s1

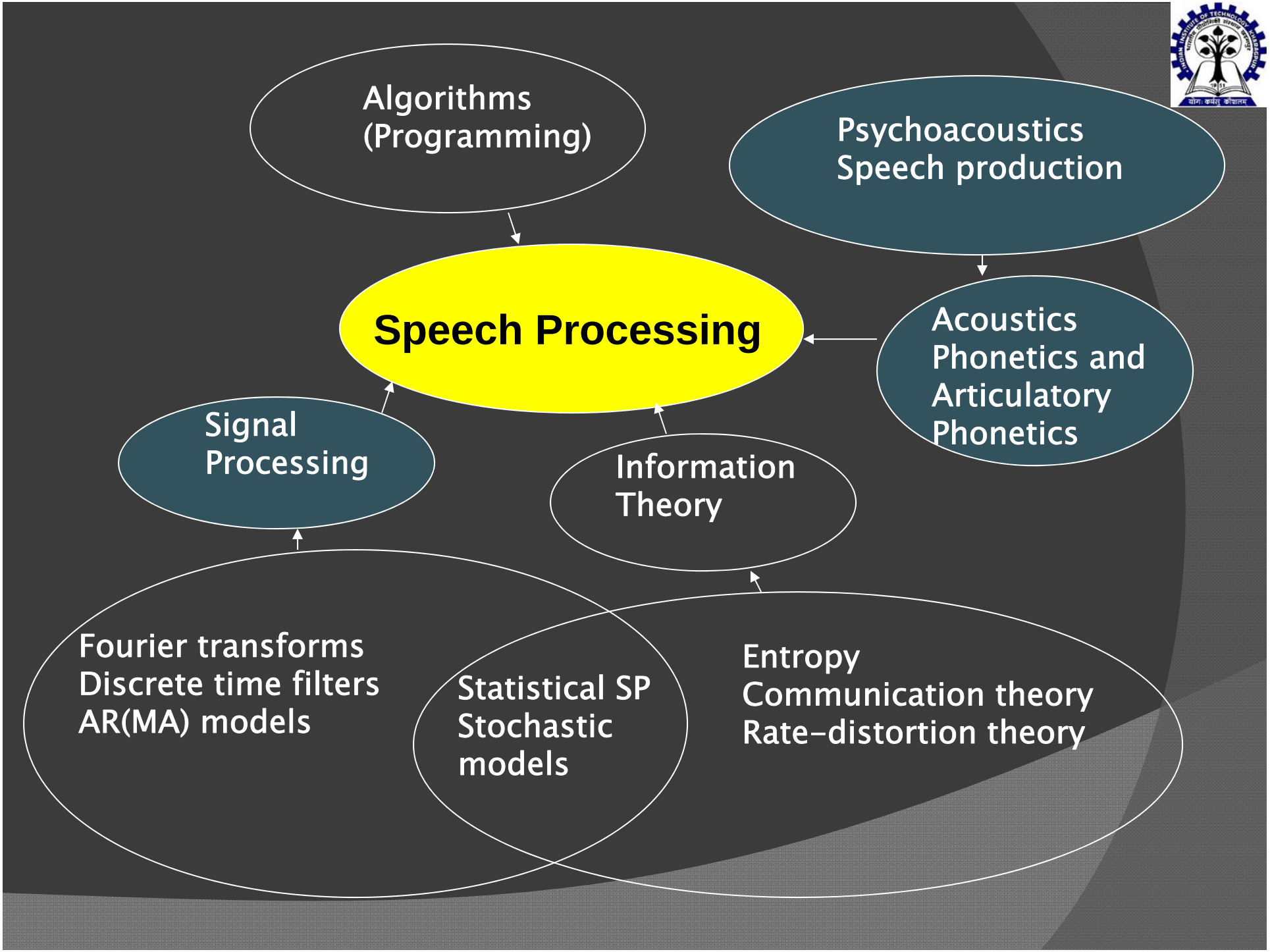
DIGITAL SPEECH PROCESSING

Dr. Shyamal Kumar Das Mandal
Assistant Professor

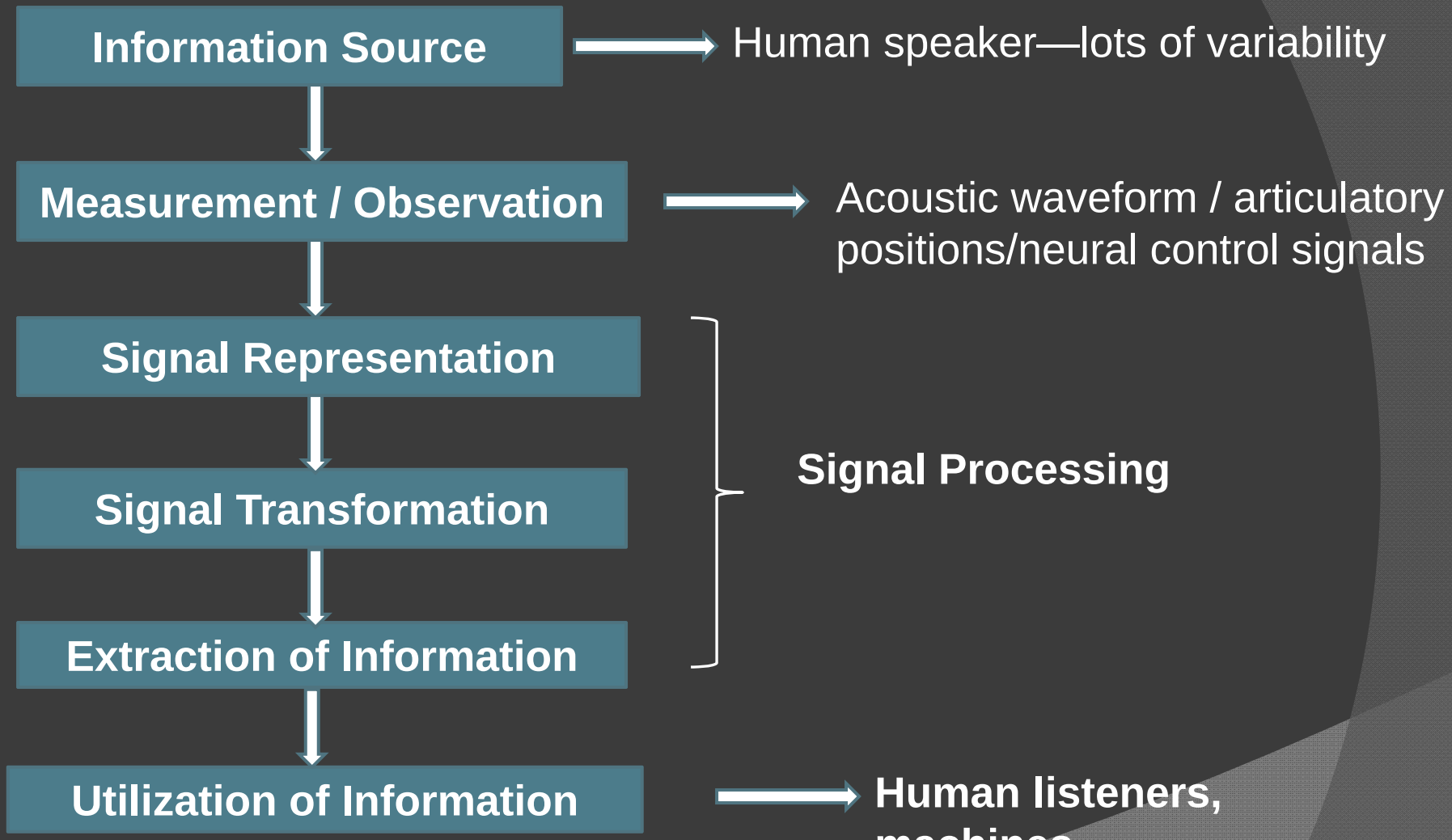
sdasmandal@cet.iitkgp.ernet.in
Centre for Educational Technology
Indian Institute of Technology, Kharagpur

Slide 1

s1 sdm, 12/14/10


Algorithms
(Programming)
Psychoacoustics
Speech production
Speech Processing
Acoustics
Phonetics and
Articulatory
Phonetics
Signal
Processing
Information
Theory
Fourier transforms
Discrete time filters
AR(MA) models
Statistical SP
Stochastic
models
Entropy
Communication theory
Rate-distortion theory

Speech Processing Model



Course coverage

- Digitization and Recording of speech signal, Review of Digital Signal Processing Concepts

