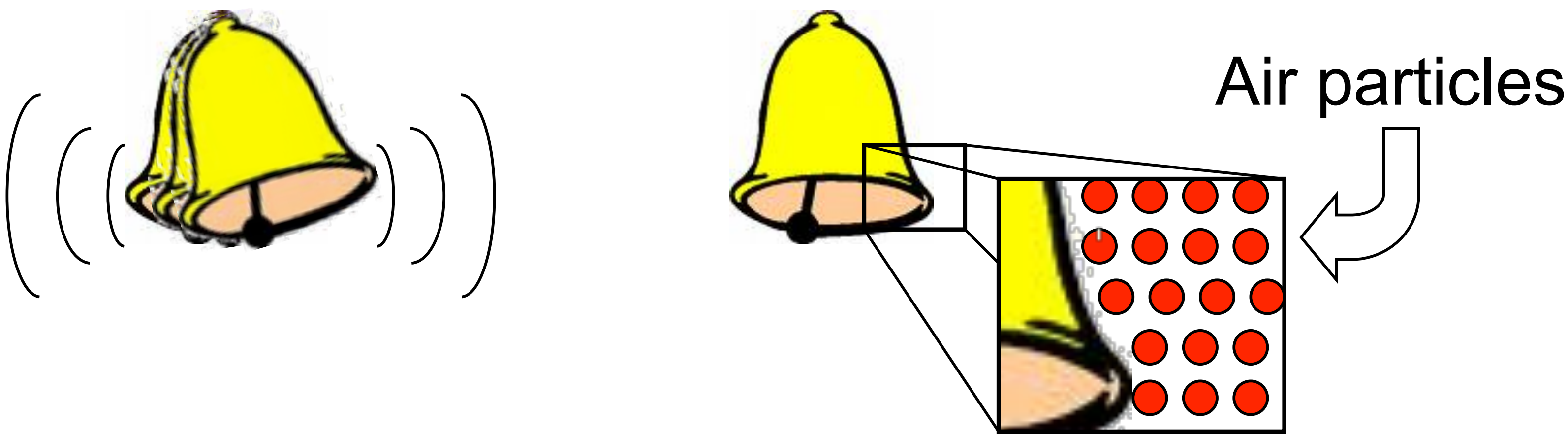
Fundamentals of Sound and Time-Frequency Representations

Juan Pablo Bello
EL9173 Selected Topics in Signal Processing: Audio Content Analysis
NYU Poly

Sound

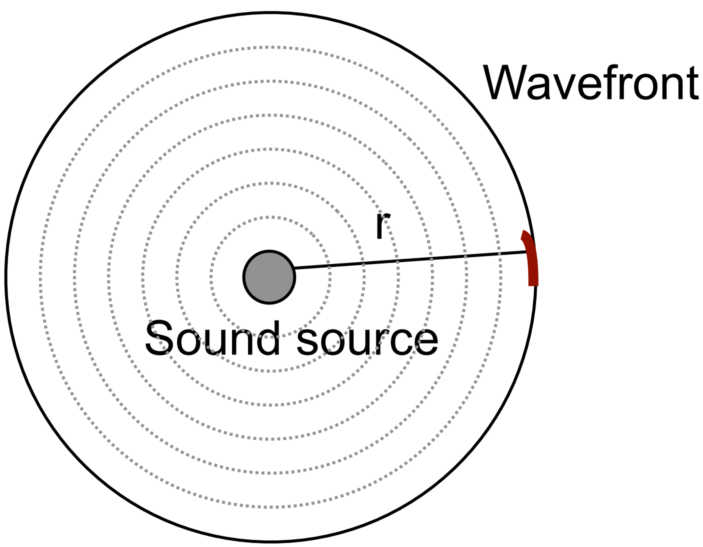
- Sound is produced by a vibrating source that causes the matter around it to move.
- No sound is produced in a vacuum - Matter (air, water, earth) must be present



- The vibration of the source causes it to push/pull its neighboring particles, which in turn push/pull its neighbors and so on.
- Pushes increase the air pressure (compression) while pulls decrease the air pressure (rarefaction)
- The vibration sends a wave of pressure fluctuation through the air

Sound power and intensity

- A source (e.g. bell) vibrates when a force (e.g. striking hammer) is applied to it.
- The force applied and the resulting movement characterize the work performed by the source ($W = F \times \Delta s$)
- Power ($P = W/t$) is the rate at which work is performed and is measured in watts.
- An omnidirectional sound source produces a 3-D longitudinal wave. The resulting wavefront is defined by the surface of a sphere ($S = 4\pi r2$), where r is the distance from the source.



The original power is distributed on the surface of the wavefront.
As r increases, the power per unit area (intensity) decreases: $I = P/S$

Intensity and SPL

- The effect of sound power on its surroundings can be measured in sound pressure levels (SPL) - much as temperature in a room relates to the energy produced by a heater.
- Both intensity (Watts/area) and sound pressure (Newtons/area) are usually represented using decibels (dB)
- dB are based on the logarithm of the ratio between two powers, thus describing how they compare ($dB = 10\log_{10}(P1/P2)$).
- This can be applied to other measures (amplitude, SPL, voltage), as long as their relationship to power is taken into account.
- In the case of intensity and SPL, the denominator of the ratio is a reference value, defined according to the quietest sound perceivable by the average person.
- Thus by convention, 0 dB corresponds to SPL = 2×10^{-5} N/m^2 or I = 10^{-12} watt/m^2

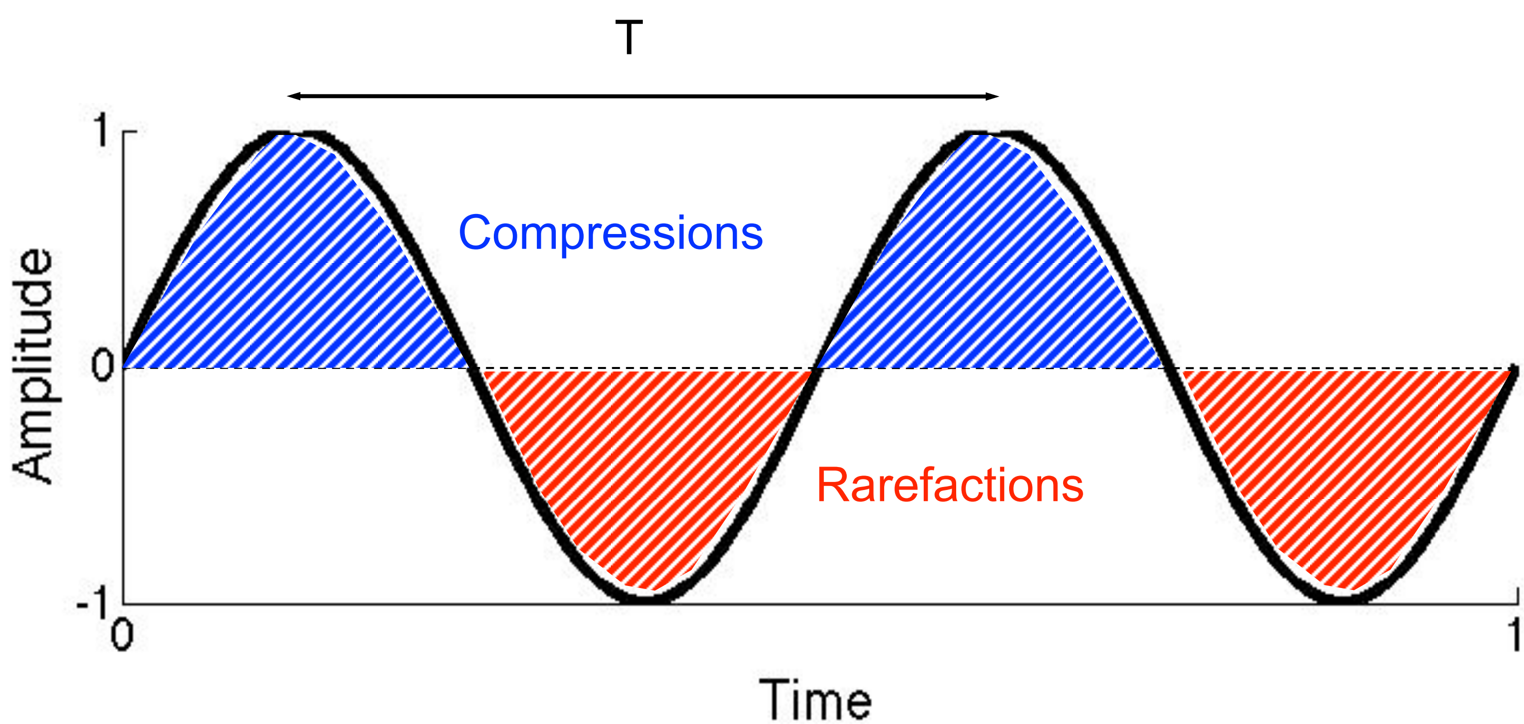
Sound waves (1)

- In sound wave motion air particles do not travel, they oscillate around a point in space.
- The rate of this oscillation is known as the frequency (f) of the sound wave and is denoted in cycles per second (cps) or hertz (Hz).
- The amount of compression/rarefaction of the air is the amplitude (A) of the sound wave.
- The distance between consecutive peaks of compression or rarefaction is the wavelength of the sound wave (λ)

Sound waves (2)

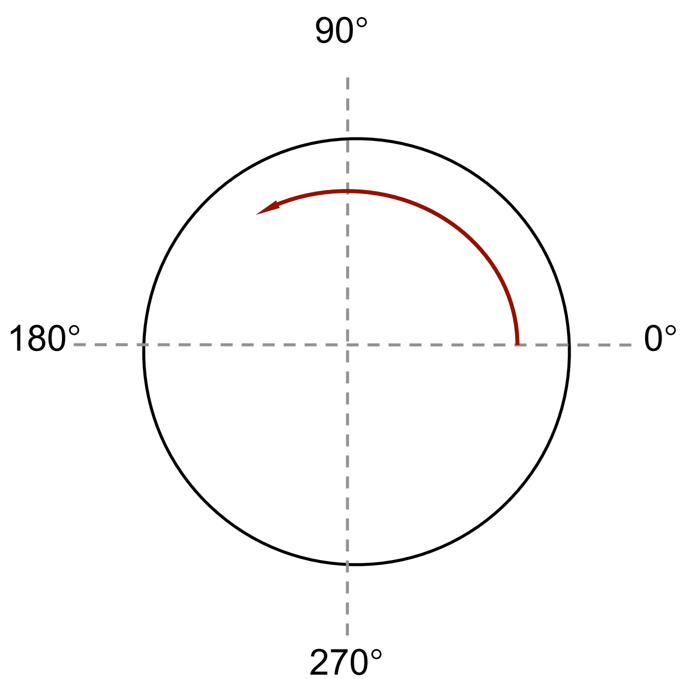
- If the frequency of the oscillation is stable, then the sound wave is periodic (with period T, and frequency f = 1/T)

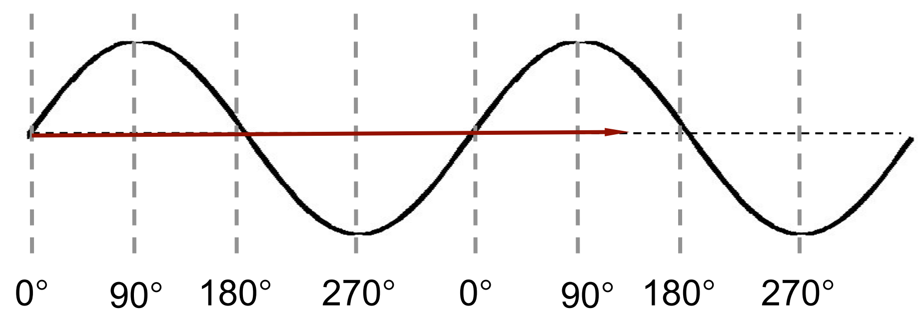
- The simplest periodic wave is a sinusoid: $x(t) = A \cdot sin(2\pi ft + \theta)$



Sound waves (3)

- Phase is a temporal offset, defined in terms of a fraction (degrees) of a complete cycle of the periodic wave.

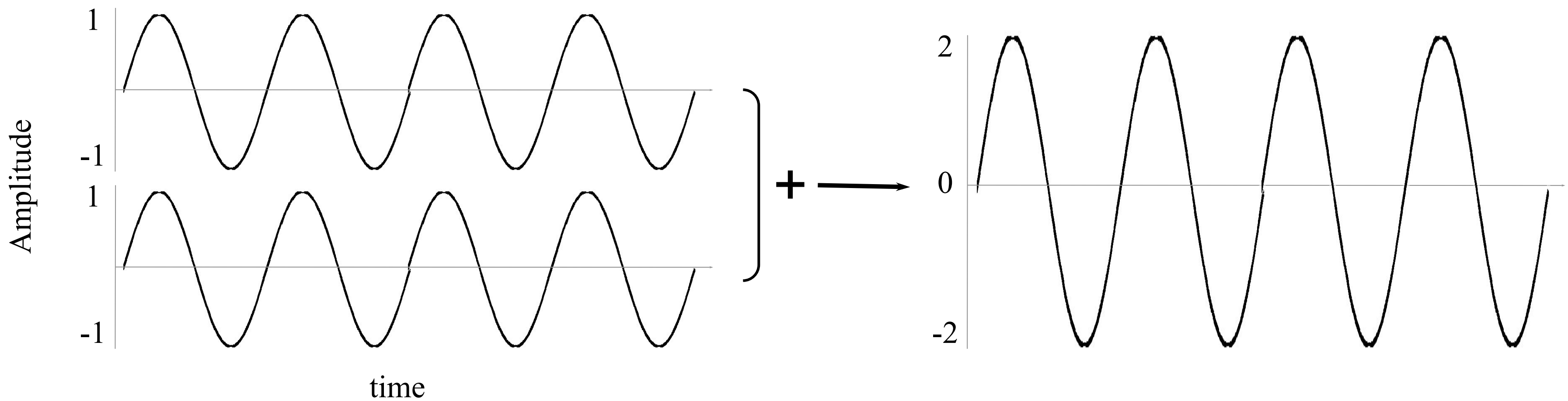




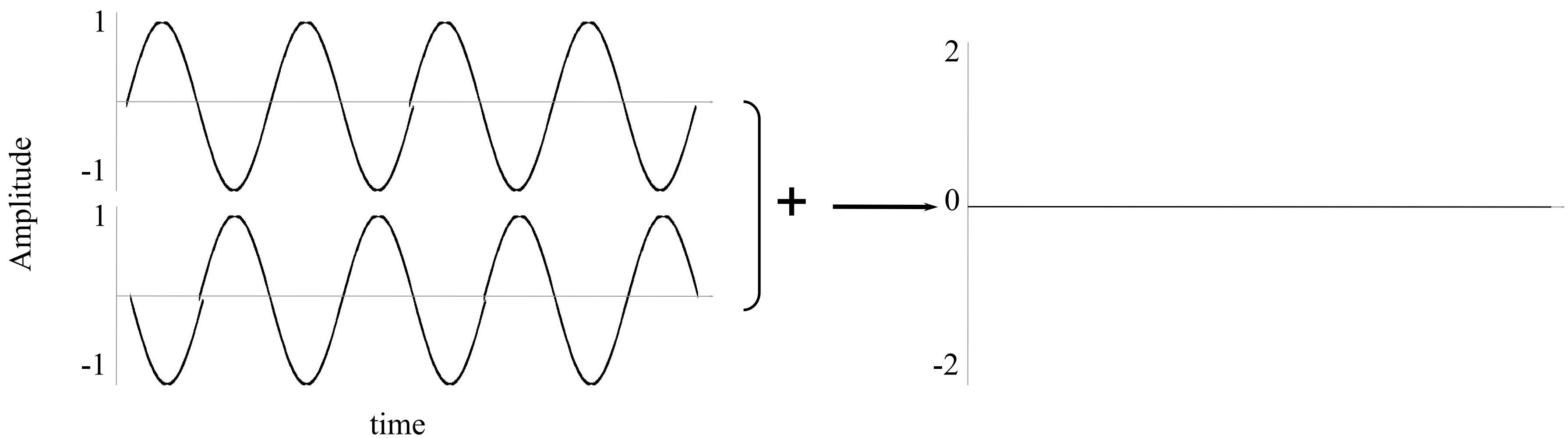
- The frequency defines the number of cycles per second, thus the time x frequency x 360° returns the (unwrapped) angular phase

Phase (1)

- In phase: cycles coincide exactly (sum duplicates amplitude)

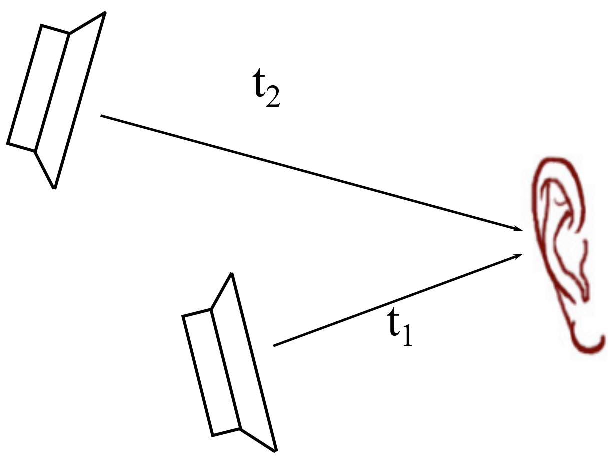


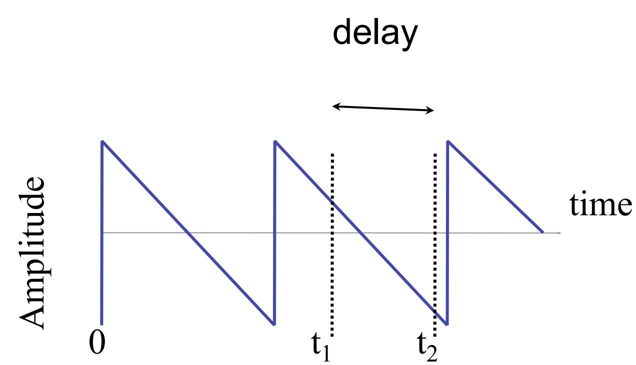
- Out of phase: half cycles are exactly opposed (sum cancels them)



Phase (2)

- There is a range of partial additions and cancellations in between those extremes
- What causes phase difference?

