Lecture Notes in Speech Production, Speech Coding, and Speech Recognition

Mark Hasegawa-Johnson
University of Illinois at Urbana-Champaign

February 17, 2000

Chapter 5

Spectral and Cepstral Distance Measures

5.1 Homomorphic Analysis

Linear systems are homomorphic to addition:

$$L[x_1(n) + x_2(n)] = L[x_1(n)] + L[x_2(n)] \tag{5.1}$$

Linear filtering is useful for analyzing a signal with two additive components, e.g. $y(n) = x(n) + \epsilon(n)$. In speech, we are often more interested in "convolutional components." For example, the speech signal can be modeled as the convolution of a source function $p(n)$, a transfer function $t(n)$, and a radiation function $r(n)$:

$$x(n) = r(n) * (t(n) * p(n)) \tag{5.2}$$

In order to analyze $x(n)$, we want a nonlinear "filtering" system which is "homomorphic to convolution," that is,

$$H[t(n) * p(n)] = H[t(n)] * H[p(n)] \tag{5.3}$$

The system $H[\bullet]$ can be written as the series connection of a transformation $D[\bullet]$, a linear system $L[\bullet]$, and the inverse transformation $D^{-1}[\bullet]$:

$$H[t(n) * p(n)] = D^{-1}\left[L\left[D[t(n) * p(n)]\right]\right] \tag{5.4}$$

where $D[\bullet]$ is the transformation which converts convolution into addition:

$$D[t(n) * p(n)] = D[t(n)] + D[p(n)] \tag{5.5}$$

$D[x(n)]$ can be written as $D[x(n)] = \hat{x}(n)$, where $\hat{x}(n)$ is defined to be the *complex cepstrum* of $x(n)$. The form of the complex cepstrum is obvious if one considers the z transforms of $x(n)$ and $\hat{x}(n)$:

$$\left.\begin{aligned} X(z) &= R(z)T(z)P(z) \\ \hat{X}(z) &= \hat{R}(z) + \hat{T}(z) + \hat{P}(z) \end{aligned}\right\} \hat{X}(z) = \log(X(z)) \tag{5.6}$$

5.2 Definitions

5.2.1 Complex Cepstrum

$$\boxed{\hat{x}(n) = \tfrac{1}{2\pi}\int_0^{2\pi} \log(X(e^{j\omega}))e^{j\omega n}d\omega} \tag{5.7}$$

$$\hat{X}(e^{j\omega}) = \log(X(e^{j\omega})) = \log|X(e^{j\omega})| + j\widehat{\arg}(X(e^{j\omega})) \tag{5.8}$$

- The function $\widehat{\arg}(X(e^{j\omega})$ is the "unwrapped phase" of X. Recall that the principal argument, $\arg(X(e^{j\omega}))$, is only defined over the range of $(-\pi, \pi]$. Such a constraint is not appropriate for the definition of cepstrum, because we require that the sum of two cepstra should still be a valid cepstrum:

$$\widehat{\arg}(X(e^{j\omega})) = \widehat{\arg}(R(e^{j\omega})) + \widehat{\arg}(T(e^{j\omega})) + \widehat{\arg}(P(e^{j\omega})) \tag{5.9}$$

  This requirement can be met by adding integer multiples of 2π to the principal argument, as necessary, in order to produce a continuous, odd function of ω; this process is known as "unwrapping" the phase (the argument is only odd if $x(n)$ is real).

- $\hat{x}(n)$ is only defined if $\log(X(z))$ is a valid Z transform, uniformly defined on the unit circle.

- If $x(n)$ is real, then $\log|X(e^{j\omega})|$ is even, and $\widehat{\arg}(X(e^{j\omega}))$ is odd, and therefore $\hat{x}(n)$ is real.

- n is sometimes called "quefrency," especially in echo analysis applications. In speech analysis, n is usually called the cepstral "lag," just as the argument of $R(n)$ is called the autocorrelation "lag."

