaliers est constante.



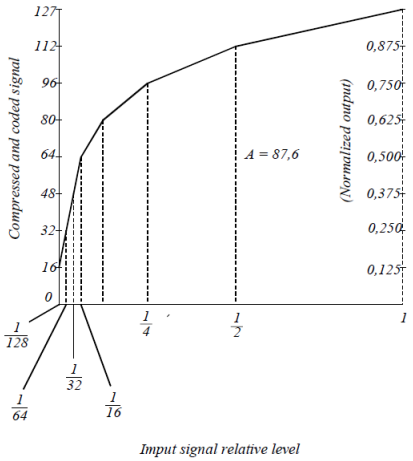

- L'erreur liée à la quantification est la différence entre la sortie discrète et l'entrée continue (cf. triangle rouge). Plus $b$ est grand, au plus l'erreur est petite.

**Détailler l'emploi de la technique de quantification uniforme pour un CD audio et calculer le débit binaire correspondant.**

- Voir image ci-dessus, On utilise $b$ = 16bits, et $44100\text{Hz} \cdot 16\text{bits} = 705600\text{bits/s}$.
  et comme on a des sons stéréo, on peut multiplier ce nombre par deux.
  Ce n'est donc pas un codeur de ce type qui est utilisé pour les CD mais plutôt un codage éliminant les composantes sonores inaudibles du fait de l'effet de masquage fréquentiel.

**Détailler l'emploi de la technique de quantification uniforme pour le codage de parole.**

**Ex : RNIS**
**Fs=8000Hz**
**8 bits Alaw**
**(G.711 norm)**
**64 kbps**




• Le signal de parole utilisant beaucoup plus de petites valeurs que de grandes, on va préférer un système de quantification non linéaire, qui permet d'être plus précis dans les petites valeurs.

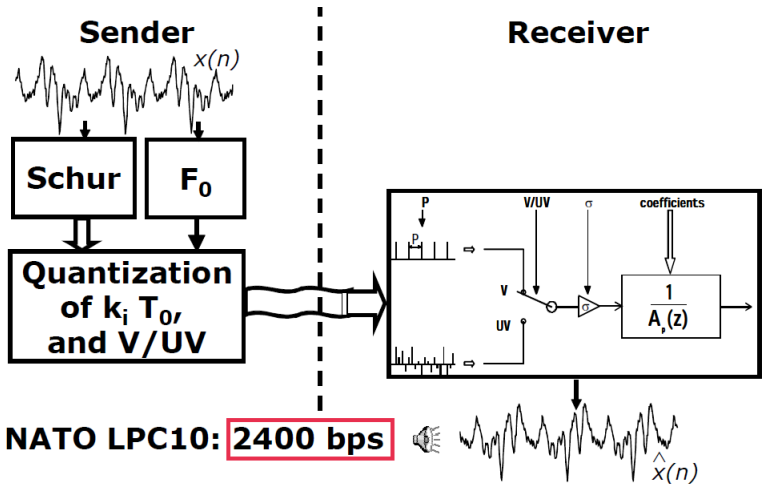
**Expliquer le principe d'une compression selon la loi A et citer une technologie qui l'exploite.**

• Cette compression permet d'obtenir une bonne qualité audio en prenant $b = 8$bits et $f = 8$kHz.
• Utilisée dans le RNIS : codage et décodage entre deux centrales téléphoniques.

## 3.2 Coders

**Schématiser un codeur par prédiction linéaire et commenter son fonctionnement.**

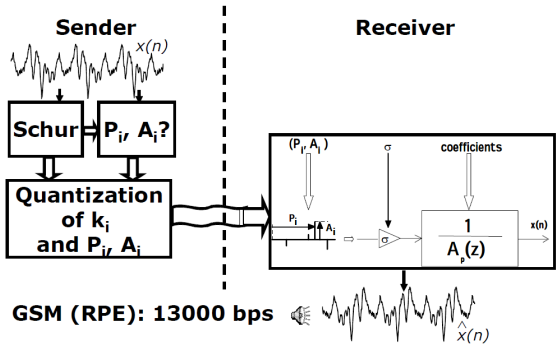
1. Algorithme de Shchur afin de devnier les coefficients de prédiction.
2. Analyse de fréquence fondamentale.
3. Décision signal voisé/non-voisé.
4. Transmission (100 fois/s) des coefficients, de la période et l'information voisé/non-voisé. Le tout en 24bits.
5. Décodage par synthétiseur LPC



**Citer la caractéristique typique de la parole codée en respectant la norme LPC10 et donner un exemple d'emploi dans la vie courante.**

- On y entend un caractère distordu métallique. De plus, une mauvaise décision voisé/non-voisé peut causer qu'un $sss$ se transforme en $zzz$ après codage/décodage.
- Les transmissions par satellites sont normalisées en LPC10.

**Schématiser un codeur MP-LPC et commenter son fonctionnement.**



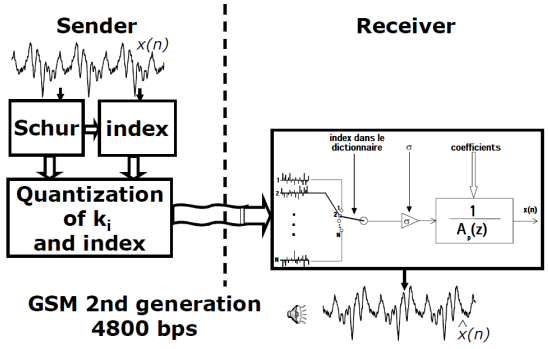

1. Algorithme de Schur.
2. Estimation de la position et de l'amplitude du modèle.
3. Transmission et quantification de ces grandeurs.
4. Décodage par synthétiseurs MP-LPC

**Justifier la dégradation de la parole lors d'une communication par GSM.**

- La norme RPE est un modèle simplifié du modèle MP-lPC. La dégradation est due au fait qu'on n'entend pas vraiment l'interlocuteur mais un synthétiseur de voix qui fait de son mieux pour reproduire le signal vocal.
- De plus, le fait de deviner les coefficients implique la résolution d'un système de 10 équations à 10 inconnues 10 fois par seconde.

**Schématiser un codeur CELP et commenter son fonctionnement.**

1. Algorithme de Schur.
2. Estimation de l'indice du dictionnaire d'excitation du modèle.
3. Transmission et quantification de ces grandeurs en 48 bits.
4. Décodage par synthétiseurs CELP

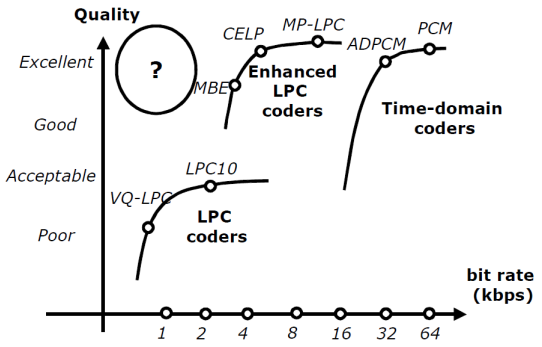


**Expliquer l'avantage que procure l'UMTS sur le GSM en ce qui concerne la transmission de la voix.**

- Il n'a besoin que de 10 bits par $0,1s$ pour les coefficients du système d'équations. Le reste disponible peut être utilisé pour stocker l'indice du dictionnaire. (1024 positions = 10 bits $\rightarrow$ 38 bits = beaucoup !)

