PART II
(short-term) Modeling of the speech signal

Definitions

- Speech processing is based on **speech models**
- Models have **parameters**, like black boxes have switches and sliders
- Parameters are **estimated** via **algorithms**
- Errors: output of model ≠ input signal
  - **Modeling** (intrinsic) errors
  - **Estimation** (extrinsic) errors
- Algorithms **minimize** errors

Families of speech models

- **Articulatory** models (parameters = position of tongue, glottis, lip opening, etc.)
- **Production** models (electrical analogy of vocal tract: combination of electrical generators and filters; parameters = switches and coef. of filters)
- **Phenomenological** models (pure signal processing techniques for modeling the speech signal or its spectrum: FFT, wavelets, time-domain processing, etc.)

Contents

- Production models
  - **Autoregressive modeling of speech**
  - Autoregressive modeling of stationary random signals
    - AR model - Estimation algorithms
    - Autoregressive estimation of speech
    - Extensions of the AR model
  - Transform-based models **(phenomenological)**
    - Short-term Fourier transform
      - Definition - Interpretation
      - Relationship with AR modeling
      - Hybrid harmonic+noise model - Estimation
    - Homomorphic (or cepstral) transform
      - Definition - Properties
      - Relationship with AR modeling
  - F0 estimation
Autoregressive modeling of speech (Fant, 50’s)

\[ G(z) = \frac{1}{(1-\alpha z^{-1})(1-\beta z^{-1})} \]

Glottal volume velocity waveform

\[ V(z) = \frac{B}{\prod_{k=1}^{K}(1 + b_1 z^{-1} + b_2 z^{-2})} \]

Volume velocity waveform

\[ R(z) = c(1 - z^{-1}) \]

Pressure waveform

Contents

- Production models
  - Autoregressive modeling of speech
  - Autoregressive modeling of stationary random signals
    - AR model - Estimation algorithms
    - Autoregressive estimation of speech
    - Extensions of the AR model
- Transform-based models (phenomenological)
  - Short-term Fourier transform
    - Definition - Interpretation
    - Relationship with AR modeling
    - Hybrid harmonic+noise model - Estimation
  - Homomorphic (or cepstral) transform
    - Definition - Properties
    - Relationship with AR modeling
- F0 estimation

Autoregressive modeling of stationary random signals

\[ G(z)V(z)R(z) \approx \sigma/A_p(z) : « \text{All pole} » \text{ model} \]

\[ x(n) \begin{cases} \mathcal{N}(0, \sigma^2) \\
\mathcal{N}(0, \sigma^2) \end{cases} \]

\[ e(n) \begin{cases} \mathcal{N}(0, \sigma^2) \\
\mathcal{N}(0, \sigma^2) \end{cases} \]

\[ \frac{B(z)}{A(z)} \]

\[ \frac{1}{A(z)} \]

Much simpler for estimation : AR model

\[ x_{AR}(n) \]

\[ x_{ARMA}(n) \]
Autoregressive modeling of stationary random signals

- Each sample \( x(n) \) is the sum of a completely (linearly) predictable component, and an « innovation » component \( e(n) \)

\[ x(n) = e(n) + \sum_{i=1}^{p} a_i x(n-i) \]

\( \Rightarrow \) Autoregressive or linear prediction (LP) model

\[ e(n) \text{ is the output of the inverse filter} \]

\[ YULE-WALKER \text{ equations} \]

\[
\sum_{j=1}^{p} \phi_j (i-j) a_j = -\phi_i (i \in \{1, \ldots, p\})
\]

\[
\begin{bmatrix}
\phi_0 (0) & \phi_1 (1) & \ldots & \phi_{p-1} (p-1) & a_1 \\
\phi_1 (1) & \phi_0 (0) & \ldots & \phi_{p-2} (p-2) & a_2 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\phi_{p-1} (p-1) & \phi_{p-2} (p-2) & \ldots & \phi_0 (0) & a_p \\
\end{bmatrix}
= -
\begin{bmatrix}
\phi_1 (1) \\
\phi_2 (2) \\
\vdots \\
\phi_p (p) \\
\end{bmatrix}
\]

\[ \Phi_p^T a = -\varphi_p^2 \]

