Reconnaissance Automatique de la parole (RAP)

Master 1 IATI
Avril 2020
Département Informatique
UBMA Université

Reconnaissance Automatique de la Parole

- *Définition*

Transformation d'un signal de parole en une séquence de symboles représentative du contenu du signal.

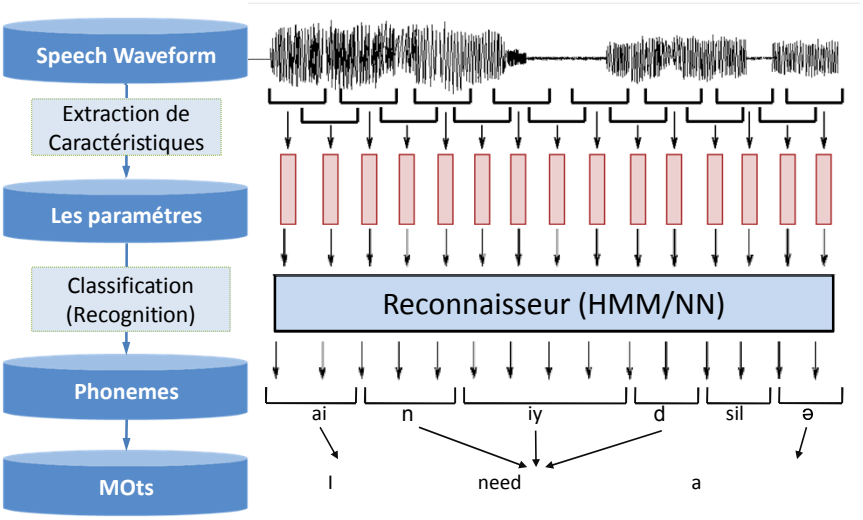
- **Objectifs**

Transformer un signal de parole en :
- Texte (dictée vocale, transcription)
- Action (commande vocale, systèmes de dialogue)
- Information indexée (annotation, indexation)

Types de RAP

Mono locuteur
Multi locuteurs
Indépendant du locuteur

Architecture de la RAP



Processus de reconnaissance Automatique de la parole (RAP)

- Le processus de la RAP comporte deux phases:
- Apprentissage
  - Prétraitement + Extraction de caractéristiques
  - Modélisation
  - Entrée: les sons de parole d'apprentissage (70 a 80%) de la base de sons
- Test
  - Prétraitement + Extraction de caracteristiques
  - Décision (prédiction)
  - Entrée: les sons de parole dde Test (30% a 20%) de la base de sons

Module Prétraitement
La Détection du Silence

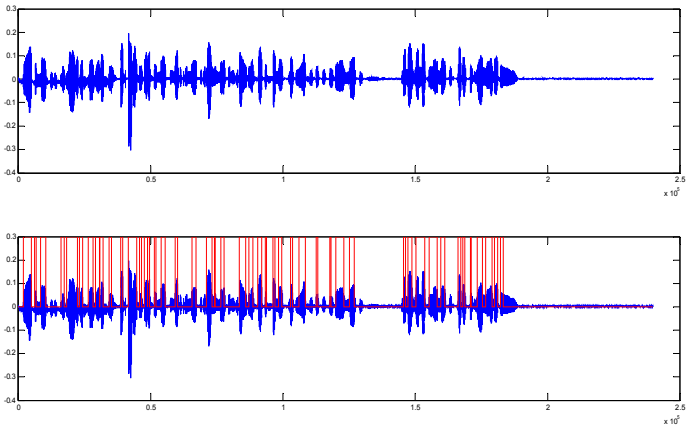
- BUT: récupérer uniquement le signal de parole éliminer le silence et diminuer le bruit.
- L'énoncé de parole d'un locuteur comporte également des trames qui ne sont pas représentatives de la voix de l'auteur de l'énoncé pas porteuses d'informations utiles et par conséquent affecterons les performances du système comme le silence ou le bruit.
- En cette étape, nous procédons à l'élimination des trames inutiles et plus précisément le silence ou les trames de faible énergie.
- Le Taux de passage par zero ZCR et l'energie sont calculées pour évaluer les zones à supprimer.