5.2.2 Cepstrum

$$\boxed{c(n) = \frac{1}{2\pi}\int_0^{2\pi} \log|X(e^{j\omega})|e^{j\omega n}d\omega} \tag{5.10}$$

$$\frac{\hat{x}(n) + \hat{x}(-n)}{2} = \frac{1}{2\pi}\int_0^{2\pi} \log|X(e^{j\omega})|e^{j\omega n}d\omega = c(n) \tag{5.11}$$

5.2.3 Example

$$x(n) = \delta(n) - \alpha\delta(n-N), \quad |\alpha| < 1 \tag{5.12}$$

$$X(z) = 1 - \alpha z^{-N} \tag{5.13}$$

$$\hat{X}(z) = \log(1-\alpha z^{-N}) = -\sum_{r=1}^{N} \frac{\alpha^r z^{-rN}}{r} \quad \text{if} \quad |\alpha z^{-N}| < 1 \tag{5.14}$$

$$\boxed{\hat{x}(n) = -\sum_{r=1}^{\infty} \frac{\alpha^r}{r}\delta(n-rN)} \tag{5.15}$$

$$\boxed{c(n) = (1/2)(\hat{x}(n) + \hat{x}(-n)) = -\sum_{r=1}^{\infty} \frac{\alpha^r}{2r}(\delta(n-rN) + \delta(n+rN))} \tag{5.16}$$

5.3 Minimum and Maximum Phase Sequences

Consider the class of signals whose spectra can be expressed as follows:

$$X(z) = G\frac{\prod_{k=1}^{N_a}(1-a_k z^{-1})\prod_{k=1}^{N_b}(1-b_k z)}{\prod_{k=1}^{N_c}(1-c_k z^{-1})\prod_{k=1}^{N_d}(1-d_k z)}, \quad |a_k|, |b_k|, |c_k|, |d_k| < 1 \tag{5.17}$$

For this class of signals, all stable minimum phase signals (all signals with $N_b = 0, N_d = 0$) are also causal, and all stable maximum phase signals ($N_a = 0, N_c = 0$) are also anti-causal. The cepstrum is:

$$\hat{x}(n) = \begin{cases} -\sum_{k=1}^{N_a} \frac{a_k^n}{n} + \sum_{k=1}^{N_c} \frac{c_k^n}{n} & n > 0 \\ \log(G) & n = 0 \\ \sum_{k=1}^{N_b} \frac{b_k^{-n}}{n} - \sum_{k=1}^{N_d} \frac{d_k^{-n}}{n} & n < 0 \end{cases} \tag{5.18}$$

- For signals in this class, $\hat{x}(n)$ is an infinite length signal, even if $x(n)$ is finite in length (the only exception is $x(n) = G\delta(n)$).
- $\hat{x}(n)$ decays exponentially fast as a function of n.
- Minimum-phase, causal sequences have causal $\hat{x}(n)$.
- Maximum-phase, anti-causal sequences have anti-causal $\hat{x}(n)$.

5.4 Recursive Formula for the Cepstral Coefficients

$$\hat{X}(z) = \log(X(z)) \tag{5.19}$$

$$\frac{d}{dz}\hat{X}(z) = \frac{1}{X(z)}\frac{d}{dz}X(z) \tag{5.20}$$

$$\left[-z\frac{d}{dz}\hat{X}(z)\right] X(z) = \left[-z\frac{d}{dz}X(z)\right] \tag{5.21}$$

$$n\hat{x}(n) * x(n) = nx(n) \tag{5.22}$$

$$\sum_{k=-\infty}^{\infty} k\hat{x}(k)x(n-k) = nx(n) \tag{5.23}$$

For $n \neq 0$, this yields

$$\boxed{\sum_{k=-\infty}^{\infty} \frac{k}{n}\hat{x}(k)x(n-k) = x(n)} \tag{5.24}$$

If $x(n)$ is minimum-phase and causal, the summation in the above equation is only non-zero for $0 \leq k \leq n$, yielding the following recursion for $\hat{x}(n)$:

$$\hat{x}(n) = \frac{x(n)}{x(0)} - \sum_{k=0}^{n-1} \frac{kx(n-k)}{nx(0)}\hat{x}(k), \quad n > 0 \tag{5.25}$$

If $x(n)$ is maximum-phase and anti-causal, the summation is only non-zero for $n \leq k \leq 0$, yielding the following formula for $\hat{x}(n)$:

$$\hat{x}(n) = \frac{x(n)}{x(0)} - \sum_{k=n+1}^{0} \frac{kx(n-k)}{nx(0)}\hat{x}(k), \quad n < 0 \tag{5.26}$$

In both cases, we have already shown that

$$\hat{x}(0) = \log(x(0)) \tag{5.27}$$

5.5 LPC Cepstrum

5.5.1 Complex Cepstrum

The cepstrum of the transfer function, $\hat{t}(n)$, can also be estimated from the LPC coefficients:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{5.28}$$

