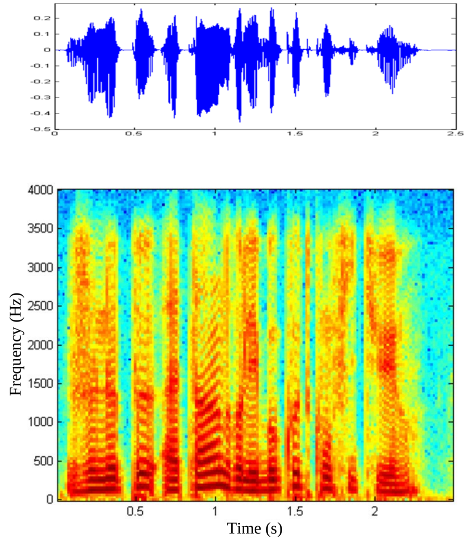
Spectral and temporal modifications

Roland Badeau,
roland.badeau@telecom-paris.fr

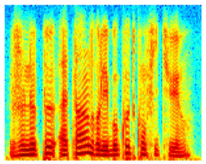
TSIA 206 - Speech and audio processing

Introduction

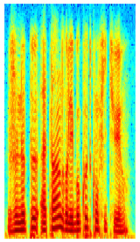


Original waveform and spectrogram

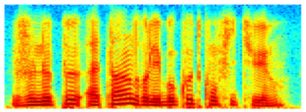
Modification of playback speed



Original sound



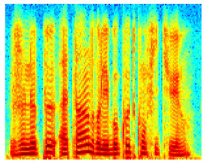
Increased speed



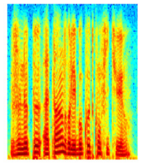
Lowered speed

Modifying playback speed impacts both time and frequency scales
Origin of the problem: $y(t) = x(\alpha t) \Leftrightarrow Y(f) = \frac{1}{|\alpha|} X(\frac{f}{\alpha})$

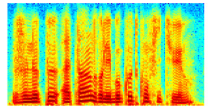
Modifications of duration and pitch



Original sound



Shorter time scale



Lower frequency scale

Goal: separately control the time and frequency scales

Outline

- Separate control of the time and frequency scales
  - Synthesis by means of wavetable sampling
  - Post-synchronization of sound and video
  - Musical post-production
- Three categories of methods:
  - Spectral methods: phase vocoder
  - Temporal methods: TD-PSOLA
  - Parametric methods: LPC, sinusoids plus noise model

Part I

Definitions

Vocal production model

- Time-varying, linear source / filter model:
  $x(t) = \int_{-\infty}^{+\infty} g(t,\tau)\, e(t-\tau)\, \mathrm{d}\tau$
- Frequency response of the filter:
  $G(t,f) = \int_{-\infty}^{+\infty} g(t,\tau)\, e^{-j2\pi f\tau} \mathrm{d}\tau = M(t,f)\, e^{j\varphi(t,f)}$
- Harmonic source: $e(t) = \sum_{k=1}^{L} e^{j\xi_k(t)}$, where $\frac{\mathrm{d}\xi_k}{\mathrm{d}t} = 2\pi f_k(t)$
- Quasi-stationarity assumption: $\xi_k(t-\tau) \simeq \xi_k(t) - 2\pi f_k(t)\tau$
- Filtered signal: $x(t) = \sum_{k=1}^{L} M(t, f_k(t))\, e^{j(\xi_k(t)+\varphi(t,f_k(t)))}$

Signal models

McAulay and Quatieri model (speech coding)
$x(t) = \sum_{k=1}^{L} A_k(t)\, e^{j\Psi_k(t)}$ where $\frac{\mathrm{d}\Psi_k}{\mathrm{d}t} = 2\pi f_k(t)$
and $A_k(t)$ and $f_k(t)$ have slow variations compared with $e^{j\Psi_k(t)}$

Serra and Smith model (music signal synthesis)
$x(t) = \sum_{k=1}^{L} A_k(t)\, e^{j\Psi_k(t)} + b(t)$
where $b(t)$ is a white noise filtered by a time-varying filter

