Lecture 5, The Vocoder, Part 2: Unvoiced Sounds

Mark Hasegawa-Johnson

ECE 537: Speech Processing Fundamentals

Outline

The Vocoder Block Diagram

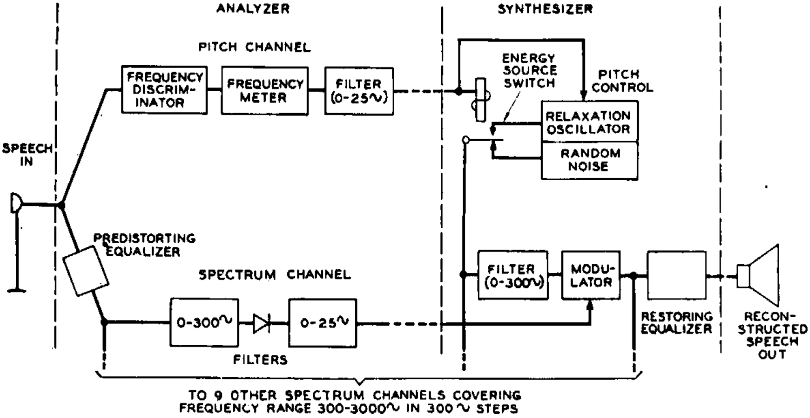


Fig. 2. Schematic arrangement of the Vocoder.

Vocoder Signals Summary

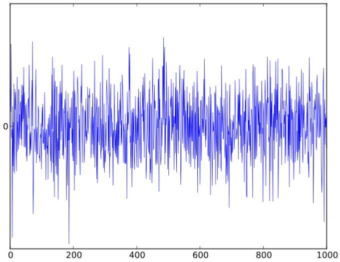
- What is the spectrum of a "relaxation oscillator"?
  - To answer this question, we need to learn about the Discrete-Time Fourier Series (DTFS).
  - What happens when you bandpass filter it?
  - What happens when you adjust its level?
- What is the spectrum of a "random noise"?
  - To answer this question, we need to learn about autocorrelation and the power spectrum.
  - What happens when you bandpass filter it?
  - What happens when you adjust its level?

Outline

Random Signals

Let's start out with a zero-mean random signal, $x[n]$.

- Random signal: each $x[n]$ is a random number.
- Zero mean: $E[x[n]] = 0$ (for all n).



←Listen→

Properties a Random Signal Might Have

- A random signal is **zero-mean** if $E[x[n]] = 0$ for all n.
- A random signal is **unit-power** if $E\left[|x[n]|^2\right] = 1$, regardless of n.
- A random signal is **white noise**, a.k.a. **uncorrelated** if $E\left[x[n]x[m]^*\right] = 0$ for all $n \neq m$.

Wide-Sense Stationary Signals

A random signal is called "wide-sense stationary (WSS)" if its mean, variance, and covariance are independent of n:

- $E[x[n]] = \mu_x$, regardless of n.
- $E\left[|x[n] - \mu_x|^2\right] = \sigma_x^2$, regardless of n.
- The **autocorrelation** and **autocovariance** of a WSS signal are defined to be

$$R_{xx}[m] = E\left[x[n]x^*[n-m]\right]$$
$$K_{xx}[m] = E\left[(x[n] - \mu_x)\left(x[n-m] - \mu_x\right)^*\right],$$

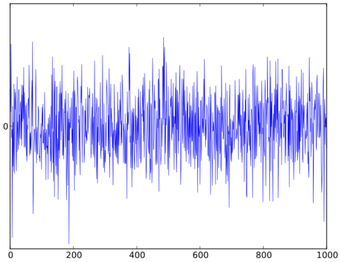
regardless of n.

Example: Zero-Mean, Unit-Variance White Noise

We'll often use zero-mean, unit-variance white noise as a building block:

- $E[x[n]] = 0$
- $R_{xx}[m] = \delta[m] = \begin{cases} 1 & m = 0 \\ 0 & m \neq 0 \end{cases}$

Note: if we add one more assumption ($x[n]$ is Gaussian), then it's also true that $x[n]$ are i.i.d.



