Speech Analysis

1. Short-Time analysis

2. Short-Time analysis in the time domain
   - Short-time Energy.
   - Short-time zero-crossing
   - Short-time autocorrelation.

3. Short-Time analysis in the frequency domain
   - Short-Time Fourier Transform

4. Linear Prediction

5. Homomorphic analysis: Cepstrum


**Short-Time Analysis**

**Introduction**

- Stationary analysis and the speech signal

\[
Q_n = \sum_{m=-\infty}^{\infty} T[x(m)w(n-m)]
\]

- Window
- Transformation
- Parameter
**Short-Time Analysis in the Time Domain**

- **Short-Time Energy**

  \[ E_n = \sum_{m=-\infty}^{\infty} \left( x(m)w(n-m) \right)^2 \]

\[ x(n) \xrightarrow{\cdot^2} x^2(n) \xrightarrow{h(n)} E_n \]

\[ h(n) = w^2(n) \]
Short-Time Analysis in the Time Domain

Short-time zero-crossing

\[ RZ_n = \frac{1}{N} \sum_{m=-\infty}^{\infty} \left\{ \frac{1}{2} |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]| w(n-m) \right\} \]

where

\[ \text{sgn}[x(n)] = \begin{cases} 
1 & x(n) \geq 0 \\
-1 & x(n) < 0 
\end{cases} \]

\[ 0 \leq RZ_n < \frac{N-1}{N} \]

\[ RZ_n = \frac{2F}{F_s} \]
Short-Time Analysis in the Time Domain

- **Short-Time Autocorrelation**

\[
R_n(k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)x(m-k)w(n-m-k)
\]

\[
x(n) \rightarrow x(n) \rightarrow h_k(n) \rightarrow R_n(k)
\]

\[
h_k(n) = w(n)w(n-k)
\]

Applications:
- Pitch Estimation
- Linear Prediction

Different for each \(k\)
Short-Time Analysis in the Frequency Domain

- Short-Time Fourier Transform

\[ X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega m} \]
Linear Prediction Coding (LPC)

- Representation of the spectral characteristics in a precise and efficient form
- A few parameters, Simple calculations
- Linear Prediction
  \[ \hat{s}(n) = -\sum_{i=1}^{p} a_is(n-i) \]
- Prediction Error
  \[ e(n) = s(n) - \hat{s}(n) \]

\[ P(z) = -\sum_{i=1}^{p} a_iz^{-i} \quad A(z) = 1 + \sum_{i=1}^{p} a_iz^{-i} \]
\[ \hat{s}(n) = -\sum_{i=1}^{p} a_i s(n - i) \quad e(n) = s(n) - \hat{s}(n) \]

**Minimum Mean Square Error Criterion**

\[
E\{e^2(n)\} = E\left\{ s(n) + \sum_{i=1}^{p} a_i s(n - i) \right\}^2 \]

\[
\frac{\partial E\{e^2(n)\}}{\partial a_k} = 2 E\left\{ \left[ s(n) + \sum_{i=1}^{p} a_i s(n - i) \right] s(n-k) \right\} =
\]

\[ = E \left\{ s(n)s(n-k) + s(n-k)\sum_{i=1}^{p} a_i s(n - i) \right\} = 0 \quad k = 1, 2, ..., P \]

\[
E\{s(n)s(n-k)\} = -\sum_{i=1}^{p} a_i E\{s(n-i)s(n-k)\} \iff R_s(k) = -\sum_{i=1}^{p} a_i R_s(i-k) \quad k = 1, ..., P \]
**Coefficients computation**

**Yule-Walker Equations**

\[
\begin{bmatrix}
R_s(0) & R_s(1) & \ldots & R_s(N-1) \\
R_s(1) & R_s(0) & \ldots & R_s(N-2) \\
\vdots & \vdots & \ddots & \vdots \\
R_s(p-1) & R_s(p-2) & \ldots & R_s(0)
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
\vdots \\
a_p
\end{bmatrix}
= -
\begin{bmatrix}
R_s(1) \\
R_s(2) \\
\vdots \\
R_s(p)
\end{bmatrix}
\]

\[
R_s a = -B_s \Rightarrow a = -R_s^{-1} B_s
\]

