SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Machine Listening

Lecture 2
- Acoustic Feature Extraction
- Speech/Music Analysis

SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Today's Topics

- Quick recap: spectral interpolation, zero padding, spectral leaking, windowing
- Low-level audio features
- Spectral features
- Tonal audio features
- Speech analysis
- Music analysis

Frequency Resolution

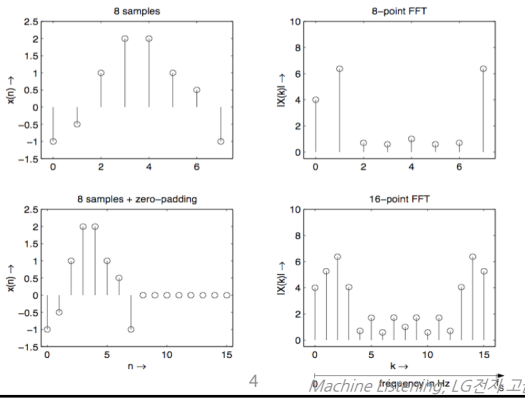
- Determined by how many sinusoids are used to describe a spectrum
- In *N*-point DFT or FFT, the frequency resolution is given by

$$\Delta f = f_s / N$$

- Intuitively, we can have finer frequency resolution if we increase *N*
- However, this results in poorer temporal resolution => tradeoff between frequency[spectral] and time[temporal] resolution

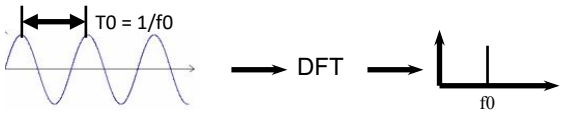
Zero-Padding

- A possible solution to increase frequency resolution without increasing *N* - i.e., without losing time resolution
- Add zero-valued samples - thus doesn't change spectrum itself - to yield better spectral resolution



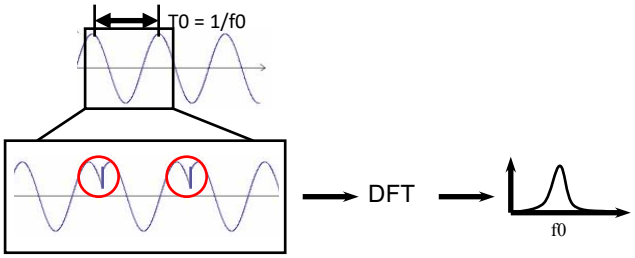
DFT of a Sinusoid

- In theory the DFT of a pure sinusoid results in a single sharp line at the frequency of the sinusoid



Spectral Leaking

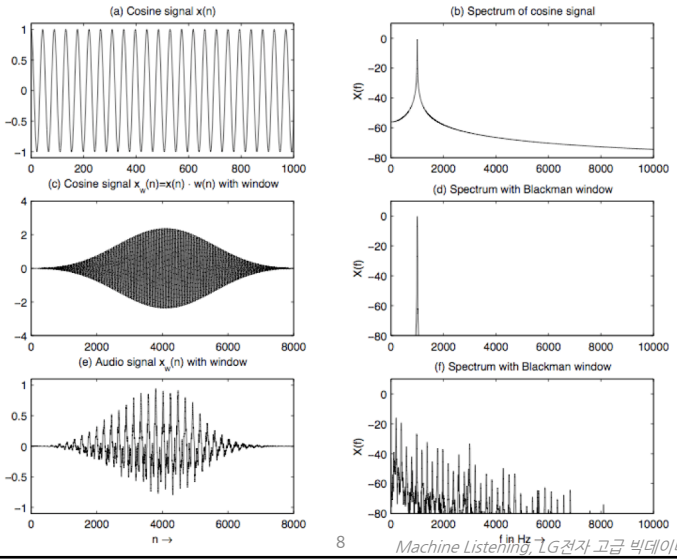
- In practice, unless we perform f0-synchronous analysis, there are discontinuities (sharp changes) at the segment boundaries that introduce some noise. Thus the spectral line around f0 is smeared.
- This is known as *spectral leaking*



Windowing

- In order to reduce spectral leaking, we need to avoid abrupt changes between the segment boundaries
- This is done by multiplying a *window* whose amplitude gradually reaches zero at both ends, thus guaranteeing the continuity of a segmented signal when repeated
- Possible windows are: rectangular, hamming, hann, Blackman, Gaussian, and so on.

Windowing Example



Audio Features

Audio Feature Extraction

- Raw audio samples are:
  - Noisy
  - Redundant
  - Computationally inefficient
  - Not a good input to MIR systems
- Need to convert them to more robust, compact yet meaningful representation called *audio features* or *acoustic features*

Audio Features

- Spectral features
  - Spectral low-level features: spectral centroid, spectral flatness measure, spectral flux, etc.
  - Spectral envelope: LPCs, MFCCs, etc.
- Temporal features
  - ZCR (zero crossing rate), tempo histogram, novelty function, etc.
- Tonal features
  - PCP (pitch class profile), chromagram, tonal centroid, etc.

Spectral Low-level Features

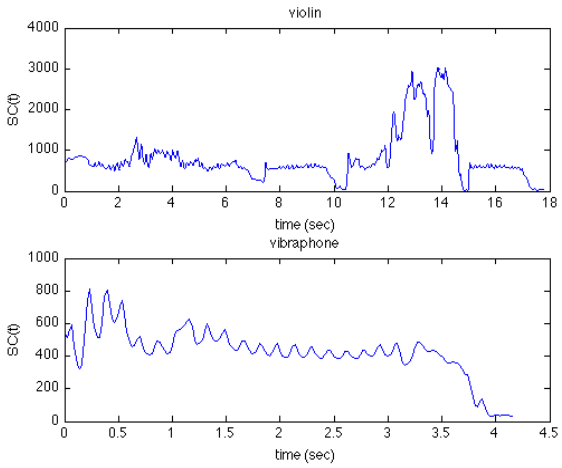
Spectral Centroid (SC)

$$SC = \frac{\sum_{k=0}^{N/2} f_k |X(k)|^2}{\sum_{k=0}^{N/2} |X(k)|^2},$$

where f_k is the center frequency of kth bin and |X(k)| is the DFT

- Characterizes the center of gravity of the (power) spectrum
- Usually associated with the "sharpness / dullness (or brightness / darkness)" of a sound

Spectral Centroid: Example



(a) violin

(b) vibraphone

Temporal evolution of the spectral centroid for two instrumental sounds
(a) violin and (b) vibraphone

Spectral Spread (SS)

$$SS = \sqrt{\frac{\sum_{k=0}^{N/2} (f_k - SC)^2 |X(k)|^2}{\sum_{k=0}^{N/2} |X(k)|^2}}$$

- Measure of the average spread of the spectrum in relation to its centroid
- Noisy, broadband signals have high SS while tonal sounds show lower SS

Spectral Flatness Measure (SFM)

$$SFM_b = \frac{\sqrt[N_b]{\prod_{k_b} |X(k_b)|^2}}{\frac{1}{N_b} \sum_{k_b} |X(k_b)|^2}, \quad k_b = k_l, k_l + 1, \ldots, k_u$$

where N_b is the number of spectral bins in a subband or
$N_b = k_u - k_l + 1$

- Reflects how "flat" a signal's power spectrum is
- Calculated as the ratio of the geometric mean and the harmonic mean
- Usually computed *per* spectral band (critical bands or bark bands, etc.), thus SFM_b
- Flatness for the whole spectrum is the average of the subband measures

Harmonic Spectral Centroid (HSC)

$$HSC = \frac{\sum_{h=1}^{N_h} f_h A_h}{\sum_{h=1}^{N_h} A_h}$$

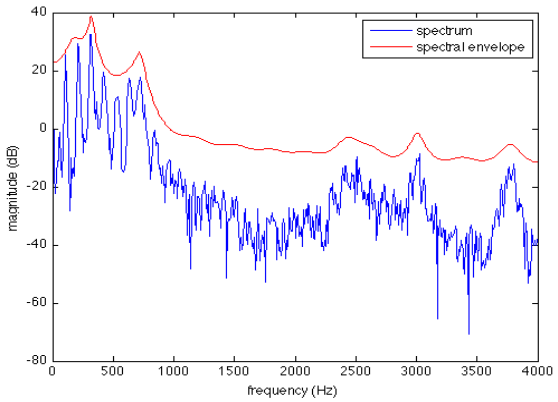
where f_h and A_h are the frequency and the amplitude of the h_{th} harmonic, respectively

- Measure of the amplitude-weighted mean of the harmonic (spectral) peaks of the spectrum
- Compared to SC, HSC focuses only on harmonic (spectral) peaks, which are more musically meaningful in general
- Harmonic spectral spread (HSS) is similarly defined

Spectral Envelope

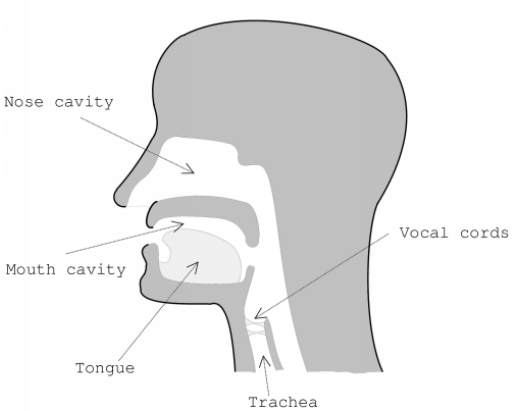
What is the Spectral Envelope?