←Listen→

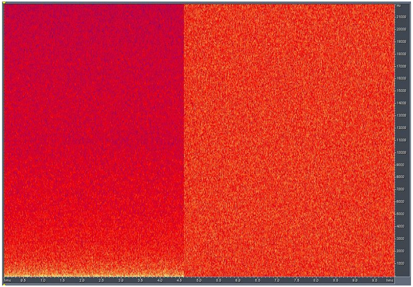
Outline

Fourier Transform of a Random Signal is a Random Vector

The Fourier Transform of a random signal is a random vector.

$$X(\omega) = \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega n}$$

- $e^{-j\omega n}$ is a constant
- $x[n]$ is random
- $X(\omega)$ is the weighted sum of the random variables $x[n]$



Spectrogram of pink noise (left) and white noise (right), shown with linear frequency axis (vertical).

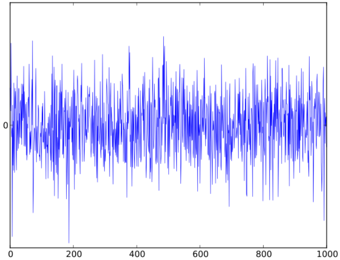
Zero-Mean Random Signal ↔ Zero-Mean Random Vector

The Fourier Transform of a zero-mean random signal is a zero-mean random vector.

$$E\left[X(\omega)\right] = E\left[\sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n}\right]$$

$$= \sum_{n=-\infty}^{\infty} E\left[x[n]\right] e^{-j\omega n}$$

$$= 0$$



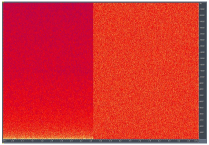
Variance of the Fourier Transform is Interesting

The magnitude-squared Fourier Transform is also a random variable, but its expected value is not zero.

$$E\left[|X(\omega)|^2\right] = E\left[\left(\sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n}\right)\left(\sum_{m=-\infty}^{\infty} x[m]e^{-j\omega m}\right)^*\right]$$

$$= \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} E\left[x[n]x^*[m]\right] e^{-j\omega(n-m)}$$

$$= \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} R_{xx}[n-m]e^{-j\omega(n-m)}$$



Spectrogram of pink noise (left) and white noise (right), shown with linear frequency axis (vertical).

Power Spectrum = Time-Normalized Expected Value of the Variance of the Fourier Transform

For most signals, the formula on the previous slide gives $E\left[|X(\omega)|^2\right] \rightarrow \infty$. To make it easier to work with, Norbert Wiener defined the power spectrum to be the time-normalized expected value of the magnitude squared Fourier transform:

$$R_{xx}(\omega) = \lim_{N\rightarrow\infty} \frac{1}{N} E\left[\left|\sum_{n=-\left(\frac{N-1}{2}\right)}^{\left(\frac{N-1}{2}\right)} x[n]e^{-j\omega n}\right|^2\right]$$

Short-Time Power Spectrum

Most practical signals are not infinite length. Instead, we usually want to just compute the Fourier transform over N samples, say, $0 \leq n \leq N-1$. In this case we can define the short-time power spectrum to be

$$R_{xx}(\omega) = \frac{1}{N} E\left[\left| \sum_{n=0}^{N-1} x[n] e^{-j\omega n} \right|^2 \right]$$

Example: Power Spectrum of White Noise

For example, consider white noise: $E[x[n]x[m]] = 0$ unless $n = m$. In this case,

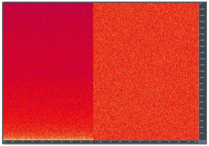
$$\begin{aligned}
R_{xx}(\omega) &= \frac{1}{N} E\left[|X(\omega)|^2\right] \\
&= \frac{1}{N} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} E\left[x[n]x^*[m]\right] e^{-j\omega(n-m)} \\
&= \frac{1}{N} \sum_{n=0}^{N-1} E\left[|x[n]|^2\right] \\
&= E\left[|x[n]|^2\right]
\end{aligned}$$

Example: Power Spectrum of White Noise

For example, consider white noise: $E[x[n]x[m]] = 0$ unless $n = m$. In this case,

$$R_{xx}(\omega) = E\left[|x[n]|^2\right]$$

This is why we call it white noise: its power spectrum is a constant, $R_{xx}(\omega) = E\left[|x[n]|^2\right]$, at every frequency. For example, for zero-mean unit-variance white noise, $R_{xx}(\omega) = E\left[|x[n]|^2\right] = \sigma_x^2 = 1$.