**Autocorrelation Matrix**

Two estimation methods:

a) **Autocorrelation Method**

A toeplitz matrix is obtained

Levinson-Durbin Algorithm

b) **Covariance Method**

A non-toeplitz matrix is obtained

Resolution using factorial decomposition (Cholesky)
**Levinson-Durbin Algorithm**

INICIO

\[ E^{(0)} = R(0) \]

\[ i: 1 \ldots p \]

\[ K_i = \frac{1}{E^{(i-1)}} \left\{ R(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j) \right\} \]

\[ a_j^{(i)} = -K_i \]

\[ j: 1 \ldots i-1 \]

\[ a_j^{(i)} = a_j^{(i-1)} - K_i a_{i-j}^{(i-1)} \]

\[ j \]

\[ E^{(i)} = (1 - K_i^2) E^{(i-1)} \]

\[ i \]

INICIO
**PARCOR Coefficients**

In order to go from a predictor of order $i$ to one of order $i+1$ it is necessary to recompute all the prediction coefficients.

The partial correlation coefficients (PARCOR) allow to obtain coefficients of order $i$ from those of order $i-1$

$$a^{(i)}_j = a^{(i-1)}_j - K_i a^{(i-1)}_{i-j}$$

Sometimes, it can turn out more convenient to work with the PARCOR.

We passed from transversal structures to lattice structures.
**PARCOR Coefficients**

\[ s(n) \xrightarrow{z^{-1}} b^0(n) \]

\[ b^i(n) \quad \text{Backward Error Order i} \]

\[ e^i(n) \quad \text{Forward Error Order i} \]

\[ K^i = \frac{\sum_{n} e^{i-1}(n)b^{i-1}(n-1)}{\sqrt{\sum_{n} \left(e^{i-1}(n)\right)^2 \sum_{n} \left(b^{i-1}(n-1)\right)^2}} \quad \text{Reflexion coefficients of the acoustic tubes model} \]
LPC Analysis

The speech is a all-pole signal (except nasal sounds)

\[ H(z) = \frac{G}{1 - \sum_{i=1}^{P} \alpha_i z^{-i}} \]

\[ H(z) = \frac{G}{A(z)} \]
LPC Analysis

Spectral envelope
LPC Analysis

Matlab code:

% "s" is a speech signal (8 kHz of sampling freq.)
% Framing and windowed
x=s(11740:11740+239);  an specific frame of the signal s
xx=x.*hamming(240);  Hamming window
% Estimation of the Power Spectral Density
X=fft(xx,1024);
Sx=abs(X).^2./1024;
% LPC analysis of order 8
[A,g]=lpc(xx,8);
% AR Power Spectral Density Estimation
Sxar=g./abs(fft(A,1024)).^2;
% error signal
e=filter(real(A),sqrt(g),x);
LPC Analysis

/B</a/, Autocorrelation method, Hamming window, P=6

SIGNAL

ERROR

SIGNAL SPECTRUM

ERROR SPECTRUM
LPC Analysis

//<m>, Autocorrelation method, Hamming window, P=3

SIGNAL

ERROR

SIGNAL SPECTRUM

ERROR SPECTRUM
LPC Analysis

Spectrum estimation with different orders

- P=2
- P=4
- P=6
- P=8
- P=12
- P=16
- P=200
LPC Analysis

Error signal for a voiced sound
LPC Analysis

Error signal for an unvoiced sound
Comparing two speech signals: Itakura distortion/distance

Assuming that $x[n]$ and $y[n]$ are AR process, let’s find a distance between both.

$$D(x, y) = \log \frac{a_y^H R_{xx} a_y}{a_x^H R_{xx} a_x}$$
Homomorphic Analysis

1. Cepstrum
2. A model for the environmental degradation
3. Perceptual representations
   - MFCC: Mel-Frequency Cepstrum Coefficients
   - PLP: Perceptual Linear Prediction
   - Transformación Bilineal
Homomorphic Analysis

The homomorphic transformation: \[ \hat{x}[n] = D(x[n]) \]
transform the convolution \[ x[n] = e[n] * h[n] \]
to an addition \[ \hat{x}[n] = \hat{e}[n] + \hat{h}[n] \]

