The Cepstrum and Homomorphic Processing

Introduction

- *The Quefrency Alanysis of Time Series for Echoes*
- "In general, we find ourselves operating on the frequency side in ways customary on the time side and vice versa"

Cepstrum

Definition

The *cepstrum* of a signal $x[n]$ is defined by

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

where $X(e^{j\omega})$ is the DTFT of $x[n]$. From (1), we can see that $c[n]$ is the inverse DTFT of the logrithm of the magnitude spectrum $\log|X(e^{j\omega})|$.

Complex Cepstrum

The *complex cepstrum* is defined by

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

Homomorphic Systems

Oppenheim defined classes of non-linear systems based on a *generalized principle of superposition*. Such systems are called *homomorphic systems*.

The Principle of Superposition

For a linear system $\mathcal{L}(\cdot)$, the principle of superposition is obeyed. That is,

$$\begin{aligned} &\mathcal{L}(x_1[n] + x_2[n]) = \mathcal{L}x_1[n] + \mathcal{L}x_2[n] \\ &\mathcal{L}(cx[n]) = c\mathcal{L}x[n] \end{aligned} \tag{3}$$

as shown in the following figure

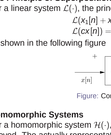

Figure: Conventional Linear System

Homomorphic Systems

For a homomorphic system $\mathcal{H}(\cdot)$, the generalized principle of superposition is obeyed. The actually representation depends on the operation of interest. In the case of convolution, we have

$$y[n] = \mathcal{H}(x_1[n] * x_2[n]) = \mathcal{H}x_1[n] * \mathcal{L}x_2[n] = y_1[n] * y_2[n] \tag{4}$$

as shown in the following figure

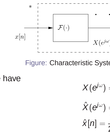

Figure: A Homomorphic System for Convolution

Canonical Homomorphic System for Convolution

A homomorphic system (for convolution) is equivalent to a cascade of three homomorphic systems depicted as follows

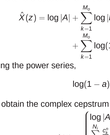

Figure: Canonical System for a Homomorphic System for Convolution

$\mathcal{D}_*(\cdot)$ is called the *characteristic system for convolution*, while $\mathcal{D}_*^{-1}(\cdot)$ is called *the inverse characteristic system for convolution*. Both are (normally) fixed systems. The linear system $\mathcal{L}(\cdot)$ is to be designed for special applications. The input-output relations of the subsystems are as follows.

$$\begin{aligned} &\hat{x}[n] = \mathcal{D}_*(x_1[n] * x_2[n]) = \mathcal{D}_*x_1[n] + \mathcal{D}_*x_2[n] = \hat{x}_1[n] + \hat{x}_2[n] \\ &\hat{y}[n] = \mathcal{L}(\hat{x}_1[n] + \hat{x}_2[n]) = \mathcal{L}\hat{x}_1[n] + \mathcal{L}\hat{x}_2[n] = \hat{y}_1[n] + \hat{y}_2[n] \\ &y[n] = \mathcal{D}_*^{-1}(\hat{y}_1[n] + \hat{y}_2[n]) = \mathcal{D}_*^{-1}\hat{y}_1[n] * \mathcal{D}_*^{-1}\hat{y}_2[n] = y_1[n] * y_2[n] \end{aligned} \tag{5}$$

Representation by DTFT

A characteristic system for homomorphic deconvolution using DTFT is

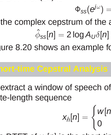

Figure: Characteristic System for Homomorphic System using DTFT

We have

$$\begin{aligned} &X(e^{j\omega}) = \sum_n x_0[n] e^{-j\omega n}, \\ &\hat{X}(e^{j\omega}) = \log\{X(e^{j\omega})\}, \\ &\hat{x}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}(e^{j\omega}) e^{j\omega n} d\omega \end{aligned} \tag{6}$$

It is easy to show that

$$z[n] = x[n] * y[n] \quad \Rightarrow \quad \hat{z}[n] = \hat{x}[n] + \hat{y}[n] \tag{7}$$

Representation by z-Transform

A similar characteristic system can be defined using z-transform instead of DTFT. See Figures 8.9 and 8.10.

