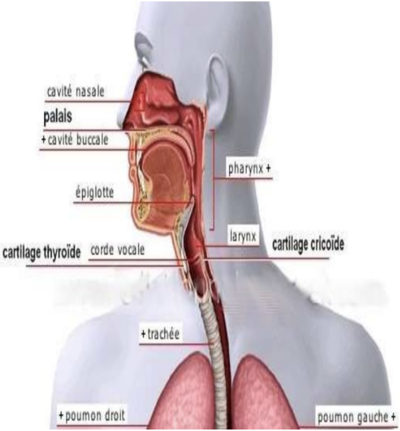
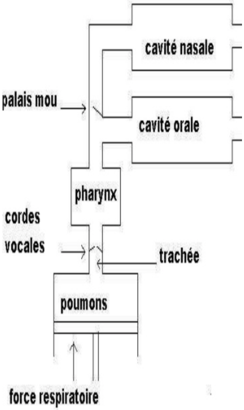
Audio Signal Processing : V. Speech processing

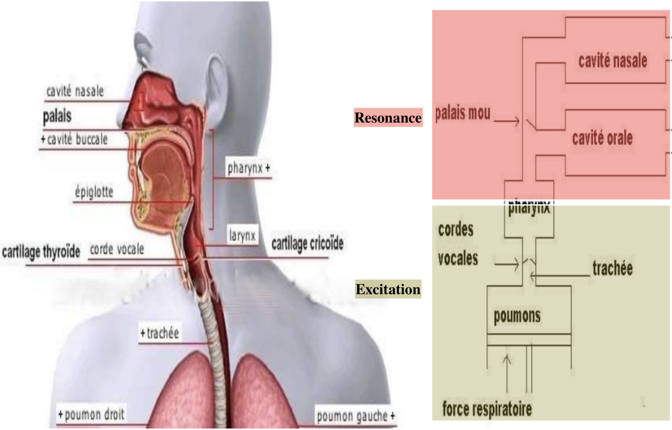
Emmanuel Bacry

bacry@ceremade.dauphine.fr
http://www.cmap.polytechnique.fr/~bacry


cavité nasale
palais
+ cavité buccale
pharynx +
épiglotte
larynx
cartilage cricoïde
cartilage thyroïde
corde vocale
+ trachée
+ poumon droit
poumon gauche +

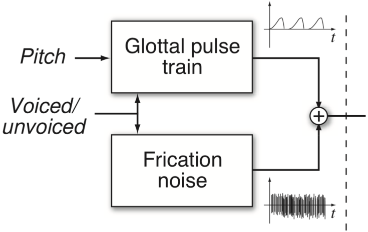

cavité nasale
palais mou
cavité orale
pharynx
cordes vocales
trachée
poumons
force respiratoire

Towards an Excitation/Resonance model



Hypothesis : Excitation and Resonance parts are independant

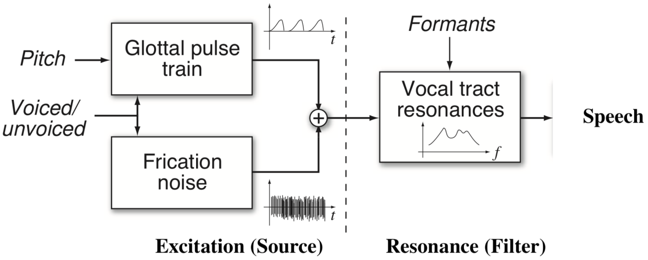
Towards an Excitation/Resonance (Source/filter) model



Excitation (Source)

The Excitation part.....