Détection Parole –Non Parole

- Détection des zones voisées/ non voisées permettent de se concentrer sur les zones de parole.
- Le taux de passage par zero
- Energie (à court terme)

Détection de Parole

signal de parole ; en rouge 1:parole 0:non parole



2- Les méthodes non paramétriques (exemple)

- **Analyse à court terme**

Énergie

$$E(M\cdot n)=\sum_{m=M\cdot n-N+1}^{M\cdot n}\left|x(m)\cdot w(M\cdot n-m)\right|^2$$

Puissance

$$P(M\cdot n)=\frac{1}{N}\sum_{m=M\cdot n-N+1}^{M\cdot n}\left|x(m)\cdot w(M\cdot n-m)\right|^2$$

Amplitude moyenne

$$M(M\cdot n)=\frac{1}{N}\sum_{m=M\cdot n-N+1}^{M\cdot n}\left|x(m)\cdot w(M\cdot n-m)\right|$$

Taux de passages par zéro

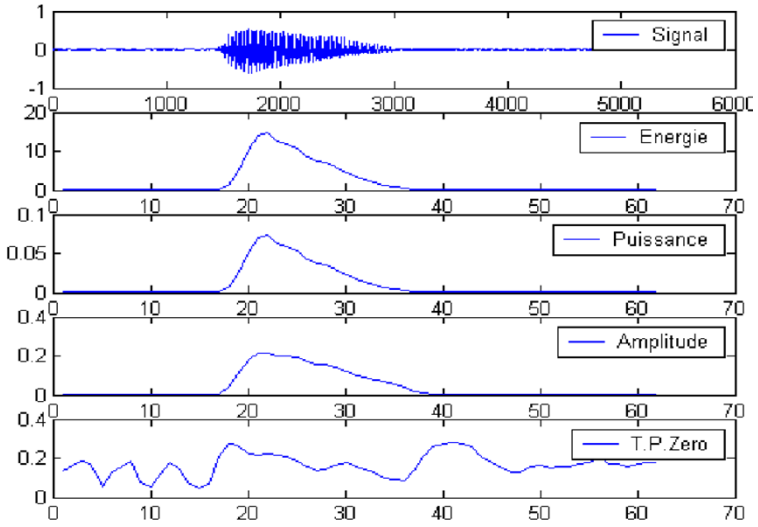
$$Z(M\cdot n)=\frac{1}{N}\sum_{m=M\cdot n-N+1}^{M\cdot n}\frac{\left|\operatorname{sgn}[x(m)]-\operatorname{sgn}[x(m-1)]\right|}{2}$$

8

2- Les méthodes non paramétriques (exemple)

- **Analyse à court terme**

exemple



Analyse acoustique du signal de parole

Microphone → Filtrage analogique → Conversion Analogique /Numérique → Calcul des cœfficients → Trame acoustique

- L'information acoustique pertinente du signal de parole se situe principalement dans la bande passante [50 Hz - 8 kHz] **=>**
  1. Filtrage élimine tous les composants du signal en dehors de cette bande passante
  2. La fréquence d'échantillonnage doit donc au moins être égale à 16 kHz (seulement 8 kHz signal de ligne téléphonique)
  3. Un calcul des coefficients : Une fois le signal de parole échantillonné et numérisé les méthodes d'analyses acoustiques le traitent par bloc d'échantillons de longueur fixe (20 à 40 ms)

=> Résultat : une suite d'observations; chaque observation est un vecteur de coefficients acoustiques associés à la trame paramétrisée ou trame acoustique.

Remarque : Les deux premières étapes sont communes à la plupart des méthodes d'analyse acoustique de parole

Modules Extraction de Caractéristique

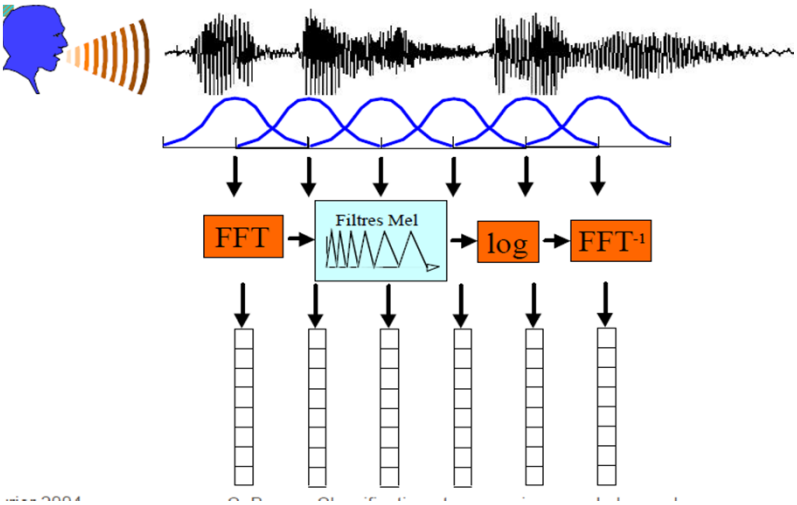
- extraire les paramètres pertinents du signal de parole tels que :