Complete analysis / modification / synthesis system:

- estimation of the deterministic components
- linear interpolation of amplitudes and cubic interpolation of phases
- subtraction of the deterministic part to get $b(t)$
- transformation of each of the two components
- re-synthesis

Scales modifications

Duration modification

- Temporal distortion function: $\tau = T(t)$
- Modified signal: $y(\tau) = \sum_{k=1}^{L} A_k(T^{-1}(\tau))\, e^{j\phi_k(\tau)}$
- Preservation of the frequencies: $\phi_k(\tau) = 2\pi \int_0^\tau f_k(T^{-1}(u))du$

Pitch modification

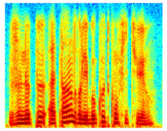
- Spectral compression rate: $\alpha(t)$
- Modified signal: $y(t) = \sum_{k=1}^{L} A_k(t)\, e^{j\Phi_k(t)}$
- Frequencies modification: $\Phi_k(t) = 2\pi \int_0^t \alpha(u) f_k(u) du$

Reciprocity

- temporal distortion T plus temporal re-scaling T^{-1}
  $\Leftrightarrow$ pitch modification of rate $\alpha(t) = T'(t)$

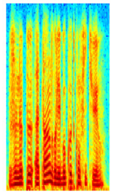
Equivalence of the two modifications

- Duration modification (shorter time scale)



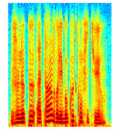
Original sound

$\longrightarrow$



Increased speed

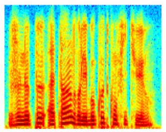
$\longrightarrow$



Lower frequency scale

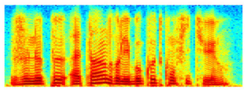
Equivalence of the two modifications

- Pitch modification (lower frequency scale)



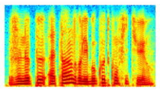
Original sound

$\rightarrow$



Lowered speed

$\rightarrow$

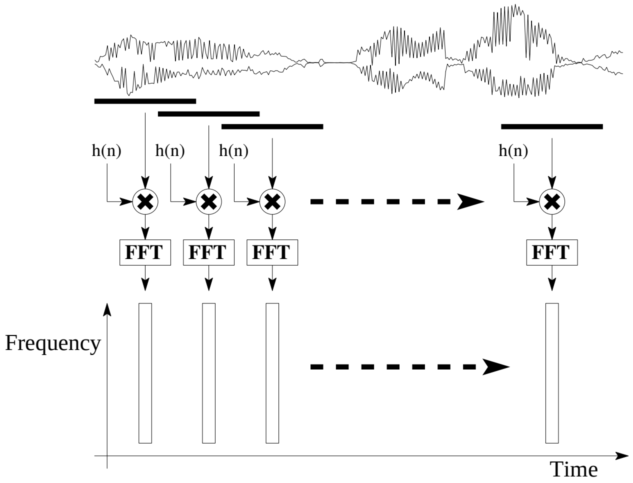


Shorter time scale

Part II

Short time Fourier transform

Principle diagram



Short time Fourier transform

Definition: $\widetilde{X}(t_a,\nu) = \sum_{n\in\mathbb{Z}} x(n+t_a)\, w_a(n)\, e^{-j2\pi\nu n}$, where

- the analysis window $w_a(n)$ is finite, real and symmetric
- the analysis times t_a are indexed by an integer u