- Human Speech production, Acoustic Phonetics and Articulatory Phonetics, Different categories speech sounds and Location of sounds in the acoustic waveform and spectrograms

- Uniform Tube Modeling of Speech Production, Speech Perception

- Time Domain Methods in Speech Processing, Analysis and Synthesis of Pole-Zero Speech Models

- Short-Time Fourier Transform, Analysis:- FT view and Filtering view, Synthesis:-Filter bank summation (FBS) Method and OLA Method

- Features Extraction and Extraction of Fundamental frequency

- Speech Prosody, Speech Prosody Modeling (Fujisaki Model)

- Overview of Speech based Applications development (TTS, ASR and spoken language acquisition)

Course objective

- Categories and labeling of different speech sound for a given speech signal based on waveform and spectrographic view
- Explain the psychoacoustic properties of speech perception and production
- Design the Uniform tube model for speech sound production and implement it based on discrete time modeling
- Extract the fundamental frequency of speech signal based on time domain and frequency domain method
- Extract spectral parameters and time domain parameters of speech signal for speech technology application
- Design an simple TTS and ASR system.
- Explain the prosodic structure of spoken language and design F_0 contour modeling based on Fujisaki Model

Review of DSP Concepts

Concept of frequency in continuous-time and Discrete-time signal

Continuous sinusoidal Time Signal $x_a(t) = A\cos(\Omega t + \theta)$

- For every fixed value of the frequency F *x(t)* is periodic
- Continuous time sinusoidal signal with distinct frequencies are themselves distinct.
- Increase the frequency result in increase in the rate of oscillation of the signal → more period are included.

Complex exponent from $x_a(t) = Ae^{j(\Omega t+\theta)}$

Discrete time Sinusoidal

$$x(n) = A\cos(\omega n + \theta)$$

- A discrete time signal is periodic if its frequency f is a rational number
- Discrete time sinusoidal whose frequency are separated by an integer multiple of 2π are identical
- The highest rate of oscillation in a discrete time sinusoidal is attained when w= π or(- π).

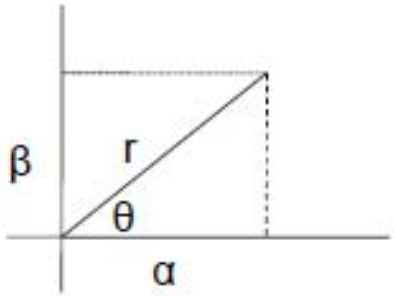
Continuous time signal	Discrete time signal
$\Omega = 2\pi F$ Ω is in Radians/sec and F is in Hz $-\infty < \Omega < \infty$ $-\infty < F < \infty$	$\omega = 2\pi f$ $\omega$is in Radians/sample and f is in cycles/sample $-\pi < \omega < \pi$ $-1/2 < f < 1/2$
$\Omega = \omega / T_s \qquad F = f.F_s$	$\omega = \Omega.T_s \quad f = F / F_s$
Where F_s is the sampling frequency and $T_s=1/F_s$	