- Les Coefficients Cepstraux MFCC (Mel Frequency Cepstral Coeficient)

- Les coefficients de prédiction linéaires LPCC

Extraction de Caractéristiques [Baras, 2004]

Les Coefficients Cepstraux MFCC (Mel Frequency Cepstral Coeficient)



para métrisation

Différents niveaux de paramétrisation

- Niveau mot :
  - Durée du mot
  - Énergie du mot
- Niveau phonétique :
  - Durée du phonème, Énergie du phonème
  - Taux de passage par zéro
  - Fréquence fondamentale du phonème, Formants
- Niveau acoustique :
  - Mel Frequency Cepstral Coeficients MFCC
  - Linear predictive Model LPCCs
  - Énergie

Les étapes de la RAP: **La Modélisation**

Elle consiste à créer des modèles représentatives des unités de Langage par des approches de la reconnaissance des formes.

Reconnaissance des mots isolées:

Modèle de mots.

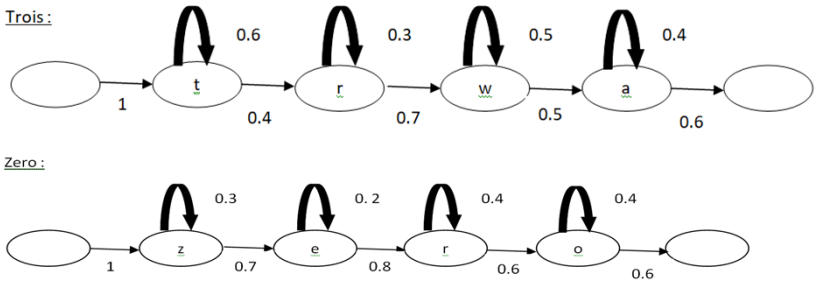
Reconnaissance de la Parole Continue:

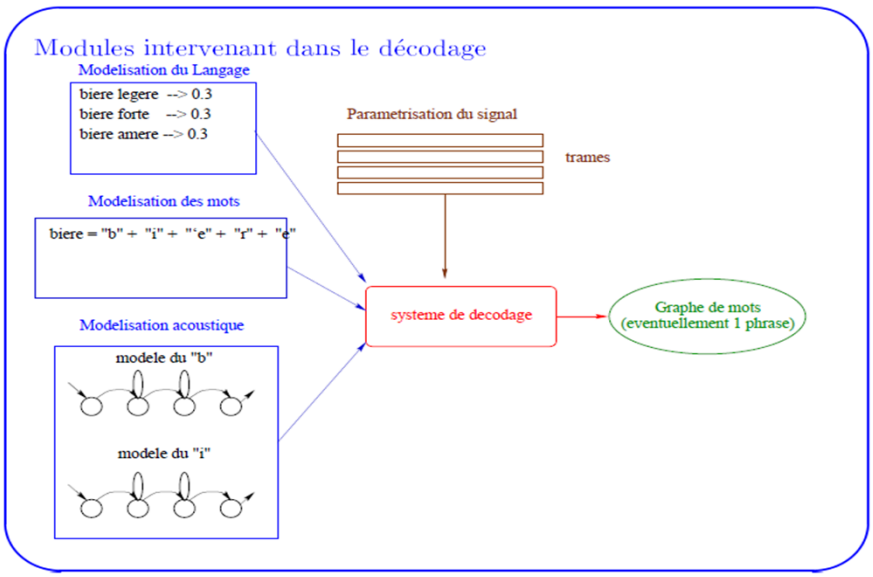
Modèle de phonèmes .

Les approches de Modélisation

- Les Méthodes Statistiques
- Les Modèles de Markov Cachées HMM (Hiden Markov Models
- Dynamic Time warping DTW alignement temporel des mots (cas de reconnaissance de mots isolées)
- Quantification Vectorielle VQ
- Les Réseaux de Neuronnes

