Lecture notes 10 – Monday 7/10

Summary of last lecture
Acoustical model of speech production
- glottal excitation: random noise (voiceless) or impulse train & glottal pulse shape (voiced)
- vocal tract filter, concatenation of lossless tubes, modeled as an all-pole IIR filter
Linear predictive coding (LPC)
- predict current sample from past samples
- order of LPC, depends on signal, usually $1/kHz \times 2–4$ to model glottal effects and such.

The cepstrum
The word *cepstrum* is a play on the word spectrum, created by reversing the first syllable of the word. The term was coined by the inventors —Bogert, Healy and Tukey— in their paper “The Quefrency analysis of time series for echoes”. Originally designed to analyze echoes, today the cepstrum has taken a dominant place in speech recognition. In today’s lecture, we will study the cepstrum and its properties.

Homomorphic processing
The cepstrum is one type of ‘homomorphic’ transformation. A homomorphic transformation

$$\hat{x}[n] = D\{x[n]\}$$

is a transformation that converts a convolution,

$$x[n] = e[n] * h[n],$$

into a sum

$$\hat{x}[n] = D\{x[n]\}
= D\{e[n] * h[n]\}
= D\{e[n]\} + D\{h[n]\}
= \hat{e}[n] + \hat{h}[n]$$

We will see that the cepstrum, as a homomorphic transformation, allows us to separate the glottal excitation $e[n]$ from the vocal tract filter $h[n]$. This separation is possible because we can find a value $N$ such that the cepstrum of the filter $\hat{h}[n] \approx 0$ for $n \geq N$ and the cepstrum of the excitation $\hat{e}[n] \approx 0$ for $n < N$. With this assumption, we can approximately recover both sequences $e[n]$ and $h[n]$ from the cepstrum $\hat{x}[n]$ by ‘homomorphic filtering’. This is illustrated by the figures below.
Top row: a speech signal and its spectrum
Left: cepstrum of the speech signal

Left: selecting the glottal excitation cepstral coefficients

Left: selecting the vocal tract cepstral coefficients

Bottom rows: excitation signal and vocal tract impulse response and their spectra.
Definition of the cepstrum
The cepstrum transforms a convolution to an addition in three steps. Suppose \( x[n] = e[n] * h[n] \):

1. Take the DTFT of (a windowed version of) \( x[n] \). This results in the spectrum \( X(\omega) \) of the signal \( x[n] \), and transforms the convolution (in time) into a multiplication (in frequency):
   \[
   X(\omega) = E(\omega)H(\omega)
   \]

2. Take the log of \( X(\omega) \) (usually natural log, base \( e \)); this results in the log-spectrum of \( x[n] \), and transforms the multiplication into an addition:
   \[
   \log X(\omega) = \log E(\omega)H(\omega) = \log E(\omega) + \log H(\omega)
   \]

3. Take the inverse DTFT of the result.

Recall that the spectrum \( X(\omega) \) is complex. There are two version of the cepstrum defined: the real cepstrum and the complex cepstrum. In the real cepstrum, the real logarithm is taken of the magnitude of the spectrum, \( |X(\omega)| \), in step 2. In the complex cepstrum, the complex logarithm is taken of the spectrum in step 2.

The details of the complex logarithm are not important to us. But what is important is that if we want to reconstruct a signal from a (modified) cepstrum, we must use the complex cepstrum. In the real cepstrum, the phase information of \( X(\omega) \) is lost, and reconstruction is only possible under a “minimal phase” assumption that usually does not give the desired reconstruction result. In speech recognition, reconstruction is not required and therefore the simpler real cepstrum is used.

Cepstral features of voiced speech
For the example voiced segment of speech shown in the figure below,

the cepstrum is shown in the following figure:
In the cepstrum of this voiced segment of speech, we can identify the following features:
  1. High energy content around low quefrencies
  2. Decaying towards higher quefrencies
  3. Distinct peak around fundamental quefrency

The first feature is caused by the resonances in the vocal tract. The second feature can be attributed to a property of the cepstrum of an impulse train. It can be shown that the cepstrum of an impulse train is a decaying function. The third feature is caused by the echo detection capabilities of the cepstrum. The repeated glottal periods are detected as echoes.

**Separating the glottal excitation from the vocal tract**

To separate the glottal excitation from the vocal tract response, we must choose a value $N$ such that the vocal tract response has most of its energy in the quefrencies less than $N$, and the glottal excitation has most of its energy in the quefrencies above $N$. Speech recognition researchers have shown empirically that a value of $N$ between 12 and 20 is a good choice to separate glottal excitation from the vocal tract response, depending on the sampling rate and whether ‘frequency warping’ is done. An example of this separation is shown in the figures on page 2, where $N = 15$.

**Frequency warping: the Mel-frequency cepstrum**

The Mel-frequency cepstrum differs from the real cepstrum in that it uses a non-linear frequency scale, which approximates the behavior of the auditory system.

The Mel-frequency cepstrum works as follows:
1. Given a (windowed) input signal $x[n]$, compute its $N$-point DTFT $X[k]$, for $k = 0, \ldots, N-1$
2. Construct $M$ filterbank outputs from the $N$-point DTFT by multiplying the $N$-point DTFT with triangular filters, as shown below
3. Compute the log of the energy at the output of each filter of the filterbank, as

$$S[m] = \log \left[ \sum_{k=0}^{N-1} |X[k]|^2 H_m[k] \right]$$
4. Take a modified inverse DTFT of the result

\[ \hat{x}[n] = \sum_{m=0}^{M-1} S[m] \cos(\pi n (m - \frac{1}{2}) / M) \]

The Mel-frequency cepstrum is used extensively as a feature vector for speech recognition systems. For speech recognition, only the first 13 cepstrum coefficients are used, and are referred to as the Mel-frequency cepstral coefficients (MFCC).

Note that the Mel-frequency cepstrum is no longer a homomorphic transformation, due to the computation of the weighted log-energy in step 3. In practice, however, the MFCC representation is approximately homomorphic for filters that have a smooth frequency response. The advantage of the MFCC representation is that the filter energies are more robust to noise and spectral estimation errors.