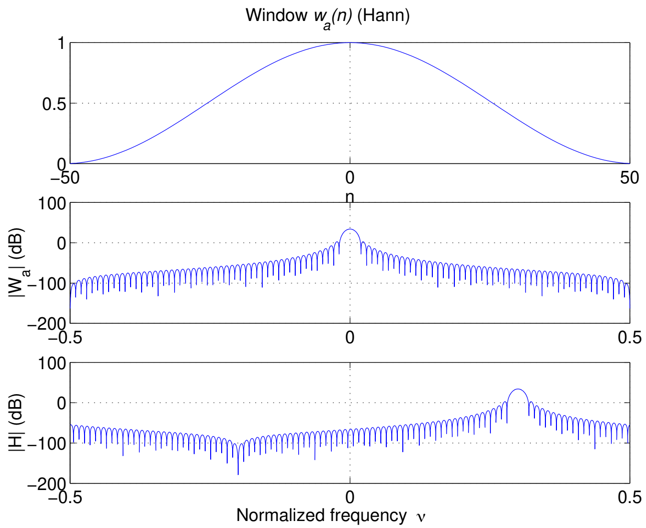
Interpretation: band-pass convention

- $\widetilde{X}(t_a,\nu_p) = [x \star h](t_a)$ where $h(n) = w_a(-n)\, e^{j2\pi\nu_p n}$
- the FT $h(n)$ is $H(e^{j2\pi\nu}) = W_a\left(e^{j2\pi(\nu_p-\nu)}\right)$

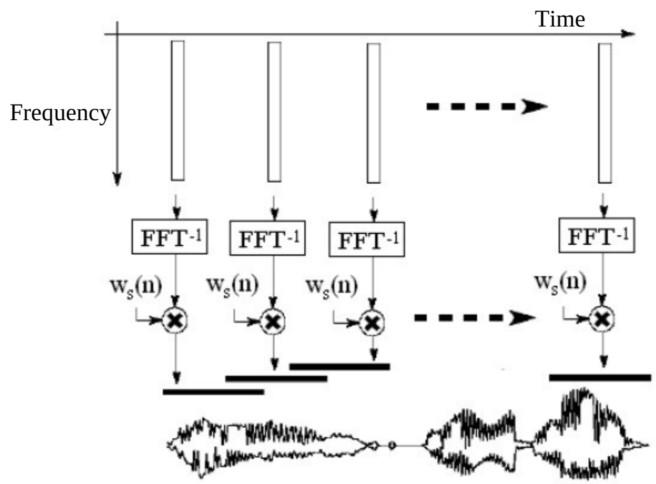
Discrete version of the STFT: let $\nu_p = \frac{p}{N}$

- $\widetilde{X}(t_a,\nu_p) = \sum_{n=0}^{N-1} x(n+t_a)\, w_a(n)\, e^{-j2\pi\frac{pn}{N}}$
- the length of the analysis window must be $\leq N$

Equivalent band-pass filter



Synthesis diagram



Signal reconstruction

Perfect reconstruction condition ($t_s = t_a$ and $Y = \widetilde{X}$)

- Overlap-add (OLA) synthesis

$$y(n) = \sum_u w_s(n - t_s(u))\, y_w(n - t_s(u), t_s(u))$$

$\text{supp}(w_s) \subset [0, N-1]$,
$y_w(n, t_s(u)) = \frac{1}{N}\sum_{p=0}^{N-1} Y(t_s(u), \nu_p)\, e^{j2\pi\nu_p n}$

- sufficient condition: $\sum_u w_a(n - t_a(u))\, w_s(n - t_a(u)) \equiv 1$

Modifications and problems raised:

- Modification of the amplitudes and phases of the STFT
- $t_a \longrightarrow t_s$, $\widetilde{X}(t_a(u), \nu_p) \longrightarrow Y(t_s(u), \nu_p)$
- Difficulty: Y is generally not the STFT of a signal
- Re-synthesis from a sinusoidal model

Part III

Phase vocoder

Instantaneous frequency

- McAulay and Quatieri model: $x(t) = \sum_{k=1}^{L} A_k(t)\, e^{j\Psi_k(t)}$
- **Quasi-stationarity assumption**: $\forall n \in \{0 \dots N-1\}$

$$\begin{cases} A_k(n + t_a) & \simeq & A_k(t_a) \\ \Psi_k(n + t_a) & \simeq & \Psi_k(t_a) + 2\pi f_k(t_a) n \end{cases}$$