Spectrogram of pink noise (left) and white noise (right), shown with linear frequency axis (vertical).

Outline

Power Spectrum of a WSS Signal

Remember that WSS signals have an autocorrelation function that doesn't depend on n:

$$R_{xx}[m] = E\left[x[n]x^*[n-m]\right]$$

For a WSS signal, it's possible to use a dramatic shortcut to compute the power spectrum:

$$\begin{aligned}
R_{xx}(\omega) &= \frac{1}{N}E\left[\left|\sum_n x[n]e^{-j\omega n}\right|^2\right] \\
&= \frac{1}{N}E\left[\left(\sum_n x[n]e^{-j\omega n}\right)\left(\sum_{n-m} x[n-m]e^{-j\omega(n-m)}\right)^*\right] \\
&= \frac{1}{N}\sum_n \sum_m E\left[x[n]x^*[n-m]\right]e^{-j\omega m} \\
&= \sum_m R_{xx}[m]e^{-j\omega m}
\end{aligned}$$

Power Spectrum of a WSS Signal

Let me just repeat that, since it's the most important formula today.

$$R_{xx}(\omega) = \sum_m R_{xx}[m]e^{-j\omega m}$$

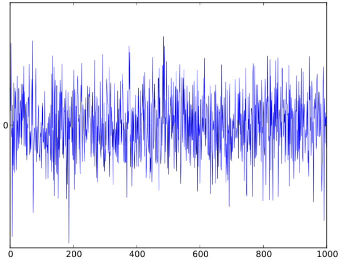
Example: Zero-Mean, Unit-Variance White Noise

For example, consider white noise:

$$R_{xx}[m] = \delta[m] = \begin{cases} 1 & m = 0 \\ 0 & m \neq 0 \end{cases}$$

So its power spectrum is

$$R_{xx}(\omega) = \mathcal{F}\left\{\delta[m]\right\} = 1$$



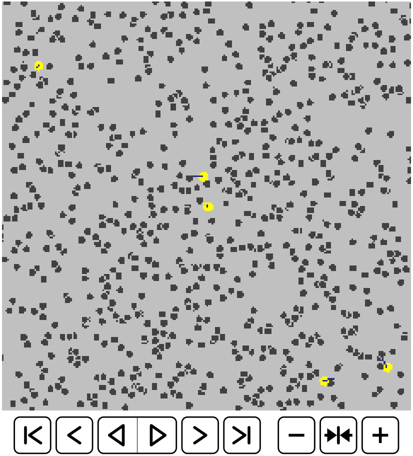
https://upload.wikimedia.org/wikipedia/commons/9/98/White-noise-sound-20sec-mono-44100Hz.ogg←Listen→

Example: Brownian Motion

Brownian motion, as shown in the video, is motion with independent random increments, i.e., if $x[n]$ is the position and $v[n]$ is an independent increment, then

$$x[n] = ax[n-1] + bv[n]$$

Natural Brownian motion uses $a = b = 1$, but if we want a WSS signal, we need to use $b^2 = 1 - a^2$.