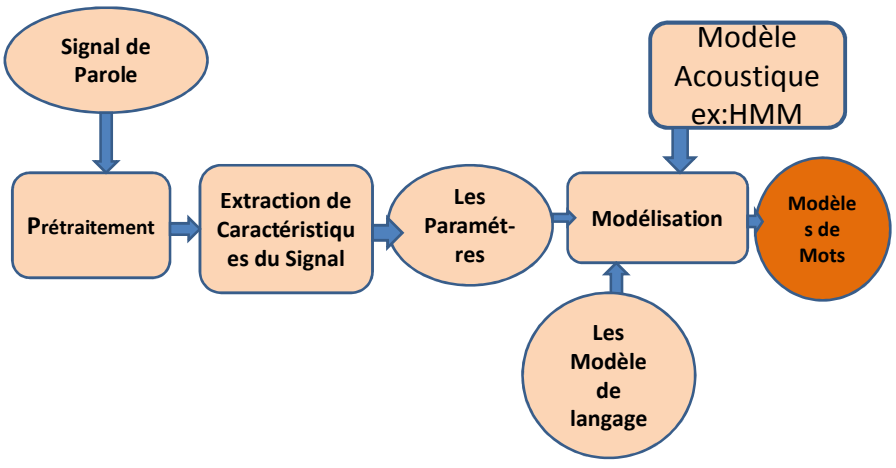
Les Modeles de Markov Chachés (hidden Markov Models) HMM





Bruno Jacob – 2001 DEA Mans

Apprentissage en Reconnaissance RAP



Modèle de Langage

- A un instant donné, tous les mots n'ont pas la même probabilité de présence :
  - Le petit chat boit du …
- Grammaires probabilistes : toutes les phrases sont possibles mais avec des probabilités différentes
- Grammaires à états finis : partition binaire des séquences de mots en « séquences possibles » et « séquences impossibles »

La Grammaire

- **Systèmes à simple commande**
  - La grammaire est une liste de mots.
- **Systèmes à dialogues**
  - La grammaire est un ensemble fini de phrases et de mots (*grammaire à états finis)*

Systèmes à dictée vocale

- La grammaire est *stochastique, elle est définie en terme de probabilités qu'un mot* (bigram) ou qu'un ensemble de mots (n-gram) précèdent un autre mot.

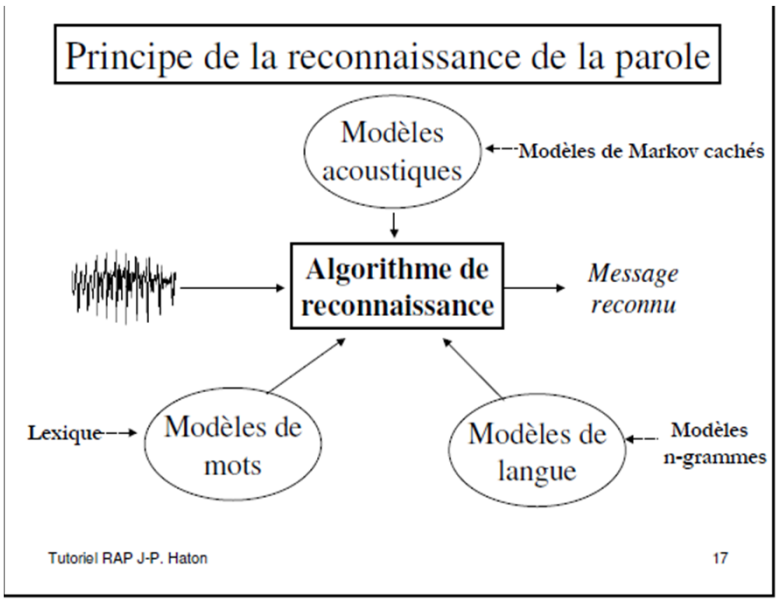
Garammaire exemple

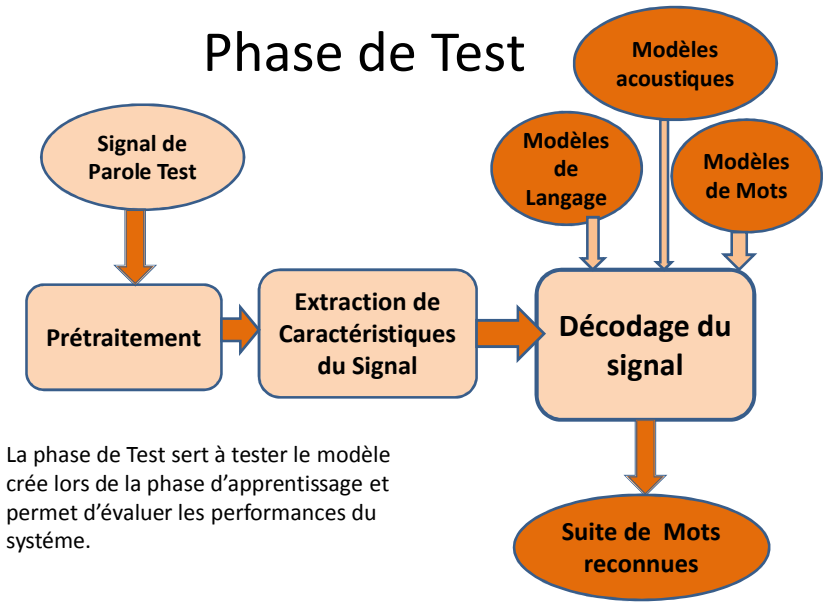
- $chiff = sifr | wahid |ithnane | thalatha | arbaa | khamsa | sita | sabaa | Thamania | tisaa;
- $pause = pause ;$prenom = [Ali [pause]] Reda | salim | [leila [pause]] salim ;
- $numTel = $chiff [$pause] $chiff [$pause] $chiff [$pause] $chiff [$pause] $chiff [$pause] $chiff;
- (SENT-START ($pause (composer | appeler) [$pause] le [$pause] $numeTel [$pause])|