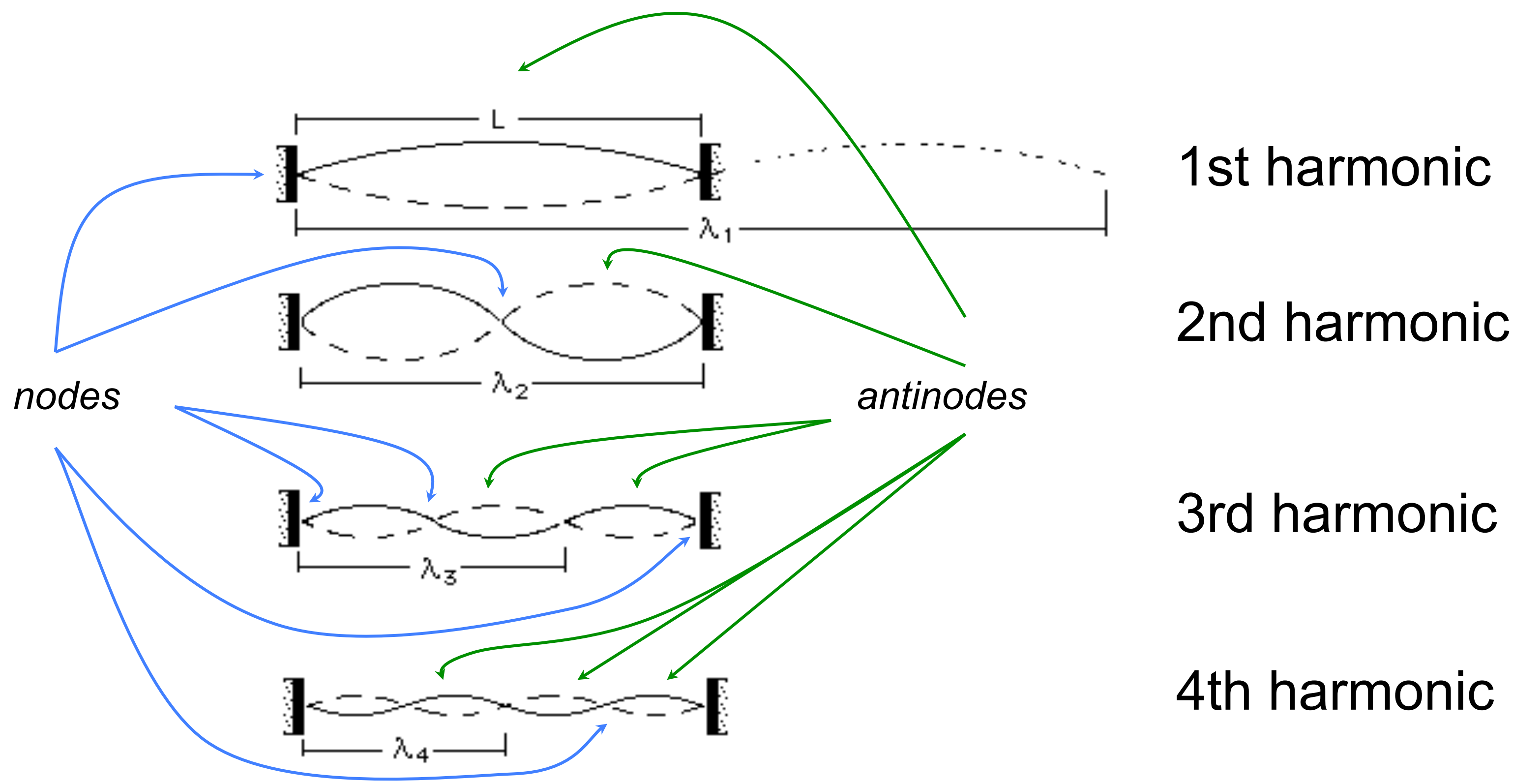




- The phase difference depends on the time deviation and the wave's frequency

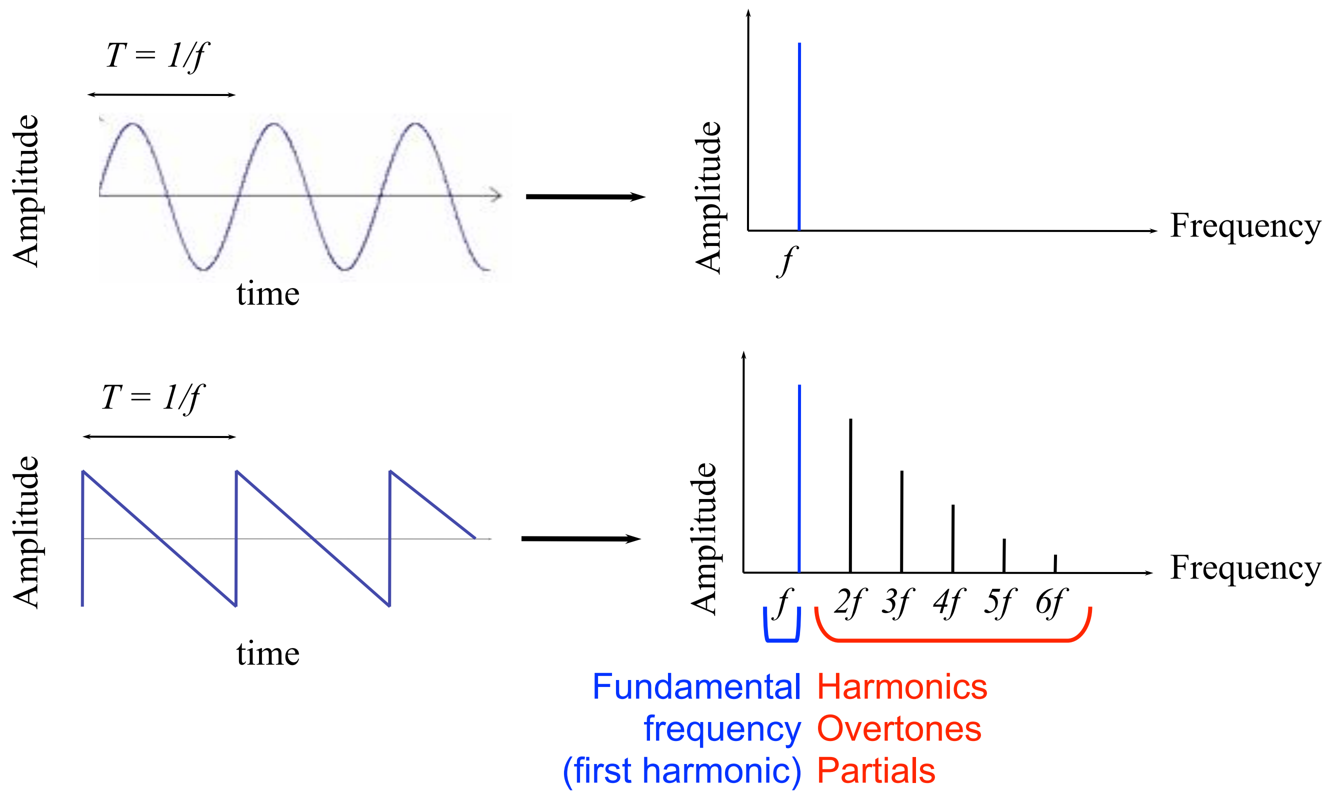
Types of sounds (1)

- Sinusoids are only one possible type of sound corresponding to the simplest mode of vibration, producing energy at only one frequency
- Most sources are capable of vibrating in several harmonic modes at the same time, generating energy at different frequencies



Types of sounds (2)

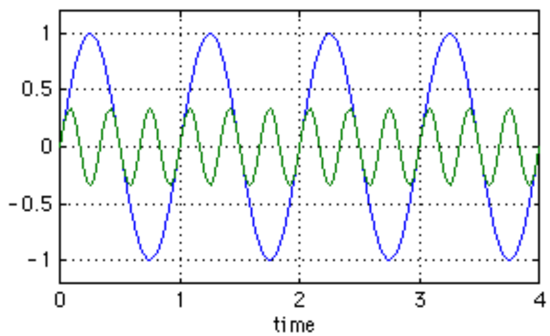
- Harmonics (or Overtones or Partials) are frequency components that occur at integer multiples of the fundamental frequency
- Their amplitude variations determine the timbre of the sound



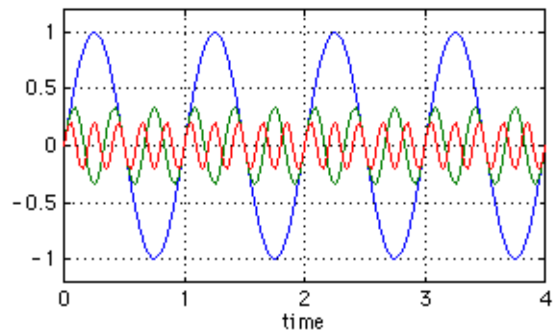
Types of sounds (3)

- Example: Square wave - only odd harmonics (even are missing). Amplitude of the nth harmonic = 1/n

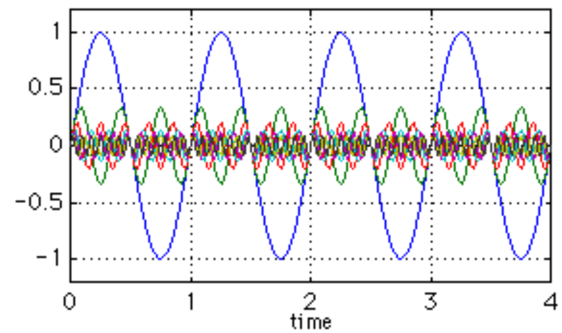
Two Component Recipe for a "Square Wave"



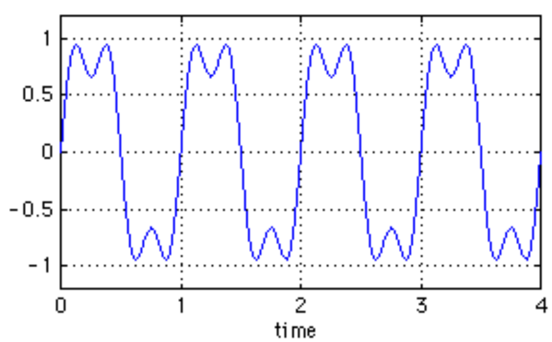
Three Component Recipe for a "Square Wave"



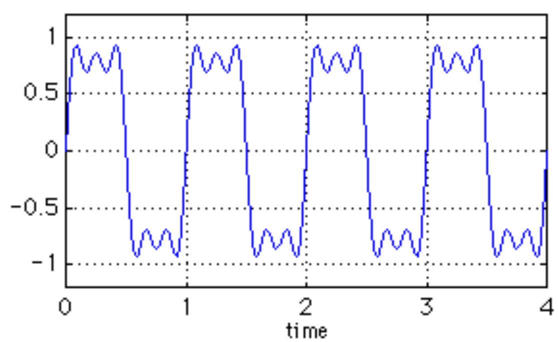
7-Component Recipe for a "Square Wave"



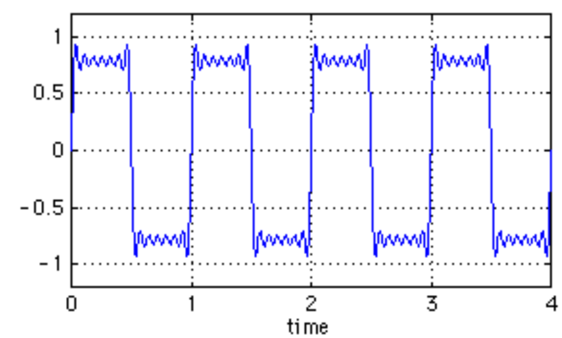
"Square Wave" (Two Components)



"Square Wave" (Three Components)

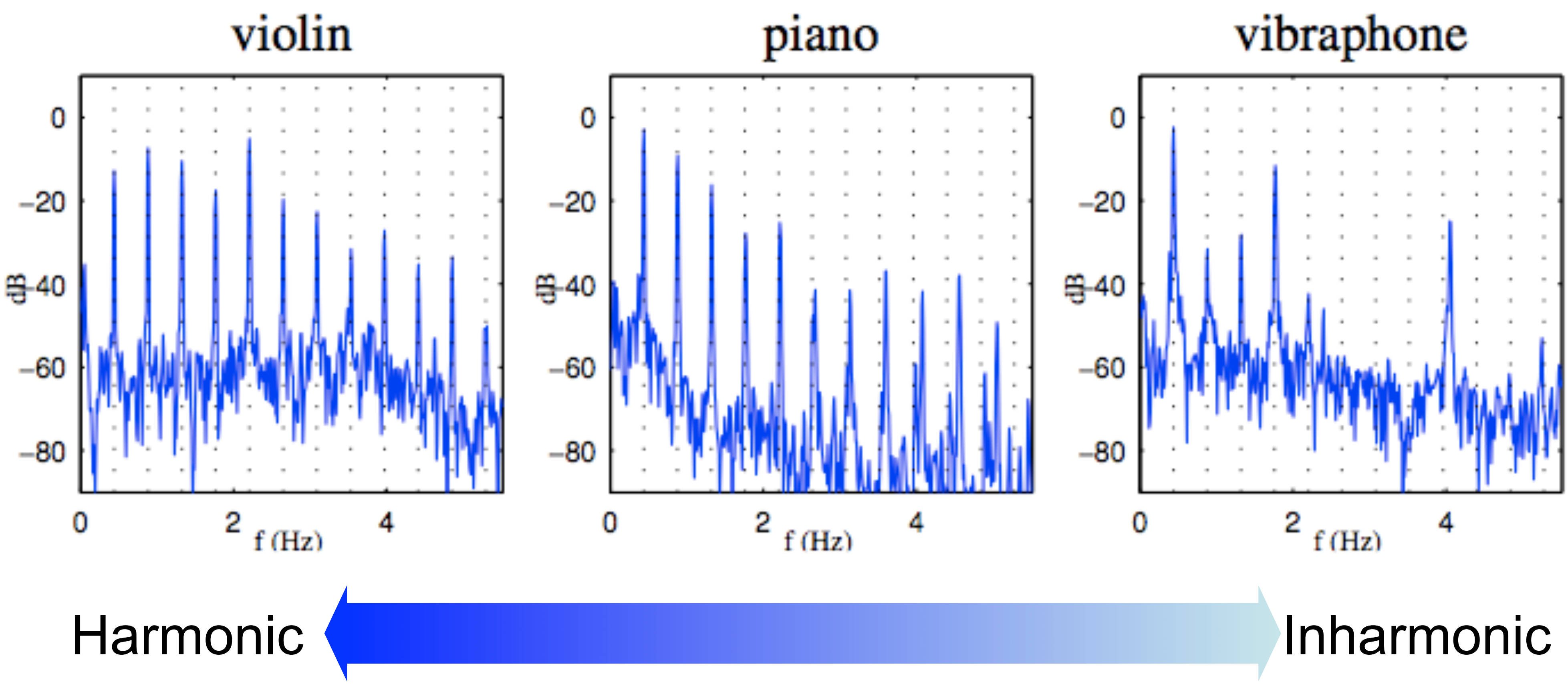


"Square Wave" (Seven Components)



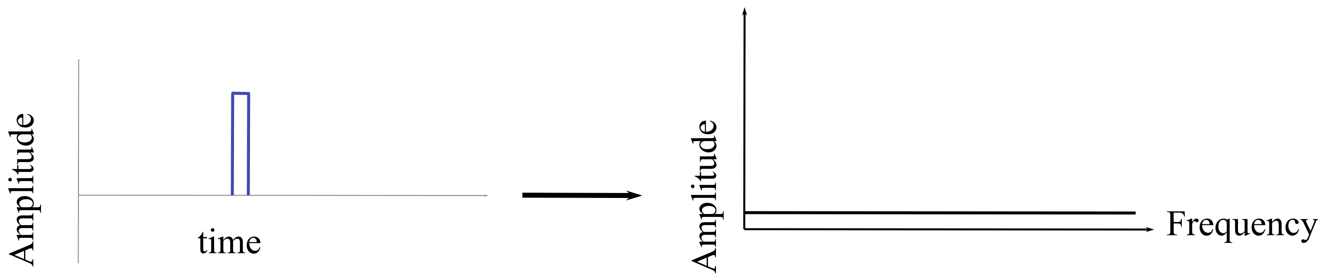
Types of sounds (4)

- Most natural pitched sounds also present overtones which are not integer multiples of the fundamental.
- These are known as inharmonic partials

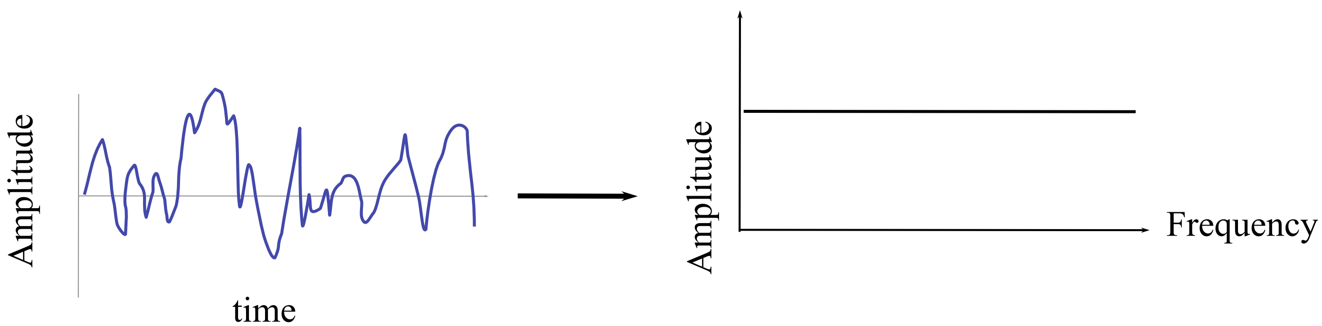


Types of sounds (5)

- Non-periodic sounds have no pitch and tend to have continuous spectra, e.g. a short pulse (narrow in time, wide in frequency)

