Harmonic Sound : e.g. violin, flute, french horn, voice, etc.

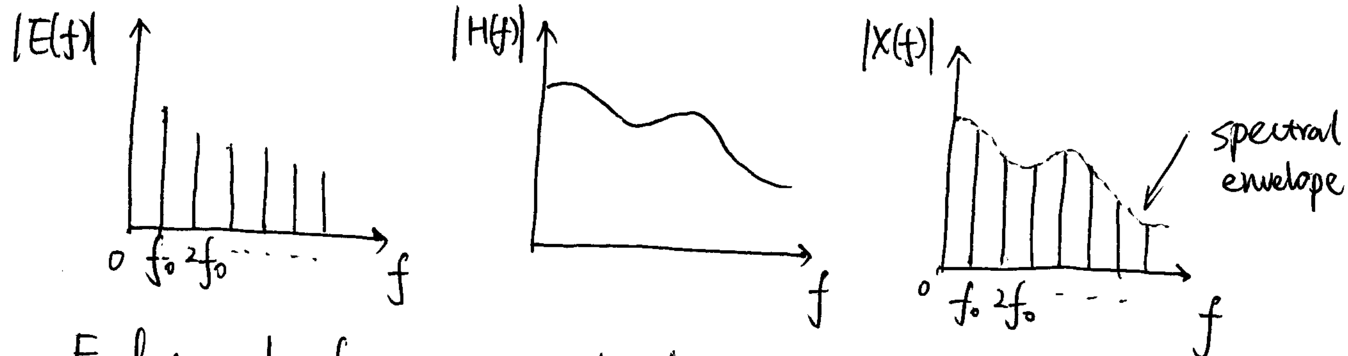
Violin : String vibration ⟹ Body resonance ⟹ Sound production

(Source / excitation) (filter) (Signal)

time domain $e(t) \ast h(t) = x(t)$

$h(t)$: impulse response of linear filter

freq. domain $E(f) \times H(f) = X(f)$



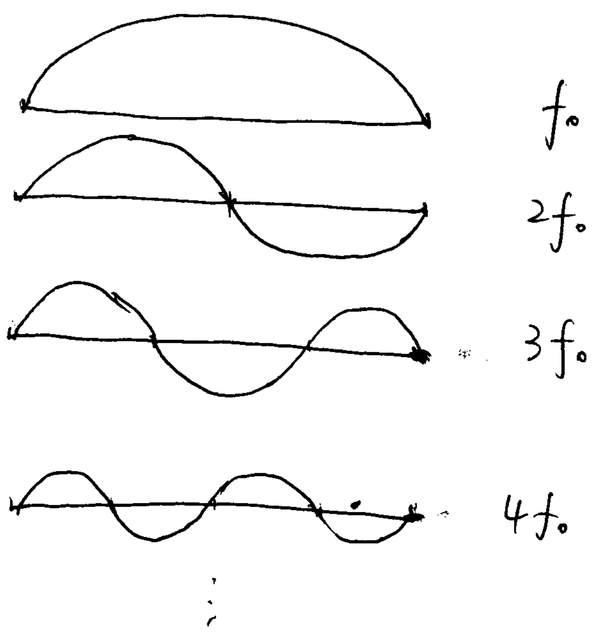
f_0 : Fundamental frequency, related to perception "pitch"

nf_0 : Harmonics

String vibration

Standing waves.

a periodic impulse train



Timbre: the thing that people use to discriminate two sounds with the same pitch, loudness, and duration.

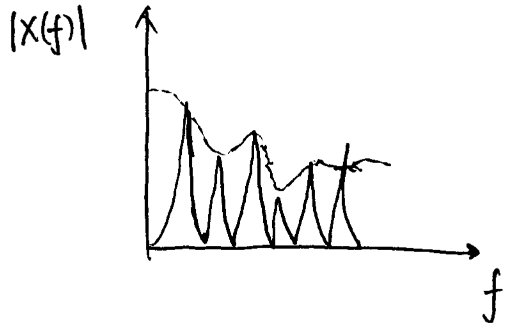
What is timbre exactly?