The **homomorphic system** will be composed by:

\[ \begin{align*}
 & \text{D[ ]} & \text{+} & \text{L[ ]} & \text{+} & \text{D}^{-1}[ ] \\
 & x[n] & \star & \hat{x}[n] & \star & \hat{y}[n] \\
\end{align*} \]
Homomorphic Analysis

Analysis systems

\[ D[\cdot] \]

\[ Z[\cdot] \]
\[ E[z] \quad H[z] \]
\[ \ln(E[z]) + \ln(H[z]) \]
\[ \hat{e}[n] + \hat{h}[n] \]
Homomorphic Analysis

Synthesis System

\[
\begin{align*}
Z[n] & \quad \hat{e}[n] + \hat{h}[n] \\
\ln(E[z]) + \ln(H[z]) & \quad E[z] \cdot H[z]
\end{align*}
\]

\[
\begin{align*}
e^{(\cdot)} & \quad \ast
\end{align*}
\]
Homomorphic Analysis

SPECTRUM  ➔  CEPSTRUM

Complex Cepstrum

\[ \hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln[X(e^{j\omega})] e^{j\omega n} \, d\omega \]

but

\[ \ln[X(e^{j\omega})] = \ln|X(e^{j\omega})| + j\text{arg}[X(e^{j\omega})] \]

\[ \hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln|X(e^{j\omega})| e^{j\omega n} \, d\omega + \frac{j}{2\pi} \int_{-\pi}^{\pi} \text{arg}[X(e^{j\omega})] e^{j\omega n} \, d\omega \]

If \( x[n] \) is real, the complex cepstrum is real
Homomorphic Analysis

Real Cepstrum

\[ c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln|X(e^{j\omega})|e^{j\omega n} d\omega \]

The real cepstrum is the even part of the complex cepstrum

\[ c[n] = \frac{\hat{x}[n] + \hat{x}[-n]}{2} \]

Independent variable n → Quefrency → time
Homomorphic Analysis

Cepstrum of the speech signal

Excitation $e[n]$: 
- Periodic sequence of pulses
- White noise

Filter: 
- Rational transfer function ARMA
Homomorphic Analysis

Cepstrum of an ARMA system (poles-zeros)

\[
H[z] = \prod_{k=1}^{M_i} (1-a_k z^{-1}) \prod_{k=1}^{M} (1-u_k z^{-1}) \prod_{k=1}^{N_i} (1-b_k z^{-1}) \prod_{k=1}^{N} (1-v_k z^{-1})
\]

Taking logarithms

\[
\ln[H(z)] = \ln[A] + \ln[z^D] + \sum_{k=1}^{M_i} \ln(1-a_k z^{-1}) + \sum_{k=1}^{M} \ln(1-u_k z^{-1}) - \sum_{k=1}^{N_i} \ln(1-b_k z^{-1}) - \sum_{k=1}^{N} \ln(1-v_k z^{-1})
\]
**Homomorphic Analysis**

Using the Taylor series

\[
\ln(1-x) = - \sum_{n=1}^{\infty} \frac{x^n}{n}
\]

Thus, the term associated with the inner poles

\[
\sum_{k=1}^{N} \ln(1-b_k z^{-1}) = - \sum_{n=1}^{\infty} \left( \sum_{k=1}^{N} \frac{b^n_k}{n} \right) z^{-n}
\]

Sequence over the Z transform is computed
Homomorphic Analysis

The term $\ln[z^D]$ gives information about the time beginning and usually we don’t take into account.

$$
\hat{h}[n]=\begin{cases}
\ln[|A|]+j\sigma & n=0 \\
\sum_{k=1}^{N_i} b_k^n - \sum_{k=1}^{M_i} a_k^n & n>0 \\
\sum_{k=1}^{M_o} u_k^{-n} - \sum_{k=1}^{N_o} v_k^{-n} & n<0
\end{cases}
$$

$$
\begin{align*}
\sigma &= \pi \text{ si } A<0 \\
\sigma &= 0 \text{ si } A>0
\end{align*}
$$
Homomorphic Analysis

Properties

1. If the filter does not have poles and zeros outside the unit circle (minimum phase system)

\[ \hat{h}[n] = 0 \quad \text{para } n < 0 \]