- Then $\widetilde{X}(t_a(u), \nu_p) = \sum_{k=1}^{L} A_k(t_a)\, e^{j\Psi_k(t_a)}\, W_a\left(e^{j2\pi(\nu_p - f_k(t_a))}\right)$
- Let f_c be the cutting frequency of the low-pass filter $w_a(n)$
- **Narrow band condition**: $\exists!$ l such that $|\nu_p - f_l(t_a)| \leq f_c$
Interpretation (harmonic spectrum): $N \geq \frac{4}{f_0}$
- Then $\widetilde{X}(t_a(u), \nu_p) = A_l(t_a)\, e^{j\Psi_l(t_a)}\, W_a\left(e^{j2\pi(\nu_p - f_l(t_a))}\right)$
$\Rightarrow$ the STFT permits us to estimate phases $\Psi_l(t_a)$ modulo 2π

Overlap condition

Removing the phase ambiguity modulo 2π:

- Phase difference between two successive times:
$\Delta\Phi_p = 2\pi(f_l(t_a) - \nu_p)\Delta t_a(u) + 2\pi\nu_p\Delta t_a(u) + 2n\pi$
- Minimal overlap condition: $f_c\,\Delta t_a(u) < \frac{1}{2}$
Interpretation (Hann window): $f_c = \frac{2}{N} \Rightarrow \Delta t_a < \frac{N}{4}$
- $\exists!$ n such that $|\Delta\Phi_p - 2\pi\nu_p\Delta t_a(u) - 2n\pi| < \pi$

Estimation of the instantaneous frequency $\forall p \in \{0 \dots N-1\}$

1. computation of the STFT at two successive times $\longrightarrow \Delta\Phi_p$
2. computation of $Q(n_0) = \Delta\Phi_p - 2\pi\nu_p\Delta t_a - 2n_0\pi$ such that $|Q(n_0)| < \pi$
3. computation of instantaneous frequency $f_l(t_a) = \nu_p + \frac{Q(n_0)}{2\pi\Delta t_a}$

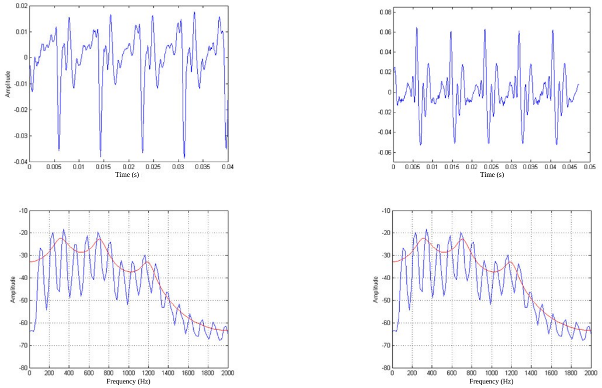
Duration modification

Unwrapping of the instantaneous phases for a distortion $T(t)$

Modification algorithm:

1. computation of the STFT and of $f_l(t_a(u))$ in each channel
2. computation of the new synthesis time $t_s(u) = T(t_a(u))$
3. computation of the synthesis instantaneous phase
   $\Phi_s(t_s(u+1), \nu_p) =$
   $\Phi_s(t_s(u), \nu_p) + 2\pi f_l(t_a(u))\,(t_s(u+1) - t_s(u))$
4. computation of the synthesis STFT at $u+1$
   $\widetilde{Y}(t_s(u+1), \nu_p) = A_p(t_a(u+1))\, e^{j\Phi_s(t_s(u+1),\nu_p)}$

Influence of the initial phases



Original sound

Synthesis with random phases

Modifying the initial phases changes the waveform, but neither the spectrum nor perception

Pitch modification

Temporal re-sampling method

1. time stretching of rate $T(t) = \int_0^t \alpha(u)du$
2. temporal re-scaling of rate $T^{-1}(\tau)$

Spectral re-sampling method

1. Linear interpolation of the analysis STFT
   - $\alpha(t_a) > 1$: information loss in high frequencies
   - $\alpha(t_a) < 1$: spectral completion in high frequencies
2. re-synchronization of the phases in the re-synthesis