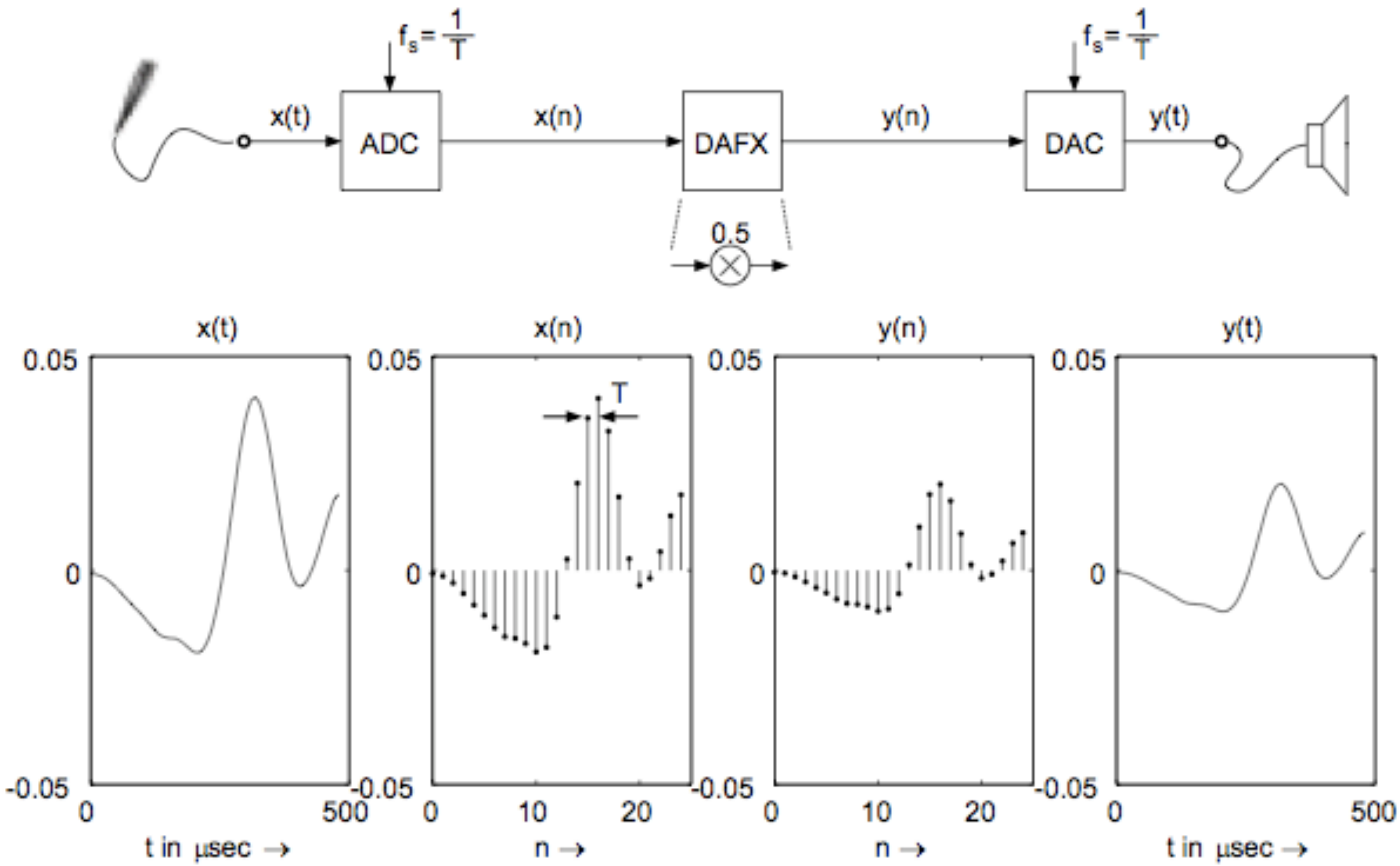


- The most complex sound is white noise (completely random)

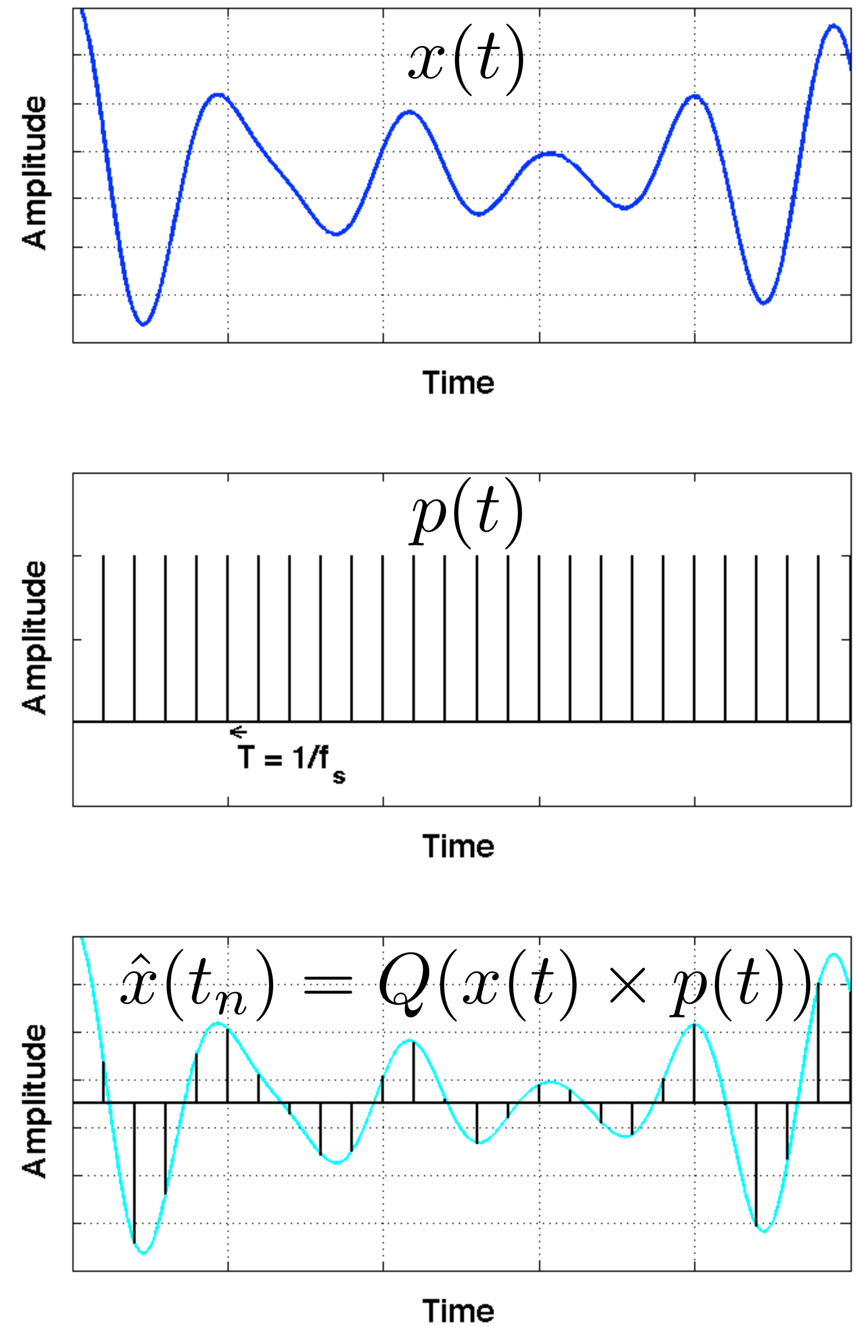


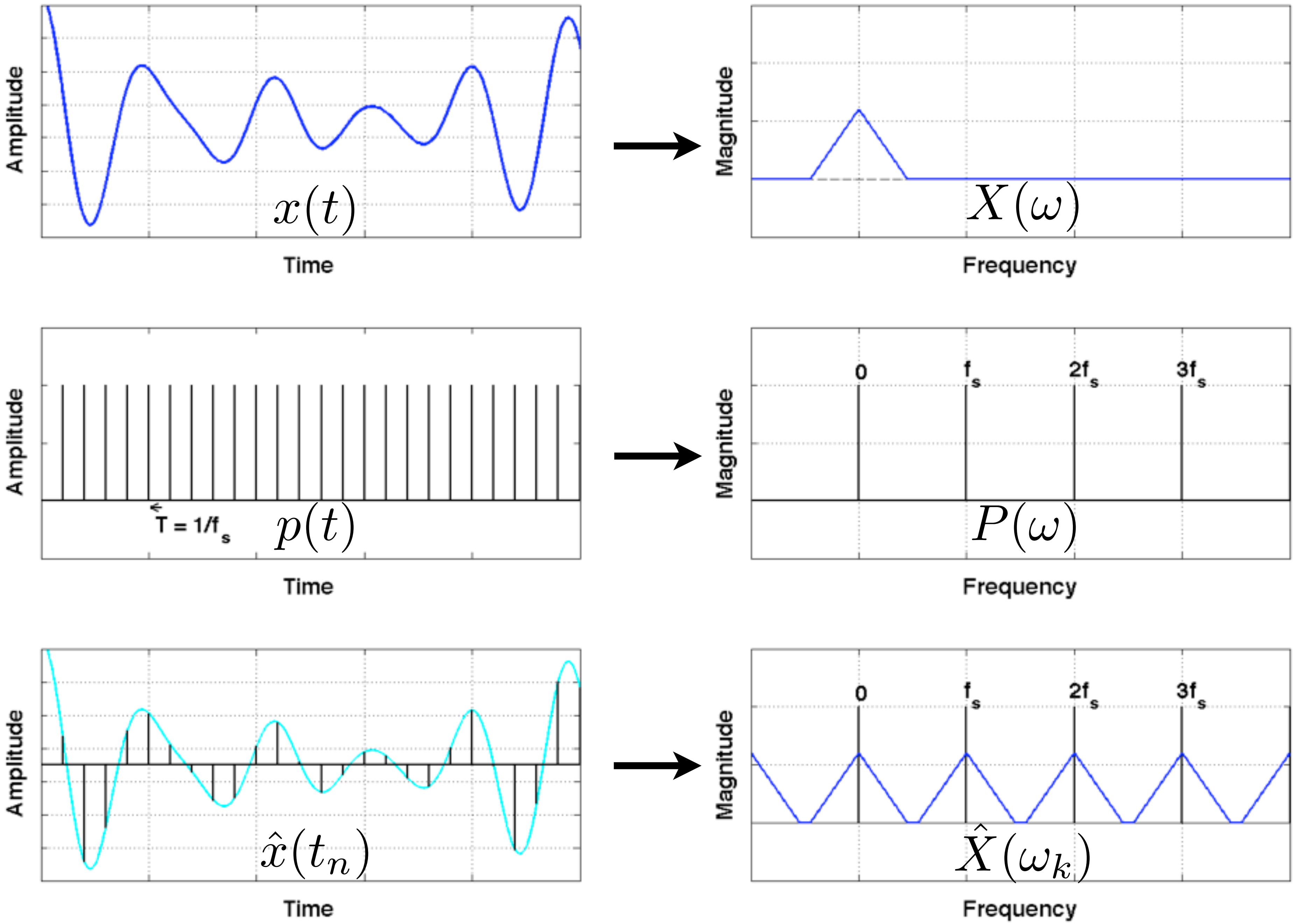
- The more complex, the noisier the sound is

Digital audio

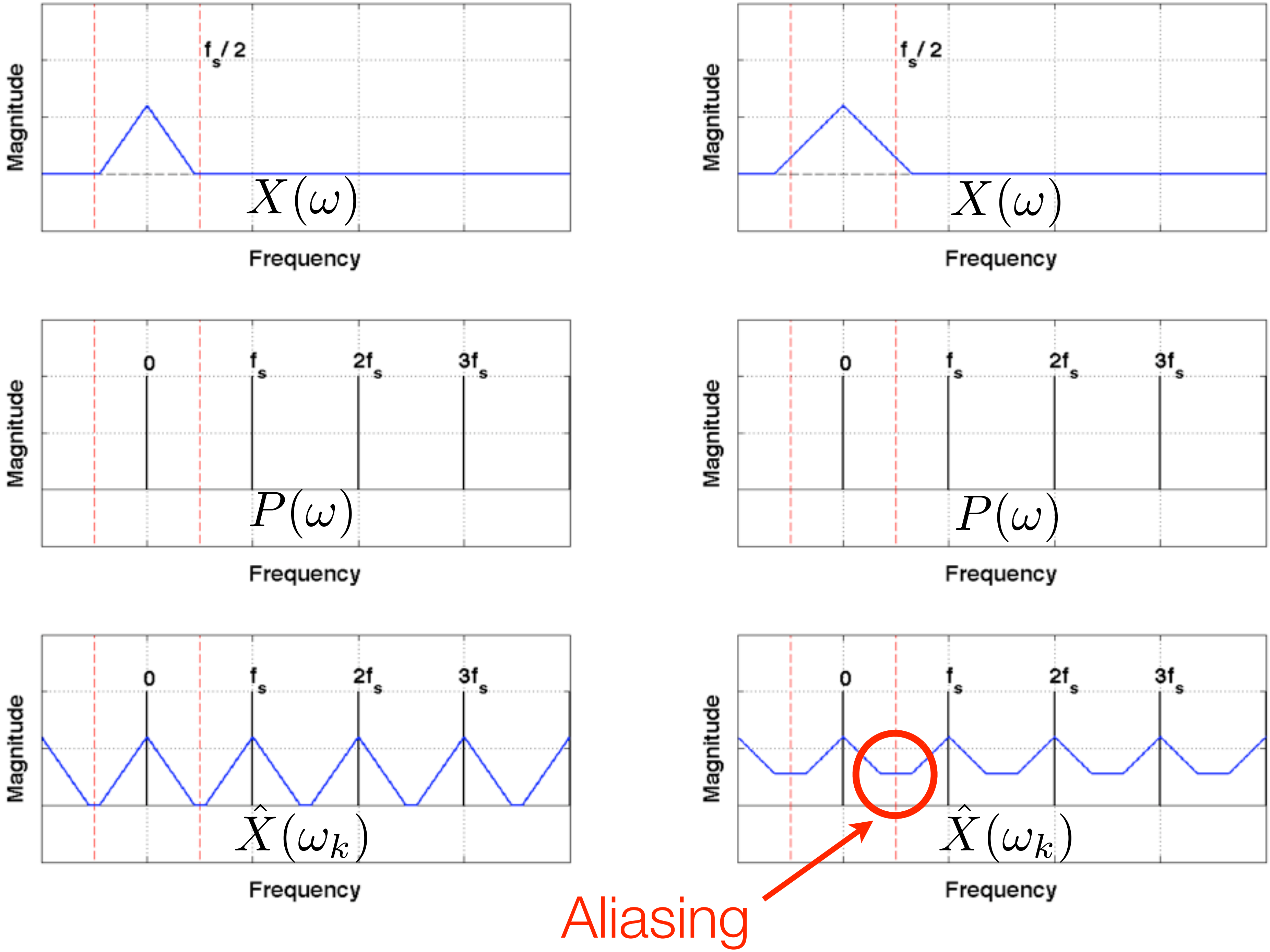


Discrete Signal and Sampling



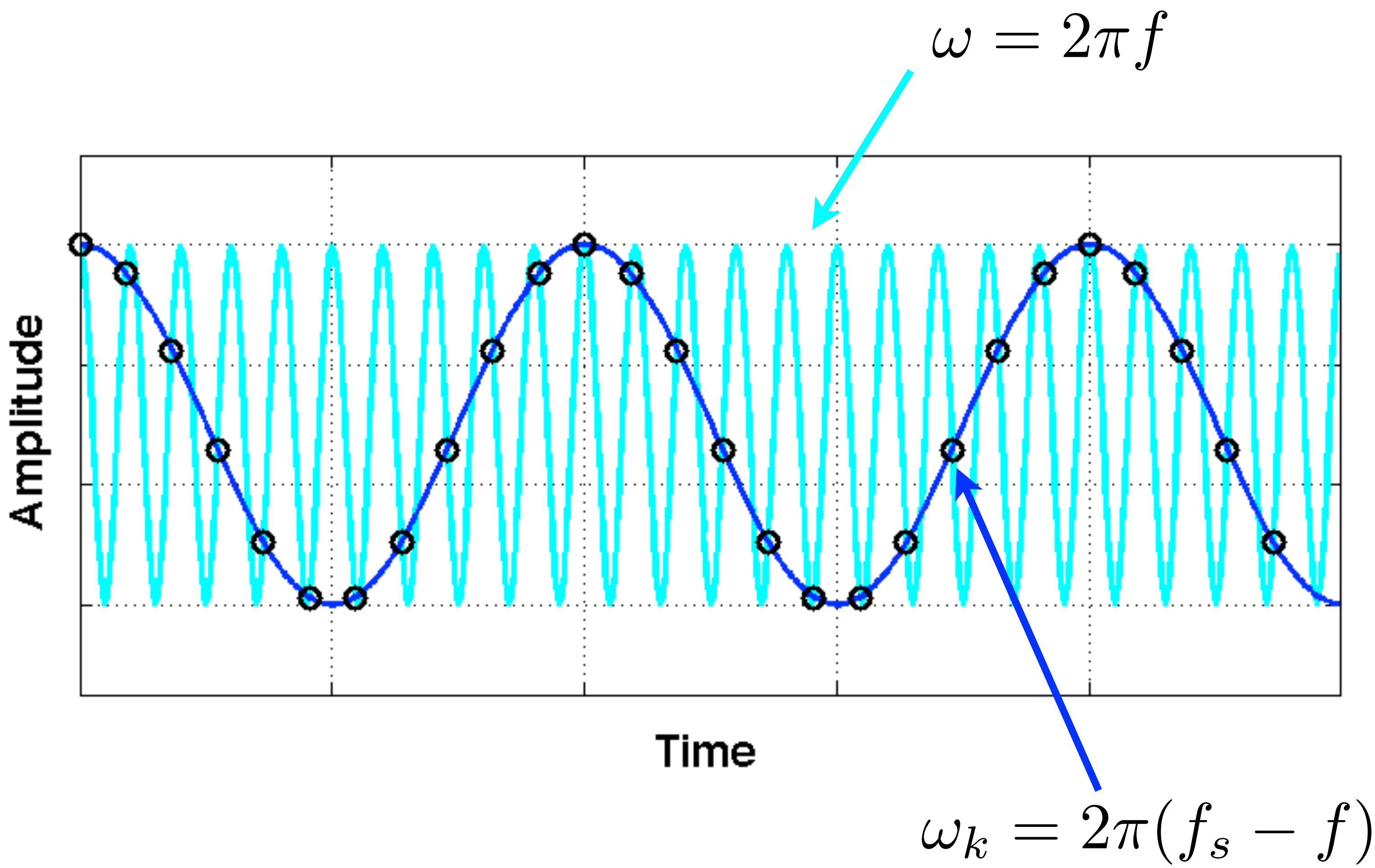


Amplitude
$x(t)$
Time
Magnitude
$X(\omega)$
Frequency
Amplitude
T = 1/f_s
$p(t)$
Time
Magnitude
0
f_s
2f_s
3f_s
$P(\omega)$
Frequency
Amplitude
$\hat{x}(t_n)$
Time
Magnitude
0
f_s
2f_s
3f_s
$\hat{X}(\omega_k)$
Frequency



Magnitude
Frequency
fs / 2
X(ω)
0
fs
2fs
3fs
P(ω)
X̂(ωk)
Aliasing

Aliasing



Discrete Fourier Transform (DFT)

$X(\omega_k) \equiv \sum_{n=0}^{N-1} x(t_n) e^{-j\omega_k t_n}$

$x = \text{input signal}$
$t_n = \frac{n}{fs} = \text{discrete time (s)},\ n \geq 0 \text{ is an integer}$
$fs = \text{sampling rate (Hz)}$
$X = \text{spectrum of } x$
$\omega_k = k\Omega = \text{discrete frequency (rad/s)},\ k \geq 0 \text{ is an integer}$
$\Omega = 2\pi(\frac{fs}{N}) = \text{frequency sampling interval (rad/s)}$
$N = \text{number of time/frequency samples}$