① spectral envelope.

② temporal dynamics.

Question: How do we extract and represent spectral envelope from the final sound production?

First thought: direct calculation from magnitude spectrum



Connect spectral peaks (ideally should represent harmonics) with a smooth curve.

Problem: ① the curve can be sensitive to the peaks. (How do we make sure the peaks are all correct, i.e., represent harmonics?) (What if the signal is not harmonic?)

② Represent the curve non-parametrically? A vector with very high dimension.

Idea: Calculate and represent spectral envelope with a parametric representation.

Linear Predictive Coding. (Time domain representation of spectral envelope)

Assume current singal is a linear combination of past signals plus the excitation signal. (Auto-regressive model)

$$x[n] = \sum_{k=1}^{p} a_k \underbrace{x[n-k]}_{\text{past signal}} + \underbrace{e[n]}_{\text{excitation (periodic impulse train)}}$$

By Z-transform

$$X(z) = \sum_{k=1}^{p} a_k X(z) z^{-k} + E(z)$$

$\therefore$ transfer function $H(z) = \frac{X(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$

(Resonance filter frequency response) (all-pole model)

We can use $\{a_k\}_{k=1,\cdots,p}$ to represent the filter, i.e., the spectral envelope.

p is the model order.

Compared to the signal, the filter changes relatively slowly.

For musical instruments, it doesn't change basically, as the instrument body doesn't change.

For speech, it changes little within 20ms.

We perform short-time analysis of the signal to estimate a_k.

$$e[n] = x[n] - \sum_{k=1}^{p} a_k x[n-k]$$

Remember $e[n]$ is periodic impulse train, its value is small most of the time

It's resonable to minimize the mean square of $e[n]$ within a short time

$$\mathcal{E} \triangleq \sum_n e^2[n] = \sum_n \left(x[n] - \sum_{k=1}^{p} a_k x[n-k] \right)^2$$

$$\frac{\partial \mathcal{E}}{\partial a_k} = 2 \sum_n \left(x[n] - \sum_{i=1}^{p} a_i x[n-i] \right) \cdot \left(- x[n-k] \right) = 0$$

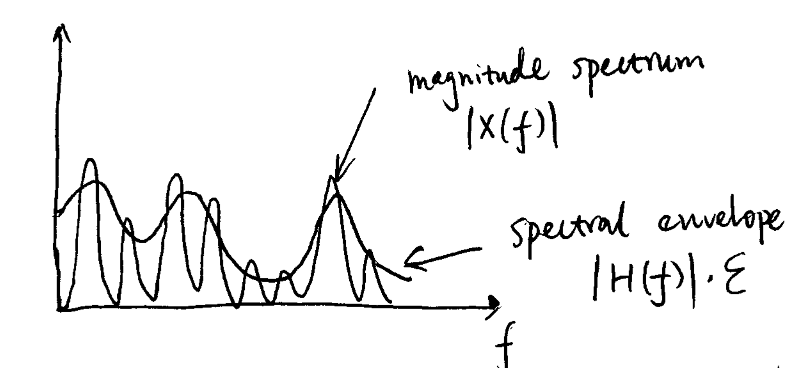
$$\therefore \sum_n x[n]\, x[n-k] = \sum_{i=1}^{p} a_i \sum_n x[n-i]\, x[n-k] \quad \text{for } k=1,\cdots,p.$$

p equations, p unknowns.

Let $\varphi[i,k] = \sum_n x[n-i]\, x[n-k]$ Autocorrelation. $R(|i-k|)$

then we have $\varphi[0,k] = \sum_{i=1}^{p} a_k \varphi[i,k]$ for $k=1,\cdots,p$

frequency response of $H(z)$



$\mathcal{E} \cdot |H(f)|$ becomes less smoother and closer to $|X(f)|$ if increase p.

