Traitement Automatique de la Parole

N7 3SN-M

Jérôme Farinas
jerome.farinas@irit.fr

Institut de Recherche en Informatique de Toulouse
Université de Toulouse

10 janvier 2025

TOULOUSE INP N7 UNIVERSITÉ DE TOULOUSE

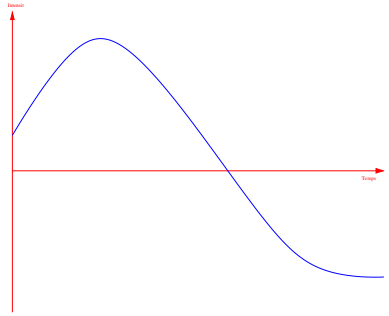
Deuxième partie II

Paramétrisation du signal

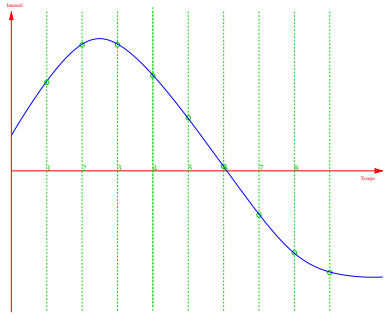
Plan

Numérisation du signal

- Objectif : reproduire un signal sonore analogique dans un fichier informatique numérique



Fréquence d'échantillonnage



- Le signal est échantillonné avec un pas de mesure $T = 1/F_{ech}$

Fréquence d'échantillonnage

Théorème de Shannon

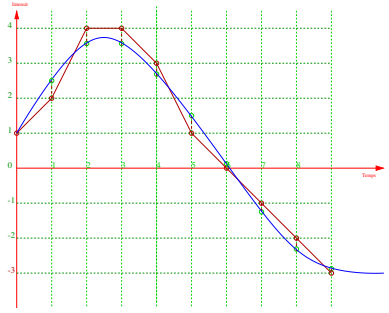
La fréquence d'échantillonnage assurant un non repliement du spectre doit être supérieure à 2 fois la fréquence haute du spectre du signal analogique.

$$F_{ech} = 2 * F_{max}$$

- Signal audio pleine bande à un spectre s'étalant de 20 Hz à 20 kHz
- Bandes de fréquences couramment utilisées pour réduire F_{ech} :

	Spectre du signal	F_{ech}	Applications
Qualité téléphonique	[300-3400 Hz]	8 kHz	Téléphonie
Qualité "bande élargie"	[50-7000 Hz]	16 ou 22 kHz	PC, audio-conférence (ADPCM)
Haute qualité en radiodiffusion	[50 - 15 000 Hz]	32 kHz	DAB, NICAM
Qualité "Hi-Fi"	[20 - 20 000 Hz]	44.1 ou 48 kHz	CD Audio, Studio numérique, DAT

Quantification



- Le signal numérisé est représenté en rouge, l'erreur de quantification en pointillé rouge

Quantification

- Nombre n de bits utilisés pour le codage des échantillons
- Erreur de quantification maximale proportionnelle à $\left(\frac{1}{2}\right)^n$
- Dans le cas d'une quantification linéaire (pas de quantification constant sur toute la plage de conversion), on exprime l'erreur due à la quantification sous la forme d'un rapport Signal à Bruit (SNR : Signal to Noise Ratio) dont l'expression maximale est la suivante :

$$SNR_{dB} = 6.02 * n + 1.76dB$$

	Q	SNR_{dB}
Qualité "Hi-Fi"	16 bits	98
Codage la parole, NICAM	14 bits	86
Codage son PC	8 bits	50

Fichiers informatiques

- Il existe de nombreux formats de fichiers son non compressés
- Parmi les plus courant :
  - raw : fichier binaire nu
  - wav : introduit par Microsoft
  - aiff : utilisé sur Silicon Graphics et Macintosh
  - au : format SUN
- En plus du fichier brut, ces formats encapsulent des informations complémentaires, comme la valeur de la fréquence d'échantillonnage, la quantification utilisée, le nombre de canaux...
- Il existe également des fichiers compressés : .mp3, .flac, .ogg...

Exercices

1. Calculer la taille que prendrait en mémoire 1h14 d'un enregistrement en qualité téléphonique (8 bits, 8 kHz, mono). Même question pour un enregistrement en qualité CD (16 bits, 44.1 kHz, stéréo).
2. Quel est le débit (en bits/s) d'un signal en qualité CD (16 bits, 44.1 kHz, stéréo) ?
3. Si un ordinateur possède une connection internet qui peut fonctionner à 100 ko/s, combien de temps sera nécessaire pour télécharger un fichier d'une heure quatorze qualité CD ?
4. Si le signal suivant a été quantifié sur 5 bits, quelles sont les valeurs du signal numérisé ? 010111101000100
5. Quel est le SNR en décibel d'un signal Blu-Ray Pure Audio (PCM, 24 bits, 192 kHz) ?

Exercices

1. Calculer la taille que prendrait en mémoire 1h14 d'un enregistrement en qualité téléphonique (8 bits, 8 kHz, mono). Même question pour un enregistrement en qualité CD (16 bits, 44.1 kHz, stéréo).
2. Quel est le débit (en bits/s) d'un signal en qualité CD (16 bits, 44.1 kHz, stéréo) ?
3. Si un ordinateur possède une connection internet qui peut fonctionner à 100 ko/s, combien de temps sera nécessaire pour télécharger un fichier d'une heure quatorze qualité CD ?
4. Si le signal suivant a été quantifié sur 5 bits, quelles sont les valeurs du signal numérisé ? 010111101000100
5. Quel est le SNR en décibel d'un signal Blu-Ray Pure Audio (PCM, 24 bits, 192 kHz) ?