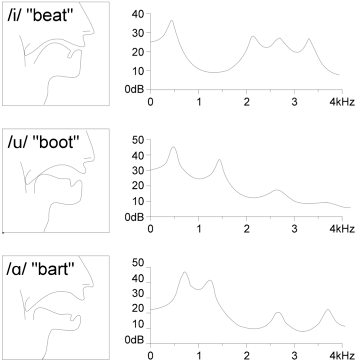
Towards an Excitation/Resonance (Source/filter) model



... followed by the Resonance part

Vowels

- Excitation : Glottal pulse train
- Pitch = train period
- Which vowel = given by resonance



source : Mike Brookes

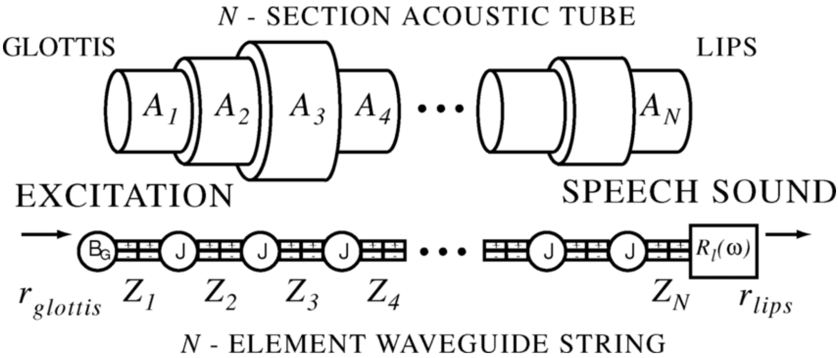
Consonant

- Plosives : brutal opening of the vocal tract
- Fricatives : constriction of the vocal tract
- Nasals
- and many more . . .

Consonant : a little game

	plosives	fricatives
palatal	?/?	?/?
labial	?/?	?/?
dental	?/?	?/?

Voiced/Unvoiced



Each tube $\Longrightarrow \simeq$ AR(2)

The Resonance model

- Sum of formant (resonance)
- Each formant $\simeq$ AR(2)
  - ω_0 : resonance frequency
  - $\Delta\omega$: band width
  - A : gain
- Resonance $\simeq$ AR(2N) filter H_{2N}
  (4-5 formants are enough for vowel recognition)
- Lips radiation : High-pass filter $1 - Z^{-1}$

A. A Glottal pulse train model

- Glottal pulse train of wave form $\simeq Ag(t)$ (with $\int g(t)dt = 1$)
- Amplitude : A
- Period : T
- Support of $g \quad << T$

$$G(t) = \sum_n Ag(t - nT) = Ag \star \sum_n \delta(t - nT)$$

A. A Glottal pulse train model

$$G(t) = Ag \star \sum_n \delta(t - nT) \Longrightarrow \hat{G}(\omega) = \frac{2\pi A}{T}\hat{g}\sum_n \delta(\omega - \frac{2\pi k}{T})$$

- Pitch is $1/T$ (fundamental frequency)
- Glottal train + Resonance (H_{2N} : AR(2N))

$$H_{2N} \star G(t) = AH_{2N} \star g \star \sum_n \delta(t - nT)$$

- Resonance approximation : we take care of g in the resonance part

$$H_{2N} \longrightarrow H_{2N} \star g$$

A. A Glottal pulse train model

A simple final model

$$G(t) = A \sum_n \delta(t - nT)$$

With only two parameters

- A : Amplitude
- T : Period

B. A Frication noise model

$$F(t) = Ah \star W(t)$$

where

- $W(t)$: is a normalized white noise
- A : Amplitude
- $h(t)$: is a low pass filter

B. A Frication noise model

$$F(t) = Ah \star W(t)$$

Frication + resonance :

$$H_{2N} \star F(t) = AH_{2N} \star h \star W(t)$$

$\Longrightarrow$ Resonance approximation : we take care of h in the resonance part

$$H_{2N} \longrightarrow H_{2N} \star h$$

B. A Frication noise model

A simple final model

$$F(t) = AW(t)$$

With a single parameter !