- \( p \) linear equations with \( p \) unknowns \( \Rightarrow O(p^3) \)?
- Toeplitz matrix \( \Rightarrow O(p^2) \) : recursively on the prediction order

**LEVINSON** SCHUR(-LEROUX-GUEGEN)

**YULE-WALKER equations**

\[ \sigma_e^2 = E[x^2(n)] = E \left[ \sum_{j=0}^{p} a_j x(n-i) \sum_{j=0}^{p} a_j x(n-j) \right] \]

\[ = \sum_{i=0}^{p} a_i a_i E[x(n-i)x(n-j)] \]

\[ = \sum_{i=0}^{p} a_i a_i \phi_i (i-j) \]

\[ \frac{\partial \sigma_e^2}{\partial a_i} = 0 \quad (i = 1, \ldots, p) \]

\[ 2\sum_{j=0}^{p} \phi_i (i-j) a_j = 0 \]

\[ \sum_{j=1}^{p} \phi_i (i-j) a_j = -\phi_i \quad (i = 1, \ldots, p) \]

- **LEVINSON algorithm**

\[ \begin{bmatrix}
\phi_0 (0) & \phi_1 (1) & \ldots & \phi_{m-1} (m-1) & a_1^{m-1} \\
\phi_1 (1) & \phi_0 (0) & \ldots & \phi_{m-2} (m-2) & a_2^{m-2} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\phi_{m-1} (m-1) & \phi_{m-2} (m-2) & \ldots & \phi_0 (0) & a_p^{m} \\
\end{bmatrix}
= -
\begin{bmatrix}
\phi_1 (1) \\
\phi_2 (2) \\
\vdots \\
\phi_m (m) \\
\end{bmatrix}
\]

Find solution for order \( m+1 \) from sol. for \( m \)?
**LEVINSON algorithm**

- **Initialization**
  
  \[ a_m(0) = 1, \quad m = 1, 2, \ldots, p \quad a_0 = \phi_s(0) = \sigma_s^2 \]

- **Recursion on** \( m=0 \ldots p-1 \)

  \[ k_{m+1} = -(1/a_m) \sum_{i=0}^{m-1} a_m(i) \cdot \phi_s(m-i) \]

  \[ a_m(i) = a_{m-1}(i) + k_m \cdot a_{m-1}(m-i) \]

  \[ a_{m+1}(m) = k_{m+1} \]

  \[ a_{m+1} = a_m(1-k_{m+1}) \]
\( k_i \): PARCOR coefficients

- Naturally appear in the LEVINSON recursion
- Can be physically interpreted as area ratios of acoustic tubes in series
- Equivalent to prediction coefficients (there is a recursion formula for \( \{a_i\} \rightarrow \{k_i\} \rightarrow \{a_i\} \))
- Transfer function of \( 1/A(z) \) much less sensitive to \( \Delta k_i \) than to \( \Delta a_i \)
- Good interpolation properties
- Correspond to the lattice filter structure for \( 1/A(z) \)
- \( 1/A(z) \) stable if \(-1 < k_i < +1\)

How good are the \( a_i \)'s or \( k_i \)'s?

\[
\phi_x(k) = \phi_{AR}(k)
\]
for \( k = 0 \ldots p \)
Contents

- Production models
  - Autoregressive modeling of speech
  - Autoregressive modeling of stationary random signals
    - AR model - Estimation algorithms
- Autoregressive estimation of speech
  - Extensions of the AR model
- Transform-based models (phenomenological)
  - Short-term Fourier transform
    - Definition - Interpretation
    - Relationship with AR modeling
  - Hybrid harmonic+noise model - Estimation
    - Homomorphic (or cepstral) transform
      - Definition - Properties
      - Relationship with AR modeling
- F0 estimation