Exercices

1. Calculer la taille que prendrait en mémoire 1h14 d'un enregistrement en qualité téléphonique (8 bits, 8 kHz, mono). Même question pour un enregistrement en qualité CD (16 bits, 44.1 kHz, stéréo).
2. Quel est le débit (en bits/s) d'un signal en qualité CD (16 bits, 44.1 kHz, stéréo) ?
3. Si un ordinateur possède une connection internet qui peut fonctionner à 100 ko/s, combien de temps sera nécessaire pour télécharger un fichier d'une heure quatorze qualité CD ?
4. Si le signal suivant a été quantifié sur 5 bits, quelles sont les valeurs du signal numérisé ? 010111101000100
5. Quel est le SNR en décibel d'un signal Blu-Ray Pure Audio (PCM, 24 bits, 192 kHz) ?

Exercices

1. Calculer la taille que prendrait en mémoire 1h14 d'un enregistrement en qualité téléphonique (8 bits, 8 kHz, mono). Même question pour un enregistrement en qualité CD (16 bits, 44.1 kHz, stéréo).
2. Quel est le débit (en bits/s) d'un signal en qualité CD (16 bits, 44.1 kHz, stéréo) ?
3. Si un ordinateur possède une connection internet qui peut fonctionner à 100 ko/s, combien de temps sera nécessaire pour télécharger un fichier d'une heure quatorze qualité CD ?
4. Si le signal suivant a été quantifié sur 5 bits, quelles sont les valeurs du signal numérisé ? 010111101000100
5. Quel est le SNR en décibel d'un signal Blu-Ray Pure Audio (PCM, 24 bits, 192 kHz) ?

Exercices

1. Calculer la taille que prendrait en mémoire 1h14 d'un enregistrement en qualité téléphonique (8 bits, 8 kHz, mono). Même question pour un enregistrement en qualité CD (16 bits, 44.1 kHz, stéréo).
2. Quel est le débit (en bits/s) d'un signal en qualité CD (16 bits, 44.1 kHz, stéréo) ?
3. Si un ordinateur possède une connection internet qui peut fonctionner à 100 ko/s, combien de temps sera nécessaire pour télécharger un fichier d'une heure quatorze qualité CD ?
4. Si le signal suivant a été quantifié sur 5 bits, quelles sont les valeurs du signal numérisé ? 010111101000100
5. Quel est le SNR en décibel d'un signal Blu-Ray Pure Audio (PCM, 24 bits, 192 kHz) ?

Plan

3. Numérisation, codage du signal
4. Analyse du signal
   - Analyse spectrale
   - Fréquence fondamentale (f0)
   - Fenêtrage
   - Analyses temps/fréquence
   - Echelles perceptives
   - Analyse cepstrale
5. Paramétrisation classique de la parole
   - Paramètres MFCC
   - Paramètres LPCC
   - Paramètres PLP
   - Comparaison MFCC/LPCC/PLP/RASTA-PLP
6. Paramétrisations avancées de la parole
   - Normalisation et transformation des paramètres
   - Normalisation par rapport au locuteur
   - Paramètres discriminants
   - Représentation par sous-espaces latents

Plan

3. Numérisation, codage du signal
4. Analyse du signal
   - Analyse spectrale
   - Fréquence fondamentale (f0)
   - Fenêtrage
   - Analyses temps/fréquence
   - Echelles perceptives
   - Analyse cepstrale
5. Paramétrisation classique de la parole
   - Paramètres MFCC
   - Paramètres LPCC
   - Paramètres PLP
   - Comparaison MFCC/LPCC/PLP/RASTA-PLP
6. Paramétrisations avancées de la parole
   - Normalisation et transformation des paramètres
   - Normalisation par rapport au locuteur
   - Paramètres discriminants
   - Représentation par sous-espaces latents

Décompositon du signal en séries de Fourier

Théorème de Dirichlet

Soit $s : \mathbb{R} \to \mathbb{C}$ avec un nombre fini de discontinuités
s est périodique de période $T_0 = \frac{1}{f_0} = \frac{2\pi}{w_0}$
alors

$$s(t) = \sum_{k=0}^{+\infty} [a_k \cos kw_0 t + b_k \sin kw_o t]$$

avec $a_k = \frac{w_0}{2\pi} \int^{\Delta} s(t) \cos(kw_0 t) dt$ et $b_k = \frac{w_0}{2\pi} \int^{\Delta} s(t) \sin(kw_0 t) dt$
et Δ l'intervalle de longueur T_0

Transformée de Fourier I

Transformée de Fourier

$$S(w) = \int_{-\infty}^{+\infty} (s(t)\cos(wt) - i.s(t)\sin(wt))dt = \int_{-\infty}^{+\infty} s(t)e^{-iwt}dt$$ avec $s(t)$ un signal réel continu

- $s(t)$ est remplacé par son spectre complexe $S(w)$
- cette opération est réversible :

Transformée de Fourier inverse

$$s(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} S(w)e^{iwt}dw$$

- la transformée de Fourier fournit deux descriptions duales, temporelle et fréquentielle d'un signal

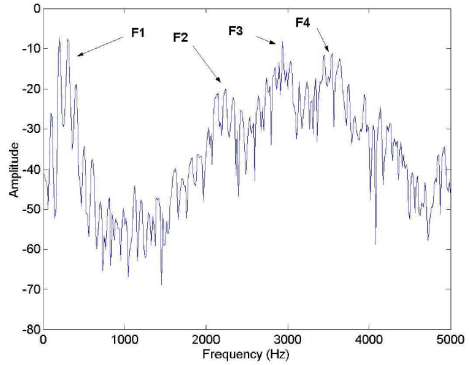
Transformée de Fourier II

- le spectre complexe peut s'écrire :
  $S(w) = R(w) + i.I(w) = A(w)e^{i\phi(w)}$
  où :
  - $A^2(w) = R^2(w) + I^2(w)$ est le spectre de puissance du signal $s(t)$
  - $\phi(w) = \arctan(I(w)/R(w))$ le spectre de phase
- La transformée de Fourier permet de décrire un signal continu dans l'espace des fréquences