- Spectral envelope is a smoothed version of the spectrum that preserves its overall shape while neglecting its fine spectral structure



Human Speech System

- Vocal cords act as an oscillator, which generates a spectrally rich *source* signal
- Everything else is *filter*: vocal tract, mouth/nose cavity, tongue
- Thus called "source-filter" model
- These filters define the shape of the spectral envelope

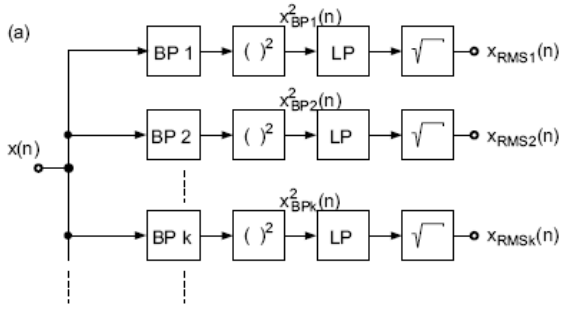


Spectral Envelope Estimation

- A few popular techniques to estimate the spectral envelope
  - Channel vocoder: estimates the amplitude of the signal within several frequency bands
  - Linear prediction: estimates the parameters (or filter coefficients) of a filter that approximates the spectrum
  - Cepstrum analysis: inverse-FFT the log-spectrum and low-pass filters it to obtain the envelope

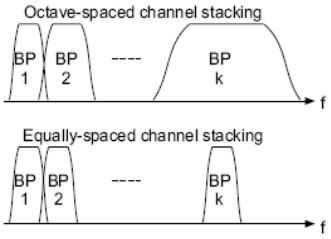
Channel Vocoder (1)

- Filters a signal with a bank of bandpass filters
- Calculates RMS of each bandpassed signal
- The more filters used, the finer spectral envelope estimated



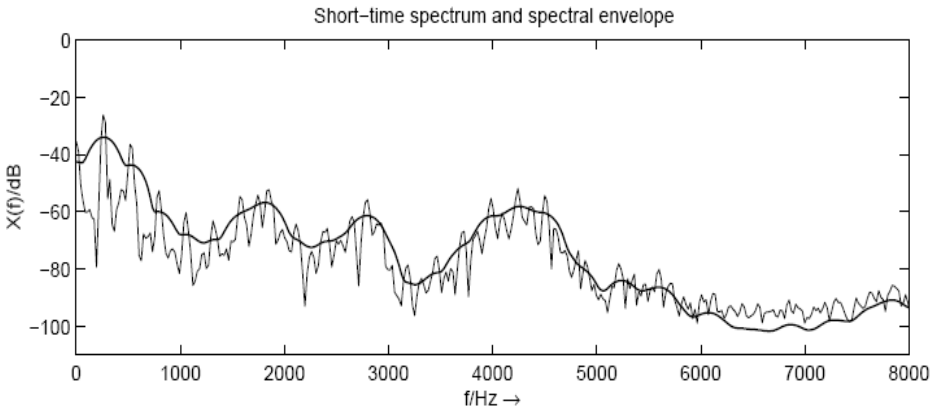
Channel Vocoder (2)

- In the frequency domain, multiply the spectrum with the filters' frequency response and square-root the sum of each filter's output



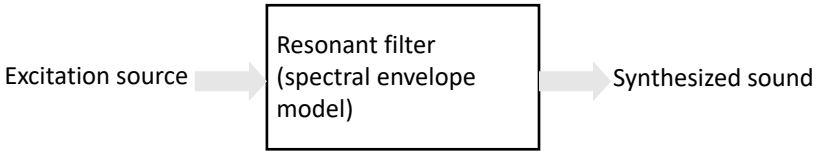
- The filterbank can be either linearly or logarithmically spaced (e.g., constant-Q or Mel-scale filter bank)

Channel Vocoder Example



Linear Predictive Coding (1)

- Linear predictive coding is a source-filter model that approximates the way a sound is generated as an excitation (a pulse train or noise) passing through an all-pole resonant filter



- Widely used in speech and music applications
- Reduces large amount of data (e.g., N samples) to a few filter coefficients while preserving the overall shape

Linear Predictive Coding (2)

- The nth sample $x(n)$ is extrapolated, i.e., predicted by a linear combination of p past samples:

$$x(n) \approx \hat{x}(n) = \sum_{k=1}^{p} a_k x(n-k)$$

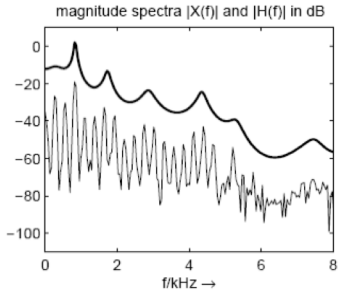
- The residual error is given by

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=1}^{p} a_k x(n-k)$$

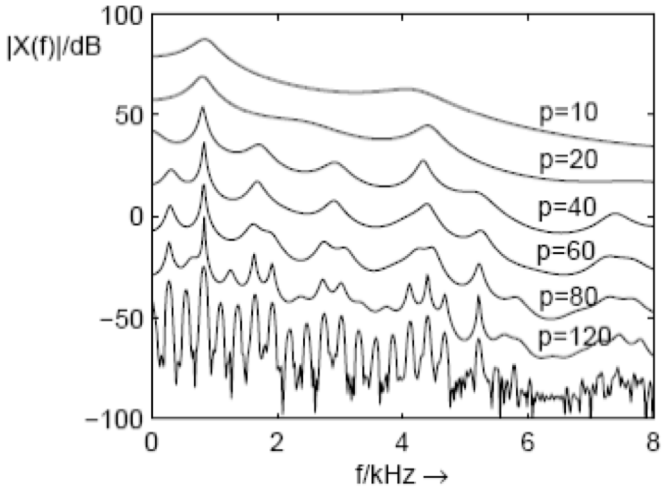
and we want to minimize this error

Linear Predictive Coding (3)

- The parameters a_k's are called linear predictive coefficients (LPCs)
- The filter represented by these coefficients is a resonant filter and its frequency response represents the spectral envelope
- The higher the filter order p, the closer the approximation is to the signal's spectrum



LPC Order and Approximation

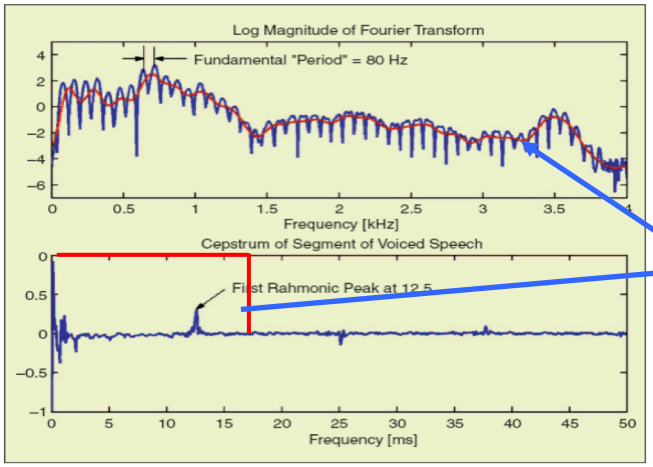


Cepstrum Analysis

- Cepstrum is the result of taking the FFT of the log-spectrum as if it were a *signal*
- Measures the rate of change in different spectral bands
- The name cepstrum was coined by Bogert *et al.* (1963) by reversing the first four letters of the spectrum (similarly for quefrency alanysis and liftering, etc.)
- For a real signal $x(n)$, the real cepstrum is calculated as follows:

$$c_R(n) = IFFT\left(\log(|X(k)|)\right)$$

Cepstrum & Spectral Envelope



By low-pass "liftering" the cepstrum we obtain the spectral envelop the signal

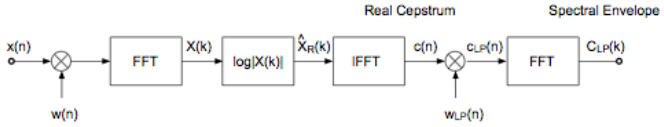
Spectral Envelope Estimation

- Using a low-pass window of the form:

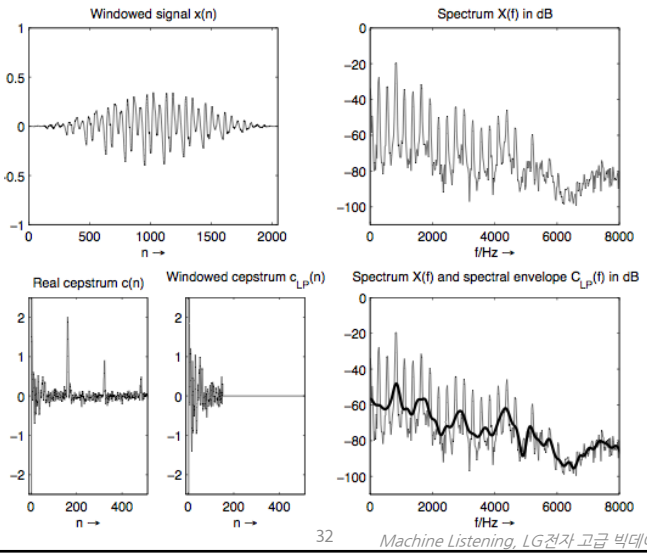
$$\omega_{LP}(n) = \begin{cases} 1 & n = 0, N_1 \\ 2 & 1 \le n \le N_1 \\ 0 & N_1 < n \le N-1 \end{cases}$$

we can low-pass the cepstrum and obtain the spectral envelope by:

$$c_{LP}(n) = c_R(n) \cdot \omega_{LP}(n)$$
$$C_{LP}(k) = FFT[c_{LP}(n)]$$

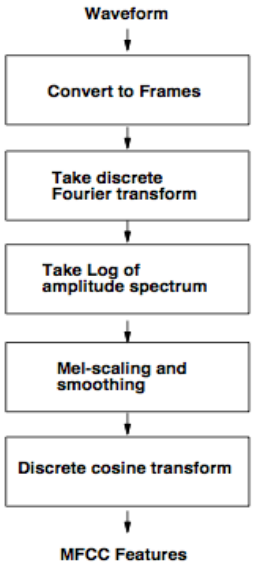


Cepstrum Example



MFCC

- Mel-frequency Cepstral Coefficients (MFCCs) are a variation of cepstrum, motivated by human perception (Logan, 2000)
- Most extensively used in speech and music applications (e.g., speech recognition, genre classification, instrument recognition, etc.), due to its ability to compactly represent the spectral characteristics (just ~13 coefficients)