(< $pause appeler [$pause] $prenom > [$pause]) SENT-END)

Format de la grammaire de Java (JSGF)

- **#JSGF V1.0**
- **public <basicCmd> =<startPolite><command><endPolite>;**
- **<command> = <action> <object>;**
- **<action> = open | close | delete | move;**
- **<object> = [the | a] (window | file | menu);**
- **<startPolite> = (please | kindly | could you | oh**
- **mighty computer) *;**
- **<endPolite> = [please | thanks | thank you];**





La phase de Test sert à tester le modèle crée lors de la phase d'apprentissage et permet d'évaluer les performances du systéme.

La Classification

- Le But de cette étape est de reconnaitre les mots prononcées (phonémes).
- Etant donné le signal de parole test
- Etant donné les Modéles de mots M1,M2,…Mn déjà construits
- L'algorithme de reconnaissance efféctue le décodage donc fournie les suites de mots reconnues.

La Classification ou Décodage

- Le signal de parole est l'**observation. On le note X.**
- **X est le résultat de l'extraction de** caractéristiques du signal de parole, càd une séquence de *vecteurs acoustiques.*
- **M est le *modèle associé au mot ou à la séquence de mots reconnue.* {M0, M1, …, MI} est l'ensemble des modèles associés aux mots** constituant le vocabulaire du système.
- **But: choisir le modèle optimal pour minimiser le nombre d'erreurs de** classifications.
- La théorie de Bayes nous dit

$$M = argmax\ P(Mj \mid X)$$

Les Performances de la RAP

- Dans le cas où l'interprétation est nécessaire, les performances sont mesurées par deux taux:

- Le taux d'erreur : performance de la phase de reconnaissance de parole

– Le taux de fausse interprétation: performance de la phase d'interprétation

Types d'erreur

- Si on fait l'hypothèse que l'utilisateur dit quelque chose qui est dans la grammaire (in-grammar), il y a trois types d'erreur:

– Insertion

– Substitution

- Ces erreurs sont pondérées et sommées pour calculer un taux d'erreur représentatif des performances du système

Type d'erreur

- Les Performances sont évaluées suite à la reconnaissance.
- Si l'utilisateur dit quelque chose hors de la grammaire, le système doit être capable de **rejeter l'hypothèse.**
- **Fausse Acceptation**
- **Faux Rejet**
- **Le rapport entre les mots correctement reconnus et le nombre total de mots.**

Type d'erreurs

Types d'erreur	Phrase dite	Phrase reconnue	Décision
Fausse Acceptation d'une phrase hors-grammaire	« Abracadabra »	yes	S > T
Faux Rejet d'une phrase in-grammar	« Yes please »	REJECT	S < T
Fausse reco (Substitution) d'une phrase in-grammar	« Yes please »	no thank you	S > T

Grammaire

[Yes ? Please]

[No? (thank you)]

Référence

- Cours : Reconnaissance de la parole
- **Ivan Magrin-Chagnolleau, CNRS**
- **Laboratoire Dynamique Du Langage**
- **ivan@ieee.org**

Partie 1:

- **Extraction de Caractéristiques**

Extraction de caractéristiques

- Extraire des informations qui permettent de mieux séparer les sons.
- Efficacité en terme de rapidité de calcul et de réduction de la quantité d'information à traiter

Les algorithmes les plus utilisés sont:

- *mel frequency cepstral coeficients (mfcc)*
- *linear predictive cepstral coeficients (lpcc)*

Prétraitement et Paramétrisation

- Le système de paramétrisation utilise, en entrée, le signal de parole et retourne, en sortie, des vecteurs de paramètres à intervalle de temps régulier.
- Le signal de parole n'est pas directement utilisable à cause de sa grande complexité « grande diversité d'information » et de son caractère redondant.
- Le but de la Paramétrisation est d'extraire l'information pertinente pour la tâche proposée.