Complex Signal

$$x[n] = (\alpha + j\beta)^n u[n] = (re^{j\theta})^n u[n]$$

$$r = \sqrt{\alpha^2 + \beta^2}$$

$$\theta = \tan^{-1}(\beta / \alpha)$$



$$x[n] = r^n e^{j\theta n} u[n]$$

r^n is a dying exponential

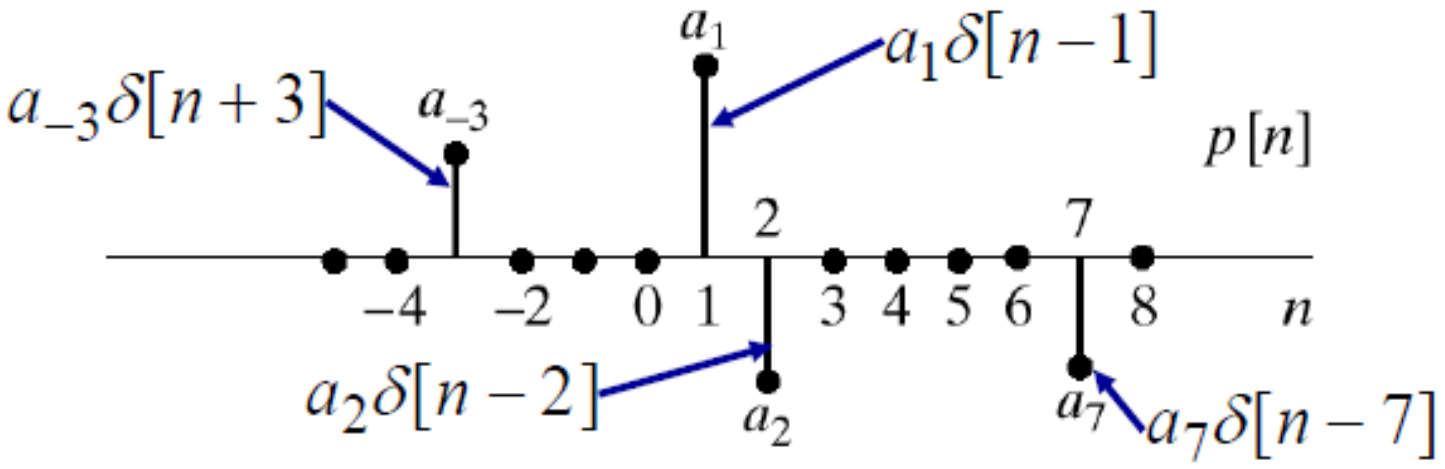
$e^{j\theta n}$ is a linear phase term

Impulse Representation of Sequences

$$x[n] = \sum_{k=-\infty}^{\infty} x[k]\delta[n-k]$$

A sequence, a function

Value of the function at k



$$p[n] = a_{-3}\delta[n+3] + a_1\delta[n-1] + a_2\delta[n-2] + a_7\delta[n-7]$$

Classification of Discrete time signal

Energy signal and power signal: If E is the energy of a signal x[n]

$$E \equiv \sum_{n=-\infty}^{\infty} \left|x[n]\right|^2$$

If E is finite then x(n) is called an energy signal

Many signal possesses infinite energy, have a finite average power P. The average power define as:

$$P = \lim_{N->\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \left|x[n]\right|^2$$

If P is finite and nonzero then the signal is called a power signal

Periodic and aperiodic signal: a signal x[n] is periodic with period N if and only if x[n+N]=x[n] for all n

The smallest value of N for which holds the above equation is called fundamental period.

If there is no value of N that satisfies the above equation then the signal is called aperiodic signal.

Symmetric and antisymmetric signal: a real-valued signal x[n] is called symmetric if x[-n]=x[n]

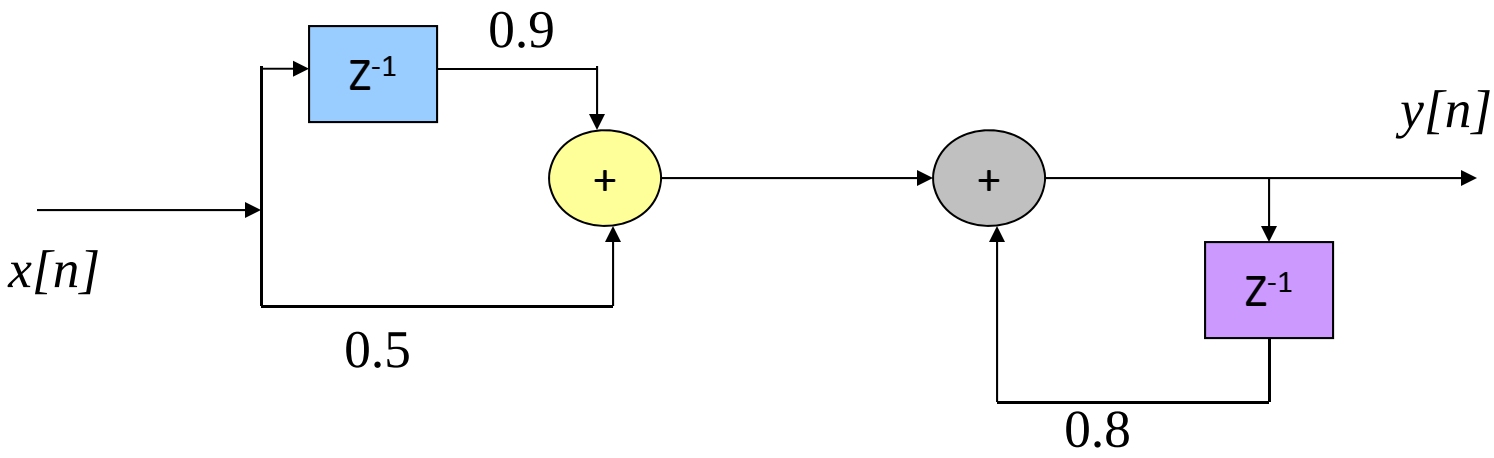
On the other hand a signal x[n] is called antisymmetric if x[-n]=-x[n]