The LPC cepstrum $\hat{h}(m)$ is the inverse transform of $\log H(z)$:

$$\begin{aligned}
\hat{h}(m) &= \mathcal{Z}^{-1}\left(\log H(z)\right) && (5.29)\\
&= \mathcal{Z}^{-1}\left(\log G - \log A(z)\right) && (5.30)\\
&= \log G\delta(n) - \hat{a}(m) && (5.31)\\
&= \begin{cases} 0 & n<0 \\ \log(G) & n=0 \\ \hat{a}(m) & n>0 \end{cases} && (5.32)
\end{aligned}$$

Since $H(z)$ is minimum-phase, $\hat{h}(n)$ can be calculated from the log-spectrum of $H(z)$:

$$c(n) = \frac{1}{2\pi}\int_0^{2\pi} \log|H(e^{j\omega})|e^{j\omega n}d\omega \tag{5.33}$$

$$\hat{h}(n) = \begin{cases} 0 & n<0 \\ \log(G) & n=0 \\ 2c(n) & p\geq n>0 \end{cases} \tag{5.34}$$

Alternatively, since $A(z)$ is minimum-phase, $\hat{a}(m)$ is causal, and therefore $\hat{h}(m)$ is also a causal sequence. The form of $\hat{a}(m)$ can be computed from the following recursion, which can be derived by differentiating $\log A(z)$:

$$n\hat{a}(n) * a(n) = na(n) \tag{5.35}$$

$$\hat{h}(n) = \begin{cases} 0 & n<0 \\ \log(G) & n=0 \\ \alpha_n + \sum_{k=1}^{n-1}\frac{k}{n}\alpha_{n-k}\hat{h}(k) & p\geq n>0 \end{cases} \tag{5.36}$$

Notice that the first $p+1$ cepstral coefficients ($0\leq n\leq p$) contain a complete description of the transfer function; $\hat{h}(n)$ for larger n can be computed recursively from the first $p+1$ values of $\hat{h}(n)$.

5.5.2 LPC Power Cepstrum

In speech recognition, the cepstrum we work with most often is the inverse transform of $|H(z)|^2$:

$$\begin{aligned}
c_m &= \mathcal{Z}^{-1}\left(\log|H(z)|^2\right) && (5.37)\\
&= \mathcal{Z}^{-1}\left(\log G^2 - \log A(z) - \log A(z^{-1})\right) && (5.38)\\
&= \log G^2\delta(n) - \hat{a}(m) - \hat{a}(-m) && (5.39)\\
&= \log E_{min}\delta(n) - \hat{a}(m) - \hat{a}(-m) && (5.40)\\
& && (5.41)
\end{aligned}$$

Since $\hat{a}(n)$ is causal,

$$c_m = \begin{cases} \hat{h}(m) & m>0 \\ \hat{h}(m)/2 & m=0 \\ \hat{h}(-m) & m<0 \end{cases} \tag{5.42}$$

5.5.3 How is the LPC Cepstrum Usually Used?

1. Calculate LPC coefficients using autocorrelation method.
2. Convert into cepstral coefficients c_m, $0 \leq m$.
3. Calculate distances using cepstral coefficients. When deriving distance formulas, we must remember that c_m is an even sequence!!

5.6 Review

5.6.1 Complex Cepstrum

- The "complex cepstrum" $\hat{x}(n)$ is a *real* sequence. It is called the complex cepstrum because it has a complex Z transform.
- The complex cepstrum is defined to be

$$\boxed{\hat{x}(n) = \frac{1}{2\pi}\int_0^{2\pi} \log(X(e^{j\omega}))e^{j\omega n}d\omega} \tag{5.43}$$

- The complex cepstrum satisfies the equation

$$\boxed{\sum_{k=-\infty}^{\infty} k\hat{x}(k)x(n-k) = nx(n)} \tag{5.44}$$

5.6.2 Cepstrum

- The cepstrum $c(n)$ is a real sequence with a real Z transform.
- The cepstrum can be defined

$$\boxed{c(n) = \frac{1}{2\pi}\int_0^{2\pi} \log|X(e^{j\omega})|e^{j\omega n}d\omega} \tag{5.45}$$

- The cepstrum is the even part of the complex cepstrum

$$\boxed{c(n) = \frac{\hat{x}(n)+\hat{x}(-n)}{2}} \tag{5.46}$$

5.6.3 Signals with Rational Z Transforms

If $X(z)$ has the form

$$X(z) = G\frac{\prod_{k=1}^{N_a}(1-a_k z^{-1})\prod_{k=1}^{N_b}(1-b_k z)}{\prod_{k=1}^{N_c}(1-c_k z^{-1})\prod_{k=1}^{N_d}(1-d_k z)}, \qquad |a_k|,|b_k|,|c_k|,|d_k| < 1 \tag{5.47}$$

then the cepstrum is

$$\hat{x}(n) = \begin{cases} -\sum_{k=1}^{N_a}\frac{a_k^n}{n} + \sum_{k=1}^{N_c}\frac{c_k^n}{n} & n > 0 \\ \log(G) & n = 0 \\ \sum_{k=1}^{N_b}\frac{b_k^{-n}}{n} - \sum_{k=1}^{N_d}\frac{d_k^{-n}}{n} & n < 0 \end{cases} \tag{5.48}$$

If $X(z)$ is minimum-phase ($N_b = 0, N_d = 0$), then

- If $x(n)$ is stable, it must be causal.