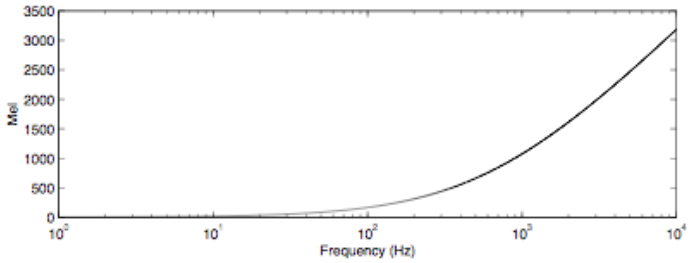
Mel Scale

- The Mel scale is a non-linear perceptual scale of pitches judged to be equidistant
- Approximately linear below 1 kHz and logarithmic above
- 1 kHz corresponds to 1000 Mel (reference point)
- With the Mel scale, a 1000-Mel tone should sound as twice as high as a 500-Mel tone (this is not true with linear frequency Hz)

Mel vs. Linear Frequency

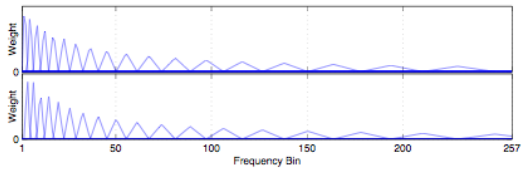
- The relation between Mel and Hz is given by

$$m = 1127.01048 \log(1 + f/700)$$
$$f = 700(e^{m/1127.01048} - 1)$$

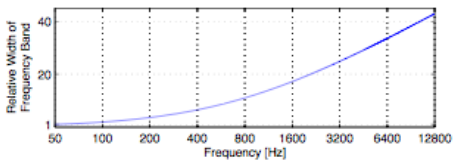


Mel-frequency Spectrum

- To convert a linear spectrum to Mel we can use a filterbank of overlapping triangular windows:



- Such that the width d of each window increases according to the Mel scale, and the height of each triangle is $2/d$

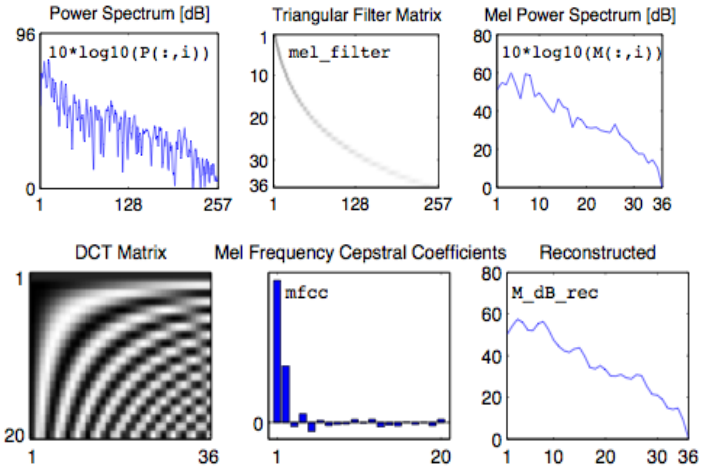


Decorrelation of Mel-scale Spectrum

- The resulting Mel-scale spectral vectors are highly correlated with each other; i.e. highly redundant
- Thus a more efficient representation of the log-spectrum can be obtained by applying a transform that decorrelates those vectors (Rabiner and Juang, 93)
- This decorrelation is commonly approximated by means of the Discrete Cosine Transform (DCT)
- The DCT is similar to a DFT but only for real numbers. It has the property that most of its energy is concentrated on a few initial coefficients (thus effectively compressing the spectral info)

$$X_{DCT}(k) = \sqrt{\frac{2}{N}} \sum_{N=0}^{N-1} x(n) \cos[\frac{\pi}{N}(n + \frac{1}{2})k]$$

Fast Computation of MFCCs



MFCCs roughly model certain characteristics of human auditory perception: the non-linear perception of loudness and frequency and spectral masking (Pampalk, 2006)

Tonal Features

Tonality

- Very important attribute in (Western) tonal music
- Explain the relationship among the tones
- Several musical attributes are closely related with tonality
  - Scale
  - Key
  - Pitch
  - Interval
  - Chord

Tonal Features vs. Spectral Features

- Spectral features
  - Good for describing certain spectral characteristics (e.g., sharpness, noisiness, etc.)
  - Good for representing sonic texture or timbre by capturing overall frequency magnitude response (e.g., LPCs, cepstral coefficients, MFCCs)
  - Not good for tonal analysis: pitch- or tone-relevant information gets lost
- Tonal features retain tonal structure in musical audio
  - Tonal relations
  - Interval relations

Constant-Q Transform

Constant-Q Transform

- In DFT, the center frequency f_k of the frequency bin is given by

$$f_k = \frac{f_s}{N} k, \quad k = 0,1,...,N-1$$

  where f_s is the sampling rate and N is the DFT size
- Therefore, all the frequency bins are linearly spaced
- However, musical scale as well as human hearing mechanism are *logarithmic*
- Brown proposed the constant-Q transform whose frequency resolution conforms to equal-tempered scale (1990)
- Well suited for pitch-related analysis

Constant-Q Transform (cont'd)

- In constant-Q transform the kth spectral frequency is defined as

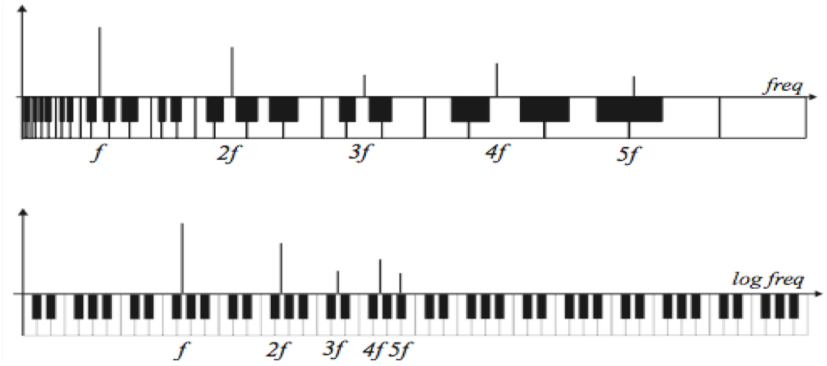
$$f_k = (2^{1/B})^k f_{\min}, \quad k = 0,1,...,N-1,$$

  where B is the number of bins in an octave and f_{min} is the minimum frequency set by user
- It is called "constant-Q" because Q or "quality factor" is constant along the frequency axis, which is defined as

$$Q = \frac{f_k}{f_w}, \quad k = 0,1,...,N-1,$$

  where f_w is the filter width

Linear vs. Logarithmic Frequency



Computation of CQ Transform

- CQ transform can be obtained from the DFT using logarithmically-spaced filterbank

$$X_{cq}(k) = \frac{1}{N(k)} \sum_{n=0}^{N(k)-1} x(n) w(n,k) e^{-j2\pi Q n / N(k)}$$

$$N(k) = f_s Q / f_k$$

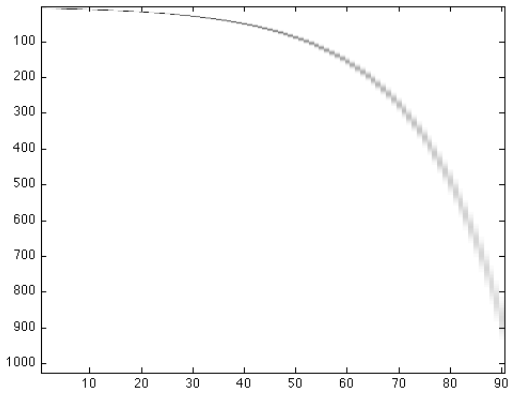
- That uses a variable window length to obtain more resolution at lower frequencies and less at higher (logarithmic distribution of bins in frequency)

Computation of CQT (cont'd)

- CQ transform can be efficiently computed using a *CQ kernel* which is a 2-d matrix that maps the DFT to the CQT

N-point DFT vector	x	NxM CQ kernel matrix	=	M-point CQT vector

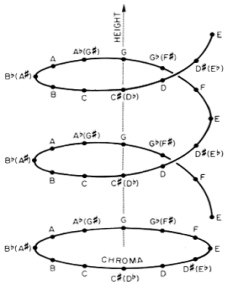
CQ Kernel Matrix

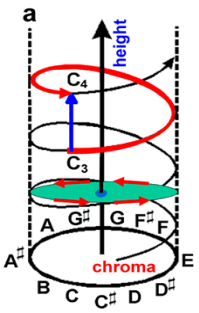


Chroma

Pitch Helix

- The pitch helix is a pitch space where linear pitch is wrapped around a cylinder, thus modeling the special relationship that exists between octave intervals





- Two dimensions
  - Height: naturally organizes *absolute* pitches from low to high
  - Chroma: represents the inherent circularity of pitch (*relative* relationship between pitch classes)

Chroma (aka Pitch Class Profile)

- Good for describing relative pitch relationship, disregarding absolute pitch height
- Very useful for harmony analysis, key and chord, in particular
- A key and/or a chord can be described as a function of its pitch classes
- Almost universal feature for key/chord estimation applications
- Chroma as audio feature first introduced by Fujishima (1999)

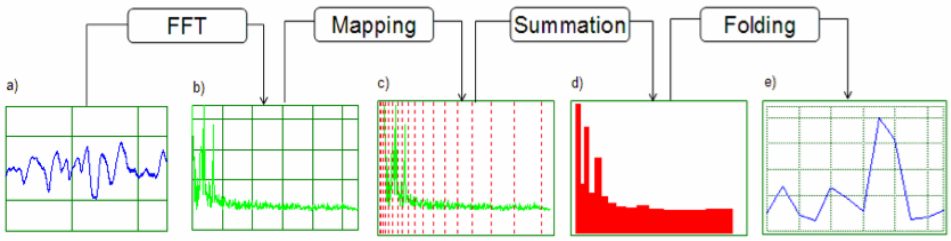
Computation of Chroma

- Easily computed from the constant-Q transform by collapsing it to an octave, or