2. In minimum phase systems the complex cepstrum is totally defined by its real cepstrum

\[ \hat{h}[n] = \begin{cases} 0 & n < 0 \\ c[n] & n = 0 \\ 2c[n] & n > 0 \end{cases} \]

3. The cepstrum is an infinite sequence but decays like $1/n \rightarrow$ concentrated in the origin
Homomorphic Analysis

In the case of an AR filter (all-poles)

\[ H(z) = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}} \]

Gain and coefficients are calculated using Linear Prediction

\[ \hat{h}[n] = \begin{cases} 0 & n < 0 \\ \ln[G] & n = 0 \\ a_n + \sum_{k=1}^{n-1} \left( \frac{k}{n} \right) \hat{h}[k] a_{n-k} & 0 < n \leq p \\ \sum_{k=n-p}^{n-1} \left( \frac{k}{n} \right) \hat{h}[k] a_{n-k} & n > p \end{cases} \]
Homomorphic Analysis

Cepstrum of a periodic signal (windowed)

\[ e[n] = \sum_{k=0}^{M-1} \alpha_k \delta[n-kN] \]

Z transform

\[ E[z] = \prod_{k=1}^{M} (1-a_k z^{-N}) \]

Cepstrum

\[ \hat{e}[n] = \sum_{m=1}^{M} \frac{a_k^m}{m} \delta(n-mN) \quad m > 0 \]
Homomorphic Analysis

Limit case: infinite train of unit impulses

\[ e[n] = \sum_{k=0}^{\infty} \delta(n-kN) \]

\[ \hat{e}[n] = \sum_{m=1}^{\infty} \frac{1}{m} \delta(n-mN) \]

Properties:

1. Is non zero only in integer multiples of N, being zero at the origin
2. Has a decay of 1/m with the time
Homomorphic Analysis

Short-Time Cepstrum of the speech signal: Computed using the short-time Fourier Transform

If \( x[n] \) is the windowed sequence

\[
X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N} \quad 0 \leq k < N
\]

\[
\hat{x}_p[n] = \frac{1}{N} \sum_{k=0}^{N-1} \ln X[k] e^{j2\pi kn/N} = \sum_{r=-\infty}^{\infty} \hat{x}[n+rN]
\]

To reduce aliasing \( \rightarrow N \) large
Homomorphic Analysis
Short-Time Cepstrum for man voice

Filter

Pitch
Homomorphic Analysis

Short-Time Cepstrum for a woman's voice

Speech Technologies – Speech Analysis
Homomorphic Analysis

Separation excitation-filter by homomorphic filtering

Filtering → Lifting: homomorphic filtering

![Graph showing cepstrum real vs. samples with labels for Filter and Excitation.]
Homomorphic Analysis

研究表明，在语音分析中使用霍马多摩分析方法可以提高语音处理的准确性和鲁棒性。以下是一个示例：

左图：原始语音信号

右图：霍马多摩分析后的结果

通过霍马多摩分析，可以有效地提取语音信号的特征，从而实现更精确的语音识别和处理。
Homomorphic Analysis

Unvoiced segment

Speech Technologies – Speech Analysis
Homomorphic Analysis

Homomorphic filtering for blind deconvolution

Speech signal with an unknown convolutive distortion

\[ x[n] = s[n] \times h[n] \]

Suposition: the filter \( h[n] \) is time invariant

In the cepstral domain

\[ \hat{x}[n] = \hat{s}[n] + \hat{h}[n] \]

Taking the mean

\[ E(\hat{x}[n]) = E(\hat{s}[n]) + E(\hat{h}[n]) = E(\hat{s}[n]) + \hat{h}[n] \]

Mean sustraction

\[ \hat{x}[n] - E(\hat{x}[n]) = \hat{s}[n] - E(\hat{s}[n]) \]

The filter disappear, but.....
Homomorphic Analysis

Homomorphic filtering for blind deconvolution

... The mean cepstrum of the speech signal is also removed

\[ E(\hat{s}[n]) \]

which it is related to the long term average power spectral density of the speech signal:

- lowpass behavior influenced to a great extent by the glotal pulse

The short-time analysis it is influenced by the window. Windows of length much greater are required than the length of the impulsive response to remove.