- A : Amplitude

- Excitation : Frication
- Resonance : AR(N) filter

$\Longrightarrow$ **AR Processes**

Definition of AR(N) process $X[n]$

$$X[n] + \sum_{k=1}^{N} a_k X[n-k] = W[n]$$

where

- $W[n]$ is a white noise of variance σ^2
- $\{a_k\}_{k\in[1,N]} \in \mathbb{R}^N$ $(a_0 = 1)$

$$\Longleftrightarrow$$

$$a \star X[n] = W[n]$$

$$a \star X[n] = W[n]$$

A first important question :

- Is there a stationary solution ?

$$a \star X[n] = W[n]$$

A first important question :

- Is there a stationary solution ?

YES : if the inverse filter of a is stable

$\Longleftrightarrow$ All the zeros of $\hat{a}(Z)$ are such that $|Z| < 1$

$$a \star X[n] = W[n]$$

A second important question :

- How do we manage initialization ?

$$a \star X[n] = W[n]$$

A second important question :

- How do we manage initialization ? (stationarity ?)

Theorem if $\{Y[n]\}_n$ is a process such that

$$h \star Y[n] = 0$$

where $h[n]$ is a FIR filter, then

$$\lim_{n \to +\infty} Y[n] = 0 \quad \text{iff} \quad \hat{h}(Z_i) = 0 \Leftrightarrow |Z_i| < 1$$

AR processes estimation ? The Yule-Walker system for AR(N) processes

$$\Longrightarrow$$

The Yule-Walker system for AR(N) processes

The Yule-Walker system for AR(N) processes

$$\begin{pmatrix} R_X[0] & R_X[1] & \dots & R_X[N-1] \\ R_X[1] & R_X[0] & \dots & R_X[N-2] \\ R_X[2] & R_X[1] & \dots & R_X[N-3] \\ \dots & \dots & \dots & \dots \\ R_X[N-2] & R_X[N-3] & \dots & R_X[1] \\ R_X[N-1] & R_X[N-2] & \dots & R_X[0] \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \dots \\ a_{N-1} \\ a_N \end{pmatrix} = \begin{pmatrix} R_X[1] \\ R_X[2] \\ R_X[3] \\ \dots \\ R_X[N-1] \\ R_X[N] \end{pmatrix}$$

Levinson Durbin algorithm $O(N^2)$

$\Longrightarrow \{a_k\}_{k\in[1,N]}$ **estimation**

Variance estimation

$$\sigma^2 = R_X[0] - \sum_{k=1}^{N} a_k R_X[k]$$

If we only have access to a single realization of $X[n]$

- $X[n] \longrightarrow x[n]$
- $R_X[k] \longrightarrow r_x[k] = \frac{1}{P}\sum_{p=0}^{P-1} x[p]x[p+k]$

If we only have access to a single realization of $X[n]$

$$\begin{pmatrix} r_x[0] & r_x[1] & \dots & r_x[N-1] \\ r_x[1] & r_x[0] & \dots & r_x[N-2] \\ \dots & \dots & \dots & \dots \\ r_x[N-2] & r_x[N-3] & \dots & r_x[1] \\ r_x[N-1] & r_x[N-2] & \dots & r_x[0] \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_{N-1} \\ a_N \end{pmatrix} = \begin{pmatrix} r_x[1] \\ r_x[2] \\ \dots \\ r_x[N-1] \\ r_x[N] \end{pmatrix}$$

$$\sigma^2 = r_x[0] - \sum_{k=1}^{N} a_k r_x[k]$$

Linear prediction problem :

$\{x[n]\}_n$ is a signal, what are the optimal coefficients $\{a_k\}_{k\in[1,N]}$ that allows the best prediction of $x[n]$ from $\{x[n-1],\ldots,x[n-N]\}$, i.e.,

$$\tilde{x}[n] = -\sum_{k=1}^{N} a_k x[n-k], \quad \text{with} \quad \sum_n |\tilde{x}[n] - x[n]|^2 \text{ minimum}$$