## 3.3 Conclusion

**Positionner sur un schéma qualité/débit les techniques de codage de parole.**




**Citer une restriction importante à opérer sur le signal de parole traité, pour espérer un jour coder la parole avec un très bas débit (moins de 200 bits/s).**

- Il faudrait que les codeurs soient des codeurs phonétiques. C'est-à-dire qu'ils doivent savoir quelle langue ils codent/décodent. Ce qui est pour le moment impossible à automatiser.

**Prédire l'avenir de la recherche en codage de parole et justifier cette prédiction.**

- Tout ce qui devait être trouvé en codage de la parole a déjà été trouvé. la seule manière d'améliorer les techniques existantes seraient de surmonter la restriction citée ci-dessus.

# 4 Automatic Speech Recognition

## 4.1 Notes du présentiel

## 4.2 Automatic Speech Recognition (ASR)

**Justifier, sur le plan économique, le développement des systèmes de reconnaissance de parole.**

• Les sociétés de télécomunication sont les sponsors principaux de la recherche dans la reconnaissance automatique de la parole. En effet, ces sociétés s'intéressent à offrir des services payants à leurs clients en impliquant le moins d'employés possible. L'automatisation de la reconnaissance de la parole permet ce genre de services.

- Secrétaire virtuelle : pouvoir parler à un robot qui prend des rendez-vous, qui vous connait, ...
- Traduction automatique : pouvoir parler en anglais et l'ordinateur le répète en mandarin.

• L'armée américaine a subsidié aussi cette recherche dans le but de pouvoir traduire l'anglais américain vers de dialectes arabes.

**Donner trois domaines d'application de la reconnaissance de parole, ainsi qu'un exemple pratique pour chacun de ces domaines.**

- Bureau : contrôle vocal d'un ordinateur, d'une prise de note. (Un peu abandonné...)
- Business : gérer des stocks, en disant ce qu'il manque/reste à un ordinateur. (Un peu abandonné...)
- Médical : système de reconnaissance de mot clé pour rédaction automatique de rapport (Encore en développement actif !).
- Traduction : traduire d'une langue en une autre rien qu'en entendant la parole. (Encore en développement actif !).

**Expliquer pourquoi la coarticulation représente un important défi de la reconnaissance de parole.**

- Rappel – coarticulation : lorsqu'on prononce deux fois le même son dans deux mots différents, ils seront différents. Car le son dépend du son émis précédemment et des son à émettre prochainement.
- On peut donc avoir du mal à reconnaître les sons, car les zones de prononciation des phonèmes se recouvrent. De plus, selon la personne qui parle, les zones de recouvrements changent selon l'orateur.

Les autres défis sont :

- Différence des conduits vocaux
- Genre, dialecte

- Différence entre les langues
- Bruit ambiant, dû à l'environnement ou bruit humain.

**Citer la plus importante faiblesse des systèmes de reconnaissance de parole actuels.**

- La robustesse au bruit : capacité à ne pas considérer le bruit comme un problème.

**Citer et définir les deux principaux types de bruit, donner deux exemples par type.**

- Bruit additif : bruit s'ajoutant au signal de la parole (applaudissement, bruit dans le speaker du téléphone, ...). Le cerveau reconstitue de lui-même les parties qui n'ont pas été entendues.
- Bruit convolutif : bruit qui est dû à l'environnement se trouvant entre le conduit vocal de l'orateur et le conduit auditif du destinataire (réverbérations, distance, ...). Dans un sens, on change le filtre, la fonction de transfert de la parole. Les bruit dûs à la qualité du canal téléphonique sont des bruits de convolution car le signal est envoyés par une ligne différente entre chaque appel.
- Bruit inter-orateur : bruit dû au stress, à l'âge, à l'humeur, à l'effet Lombard...

**Donner une idée de ce qu'est l'effet Lombard.**

- L'effet Lombard : le cerveau est constamment en train d'écouter le bruit ambiant afin de savoir si ce que l'on va dire va bien être transmis au destinataire ou pas.
  Il y a donc une boucle de rétroaction (feedback) qui va permettre au cerveau d'adapter la transmission de certains sons selon le bruit ambiant.

## 4.3 ASR Systems

**Énoncer les contraintes qui régissent le choix d'un système de reconnaissance de parole.**

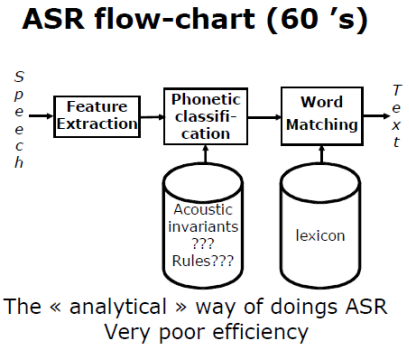
- Dépendant ou indépendant du locuteur.
- Type de reconnaissance :
  - Reconnaissance par mot isolé → Mettre des blancs volontairement entre les différents mots pour faciliter la segmentation de la phrase
  - Reconnaissance par mots connectés → Apparent à la reconnaissance par mots clés.
  - Reconnaissance de parole continue → Le plus compliqué car doit diviser la phrase en mots
  - Reconnaissance de mots clés → Ce qui est fait intuitivement quand on écoute une langue étrangère.
- Taille du vocabulaire connu. Petit : 100 mots; Moyen : 5000 mots; Grand : 50000 mots.
- Perplexité : le nombre de mots moyen qui pourrait suivre un mot.

- Contraintes de robustesse (conditions de laboratoire à conditions réelles) :
  - Bruit ambiant.
  - Type et qualité du microphone.
  - Position du micro.

**Positionner la complexité d'un système de reconnaissance de parole par rapport à un autre en fonction de leurs contraintes.**

| | *Isolated* | *Connected* | *Continuous* |
|---|---|---|---|
| ***Speaker dependent*** | small **1**<br>large **4** | small **4**<br>large **5** | small **5**<br>large **6** |
| ***Multi speaker*** | small **2**<br>large **4** | small **4**<br>large **5** | small **6**<br>large **7** |
| ***Speaker independent*** | small **3**<br>large **5** | small **4**<br>large **8** | small **5**<br>large **10** |

**Schématiser et commenter les principes de fonctionnement des trois grandes familles de reconnaisseurs de parole étudiés depuis les années 60.**

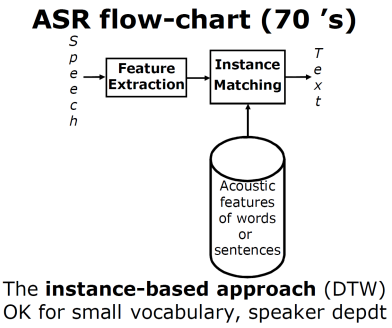


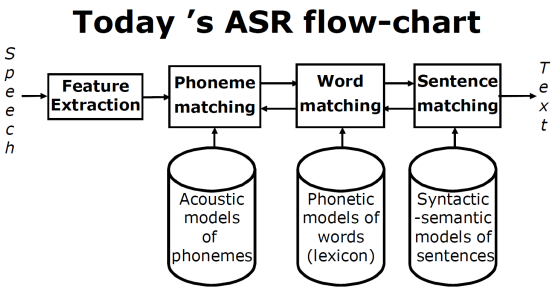

• Extraire les formants par LPC et en déduire le son produit. Ce reconnaisseur était régit par des règles de type *Si, Alors*.

• Idée: implémenter l'expertise d'un spécialiste de la parole qui sait lire un spectrogramme. De bonnes règles n'ont jamais été trouvées.