Here we use "energy matching criterion" to define/draw the spectral envelope.

i.e. the impulse response of the model has the same energy as the observed signal $x[n]$.

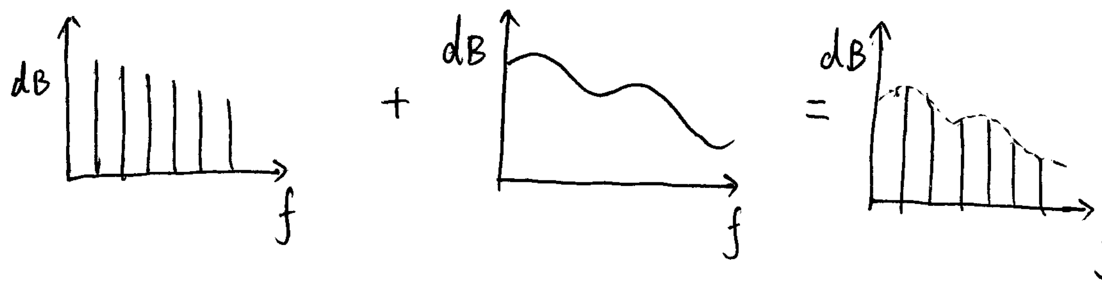
The error signal $e[n]$ is ~~a~~ our estimate of the excitation signal, where the resonance filter effect has been removed. Also called "whitened signal". Its spectrum has a flat envelope.

Cross synthesis: Combine $x_1[n]$'s excitation signal with $x_2[n]$'s filter to synthesize new signal $x_3[n]$.

1. Calculate $H_1(z)$ of $x_1[n]$ using LPC.
2. Calculate $e_1[n]$ by removing effect of $H_1(z)$
3. Calculate $H_2(z)$ of $x_2[n]$ using LPC
4. Filter $e_1[n]$ with $H_2(z)$ to synthesize $x_3[n]$.

Cepstrum (cepstral domain representation of spectral envelope).

Basic idea:



$$e(t) \quad * \quad h(t) \quad = \quad x(t)$$

$$\Rightarrow \quad E(f) \quad \times \quad H(f) \quad = \quad X(f)$$

$$\Rightarrow \quad |E(f)| \quad \times \quad |H(f)| \quad = \quad |X(f)|$$

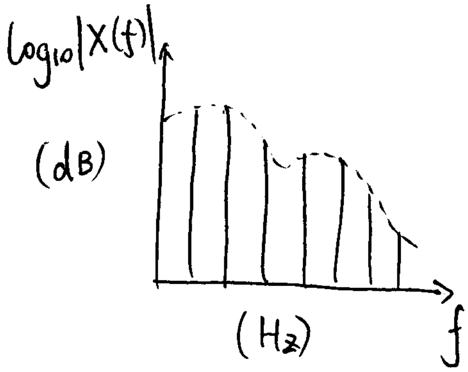
$$\Rightarrow \quad 20\log_{10}|E(f)| + 20\log_{10}|H(f)| = 20\log_{10}|X(f)|$$

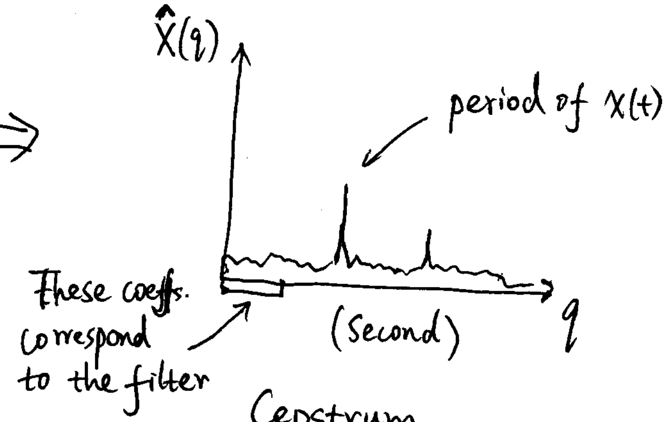
View these spectra as ~~time do~~ "frequency-domain signal"

Excitation	Filter	Outcome
periodic, high freq.	low freq.	mixture

If we perform Fourier analysis on ~~X(t)~~. $20\log_{10}|X(f)|$ and transform it to a new domain, low coefficients will correspond to excitation, while high coefficients will correspond to the filter!