Paramétrisation

- Cette étape permet donc de transformer un signal de parole en une suite de vecteurs appelés trames.
- Le calcul des paramètres acoustiques est ainsi réalisé en glissant avec une cadence régulière (ex : 10ms) une fenêtre de pondération d'une longueur bien définie sur tout le signal. Généralement, la longueur de la fenêtre de pondération peut varier de 20ms à 30ms.
- De tous les types de fenêtrages en traitement de signal de parole (Hamming, Hanning, Blackman, etc.), celui de Hamming est le plus utilisé. Chaque fenêtre nous permet d'avoir une trame. Les trames obtenues sur tout le signal de parole sont traitées par la suite afin de produire les vecteurs de paramètres acoustiques [KHA, 2002].

Les paramètres issus de l'analyse spectrale

- Dans la littérature, il existe trois grandes familles de paramètres :
- La paramétrisation spectrale est utilisée. Elle représente les caractéristiques physiques de l'appareil phonatoire de chaque individu [CHA, 1997] [HOM, 1995] [REY, 1994].
- Les principaux paramètres de l'analyse spectrale utilisés en RAP sont les coefficients de prédiction linéaire et leurs différentes transformations (LPC (Linear Predictive Coefficients), LPCC (Linear Predictive Cepstral Coefficients), …),
- Les coefficients MFCC (Mel Frequency Cepstral Coefficients)

Les Coefficients Cepstraux de Prédiction Linéaire

- La prédiction linéaire est une technique, issue de l'analyse de la production de la parole [BOI, 2000].
- Il en résulte qu'un échantillon de parole émis à l'instant t peut être estimé à partir des échantillons de parole précédents.
- Les coefficients de prédiction linéaire (Linear Prediction Coefficients – LPC) obtenus sous cette hypothèse Peuvent être utilisés pour calculer des coefficients cepstraux LPCC (Linear Prediction Cepstral Coefficients – LPCC) [TRE, 1982].

Les paramètres prosodiques

- Le terme « paramètres prosodiques » réunit l'énergie, la durée [VAN, 1994] et la fréquence fondamentale (ou pitch) [ATA, 1972].
- Ces paramètres sont souvent associés aux paramètres de l'analyse spectrale (surtout l'énergie). C'est aussi le cas pour la durée et pour la fréquence fondamentale (1).

Les paramètres dynamiques

- Le vecteur de paramètres résultant des paramétrisations précédemment évoquées peut être complété par le vecteur correspondant aux dérivées du premier et second ordre de ces paramètres. Calculées à partir de plusieurs trames adjacentes.
- Ces dérivées permettent d'introduire une information concernant le contexte temporel d'une trame courante [FRE, 2000].
-

Les paramètres MFCCs (Mel Frequency Cepstral Coefficient

- Ces coefficients font partie des paramètres les plus couramment utilisés en traitement de la parole [FUR, 1981] [BIM, 2004]. Ils sont obtenus par une analyse fréquentielle du signal et l'utilisation de bancs de filtres qui permettent de rapprocher l'information extraite de celle perçue par une oreille humaine.

Les coefficients MFCC

- La procédure de calcul des MFCC est la suivante :

– **Decoupage en Trame** (stationnarité), chevauchement (éviter les transitions brusques de trame en trame).

– **Pré-accentuation** (pour donner plus d'énergie puis **fenêtrage** de Hanning (pour la continuité aux bords) ou Hamming.

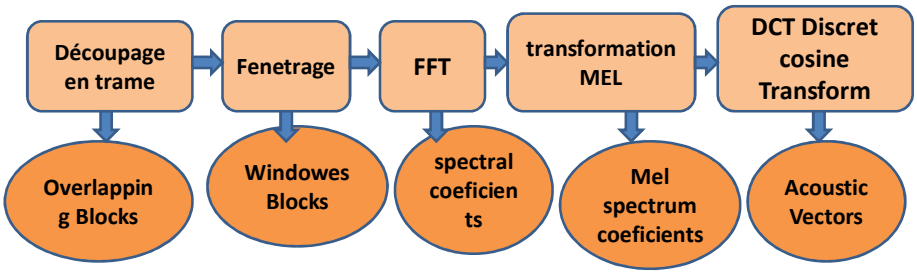
– **Calcul de la TFD sur chaque trame**.

– **Filtrage par un banc de filtres triangulaires répartis le long de l'echelle de Mel.**

– **Calcul du logarithme du module de l'energie** en sortie du banc de filtres.

– **Application de la Transformée en Cosénus Discréte** (joue le rôle d'une TFD inverse).Seuls les premiers coefficients sont conservés.