Transformée de Fourier Discrète

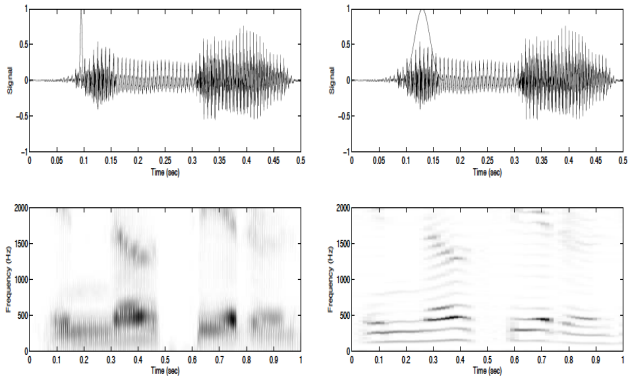
- cas des signaux discrets
- TFD : $s(n) \rightarrow S(w) = \sum_{n=0}^{N-1} s(n)e^{-iwn}$
- algorithme de calcul rapide : FFT (Fast Fourier Transform)
  - principe de base : calculer un ensemble de TFD pour des valeurs inférieures à N et combiner les résultats obtenus
  - complexité : $N^2 \rightarrow N \log_2 N$

Illustration Spectre



spectre de la voyelle [i]

Spectre bande large/étroite



Plan

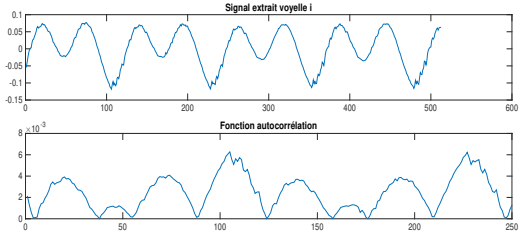
Fréquence fondamentale (f0)

- fréquence de vibration des cordes vocales
- correspond à la hauteur de la voix
- intonation (mélodie) = variation dans le temps de la f0
- fréquence de variation moyenne (en Hz) :
  - 100-150 : hommes
  - 150-250 : femmes
- cette valeur moyenne varie tout au long de la vie (hommes et femmes)
- lié à la physiologie des cordes vocales (taille 60% plus grande chez les hommes que les femmes)

Extraction f0 : méthode autocorrélation

- algorithme de type corrélation [Hess 1993, pp. 351-356]
- hypothèse : signal stationnaire

$$f_{autoc}(\tau) = \frac{1}{n}\sum_{i=1}^{n-\tau} s_i s_{i+\tau}$$



Extraction f0 : méthode AMDF

- fonction de distance [Miler et Weibel 1956]
- critère de variation d'amplitude à court terme (Average Magnitude Difference Fonction)

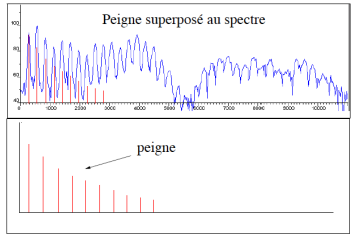
$$f_{AMDF}(\tau) = \frac{1}{n-\tau}\sum_{i=1}^{n-\tau} |s_i - s_{i+\tau}|$$

- ambiguïté entre les pics T_0 , $2 * T_0$... atténué par la non stationnarité du signal
- résiste aux ambiguïté d'octave
- rapidité de calcul

Extraction f0 : peigne spectral

- intercorrélation entre le spectre du signal et une série d'harmoniques d'impulsions de Dirac d'amplitude normalisée (« peigne ») [Martin 1981]

$$f_{peigne}(\omega) = \sum_{i=1}^{n(\omega)} \alpha_i |S(i * \omega)|$$



Fenêtrage

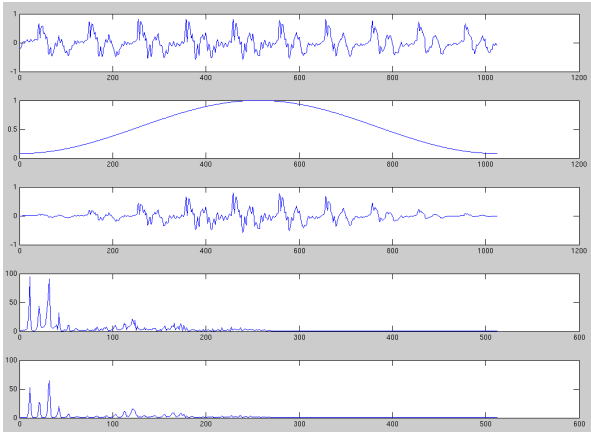
- signal quasi stationnaire sur fenêtres de 10 à 30 ms
- pour limiter effets de bord (phénomène de Gibbs) → pondération par fenêtre temporelle aplatie aux extrémités

Fenetre de Hamming

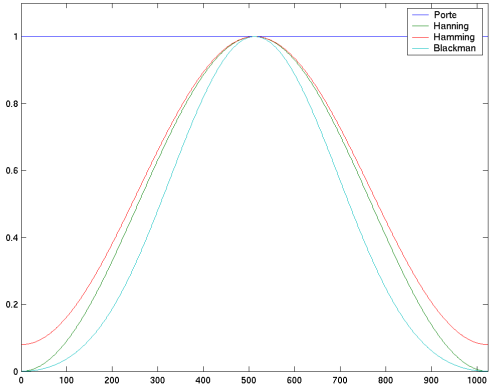
$$h(n) = \begin{cases} 0,54 - 0,46 * cos(2\pi \frac{n}{N-1}) \; si 0 \leq n \leq N-1 \\ 0 \; sinon \end{cases}$$

où N est la taille de la fenêtre en échantillons.