• Approche basée sur des exemples. On stocke la suite des vecteurs donnés par l'analyse LPC. Ces vecteur sont soumis à une comparaison par rapport à un fichier exemple. Si le résultat de la comparaison est bon, on admet que c'est le même mot.

• Demande un petit vocabulaire et d'être dépendant du locuteur.






**Phoneme-based approach using statistical models** (HMM or HMM/ANN) for acoustics and linguistics: Large vocabulary, speaker indepdt

• Approche basée sur des modèles statistiques. On compare les vecteurs donnés par LPC et on les compare avec un modèle qui donne un degrés de certitude de match.

• Ceci est d'abord réalisé pour des phonèmes et ensuite pour des mots, et enfin pour des phrases. C'est cette reconnaissance de phrases qui va émettre une conclusion.

• Les deux derniers blocs vont chercher le meilleur chemin parmi ceux qui sont possibles grâce à l'analyse précédente.

**Expliquer quels types de connaissances embarque un système de reconnaissance basée sur des modèles statistiques.**

- Connaissances acoustiques : quel son a du sens pour la langue.
- Connaissances lexicales : quel mot fait ou ne fait pas parti de la langue.
- Connaissances sémantiques/syntaxiques : quelle phrase a du sens dans la langue.

**Préciser les contextes d'utilisations respectifs (en fonction des contraintes qui régissent le choix d'un système de reconnaissance) de la de la reconnaissance basée sur des modèles et de celle basée sur des modèles statistiques.**

- Modèle: besoin de stocker tous les exemples en mémoire afin de les comparer un à un. Fonctionne bien avec un petit vocabulaire et en étant dépendant du locuteur.
- Statistiques: besoin d'un seul modèle statistique pour tout reconnaitre (ou essayer du moins). Correspond aux contraintes grand vocabulaire et indépendant de l'interlocuteur.

## 4.4 Instance-based ASR

**Préciser le problème de base d'un système de reconnaissance basé sur des exemples et présenter la solution généralement employée.**

• Le principe est de calculer le résultat d'une fonction $D$ de distance entre l'entrée et tous les exemples et on va associer l'entrée à l'exemple qui a donné le plus petit résultat pour $D$.

- Unknown utterance $\boldsymbol{X}=\{x_1,x_2,\ldots,x_N\}$
  (with $x_l=[x_{l1},x_{l2},\ldots,x_{ld}]^T$)
  Known utterances $\boldsymbol{Y^1}=\{y^1{}_1,y^1{}_2,\ldots,y^1{}_{J(1)}\}$
  $\boldsymbol{Y^2}=\{y^1{}_1,y^1{}_2,\ldots,y^1{}_{J(2)}\}$
  …
  $\boldsymbol{Y^K}=\{y^1{}_1,y^1{}_2,\ldots,y^1{}_{J(M)}\}$
- Compute $D(\boldsymbol{X},\boldsymbol{Y^k})$ for $k=1\ldots M$
- *Recognize:*

$$\boldsymbol{X}=\boldsymbol{Y^{best}}$$
with $D(\boldsymbol{X},\boldsymbol{Y^{best}})\leq D(\boldsymbol{X},\boldsymbol{Y^k})$ for $k=1\ldots M$

- Problème de base : il est impossible de reproduire la même durée d'élocution pour un même mot entre deux élocutions différentes → On a pas le même nombre de vecteurs entre l'entrée et les exemples.

**Préciser la notion de distance locale et de distance globale dans le cadre d'un système de reconnaissance basé sur des exemples.**

- Distance locale : distance qui sépare un quelconque vecteur de l'entrée avec un quelconque vecteur d'un exemple. Ces vecteurs sont de même tailles (car même analyse à l'origine). Plus simple à calculer : distance euclidienne. Il en existe d'autre ciblée au TDP.
- Distance globale : On fait correspondre les vecteurs de l'entrée avec les vecteurs du mot de référence, et on calcule la distance entre ces correspondance et une courbe de base.

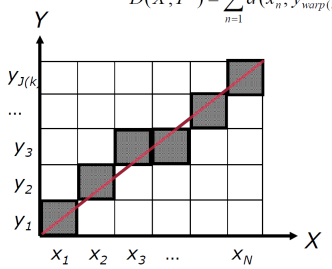
**Schématiser la solution la plus adéquate au problème du calcul de la distance globale dans le cadre de la reconnaissance basé sur des exemples, lorsque les mots à reconnaître sont longs.**

• Fait correspondre chaque vecteur de l'entrée avec chaque vecteur d'un exemple et calcule la distance entre cette correspondance et une correspondance linéaire "parfaite".

• Pas réaliste car suppose une grande probabilité de correspondance parfaite, ou du moins un allongement uniforme entre l'entrée et l'exemple ce qui n'est pas possible car des sons comme $d$ ne s'allonge pas.

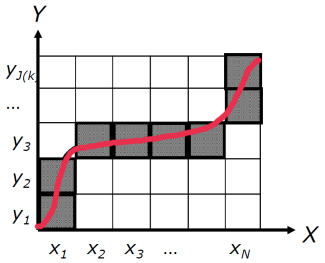
- Linear time warping

$$D(X,Y^k) = \sum_{n=1}^{N} d(x_n, y^k_{warp(n)})$$




- Not realistic for long words or phrases

- **Non linear** time warping




- Best path? **DYNAMIC TIME WARPING (DTW)**

• Cette fois, on suppose que certaines zones de l'entrée sont répétées plus longtemps que dans l'exemple et que d'autres zones durent le même temps.

**Expliquer conceptuellement l'algorithme DTW.**

- DTW : On examine toutes les courbes de référence possible, et pour tous ces chemins, on calcule la distance globale avec la correspondance expérimentale et on garde la courbe qui représente la distance globale minimale.

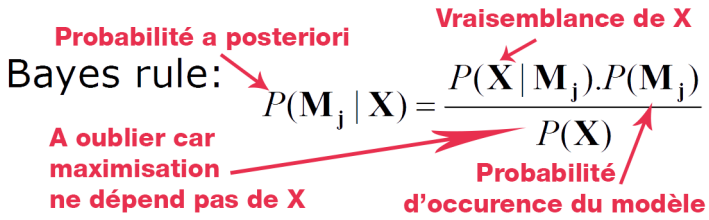
## 4.5 Model-based ASR

**Présenter la classification bayesienne et définir les termes de la règle de Bayes dans le cadre de la reconnaissance basée sur des modèles.**

- Le principe est de supposer que l'entrée correspond à un des modèle présent en mémoire. Ce faisant, on suppose que l'ensemble des sons qu'il est possible de reconnaitre est fini, ce qui n'est en réalité pas le cas. Ensuite, le but est de calculer, pour chaque modèle, la probabilité que l'entrée corresponde à ce modèle.

- On parle de classification Bayésienne lorsqu'on utilise une classification *a posteriori*, c'est à dire lorsqu'on compare ce qui a voulu être dit par rapport à ce qui a été dit.

- Le problème est qu'il est impossible en pratique de connaître cette probabilité. On va donc utiliser la règle de Bayes.




**Développer la règle de Bayes afin d'obtenir une expression exploitable de la probabilité à posteriori de prononciation d'une phrase, justifier chaque étape de ce développement.**

$$\max P(\mathbf{M}|\mathbf{X}) = \max [P(\mathbf{X}|\mathbf{M}) \cdot P(\mathbf{M})]$$

La vraisemblance de X se calcule selon le nombre de transcriptions possibles (façon de dire) du modèle et de la probabilité de chaque transcription ($P(P_i|M)$). $P_i$ = suite de phonèmes.

