Homomorphic Processing of Speech
ES 157/257: Speech and Audio Processing

Patrick J. Wolfe

Teaching Fellow: Prabahan Basu

Division of Engineering and Applied Sciences
Department of Statistics
Harvard University

Lecture 11

Based on the book and teaching materials of Tom Quatieri. Copyright © by Pearson Education.
Outline
Today: Chapter 6 pp. 276-296

1. Review from Last Time (Chapter 6 pp. 253-269)

2. Short-time Homomorphic Analysis
   - Periodic Sequences
   - Homomorphic Filtering of Speech

3. Analysis / Synthesis Structures
   - General Framework
   - Effect of Phase

4. Contrasting Linear Prediction and Homomorphic Filtering
Last Time
Homomorphic Signal Processing

- Introduction to Homomorphic Processing
  - Generalized superposition
  - Properties of the complex logarithm
  - Phase unwrapping

- Properties of the Complex Cepstrum
  - Rational z-transforms and periodic sequences
  - Minimum / maximum phase sequences
  - A useful representation for deconvolution

- Administrivia
  - Homework #3 posted online due Thursday 03/12
  - Midterm #1 in class Tuesday 03/17: Chapters 1-6
We have studied the application of homomorphic processing to the signal of the form \( x[n] = p[n] \ast h[n] \) where \( p[n] \) and \( h[n] \) are sequences representing the source and the filter.

In practice, the speech waveform is multiplied by a window \( w[n] \) yielding: \( s[n] = w[n](p[n] \ast h[n]) \).

Ideally, we want \( s[n] \) to be close to the convolutional model: \( (w[n]p[n]) \ast h[n] \). In this case, the complex cepstrum \( \hat{s}[n] \) is approximately: \( \hat{s}[n] \approx \hat{p}[n] + \hat{h}[n] \) where \( \hat{p}[n] \) and \( \hat{h}[n] \) are the complex cepstra of \( (w[n]p[n]) \) and \( h[n] \).

If the pitch period \( P \) were large enough so that the sequences \( h[n - kP] \) don’t overlap and \( w[n] \) were constant over each one, this approximation would be exact.
In general it can be shown that complex cepstrum of a windowed sequence is given exactly by:

\[ \hat{s}[n] = \hat{p}[n] + D[n] \sum_{k=-\infty}^{+\infty} \hat{h}[n - kP] \]

- \( D[n] \) is a weighting function concentrated at \( n = 0 \) and dependent on the window \( w[n] \)
- We can impose conditions on \( w[n] \) so that the approximation \( \hat{s}[n] \approx \hat{p}[n] + \hat{h}[n] \) is reasonable.
- For Hamming and Hanning windows these conditions reduce to:
  - Length of \( w[n] \) should be on the order of 2-3 pitch periods
  - The window should be time-aligned with \( h[n] \) to avoid any phase contribution
The complex cepstrum of a windowed periodic sequence contains the impulse response of the vocal tract filter $\hat{h}[n]$ repeated at the pitch period rate.

The impulse response replicas are weighted by a distortion function $D[n] = w[n]$ in this special case.
Sensitivity of Phase Estimate to Analysis Window
Effects of time-misalignment and window length

a. Unwrapped phase of artificial vocal tract impulse response
b. Unwrapped phase of periodic waveform with $sinc^2$ window time-aligned to the system impulse response
c. Same as (b) with window displacement
d. Same as (b) but with Hamming window 2 pitch periods in length
e. Same as (d) with Hamming window 3.9 pitch periods in length
Homomorphic Filtering Example 1/3

- We will consider the, by now familiar, model for speech in which the glottal flow is represented by two maximum-phase poles and a mixed-phase vocal tract impulse response.

- A goal in the analysis of voiced speech is to separate the impulse response of the vocal tract $h[n]$ from the source $p[n]$, using a short-time segment of speech that is extracted by an appropriate window.

- We will consider homomorphic filtering of a speech waveform from a female speaker with an average pitch period of about 5ms.