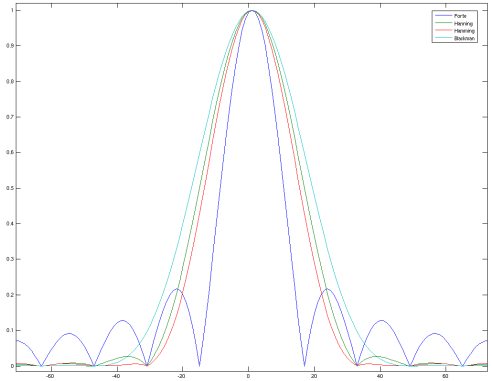
Illustration du fenêtrage de Hamming



Fenêtrages : représentation temporelle

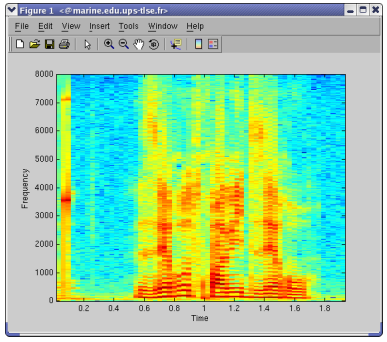


Fenêtrages : représentation fréquentielle



Spectrogramme

```
specgram(signal,1024,fs,hamming(1024)) ;
```



Echelles perceptives

- Echelle MEL :

$$M = \frac{1000}{\log 2} \log(1 + \frac{F}{1000})$$

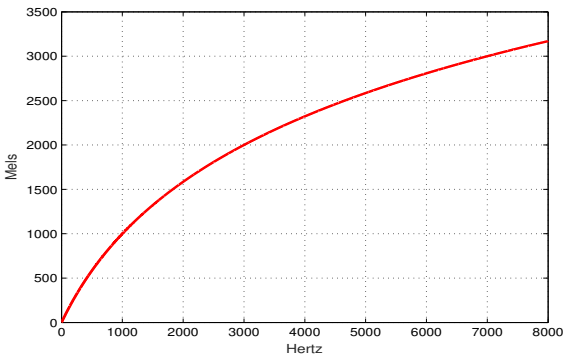
avec F en Hz et M en Mel

- Echelle Bark [Hartmann 97] :

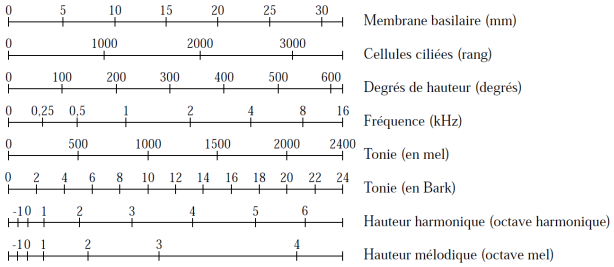
$$B = \frac{26,81 * F}{1960 + F} - 0,53$$

avec F en Hz et B en $Bark$
facteurs de correction :
  - si $B < 2$ alors $B' = B + 0,15 * (2 - B)$
  - si $B > 20,1$ alors $B' = B + 0,22 * (B - 20,1)$

Illustration fonction MEL



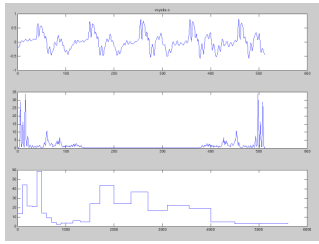
Illustrations echelles perceptives



source : Zwicker 1981

Illustration canaux MEL

Fréquences de coupure de la pondération triangulaire : 0,100,200,300,400,500,600,700,800,900,1000,1150,1300,1500,1700,2000, 2350,2700,3100,3550,4000,4500,5050,5600,6200,6850,7500 *Hz*



1. Fenêtre d'analyse (512 points de signal)
2. Module du spectre sur 256 points
3. Canaux d'énergies dans les 24 bandes réparties selon les canaux MEL

Plan

Calcul du cepstre

Considérons que le signal de parole est issu de la convolution de $f(t)$ (source) et du conduit vocal $c(t)$:

$$s(t) = f(t) \otimes c(t) \quad (1)$$

Passage dans le domaine fréquentiel avec une transformée de Fourier :

$$\hat{s}(w) = S(w) = \hat{f}(w) * \hat{c}(w) \quad (2)$$

Passage en log :

$$\log \hat{s}(w) = \log \hat{f}(w) + \log \hat{c}(w) \quad (3)$$

Retour dans le domaine temporel en appliquant une transformée de Fourier inverse :

$$\widehat{\log \hat{s}}(t) = \hat{\hat{s}}(t) = \hat{\hat{f}}(t) + \hat{\hat{c}}(t) \quad (4)$$

Illustration cepstre I

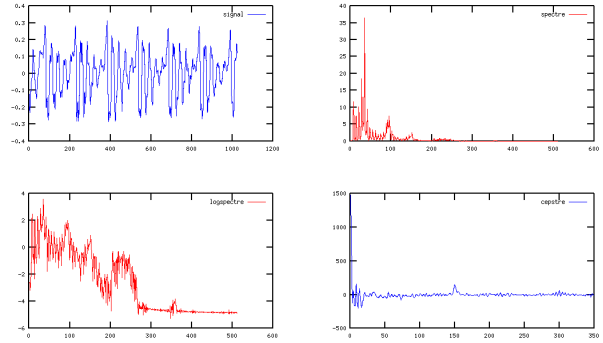
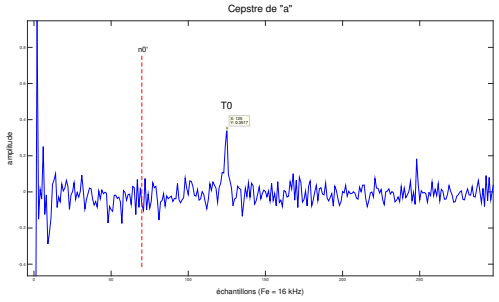
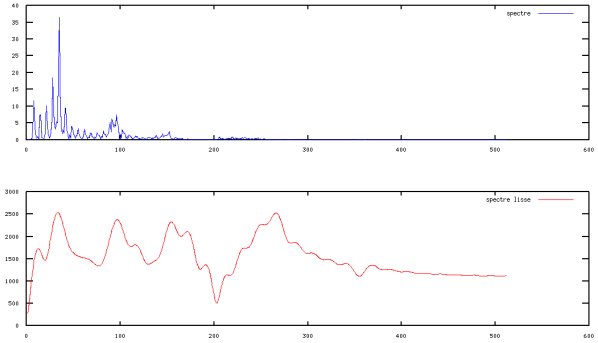


Illustration cepstre II



Lissage du spectre



Débruitage (1)