Discrete time system

Discrete time system is a device or algorithm that operate on a discrete time signal, called the input according to some well define rule, to produce another discrete time signal called output of the system.

$$y[n]=H[x[n]]$$

$$y[n]=0.8y[n-1]+0.5x[n]+0.9x[n-1]$$



Classification of discrete system

• Static and Dynamic system

A discrete time system is called static or memory less it its output at any instant n depends at most on the input sample at the same time, but not on past or future samples of the input. In any other case the system is said to be dynamic or to have memory.

• Time invariant and time variant system

A system is called time invariant if its input-output characteristics do not change with time.

A relaxed system H is time invariant or shift invariant if and only if

$x[n] \xrightarrow{H} y[n]$ implies that $x[n-k] \xrightarrow{H} y[n-k]$

If the output y[n-k] not = y[n-k] even for one value of k the system is time variant

•**Linear and non-linear system:** A linear system is one that satisfies the superposition principle. The principle of superposition requires that the response of the system to a weighted sum of signal is equal to the corresponding weighted sum of the response of the system to each of the individual input signal.

• **Causal and non-causal system**: A system is said to be causal if the output of the system at any time n depends only on present and past input, but not depend on future inputs.

$y[n]=F[x[n],x[n-1],x[n-2]\ldots\ldots\ldots x[n-k]]$ where F is any function

If a system does not satisfy the above condition then the system is called Non-causal system.

•**Stable and unstable system:** An arbitrary relaxed system is said to be bounded input-bounded output(BIBO) stable if and only if every bounded input produces a bounded output. x[n], y[n] are bounded is simply translated mathematically to mean that there exist some finite numbers say M_x, M_y such that

$$|x[n]| \leq M_x \leq \infty \qquad |y[n]| \leq M_y \leq \infty$$

for all n

Recursive and Non-recursive discrete system.

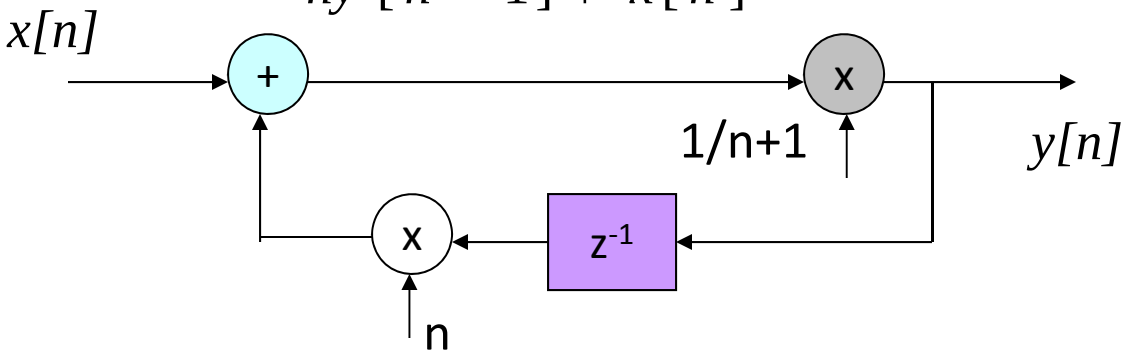
$$y[n] = \sum_{k=-\infty}^{\infty} h[k]x[n-k]$$

Cumulative average of signal x(n)

$$y[n] = \frac{1}{n+1}\sum_{k=0}^{n} x[k]$$

$$(n+1)\,y[n] = \sum_{k=0}^{n-1} x[k] + x[n]$$

$$= ny[n-1] + x[n]$$



A system whose output y[n] at time n depends on any number of past output values y[n-1], y[n-2] …. Is called recursive system