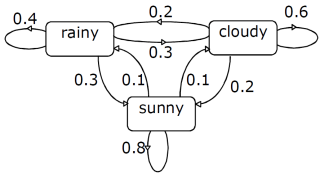
$$P(\mathbf{X}|\mathbf{M}) = P(\mathbf{X}|\mathbf{P_1})P(\mathbf{P_1}|\mathbf{M}) + P(\mathbf{X}|\mathbf{P_2})P(\mathbf{P_2}|\mathbf{M}) \\ \ldots + P(\mathbf{X}|\mathbf{P_L})\, P(\mathbf{P_L}|\mathbf{M})$$

**Citer les probabilités estimées par les modèles acoustique, phonétique et de la langue.**

$P(\boldsymbol{X}|\boldsymbol{P}_l)$ ← **Acoustic model**
$P(\boldsymbol{P}_l|\boldsymbol{M})$ ← **Phonetic model**
$P(\boldsymbol{M})$ ← **Language model**

## 4.6 Acoustic Model : Markov Chain

**Décrire le principe des chaînes de Markov.**




- Automate déterministe fini et stochastique avec des probabilités liés à la transition entre ces états. Une suite de phonème pourrait être suivie dans un tel modèle.
- Problème : Il y a plusieurs suites de phonèmes possible pour exprimer la même chose.

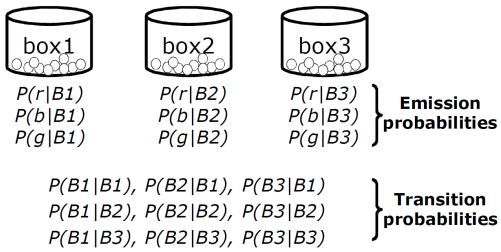
**Expliquer pourquoi une chaîne de Markov ne permet pas de calculer directement une probabilité dans un modèle acoustique en reconnaissance de parole basée sur des exemples.**

- Dans le modèle de Markov, la probabilité d'un état donné dépend uniquement de la probabilité de la transition vers lui depuis l'état précédent.
  De plus, l'ensemble des états est observable, il n'y a aucune indétermination sur ces états.
- Or, le signal de parole est constitué de phonème, et la coarticulation fait qu'il n'est pas possible de savoir avec précision quel son a voulu être prononcé (→ pas observable)
  De plus, il existe plusieurs façon de prononcer une même phrase et un exemple devrait donc avoir plusieurs chaînes de Markov qui lui est associé.

## 4.7 Acoustic Model : Hidden Markov Model (HMM)

**Décrire le fonctionnement des modèles de Markov cachés sur base d'un exemple.**

- On a trois boites avec des boules de couleurs différentes. On peut donc calculer la probabilité d'aller chercher une boule d'une certaine couleur selon ce qui a déjà été pris.



$$P(r,b,r|Model) = P(r,b,r|B1,B1,B1)\,P(B1,B1,B1) + P(r,b,r|B1,B1,B2)\,P(B1,B1,B2) + P(r,b,r|B1,B1,B3)\,P(B1,B1,B3) + P(r,b,r|B1,B2,B1)\,P(B1,B2,B1) + P(r,b,r|B1,B2,B2)\,P(B1,B2,B2) + \dots + P(r,b,r|B3,B3,B3)\,P(B3,B3,B3)$$

$$\text{with} \quad P(r,b,r|Bi,Bj,Bk) = P(r|Bi)\,P(b|Bj)\,P(r|Bk)$$

$$P(Bi,Bj,Bk) = P(Bi)\,P(Bj|Bi)\,P(Bk|Bj)$$

• On a donc un grand nombre de combinaison qui rendent la suite "rouge, bleu, rouge" possible.

• Mais, l'expérimentateur peut décider à tout moment de changer de boite. On veut donc savoir la probabilité d'une suite de couleur étant donné cette *double indétermination*.

**Différencier les probabilités d'émission et de transition dans les modèles de Markov cachés.**

- Probabilité d'émission : probabilité qu'un évènement se produise dans un état.
  → Probabilité d'une couleur prise dans une boite donnée.
- Probabilité de transition : probabilité qu'une transition d'état se produise.
  → Probabilité que l'on change de boite pour prendre une boule.

**Définir en quoi les modèles de Markov cachés sont-ils doublement stochastiques.**

- Pour l'exemple, on doit jouer à la fois sur l'inconnue couleur mais aussi sur l'inconnue boite choisie. On doit donc prendre en compte les probabilités d'émission et les probabilités de transition.

**Citer et définir les trois types de problèmes à traiter pour utiliser les modèles de Markov cachés.**

- Estimation : quelle est la probabilité de l'occurrence des données étant donné le modèle.
- Entraînement : comment calculer les probabilité d'émission et de transition.
- Décodage : étant donné une séquence d'observations (boules), quelle est la suite d'états (boite) la plus probable qui mène à cette séquence d'observation.

**Citer les deux algorithmes utilisés pour résoudre le problème d'estimation des modèles de Markov cachés.**

- Baum-Welch : prend en compte tous les chemins possibles dans le calcul.
- Viterbi : approxime le calcul de probabilité en ne prenant compte que du chemin le plus probable.

**Illustrer par un exemple l'entraînement d'un modèle de Markov caché.**

Sur l'exemple des boules et des boites :

1. On prend une longue séquence de boules.

2. On suppose qu'on connait les probabilités d'émission et de transition, et si on ne les connaît pas, on les invente au hasard.
3. On décode la séquence → On estime de quelle boite provient une boule de la séquence.
4. On ré-estime les probabilités d'émission et de transition

On répète jusqu'à stabilisation du système (les probabilités ne changent plus entre deux itérations).

**Énoncer le principe général de l'algorithme EM permettant l'entraînement des modèles de Markov cachés.**

1. On prend une longue séquence de données sur lesquelles travailler.
2. On suppose qu'on connait les probabilités d'émission et de transition, et si on ne les connaît pas, on les invente au hasard.
3. On décode la séquence → On estime de quelle état provient un élément de la séquence.
4. On ré-estime les probabilités d'émission et de transition

On répète jusqu'à stabilisation du système

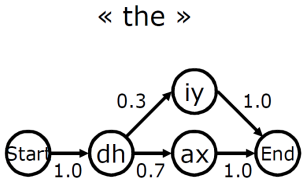
## 4.8 Phonetic Model

**Citer la technique utilisée dans les modèles phonétique pour calculer la probabilité d'une transcription phonétique étant donné une suite de mots.**

• On utilise des chaînes de Markov. Chaque état de l'automate est lui même un automate

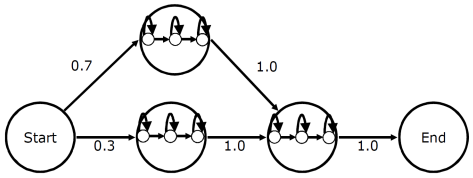
**Donner un exemple illustrant les limites des modèles phonétiques.**

• On ne travaille qu'avec des probabilités a priori, on ne tient pas compte du mot qui suit. Ce qui biaise les résultats de ces modèles.