La combinaison des signaux parole et bruit est linéaire dans le domaine temporel :

$$y(t) = x(t) + b(t)$$

Fenêtrage et transformation de Fourier :

$$Y(w) = X(w) + B(w)$$

L'additivité n'est plus conservée lors du passage au spectre de puissance :

$$|Y(w)| = |X(w)| + |B(w)| \cos(\phi)$$

où ϕ est le déphasage entre le bruit et la parole propre.
L'hypothèse $\cos(\phi) \backsimeq 1$ est souvent faite.

$$|Y(w)| \backsimeq |X(w)| + |B(w)|$$

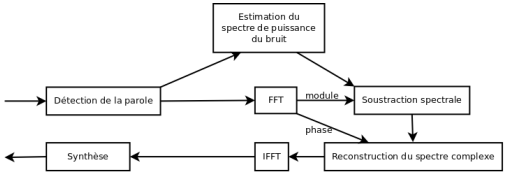
Débruitage (2)

La soustraction spectrale revient donc a calculer une estimée du spectre de puissance du bruit $|\hat{B}(w)|$ puis à soustraire cette estimée du spectre du signal observé.

$$|\hat{X}(w)| = |Y(w)| - |\hat{B}(w)|$$

Le spectre débruité peut alors être utilisé :

- pour construire des paramètres MFCC
- pour reconstruire le signal temporel de la parole débruitée.



Plan

Plan

Paramètres MFCC

- Mel Frequency Coefficient Cepstra
- Coefficients extraits sur environ 20-40 ms (25 : valeur standard)
- Schéma général :
  1. Préaccentuation des aigus (filtre $H(z) = 1 - 0.98z^{-1}$)
  2. Fenêtrage de Hamming
  3. Transformée de Fourier Rapide pour obtenir le Spectre de puissance (sur 512 points)
  4. Pondération triangulaire suivant l'échelle Mel (26 canaux)
  5. Passage en log de tous les canaux
  6. Transformée en Cosinus Discrète (permet de décorréler les canaux)
  7. Garder les coefficients 2 à 13

Chaîne MFCC

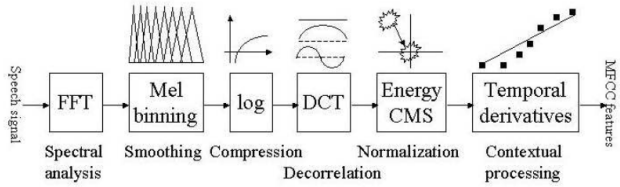
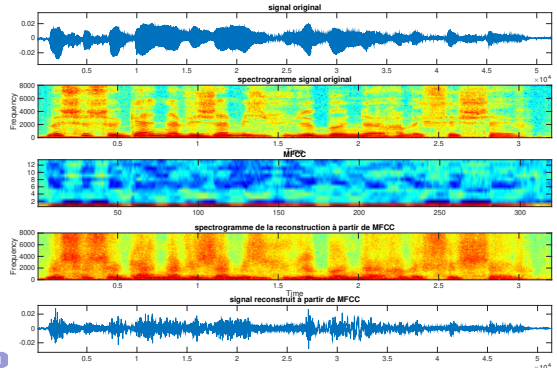
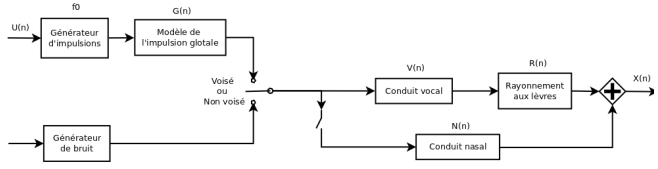


Illustration MFCC



Plan

Modèle de production I



Hypothèses :

- le larynx se comporte comme un filtre passe bas d'ordre 2 :

$$G(z) = \frac{A}{(1 + \alpha z^{-1})(1 + \beta z^{-1})}$$

Modèle de production II

- le conduit vocal peut-être assimilé à une succession de tubes acoustiques élémentaires. La modélisation se fera par une cascade de résonateurs de deuxième ordre dont la fonction de transfert est de la forme :

$$V(z) = \frac{B}{\prod_{k=1}^{K} (1 + b_{1k}.z^{-1})(1 + b_{2k}z^{-2})}$$

- le rayonnement aux lèvres peut-être caractérisé par la fonction de transfert suivante :

$$R(z) = \sigma(1 - z^{-1})$$

Modèle de production III

La fonction de transfert du système global est le produit de ces fonctions de transfert :

$$H(z) = \frac{X(z)}{U(z)} = G(z)\,V(z)\,R(z)$$

$$H(z) = \frac{AB\sigma\left(1-z^{-1}\right)}{\left(1+\alpha z^{-1}\right)\left(1+\beta z^{-1}\right)\prod_{k=1}^{K}\left(1+b_{1k}.z^{-1}+b_{2k}z^{-2}\right)}$$

En faisant l'hypothèse supplémentaire que l'un des pôles $G(z)$ est voisin de l'unité, on obtient l'approximation de la fonction de transfert $H(z)$:

$$H(z) = \frac{\mu}{\left(1+\alpha z^{-1}\right)\prod_{k=1}^{K}\left(1+b_{1k}.z^{-1}+b_{2k}z^{-2}\right)} = \frac{\mu}{A(z)}$$

avec :

$$A(z) = \left(1+\alpha z^{-1}\right)\prod_{k=1}^{K}\left(1+b_{1k}.z^{-1}+b_{2k}z^{-2}\right)$$

Modèle de production IV

En développant cette relation :

$$A(z) = 1 + \sum_{k=1}^{2K-1} a_k z^{-k}$$

Donc :

$$H(z) = \frac{X(z)}{U(z)} = \frac{1}{1+\sum_{k=1}^{p} a_k z^{-k}}$$

avec $p = 2K-1$ l'ordre du filtre.
Cherchons à déterminer l'expression de l'échantillon de signal x à l'instant n à l'aide de la fonction de transfert en prenant $\mu = 1$:

$$\frac{X(z)}{U(z)} = \frac{1}{1+\sum_{k=1}^{p} a_k z^{-k}}$$

Modèle de production V

$$X(z)\left(1+\sum_{k=1}^{p} a_k z^{-k}\right) = U(z)$$

$$X(z) + \sum_{k=1}^{p} a_k z^{-k} X(z) = U(z)$$

$$x(n) + \sum_{k=1}^{p} a_k x(n-k) = u(n)$$

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + u(n)$$

l'échantillon de signal x à l'instant n peut-être déterminé par les p échantillons antérieurs à n moyennant la connaissance du processus u (excitation qui génère le signal x). Ce modèle de production de signal est appelé autorégressif.