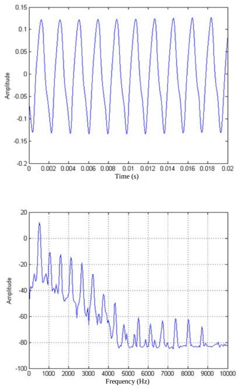
Problem in speech processing: "Donald Duck" effect

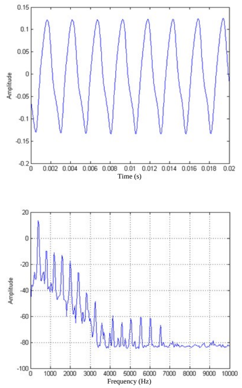
- spectral envelope estimation (LPC) and "whitening"
- pitch modification, then inverse filtering

Part IV

Processing specific to speech

Time-frequency reciprocity

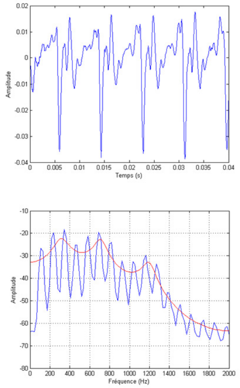


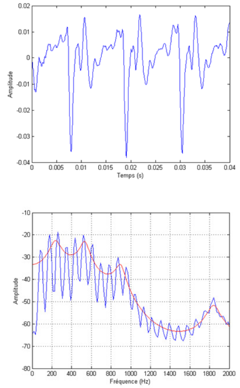


Original sound — Lower frequency scale

A piano sound still sounds natural after changing the frequency scale

Time-frequency reciprocity

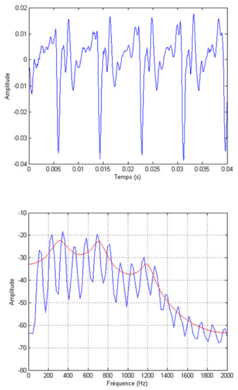


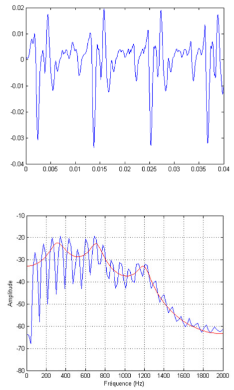


Original sound — Lower frequency scale

Voiced speech sound seems unnatural after changing frequency scale
Explanation: spectral envelope is distorted with the harmonics