**Préciser comment les modèles phonétiques intègrent un traitement individualisé pour chaque phonème et signaler les avantages de cette technique.**

• Si chaque phonème a sa propre chaîne de Markov, la chaîne de Markov du mot devient alors un Markov caché.
• Le but d'une telle pratique est de permettre à un modèle une très grande diversité en n'utilisant qu'un seul grand modèle, ce qui permet de réaliser en une fois l'entrainement du modèle.




## 4.9 Language Model

**Citer et définir les trois problèmes des modèles de la langue.**

- Entraînement : comment estimer les valeurs des paramètres des modèles
- Estimation : comment estimer la probabilité d'une phrase dans la modèle de la langue.
- Décodage : le nombre de phrase d'une langue étant infini, il est impossible de fournir une liste exhaustive de toutes les phrases prononçable. On va donc essayer de deviner le mot qui suivra le mot précédent afin d'essayer de discrétiser les cas possibles.

**Exprimer de trois façons différentes la probabilité de prononcer une phrase donnée en fonction des mots qui la composent et commenter l'utilisation de ces expressions dans les modèles de la langue.**

$$P(\mathbf{M}) = P(\mathbf{W_1}, \mathbf{W_2}, ..., \mathbf{W_K})$$
$$= \prod_{k=1}^{K} P(\mathbf{W}_k \mid \mathbf{W}_{k-1}, \mathbf{W}_{k-2}, ..., \mathbf{W}_1)$$

• Part du principe que chaque mot permet de deviner un ensemble de mot qui pourrait le suivre.
• Impossible à réaliser dans la réalité...

• Part du principe que certains couples de mots reviennent souvent et que d'autres ne veulent rien dire ou sont très rares.
• S'entraine à partir d'un très grand corpus de textes. S'estime et se décode grâce à une grande base de données dans laquelle le corpus a été entré → Tend à éliminer des couples considérés comme "improbables" alors qu'ils se retrouvent peut être souvent dans le langage parlé.

**Word-pair model**

$$P(\mathbf{M}) = 1 \quad \text{if} \quad \exists(\mathbf{W}_k, \mathbf{W}_{k-1}) \quad \text{for all } k$$
$$= 0 \quad \text{otherwise}$$

**n-gram models**

$$P(\mathbf{M}) = \prod_{k=1}^{K} P(\mathbf{W}_k \mid \mathbf{W}_{k-1}, \mathbf{W}_{k-2}, ..., \mathbf{W}_1)$$
$$= \prod_{k=1}^{K} P(\mathbf{W}_k \mid \mathbf{W}_{k-1}, \mathbf{W}_{k-2}, ..., \mathbf{W}_{k-n})$$

• Essaie de deviner un mot selon les n mots qui le précède. Un tri-gram (n=2) essaie de deviner le mot suivant ses deux prédécesseurs.

**Donner une valeur réaliste au nombre n du modèle n-gramme et préciser les conséquences liées à ce choix.**

• On peut se limiter à des suites de 3 mots. Mais il arrivera de tomber sur une suite de trois mots qui existe dans la langue mais qui ne se retrouve dans aucun texte.

**Citer le problème du modèle n-gramme et proposer deux techniques pour le résoudre.**

• Même si on se limite à trois mots, il existera des suites de 3 mots qui sont très rare et qui seront donc considérées comme incorrectes.

- Back-off: un premier principe est alors de retomber sur deux suites de deux mots, que l'on va pondérer afin de ne pas considérer la suite de mots comme improbable.
- Part-of-speech: on se réfère à la nature syntaxique des mots précédents plutôt que de la suite même des mots.

**Commenter la qualité actuelle des systèmes exploitant les modèles de la langue.**

• Tout le monde sait que les n-gram sont des approches imparfaites. Mais chaque fois qu'on a essayé d'aller plus loin, de préciser ces probabilités, on a fait baisser le taux de reconnaissance.

## 4.10 ASR Conclusion

**Positionner le taux d'erreur d'un système de reconnaissance de parole en fonction du type de reconnaissance, de la tâche, de la dépendance du locuteur, de la taille du vocabulaire.**

| Type | Task | Mode | Vocabulary | error rate |
|---|---|---|---|---|
| ***Isolated words*** | Equiprobable words | Sp. Depdt | 10 digits | 0% |
| | | Sp. Indepdt | 39 ascii | 4.5% |
| | | Sp. Indepdt | 1109 basic English | 4.3% |
| | | Sp. Indepdt | 10 digits | 0.1% |
| | | Sp. Indepdt | 39 ascii | 7.0% |
| | | Sp. Indepdt | 1218 names | 4.7% |
| ***Connected words*** | Sequence of digits id. | Sp. Depdt | 10 digits | 0.1% |
| | | Sp. Indepdt | 11 digits | 0.2% |
| | Flight reservation | Sp. Depdt | 129 words | 0.1% |
| ***Continuous speech*** | Ressource management (perplexity 60) | Sp. Indepdt | 991 words | 3.0% |
| | Airline travel information system (perplexity 25) | Sp. Indepdt | 1800 words | 3.0% |
| | Wall street journal (perplexity 145) | Sp. Indepdt | 20000 words | 12.0% |

**Expliquer la différence d'erreur de reconnaissance entre la lecture d'un journal en laboratoire et la parole de la vie réelle.**

• La parole de la vie réelle comporte du bruit de fond, de bruit gutturaux, de la toux, ... Cela implique un taux d'erreur supplémentaire d'environ 30%.

• Current issues :
**Robustness Spkr adaptation Language models**

**Souligner l'importance du modèle de la langue en reconnaissance de parole.**

• Si on fait sauter le modèle de la langue, le taux d'erreur grimpe fortement. Dés lors, on peut imaginer grappiller les derniers pourcentages d'erreur en améliorant le modèle de la langue.

**Citer les champs d'investigation actuellement explorés en reconnaissance de parole.**

- Robustesse .
- Modèle de la langue .
- Adaptation au locuteur : plus on utilise le système et plus il connait le locuteur et moins il fait d'erreur.

**Justifier qu'un être humain reste plus performant qu'un ordinateur pour reconnaître la parole.**

• L'être humain est capable à la fois de reconnaître ce qui a été dit mais aussi de comprendre ce qui a été dit. Ce qui est pour le moment assez laborieux pour une machine.
• On peut donc prévoir que si une machine est capable de comprendre ce qui a été dit, alors, le taux d'erreur de reconnaissance de la parole serait facilement diminué.

# 5 Speech Synthesis

## 5.1 Notes du présentiel

### Domaines d'application

- Beaucoup de domaines ont été abandonnés comme le renseignement par téléphone. Le plus gros domaine reste l'aide aux handicapés et la recherche.

### En quoi c'est difficile ? Phonétisation - Intonation - Durée

- Problèmes liés à la langue (principalement langue française) comme l'homographe hétérogène. ce sont des mots qui s'écrivent de la même manière mais se prononcent de manière différente. Il y a aussi des assimilations de nasalité (des phonèmes que l'on ne prononce pas correctement), ...
- L'intonation présente des problèmes : comment modéliser une courbe de pitch ? Cette courbe est un tout car la moindre modification s'entend, mais chaque personne le dira différemment. L'accent tonique est également un problème.
- La durée d'un même phonème à deux endroit différents d'une phrase n'aura pas du tout la même durée.

### En quoi c'est difficile ? Traitement de signal

- Coarticulation pose un gros problème...

### Historique

- Première machines parlantes en 1791 et machine mécanique plus tard ... Mais que pour des sons de base.
- Machine à parler électrique de 1936 avec 10 touches qui représentent les formants et pédale de pitch, etc... C'est bien, mais rien d'automatisé.