Extraction de caractéristiques
Les coefficients MFCC



Extraction de coeficients Mel Frequency Cepstral Coefficients

MFCC
Découpage en trames

- Après l'acquisition de la parole, cette opération découpe le signal de parole d'un point de vue pratique en trames de taille fixe « de 20 ms » reparties de façon uniforme le long du signal « toutes les 10 ms », c'est ce que nous appelons le chevauchement.
- L'objectif de cette opération c'est « la stationnarité »
- il a été démontré que les propriétés du conduit vocal peuvent être considérée comme invariantes sur une petite durée égale à 30 ms.
- Il est à noter que des tranches trop courtes ne permettent pas l'analyse spectrale
- Des tranches trop longues « plus de 30 ms » risquent de tomber sur des parties non stationnaires du signal.

Prétraitements
La Détection du Silence

- L'énoncé de parole d'un locuteur comporte également des trames qui ne sont pas représentatives de la voix de l'auteur de l'énoncé pas porteuses d'informations utiles et par conséquent affecterons les performances du système comme le silence ou le bruit.
- En cette étape, nous procédons à l'élimination des trames inutiles et plus précisément le silence ou les trames de faible énergie.
- Le Taux de passage par zero ZCR et l'energie sont calculées pour évaluer les zones à supprimer.

Detection Parole –Non Parole

- Détection des zones voisées/ non voisées permettent de se concentrer sur les zones de parole.
- Le taux de passage par zero
- Energie (à court terme)

2- Les méthodes non paramétriques (exemple)

- **Analyse à court terme**

Énergie

$$E(M\cdot n)=\sum_{m=M\cdot n-N+1}^{M\cdot n}\left|x(m)\cdot w(M\cdot n-m)\right|^2$$

Puissance

$$P(M\cdot n)=\frac{1}{N}\sum_{m=M\cdot n-N+1}^{M\cdot n}\left|x(m)\cdot w(M\cdot n-m)\right|^2$$

Amplitude moyenne

$$M(M\cdot n)=\frac{1}{N}\sum_{m=M\cdot n-N+1}^{M\cdot n}\left|x(m)\cdot w(M\cdot n-m)\right|$$

Taux de passages par zéro

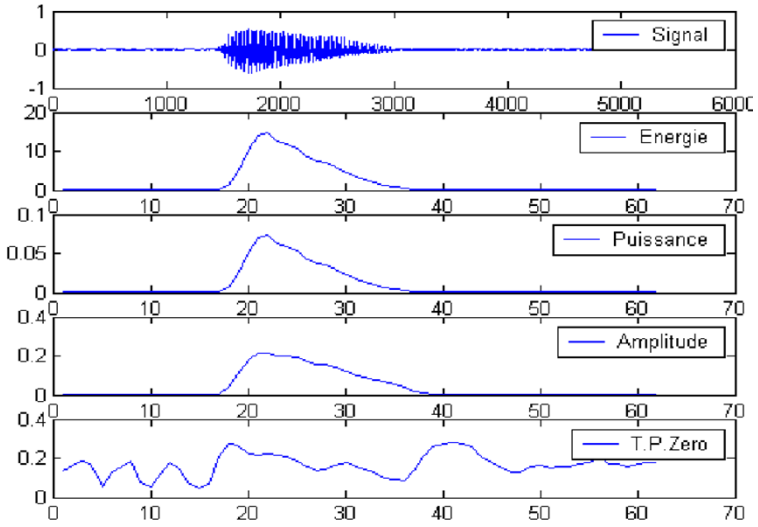
$$Z(M\cdot n)=\frac{1}{N}\sum_{m=M\cdot n-N+1}^{M\cdot n}\frac{\left|\operatorname{sgn}[x(m)]-\operatorname{sgn}[x(m-1)]\right|}{2}$$

46

2- Les méthodes non paramétriques (exemple)

