Topic 6

Timbre Representations

We often say…

- “that singer’s voice is magnetic”
- “the violin sounds bright”
- “this French horn sounds solid”
- “that drum sounds dull”

- What aspect(s) of sound do these words describe?
  – Pitch? Loudness? Harmonicity?

We can easily…

- Recognize a friend’s voice from only a few words
- Distinguish the sound of clarinet from oboe, even if they play the same note with the same loudness and duration

Oboe Clarinet

- What physical properties of sound do we use?

Timbre (tone quality, tone color)

"That attribute of auditory sensation in terms of which a subject can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar."

---- ANSI, 1960.

- OK, but…, what is timbre?
- What physical properties does timbre refer to?

Timbre and Physics

- "Quality of tone [timbre] should depend on the manner in which the motion is performed within the period of each single vibration"

---- Helmholtz, 1877.

- "Timbre depends primarily upon the spectrum of the stimulus, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus."

---- ANSI, 1960.

Examples

- Spectral energy distribution
  - The clarinet and oboe example
- Attack (onset)

Without attack

With attack

- Temporal evolution

Time reverse

Timbre and Sound Synthesis

- Sound synthesizers use some of the previously mentioned attributes to synthesize instrument sounds

- Somewhat similar to the real instrument, but not quite

The concept of timbre is still vague

- "The word timbre...is empty of scientific meaning, and should be expunged from the vocabulary of hearing science."

  ---- Keith Martin, PhD thesis, 2000.

- But, it's worth figuring it out, at least partially, if we want to design computational systems to recognize timbre

Physical vs. Psychological

Frequency	Pitch	Low - high
Intensity	Loudness	Soft - loud
?	Timbre	Warm Bright Rough Violin-like …

Question

- How to find out what attributes contribute to the diversity of timbre?
  - Randomly choose some attribute (or their combinations) and change it, and then see if the timbre is significantly changed?
  - So many attributes and combinations
  - Doesn’t sound efficient

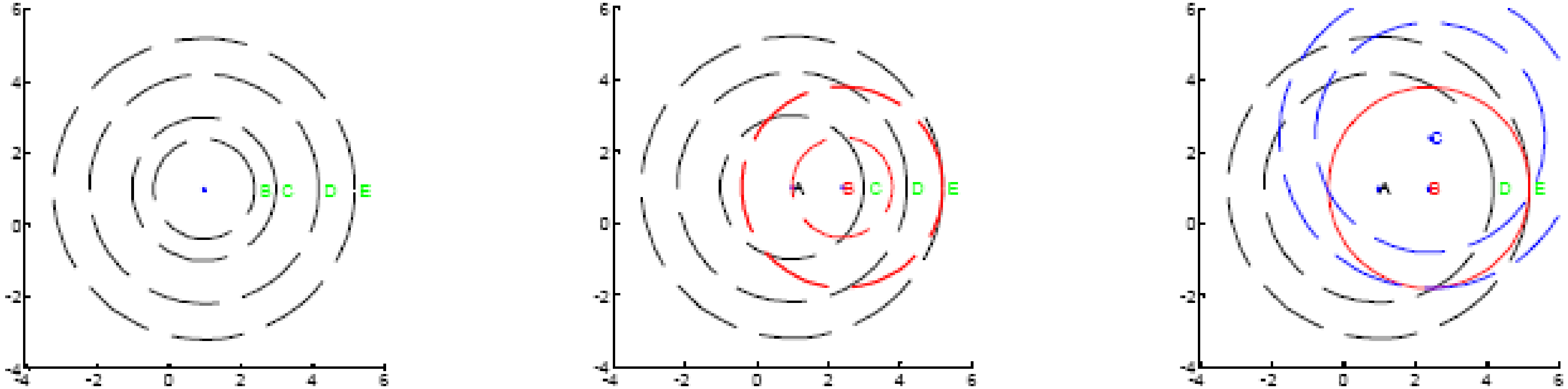
Induction from Observations

- Collect a number of sounds with different timbre
- Ask a number of people to rate the timbre similarity/distance between the sounds
- Embed the similarity/distance matrix into a low dimensional space
- Observe/listen to the change of sound along some dimensions

Multidimensional Scaling (MDS)

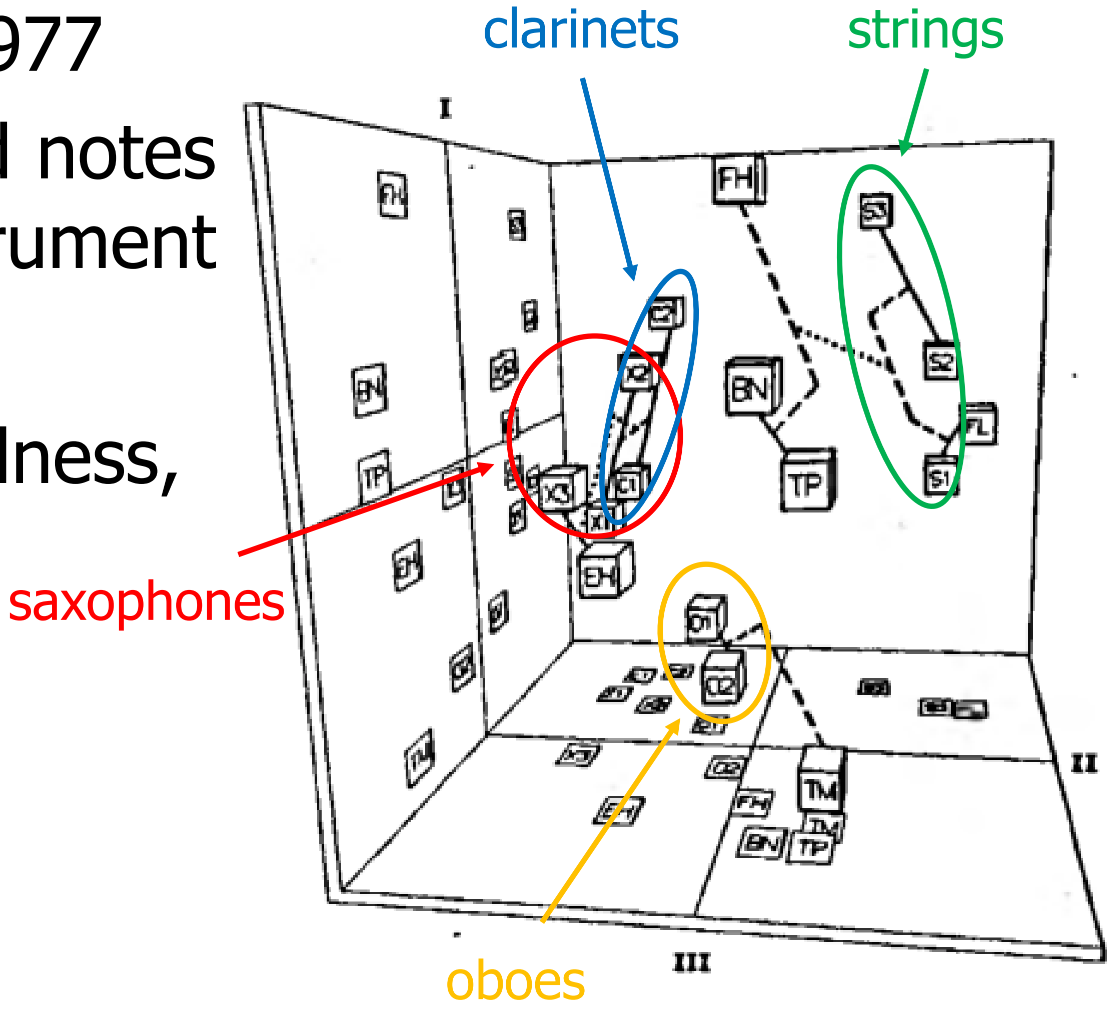
- We have got a distance matrix between objects
- Put objects into a low dimensional space such that the distances are (approximately) preserved