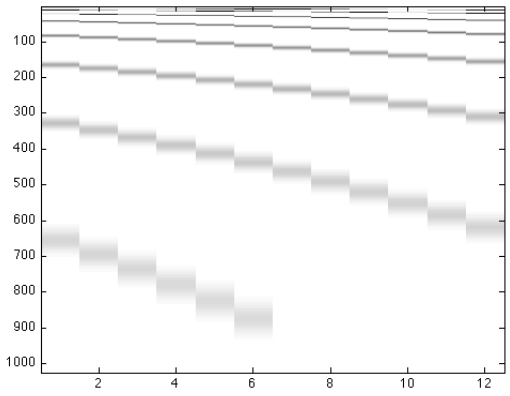
$$Chroma(b) = \sum_{m=0}^{M-1} \left| X_{CQ}(b + mB) \right|,$$

where $X_{CQ}(k)$ is the CQ transform,
M is the total number of octaves of interest,
B is the number of chroma bins in an octave,
and $b = 1,2,\ldots,B$ is the chroma bin index

Computation of Chroma (cont'd)



Chroma Kernel Matrix

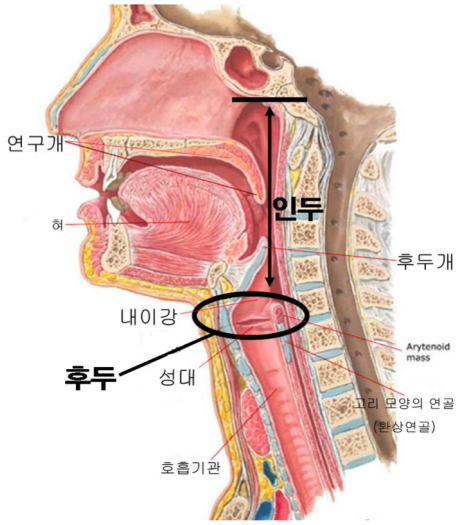


Speech Analysis

Acoustics of speech sounds

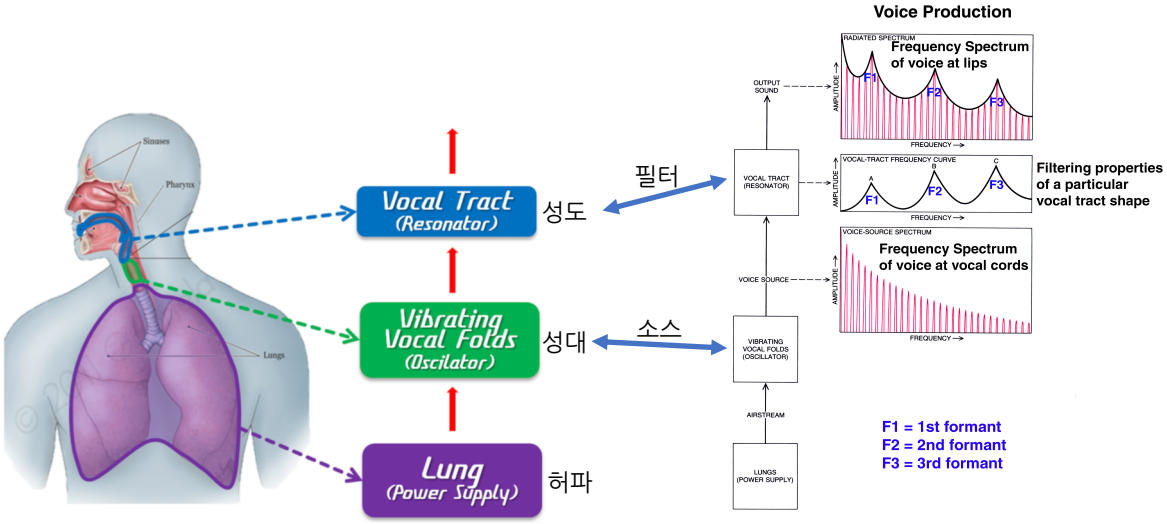
- 음성 발성: 조음 메커니즘
  - 모음
  - 자음
- 음성 지각: 음향학적 패턴으로서의 음성
  - 스펙트럼 큐 spectral cues
  - 포만트 formants
  - 음고 pitch
  - 시간 도메인 큐 time domain cues
  - 포만트 전이 formant transition
  - 성대진동 개시시간 voice onset time
  - 소음 및 무음 noise and silence

Human vocal tract

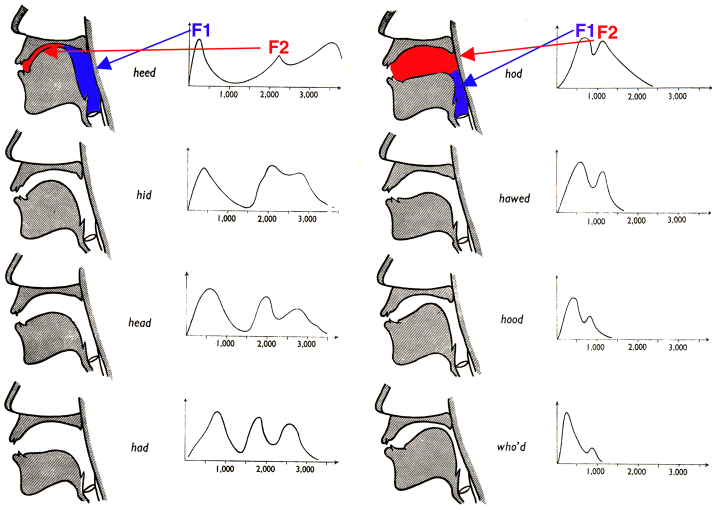


Human articulation

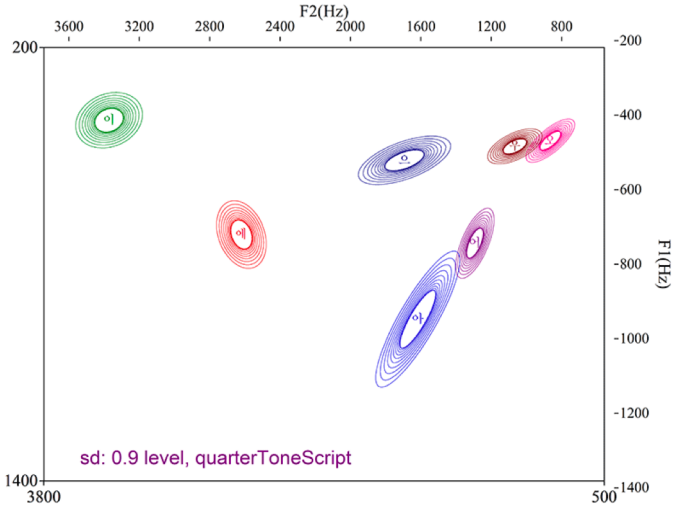
Source-filter model of speech production



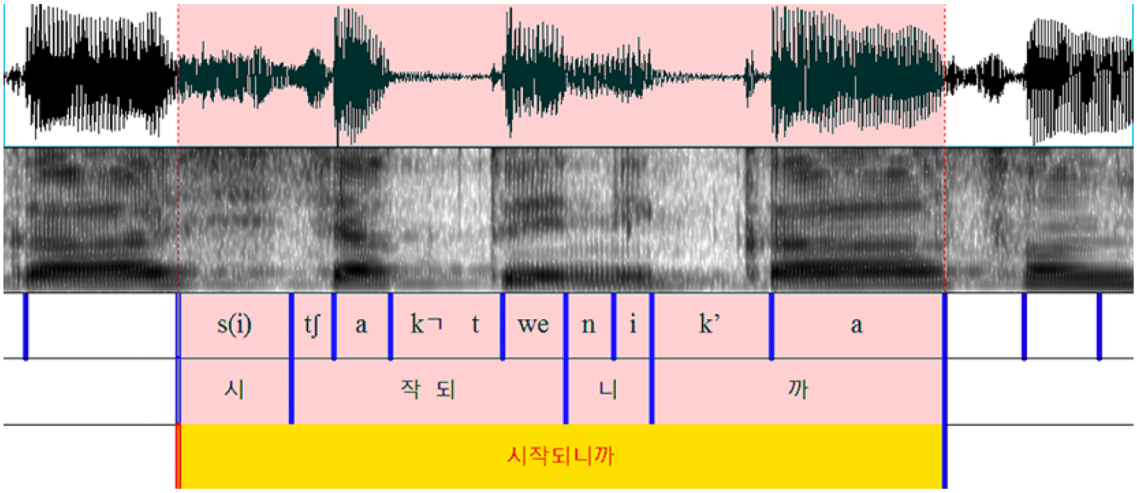
Formants



Formants of Korean vowels

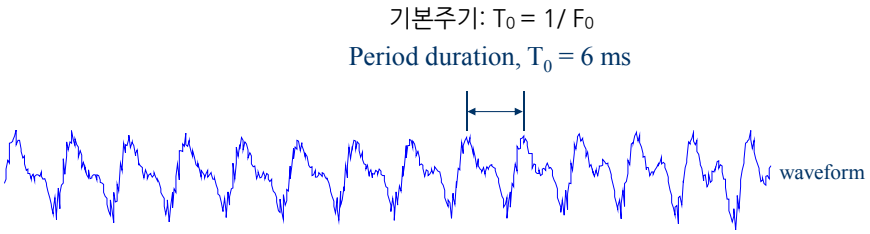


Speech waveform and spectrogram

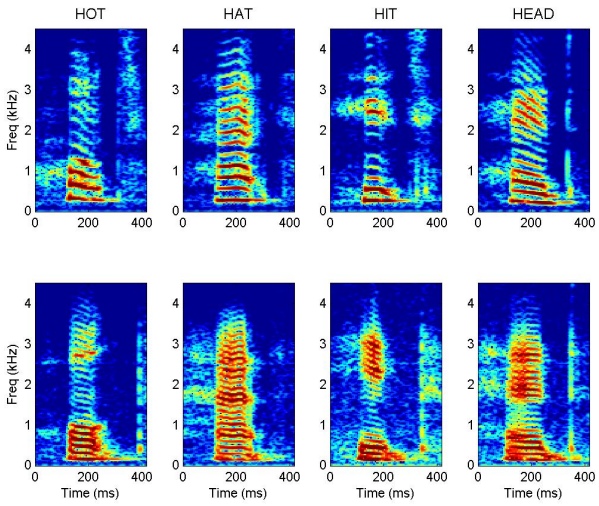


Pitch

- 성대의 진동수에 의해 결정되는 기본 주파수(F_0; fundamental frequency)
- 남성의 평균 음고: 80~200 Hz
- 여성의 평균 음고: ~400 Hz