Example: Brownian Motion

Suppose we assume $v[n]$ is zero-mean unit-variance white noise, $b^2 = 1 - a^2$, and $x[n] = ax[n-1] + bv[n]$, so that

$$E\left[x[n]x[n-1]\right] = E\left[\left(ax[n-1] + bv[n]\right)x[n-1]\right] = a$$

$$E\left[x[n]x[n-2]\right] = E\left[\left(a^2x[n-2] + abv[n-1] + bv[n]\right)x[n-2]\right] = a^2$$

$$\vdots$$

$$R_{xx}[m] = a^{|m|}$$

The power spectrum is

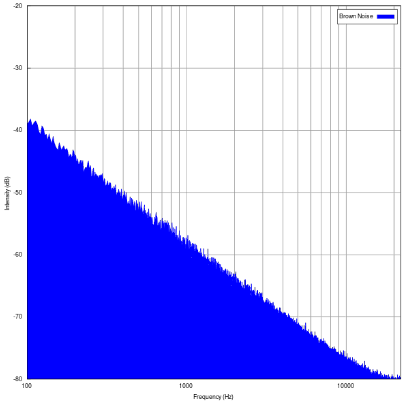
$$R_{xx}(\omega) = \mathcal{F}\left\{a^{|m|}\right\} = \frac{b^2}{\left|1 - ae^{-j\omega}\right|^2} = \frac{1}{\mathcal{O}\left\{\omega^2\right\}}$$

- Impulse trains and white noise both have flat spectra.

$$|X_k|^2 = R_{xx}(\omega) = 1$$

- Square waves and Brownian motion both have Brownian spectra.

$$|X_k|^2 = R_{xx}(\omega) = \frac{1}{\mathcal{O}\left\{\omega^2\right\}}$$



CC-SA 3.0, `https://commons.wikimedia.org/wiki/File:Brown_noise_spectrum.svg`

Short-Time Autocorrelation

Autocorrelation isn't a function of n, so it doesn't hurt if we average it over many samples of n:

$$R_{xx}[m] = \frac{1}{N}\sum_{n=0}^{N-1} R_{xx}[m] = \frac{1}{N} E\left[\sum_{n=0}^{N-1} x[n]x^*[n-m]\right]$$

$$= \frac{1}{N} E\left[x[m] * x^*[-m]\right]$$

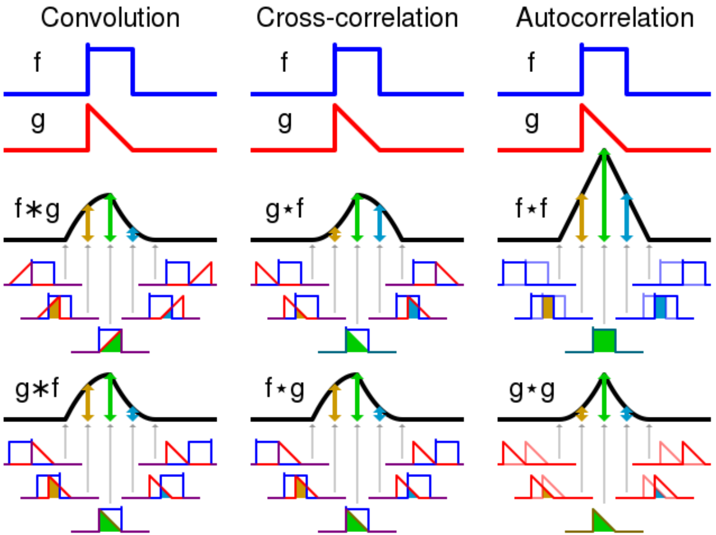
- **Convolution:** Flip, shift, multiply, and add:

$$x[m] * h[m] = \sum_n x[n]h[m-n]$$

- **Correlation:** DON'T flip. Just shift, multiply and add:

$$x[m] * h^*[-m] = \sum_n x[n]h^*[n-m]$$

Correlation: Shift, Multiply and Add



Outline

Facts about convolution

Convolution is commutative:

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

It is also associative:

$$g[n] * (x[n] * h[n]) = (g[n] * x[n]) * h[n]$$

The Vocoder Block Diagram

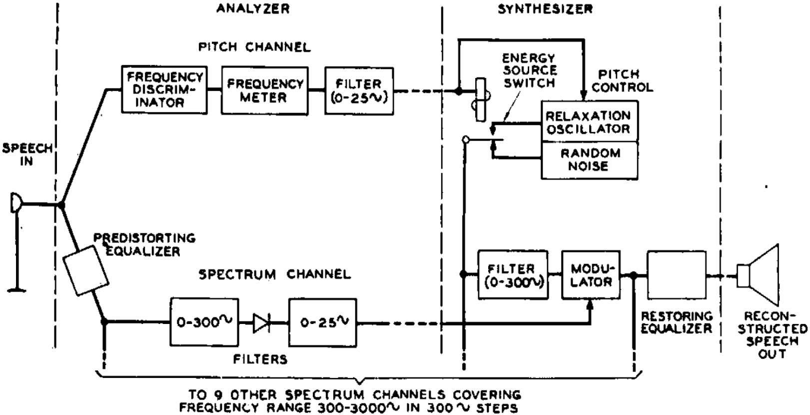


Fig. 2. Schematic arrangement of the Vocoder.

Spectrum of a Bandpass-Filtered Noise

Suppose

$$y[n] = h[n] * x[n]$$

The autocorrelation of $y[n]$ is defined to be $R_{yy}[m] = E\left[y[n]y^*[n-m]\right]$. But remember we can estimate it using the **short-time autocorrelation:**

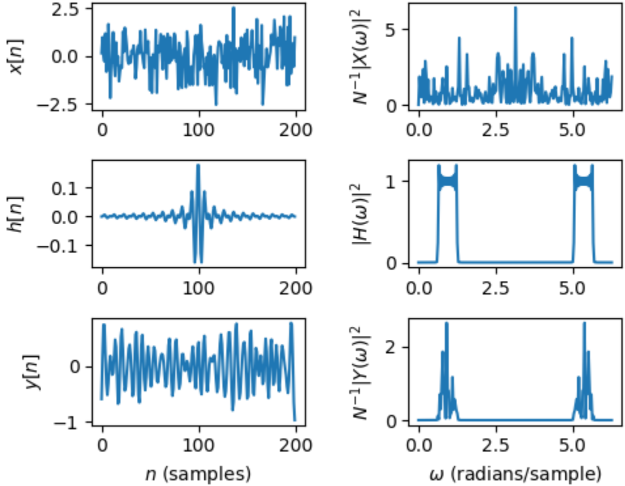
$$\begin{aligned} R_{yy}[n] &= \frac{1}{N} E\left[y[n] * y^*[-n]\right] \\ &= \frac{1}{N} E\left[h[n] * x[n] * h^*[-n] * x^*[-n]\right] \\ &= \frac{1}{N}\left(h[n] * h^*[-n] * E\left[x[n] * x^*[-n]\right]\right) \\ &= h[n] * h^*[-n] * R_{xx}[n] \end{aligned}$$

Spectrum of a Bandpass-Filtered Noise

$$R_{yy}[n] = h[n] * h^*[-n] * R_{xx}[n]$$

$$R_{yy}(\omega) = |H(\omega)|^2 R_{xx}(\omega)$$

Unvoiced Speech, Step 2: Bandpass Filter the White Noise



Outline

The Vocoder Block Diagram

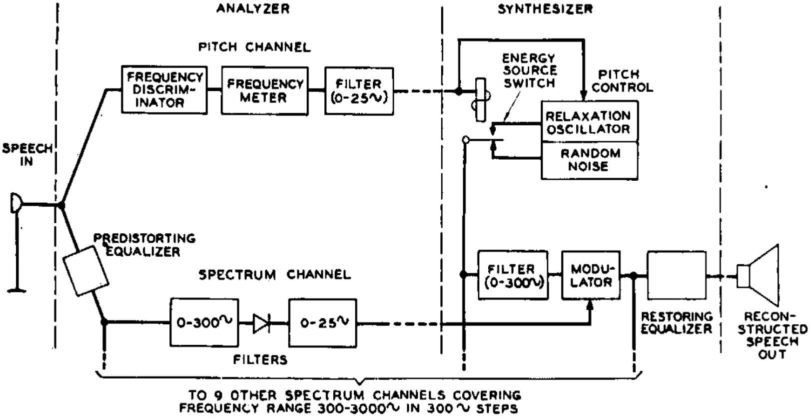


Fig. 2. Schematic arrangement of the Vocoder.