	A	B	C	D	E
A	0	1.4	2	3.2	4.2
B		0	1.4	2.8	2.8
C			0	4.2	3.2
D				0	4
E					0



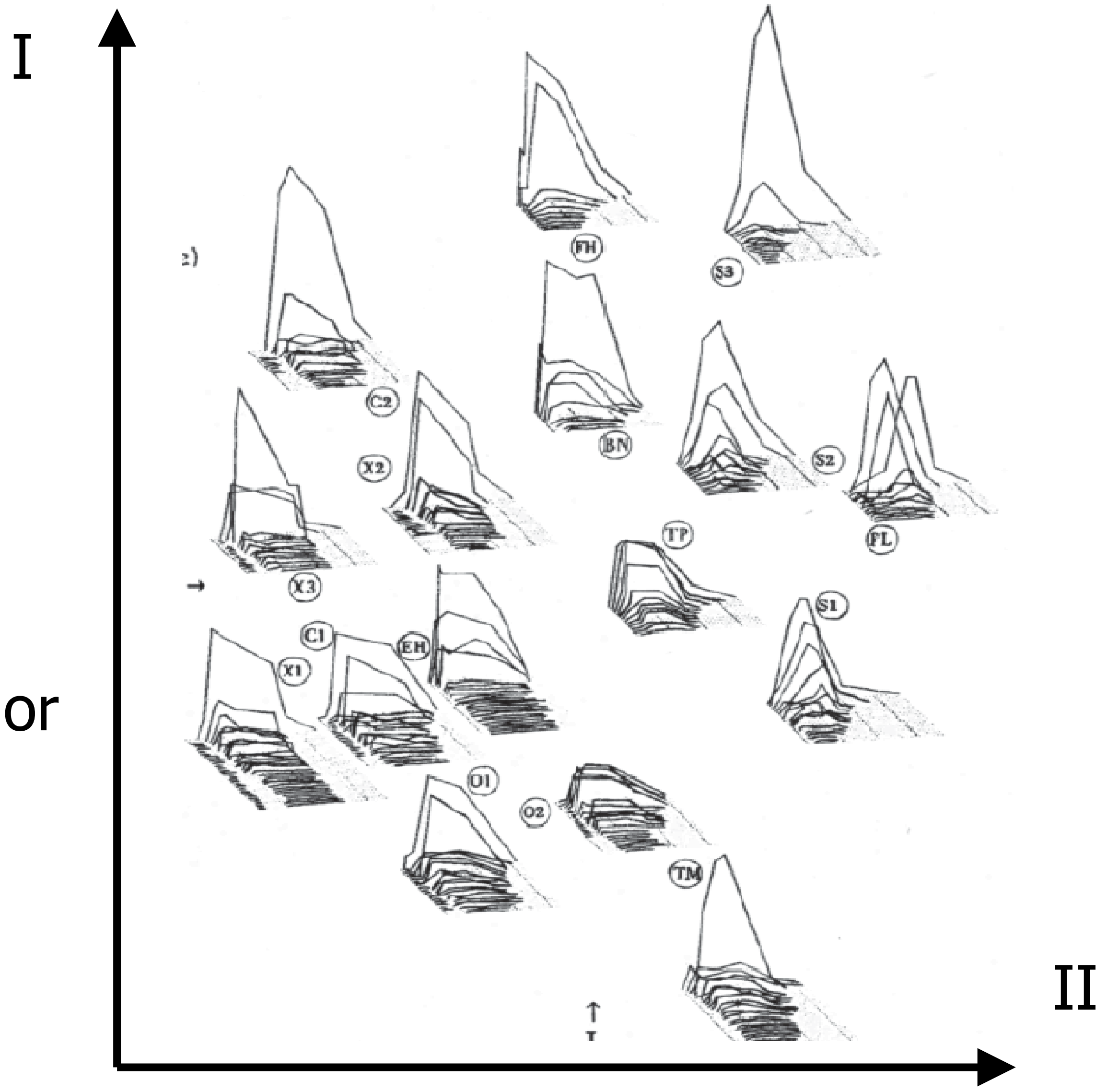
MDS for Timbre

- John M. Grey, 1977
- 16 resynthesized notes by different instrument
- Same pitch, loudness, and duration
- 35 listeners