Pitch and its harmonics



Hearing through eyes? McGurk effect

https://www.youtube.com/watch?v=aFPtc8BVdJk

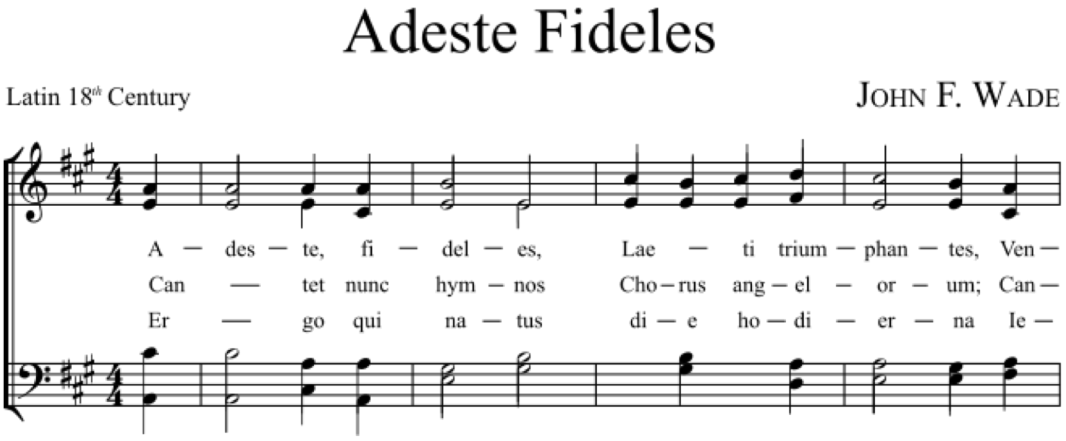
Music Analysis

What is music?

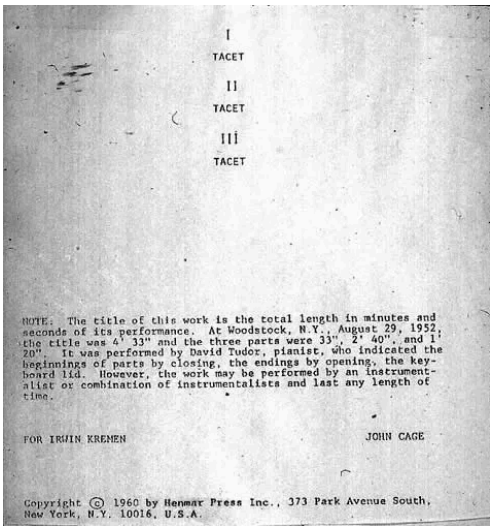
What is music?

- 목소리 또는 악기(또는 둘 다)를 이용하여 형식, 조화 및 감정 표현의 아름다움을 만들어내는 방식으로 결합된 소리 - Oxford Dictionaries
- 음악은 말로 표현할 수 없고 침묵할 수 없는 것을 표현한다 - Victor Hugo
- 음악은 세고 있다는 것을 모르는 영혼의 숨겨진 산술 운동이다 - Gottfried Wilhelm von Leibniz
- 음악은 정돈된 소리이다 - Edgar Varese

Music is organized sound.



4’ 33” by John Cage


I
TACET
II
TACET
III
TACET

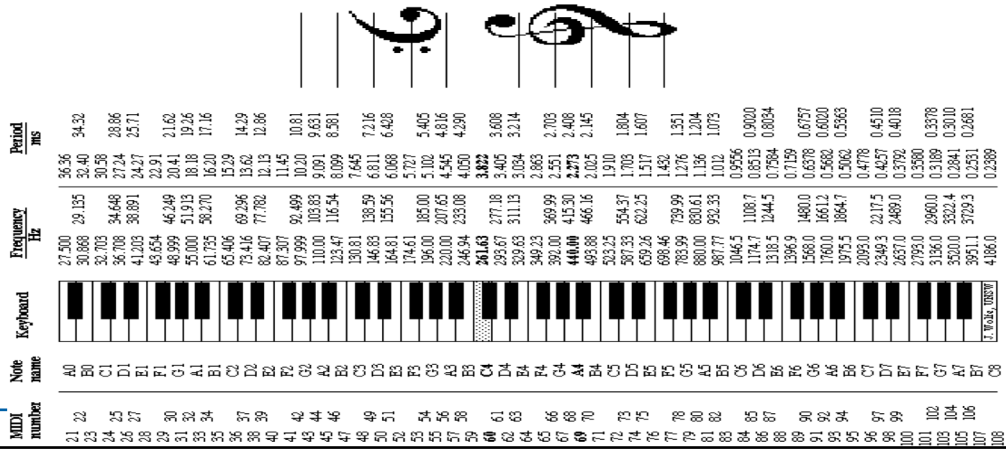
NOTE: The title of this work is the total length in minutes and seconds of its performance. At Woodstock, N.Y., August 29, 1952, the title was 4' 33" and the three parts were 33", 2' 40", and 1' 20". It was performed by David Tudor, pianist, who indicated the beginnings of parts by closing, the endings by opening, the keyboard lid. However, the work may be performed by an instrumentalist or combination of instrumentalists and last any length of time.

FOR IRWIN KREMEN

JOHN CAGE

Physical vs. musical properties

- 진폭(amplitude) vs. 음량(loudness or velocity)
- 주파수(frequency) vs. 음고(pitch)

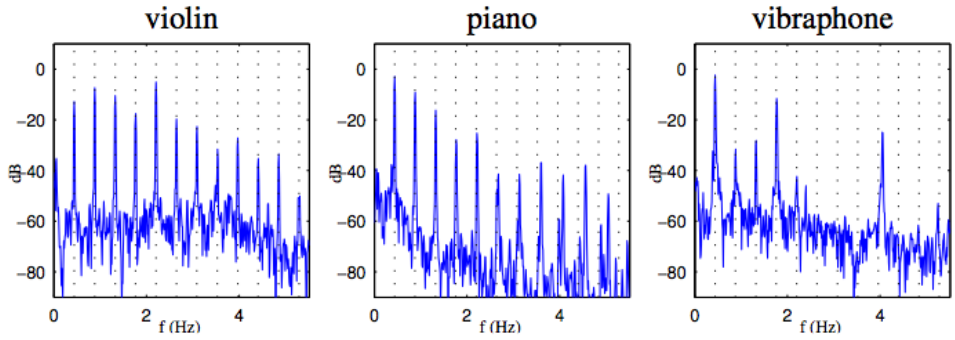


Musical attributes

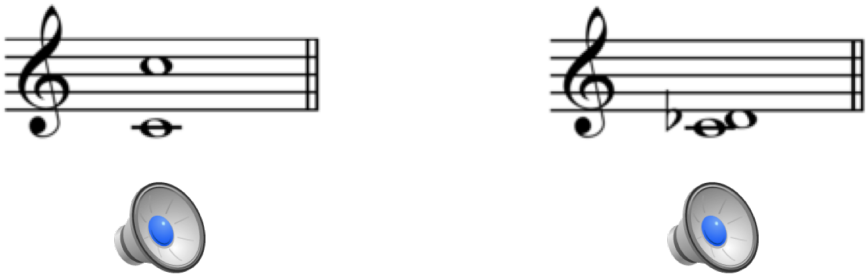
- 멜로디(melody): 주어진 문화적 관례 및 제약 조건에 따라 순차적으로 배열된 음고가 있는 소리
- 화성(harmony): 두개 이상의 음을 동시에 결합하여 코드를 생성하고, 연속적으로 코드 진행을 생성
- 리듬(rhythm): 규정된 강약 요소의 연속에 의해 표현되는 움직임

Harmonics

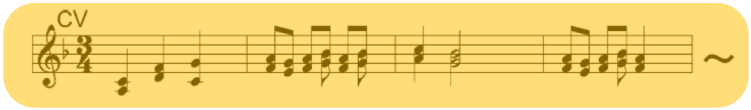
- 배음: 대부분의 악기(사람 목소리 포함)는 음고에 해당하는 주파수 뿐만 아니라 이의 정수배에 해당하는 주파수 성분을 생성한다.

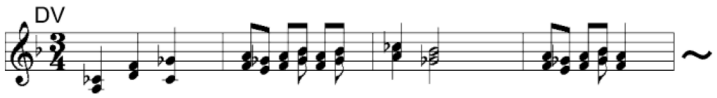


Consonance vs. dissonance

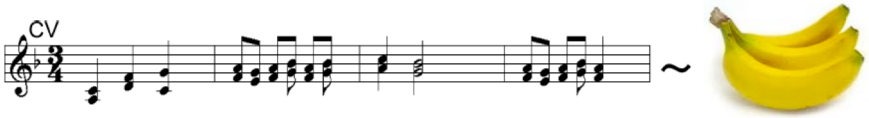


Preference for consonance





Why we prefer consonance

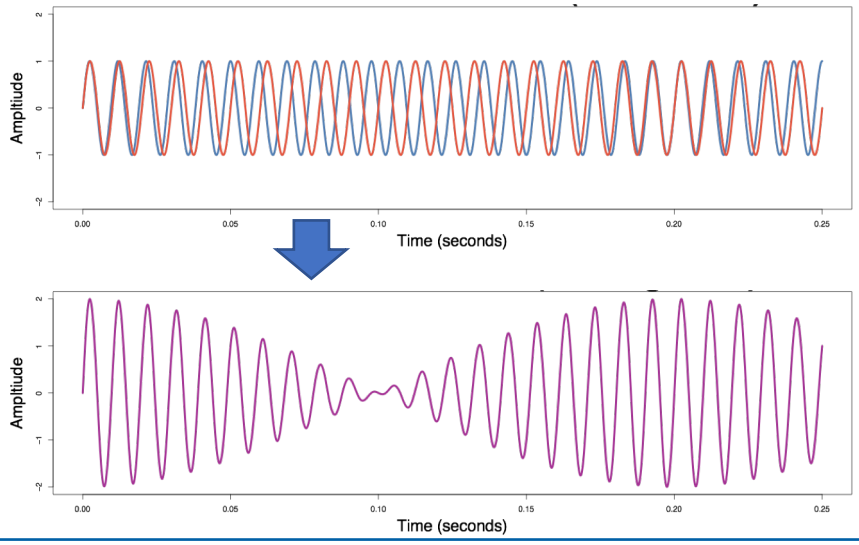


Acoustic dissonance: beating

- 불협화음을 만들어내는 물리적 특성
- 맥놀이(beating): 주파수가 가까운 두 신호로 인해 생기는 진폭 변조 현상(=roughness)