Solving linear prediction $\iff$ solving Yule-Walker :

$$\begin{pmatrix} r_x[0] & r_x[1] & \dots & r_x[N-1] \\ r_x[1] & r_x[0] & \dots & r_x[N-2] \\ \dots & \dots & \dots & \dots \\ r_x[N-2] & r_x[N-3] & \dots & r_x[1] \\ r_x[N-1] & r_x[N-2] & \dots & r_x[0] \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_{N-1} \\ a_N \end{pmatrix} = \begin{pmatrix} r_x[1] \\ r_x[2] \\ \dots \\ r_x[N-1] \\ r_x[N] \end{pmatrix}$$

$$\sigma^2 = r_x[0] - \sum_{k=1}^{N} a_k r_x[k]$$

- Excitation : Glotal pulse train $A\sum_k \delta[n+kP]$
- Resonance : AR(N) filter

$\Longrightarrow$ **Estimation ?**

The model

$$x[n] = -\sum_{k=1}^{N} a_k x[n-k] + e[n]$$

where

$$e[n] = \sum_{k} \delta[n+kP]$$

Goal : we want to prove that

$$\text{Linear prediction solution} \simeq \{a_k\}_{k\in[1,N]}$$

Conclusion

- Excitation : Glotal pulse train $A \sum_k \delta[n + kP]$
- Resonance : AR(N) filter

If P is "large enough" then solving Yule-Walker leads to the AR(N) coeffiients estimation

Intuition ?

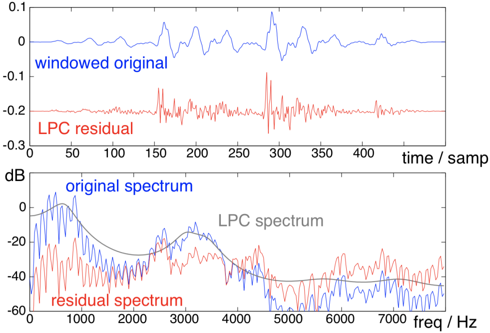
0. Choose N (order of the AR filter)

Then, on sliding windows

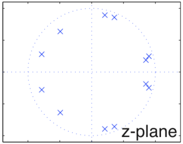
1. Excitation : Noise and/or Pulse train
    - P : period of Pulse train (using e.g., autocovariance function)
2. Solving Yule-Walker

- $\{a_k\}_{k \in [1,N]}$ estimation
- σ^2 : variance of noise
- A : amplitude of pulse train

LPC spectrum on a windowed signal



LPC poles



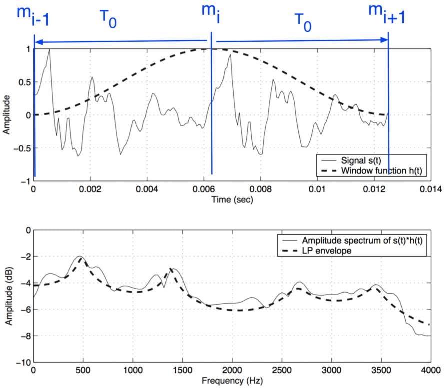
Applications

- Analysis-Synthesis (coding-transmission)
  e.g., low-bit rate coding : LPC-10 (2400 bits/s)
  - $N = 10$ (5 formants)
  - $Fs = 8kHz$
  - window size $K = 180$
- Recognition/classification
- Modification

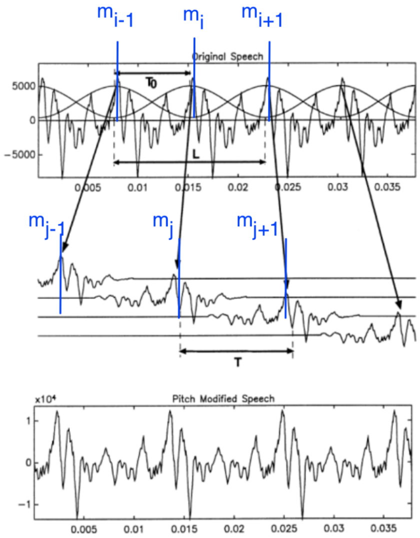
P-SOLA : Pitch-Synchronous Overlapp Add

Analysis :