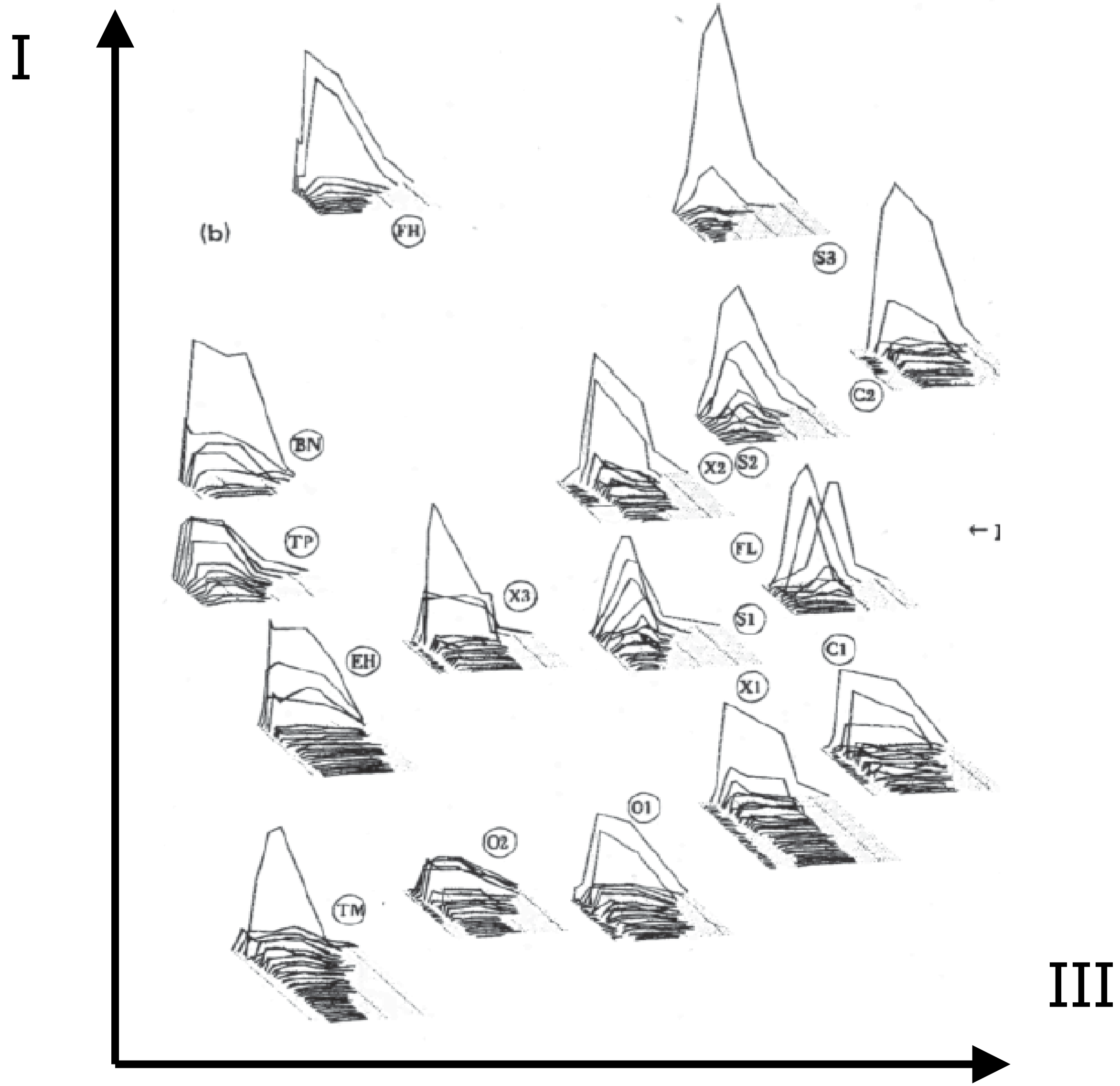
Dimensions I and II

- Dimension I corresponds to spectral energy distribution
- Dimension II corresponds to spectral fluctuation or synchronicity



Dimensions I and III

- Dimension I corresponds to spectral energy distribution
- Dimension III corresponds to the presence of inharmonic energy during attack



Human Instrument Classification

Confusion matrix

	Response															
Stimulus	O1	O2	EH	BN	C1	C2	X1	X2	X3	FL	TP	FH	TB	S2	S1	S3
O1	173	82	35	4	8	5	10	6	3	–	8	–	6	6	5	2
O2	115	218	24	3	1	–	2	–	–	1	–	–	–	–	–	–
EH	40	38	248	12	–	–	5	3	3	–	1	–	8	2	4	1
BN	1	4	8	305	–	–	–	–	–	–	14	26	9	–	–	–
C1	1	–	–	–	294	60	8	6	–	–	–	–	–	–	–	1
C2	–	–	2	–	77	258	10	12	6	–	1	–	–	–	–	2
X1	1	–	2	3	1	2	229	86	39	–	1	–	3	–	–	–
X2	1	–	2	3	6	8	67	231	39	1	–	1	–	–	–	–
X3	6	9	29	4	3	2	30	42	236	–	1	–	3	–	–	–
FL	–	–	–	–	–	–	–	–	–	358	–	–	–	5	8	1
TP	1	–	5	5	–	–	–	–	–	–	342	4	7	1	–	–
FH	–	–	2	1	–	–	–	–	–	5	7	356	–	–	–	–
TB	3	4	1	–	–	–	1	–	–	1	9	–	346	–	1	–
S2	–	–	–	–	1	–	–	–	–	3	–	–	–	267	74	24
S1	6	2	3	6	1	–	–	–	–	7	2	–	1	57	263	9
S3	–	–	–	1	2	–	–	–	–	1	–	2	–	26	15	320

- Human improves classification performance after practice, i.e., our ears can figure out what aspects of sounds are related to timbre.

Limitations of Grey'77

- Very few notes
- The notes are resynthesized. Not real.
- Only one pitch and loudness

- Didn't look at timbre consistency of notes played by the same instrument

Timbre Definition Revisit

"That attribute of auditory sensation in terms of which a subject can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar."

---- ANSI, 1960.

- Doesn't mention the role timber plays in cases where pitch and/or loudness are different.
  - Two notes played by the same instrument have similar timbre, even if they have different pitch and/or loudness.

Timbre Features

- Physical attributes of sounds that represent timbre
- Easy to calculate from the signal
- Can discriminate different sound sources (e.g., musical instruments, talkers)
- Approximately invariant to pitch/loudness changes for the same source

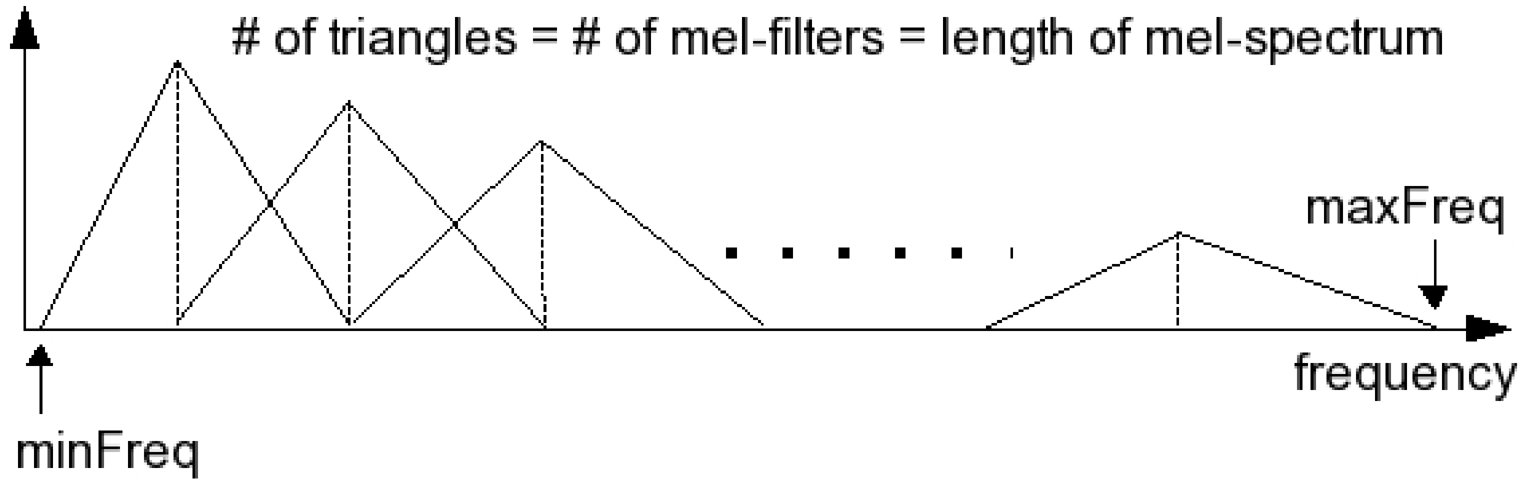
Time-domain Features

- RMS
  - Used to discriminate silence/non-silence

- Zero crossing rate (ZCR)
  - How often the time-domain signal changes its sign
  - Describes the amount of high-frequency energy
  - Correlates strongly with spectral centroid
  - Quite discriminative for percussion instruments

$$ZCR(n) = \frac{1}{2N}\sum_{i=1}^{N}\left|\mathrm{sign}(x[n+i]) - \mathrm{sign}(x[n+i-1])\right|$$

Mel Filter Bank



- Filters spaced equally in the log of the frequency.
- Mels are (more or less) related to frequency by…

$$\mathrm{Mel} = 2595 \log_{10}(1 + \frac{f}{700})$$

- Edge of each filter = center frequency of adjacent filter
- Typically, 40 filters are used

Spectral Features

- Can be calculated from either the linear frequency magnitude spectrum, or the mel-scale filter bank response.