Modéle Autorégressif de parole

Modèle AR

Le signal à l'instant n peut être prédit à partir des instants précédents

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + u(n)$$

$$x(n) = -a_1 x(n-1) - a_2 x(n-2) - a_3 x(n-3) - \ldots - a_P x(n-P) + u(n)$$

Pour la parole on adopte généralement un ordre p égal à 10.

Paramètres LPCC

- Estimation du modèle AR

$$x(n)+a_1 x(n-1)+a_2 x(n-2)+a_3 x(n-3)+\ldots+a_P x(n-P) = u(n)$$

  - avec $(a_1, a_2, ..., a_p)$ les coefficients de prédiction linéaire (LPC)
  - $u(n)$ erreur de prédiction, $u(n) \rightsquigarrow N(0, \sigma^2)$
- Equations de Yule-Walker : $R_i = E(x_n x_{n-i})$

$$\begin{bmatrix} R_0 & R_1 & R_2 & \cdots & R_p \\ R_1 & R_0 & R_1 & & \vdots \\ R_2 & R_1 & R_0 & & \vdots \\ \vdots & & & \ddots & \vdots \\ R_p & \cdots & \cdots & \cdots & R_0 \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

- LPC + cepstre = LPCC

Plan

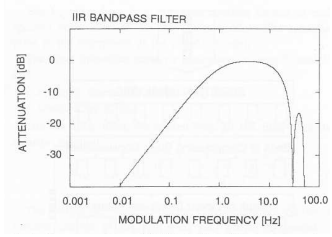
Paramètres PLP

- Perceptual Linear Predictive analysis
- Paramètres extraits toutes les 20 ms
- Schéma général :
  1. Fenêtrage de Hamming
  2. Transformée de Fourier Rapide pour obtenir le Spectre de puissance (sur 256 points)
  3. Extraction de bandes critiques suivant échelle Bark
  4. Pré-accentuation perceptuelle (courbe isosonie)
  5. Compression d'intensité en sonie $|.|^{1/3}$
  6. Transformée de Fourier inverse
  7. Analyse par prédiction linéaire (récursion de Durbin et récursion cepstrale)

Filtrage RASTA

RelAtive SpecTral Analysis

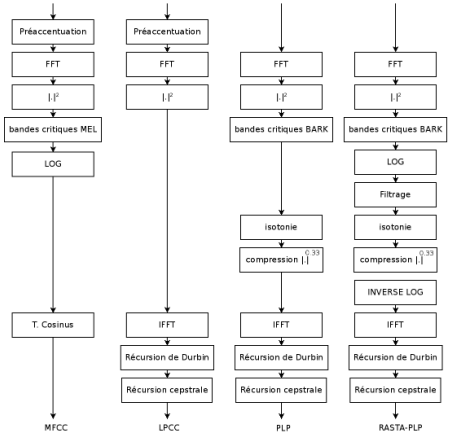
Eliminer les variations trop lentes ou trop rapides par filtrage sur le spectre d'amplitude afin de ne conserver que les variations liée au signal vocal produit par les humains (variations autour de 4 Hz)



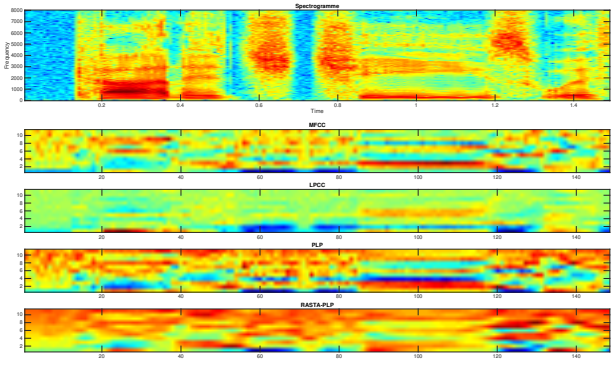
Réponse en fréquence d'un filtre RASTA qui sera appliqué à chaque bande critique extraite du signal

Plan

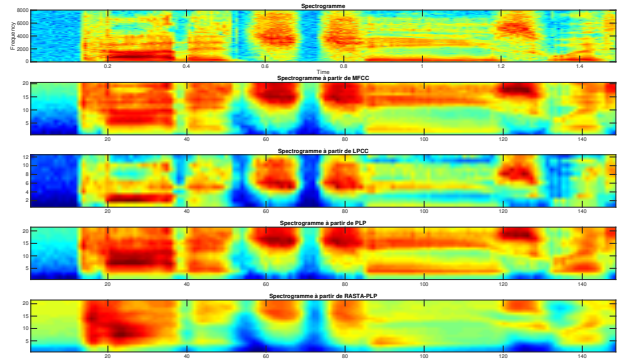
Comparaison MFCC/LPCC/PLP/RASTA-PLP



Comparaison paramètres



Comparaison spectrogrammes



Exercices

1. Quelle est la longueur en secondes d'un signal ($F_{ech} = 16000Hz$) contenant 35867 échantillons ?
2. Quelle est la durée en milisecondes d'une trame de 512 échantillons ? ($F_{ech} = 16KHz$)
3. Sur un signal échantillonné à $16KHz$, combien faut-il d'échantillons pour constituer une trame de 10 ms ?
4. En utilisant une fenêtre glissante (recouvrement sur la moitié) de 256 échantillons pour calculer un vecteur d'observation, combien de vecteurs va-t-on obtenir sur le signal de la question 1 ?
5. Si on utilise une paramétrisation constituée : de l'énergie, de 12 MFCC, de leur dérivées premières et seconde, quelle sera la dimention des vecteurs d'observation ?