- $\hat{x}(n)$ is causal.
- $\hat{x}(n)$ is an infinite length sequence, which decays exponentially fast as $n \to \infty$.
- If $x(n)$ is causal, $\hat{x}(n)$ can be calculated using the recursion

$$\hat{x}(n) = \frac{x(n)}{x(0)} - \sum_{k=0}^{n-1} \frac{kx(n-k)}{nx(0)} \hat{x}(k), \quad n > 0 \tag{5.49}$$

If $X(z)$ is maximum-phase ($N_b = 0, N_d = 0$), then

- If $x(n)$ is stable, it must be anti-causal.
- $\hat{x}(n)$ is anti-causal.
- $\hat{x}(n)$ is an infinite length sequence, which decays exponentially fast as $n \to -\infty$.
- If $x(n)$ is anti-causal, $\hat{x}(n)$ can be calculated using the recursion

$$\hat{x}(n) = \frac{x(n)}{x(0)} - \sum_{k=n+1}^{0} \frac{kx(n-k)}{nx(0)} \hat{x}(k), \quad n < 0 \tag{5.50}$$

5.7 Computational Considerations

Suppose we calculate an approximate cepstrum $\hat{x}_p(n)$ by inverse transforming the log DFT:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi kn}{N}} \tag{5.51}$$

$$\hat{X}_p(k) = \log(X(k)) \tag{5.52}$$

$$\hat{x}_p(n) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}(k) e^{\frac{j2\pi kn}{N}} \tag{5.53}$$

$$\hat{x}_p(n) = \sum_{r=-\infty}^{\infty} \hat{x}(n + rN) \tag{5.54}$$

Since $\hat{x}(n)$ is an infinite-length sequence, it is impossible to avoid aliasing.

5.8 Source-Filter Analysis

Suppose that we want to separate the transfer function of a speech signal, $t(n)$, from the periodic source spectrum $q(n) = r(n) * p(n)$:

$$x(n) = t(n) * (r(n) * p(n)) \equiv t(n) * q(n) \tag{5.55}$$

This can be done using the cepstrum, if we know a bit about the signals involved.

5.8.1 Cepstrum of a Periodic Signal

Assume that $q(n)$ has the following form:

$$q(n) = \sum_{r=0}^{\infty} \alpha^r \delta(n - rn_0), \quad |\alpha| < 1 \tag{5.56}$$

$$Q(z) = \frac{1}{1 - \alpha z^{-n_0}} \tag{5.57}$$

$$\log(Q(z)) = -\log(1 - \alpha z^{-n_0}) \tag{5.58}$$

$$\hat{q}(n) = \sum_{r=1}^{\infty} \frac{\alpha^r}{r} \delta(n - rn_0) \tag{5.59}$$

In words, $\hat{q}(n)$ is a decaying impulse train with period n_0. As $\alpha \to 1$, the rate of decay approaches $1/r$.

5.8.2 Cepstrum of the Transfer Function

If the transfer function is a minimum-phase function of the form

$$T(z) = G \frac{\prod_{k=1}^{N_a}(1 - a_k z^{-1})}{\prod_{k=1}^{N_c}(1 - c_k z^{-1})} \tag{5.60}$$

Then

$$\hat{t}(n) = \begin{cases} -\sum_{k=1}^{N_a} \frac{a_k^n}{n} + \sum_{k=1}^{N_c} \frac{c_k^n}{n} & n > 0 \\ \log(G) & n = 0 \\ 0 & n < 0 \end{cases} \tag{5.61}$$

$\hat{t}(n)$ decays at least as quickly as $\frac{r_{max}^n}{n}$, where $r_{max} = \max(|a_k|, |c_k|)$.