Short-term analysis of speech

- Speech is not stationary (if it were, information \( \equiv 0 \))
- "Pseudo-stationary" on 30 ms
- Frame-by-frame (typ.: \( N=30\text{ms}/L=10\text{ms} \))

- ex: \( F_s=10\text{kHz} \), \( L=100 \) samples, \( N=300 \) samples

Yule-Walker equations (again)

\[
E_p = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{i=0}^{p} a_i \sum_{n=-\infty}^{\infty} x(n-i) \sum_{j=0}^{\infty} a_j x(n-j) = \sum_{i=0}^{p} a_i \sum_{n=-\infty}^{\infty} [x(n-i)x(n-j)] = \sum_{i=0}^{p} a_i a_j r_s(i-j)
\]

- Minimize the energy \( E_p \) of the prediction error \( e(n) \)
- Same formalism as in the stationary case, applied here to a signal which takes non-zero values only for \( n=0..N-1 \)
- Same Yule-Walker equations, except \( \phi \) is estimated as \( r_s \)

\[
2 \sum_{i,j=0}^{p} r_s(i-j) a_i = 0 \quad (i=1...p)
\]

\[
\sum_{i,j=0}^{p} r_s(i-j) a_i = -r_s(0) \quad (i=1...p)
\]

- Toeplitz matrix \( \Rightarrow O(p^2) \) using Levinson or Schur (still slightly faster than Levinson)
- NB : this is actually called the « autocorrelation approach »; in contrast, the so-called « covariance approach » is based on a different expression of the energy of \( e(n) \) and leads to a non-Toeplitz matrix...
In practice

• Sampling frequency
  - telephone speech: 8 kHz
  - speech recognition: 10 kHz
  - speech synthesis: 16 kHz
  - multimedia applications 11.25 kHz, 22.5 kHz, et 44.1 kHz

• Prediction order \( p \)
  - 2 poles for shaping the glottal velocity volume waveform
  - 2 poles (1 resonator) per formant (≈ per kHz of bandpass)
  \[ p_{opt} = 2 + F_s \]

In practice

• (Pre-accentuation: filter \( P(z) = 1 - \mu z^{-1} \))
• (Hamming) Weighting window
  - prevent \( e(n) \) from naturally taking high values at the beginning and at the end of the frame
  - let each set of prediction coefficients be more representative of the 10 central ms
  \[ w(n) = 0.54 + 0.46\cos\left(2\pi \frac{n}{N}\right) \]

• Computational load

<table>
<thead>
<tr>
<th></th>
<th>( N=300, p=10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighting</td>
<td>( N )</td>
</tr>
<tr>
<td>Autocorrelation</td>
<td>( (N - p / 2) (p + 1) )</td>
</tr>
<tr>
<td>Schur</td>
<td>( p (p + 1) )</td>
</tr>
<tr>
<td></td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>3245</td>
</tr>
<tr>
<td></td>
<td>110</td>
</tr>
</tbody>
</table>
Contents

• Production models
  – Autoregressive modeling of speech
  – Autoregressive modeling of stationary random signals
    • AR model - Estimation algorithms
  – Autoregressive estimation of speech
  – Extensions of the AR model
• Transform-based models (phenomenological)
  – Short-term Fourier transform
    • Definition - Interpretation
    • Relationship with AR modeling
    • Hybrid harmonic+noise model - Estimation
  – Homomorphic (or cepstral) transform
    • Definition - Properties
    • Relationship with AR modeling
• F0 estimation

Problem with pseudo-stationarity

• \( r_x \) means nothing for plosives \( \Rightarrow \) meaning of \( \{a_i\} \)?
• Explosions of plosives last 1-2 ms; they are synthesized on 10 ms at best
• In practice, the ear is much less sensitive to the spectrum of highly transient sounds

LP Speech synthesis

• Parameters are changed every 10 ms

\[
V/U \rightarrow s \text{ coefficients} \rightarrow \frac{1}{A_i(z)}
\]

• \( e(n) \) is replaced by a pulse sequence or by white noise \( (\mu=0, \sigma=1) \), and amplified with \( \sigma=\sigma_e \)

Problem with anti-formants

ARMA would be better for nasal sounds
Problem with mixed voicing

Voiced fricatives: partly voiced, partly unvoiced

- Decreasing modeling errors
  - If the prediction error $e(n)$ was used as input for $1/A(z)$: $e_r(n) = 0$
  - In practice, the excitation waveforms used are a very rough approximation of $e(n)$
  - Find more realistic excitation waveforms?