- From now on, let $X[k]$ be either a linear frequency scale magnitude spectrum or a mel-scale filter bank response.

Spectral Features

- Spectral centroid

$$C_f = \frac{\sum_k kX[k]}{\sum_k X[k]}$$

- Spectral spread

$$S_f^2 = \frac{\sum_k \left(k - C_f\right)^2 X[k]}{\sum_k X[k]}$$

Spectral Features

- Spectral skewness
  - How asymmetric of the frequency distribution around the spectral centroid

$$\gamma_1 = \frac{\sum_k \left(k - C_f\right)^3 X[k]}{S_f^3 \sum_k X[k]}$$

- Spectral kurtosis
  - The peakiness of the frequency distribution

$$\gamma_2 = \frac{\sum_k \left(k - C_f\right)^4 X[k]}{S_f^4 \sum_k X[k]}$$

Spectral Features

- Spectral flatness
  - How flat (i.e., “white-noisy”) the spectrum is

$$SFM = 10\log_{10}\left(\frac{\left(\prod_{k=1}^{K} X[k]\right)^{1/K}}{\frac{1}{K}\sum_{k=1}^{K} X[k]}\right)$$

- Spectral irregularity
  - The jaggedness of the spectrum

$$SI = \frac{\sum_k (X[k] - X[k+1])^2}{\sum_k X[k]^2}$$

Spectral Features

- Spectral roll-off
  - The frequency index R below which a certain fraction γ of the spectral energy resides

$$\sum_{k=1}^{R} X[k]^2 \geq \gamma \sum_{k} X[k]^2$$

- Spectral flux (delta spectrum magnitude)
  - Measure of local spectral change

$$SFX(t) = \sum_{k} \left(\frac{X_t[k]}{\sum_k X_t[k]} - \frac{X_{t-1}[k]}{\sum_k X_{t-1}[k]} \right)^2$$

Harmonic Features

- Inharmonicity
  - Average deviation of spectral components from perfect harmonic positions

$$IH = \frac{2}{F_0} \times \frac{\sum_{h=1}^{H} |f_h - hF_0| \times a^2(h)}{\sum_{h=1}^{H} a^2(h)}$$

- Odd-to-even ratio

$$OER = \frac{\sum_{h\ \mathrm{odd}} a^2(h)}{\sum_{h\ \mathrm{even}} a^2(h)}$$

Harmonic Features

- Tristimulus
  - Relative weights of low and high harmonics

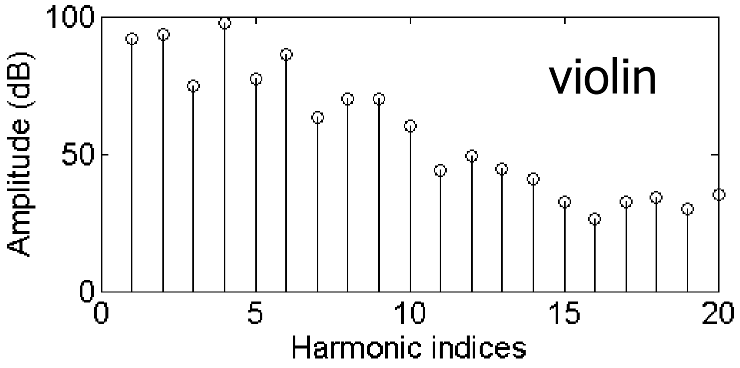
$$T1 = \frac{a^2(1)}{\sum_{h=1}^{H} a^2(h)}$$

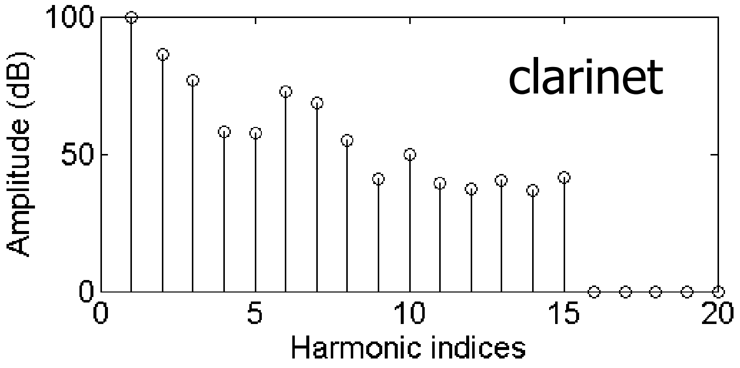
$$T2 = \frac{a^2(2) + a^2(3) + a^2(4)}{\sum_{h=1}^{H} a^2(h)}$$

$$T3 = \frac{\sum_{h=5}^{H} a^2(h)}{\sum_{h=1}^{H} a^2(h)}$$

Harmonic Features

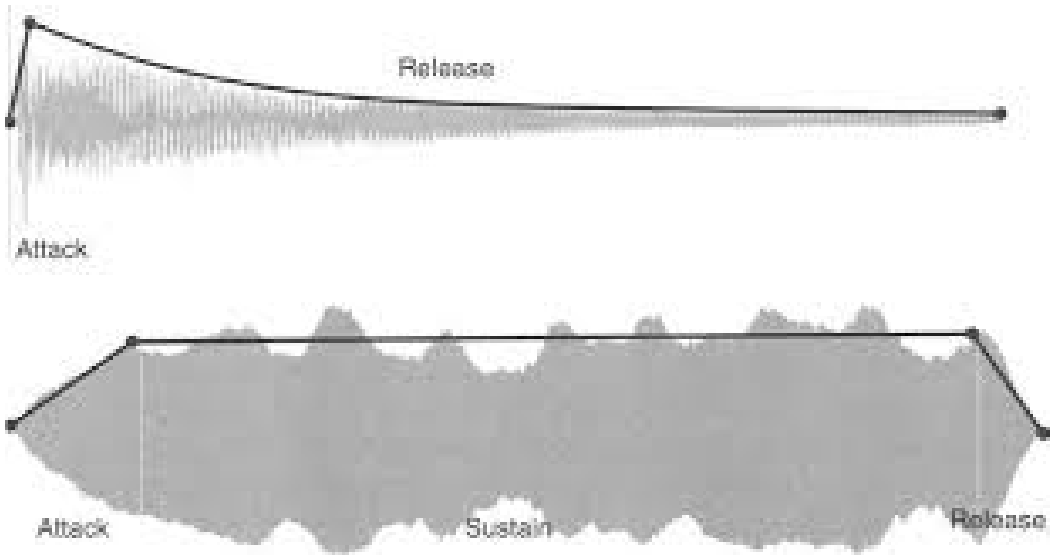
- Harmonic structure
  - Relative normalized amplitudes (dB) of harmonics