Computation of the Complex Cepstrum

Rational z-Transforms

A rational z-transform can be represented by

$$X(z) = \frac{A\prod_{k=1}^{M_i}(1 - a_k z^{-1})\prod_{k=1}^{M_o}(1 - b_k^{-1} z^{-1})}{\prod_{k=1}^{N_i}(1 - c_k z^{-1})} \tag{8}$$

where a_k is a zero inside the unit circle, c_k is a pole inside the unit circle, and b_k^{-1} is a zero outside the unit circle. Taking the logarithm, we have

$$\begin{aligned} \hat{X}(z) = \log|A| &+ \sum_{k=1}^{M_i}\log|b_k^{-1}| + \log(z^{-M_i}) + \sum_{k=1}^{M_i}\log(1 - a_k z^{-1}) \\ &- \sum_{k=1}^{M_o}\log(1 - b_k z) - \sum_{k=1}^{N_i}\log(1 - c_k z^{-1}) \end{aligned} \tag{9}$$

Using the power series,

$$\log(1 - a) = -\sum_{n=1}^{\infty}\frac{a^n}{n}, \quad |a| < 1, \tag{10}$$

we obtain the complex cepstrum

$$\hat{x}[n] = \begin{cases} \log|A| + \sum_{k=1}^{M_i}\log|b_k^{-1}|, & n = 0 \\ \sum_{k=1}^{M_i}\frac{a_k^n}{n} - \sum_{k=1}^{N_i}\frac{c_k^n}{n}, & n > 0 \\ -\sum_{k=1}^{M_o}\frac{b_k^{-n}}{n}, & n < 0 \end{cases} \tag{11}$$

Homomorphic Analysis of Speech Signal

Our basic assumption for speech production is based on convolution. Specifically, for voiced speech, we assume

$$s[n] = p[n] * h_v[n], \quad \text{where} \quad h_v[n] = A_v \cdot g[n] * v[n] * r[n]. \tag{12}$$

For unvoiced speech, we assume

$$s[n] = u[n] * h_u[n], \quad \text{where} \quad h_v[n] = A_u \cdot v[n] * r[n]. \tag{13}$$

In (12) and (13), $g[n]$, $v[n]$, and $r[n]$ are the glottal pulse, the vocal tract impulse response, and the radiation load response respectively. $p[n]$ is the quasi-periodic excitation, $u[n]$ is the unvoiced excitation signal with unit variance, and A_v, A_u are the gains.

Example of Voiced Speech

Figure 8.13 shows the respective time-domain functions of $g[n]$, $v[n]$, $r[n]$, $p[n]$ for a model for sustained vowel /AE/. Figure 8.14 shows the pole-zero plots of the respective z-transforms. Figure 8.15 shows the log magnitudes of the respective DTFTs. Figure 8.17 shows the respective complex cepstra. Figure 8.18 shows the speech complex cepstrum and cepstrum.

Example of Unvoiced Speech

For unvoiced speech, there is no glottal pulse. The periodic excitation is replaced by a random noise function $u[n]$. The *autocorrelation representation* is thus modeled by

$$\phi_{hh}[n] = \phi_{vv}[n] * \phi_{rr}[n] * A_v^2 \phi[n] = A_v^2 \phi_{vv}[n] * \phi_{rr}[n]. \tag{14}$$

The DTFT of (14) is

$$\Phi_{hh}(e^{j\omega}) = A_v^2 |V(e^{j\omega})|^2 |R(e^{j\omega})|^2, \tag{15}$$

so the complex cepstrum of the autocorrlation function is

$$\hat{\phi}_{hh}[n] = 2\log A_v \phi[n] + \{\hat{v}[n] + \hat{v}[-n]\} + \{\hat{r}[n] + \hat{r}[-n]\} \tag{16}$$

Figure 8.20 shows an example for unvoiced /AE/.