Helmholtz (1821-1894)

Beating by interference

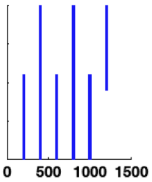


두개의 순음:
100Hz & 105Hz

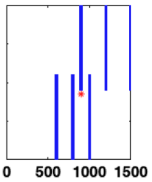
맥놀이:
100Hz + 105Hz

Intervals of two complex tones

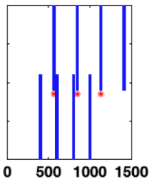
옥타브: 도-도
Octave



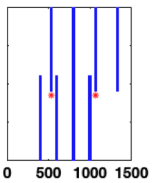
5도: 도-솔
Fifth



트라이톤: 도-파#
Tritone

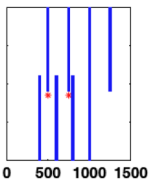


Fourth



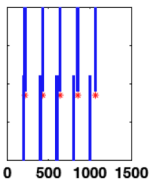
4도: 도-파

Major Third



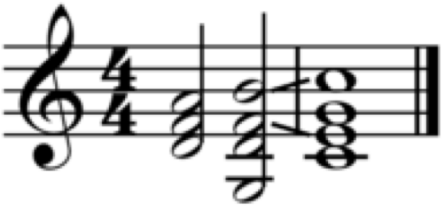
장3도: 도-미

Minor Second



단2도: 도-도#

Cadence

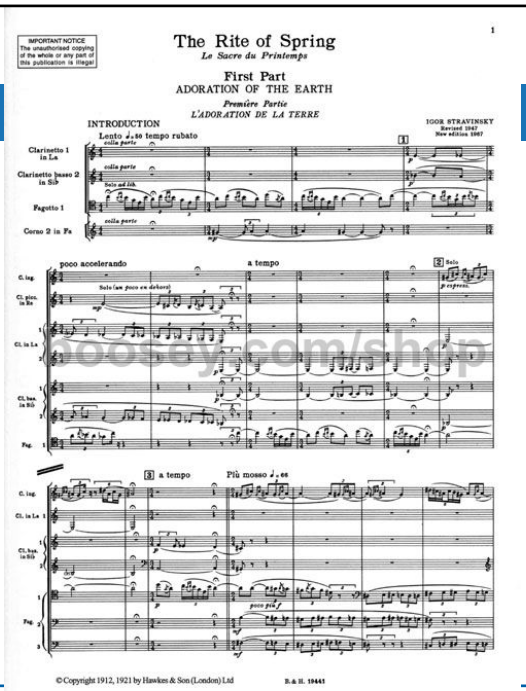


ii - V - I

SNU Convergence

MUSIC & AUDIO RESEARCH GROUP

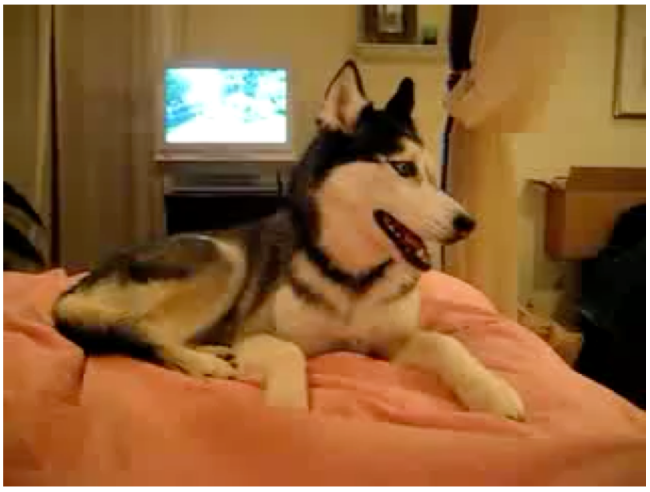
Evolution of music



SNU Convergence

MUSIC & AUDIO RESEARCH GROUP

Mishika: the talking dog



Dancing Parrot

https://www.youtube.com/watch?v=gTl1asCDOgs

Music Analysis

- Rhythmic analysis
  - Onset detection
  - Tempo estimation; Beat tracking
- Tonal analysis
  - Pitch estimation
  - Key estimation
  - Chord recognition

SNU Convergence

MUSIC & AUDIO RESEARCH GROUP

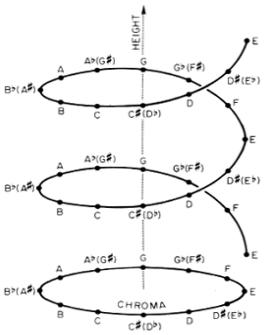
Musical Pitch

SNU Convergence

MUSIC & AUDIO RESEARCH GROUP

Pitch

- Pitch can be defined as the perceived fundamental frequency (f_0) of a sound (both represented in Hertz)
- It allows us to intuitively organize sounds on a scale from low to high



SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Pitch & Hamonicity

- Pitch is not a well defined attribute for all sounds
- Fundamental frequency can only be defined for harmonic or nearly harmonic sounds
- Broadly-speaking, western musical instruments can be divided between those that produce harmonic (or nearly harmonic) sounds and those that do not produce harmonic sounds

SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Hamonicity in Musical Instruments

Table 1: Western musical instruments which do or do not produce harmonic sounds.

Produced sounds	Instrument family	Instruments involved
Harmonic	String instruments	Piano, guitars, bowed strings (violin etc.)
	Reed instruments	Clarinets, saxophones, oboe, bassoon
	Brass instruments	Trumpet, trombone, tuba, english/french horn
	Flutes	Flute, bass flute, piccolo, organ
	Pipe organs	Flue pipes and reed pipes
	Human voice (singing)	Voiced phonemes
Not harmonic	Mallet percussions	Marimba, xylophone, vibraphone, glockenspiel
	Drums	Kettle drums, tom-toms, snare drums, cymbals

Harmonic Sounds

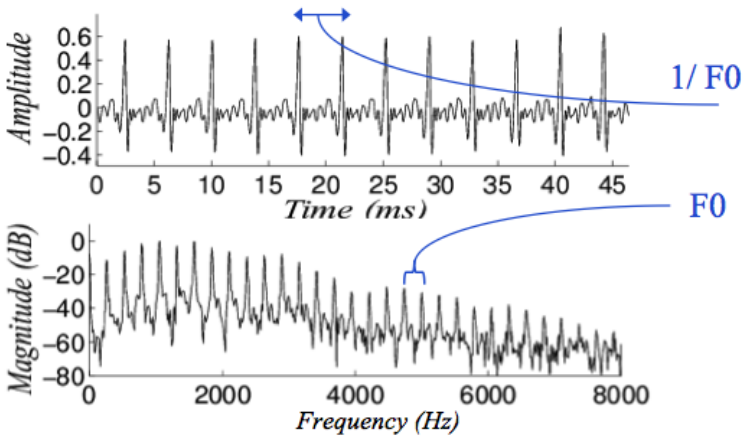
- Most common (Western) musical instruments produce harmonic sounds that contains partials at frequencies

$$f_k = kf_0,$$

where $k = 1 \ldots K$ and f_0 is the fundamental frequency (i.e. pitch)

SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

Harmonic Sounds - Example



Inharmonicity in Harmonic Sounds

- Some "harmonic" instruments are, in fact, slightly inharmonic
- This is the case for plucked or struck string instruments (e.g. piano)
- In these cases kth partial's frequency is given as:

$$f_k = k f_0 \sqrt{1 + \beta k^2},$$

$$\beta = \frac{\pi^3 E d^4}{64 T L^2},$$

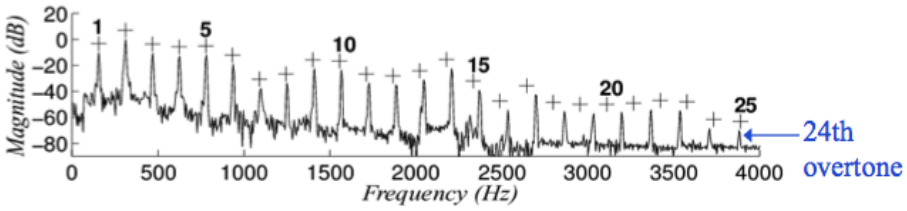
where β is the inharmonicity coefficient, E is the inverse elasticity of the medium, d is the diameter, T is the tension and L is the length of the string

- Inharmonicity is larger for steel strings than for nylon strings
- The pitch is still perceived to be equal to f_0

SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

Example - Piano Sound

Spectrum of a piano sound. Ideal partial frequencies marked with +

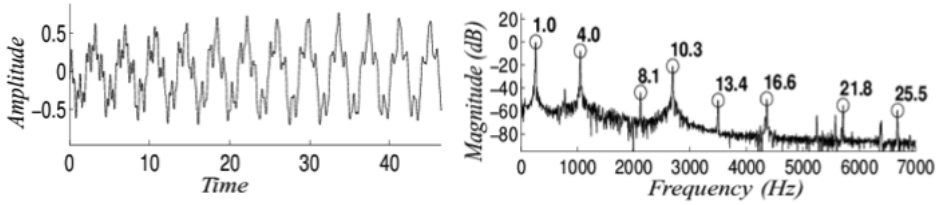


SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Inharmonic Sounds

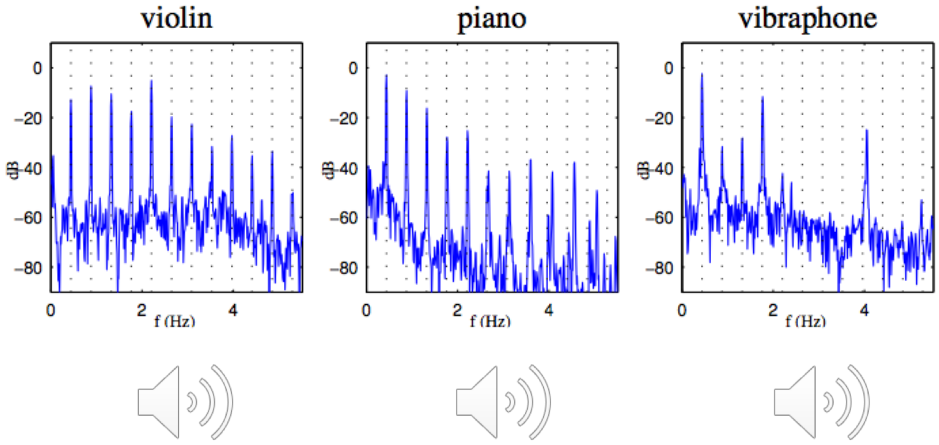
Waveform and spectrum of a vibraphone sound. Pitched but not harmonic



SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Harmonic vs. Inharmonic Sounds



SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Monophonic vs. Polyphonic

- Two broad cases for pitch estimation:
  - Monophonic music: when only one pitch is present at the time
  - Polyphonic music: when two or more pitches exist simultaneously
- Main difference with speech signals: although in music pitch information is more stable, the polyphony makes the analysis and processing of tonal contents more complex
- Polyphony: the probability of different peaks lying in the same frequency bin increases

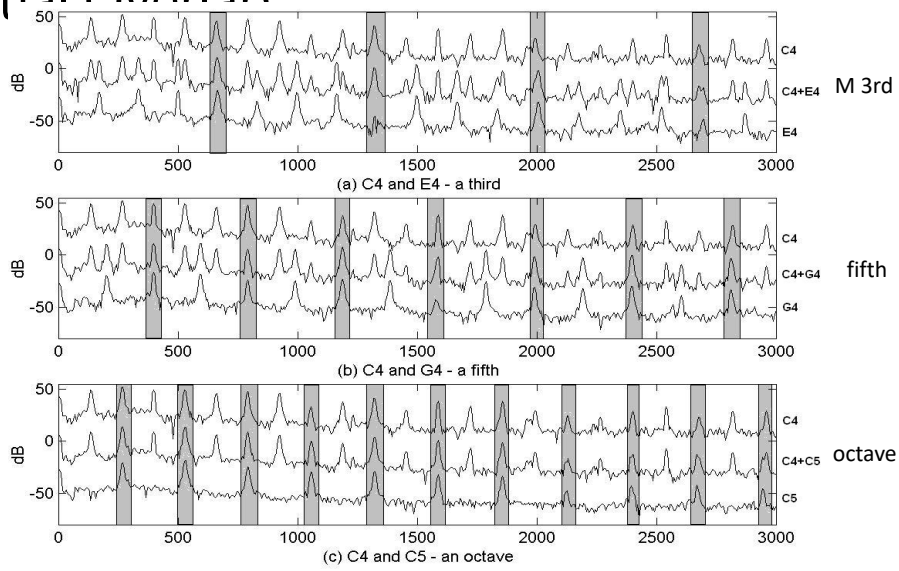
SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Harmonicity in Polyphony

- Harmonicity when two harmonic instruments are present:
  - Fundamental frequency ratio: $f_1/f_2 = a/b$, where a and b are positive integers
  - Then every b^{th} partial of Sound 1 overlaps every a^{th} partial of Sound 2
  - Common intervals in western music: octaves (1:2), fifths (2:3), fourths (3:4), major/minor thirds (4:5/5:6), etc.

Partial Overlap



SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

Pitch Estimation

Pitch Estimation Methods

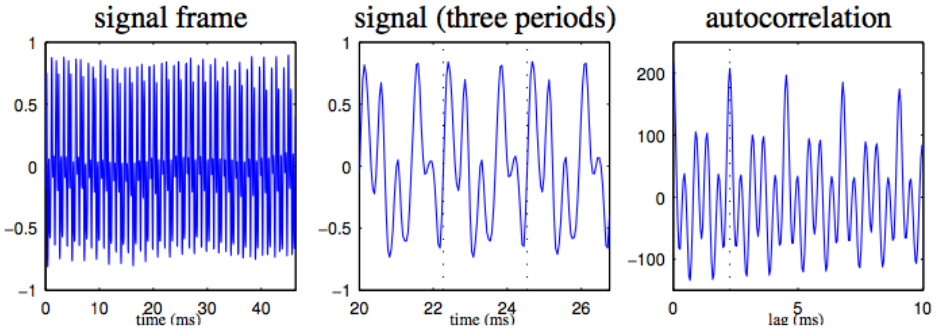
- Three different categories
  - Time-domain periodicity analysis methods
  - Frequency-domain periodicity analysis methods
  - Auditory model-based methods
- These methodologies have been used both in the monophonic and polyphonic case
- They exploit different sets of information in the signal

SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

Time-domain Methods

Autocorrelation Function (1)

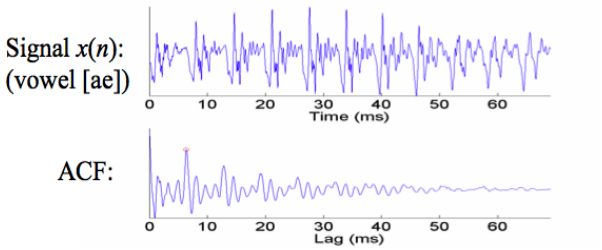
- The autocorrelation function (ACF), useful in identifying repeating patterns in a signal segment $x(n)$, is defined as: $r(\tau) = \frac{1}{N}\sum_{n=0}^{N-\tau-1} x(n)x(n+\tau)$



SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

Autocorrelation Function (2)

- For monophonic signals, the maximum peak of $r(\tau)$ for positive lags (not including the zero lag value) is usually considered to be at the fundamental period $\tau_0 = 1/f_0$



- In many cases other peaks appear at multiples and sub-multiples of the fundamental frequency, sometimes with competing amplitudes

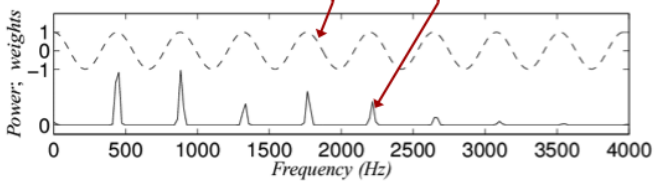
Autocorrelation Function (3)

- The ACF can also be calculated by means of the DFT as:

$$r(\tau) = IFFT\left\{\left|FFT[x(n)]\right|^2\right\}$$ --> much faster!

- This is equivalent to calculating the correlation between the power spectrum and a sinusoidal spectral template:

$$r(\tau) = \frac{1}{K}\sum_{k=0}^{K-1}\left[\cos\left(\frac{2\pi\tau k}{K}\right)X(k)^2\right]$$

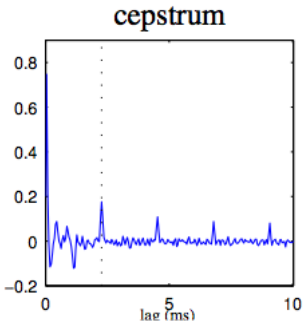


Cepstrum Analysis

- From this perspective, cepstrum analysis is actually closely-related to the ACF:

$$c(\tau) = IFFT\left\{\log\left(\left|FFT[x(n)]\right|\right)\right\}$$

- It only implies a change of the $|\,|^2$ operand for a $\log(|\,|)$ operand
- Low power partials are emphasized while becoming more sensitive to noise



Difference Function (1)

- The ACF tends to decrease for large values of τ. This for large τ_0 results in f_0/m type errors (inverse octave)
- An alternative method, known as YIN (de Cheveigne, 2002), considers the difference function of the signal:

$$d(\tau) = \sum_{n=0}^{N-1} (x(n) - x(n+\tau))^2$$

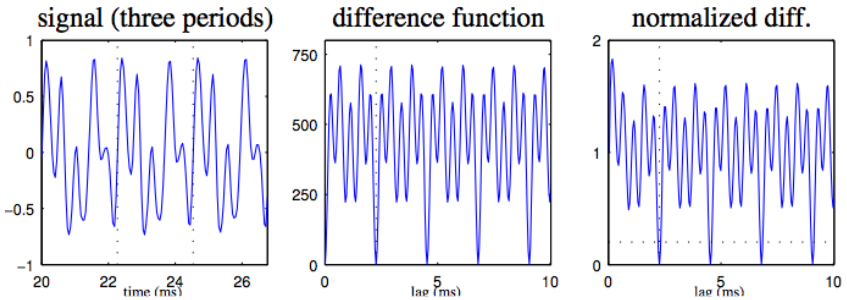
- For periodic signals, this function becomes zero (cancels itself) at zero-lag, τ_0 and its multiples
- To avoid zero-lag bias, a normalized functions is used:

$$d'(\tau) = \begin{cases} 1 & \tau = 0 \\ d(\tau) \Big/ \left[(1/\tau) \sum_{n=1}^{\tau} d(n) \right] & otherwise \end{cases}$$

SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

Difference Function (2)

- The first minimum of d' below a certain pre-specified threshold indicates the position of τ_0
- Exact pitch estimation is achieved using quadratic interpolation
- YIN demonstrates robustness when compared to ACF methods



Extension to Multipitch Estimation

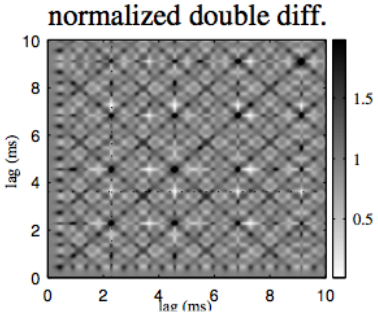
- To extend YIN to polyphonies we consider the double difference function:

$$d(\tau_1,\tau_2)=\sum_{n=0}^{N-1}(x(n)-x(n+\tau_1)-x(n+\tau_2)+x(n+\tau_1+\tau_2))^2$$