Temporal Features

- Amplitude envelope



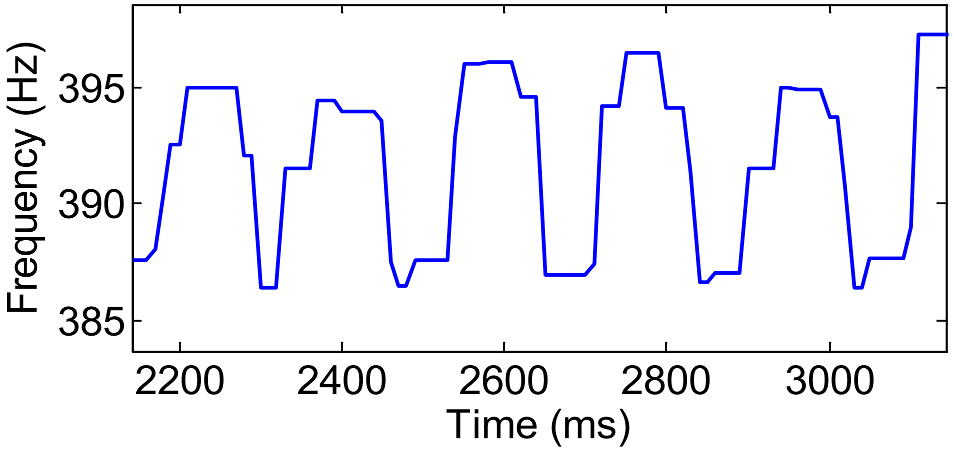
- Attack time

$$LAT = \log_{10}(t_{80} - t_{20})$$

Temporal Features

- Vibrato rate and depth
  - How fast and how much the pitch changes

Pitch contour of a violin note



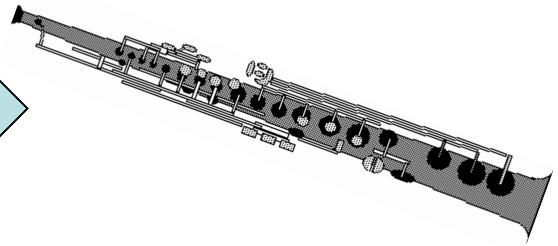
  - Around 5-6Hz
  - How to calculate its period and amplitude?

Temporal Features

- Tremolo
  - Amplitude changes periodically
  - Perform FFT on the RMS contour

Source-Filter Model

Source Signal $x(t)$ → **Filter** $h(t)$ → **Output Signal** $y(t)$

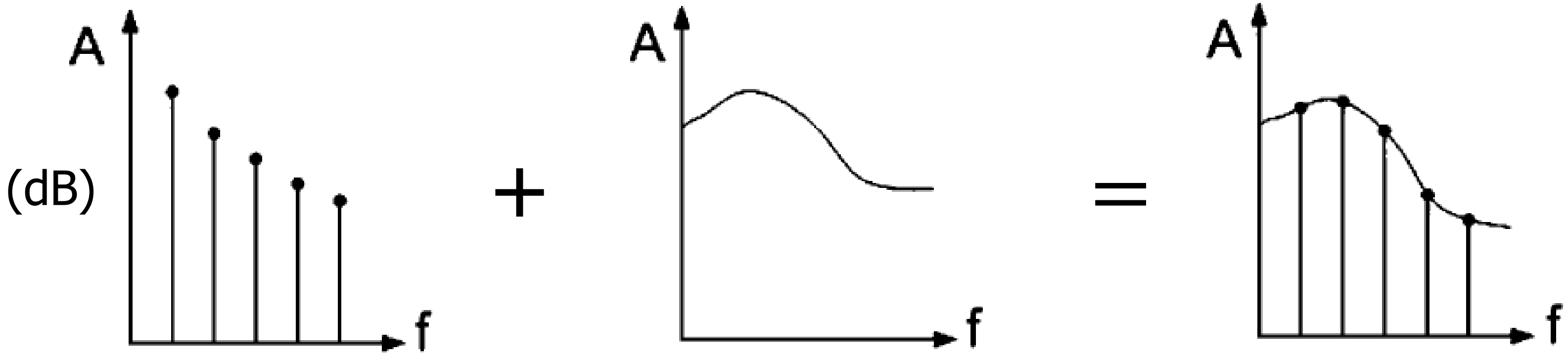


- Filtering is convolution in time domain, i.e., multiplication in frequency domain.

$$x(t) * h(t) = y(t)$$
$$X(f) \times H(f) = Y(f)$$
$$|X(f)| \times |H(f)| = |Y(f)|$$

Harmonic Sounds

Source Signal $x(t)$ → **Filter** $h(t)$ → **Output Signal** $y(t)$



Excitation spectrum (harmonic comb)

Resonance spectrum (spectral envelope)

Harmonic sound spectrum

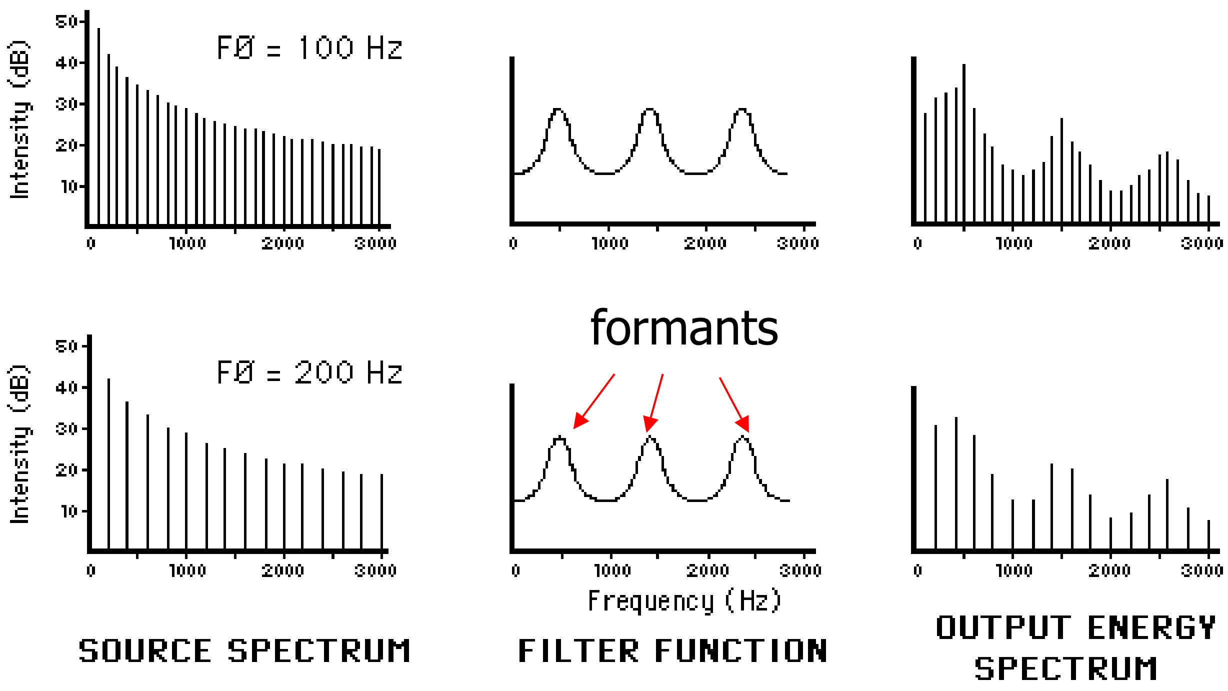
- For log-amplitudes, multiplication becomes addition