Simple form :

$X(k) \equiv \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N}$

Discrete Fourier Transform (DFT)

This can also be written as:

$$X(k) = \langle x(n), s_k(n) \rangle$$

Which can be formulated as a matrix multiplication:

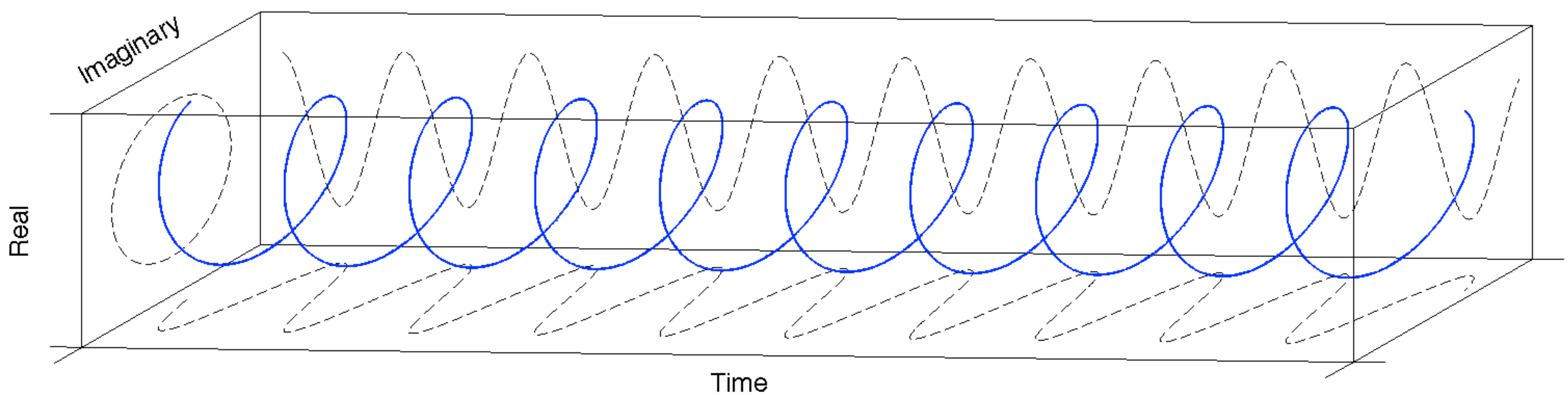
$$\begin{bmatrix} X(0) \\ X(1) \\ \vdots \\ X(N-1) \end{bmatrix} = \begin{bmatrix} s_0^*(0) & s_0^*(1) & \cdots & s_0^*(N-1) \\ s_1^*(0) & s_1^*(1) & \cdots & s_1^*(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ s_{N-1}^*(0) & s_{N-1}^*(1) & \cdots & s_{N-1}^*(N-1) \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(N-1) \end{bmatrix}$$

Discrete Fourier Transform (DFT)

where,

$$s_k(n) = e^{j2\pi nk/N} = \overbrace{cos(2\pi nk/N)}^{\text{Real part}} + \overbrace{jsin(2\pi nk/N)}^{\text{Imaginary part}}$$

is the set of the sampled complex sinusoids with a whole number of periods in N samples (Smith, 2007).



Discrete Fourier Transform (DFT)

The N resulting $X(k)$ are complex-valued vectors $X_R(k) + jX_I(k)$ such that, $\forall k = 0, 1, \cdots, N-1$:

$$|X(k)| = \sqrt{X_R^2(k) + X_I^2(k)}$$

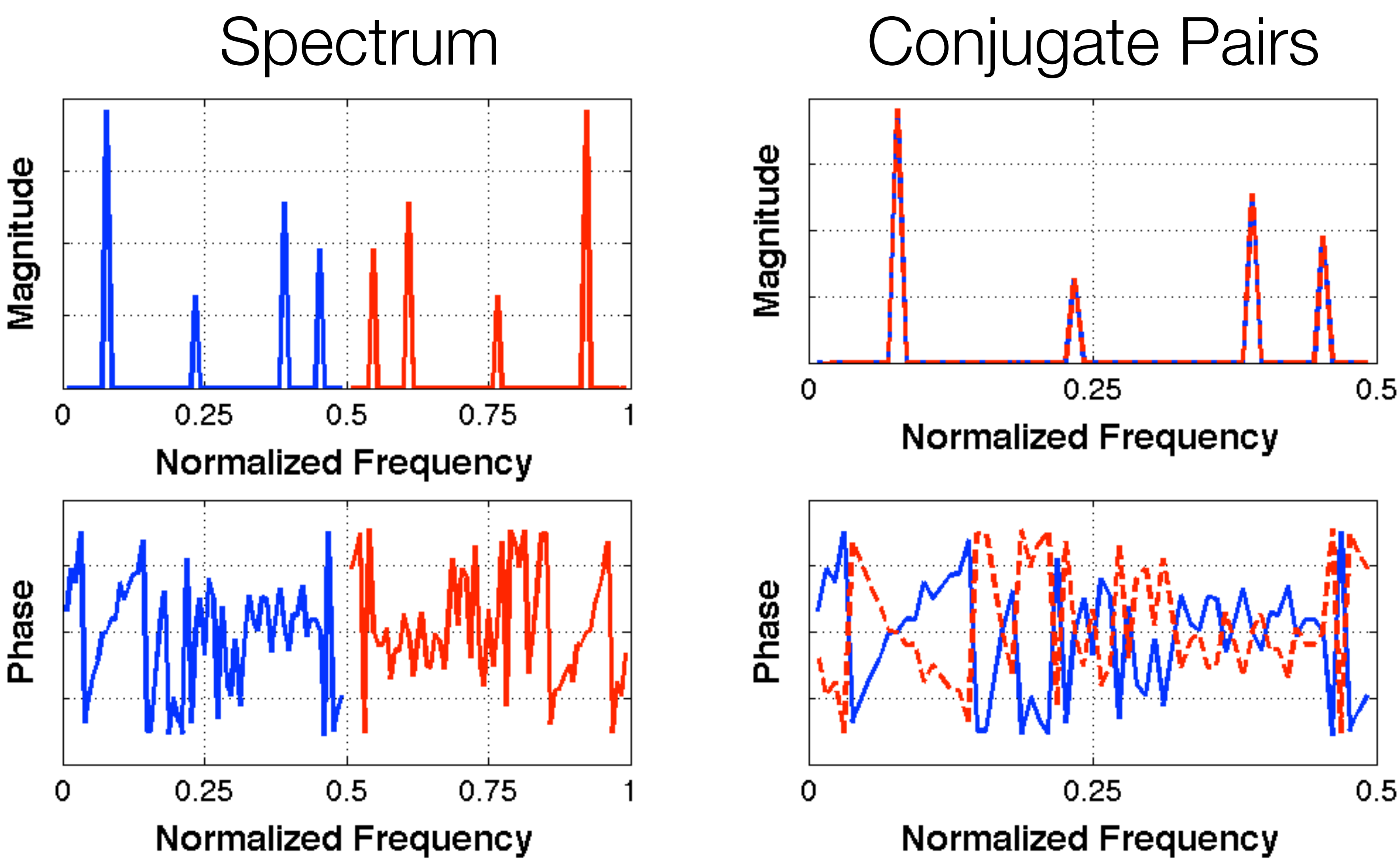
$$\angle X = \phi(k) = tan^{-1} \frac{X_I(k)}{X_R(k)}$$

Furthermore, if $x(n)$ is real-valued, then:

$$X(k) = X^*(N-k)$$

The DFT of an audio signal is half-redundant!

Discrete Fourier Transform (DFT)



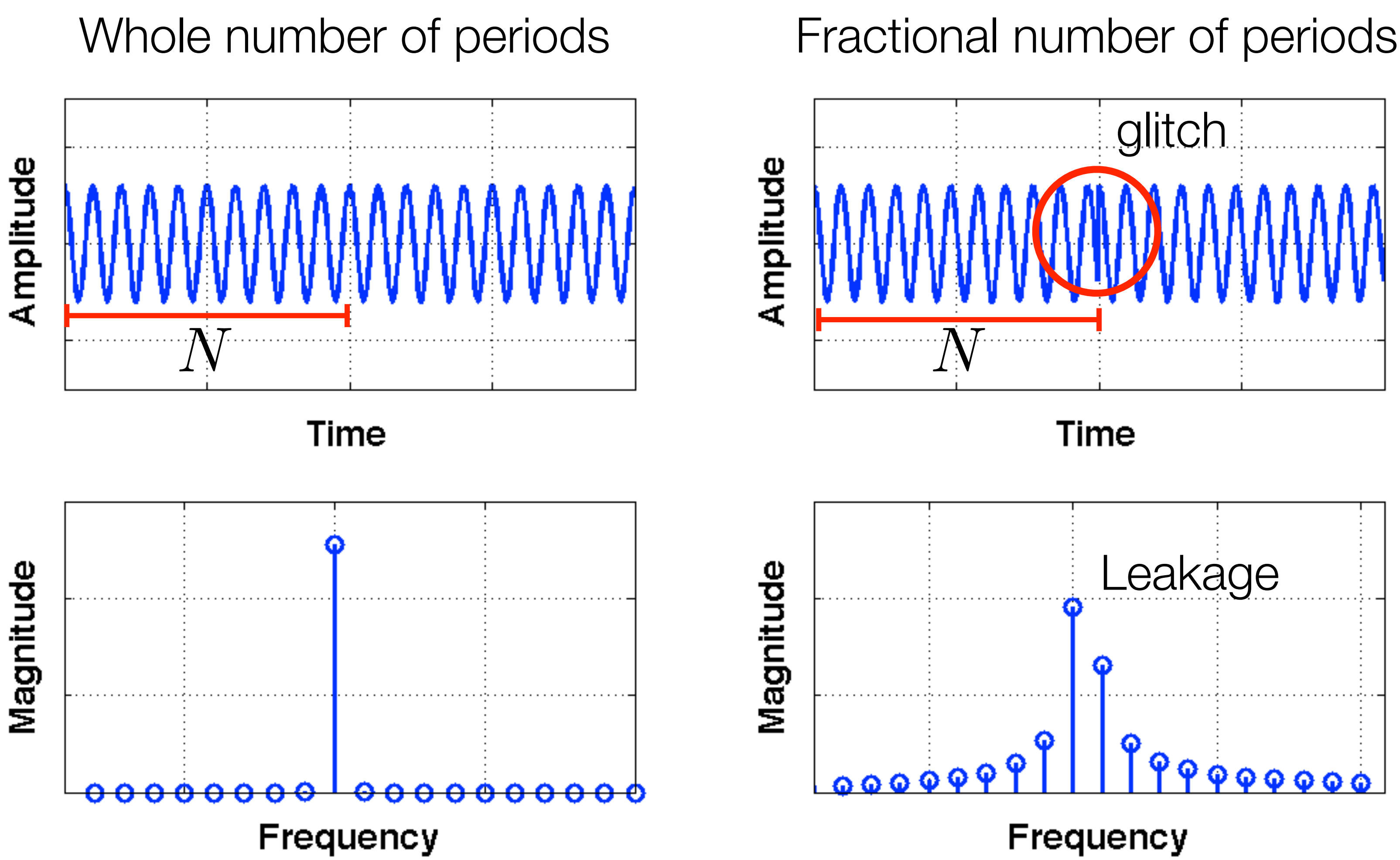
The IDFT and FFT

- The Inverse DFT (IDFT) is defined as:

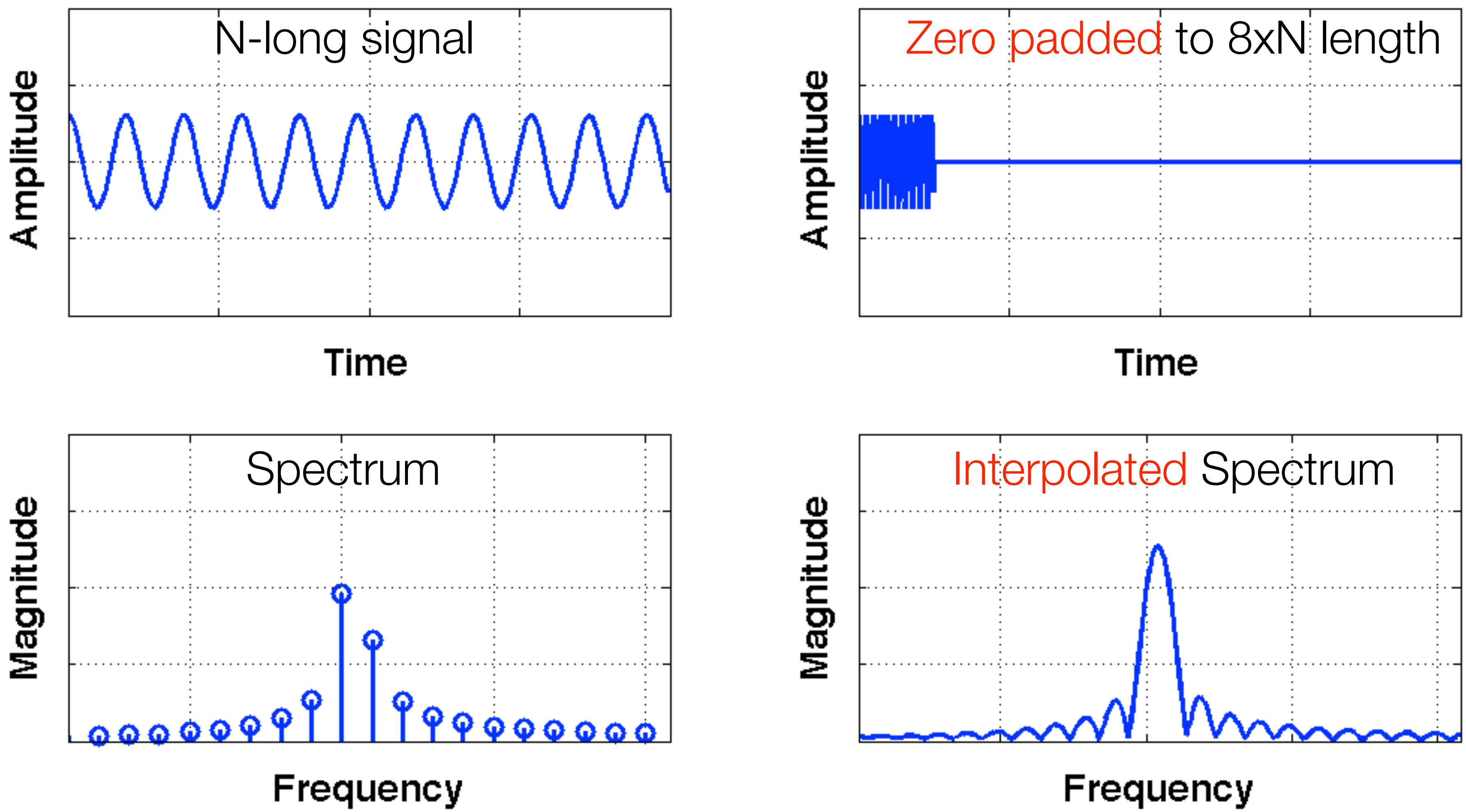
$$x(n) \equiv \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j2\pi nk/N}, \quad n = 0, 1, \cdots, N-1$$

- The DFT needs on the order of N^2 operations for its computation.
- The Fast Fourier Transform (FFT) is an efficient implementation of the DFT, that only requires on the order of $N\log_2 N$ operations when N is a power of 2.
- The FFT is so fast that it can be used to efficiently perform time-domain operations such as convolution.