Adding Random Signals

Suppose that $x[n]$ and $y[n]$ are two uncorrelated random signals, and we add them together:

$$z[n] = ax[n] + by[n]$$

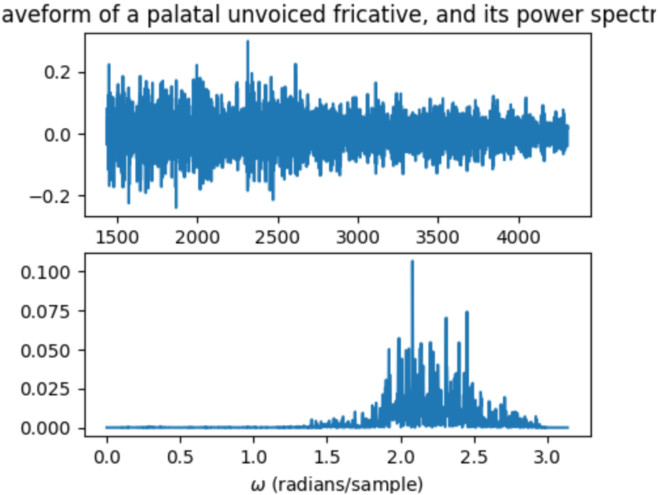
What are the autocorrelation and power spectrum of $z[n]$?

$$\begin{aligned} R_{zz}[m] &= E\left[z[n]z^*[n-m]\right] \\ &= E\left[\left(ax[n] + by[n]\right)\left(a^*x^*[n-m] + b^*y^*[n-m]\right)\right] \\ &= |a|^2 R_{xx}[m] + |b|^2 R_{yy}[m], \end{aligned}$$

and

$$R_{zz}(\omega) = |a|^2 R_{xx}(\omega) + |b|^2 R_{yy}(\omega)$$

Real Speech



→Listen←

How to Synthesize a Fricative or a Stop Burst

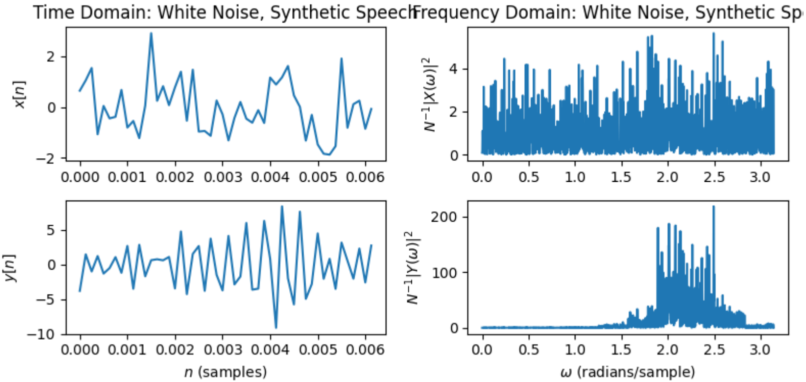
Start with white noise,

$$R_{xx}(\omega) = 1$$

Filter by a set of 10 bandpass filters $H_l(\omega)$, each about 300Hz wide, then adjust the amplitude of each one (A_l) to match the amplitude of the speech signal in the same band:

$$R_{yy}(\omega) = \sum_{l=1}^{10} A_l \sum_{k=0}^{N-1} |H_l(\omega)|^2 \, R_{xx}(\omega)$$

Synthetic Speech



→Listen←

Outline

Conclusions: How to scale the bands of a power spectrum to make fricatives

1. White noise has an autocorrelation of $R_{xx}[m] = \delta[m]$, and a power spectrum of $R_{xx}(\omega) = 1$.
2. Convolution:
$$y[n] = h[n] * x[n] \quad \leftrightarrow \quad R_{yy}[m] = h[m] * h^*[-m] * R_{xx}[m]$$
3. Linearity:
$$z[n] = ax[n] + by[n] \quad \leftrightarrow \quad R_{zz}[n] = |a|^2 R_{xx}[n] + |b|^2 R_{yy}[n]$$