Example
A Model for Environmental Degradation

In a room (vehicle):
- Additive noise from several sources
- Multipath noise: convolutional distortion
A Model for Environmental Degradation

Respuesta impulsional habitaculo coche

Original

Filtered
A Model for Environmental Degradation

Respuesta Frecuencial

dB

Hz

0 1000 2000 3000 4000
A Model for Environmental Degradation

Degraded Signal

Time

Frequency
La señal de voz en general puede venir degradada por distorsiones aditivas y convolutivas:

Ruido aditivo de fuentes independientes

Distorsión de canal

\[ x[n] = s[n] \cdot h[n] + r[n] \]

\[
X(e^{j\omega}) = S(e^{j\omega}) \cdot H(e^{j\omega}) + R(e^{j\omega}) \cdot \cos(\theta(\omega)) + 2 \cdot S(e^{j\omega}) \cdot H(e^{j\omega}) + R(e^{j\omega}) \cdot \cos(\theta(\omega))
\]

Valor medio cero
A Model for Environmental Degradation

\[ |X(e^{j\omega})|^2 \approx |S(e^{j\omega})|^2 |H(e^{j\omega})|^2 + |R(e^{j\omega})|^2 \]

Tomando logaritmos

\[ \ln |X(e^{j\omega})|^2 \approx \ln |S(e^{j\omega})|^2 + \ln |H(e^{j\omega})|^2 + \]

\[ \ln \left( 1 + \frac{|R(e^{j\omega})|^2}{|S(e^{j\omega})|^2 |H(e^{j\omega})|^2} \right) \]
A Model for Environmental Degradation

\[ \ln \left| X(e^{j\omega}) \right|^2 \approx \ln \left| S(e^{j\omega}) \right|^2 + \ln \left| H(e^{j\omega}) \right|^2 + \ln \left( 1 + \exp \left( \ln \left| R(e^{j\omega}) \right|^2 - \ln \left| S(e^{j\omega}) \right|^2 - \ln \left| H(e^{j\omega}) \right|^2 \right) \right) \]

Pasando al dominio cepstral

\[ \hat{x}[n] = \hat{s}[n] + \hat{h}[n] + g\left( \hat{r}[n] - \hat{s}[n] - \hat{h}[n] \right) \]

Donde \( g(z) \) es una función no lineal

\[ g[z] = \text{Cepstrum} \left( \ln(1 + \exp \left( \text{Cepstrum}^{-1}(z) \right) \right) \]
Perceptual Representations

MFCC: Mel-Frequency Cepstrum Coefficients

Cepstrum real sobre una escala frecuencia no lineal: Escala perceptual MEL

\[ \text{Mel}(x) = 2595 \log_{10} \left( 1 + \frac{x}{700} \right) \]
Speech Processing, Transmission and Quality aspects (STQ);
Distributed speech recognition;
Front-end feature extraction algorithm;
Compression algorithms

Input speech:
- ADC
- Offset compensation
- Framing
- PE
- W
- FFT
- MF
- LOG
- DCT

Abbreviations:
- ADC: analog-to-digital conversion
- Offset compensation
- PE: pre-emphasis
- logE: energy measure computation
- W: windowing
- FFT: fast Fourier transform (only magnitude components)
- MF: mel-filtering
- LOG: nonlinear transformation
- DCT: discrete cosine transform
- MFCC: mel-frequency cepstral coefficient
Perceptual Representations

Magnitude spectra of critical band filters

$H_i(\omega)$

$i$th conceptual critical band filter centered on frequency $F_{c,i}$

$F (kHz)$
Perceptual Representations

PLP: Perceptual Linear Prediction (Hermansky, 90)

Motivación: incluir propiedades perceptuales en el análisis LPC
Perceptual Representations

Transformación bilineal

\[ s = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \]

Célula pasatodo \( 0 < \alpha < 1 \)

Hace un mapeado en el plano complejo de la circunferencia de radio unidad sobre sí misma

\[ \Omega = \omega + 2 \arctan \left( \frac{\alpha \sin(\omega)}{1 - \alpha \cos(\omega)} \right) \]

Eligiendo adecuadamente \( \alpha \) es similar a Bark o Mel