Spectral Leakage



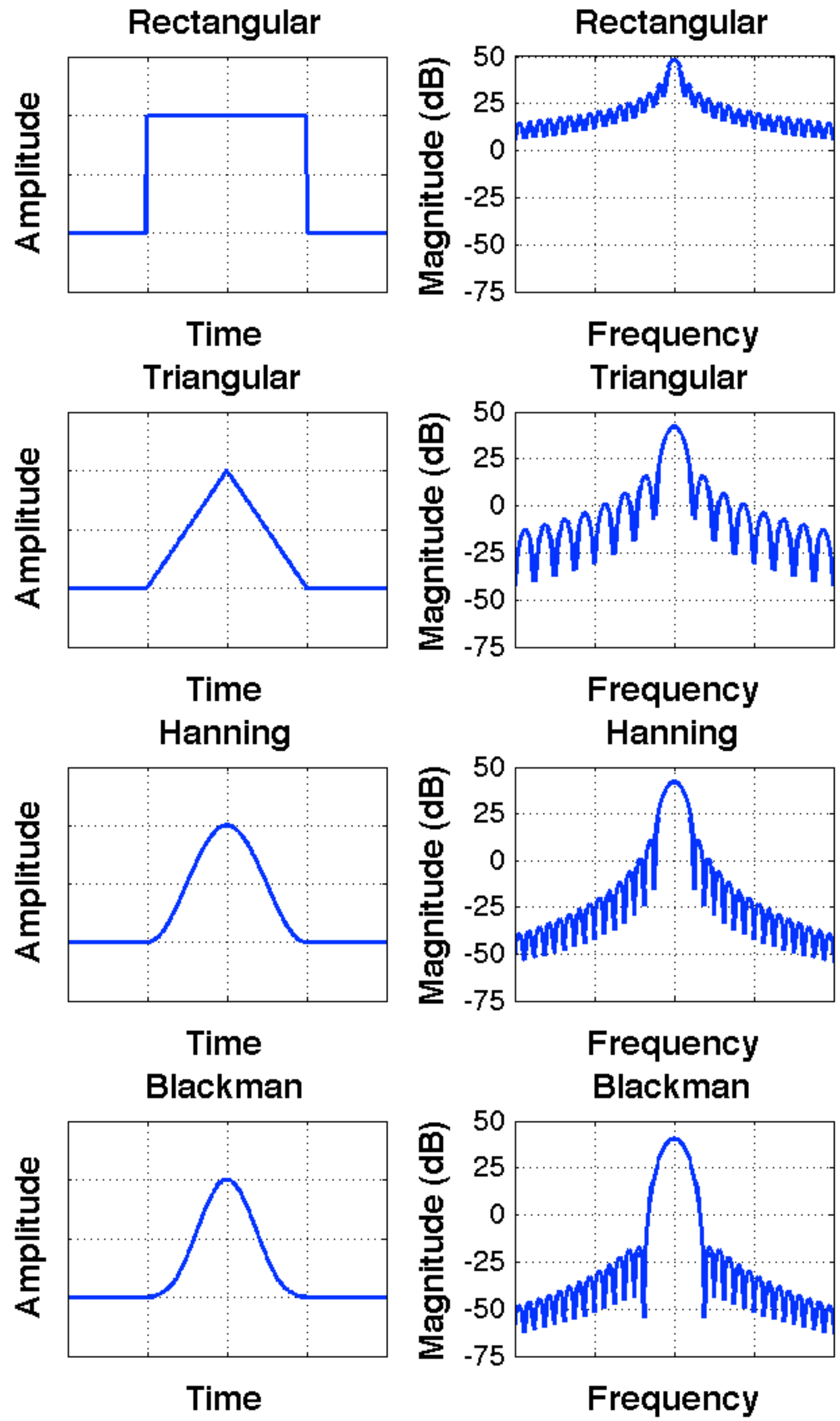
Zero padding



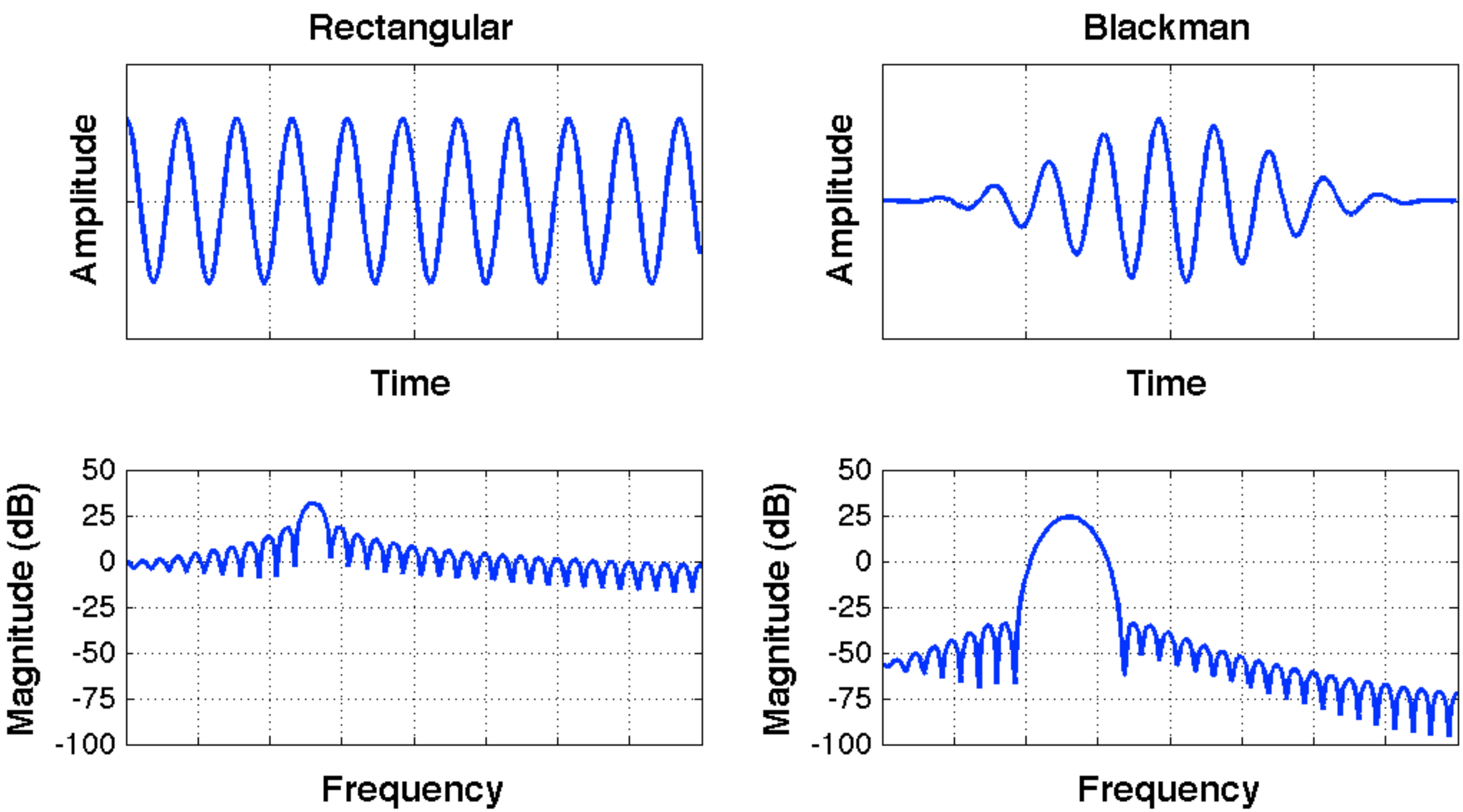
Windows

- We are effectively using a rectangular window: w(n)
- Spectrum = convolution of X(k) and W(k)
- Ideal window: narrow central lobe; strong attenuation in sidebands
- Figures show Magnitude in dB:

$$dB(X) = 20 \times log_{10}(X)$$

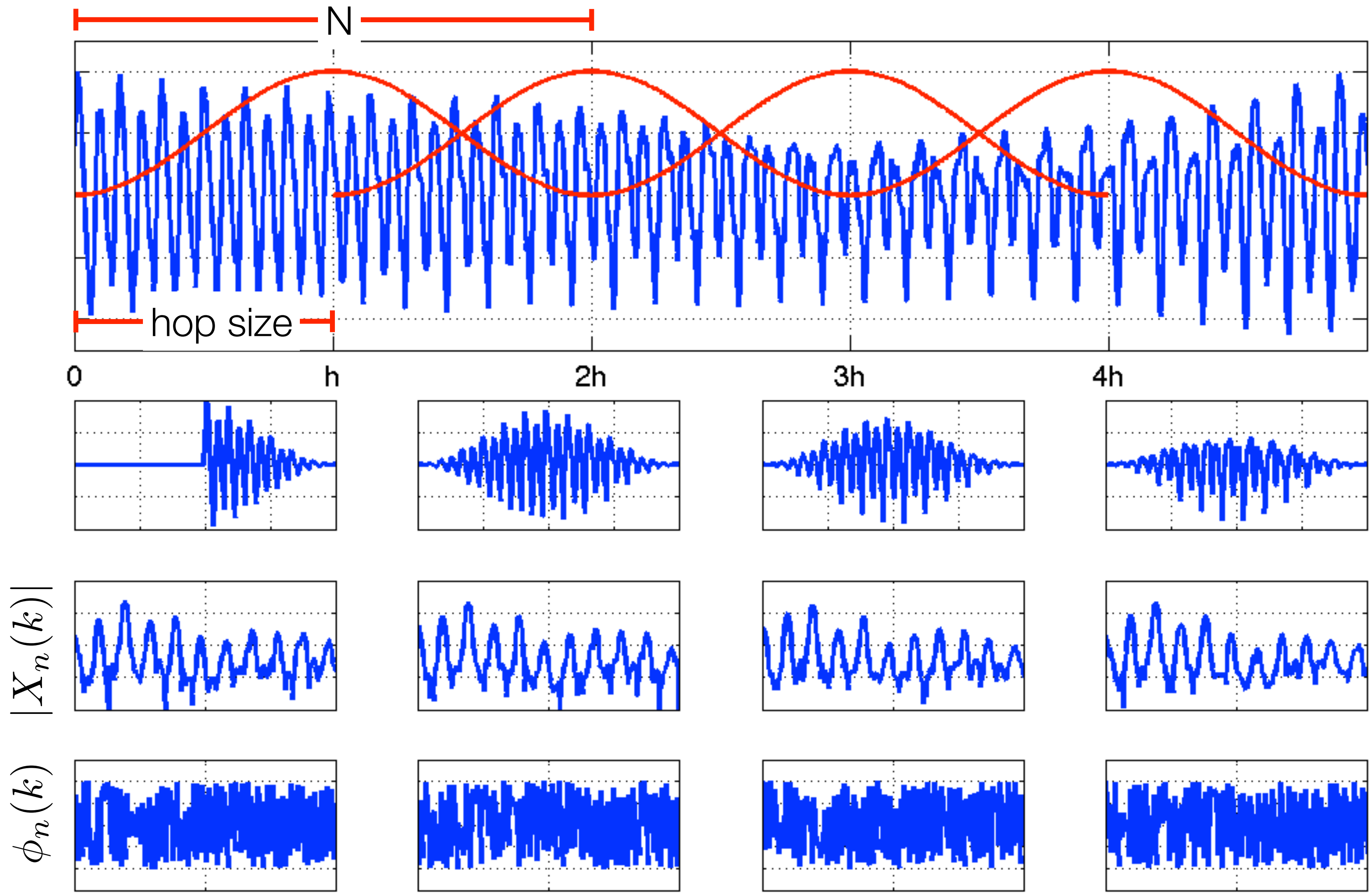


Windowing

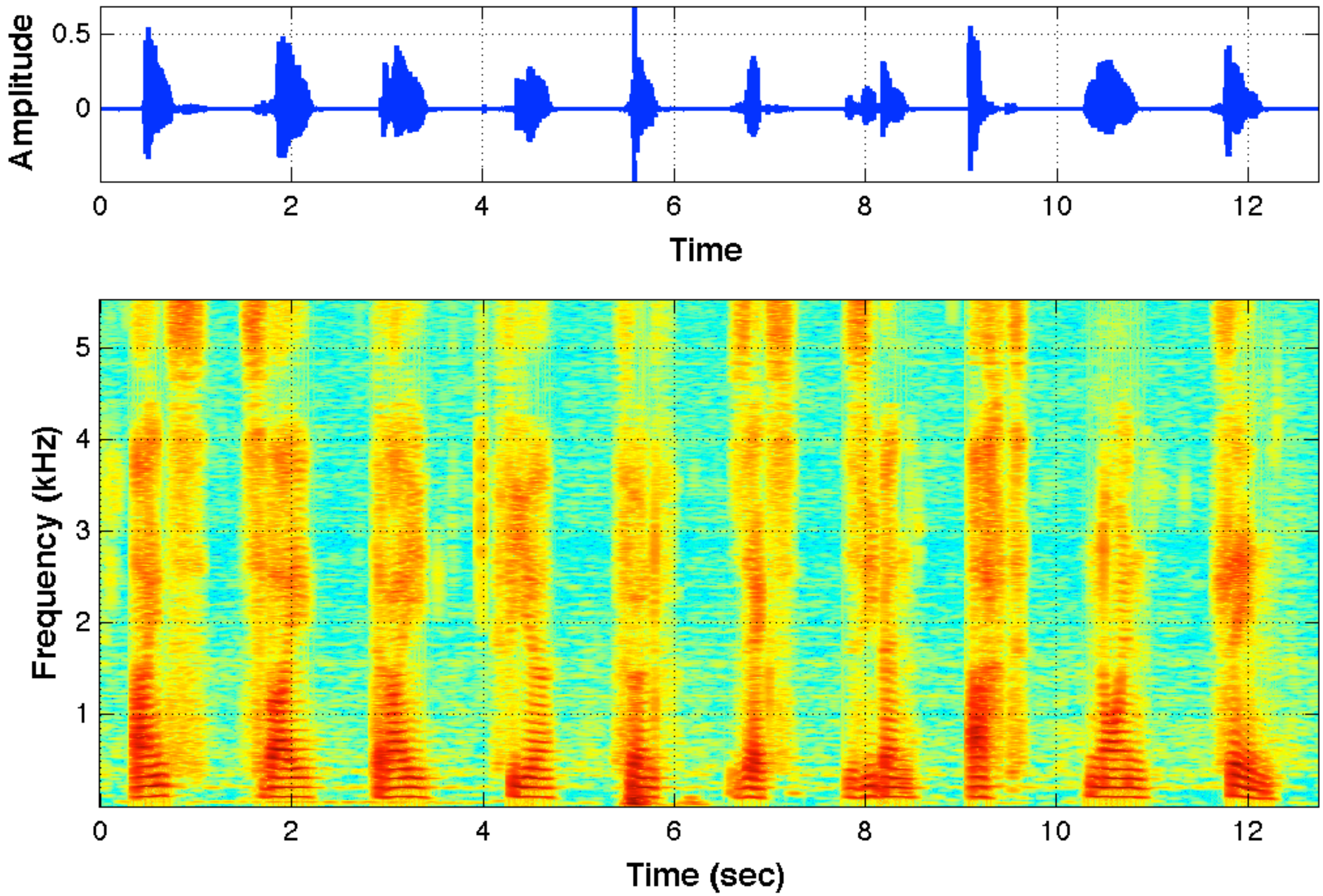


Rectangular
Blackman
Amplitude
Time
Amplitude
Time
Magnitude (dB)
50
25
0
-25
-50
-75
-100
Frequency
Magnitude (dB)
50
25
0
-25
-50
-75
-100
Frequency

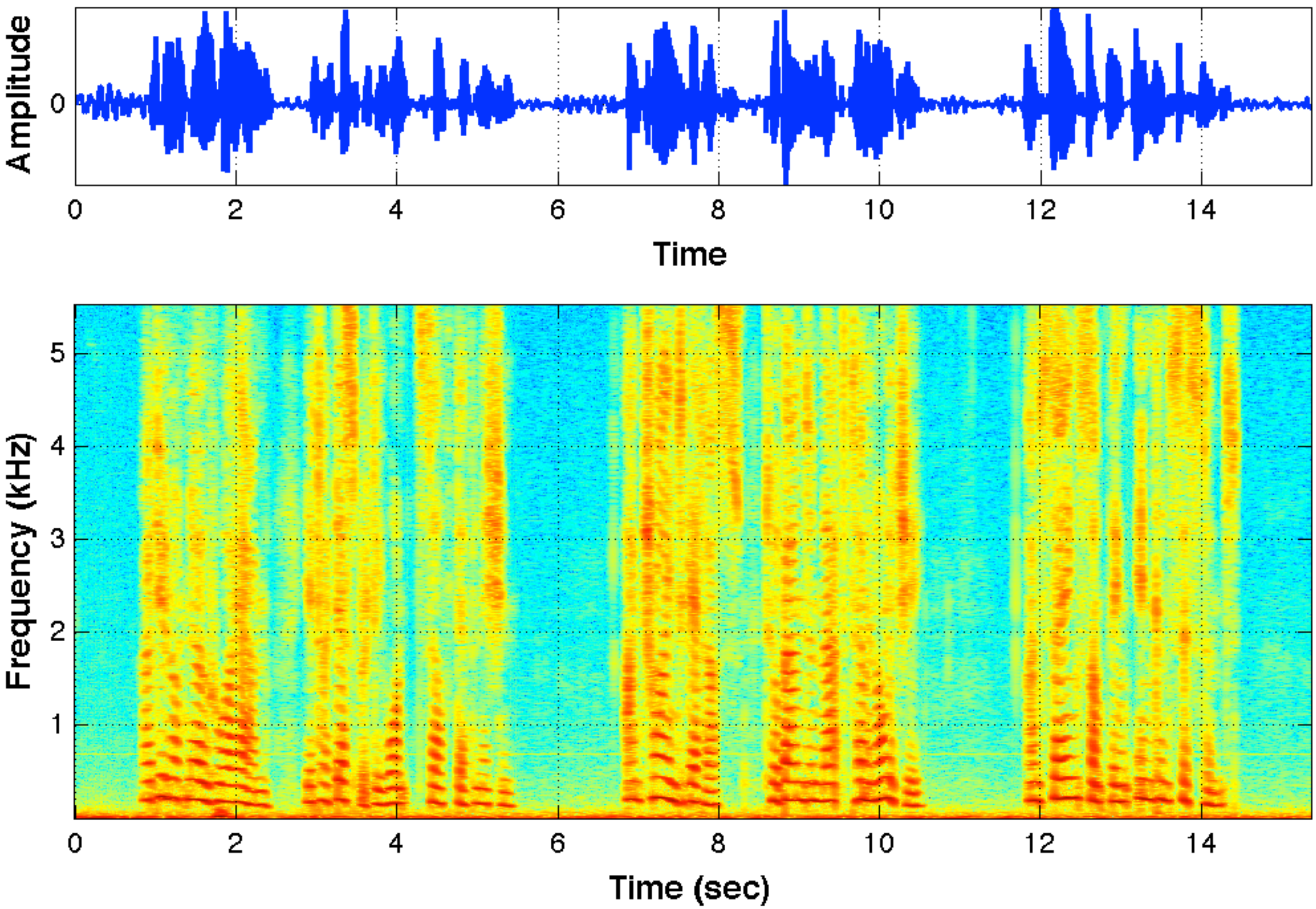
Short-Time Fourier Transform (STFT)