Pitch modification of speech

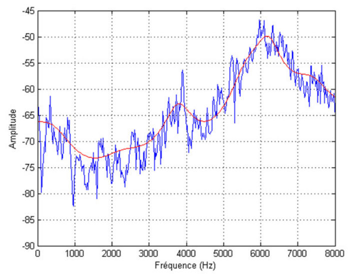


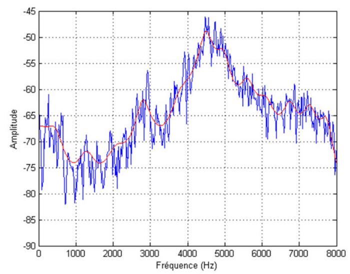


Original sound — Pitch shifting

Natural pitch shifting of speech keeps spectral envelope unchanged

Case of unvoiced sounds

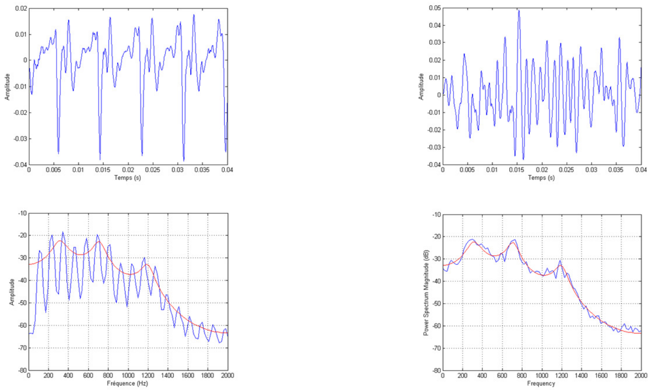




Original sound — Lower frequency scale

The spectral envelope of unvoiced sounds should not be changed

Timbre and spectral envelope



Original voiced sound — Synthetic noise with same envelope

The spectral envelope characterizes the timbre of speech sounds

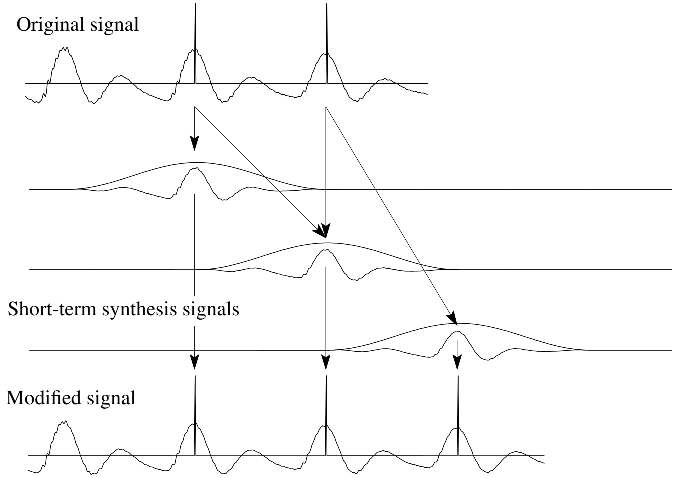
Pitch modification

- Voiced sounds:
  - modify the fundamental frequency
- Voiced/unvoiced sounds:
  - leave the spectral envelope unchanged
- Use of the vocoder
  1. Signal whitening by filtering (LPC analysis)
  2. Frequency scale modification
  3. Inverse filtering
- Methods specific to monophonic speech signals
  - Voiced/unvoiced segmentation
  - Pitch estimation on the voiced frames

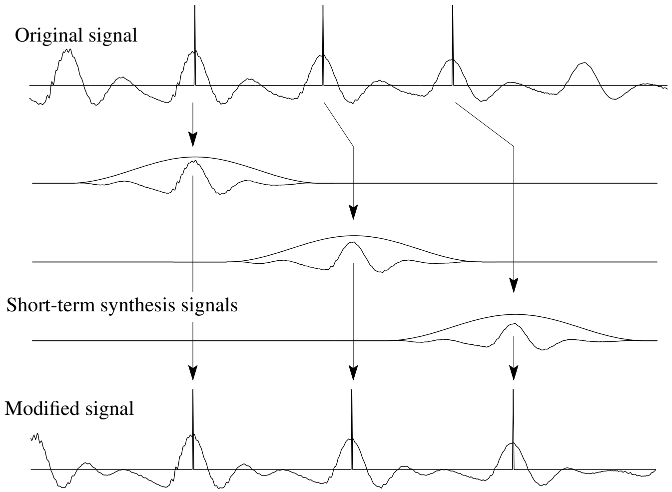
Part V

TD-PSOLA

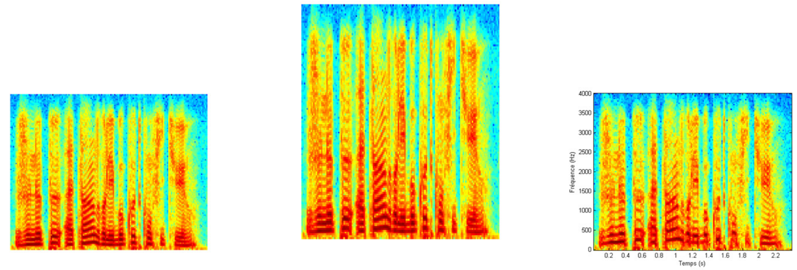
Temporal modifications



Spectral modifications



Example of pitch modification



Original sound | Phase vocoder | PSOLA

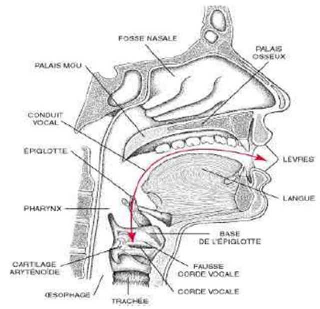
Contrary to the phase vocoder, PSOLA performs pitch shifting without modifying the spectral envelope