  - Premiers TTS : création de formants.
  - Deuxième type de TTS : On colle des diphones[1] ensemble.
  - Troisième et dernier type : système automatique à très grande BDD qui cherche les meilleurs diphones pour le mot à dire.

[1] Moitié de son (couple de phonème) qui permette de résoudre le problème de coarticulation en coupant le son après la première moitié du premier phonème et avant la deuxième moitié du deuxième phonème

## 5.2 Text-To-Speech Synthesis

**Citer trois applications de la synthèse de parole en télécommunication.**

- Téléphone
- Multimédia
- Communication homme-machine

**Positionner le besoin en interactivité pour les applications actuelles en téléphonie.**

- On voudrait savoir qui appelle, avoir assistant virtuel, annuaire inversé, ou encore avoir accès à des informations par une demande vocale (application obsolète aujourd'hui).

**Donner un exemple d'application de la synthèse de parole dans le domaine du multimédia.**

- Jeux interactifs, livre parlant, ...

**Justifier l'intérêt de la synthèse de parole pour la communication homme-machine.**

- Pouvoir contrôler, ou du moins communiquer avec des machines avancées.

**Préciser en quoi la synthèse de parole apporte une aide aux handicapés.**

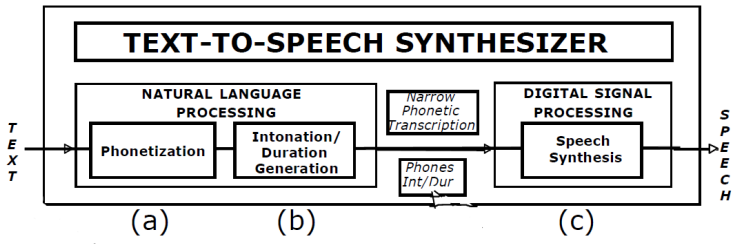
- C'est le domaine d'application le plus touché par la synthèse de la parole. Ces systèmes permettent aux personnes dans l'incapacité de parler de communiquer.

**Expliquer pourquoi la synthèse de parole est d'une importance fondamentale pour les expériences scientifiques sur le langage naturel.**

- Elle permet de mieux comprendre la façon dont nous parlons.

## 5.3 TTS Diagram / Phonetization

**Schématiser un synthétiseur de parole TTS.**



**Définir le rôle des linguistes informaticiens.**

- Établir un pont entre le traitement du langage naturel et les applications informatiques. en utilisant les règles et connaissances des linguistes, ils s'occupent du passage de (a) et (b) à (c).

**Identifier la première difficulté de l'étape de phonétisation et présenter quatre exemples supplémentaires de difficultés illustrant la complexité de cette étape.**

- On pourrait penser qu'il suffit de transformer tous les lettres présentes dans le texte en phonèmes, mais on trouve alors plusieurs problèmes.
  - Homographes-Hétérophones : mots s'écrivant de la même manière mais se prononçant différemment.
  - Les liaisons phonétiques : certaines liaisons sont obligatoires, d'autres facultatives, et d'autres interdites
  - Assimilation de nasalité : phonèmes qui sonnent différemment selon les phonèmes qui le précèdent ou qui le suivent (contraintes physiques des cordes vocales).

| Problem | Example | Level | Information |
|---|---|---|---|
| ***Assimilation*** | nasality or sonority assimimation, vocalic harmonization | word/sentence | reading style, pronunciation of neighbors |
| ***Heterophonic homographs*** | **the**, record, contrast , read, est, couvent, portions, etc. | word | part-of-speech, meaning (rare) |
| ***Schwa deletion*** | table rouge, je ne te le redirai pas | sentence | syntactic articulation, pronunciation of neighbors, speaking style |
| ***Phonetic liaisons*** | très utile, deux à deux, plat exquis | sentence | syntactic articulation, |
| ***New words*** | proopiomelancortin | word | spelling analogy |
| ***Proper names*** | *your name here ...* | word | morphology, analogy |

- À noter qu'il n'est pas nécessaire de comprendre la phrase pour pouvoir la phonétiser; L'exemple de M. Lewis le prouve.

## 5.4 Intonation / Coarticulation

**Préciser l'utilité des fréquents mouvements intonatifs dans le langage naturel.**

- Ils permettent au receveur de bien séparer les groupes de mots que l'envoyeur prononce.

**Préciser le problème rencontré lors de la modification artificielle de la courbe intonative.**

- La courbe représente un tout, le moindre changement pourrait rendre toute la courbe invalide.
  Cependant, la même phrase ne sera jamais prononcée deux fois avec la même courbe de pitch.

**Citer les deux autres rôles majeurs de l'intonation.**

- Mettre en évidence certains mots → Porter une information différente.
- Permettre à l'auditeur de savoir quand la phrase est terminée.

**Caractériser les durées des phonèmes.**

- Non constante, et différente selon la position du phonème dans la phrase et même dans le mot. Elle est donc liée à l'intonation et est assez compliquée à formaliser.
  Pour les voyelles, cette durée est corrélée avec la fréquence fondamentale.

**Citer les connaissances nécessaires pour appliquer une intonation à une phrase.**

- Syntaxe : connaître la nature du mot et sa position permet d'avoir beaucoup d'information sur le mot, y comprit sa phonétisation, intonation et durée.
- Sémantique : comprendre la phrase permet de compléter la connaissance.
- Pragmatique : sens sous-entendu dans la phrase par l'intonation par exemple.

**Justifier que la synthèse de parole tente de reproduire la coarticulation.**

- Chaque phonème se prononçant différemment et à une vitesse différente, on ne sait pas définir un unique son par phonème. Dés lors, il faut pouvoir imiter ces différences.

**Résumer les défis et contraintes auxquels doit répondre la synthèse de parole.**

- Phonétisation correcte.
- Générer une prosodie (intonation et durée et le rythme);
- Produire une suite de phonème coarticulés.
- Contraintes de coût, de maintenance, de calcul et d'adaptation à la langue.

## 5.5 TTS techniques

**Résumer le principe de fonctionnement de la machine de Von Kempelen.**

- On simule l'appareil vocale dans une machine mécanique avec un souffleur, une hanche (pour simuler les cordes vocales), une bouche molle et un souffleur permettant les explosions de certains phonèmes.

**Résumer le principe de fonctionnement du Voder d'Omer Dudley.**

- Cette fois ci, un opérateur a accès à un clavier composé de 10 touches, et de plusieurs autres clapets supplémentaires (pitch, énergie, ...). On peut faire le rapprochement entre les 10 paramètres du modèle LPC mais en analogique.

**Exposer l'idée de la synthèse par formants et justifier son nom.**

- Comme certains spécialistes savent lire des spectrogramme, l'idée est de formaliser cette expertise, inversée, dans le synthétiseur pour qu'il puisse créer les formants correspondant au mot à prononcer.

**Caractériser la parole générée lors d'une synthèse par formants.**

- Métallique et entrecoupée. Tout sauf naturel. Néanmoins utilisée par les handicapés à l'époque.