Plan

Plan

Normalisation et transformations des paramètres

- CMS sur les phrases et CVN sur les locuteurs
- Δ et $\Delta\Delta$ [Furui 1986] ont été remplacé par une matrice de projection linéaire : analyse discriminante linéaire (LDA) [Fukunaga 1990]
- les paramètres LDA peuvent être également modifiés par rotation des covariances : transformation Semi Tied Covariance (STC) [Gales 1998]
- LDA+STC reduit le taux erreur sur les mots (WER) de 10-15% sur un système grand vocabulaire par rapport à Δ et $\Delta\Delta$

Soustraction moyenne cepstrale (CMS)

- soustraction de la moyenne du cepstre (« Cepstral Mean Soustraction ») sur chaque trame du segment de parole considéré (par exemple fenêtre glissante d'une à deux secondes)
- normalisation simple qui permet de rendre plus robuste les paramètres de parole
- permet de s'affranchir des bruits additifs dans le domaine log-spectral (et donc multiplicatif dans le domaine cepstral)
- permet d'apporter une meilleure stabilité temporelle au signal

Normalisation variance cepstrale (CVN)

- Cepstral Variance Normalisatioin (CVN)
- normalisation de chaque trame acoustique pour rendre la variance unitaire

$$\hat{x} = \sqrt{\Lambda^{-1}}.(x - m)$$

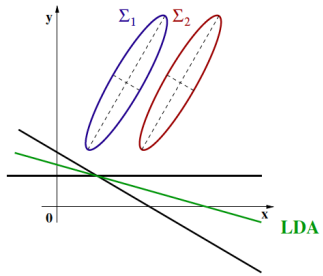
avec m le vecteur cepstral moyen, et $\Lambda^{-1} = [1/\sigma_1^2, \ldots, 1/\sigma_n^2]$ l'inverse de la matrice de covariance diagonale calculée sur une fenêtre glissante de taille finie (par exemple 1 seconde)

- permet de réduire la sensibilité au bruit additif, s'adapte rapidement aux changements de bruit
- doit être appliqué sur une durée la plus longue possible : par exemple sur le tour de parole d'un locuteur

Modélisation des changements temporels

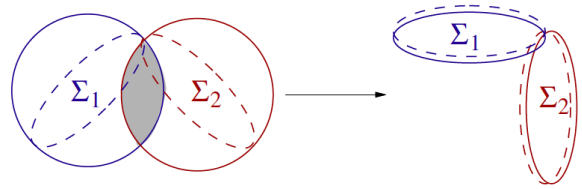
- Contruction d'un supervecteur qui concatène N (par exemple 9) trames consécutives
- Projection de ce supervecteur sur une dimension plus faible
- Utilisation de transformations linéaires :
  - pour simuler l'estimation par moindres carrés des dérivées temporelles
  - pour maximiser la discrimination entre les classes phonétiques

Analyse discriminante linéaire (LDA)



- Transformation de l'espace de représentation de façon à discriminer au maximum les classes
- Meilleurs résultats avec des classes correspondantes aux unités des HMM

Transformation par rotation de la matrice de covariance (STC)



- Problème : les paramètres finaux seront modélisés avec des matrices de covariances diagonales
- la transformation Semi Tied Covariance (STC) effectue une rotation de telle façon à rendre plus valide la modélisation [Gales 1999]

Plan

Normalisation par rapport au locuteur

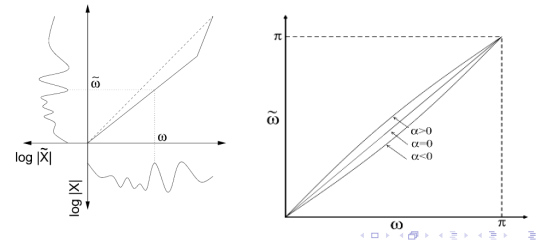
- Variations intra et inter locuteurs
- La normalisation des paramètres des locuteurs cherche à obtenir une paramétrisation canonique en éliminant la variabilité inter locuteur
  - distorsion de l'axe des fréquences pour se caler sur la longueur du conduit vocal de locuteurs de référence (VTLN)
  - transformation afine des paramètres pour maximiser la vraisemblance du modèle de locuteur courant (fMLLR)
  - Gaussianisation des paramètres
- VTLN+fMLLR peut réduire le WER d'un système de reconnaissance de la parole de 20-30%

Normalisation du conduit vocal

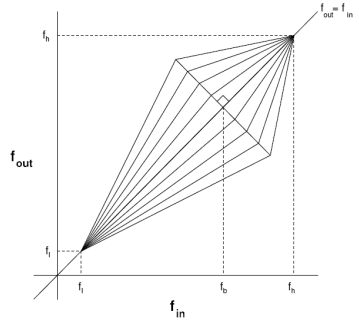
Vocal Tract Length Normalisation (VTLN)

La normalisation du conduit vocal est une technique de normalisation des locuteurs qui réduit l'influence de la longueur du conduit vocal. Il s'agit d'une transformation linéaire dans le domaine cepstral.

$$\omega \rightarrow \tilde{\omega} = g_\alpha(\omega)$$



Fonction VTLN linéaire par morceaux



- fonction linéaire par morceaux avec une distorsion de $\pm 20\%$

Adaptation fMLLR

Adaptation des paramètres à la voix d'un locuteur

feature space Maximum Likelihood Linear Regression

- appliquer une transformation affine sur les paramètres

$$\hat{\mu} = A\mu + b$$

- estimation de A par maximisation de la vraisemblance du modèle acoustique
- application de la transformation A sur la matrice de covariance

Plan

Paramètres discriminants

- Paramètres minimisation l'erreur sur les phones (fMPE) [Povey 2005]
  - fMPE et MPE peut réduire le WER de 25%
- Paramètres « Bottleneck » extraits d'un réseau de neurone à 5 couches avec une couche d'étranglement [Grezl 2007]

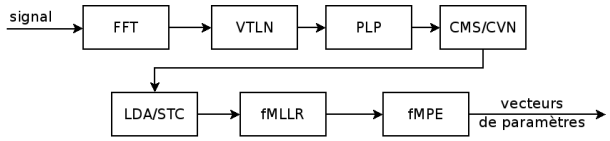
Paramètres minimisant l'erreur sur les phones (fMPE)