Part VI

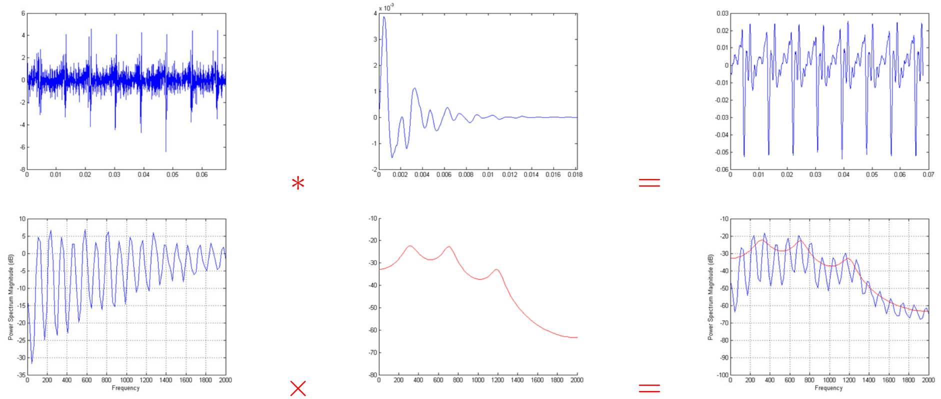
Auto-regressive models

Speech production mechanism

- Voiced sounds: vibration of the vocal cords filtered by the vocal tract
- Unvoiced sounds: turbulent noise filtered by the vocal tract

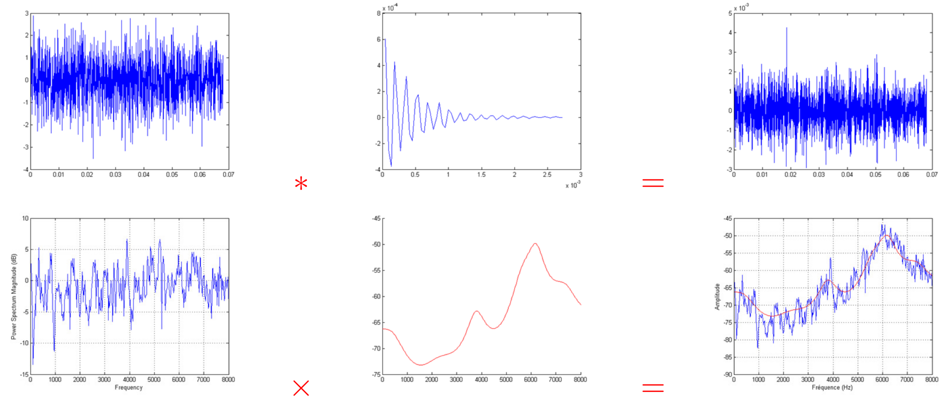


Production of voiced sounds



Glottal pulses   Vocal tract   Voiced sound

Production of unvoiced sounds



Turbulent noise   Vocal tract   Unvoiced sound

Signal model

- The vocal tract is modeled by an AR filter

$$h(z) = \frac{1}{1 + a_1 z^{-1} + \ldots + a_p z^{-p}}$$

  estimated by linear prediction (LPC analysis)
- Source model depending on the voiced / unvoiced case
  - The glottal pulse train is modeled by an impulse train of period T

$$s(t) = \sum_n \delta(t - nT)$$

  - The turbulent noise is modeled by a white noise

Synthesis with auto-regressive models

- Synthesis without modification
  - by overlap/add of the time frames
  - convolution of the source with the filter on every frame
- Synthesis with modification
  - Duration modification
    - Synthesis of a source of appropriate length
  - Pitch modification
    - Unvoiced frames: unchanged
    - Voiced frames: the period of the impulse train is changed