- A 15ms Hamming window aligned roughly with $h[n]$ is to be used.
Homomorphic Filtering Example 2/3

a. Time-domain waveform $s[n]$ and aligned window
b. Complex cepstrum of windowed speech and low-pass lifter
c. Log-magnitude spectrum of $s[n]$ (thin) and of the impulse response estimate (thick)
d. Impulse response estimate from low-pass quefrency liftering
e. Spectral unwrapped phase of $s[n]$ (thin) and of the impulse response estimate (thick)
f. Estimate of windowed impulse train from high-pass quefrency liftering
Homomorphic Filtering Example 3/3

a. Deconvolved maximum-phase component of vocal tract impulse response
b. Deconvolved minimum-phase component
c. Log-magnitude spectra of the maximum-phase (dashed) and minimum-phase (solid) components.
d. Convolution of the maximum- and minimum-phase components
Some Properties of Homomorphic Filtering

- We have looked at homomorphic deconvolution through the lens of spectral smoothing.
- Low-quefrency liftering separates slowly-varying and rapidly varying components of the spectral log magnitude and phase.
- Harmonics are not favored explicitly as during linear prediction analysis (recall our discussion of spectral distortion measures).
- Low pass lifter is centered about the origin and has a rough extent of a pitch period, resulting in greater smoothing with increasing pitch (decreasing pitch period).
- Excessive smoothing leads to broadening of the formant peaks resulting in a muffled quality of sound (see Hw 3: #6.6).
Homomorphic Filtering of Unvoiced Speech

A key difference between the homomorphic filtering of voiced and unvoiced speech is that in the latter the source and filter components overlap in the low-quefrency region.

White noise is supported over the entire quefrency axis.

Nonetheless, we can still think of homomorphic filtering as smoothing of the fluctuations in $\log[X(\omega)]$.

On the other hand, the estimation of unwrapped phase is sensitive to small perturbations in spectral nulls, leading to poor estimates of phase.

Since the phase of the impulse response is not considered perceptually significant, we may still apply this method.
Homomorphic Analysis/Synthesis 1/2

- In speech analysis underlying parameters of the speech model are estimated. We have already seen how to do this using methods of linear prediction.
- We have also seen that a non-parametric estimate of the vocal tract impulse response may be obtained from the complex cepstrum.
- Cepstral methods may be used to estimate pitch and voicing as well.
  - Presence of a peak in the high-quefrecy cepstrum indicates voicing.
  - If a peak is present, then its location determines pitch.
- Given these estimates we may re-synthesize the original waveform as we now describe...
Homomophic Analysis/Synthesis 2/2

a. Analysis: $x[n]$ is the original waveform, $l[n]$ is the cepstral lifter, and $\hat{h}[n]$ is the estimated cepstral representation of the vocal tract impulse response

b. Synthesis: Pitch and voicing estimates used to construct $p[n]$ the excitation waveform
Phase in Homomorphic Synthesis 1/2

Four possibilities

- **Minimum-Phase Synthesis**
  - Uses only right side of the liftered complex cepstrum
  - Can be performed with real cepstrum (i.e. without phase information)

- **Maximum-Phase Synthesis**
  - Uses only left side of the liftered complex cepstrum

- **Zero-Phase Synthesis**
  - Uses the right or left side of the liftered complex cepstrum, subsequently symmetrized about 0
  - Results in phase curve that is 0 over all frequencies

- **Mixed-Phase Synthesis**
  - Uses the entire liftered complex cepstrum
  - Care must be taken with alignment issues
Phase in Homomorphic Synthesis 2/2

Perceptual Effects: Play Me

- **a.** Original waveform and window.
- **b.** Mixed-phase synthesis. Reduces the “buzziness” of minimum-phase reconstruction.
- **c.** Zero-phase synthesis. Sounds very “muffled”.
- **d.** Minimum-phase synthesis. More natural than either maximum- or zero-phase synthesis, but has “buzzy” quality.
Properties of Linear Prediction

- Parametric deconvolution technique
- Based on an all-pole representation (though zeros can be represented by a large number of poles)
- Sharp smooth resonances corresponding to an all-pole model
- Constrained (through the autocorrelation method) to provide a minimum-phase estimate of the vocal tract impulse response
- Though “crisp”, the resultant sound is sometimes characterized as “buzzy” or “mechanical”
Properties of Homomorphic Filtering

- Non-parametric (transform-based) deconvolution technique
- Both poles and zeros can be represented
- Wider spurious resonances consistent with the spectral smoothing interpretation of cepstral liftering
- Can provide a minimum-phase or a mixed-phase estimate of the vocal tract impulse response by using the complex cepstrum
- Though more “natural”, than its counterpart in linear prediction, the resultant sound is sometimes characterized as “muffled”
Shortcomings of Both Methods

- Distortion with increasing pitch
  - Aliasing of vocal tract impulse response occurs in the cepstrum and autocorrelation function
  - Can perhaps be better controlled in the complex cepstrum by pitch-adaptive liftering methods

- Time-domain windowing alters the assumed speech model
  - Autocorrelation method predicts nonzero values outside window interval
  - Convolutional model distorted in homomorphic filtering

- Model selection
  - Selecting the number of poles in representation
  - Selecting the length of the cepstral lifter