Multipulse Linear Prediction

MP-LPC = LP model excited with small number of pulses, whose positions and amplitudes have to be adjusted

Code-Excited Linear Prediction

CELP = LP model excited with a real excitation signal, taken from a list (=codebook) of typ. 1024 or 2018

Code-Excited Linear Prediction

- MP-LPC
- CELP
Contents

- Production models
  - Autoregressive modeling of speech
  - Autoregressive modeling of stationary random signals
    - AR model - Estimation algorithms
    - Autoregressive estimation of speech
    - Extensions of the AR model
- Transform-based models (phenomenological)
  - Short-term Fourier transform
    - Definition - Interpretation
    - Relationship with AR modeling
    - Hybrid harmonic+noise model - Estimation
  - Homomorphic (or cepstral) transform
    - Definition - Properties
    - Relationship with AR modeling
- F0 estimation

Short-term Fourier Transform

- Continuous: \( x(t) \leftrightarrow X(t, \omega) = \int_{-\infty}^{\infty} x(\tau)w(t-\tau)e^{-j\omega \tau} \, d\tau \)

- Discrete: \( x(n) \leftrightarrow X(n, k) = \sum_{i=0}^{N-1} x(i)w(n-i)e^{-jk\omega n} \)

- Short-term spectral density function:
  \( S_s(n, \phi) = |X(n, k)|^2 \) with \( \phi = k \frac{2\pi}{N} \)

- Weighting window \( w(n) \)
  - if too long, averaging occurs (+stationnarity?)
  - if too short, frequency resolution falls down
  - shape?
Narrow-band: voiced

Narrow-band: unvoiced

Contents

- Production models
  - Autoregressive modeling of speech
  - Autoregressive modeling of stationary random signals
    - AR model - Estimation algorithms
  - Autoregressive estimation of speech
  - Extensions of the AR model
- Transform-based models (phenomenological)
  - Short-term Fourier transform
    - Definition - Interpretation
  - Relationship with AR modeling
  - Hybrid harmonic+noise model - Estimation
  - Homomorphic (or cepstral) transform
    - Definition - Properties
    - Relationship with AR modeling
- F0 estimation

STFT ↔ AR model?

\[
\begin{align*}
\phi_s(k) & \leftrightarrow S_s(\phi) = \sum_k \phi_s(k)e^{j\phi} \\
\phi_{uAR}(k) & \leftrightarrow S_{uAR}(\phi) \\
S_{uAR}(\phi) & = S_u(\phi) \left| \frac{\sigma}{A(e^{j\phi})} \right|^2 = \left| \frac{\sigma}{A(e^{j\phi})} \right|^2 \\
\phi_{uAR}(k) & = \phi_s(k) \quad \text{for } k = 0,1,\ldots, p \\
\end{align*}
\]

by definition

\[
\left| \frac{\sigma}{A(e^{j\phi})} \right|^2 \quad \text{approximates the envelope of } S_s(\phi)
\]

NB: this is also applicable to short term estimations
**Contents**

- Production models
  - Autoregressive modeling of speech
  - Autoregressive modeling of stationary random signals
    - AR model - Estimation algorithms
  - Autoregressive estimation of speech
  - Extensions of the AR model
- Transform-based models *(phenomenological)*
  - **Short-term Fourier transform**
    - Definition - Interpretation
    - Relationship with AR modeling
    - **Hybrid harmonic+noise model - Estimation**
  - Homomorphic *(or cepstral)* transform
    - Definition - Properties
    - Relationship with AR modeling
- F0 estimation