- **Analyse à court terme**

exemple



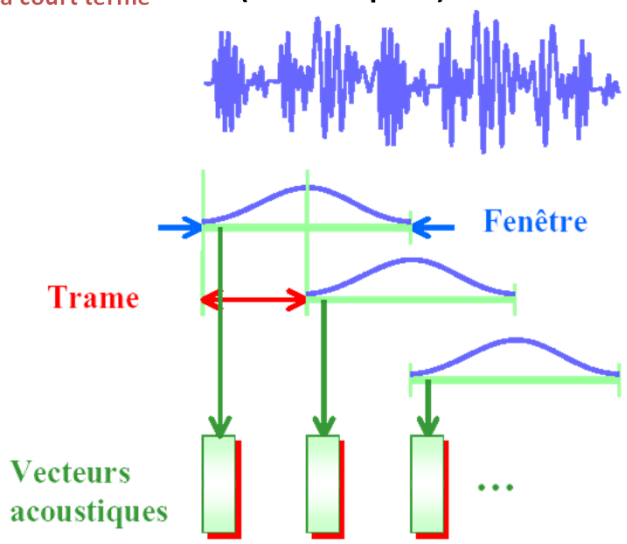
47

Le fenêtrage

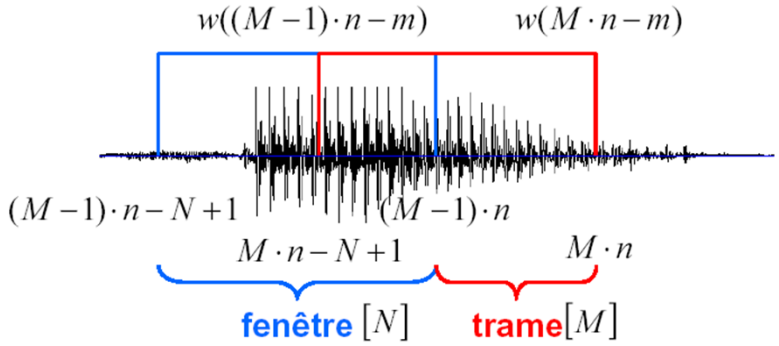
- La phase de découpage en trames à causé des hautes fréquences artificielles, comme nous avons fait passer le signal par un filtre rectangulaire pour en extraire une trame, des déformations spectrales au début et à la fin de chaque trame sont apparu.
- Pour remédier à ces déformations les extrémités des trames ne seront pas coupées directement mais progressivement. Pour cela, nous utilisons la fenêtre de Hamming.
- Par la suite on multiplie cette fonction par le signal à transformer, on minimise ainsi la distorsion spectrale crée par le recoupement.
-

2- Les méthodes non paramétriques (exemple)

- **Analyse à court terme**



2-Analyse à court terme (exemple)



Fenêtre - nombre d'échantillons utilisés pour calculer les paramètres de la trame

Trame - nombre d'échantillons pour lesquels un ensemble de paramètres est valable

Transformée de Fourier rapide

- Cette étape concerne le passage du signal de parole du domaine temporel au domaine fréquentiel, ce passage est réalisé en utilisant la transformé de Fourrier rapide FFT « pour Fast Fourrier Transform ».
- La FFT est appliquée à la trame pour en ressortir la magnitude, on obtient donc le spectre. Aussi, nous appliquons la FFT à chaque trame et aurons comme résultat le spectre d'énergie.

2- Analyse spectrale à court terme (exemple)

- Transformée de Fourier à court terme

$$X(M\cdot n,k)=\frac{1}{N}\sum_{m=M\cdot n-N+1}^{M\cdot n}x(m)\cdot w(M\cdot n-m)\cdot e^{-j(2\pi/N)mk} \qquad k=0,1,...,N$$

- Les propriétés de la transformée de Fourier à court terme dépendent beaucoup du choix de la fonction fenêtre
- La longueur de la fenêtre doit d'une part être suffisante pour assurer une bonne résolution fréquentielle; d'autre part elle doit être limitée si l'on veut suivre fidèlement l'évolution dans le temps du spectre vocal.
- Ces deux exigences sont contradictoires.

scgwww.epfl.ch/JavaSpeechLab2

L'échelle Mel

- Les MFCC s'obtiennent en utilisant, pour le calcul du spectre, une échelle fréquentielle non linéaire tenant compte des particularités de l'oreille humaine, l'échelle des fréquences Mel.
- L'échelle Mel correspond à une approximation de la sensation psychologique de hauteur d'un son qui prend notamment en compte la caractéristique suivante : la sélectivité en fréquence est plus grande dans les graves que dans les aigus

Transformée en cosinus discrète

- Pour finir, on travaille avec le spectre, on convertit le spectre logarithmique de Mel en temps au moyen de la DCT « Discret Cosinus Transform ». Elle agit comme une transformé de fourier FFT inverse.
- Enfin, nous avons extrait les N coefficients MFCC de chaque trame.

Normalisation CMS

- Les paramètres spectraux sont normalisés par l'application d'une soustraction de la moyenne spectrale et d'une normalisation de la variance. Ce processus est appliqué sur chaque fichier d'enregistrement.
- Il a pour but de réduire les bruits stationnaires de convolution et de réduire les effets de la variation entre les environnements des sessions d'apprentissage et de tests.