- Elementary wave form (signal windowing around glottal closure)
- Hypothesis : We get the IR of the LPC filter

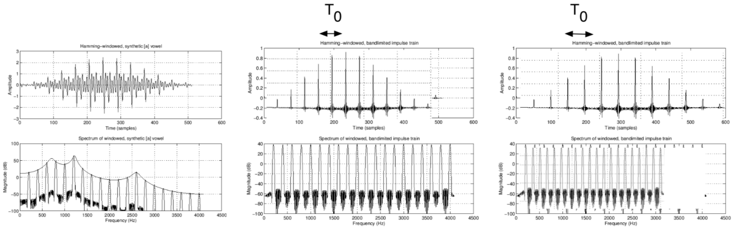


P-SOLA : Pitch-Synchronous Overlapp Add
Pitch modification (decreasing) :



LP-P-SOLA : Linear Predictive Pitch-Synchronous Overlapp Add
= deconvolution using LPC filter + P-SOLA

Pitch modification (decreasing) :



Diphone synthesis

$\longrightarrow$ Problems in phone-concatenation synthesis

- phonems are context-dependent
- coarticulation is complex
- transitions are critical to perception

$\Longrightarrow$ store transitions instead of just phonemes !

Splicing diphones together using PSOLA techniques

Different steps :

- Step 1. : The Fourier transform

$$\hat{s}(\omega) = \int s(t)e^{-i\omega t}dt$$

- Step 2. : Take the (complex) logarithm

$$\log(s(\omega))$$

- Step 3: Take the inverse fourier transform

$$C(\tau) = \int e^{i\omega\tau}\log(\hat{s}(\omega))d\omega$$

Definitions

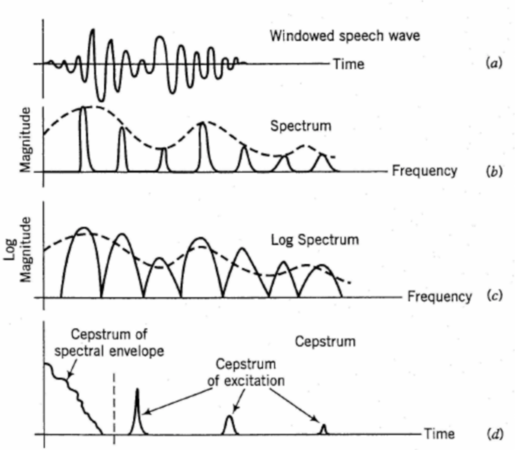
- τ : **quefrency**
- $C(\tau)$: **cepstrum**

What the hell are we using cepstrum for ?

Source/fiter model

- $s(t) = e(t) \star h(t)$
- $\hat{s}(\omega) = \hat{e}(\omega)\hat{h}(\omega)$
- $\log(\hat{s}(\omega)) = \log(\hat{e}(\omega)) + \log(\hat{h}(\omega))$
- $C_s(\tau) = C_e(\tau) + C_h(\tau)$
  - $C_e(\tau)$: most energy concentrated at high quefrency
  - $C_h(\tau)$: most energy concentrated at low quefrency

What the hell are we using cepstrum for ?