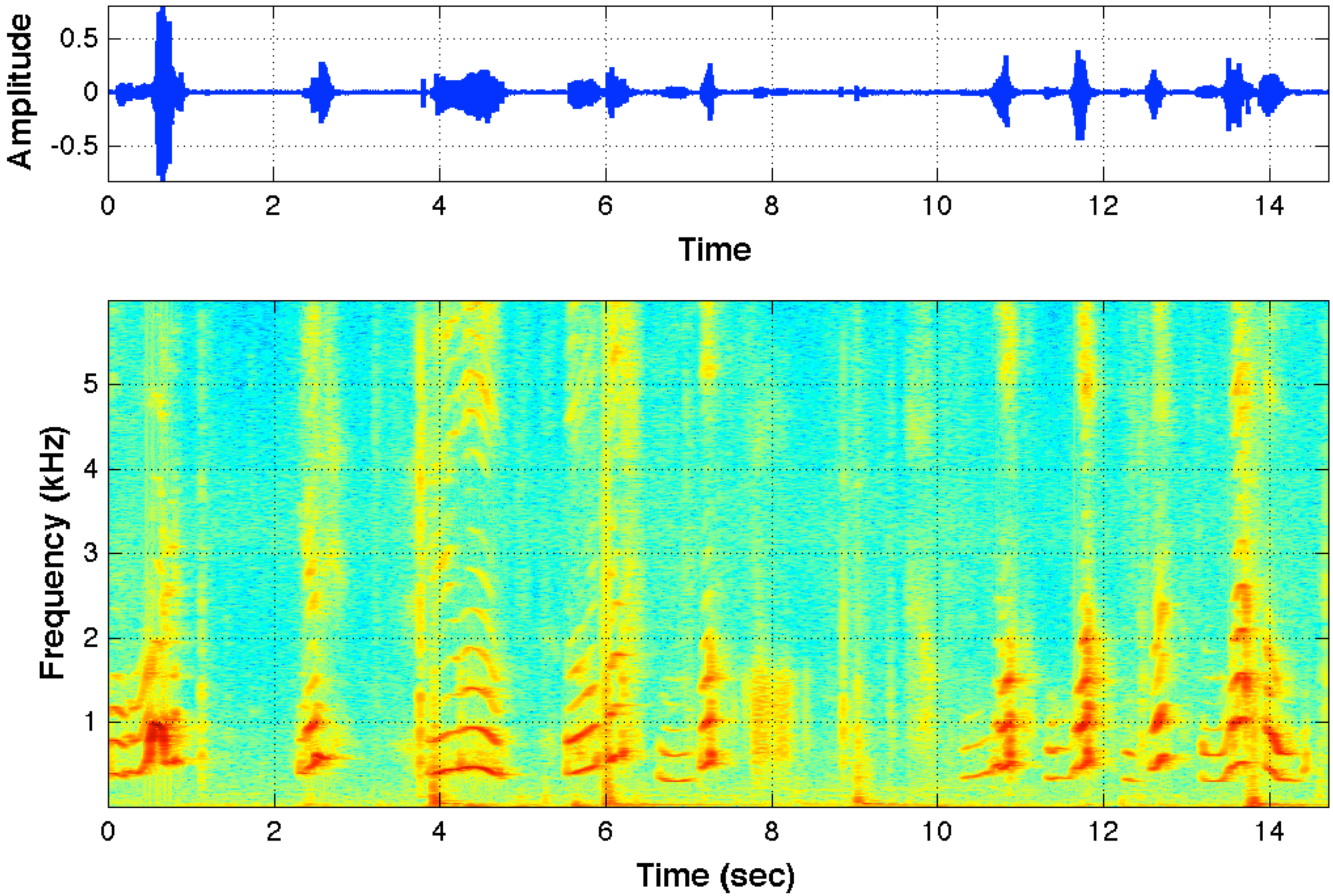
Spectrogram - male speaker



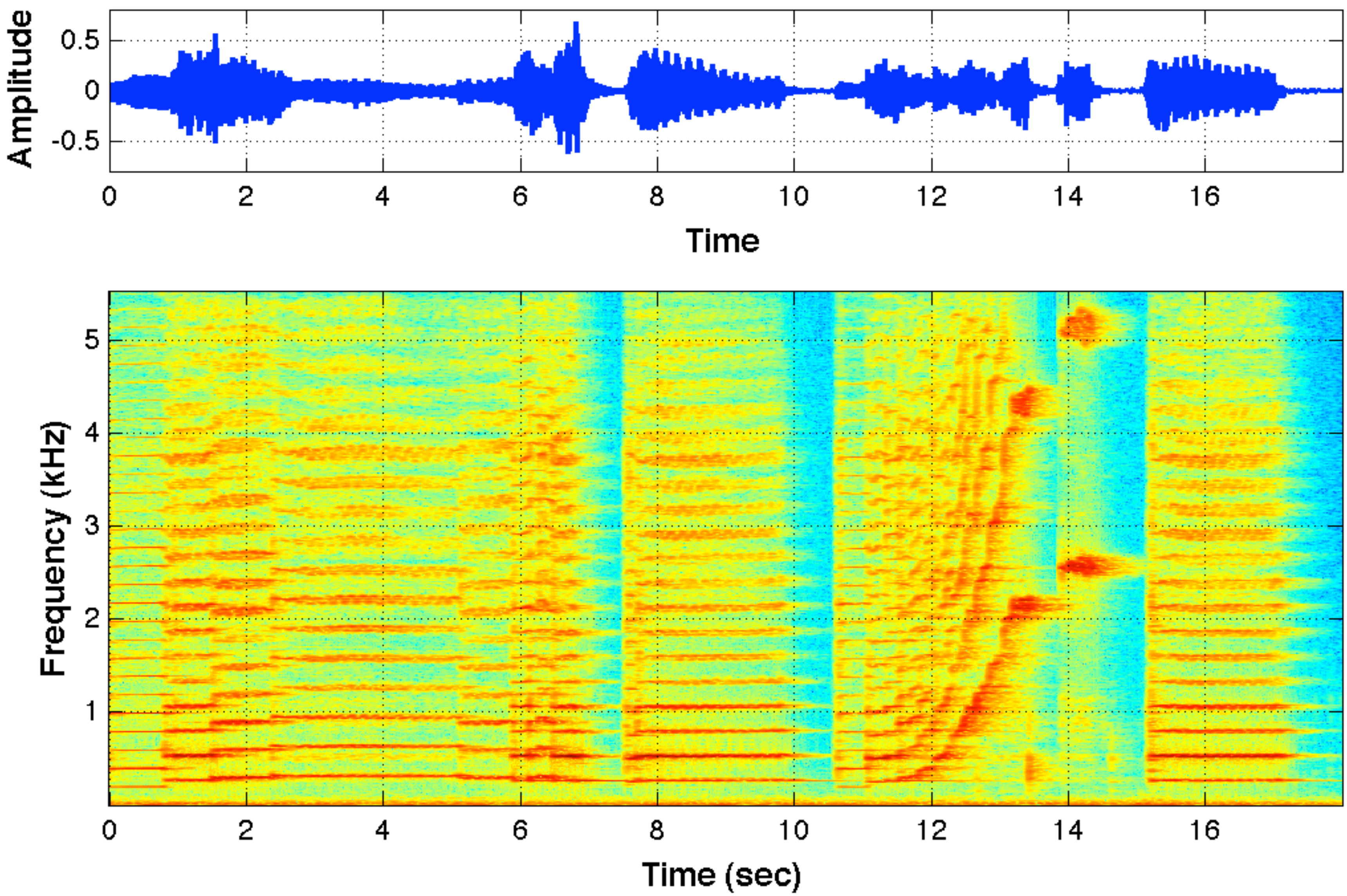
Spectrogram - female speaker



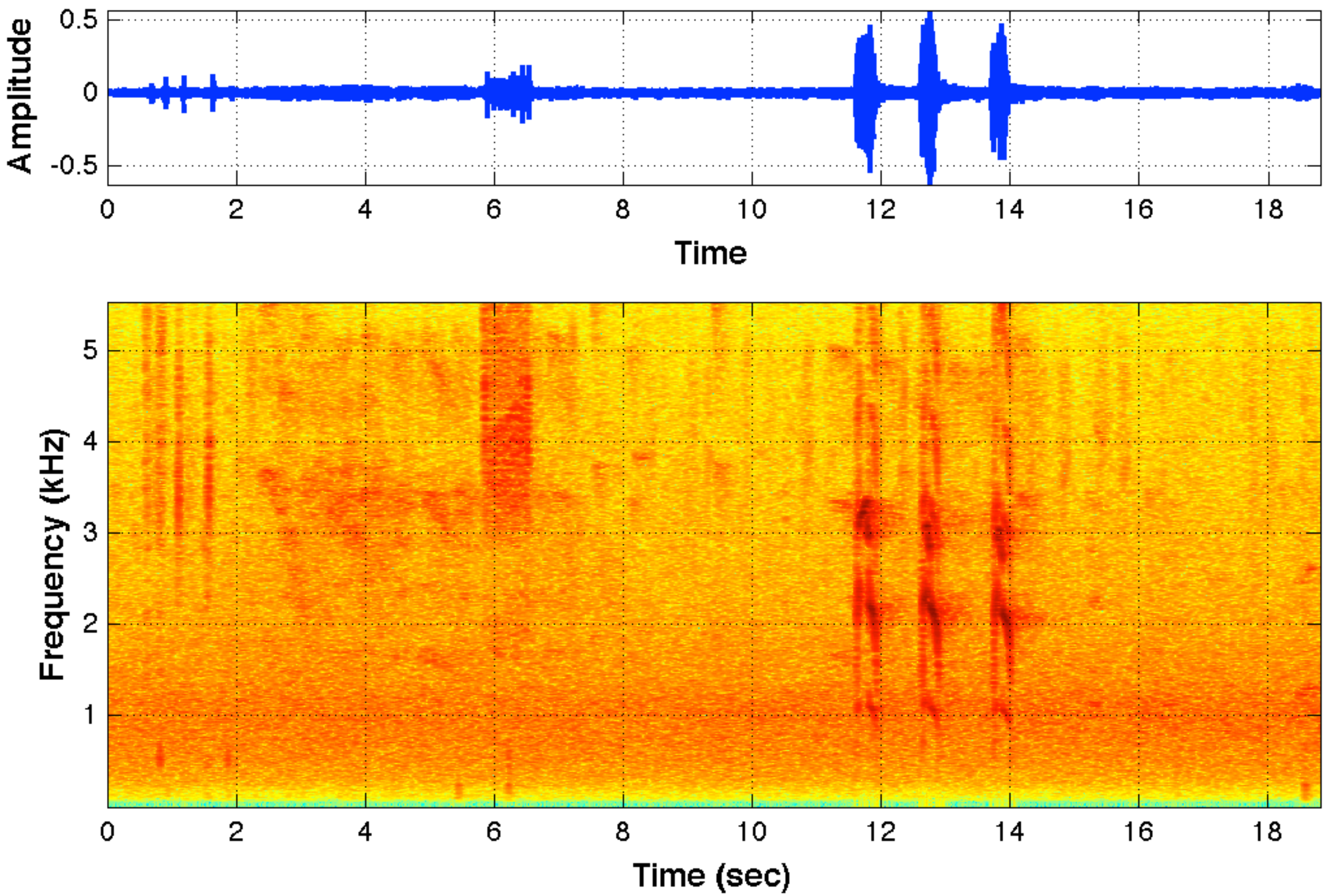
Spectrogram - baby cooing



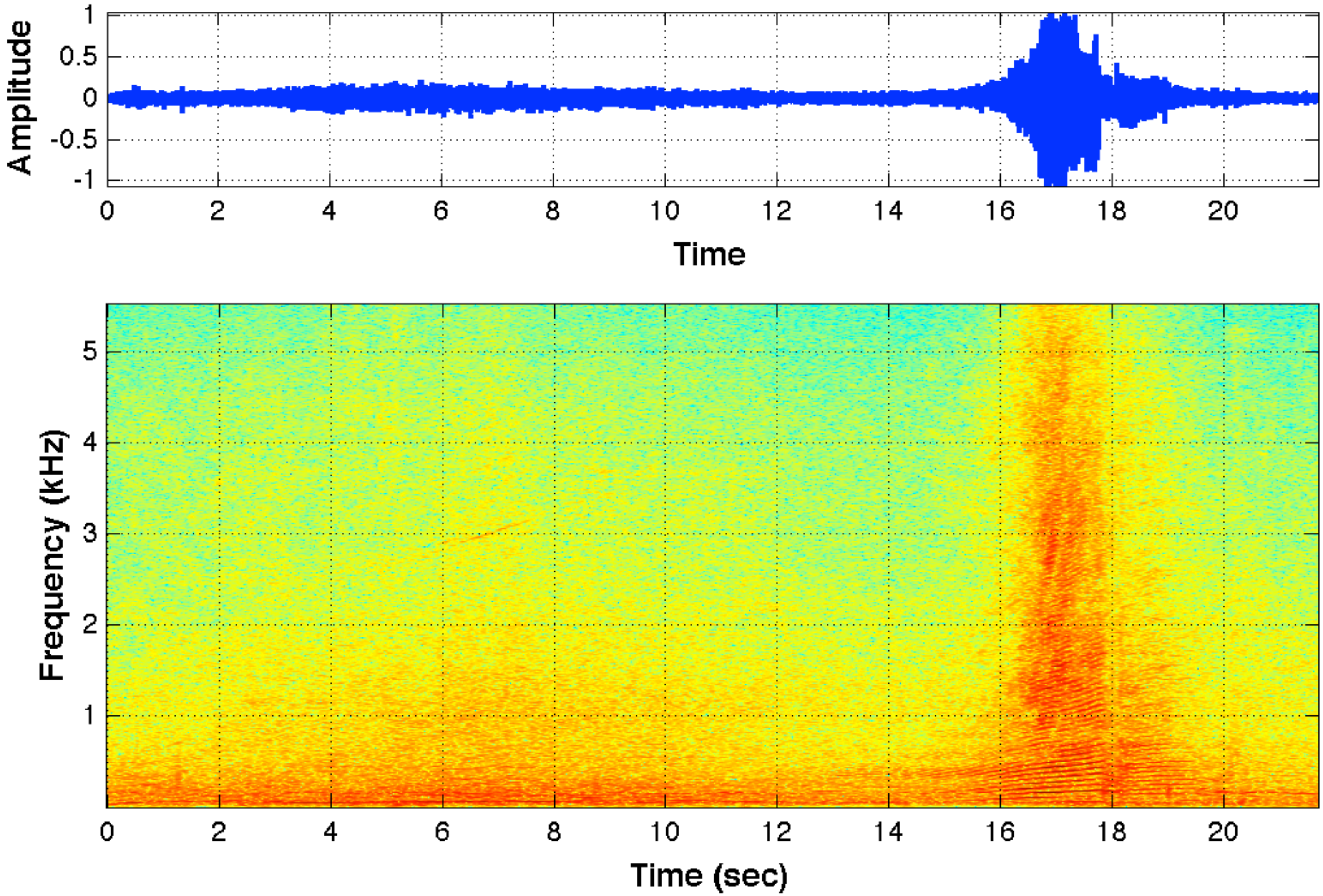
Spectrogram - violin



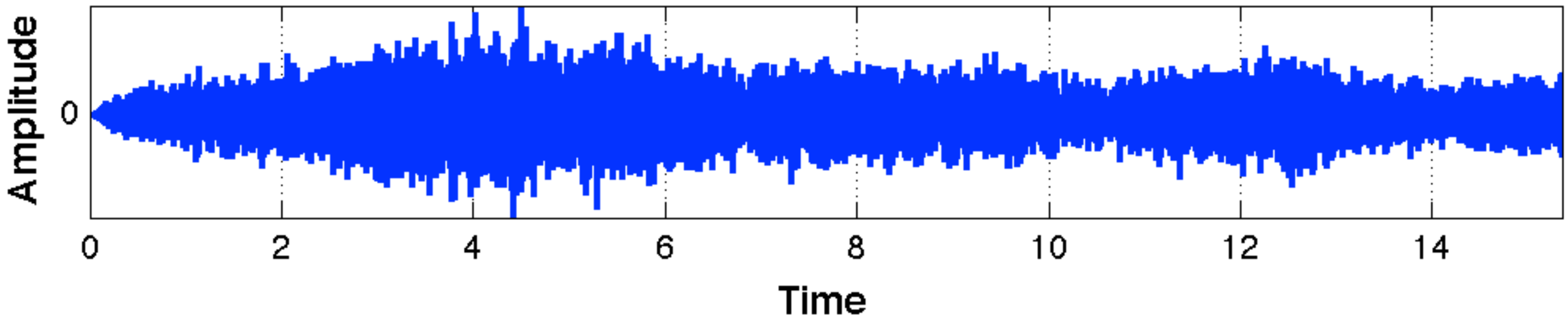
Spectrogram - birds by lakeside

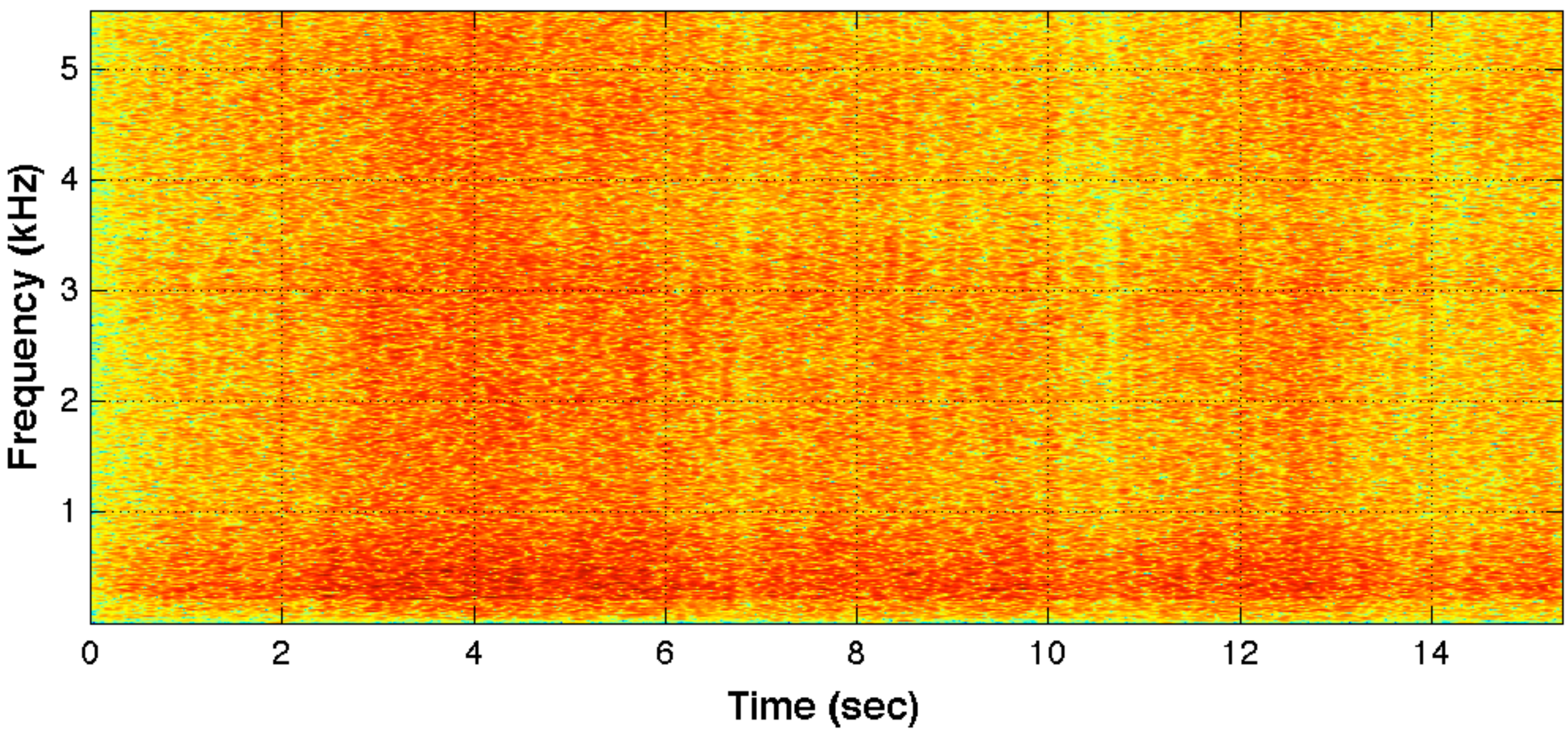


Spectrogram - street

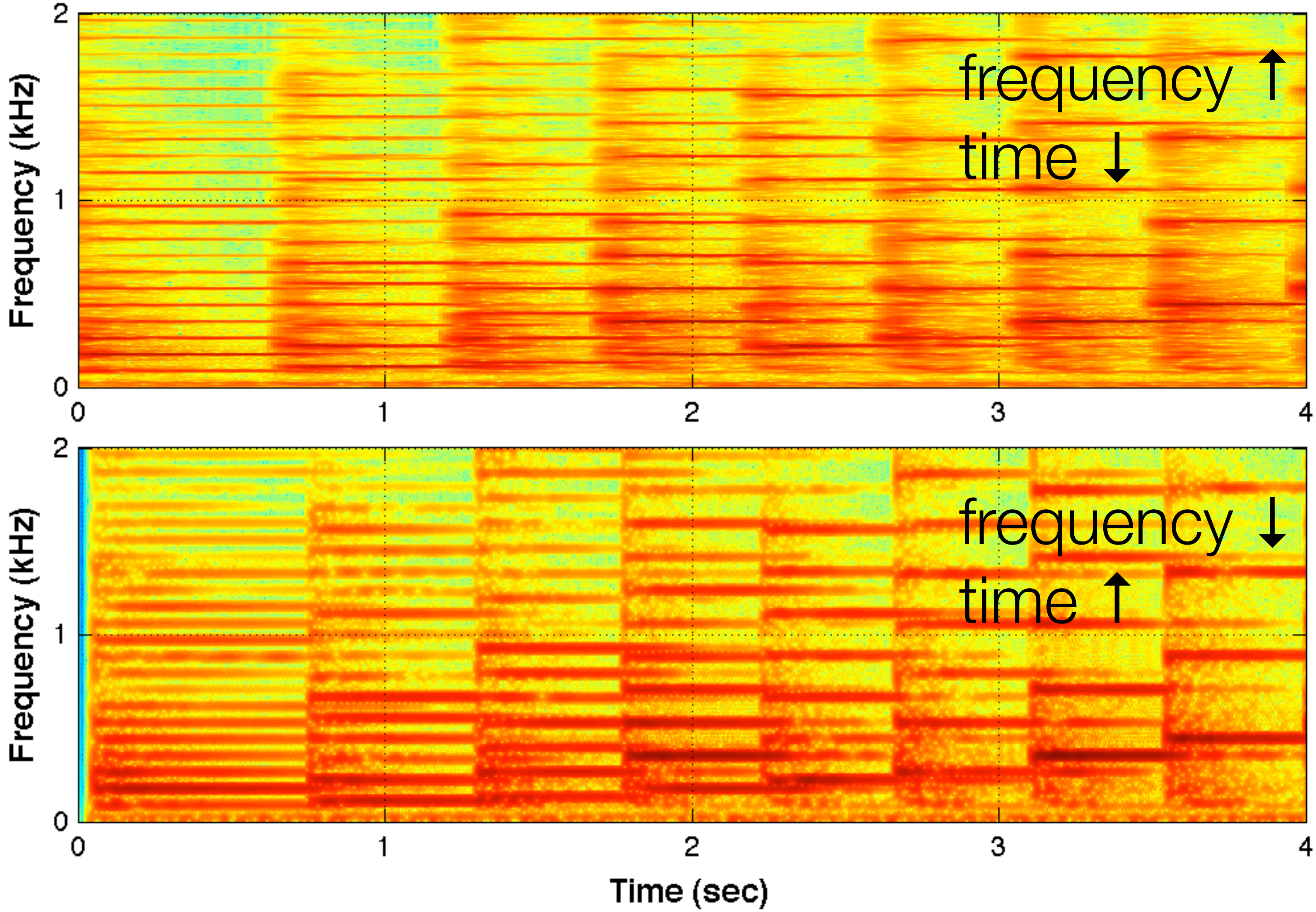


Spectrogram





Time vs Frequency Resolution



Instantaneous Frequency

The instantaneous frequency for frequency bin k at time instant mh can be defined as (Arfib et al., 2003):

$$f_i, k(m) = \frac{1}{2\pi}\frac{d\phi_k(m)}{dt} = \frac{1}{2\pi}\frac{\Delta\phi_k(m)}{h/f_s}$$