5.8.3 "Liftering" to separate source and filter

We have that

$$\hat{q}(n) = 0, \quad n < n_0 \tag{5.62}$$

$$\hat{t}(n) < \frac{r_{max}^{n_0}}{n_0}, \quad n > n_0 \tag{5.63}$$

$$\hat{x}(n) = \hat{t}(n) + \hat{q}(n) \tag{5.64}$$

So, approximately,

$$l(n)\hat{x}(n) \approx \hat{t}(n) \quad \text{if} \quad l(n) = \begin{cases} 1 & 0 \le n < n_0 \\ 0 & n \ge n_0 \end{cases} \tag{5.65}$$

$$l(n)\hat{x}(n) \approx \hat{q}(n) \quad \text{if} \quad l(n) = \begin{cases} 0 & 0 \le n < n_0 \\ 1 & n \ge n_0 \end{cases} \tag{5.66}$$

5.9 Pole-Zero Analysis

Consider the spectrum

$$V(z) = \frac{N(z)}{D(z)} = \frac{\sum_{k=1}^{q} b_k z^{-k}}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{5.67}$$

Suppose we know the signal $v(n)$, and we want to estimate the parameters b_k and a_k.

5.9.1 Estimating the Poles

Notice that, for $n > q$,

$$v(n) = \sum_{k=1}^{p} a_k v(n-k) \tag{5.68}$$

Thus, for $n > q$, $v(n)$ can be modeled using a normal covariance LPC model:

$$e(n) = v(n) - \sum_{k=1}^{p} \alpha_k v(n-k), \quad q+1 \leq n \leq N \tag{5.69}$$

$$E_n = \sum_{n=q+1}^{N} e^2(n) \tag{5.70}$$

5.9.2 Estimating the Zeros

Consider the sequence

$$n\hat{v}(n) \tag{5.71}$$

whose Z transform is

$$-z\frac{d}{dz}\hat{V}(z) = -\frac{z}{V(z)}V'(z) \tag{5.72}$$

$$= -\frac{z}{\frac{N(z)}{D(z)}}\frac{N'(z)D(z) - N(z)D'(z)}{D^2(z)} \tag{5.73}$$

$$= -z\frac{N'(z)D(z) - N(z)D'(z)}{N(z)D(z)} \tag{5.74}$$

Thus, if there is no pole-zero cancellation, the poles of $n\hat{v}(n)$ include both the poles and zeros of $v(n)$. If $V(z)$ is assumed to be minimum-phase, then the poles and zeros can be calculated as follows:

1. Find the zeros of $D(z)$ using LPC analysis of $v(n)$.
2. Find the zeros of $N(z)D(z)$ using LPC analysis of $n\hat{v}(n)$.
3. Compare the two sets, and if all of the zeros of $D(z)$ are also zeros of $N(z)D(z)$ (in other words, if there is not too much error), then the remaining zeros of $N(z)D(z)$ must belong to $N(z)$.

5.10 Log Spectral Distance

5.10.1 Power Spectrum

In much of the speech recognition work this quarter, we will define a power spectrum $S(e^{j\omega})$ which is the Fourier transform of the autocorrelation:

$$S(e^{j\omega}) = \sum_{m=-\infty}^{\infty} R(m)e^{-j\omega m} \tag{5.75}$$

Recall that if $x(n)$ is the windowed speech signal, we can write

$$R(m) = x(m) * x(-m) \tag{5.76}$$

which means that

$$S(e^{j\omega}) = X(e^{j\omega})X(e^{-j\omega}) = |X(e^{j\omega})|^2 \tag{5.77}$$

From now on, whenever the meaning is clear, we will simplify the notation of Fourier transforms as follows

$$S(\omega) \equiv S(e^{j\omega}) \tag{5.78}$$

5.10.2 Log Spectral Distance

Suppose we want to measure the "distortion" between spectra $S_1(\omega)$ and $S_2(\omega)$. One of the most widely studied distortion metrics available is the family of L_p norms, defined as

$$(d_p)^p = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\log S_1(\omega) - \log S_2(\omega)|^p \, d\omega \tag{5.79}$$

As it turns out, this is generally a pretty bad choice, because if the window length is longer than a pitch period, then d_p is dominated by differences in F_0.