Different steps :

- $\hat{s}(\omega) = \int s(t)e^{-i\omega t}dt = A(\omega)e^{i\phi(\omega)}$
- $\log(\hat{s}(\omega)) = \log(A(\omega)) + i\phi(\omega)$
- $\Re(\log(\hat{s}(\omega))) = \log(A(\omega))$
- $C(\tau) = \int e^{i\omega\tau}\Re(\log(\hat{s}(\omega)))d\omega$

Towards audio descriptors : Desribe the timber of an audio signal with few coefficients

MFCC : Mel Frequency Cepstral Coefficients

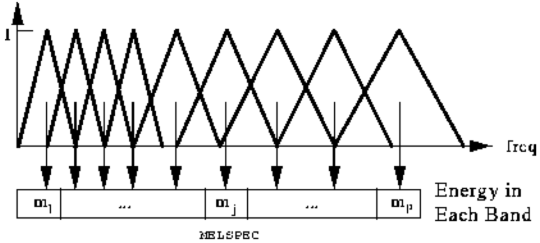
Definition : Real cepstrum computed computed on an energy spectrum after being converted in a perceptive scale

Why ? The ear has better resolution at low frequeny than high frequency

Which perceptive scales ? Mel (Bark, ERB filters, Gamma tone)

What is it used for ? Thay are the most used descriptors for audio signals

The Mel scale : This is a set of (generally) 40 triangular filters applied to the periodogram power spectral estimate



MFCC step by step

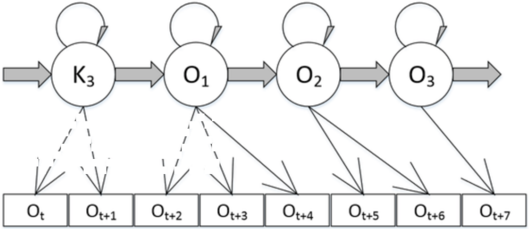
- Compute Fourier spectrum intensity $|\hat{s}(\omega)]^2$
- Compute The Mel filters
  - Number of filters (40)
  - Shape fo filters (triangle, ...)
- Compute the spectrum intensity in the Mel scale
  $S(b) = \sum_\omega |\hat{s}(\omega)]^2 H_b(\omega)$
- Take the log $\log S(b)$
- Inverse Fourier transform (or IDCT) $\log S(b)$
- Select coefficients close to 0 (generally 10-15)

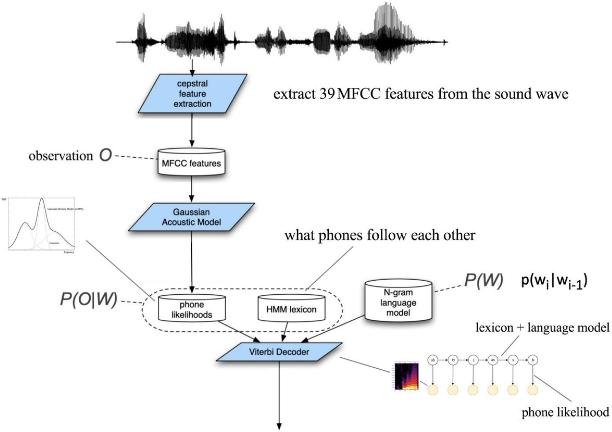
GMM : Phoneme model (using MFCC)

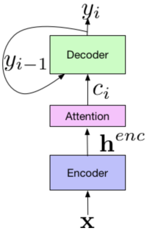
$\longrightarrow$ generally each phoneme is ut in 3 parts

- begining (O_1)
- middle (O_2)
- end (O_3)

HMM : Chaining of the different parts




cepstral feature extraction
extract 39 MFCC features from the sound wave
observation O
MFCC features
Gaussian Acoustic Model
what phones follow each other
P(O|W)
phone likelihoods
HMM lexicon
N-gram language model
P(W)
p(w_i|w_{i-1})
lexicon + language model
Viterbi Decoder
phone likelihood



Components of the LAS End-to-End Model.

State of the art speech reognition with sequence-to-sequence models, Google, 2017

- 80 Mel-filters (25ms window, shifted 10ms)
- Encoder : 5-layer LSTM with 1400 hidden units
- Attention with 4 heads
- Decoder 2-layer LSTM with 1024 units

Trained on 12.500 hours of Google search voice recording : 5.6% word error (HMM-LSTM : 6.7%)