**Hybrid harmonic/noise model**

- Harmonic component defined by the amplitudes and phases of harmonics: \(\{\text{amp}_i\}, \{\text{phase}_i\}\)
- Noise component defined by the s.d.f. of \(x_n(t)\)

**Hybrid H/N estimation**

- Rough F0 estimation *(comb-filter-like)*
- Estimate the parameters of the harmonic part

\[
e_p(t) = \int_{-\infty}^{\infty} w^2(t-\tau) |x(\tau) - x_h(\tau)|^2 d\tau = \frac{1}{2\pi} \int_{-\infty}^{\infty} [S_x(t,\omega) - S_{x_h}(t,\omega)]^2 d\omega
\]

with \(x_h(t) = \sum_{i=-N}^{N} \text{amp}_i e^{j\text{phase}_i} x(t)\) and \(a_{-i} = a^*_i\)

- Typical least squares minimization problem:

\[
Aa = b
\]

- Solution:
  \[
  A^H A a = A^H b = Ra = r
  \]
  - trivial if \(R\) is diagonal
  - Levinson if \(R\) is Toeplitz
  - Cholesky otherwise (\(R\) is always symmetric)
**Hybrid H/N estimation**

- Repeat on a grid of $F_0$ values around initial $F_0$
  - Precision on $N$th harmonic is $N \times$ precision on $F_0$
    - $F_0 = 100$ Hz; $T_0 = F_s/100$ samples
    - if error = 1 sample on $T_0$
      - $= 100 - F_s/(F_s/(100+1))$ Hz on $F_0$
      - $= 10000/(F_s+10000)$ Hz on $F_0$

- Precision required: at least 1/8th sample

- Get the s.d.f of the « noise » part with:
  \[ x_n(t) \approx x(t) - x_h(t) \]

**H/N modeling errors**

- Harmonics are found in noise (no orthogonality)
- $F_0$ is not constant on the analysis frame ⇒ broadening of harmonic lobes at high frequencies
- In real speech, noise is correlated with harmonics in the time domain

**Hybrid H/N synthesis**

- Harmonics:
  - $\sum$ of cosines
  - $\sum$ of outputs of digital oscillators whose frequencies are set to that of harmonics
  - $\sum$ of main lobes in the freq. domain + IFFT

- Noise:
  - $\sum$ of narrow band noises, obtained by (amplitude) modulating a low frequency noise with harmonics
  - frequency-domain noise matching the required s.d.f. + IFFT

--

Copyright (c)2002 Faculté Polytechnique de Mons - T. Dutoit
Contents

- Production models
  - Autoregressive modeling of speech
  - Autoregressive modeling of stationary random signals
    - AR model - Estimation algorithms
  - Autoregressive estimation of speech
  - Extensions of the AR model
- Transform-based models (*phenomenological*)
  - Short-term Fourier transform
    - Definition - Interpretation
    - Relationship with AR modeling
    - Hybrid harmonic+noise model - Estimation
  - Homomorphic (or cepstral) transform
    - Definition - Properties
    - Relationship with AR modeling
- F0 estimation

Real cepstrum

- $c_x(n)$ is the $\mathcal{g}^{-1}$ of the amplitude spectrum of $x(n)$ in neper ($\approx$ in dB)
- $c_x(n)$ can be computed with FFT$^{-1}$, provided $N_{FFT}$ is big enough (to avoid undersampling of $X(\omega)$)
- NB: cepstrum, quefrequency, liftering :-)

Homomorphic transform

- Complex cepstrum $x(n)$

- Convolution becomes summation:
  \[ x(n) = u(n) * h(n) \rightarrow X(z) = U(z)H(z) \rightarrow x(n) = u(n) + h(n) \]
- If $X(z)$ rational, then $x(n)$ falls faster than $1/n$

Applications to speech

- If $x(n)$: measure $T_0$ : easier on $c_x(n)$ than on $x(n)$
- *Measure the spectral envelope of $x(n)$: Isolate $c_h(n) \rightarrow h(n)$*
Applications to speech