## 5.6 Diphone / Unit selection-based synthesis

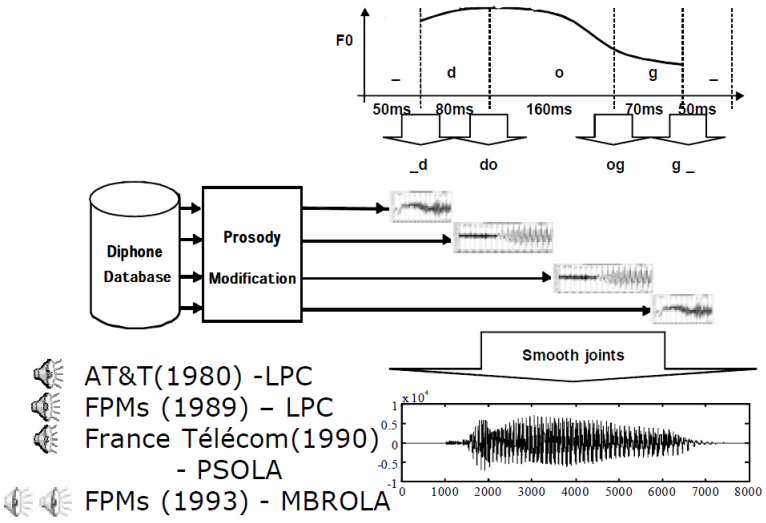
**Énoncer le principe de fonctionnement de la synthèse par diphone.**

- On colle des diphones[2] ensemble.

**Préciser comment la synthèse par diphone respecte la coarticulation du langage naturel.**

- Un système de modification de prosodie va permettre de retrouver au mieux l'intonation et la durée du signal. Et un système de lissage va permettre de lier au mieux les diphones entre eux.

**Schématiser le fonctionnement d'un synthétiseur de parole par diphone et préciser sur ce schéma les deux principaux problèmes du modèle.**



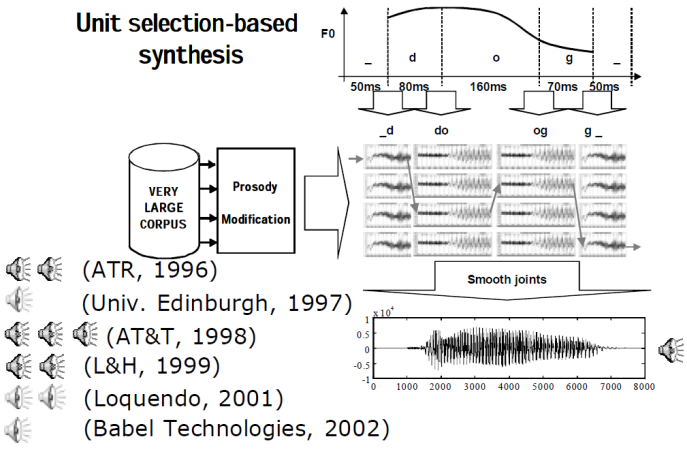

[2]Moitié de son (couple de phonèmes) qui embarque en son sein l'essentiel de la coarticulation entre les deux sons et s'obtient en coupant le son après la première moitié du premier phonème et avant la deuxième moitié du deuxième phonème

**Commenter le développement du projet MBROLA.**

- Part du principe de PSOLA qui avait comme idée que coller à des endroits différents que la base des fenêtres (en cloche) de pondérations lors de l'analyse.
  Il a été développé par M. Dutoit dans le cadre de son doctorat

**Citer la différence entre la synthèse par sélection d'unité et la synthèse par diphone.**

- Cette technique nécessite une grande base de donnée contenant plusieurs fois tous les diphones. Avec cette base, le principe est de rechercher les meilleurs diphones de la BDD en prenant en compte le diphone précédant et suivant le diphone recherché.

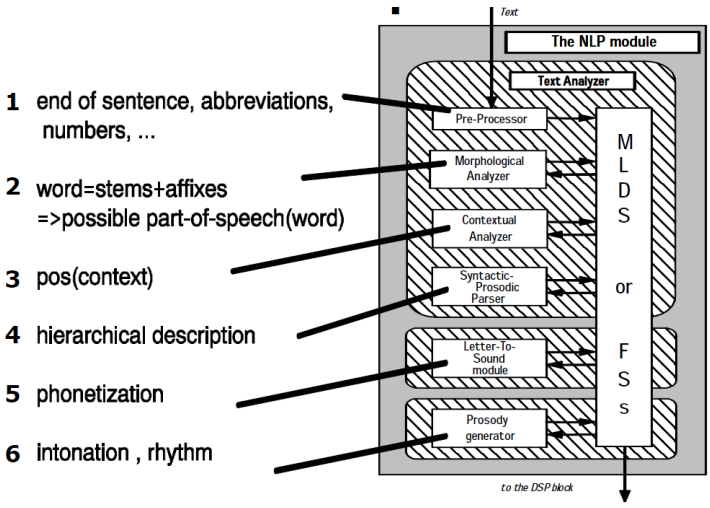



**Citer le problème qu'il restait à résoudre en synthèse par sélection d'unité.**

- La recherche des meilleurs diphones dans la base de données.

## 5.7 Pre-processing / Morphological analysis / Contextual analysis

**Citer les sous-modules d'un module de traitement du langage naturel dans un système de synthèse de parole.**



**Citer les rôles du prétraitement dans un système de synthèse de parole.**

- Détecter les nombres
- Détecter les acronymes
- Détecter les abréviations
- Détecter quel "." termine la phrase et lesquelles ne la termine pas.

**Citer le type d'outils généralement utilisés pour résoudre une grande partie des problèmes de prétraitement en synthèse de parole.**

• Grammaires régulières simples à états finis.

**Justifier l'utilité de l'analyse morphologique en synthèse de parole.**

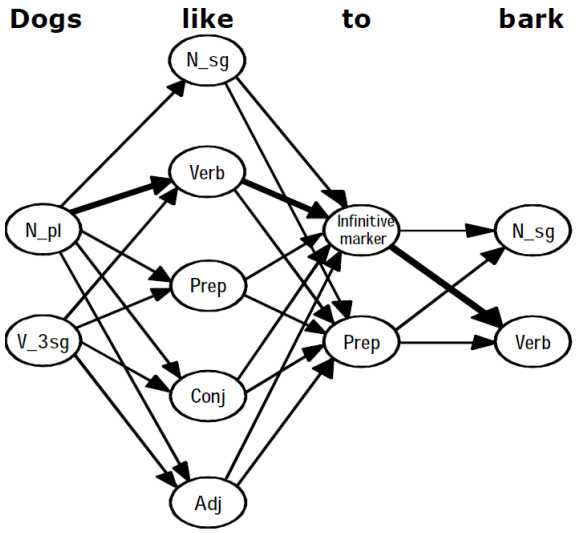
- Réduire la taille du lexique (grâce aux préfixes commun et suffixes circonstanciels).
- Aider à la prononciation lorsque la morphologie y est liée.
- Préciser les différentes natures possibles pour un même mot.
- Savoir quand l'accent tonique tombe dans un mot.

**Préciser comment est réalisée l'analyse morphologique en synthèse de parole.**

• Règles régulières, automates à états finis. Dans tous les cas, c'est très dépendant de la langue. On peut aussi utiliser une brute force sur un dictionnaire.

**Énoncer le but de l'analyse contextuelle et exposer la méthode pour y parvenir.**

• Le but est de trouver quel morphologie chaque mot a dans une phrase précise. Donc de trouver le chemin parmis un graphe contenant toutes les natures possibles du mot.
• Pour ce faire, on va utiliser la technique des n-grammes (voir partie 4 : ASR). C'est donc des modè-les probabilistes qui permettront de décider du meilleur chemin.