5.11 Cepstral Distances

5.11.1 Complex Cepstrum and Power Cepstrum

The complex cepstrum of $x(n)$ is defined to be

$$\hat{x}(m) = \mathcal{Z}^{-1}(\log X(z)) = \mathcal{F}^{-1}(\log X(\omega)) \tag{5.80}$$

The cepstrum used most often in speech recognition is something you might call the "power cepstrum" (though R&J usually just call it the "cepstrum"):

$$\begin{aligned} c(m) &= \mathcal{F}^{-1}\{\log S(\omega)\} && (5.81)\\ &= \mathcal{F}^{-1}\{\log(X(\omega)X(-\omega))\} && (5.82)\\ &= \mathcal{F}^{-1}\{\log X(\omega)\} + \mathcal{F}^{-1}\{\log(X(-\omega))\} && (5.83)\\ &= \hat{x}(m) + \hat{x}(-m) && (5.84)\\ & && (5.85) \end{aligned}$$

Properties:

- Both $\hat{x}(m)$ and $c(m)$ are real numbers.
- $c(m) = c(-m)$.
- If $X(z)$ is minimum-phase, $\hat{x}(m)$ is causal, and $c(m) = \hat{x}(m)$ for $m \geq 0$.
- If $X(z)$ is maximum-phase, $\hat{x}(m)$ is anti-causal, and $c(m) = \hat{x}(m)$ for $m \leq 0$.
- $c(m)$ is an exponentially decaying, bounded sequence.

5.11.2 Cepstral L_2 Norm

Parseval's theorem says that the L_2 spectral norm can be computed in either the frequency domain or the time domain:

$$\begin{aligned} (d_2)^2 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |\log S_1(\omega) - \log S_2(\omega)|^2 \, d\omega && (5.86)\\ &= \sum_{m=-\infty}^{\infty} (c_1(m) - c_2(m))^2 && (5.87)\\ & && (5.88) \end{aligned}$$

5.11.3 LPC Cepstrum

The LPC Power Cepstrum is the inverse transform of $\log|H(z)|^2$:

$$\begin{aligned}
\mathcal{Z}^{-1}\left\{\log\frac{G^2}{|A(z)|^2}\right\} &= \mathcal{Z}^{-1}\left\{\log G^2\right\} - \mathcal{Z}^{-1}\left\{\log|A(z)|^2\right\} && (5.89)\\
&= \mathcal{Z}^{-1}\left\{\log G^2\right\} - \mathcal{Z}^{-1}\left\{\log A(z)\right\} - \mathcal{Z}^{-1}\left\{\log A(z^{-1})\right\} && (5.90)\\
&= 2\log G\delta(m) - (\hat{a}(m) + \hat{a}(-m)) && (5.91)\\
& && (5.92)
\end{aligned}$$

Cepstrum from LPC Roots

$$\begin{aligned}
A(z) &= \prod_{i=1}^{p}(1 - r_i z^{-1}) && (5.93)\\
\log A(z) &= \sum_{i=1}^{p}\log(1 - r_i z^{-1}) && (5.94)\\
\hat{a}(m) &= -\sum_{i=1}^{p}\frac{r_i^m}{m}, \quad m \geq 1 && (5.95)\\
& && (5.96)
\end{aligned}$$

Cepstrum from LPC Coefficients

$$\begin{aligned}
\hat{A}(z) &= \log A(z) && (5.97)\\
-z\frac{d\hat{A}(z)}{dz} &= -z\frac{1}{A(z)}\frac{dA(z)}{dz} && (5.98)\\
n\hat{a}(n) * a(n) &= na(n) && (5.99)\\
& && (5.100)
\end{aligned}$$

Since we already know that $a(n)$ and $\hat{a}(n)$ are causal, the convolution in the last line can be converted into a recursive formula for $\hat{a}(n)$.

5.11.4 Cepstral Representation of Spectral Energy, Slope, and Finer Detail

The cepstrum is more interesting if you understand the "meaning" of the different coefficients. For example, $c(0)$ represents the average energy:

$$c(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log S(\omega)d\omega \tag{5.101}$$

$c(1)$ represents the spectral tilt:

$$\begin{aligned}
c(1) &= \hat{x}(1) + \hat{x}(-1) && (5.102)\\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\log X(\omega)(e^{j\omega} + e^{-j\omega})d\omega && (5.103)\\
&= \frac{1}{\pi}\int_{-\pi}^{\pi}\cos\omega\log X(\omega)d\omega && (5.104)
\end{aligned}$$

Higher cepstral coefficients represent finer-grain details of the spectral shape. For example, the coefficient $-c(2)$ represents the degree to which spectral energy is clustered around $\omega = -\pi/2$:

$$c(2) = \frac{1}{\pi}\int_{-\pi}^{\pi} \cos(2\omega)\log X(\omega)d\omega \tag{5.105}$$

5.12 Cepstral Liftering

5.12.1 Window in Time = Convolve in Frequency

Remember that windowing in time equals convolution in frequency. Suppose that $w(m)$ is a windowing sequence with spectrum $W(\omega)$; then

$$\mathcal{F}\{c(m)w(m)\} = \log S(\omega) * W(\omega) \tag{5.106}$$

If $W(\omega)$ has a low-pass filter shape, then we can smooth the log spectrum using the following procedure:

1. Convert from FFT or LPC to Cepstrum.
2. Window the cepstrum.
3. Convert back to FFT or LPC.

5.12.2 Weighted/Liftered Cepstral Distances

Remember that the L_2 distance between $S_1(\omega)$ and $S_2(\omega)$ is

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |\log S_1(\omega) - \log S_2(\omega)|^2 d\omega = \sum_{m=-\infty}^{\infty} (c_1(m) - c_2(m))^2 \tag{5.107}$$

Adding up an infinite number of cepstral samples is not practical. In practice, we usually calculate the liftered or weighted cepstral distance,

$$\begin{aligned} d_{cW}^2 &= \sum_{m=1}^{L} w^2(m)(c_1(m) - c_2(m))^2 && (5.108)\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |(\log S_1(\omega) * W(\omega)) - (\log S_2(\omega) * W(\omega))|^2 d\omega && (5.109)\\ & && (5.110)\end{aligned}$$

5.12.3 Symmetric Equivalent Window

Speech recognition often makes use of a delayed causal $w(m)$, that is, a window which is non-zero only for m strictly greater than 0. If $w(m)$ is a delayed causal window, the amount of smoothing given by cepstral liftering is slightly different from that estimated by the above formula. Suppose we define the even part of $w(m)$ to be $\tilde{w}(m)$:

$$\tilde{W}(\omega) = \Re\{W(\omega)\}, \qquad \tilde{w}(m) = \begin{cases} w(m)/2 & m > 0 \\ w(-m)/2 & m < 0 \end{cases} \tag{5.111}$$

If $w(m)$ is delayed causal, we can take advantage of the even symmetry of $c_1(m)$ to express d_{cW}^2 as a two-sided sum:

$$\begin{aligned} d_{cW}^2 &= 2\sum_{m=-L}^{L} \tilde{w}^2(m)(c_1(m) - c_2(m))^2 && (5.112)\\ &= \frac{1}{\pi}\int_{-\pi}^{\pi} |(\log S_1(\omega) * \tilde{W}(\omega)) - (\log S_2(\omega) * \tilde{W}(\omega))|^2 d\omega && (5.113)\\ & && (5.114)\end{aligned}$$

So a weighted cepstral distance is similar to the following L_2 norm:

- Smooth $\log S_1(\omega)$ and $\log S_2(\omega)$ using the smoothing spectrum $\tilde{W}(\omega) = \Re\{W(\omega)\}$.
- Calculate the L_2 distortion measure between the two smoothed log spectra.

5.12.4 Example: Rectangular Window

If $w(m)$ is a causal rectangular window covering samples 1 through L, then $\tilde{w}(m)$ is an even window of length $2L+1$:

$$w(m) = \begin{cases} 1 & m = 1, \ldots, L \\ 0 & \text{else} \end{cases}, \quad \tilde{w}(m) = \begin{cases} 1/2 & m = -L, \ldots, -1, 1, \ldots, L \\ 0 & \text{else} \end{cases} \tag{5.115}$$

$\tilde{w}(m)$ is just a rectangular window of length $2L+1$, minus the impulse $\delta(n)$. The spectrum is therefore:

$$\tilde{W}(\omega) = \frac{\sin\frac{\omega(2L+1)}{2}}{2\sin\frac{\omega}{2}} - \frac{1}{2} \approx \frac{\sin\frac{\omega(2L+1)}{2}}{2\sin\frac{\omega}{2}} \tag{5.116}$$

$$\tag{5.117}$$

where the approximation holds for large L.

5.13 Exercises

1. Consider the sequence
$$x[n] = \delta[n] - a\delta[n-1] \tag{5.118}$$
where $|a| < 1$. Suppose that we wish to approximate the complex cepstrum $\hat{x}[n]$ from samples of the logarithm of the Fourier transform:
$$\hat{x}_p[n] = \frac{1}{N}\sum_{k=0}^{N-1} \log\left(X(e^{j\frac{2\pi}{N}kn})\right) \tag{5.119}$$
Is it possible to choose N large enough such that $\hat{x}_p[n] = \hat{x}[n]$, without aliasing? If so, what is the minimum value of N? If not, what is the minimum value of N (give or take a few samples) such that
$$|\hat{x}_p[n] - \hat{x}[n]| < \left|\frac{\hat{x}[n]}{100}\right| \tag{5.120}$$
Note: You may find the following formula to be useful:
$$\log(1-x) = -\sum_{n=1}^{\infty}\frac{x^n}{n} \quad \text{if } |x| < 1 \tag{5.121}$$