where,

$$\Delta\phi_k(m) = \Omega_k h + princarg[\phi_k(m) - \phi_k(m-1) - \boxed{\Omega_k}h]$$

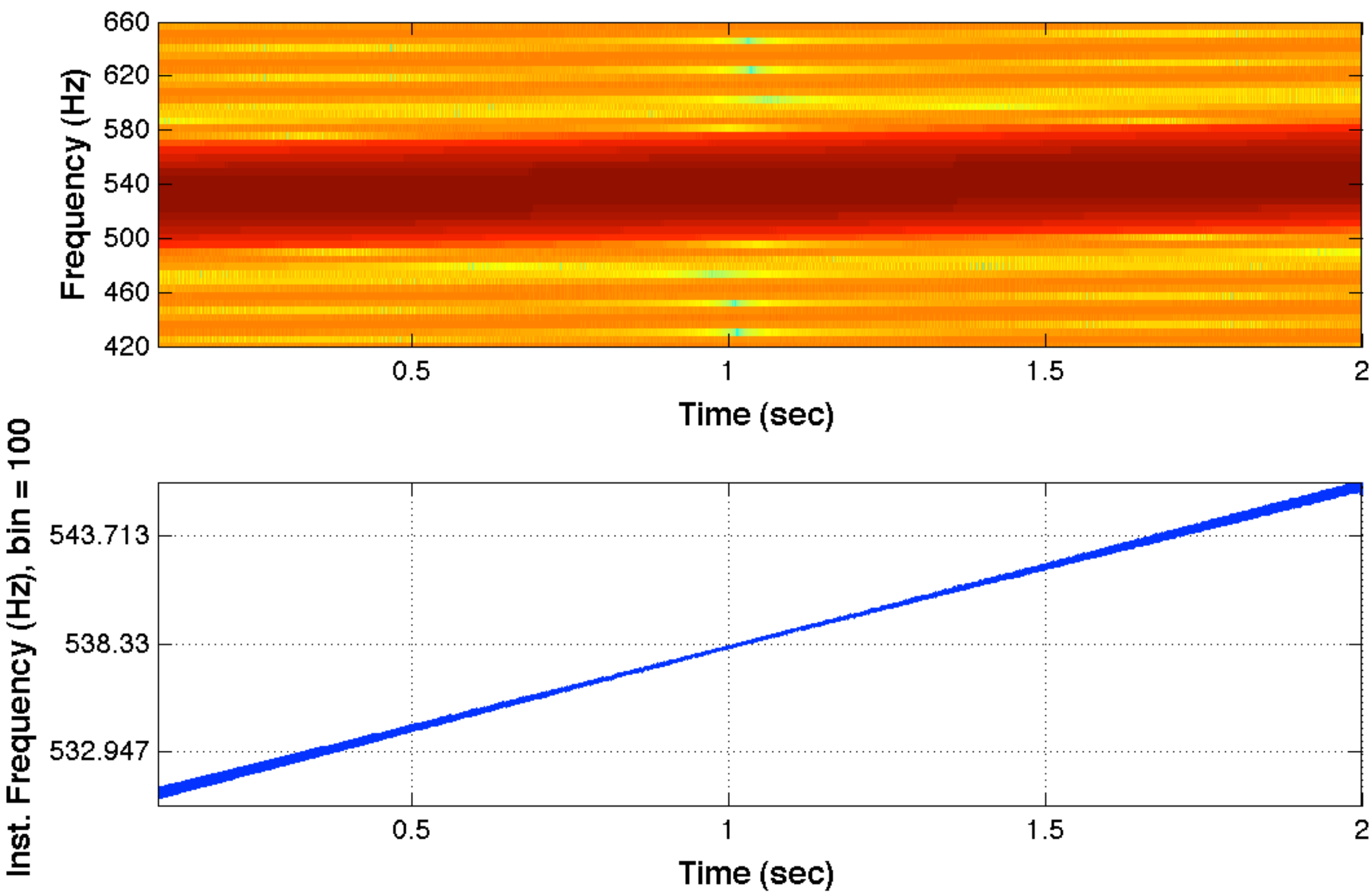
$$\frac{2\pi k}{N}$$

and

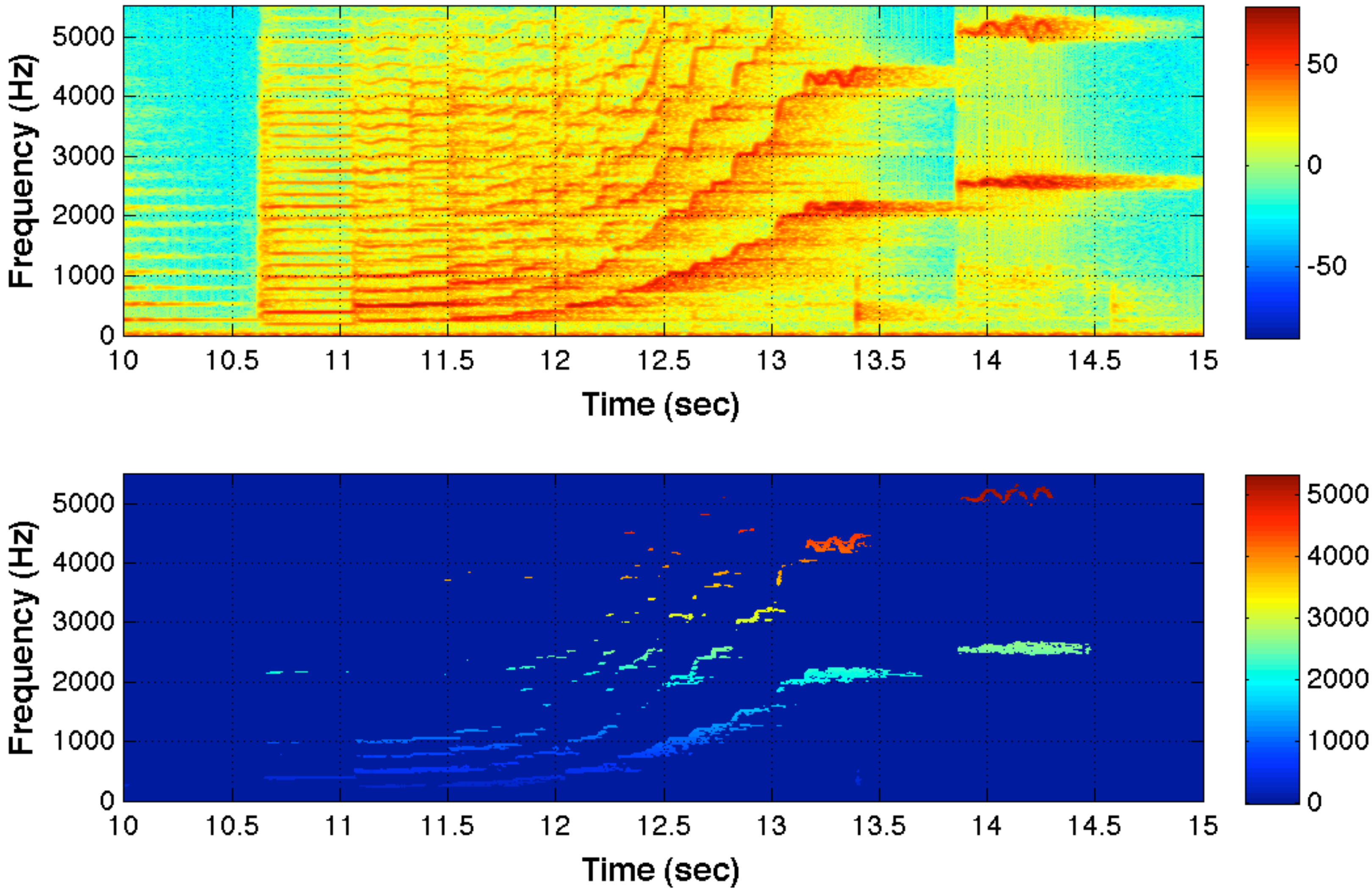
$$princarg(x) = \pi + [(x+\pi)mod(-2\pi)]$$

wraps the phase to the $(-\pi, \pi]$ range.

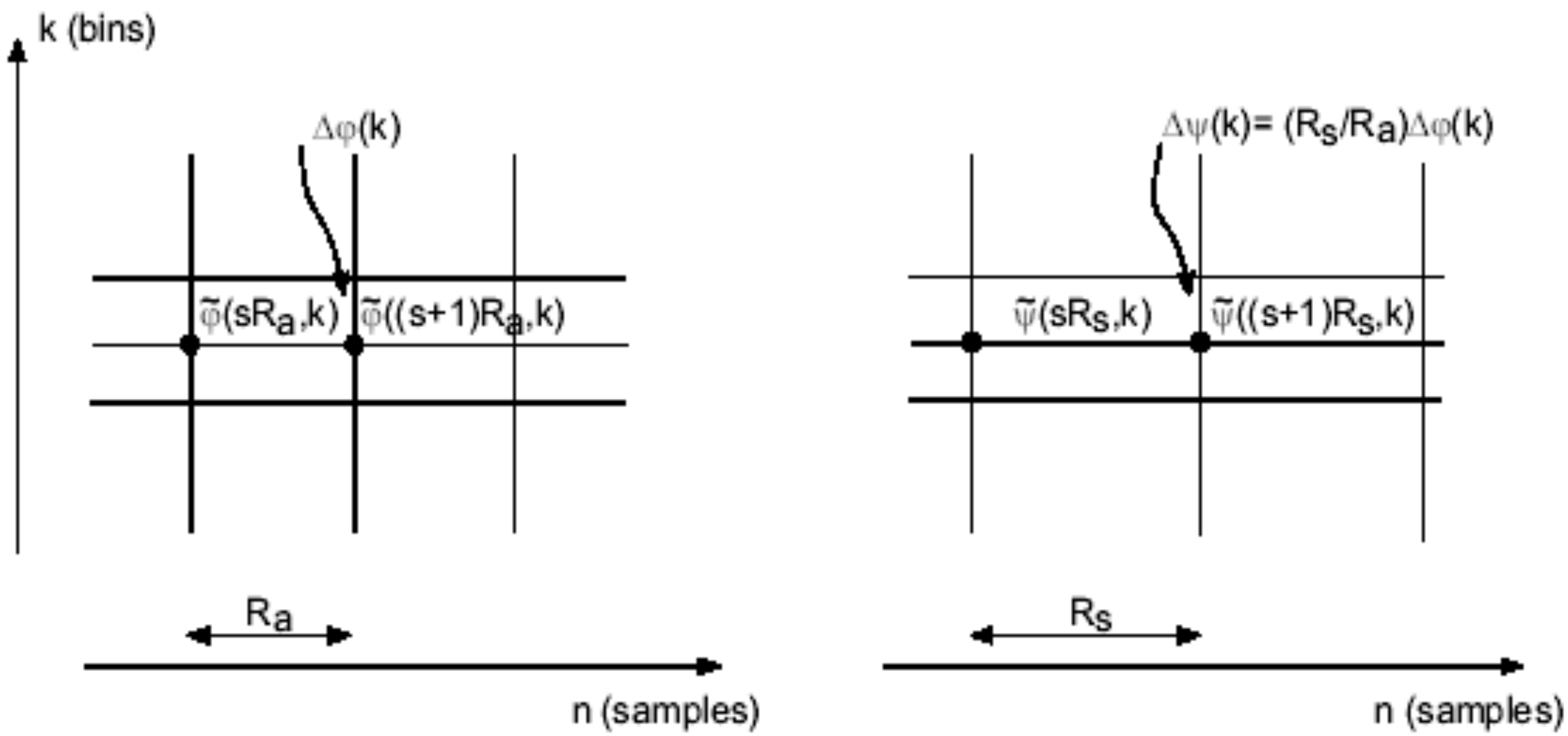
Instantaneous Frequency



Instantaneous Frequency



Time Scaling



* from DAFX book, chapter 8

- Solo guitar/polyphonic examples: (1) original, (2) standard, (3) adaptive

Sinusoidal Modeling

- The signal is approximated as a sum of time-varying sinusoidal components plus a residual:

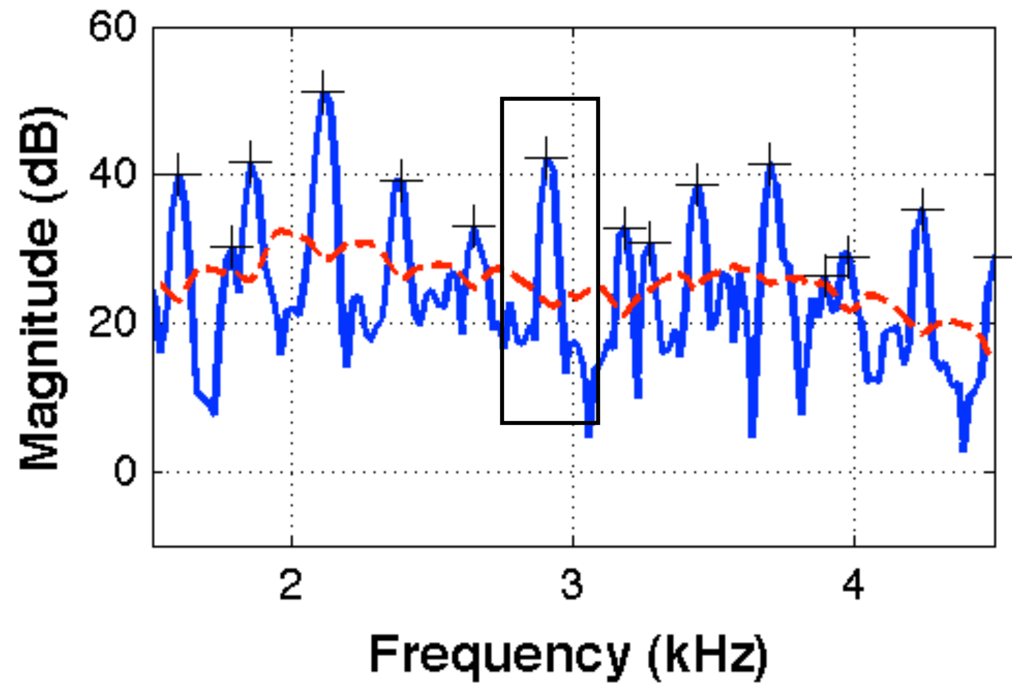
$$x(n) \approx \sum_{k=0}^{K} a_k(n) cos(\phi_k(n)) + e(n)$$