- Feature space Minimum Phone Error [Povey 2005]
- L'estimation MPE a été créé pour minimiser l'erreur sur les phonèmes des modèles acoustiques
- La transformation fMPE applique une transformation sur les paramètres pour minimiser l'erreur de reconnaissance d'une système de reconnaissance de la parole
- Chaque trame de parole est modifiée

$$\hat{x} = x_t + M^f h_t$$

avec M^f la matrice de transformation et h_t des probabilités à posteriori de Gaussiennes en haute dimension

Paramétrisation complête



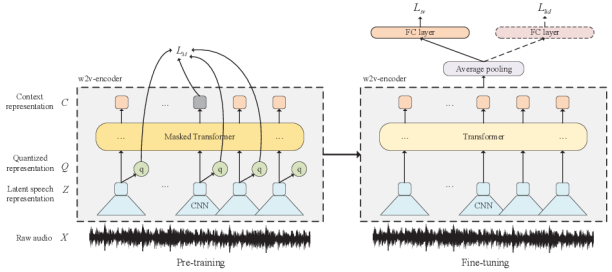
Plan

Représentation par sous-espaces latents

- La révolution des réseaux de neuronnes profonds à bouleversé la manière de représenter l'information en audio a partir de 2010.
- Il estpossible de faire entrer directement le signal de parole dans les systèmes de reconnaissance
- Il est alors possible d'obtenir une représentation à un certain niveau de profondeur des systèmes de neurones profonds qui sont optimisés pour des tâches de reconnaissance de la parole.
- On peut utiliser ces sous-espaces pour alimenter des systèmes de traitement de la parole :
  - transcription de parole
  - reconnaissance du locuteur
  - reconnaissance des émotions
  - etc.
- Quelques représentations célèbres :
  - Wav2vec
  - HuBert
  - PASE+

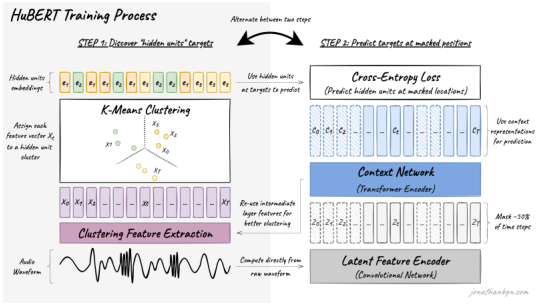
Wav2vec 2.0

- Facebook AI [Baevski, Zhou, Mohamed, Auli 2020]
- pré-entraîner un grand réseau sur des données non étiquetées pour apprendre des représentations contextuelles utiles de la séquence texte/audio
- utiliser ces représentations pour prédire le futur d'un signal audio



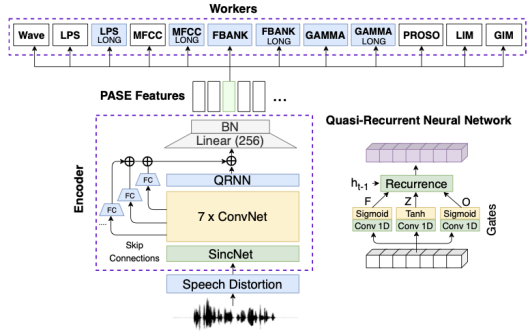
HuBERT

- Meta AI [Hsu, Bolte, Tsai, Lakhotia, Salakhutdinov, Mohamed, 2021]
- Unités cachées discrètes (Hu=HiddenUnit) pour transformer les données vocales en une structure plus "proche du langage"



PASE+

- [Ravanelli, Zhong, Pascual, Swietojanski, Monteiro, Trmal, Bengio,2020]
- robuste bruit et réverbération, nombreuses tâches de spécialisation

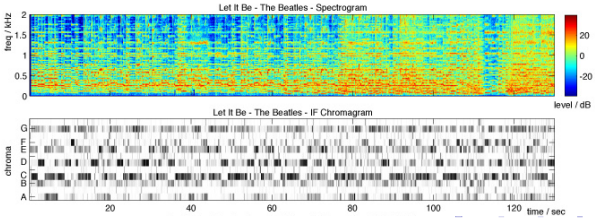


Plan

Chromagramme

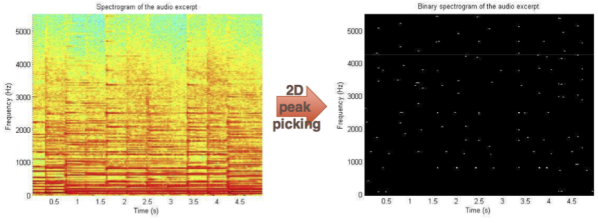
- représentation en 12 demi-tons (chroma)

- repliement sur la première octave des notes des autres octaves



Algorithme Shazam

- Extraction de points saillants dans le plan temps/fréquence
  - onset des sinusoïdes, maxima locaux
  - « constellation points »



Algorithme Shazam : empreintes

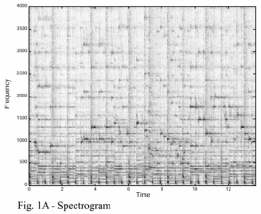


Fig. 1A - Spectrogram

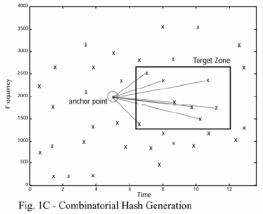


Fig. 1C - Combinatorial Hash Generation

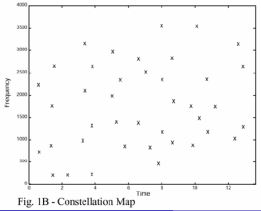


Fig. 1B - Constellation Map

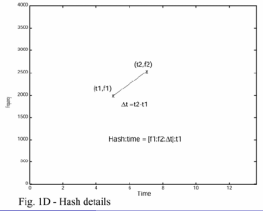


Fig. 1D - Hash details