2. Suppose that homomorphic analysis yields the following estimate of the vocal tract transfer function:
$$H(z) = \frac{G}{1 - \sum_{k=1}^{2q} a_k z^{-k}} = G\prod_{k=1}^{q}\frac{1}{(1-b_k z^{-1})(1-b_k^* z^{-1})} \tag{5.122}$$
with pole locations $b_k = r_k e^{j\theta_k}$ and $b_k^* = r_k e^{-j\theta_k}$ which are located inside the unit circle. If the sampling rate is F_s, then the formant frequencies F_k and bandwidths B_k can be estimated by:
$$\hat{F}_k = \frac{F_s\theta_k}{2\pi} \tag{5.123}$$
$$\hat{B}_k = -\frac{F_s}{\pi}\log(r_k) \tag{5.124}$$
Suppose that we suspect that all of the bandwidth estimates $\hat{B}_k$ are too large. Show that the estimated formant bandwidths are reduced, without changing the estimated formant frequencies, if we replace $H(z)$ by the following transformed spectrum:
$$\tilde{H}(z) = H\left(\frac{z}{\alpha}\right) = G\prod_{k=1}^{q}\frac{1}{(1-b_k(z/\alpha)^{-1})(1-b_k^*(z/\alpha)^{-1})} \tag{5.125}$$
where α is real and greater than unity and $|\alpha b_k| < 1$.

3. Suppose $h(n)$ in part (a) consists of a single complex pole pair of the form
$$H(z) = \frac{1}{(1-re^{j\theta}z^{-1})(1-re^{-j\theta}z^{-1})} \tag{5.126}$$
where r and θ are both real. Find expressions for the complex cepstra associated with $H(z)$ and $\tilde{H}(z)$ in this case. Find expressions for the real cepstra, and plot the real cepstra as functions of time.

4. Compute the following spectra for three different vowel segments segments, and plot the log-magnitude spectra (in dB) for frequencies between 0 and 4000Hz. You should turn in code, equations, or some combination of both which will make it clear how each spectrum was computed.

(a) Power spectrum $S(\omega)$.

(b) $S(\omega)$, smoothed using cepstral lifter $\hat{w}_1(n)$:

$$\hat{w}_1(n) = u(n-1) - u(n-L-1) \tag{5.127}$$

Choose L so that the window length is 1.5ms. What is the cutoff frequency of the magnitude lifter spectrum, $\tilde{W}_1(\omega) = |\hat{W}_1(\omega)|$?

(c) $S(\omega)$, smoothed using cepstral lifter $\hat{w}_2(n)$:

$$\hat{w}_2(n) = \hat{w}_1(n)\left(1 + \frac{L}{2}\sin(\frac{n\pi}{L})\right) \tag{5.128}$$

(d) LPC transfer function $H(\omega) = G/A(\omega)$.

(e) $H(\omega)$, smoothed using cepstral lifter $\hat{w}_1(n)$.

(f) $H(\omega)$, smoothed using cepstral lifter $\hat{w}_2(n)$.

(g) Line spectra $1/P(\omega)$ and $1/Q(\omega)$, truncated at reasonable maximum and minimum values.

5. The three vowel signals you analyzed in problem 4 are different — but how different are they? Calculate the difference between the two vowels using the following spectral distortion metrics. Turn in code and/or equations showing how each distortion metric was computed.

(a) L_2 spectral norm, calculated using log-FFT spectra.

(b) Truncated cepstral distance $d_c^2(L)$, where L is chosen as in problem 1(b).

(c) Liftered cepstral distance $d_{cW}^2(L)$, using the lifter $\hat{w}_2(n)$ defined in problem 1(c).

(d) Likelihood-ratio distortions,

$$d_{LR}\left(\frac{1}{|A_1|^2}, \frac{1}{|A_2|^2}\right) \quad \text{and} \quad d_{LR}\left(\frac{1}{|A_2|^2}, \frac{1}{|A_1|^2}\right) \tag{5.129}$$

where the subscripts 1 and 2 represent the first and second vowel.

(e) Itakura-Saito distortions,

$$d_{IS}\left(\frac{G_1^2}{|A_1|^2}, \frac{G_2^2}{|A_2|^2}\right) \quad \text{and} \quad d_{IS}\left(\frac{G_2^2}{|A_2|^2}, \frac{G_1^2}{|A_1|^2}\right) \tag{5.130}$$

(f) LSF Euclidean distance.