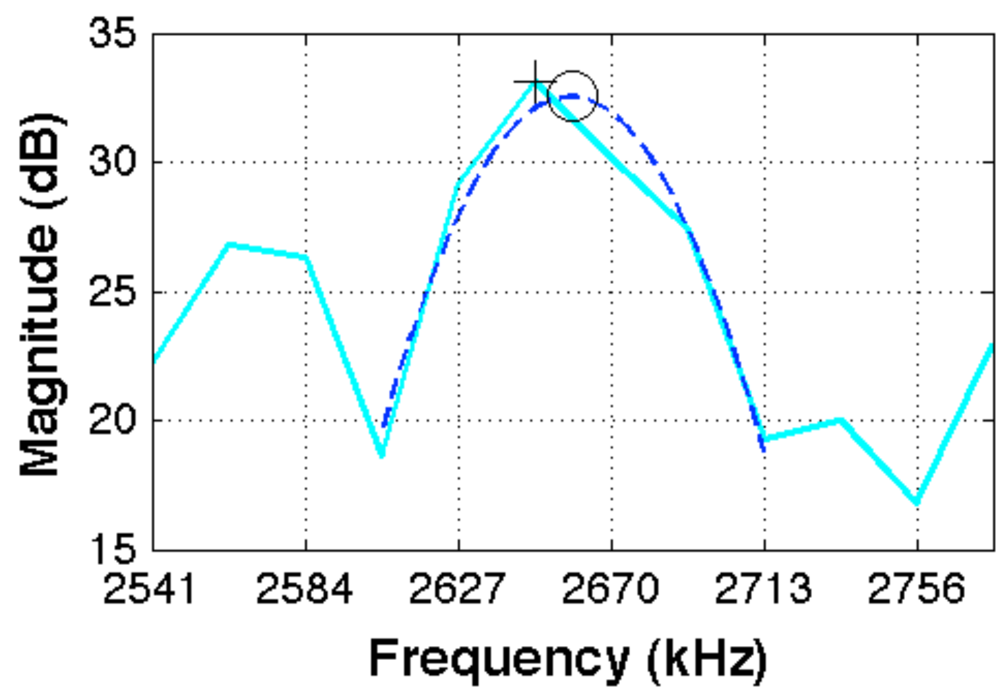
where,

$$a_k = \text{instantaneous amplitude}$$
$$\phi_k = \text{instantaneous phase}$$
$$e(n) = \text{residual (noise)}$$

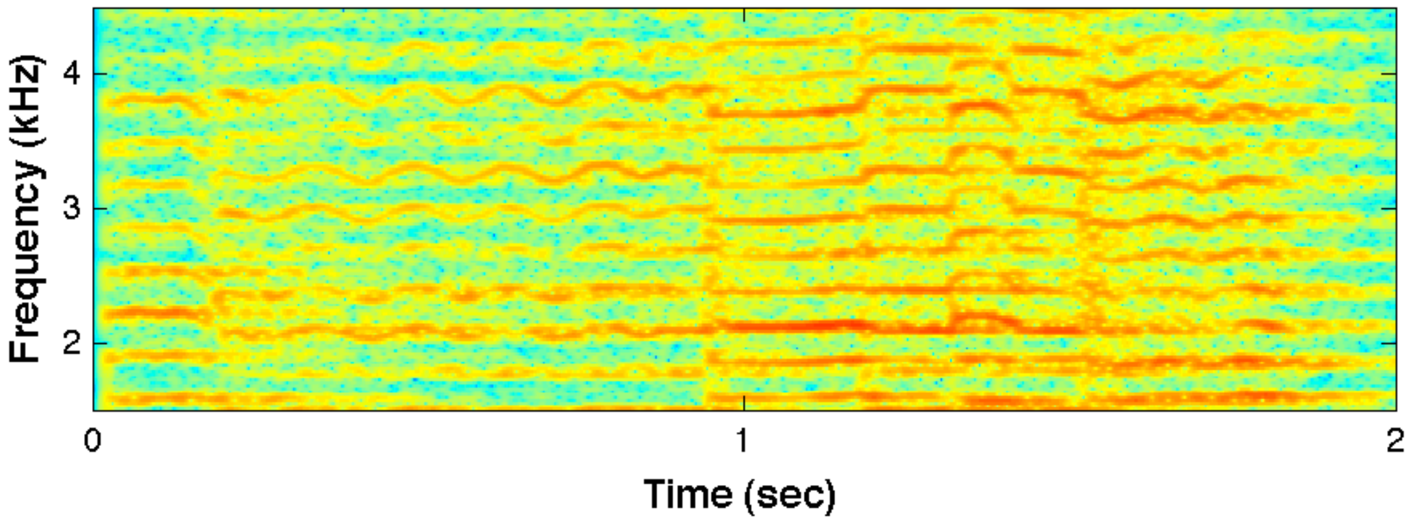
Peak picking + interpolation

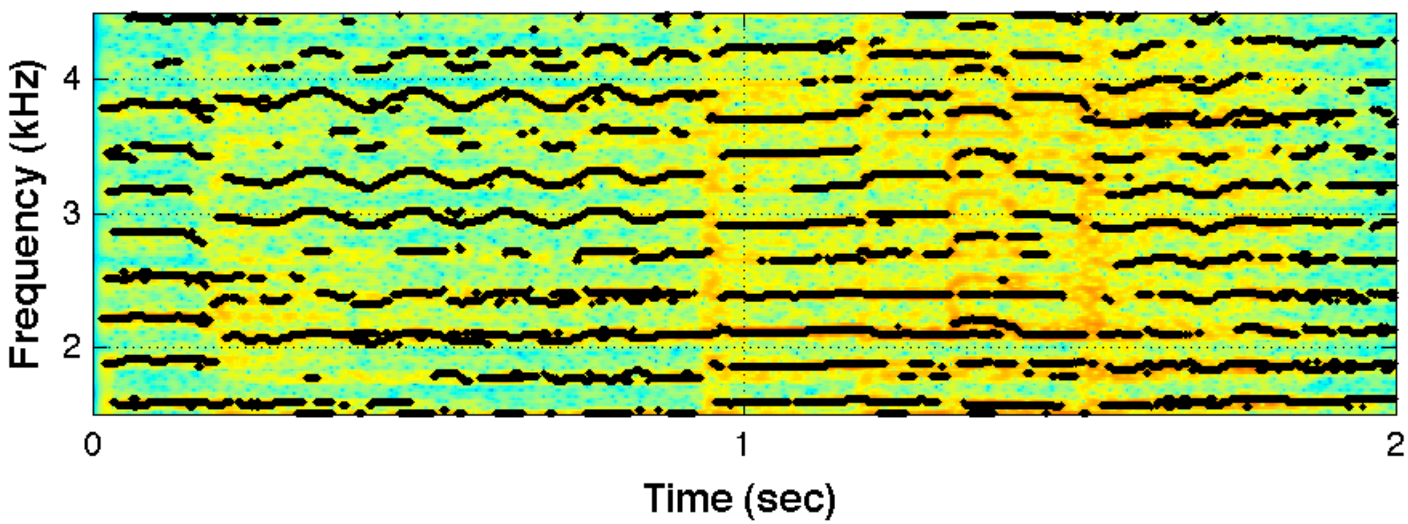
- •Sinusoidal components are peak-picked.
- •Instantaneous magnitude and phase values are obtained by interpolation.
- •Components are tracked over time



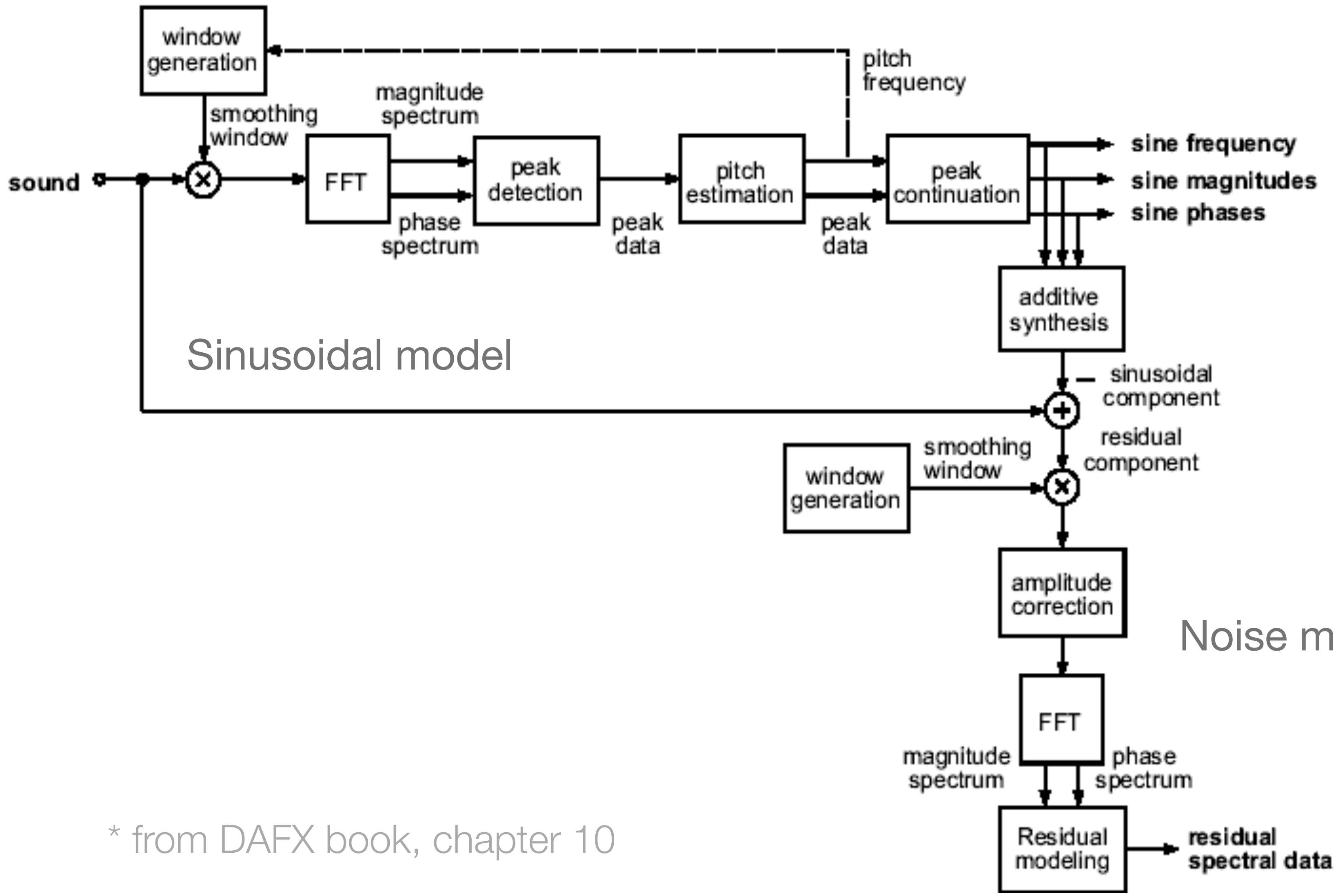


Sinusoidal Tracking





Sinusoidal Modeling

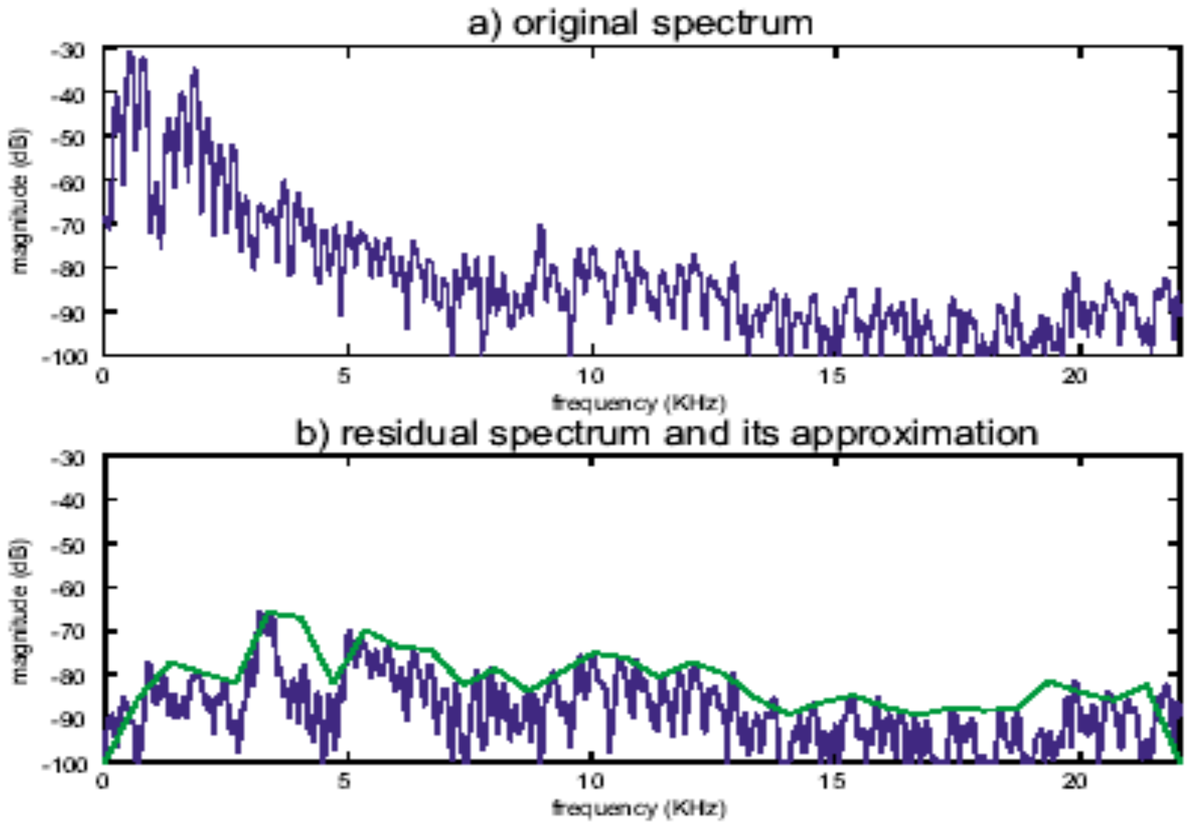


* from DAFX book, chapter 10

Noise Modeling

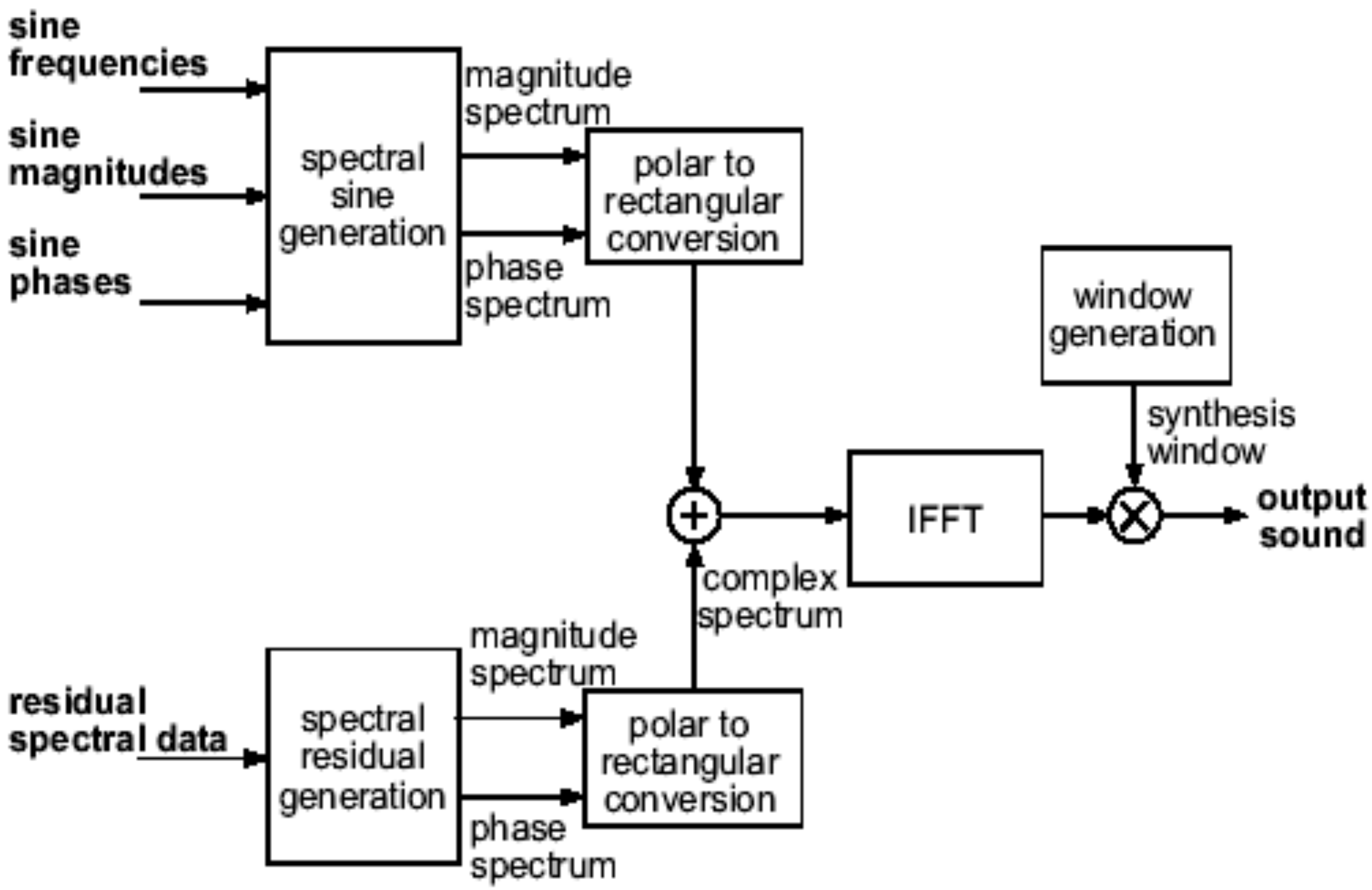
- We assume the residual to be a stochastic signal, i.e. can be described by its general spectral characteristics
- It is not necessary to maintain instantaneous phase or exact magnitudes.
- Hence, it can be modeled as the output of a time-varying filter driven by noise.
- The filter parameters encode the general spectral characteristics of the residual.
- Filter design usually involves an approximation of the spectral shape using: channel vocoder, LPC, Cepstrum, etc (see Lecture 5).

Noise Modeling



- Example approximation using max value per frequency band

Synthesis



* from DAFX book, chapter 10

Examples

http://mtg.upf.edu/technologies/sms?p=Sound%20examples

References

- Smith, J.O. “Mathematics of the Discrete Fourier Transform (DFT). 2nd Edition, W3K Publishing (2007)
- Zölzer, U. (Ed). “DAFX: Digital Audio Effects”. John Wiley and Sons (2002): chapter 8, Arfib, D., Keiler, F. and Zölzer, U., “Time-frequency Processing”; and chapter 10: Amatriain, X., Bonada, J., Loscos, A. and Serra, X. “Spectral Processing”.
- Pohlmann, K.C. “Principles of Digital Audio”. 6th Edition. McGraw Hill (2011): chapter 2.
- Loy, G. “Musimathics, Vol.2”. MIT Press (2011): chapter 3.
- Francis Rumsey and Tim McCormick (2002). “Sound and Recording: An Introduction”, Focal Press. (Chapter 1)
- Mitra, S. (2005). “Digital Signal Processing”. McGraw-Hill Science/ Engineering/Math; 3rd edition