Cepstral Mean Subtraction

If \( x(n) = h(n) \) (tel. line)

\[
\begin{align*}
  c_x(n) &= c_{\text{speech}}(n) + c_{\text{line}}(n) \\
  \mathbb{E}[c_x(n)] &= \mathbb{E}[c_{\text{speech}}(n)] + \mathbb{E}[c_{\text{line}}(n)] \\
  &= \mathbb{E}[c_{\text{line}}(n)] + k \quad (\text{estimator of } c_{\text{line}}(n))
\end{align*}
\]

\( c_{\text{CMS}}(n) = c_x(n) - \mathbb{E}[c_x(n)] \)

\( c_{\text{CMS}}(n) \) is independent from the channel

*Used in speech recognition*

Relationship with AR model

The first values of the cepstrum are characteristic of the transfer function of the vocal tract \( \Rightarrow \) must be related to \( \{a_i\} \)

\[
\begin{align*}
  \ln(\frac{1}{A(z)}) &= \sum_{n=0}^\infty x(n) z^{-n} \\
  -A'(z) / A(z) &= \sum_{n=0}^\infty n x(n) z^{-(n+1)} \\
  -\sum_{i=0}^n i a_i z^{-(i+1)} &= \left[ \sum_{i=0}^n a_i z^{-i} \right] \left[ \sum_{n=0}^\infty n x(n) z^{-(n+1)} \right] ... 
\end{align*}
\]

Contents

- Production models
  - Autoregressive modeling of speech
  - Autoregressive modeling of stationary random signals
    - AR model - Estimation algorithms
    - Autoregressive estimation of speech
    - Extensions of the AR model
- Transform-based models *(phenomenological)*
  - Short-term Fourier transform
    - Definition - Interpretation
    - Relationship with AR modeling
    - Hybrid harmonic+noise model - Estimation
  - Homomorphic (or cepstral) transform
    - Definition - Properties
    - Relationship with AR modeling
- F0 estimation
- F0 estimation
  - Short-term estimation of F0 on an analysis frame (at least 2 times the local pitch period); produces candidates for F0, with score
  - Long-term post-processing, making the ultimate choice among candidates (on a 3 frames-basis, usually), as a function of their score and taking into account constraints and correcting estimation biases
**Autocorrelation-based F0 estimation**

- Periodicity of $x$ implies a max in $r_x(k)$ for $kT_s \approx T_0$
  $$r(k) = \sum_{n=-\infty}^{\infty} x(n)x(n+k) \quad \text{for } k = K_{\text{min}}, ..., K_{\text{max}}$$

- Computational load $\gg$
- F0 doubling for females, because of $F1 \approx H2$
  $\Rightarrow$ use threshold($k$)

**Simplified Inverse Filtering Technique (SIFT)**

- Finds the max of the autocorrelation of the error signal ($\Rightarrow F1 \approx H2 : \text{flat envelope}$!)
- Decimation-interpolation for lower computational cost

**Cepstrum-based F0 estimation**

- Cepstrum is $\text{FFT}^{-1}(\log\text{Spectrum})$
- Log spectrum is flatter than Spectrum
- Cepstrum is more pulse-like than signal

---

[Diagram of SIFT process]
Estimating F0 with a comb filter

- Idea: filter signal with comb filter whose F0 varies from 70 to 500 Hz, and measure the energy of the output: \( F_0 = F(\max(\text{energy})) \)

- Usually performed in the frequency domain directly: \( F_0 = F(\max(\langle FFT(x)^2, H^2(\text{filter})\rangle)) \)

Conclusion

- LP model is good at modeling the short-term spectral envelope (id. for cepstrum)
- MP-LPC, CELP, Hybrid H/N add a better modeling of the contribution of the excitation component (the fine spectral structure)
- NB: these were models of the acoustic component of speech only...

Post-processing

- Based on Dynamic Programming:
  From the possible F0 values for successive frames, find the sequence which minimizes a cost function

- Logical filtering
  - ex: one V frame in an island of UV frames \( \Rightarrow \) UV