(Because Fourier transform is <u>linear</u> and it <u>separates low and high "freq"s</u>)





spectrum	Cepstrum
frequency	quefrency
filtering	liftering

Digital implementation: $\hat{X}[q] = \mathrm{IFFT}\left\{ \log \left| \mathrm{FFT}[x[n]] \right| \right\}$, $q = 0, \cdots, N-1$.

It has the same length of spectrum and signal, N.

The first several (e.g., 20, 40) cepstral coefficients correspond to the resonance filter. The number is called the order.

In other words, we can reconstruct spectral envelope from thes cepstral coefficients by taking FFT.

Math: $$\hat{X}[n] = \frac{1}{N} \sum_{k=0}^{N-1} a[k] e^{j2\pi kn/N}$$ where $a[k] = 20\log_{10}|X[k]|$ (symmetric)

$$= \frac{1}{N} \sum_{k=0}^{N-1} a[k] \left\{ \cos\left(\frac{2\pi kn}{N}\right) + \underbrace{j \sin\left(\frac{2\pi kn}{N}\right)}_{\text{cancelled}} \right\}$$

$$= \frac{1}{N}\left(\underbrace{a[0]}_{\text{DC}} + \underbrace{(-1)^n a\left[\frac{N}{2}\right]}_{\text{Nyquist}} \right) + \underbrace{\frac{2}{N} \sum_{k=1}^{N/2} a[k] \cos\left(\frac{2\pi kn}{N}\right)}_{\text{Positive frequency}}$$

$$= \mathrm{DCT}\{a[0 : N/2]\}$$

(Discrete cosine transform of the positive frequency range log ~~spe~~ amplitude spectrum).

Another pespective to see cepstrum.

Log-amp spectrum can be approximated by a linear combination of several sinusoids, with coefficients of a cepstral coefficients.

$$a[k] \approx c_0 + \sqrt{2} \sum_{i=1}^{p-1} c_i \cos\left(2\pi i \frac{k}{N}\right) \quad \text{, i.e.} \qquad (*)$$

$$\begin{bmatrix} a[0] \\ \vdots \\ a[N/2] \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \sqrt{2}\cos(2\pi 1 f_0) & \cdots & \sqrt{2}\cos(2\pi(p-1)f_0) \\ \vdots & & & \\ 1 & \sqrt{2}\cos(2\pi 1 f_{N/2}) & \cdots & \sqrt{2}\cos(2\pi(p-1)f_{N/2}) \end{bmatrix}}_{} \begin{bmatrix} c_0 \\ \vdots \\ c_{p-1} \end{bmatrix}$$

where $f_k = k/N$.

M: First p colums of a DCT matrix.
Colums are orthogonal.

Least-square solution:

$$\begin{bmatrix} c_0 \\ \vdots \\ c_{p-1} \end{bmatrix} = \underbrace{(M^T M)^{-1}}_{= 1/N} M^T \begin{bmatrix} a[0] \\ \vdots \\ a[N/2] \end{bmatrix} = \frac{1}{N} M^T \begin{bmatrix} a[0] \\ \vdots \\ a[N/2] \end{bmatrix}$$

∴ Cepstral coefficients are the least square solution to approximate log-amp spectrum using (*).

Mel-frequency Cepstral Coefficients (MFCC)

- broadly used in speech recognition, speaker identification, timbre modeling.
- warp freq to log-freq in spectrum before taking IFFT.