Short-time Cepstral Analysis

To extract a window of speech of length L beginning at $\hat{n}$, we define a finite-length sequence

$$x_{\hat{n}}[n] = \begin{cases} w[n]s[\hat{n} + n] & 0 \le n \le L - 1 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

The DTFT of $x_{\hat{n}}[n]$ is the short-time Fourier transform of $s[n]$

$$X_{\hat{n}}(e^{j\omega}) = \sum_{n=0}^{L-1} x_{\hat{n}}[n] e^{-j\omega n}. \tag{18}$$

(6) is modified to be

$$\begin{aligned} &X_{\hat{n}}(e^{j\omega}) = \sum_{n=0}^{L-1} x_{\hat{n}}[n] e^{-j\omega n}, \\ &\hat{X}_{\hat{n}}(e^{j\omega}) = \log\{X_{\hat{n}}(e^{j\omega})\}, \\ &\hat{x}_{\hat{n}}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}_{\hat{n}}(e^{j\omega}) e^{j\omega n} d\omega \end{aligned} \tag{19}$$

Short-time Cepstral Analysis Based on DFT

Using DFT, (19) is simplified (approximated) to

$$\begin{aligned} &X_{\hat{n}}[k] = \sum_{n=0}^{L-1} x_{\hat{n}}[n] e^{-j\frac{2\pi}{N}kn}, \quad 0 \le k \le N - 1 \\ &\hat{X}_{\hat{n}}[k] = \log\{X_{\hat{n}}[k]\}, \quad 0 \le k \le N - 1 \\ &\hat{x}_{\hat{n}}[n] = \frac{1}{N}\hat{X}_{\hat{n}}[k] e^{j\frac{2\pi}{N}kn}, \quad 0 \le n \le N - 1 \end{aligned} \tag{20}$$

It can be shown that

$$\hat{x}_{\hat{n}}[n] = \sum_r \hat{x}_{\hat{n}}[n + rN], \quad 0 \le n \le N - 1 \tag{21}$$

Homomorphic Filtering of Natural Speech

We assume that over the length of window, say L, the speech signal satisfies

$$s[n] = e[n] * h[n], \tag{22}$$

where $e[n]$ is excitation and $h[n]$ is the impulse response from glottis to lip. Furthermore, we assume that $h[n]$ is short compared to L, so

$$x[n] = w[n]s[n] = w[n](e[n] * h[n]) \approx e_w[n] * h[n], \tag{23}$$

where $e_w[n] = w[n]e[n]$.

Voiced Speech

Let $e[n]$ be a unit impulse train of the form

$$e[n] = \sum_{k=0}^{N_p - 1}\delta[n - kN_0], \tag{24}$$

where N_p is the pitch period (in samples), and N_w is the number of impulses in a window. The windowed excitation is

$$e_w[n] = w[n]e[n] = \sum_{k=0}^{N_w - 1} w_{N_0}[k]\delta[n - kN_0], \tag{25}$$

where

$$w_{N_0}[k] = \begin{cases} w[kN_0] & 0 \le k \le N_w - 1 \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

Taking the DTFT of (25), we have

$$E_w(e^{j\omega}) = \sum_{k=0}^{N_w - 1} w_{N_0}[k] e^{-j\omega kN_0} = W_{N_0}(e^{j\omega N_0}), \tag{27}$$

which is periodic in ω with period $\frac{2\pi}{N_0}$. It follows that the log spectrum

$$\hat{X}(e^{j\omega}) = \log H_v(e^{j\omega}) + \log E_w(e^{j\omega}) \tag{28}$$

has two components: a slow-varying $\log H_v(e^{j\omega})$ and a periodic $\log E_w(e^{j\omega})$. Note that as $\log E_w(e^{j\omega})$ is also periodic, and the inverse DTFT produces an impulse train with period N_0. Thus the two components in

$$\hat{x}[n] = \hat{h}_v[n] + \hat{e}_w[n] \tag{29}$$

can be distinguished.

Using DFT

The above procedure can be carried out by DFT as well. The system is depicted in Figure 8.31.

Liftering

The desired component of the input can be selected by a window function in quefrency, denoted by $l[n]$. This is also called liftering. For excitation signal, a high-pass filtering of the log spectrum is used, while for the impulse response, the low-pass filtering is used. Figure 8.32 shows an example.