$$\log_{10}|X(f)| + \log_{10}|H(f)| = \log_{10}|Y(f)|$$

Spectral envelope → timbre

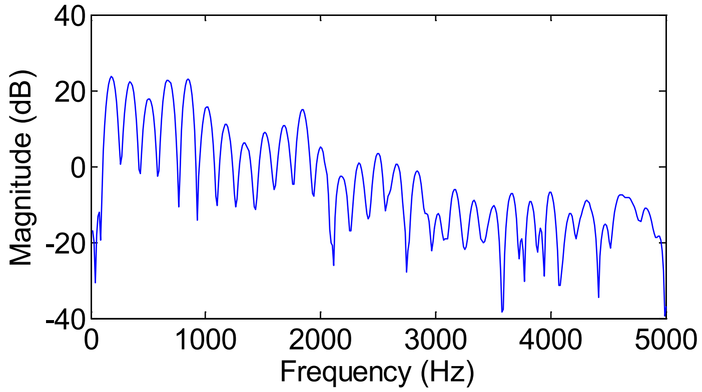
- The excitation spectrum changes with pitch
- The spectral envelope changes with the shape, material, etc. of the resonance body
  - It does not change much with pitch.

Speech production (from Haskins Lab at Yale)



How to characterize the envelope?

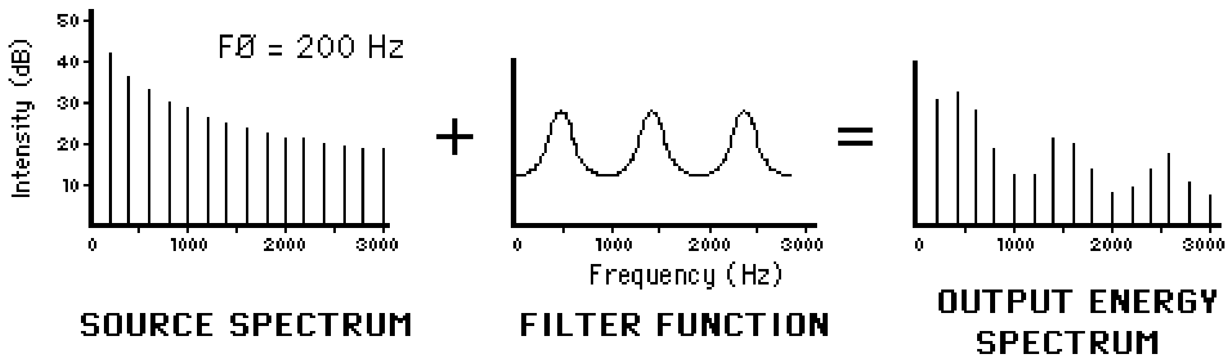
Harmonic sound magnitude spectrum



- First thought
  - Detect peaks
  - Draw a smooth line connecting the peaks
  - This line is the envelope
- How to represent the envelope?
  - Non-parameterized? Very high dimension
  - Parameterized. How?
    - Polynomial?
    - Sinusoidal?

Basic Idea of Cepstrum

- View the log-magnitude spectrum as a mixture of two signals, one high-frequency and one low frequency.



- What if we perform Fourier analysis on the mixture?
  - Fourier transform is linear!
  - Fourier transform separates low/high frequencies!
- Higher Fourier coefficients ⬄ excitation spectrum
- Lower Fourier coefficients ⬄ spectral envelope

Formal Definition of Cepstrum

- Bogert et al. 1963, heuristically

$$\text{power cepstrum} = \left|\mathcal{F}^{-1}\{\log|\mathcal{F}\{x(t)\}|^2\}\right|^2$$

- Digital version
  - Use DFT and IDFT to replace Fourier transforms.

- Why IDFT?
  - Well, it actually doesn’t matter for real signals.

IDFT or DFT? It doesn't matter.

- Remember IDFT

$$y[n] = \frac{1}{N}\sum_{k=0}^{N-1} Y[k]e^{j2\pi kn/N}$$

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

- Now, substitute $\mathrm{a}[k] = \log|X[k]|$ (symmetric, real) as $Y[k]$ into the equation

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

Discrete Cosine Transform

- The previous equation is exactly taking DCT on the positive frequency part of the log-magnitude spectrum (k=0:N/2)

- There are many types of DCT. They are basically doing the same thing. Their differences are only at some constants, DC and Nyquist components, and sometimes a half-sample phase.

Cepstral Features

- Mel-frequency Cepstral Coefficients (MFCC)
  - 1. Calculate magnitude spectrum
  - 2. Calculate the mel-scale filterbank response (e.g., 40-d)
  - 3. Take log of the filterbank response
  - 4. Perform discrete cosine transform (DCT) on the 40-d vector in 3.
  - 5. Choose the several (e.g., 15) lowest-order DCT coefficients

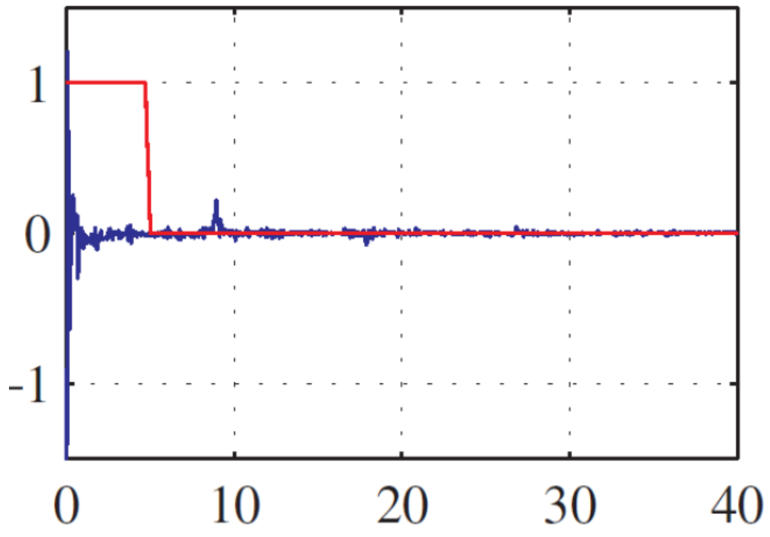
Deltas of MFCC

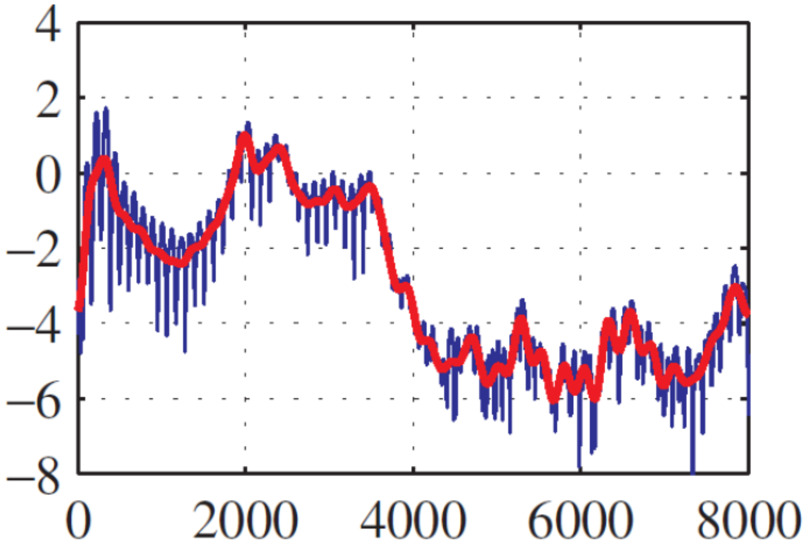
- Capture the temporal evaluation of MFCC
- Delta:
  - "velocity", the local slope. M=1 or 2.

$$\Delta\mathrm{Cep}_i(t) = \frac{\sum_{m=-M}^{M} m\mathrm{Cep}_i(t+m)}{\sum_{m=-M}^{M} m^2}$$

- Delta-delta
  - "acceleration"
- Broadly used in speech/speaker recognition, instrument recognition, etc.