- Normalized as:

$$d'(\tau_1,\tau_2)=\frac{d(\tau_1,\tau_2)}{\frac{1}{\tau_1}\sum_{n=1}^{\tau_1}d(n,\tau_2)}$$

$$d''(\tau_1,\tau_2)=\frac{d'(\tau_1,\tau_2)}{\frac{1}{\tau_2}\sum_{n=1}^{\tau_2}d'(\tau_1,n)}$$



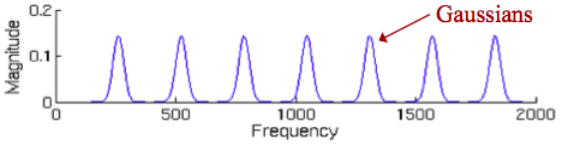
- The first minimum of d'' below a threshold is $(1/f_{0,1}, 1/f_{0,2})$
- This can be extended to multiple difference functions

SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

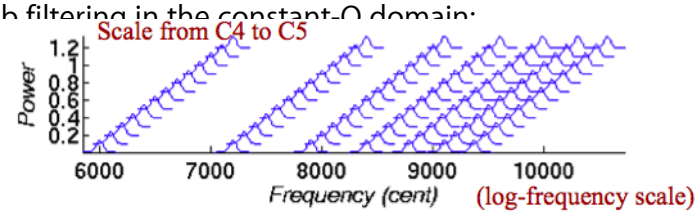
Frequency-domain Methods

Pattern Matching

- A number of approaches are based on the idea of matching patterns in the spectral domain
- Examples include Maximum-likelihood based matching (Doval and Rodet, 1991)



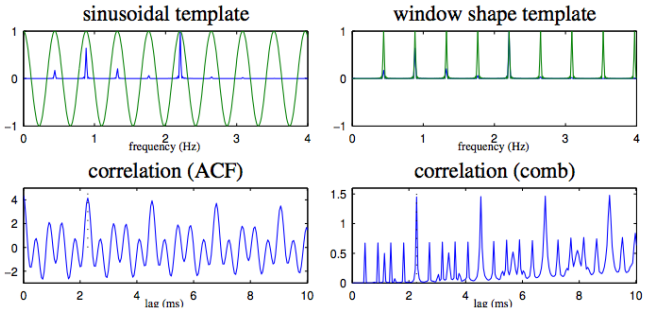
- Brown's comb filtering in the constant-Q domain:



SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

Comb Filtering

- "Comb-filtering" approaches measure the correlation between the magnitude spectrum and a harmonic pattern, much as ACF does, but more realistic



- Furthermore, windows can be spaced according to inharmonicity

Spectral Autocorrelation

- We can exploit the fact that harmonic sounds have a periodic spectrum, thus we can calculate the spectral autocorrelation:

$$\hat{r}(m) = \frac{2}{N} \sum_{k=0}^{N/2-m-1} |X(k)||X(k+m)|$$

- It has the advantage that the spectrum can be shifted without affecting the estimation (since it is independent of location)
- Works well for slightly inharmonic sounds, since intervals (although not always exactly the same) are more robust than locations

SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

Harmonic Product Spectrum (HPS)

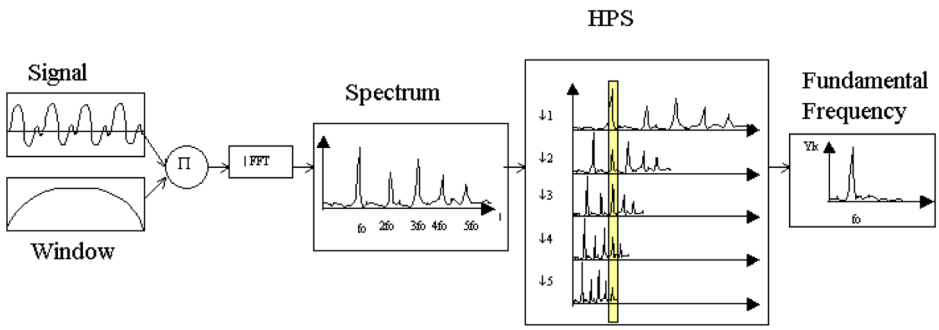
- Harmonic product spectrum (HPS) is used to estimate the fundamental frequency in monophonic speech or music signals, which are mostly harmonic
- HPS is obtained by multiplying the original magnitude spectrum and its decimated spectra by an integer number:

$$\mathrm{HPS}(k) = \prod_{m=1}^{M} |X(mk)|,$$

$$f_0 = \arg\max_k \{\mathrm{HPS}(k)\},$$

where $\mathrm{HPS}(k)$ is the harmonic product spectrum, $X(k)$ is the DFT of the signal, M is the number of harmonics to be considered, and f_0 is the estimated fundamental frequency

Calculating the HPS



SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

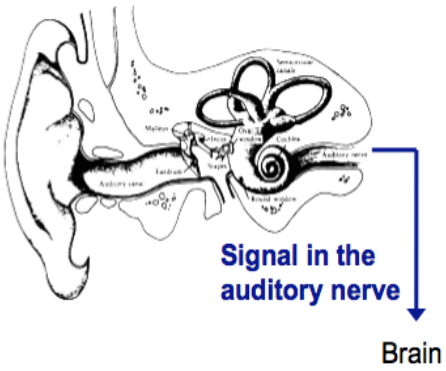
Auditory Model-based Method

SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Auditory Models

- Try to model human's hearing mechanism
- Cochlea: performs spectral analysis (typically modeled as a filterbank of linear, overlapped, band-pass filters)
- Hair cells: transform mechanical movement into neural impulses (modeled as half-way rectification, compression and low-pass filtering)
- What the brain does is not directly observable and therefore more controversial: channel periodicity analysis and data fusion across channels

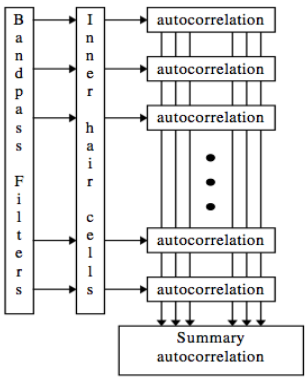


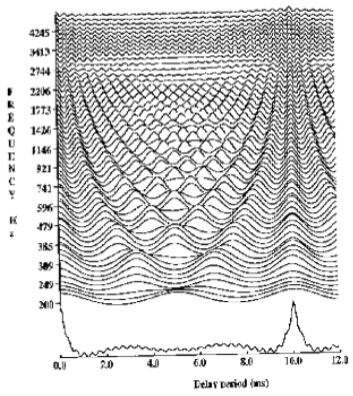
SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Summary Autocorrelation Function (1)

- We can estimate the sub-band periodicities using the ACF
- The resulting representation is known as a *correlogram*





Summary Autocorrelation Function (2)

- The sub-band information can be integrated to give a summary autocorrelation function (SACF) by summing periodogram or correlogram across all frequency bands, as proposed by Meddis (1991)

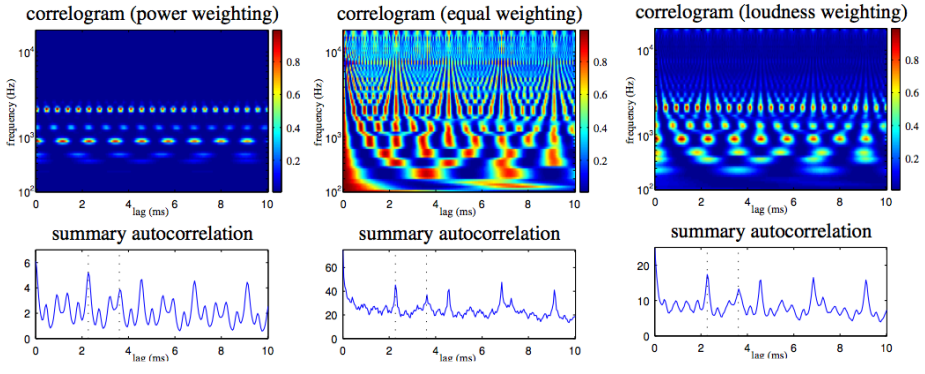
$$r_n(\tau) = \sum_{b=1}^{B} r_{bn}(\tau)$$

- Thus peaks in the SACF represent fundamental periods in the sound
- This approach is robust against band-limited noise

SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

Types of Filterbanks

- The design of the filterbank has an impact on the outlook of the correlogram and the SACF

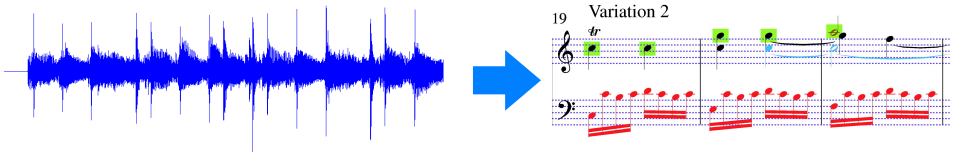


SNU Convergence

MUSIC & AUDIO RESEARCH GROUP

Automatic Music Transcription

- AMT is the process of automatically turning a recorded audio signal into a score-like music representation (e.g. MIDI, thus wav2midi)
- This narrow definition is in no way comparable to music transcription, but rather limited to the estimation of onset times, durations, velocity and (above all) pitches of the notes being played



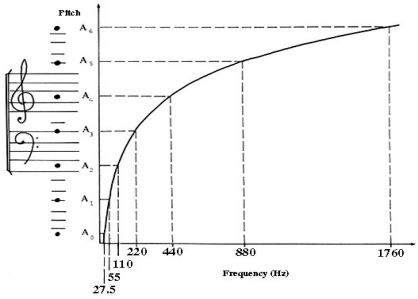
- Increased level of complexity: grouping notes into instrumental sources or even melodic lines

SNU Convergence

MUSIC & AUDIO RESEARCH GROUP

Labeling Pitch

- To create a score-like representation we need to represent pitch in terms of the (western) musical scale



- Even when a f_0 is properly estimated, its quantization into musical note values (e.g. scientific pitch notation, MIDI note number) is lossy

$$MIDI = 69 + 12 \cdot \log_2(\frac{f_0}{440})$$

Questions?