## 5.8 Syntactic-Prosodic Phrasing / Automatic Phonetization

**Offrir deux solutions pour identifier les groupes de mots d'une phrase en synthèse de parole.**

- Chinks'n chunks : un groupe de mot est composé de chinks suivi de chunks puis s'arrête pour que le prochain groupe commence avec un chink. Chink : mots fonctionnels. Chunk : mots lexicaux (noms, adjectifs, adverbes adjectivaux, ...).
- Les arbres CART : système de Machine-Learning qu'il faut entraîner avec des exemples.

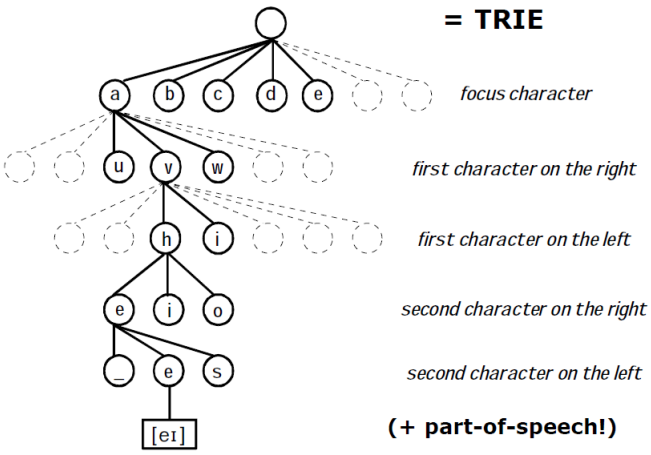
**Spécifier la méthode généralement employée pour réaliser la phonétisation automatique des mots en synthèse de parole et illustrer cette méthode par un exemple.**

• Règles réguliéres permettant de déclarer tous les cas particuliers, puis généraux. Un graphème est prononcé [...] quand / il est situé _.

```
s [z] / [éanti|hanti] _ [<V>]
s [s]/[anti|contre|impr,|prime|tourne|ultra|psycho|télé]
  _ []
s [s] / [vrai]_[em]
s [s] / [_a|para|sinu]_[e|o|y]
s [z] / [tran] _ [a|h|i]
s [z] /[<V>] _ [<V>]
```

**Présenter l'alternative à la méthode habituelle de phonétisation automatique et indiquer le point commun entre les deux méthodes.**

• Arbre de décision "TRIE" qui peuvent être entrainés par une batterie d'exemple. Pour que ca puisse fonctionner, comme pour les règles, il faut mettre des règles particulières sur la catégorie syntaxique du mot traité.



## 5.9 Prosody generation

**Commenter les premiers efforts réalisés pour générer la prosodie en synthèse de parole.**

- Consiste en 10 intonations différentes qui permettrait de pouvoir former n'importe quelle phrase dans la langue française en les combinant. Le problème est qu'on a jamais su mettre des règles sur quelle intonation utiliser à quel moment.

**Définir ce qu'est un ton et spécifier son utilisation pour caractériser un corpus.**

- Ton : mouvement voulu associé à la voyelle de chaque syllabe.
- L'idée est d'associer ces tons aux pics d'intonation présents dans une phrase afin de ne plus avoir à lire la courbe, mais juste ces unités abstraites que sont les tons.

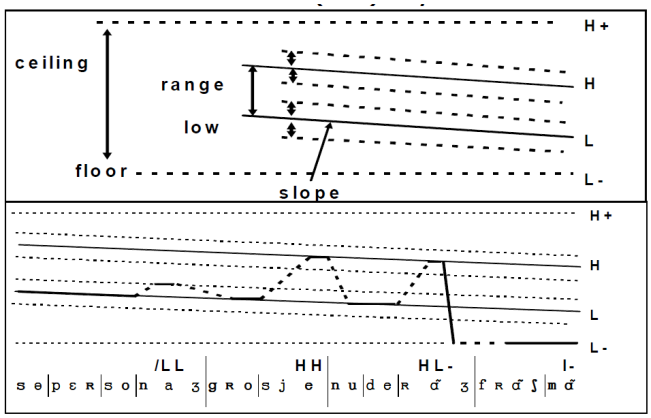
**Déterminer comment générer des tons sur base d'un texte.**

| | ce personnage grossier, te dérange-t-il |
|---|---|
| WS | . . . o . o . . o . |
| SG | (. . . -)( . -) ( . . . - ) |
| IG 1 | (. . . /LL)( . HH) ( . . . H/H) |
| IG 2 | (. . . - . HH) ( . . . H/H) |

**WS = word stress = lexical stress ← *Phonetization***
**SG = stress group**
**IG = intonation group ← *Synt.-Pros. Phrasing***
**(only one stressed syllable)**

**Déterminer comment générer par règles une courbe intonative sur base de tons.**

- Part du principe que quelque soit la phrase, il y a toujours une pente descendante.



**Déterminer comment générer par entraînement une courbe intonative sur base de tons.**

- On a une très grande base de donnée avec des phrases annotées en tons. On entraine le système avec cette base de données, et le système s'occupe ensuite de décider de la meilleure courbe possible pour la demande, à la fois pour l'intonation mais aussi pour éviter les discontinuations d'intonation.

**Établir l'analogie entre la génération de prosodie par entraînement et la synthèse par sélection d'unité et l'illustrer en schématisant un exemple.**

- Là où la synthèse par selection d'unité colle des diphones entre eux en maximisant le match et en minimisant les discontinuités, la génération de prosodie va coller des morceaux de sa base de données pour construire une courbe de pitch avec les même contraintes.

## 5.10 TTS Conclusion

**Résumer la tendance suivie depuis 1995 en synthèse de parole et donner une justification technique à cette tendance.**

- L'idée de machine learning sur corpus et de sélection automatique est à la mode depuis 1995, donc depuis l'explosion de la taille des médias numériques.
  Plus récemment, ce sont les réseaux de neurones profonds qui prennent le dessus dans la recherche. Ceux-ci ont l'inconvénient d'être très complexe et on a pas beaucoup de contrôles sur les résultats de ceux-ci.

# 6 Conclusion on Speech Processing

**Positionner l'évolution du codage de parole à l'heure actuelle.**

- On a presque trouvé tout ce qui pouvait être trouvé → On connait les normes.

**Positionner l'évolution de la reconnaissance de parole à l'heure actuelle.**

- Il y a des systèmes ouverts pour le grand public, mais avec toujours les problèmes de robustesse, de non-compréhension, ... Ceci s'est amélioré depuis 2003 mais il reste du travail.

**Positionner l'évolution de la synthèse de parole à l'heure actuelle et la comparer à celle de la reconnaissance de parole.**

- Les résultats sont très bons mais on peut faire encore mieux.

**Souligner l'importance qu'ont prise les grandes bases de données de parole dans le domaine du traitement de parole.**

- Les systèmes de Machine-Learning demandent de grands corpus. On en dispose maintenant de très grandes étiquetées en intonation, en nature de mots, ...
  Les bases de données sont généralement centralisées dans un laboratoire et on y a accès soit moyennant paiement, soit librement auprès de laboratoires.
  Il y a autant de données audio que de texte dans ces bases de données.

**Prédire l'avenir des scientifiques de la parole.**

- Se baser de moins en moins sur des règles formelles et de plus en plus sur de l'apprentissage automatique. Et sinon, on a besoin de technique d'ingénierie logicielle afin de mélanger les parties des différents spécialistes, car il est impossible d'être spécialiste dans tous les domaines de la parole en même temps.

**Émettre un avis critique sur le traitement de la parole.**

- It's up to you folks !