Liftering



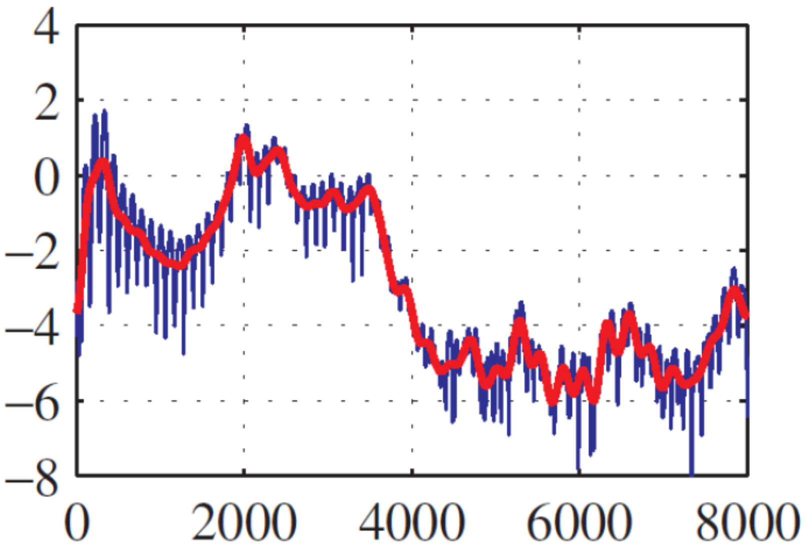


<u>Cep</u>strum | <u>Spec</u>trum
<u>Lift</u>ering | <u>Filt</u>ering
<u>Quef</u>rency | <u>Freq</u>uency

Another Explanation of Liftering

- Approximate the log-amplitude spectrum with a linear combination of several sinusoids.

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



$$\begin{pmatrix} a_0 \\ \vdots \\ a_{N/2} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & \sqrt{2}\cos(2\pi q_1 0) & \cdots & \sqrt{2}\cos(2\pi q_{p-1} 0) \\ \vdots & \ddots & & \vdots \\ 1 & \sqrt{2}\cos\left(2\pi q_1 \frac{N}{2}\right) & \cdots & \sqrt{2}\cos\left(2\pi q_{p-1}\frac{N}{2}\right) \end{pmatrix}}_{M \text{ (first } p \text{ columns of a DCT matrix)}} \begin{pmatrix} c_0 \\ \vdots \\ c_{p-1} \end{pmatrix}$$

- $q_i = i/N$ (quefrency)

Least-square Solution

$$\begin{pmatrix} c_0 \\ \vdots \\ c_{p-1} \end{pmatrix} = \underbrace{(M^T M)}_{\text{Scaled identity matrix}}{}^{-1} M^T \begin{pmatrix} a_0 \\ \vdots \\ a_{N/2} \end{pmatrix} = \frac{1}{N} M^T \begin{pmatrix} a_0 \\ \vdots \\ a_{N/2} \end{pmatrix}$$

- Columns of M are orthogonal

- The first p cepstral coefficients are the least square solution of approximating the log-amplitude spectrum using weighted sum of p sinusoids.

Question

- Can we use the previously presented features to represent a source in polyphonic audio?

- The calculation of most of those features (except harmonic features) uses the full spectrum

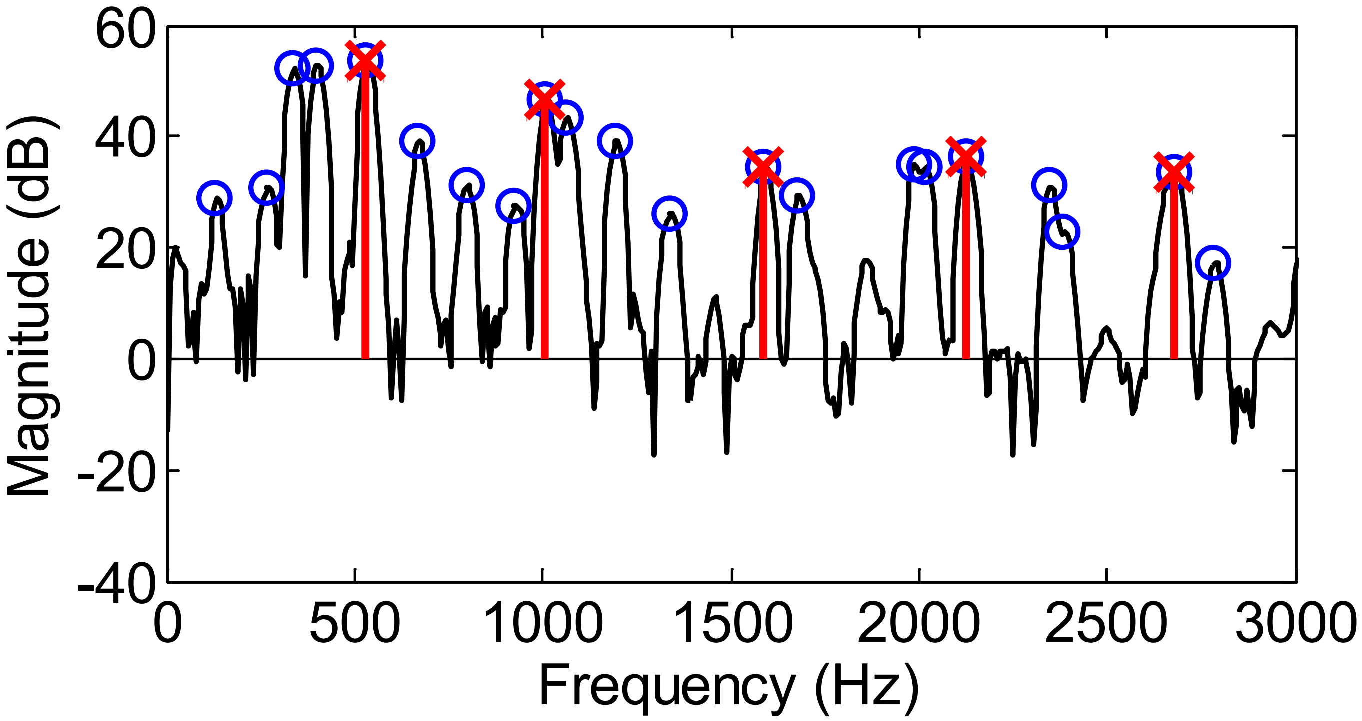
- The full spectrum of a source cannot be obtained from the mixture spectrum without source separation

Calculate features from the mixture?

- Principle
  - Find the frequency bins whose energy mostly belong to the source (i.e., observable frequencies for the source)
  - Calculate features from these frequency bins
- For harmonic sound mixtures
  - Assuming the pitch of the source is given
  - Harmonics are generally the observable frequencies
  - Calculate features from these harmonics

Harmonic Structure

- Assume the pitch of the source is given
- Detect the closest peak for each harmonic



Discrete Cepstrum (DC)

- Galas & Rodet, 1990

- Approximate the log-amplitude spectrum with a linear combination of several sinusoids, only at the observable frequencies

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

where k indexes observable frequencies.

- Least square solution of $\{c_i\}$.

Problems of DC

- The calculated cepstral coefficients tend to overfit the spectrum at observable frequencies, resulting in arbitrary values at other frequencies with huge oscillations.

- Regularized Discrete Cepstrum
  - Cappe et al., 1995
  - Regularize the smoothness of the reconstruction
  - Alleviates the problem

Uniform Discrete Cepstrum

- Duan et al., 2014
- Zero out non-observable frequencies
- Perform DCT, i.e., approximate the new log-amplitude spectrum with a linear combination of several sinusoids

$$\hat{a}[k] \approx c_0 + \sqrt{2}\sum_{i=1}^{p-1} c_i \cos\left(2\pi i \frac{k}{N}\right)$$

Where k indexes all frequencies.

- The zeros in $\hat{a}[k]$ serve as another kind of regularizer

Linear Predictive Coding

- Assuming the source-filter model.
- Assumes the current signal sample can be approximated by a linear combination of past samples and a source signal

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

- By Z transform

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

- All-pole model; autoregressive (AR) model
- $\{a_k\}$ models the resonance filter.

Estimating LPC models

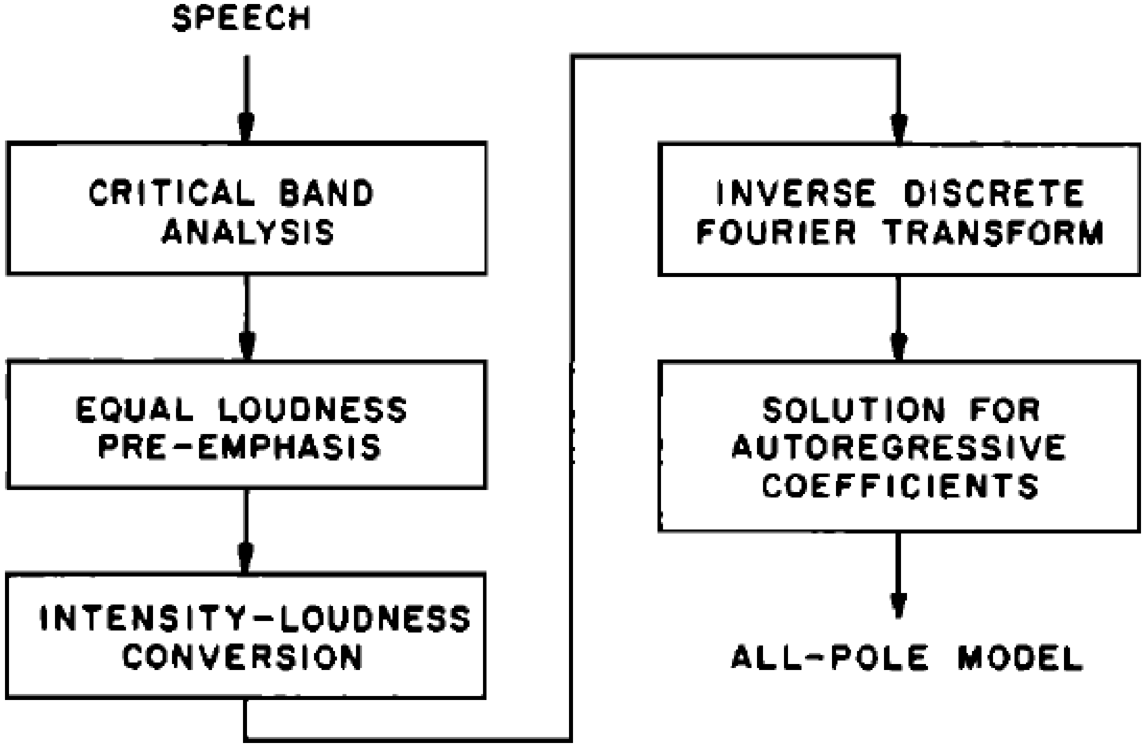
- For speech, the vocal tract (hence $\{a_k\}$) doesn't change much within about 20ms.
- Minimize the residue $e[n]$ within this range

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

- Taking derivative w.r.t. a_k, we get a system of p linear equations involving autocorrelations, i.e., Yule-Walker-Equations.

Perceptual Linear Predictive (PLP)

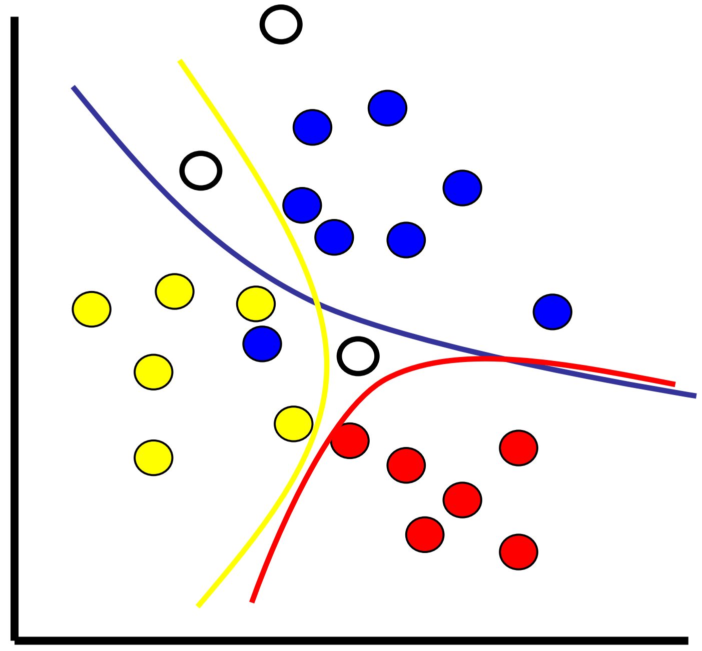
- Uses auditory models to modify LPC.



(Hermansky, 1990)

Instrument Recognition

- Training
  - Collect a bunch of notes for each instrument
  - Perform feature extraction on the notes
  - Train a classifier for each instrument
- Recognizing a note
  - Perform feature extraction on this note
  - Run each instrument classifier on it



Instrument Recognition

- Feature extraction
  - Calculate a bunch of the above-mentioned features from the audio signal
  - Stack them into a single vector (high-dimensional!)
- Feature selection
  - Which features are more useful?
  - Which features are correlated?
- Feature transformation (reduce dimensionality)
  - Principal Component Analysis (PCA), similar to MDS