Convolution

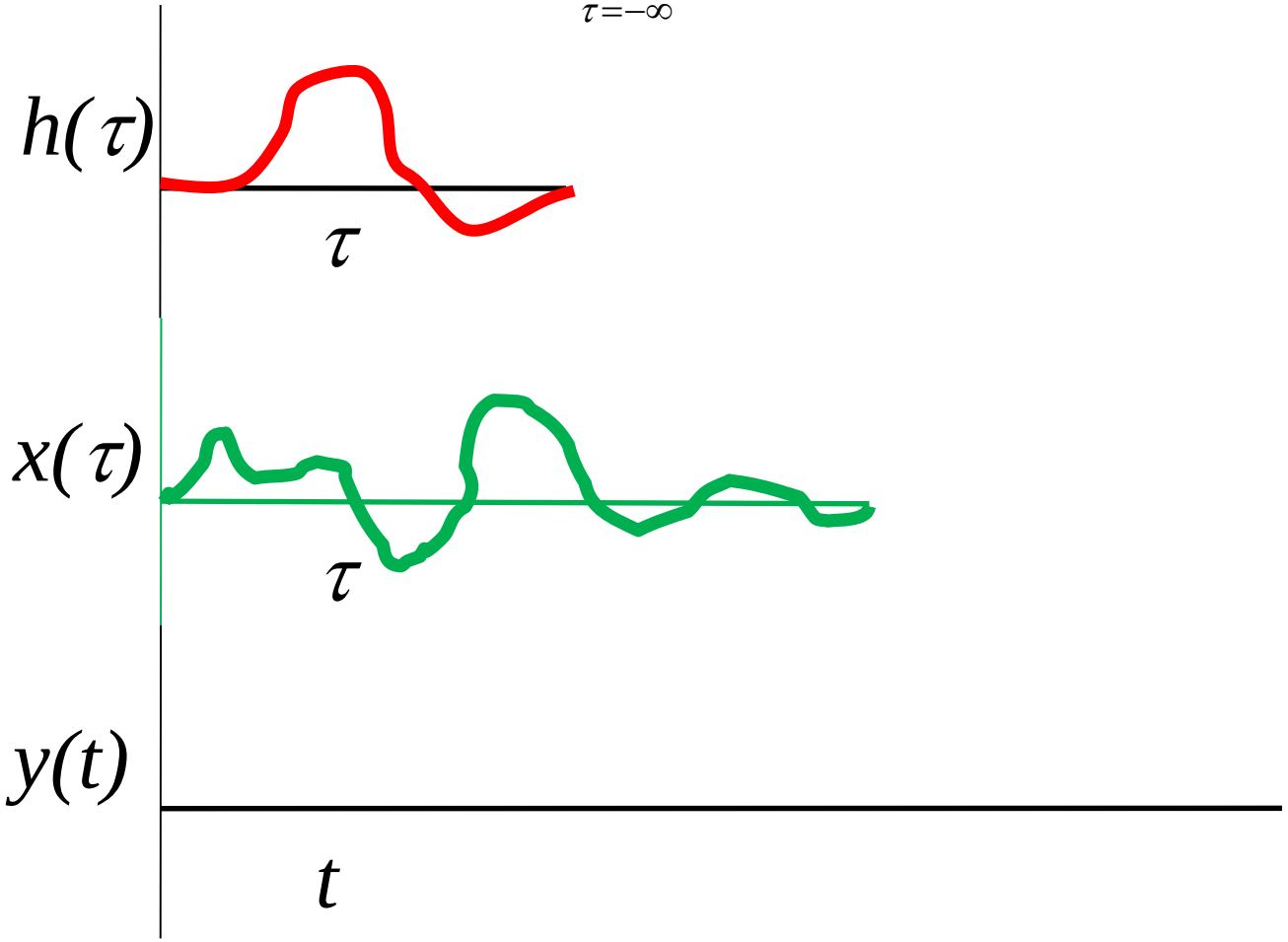
Convolution is one of the most frequently used operations in DSP. Specially in digital filtering applications where two finite and causal sequences x[n] and h[n] of lengths N_1 and N_2 are convolved

$$y[n] = h[n] \otimes x[n] = \sum_{k=-\infty}^{\infty} h[k]x[n-k] = \sum_{k=0}^{\infty} h[k]x[n-k]$$

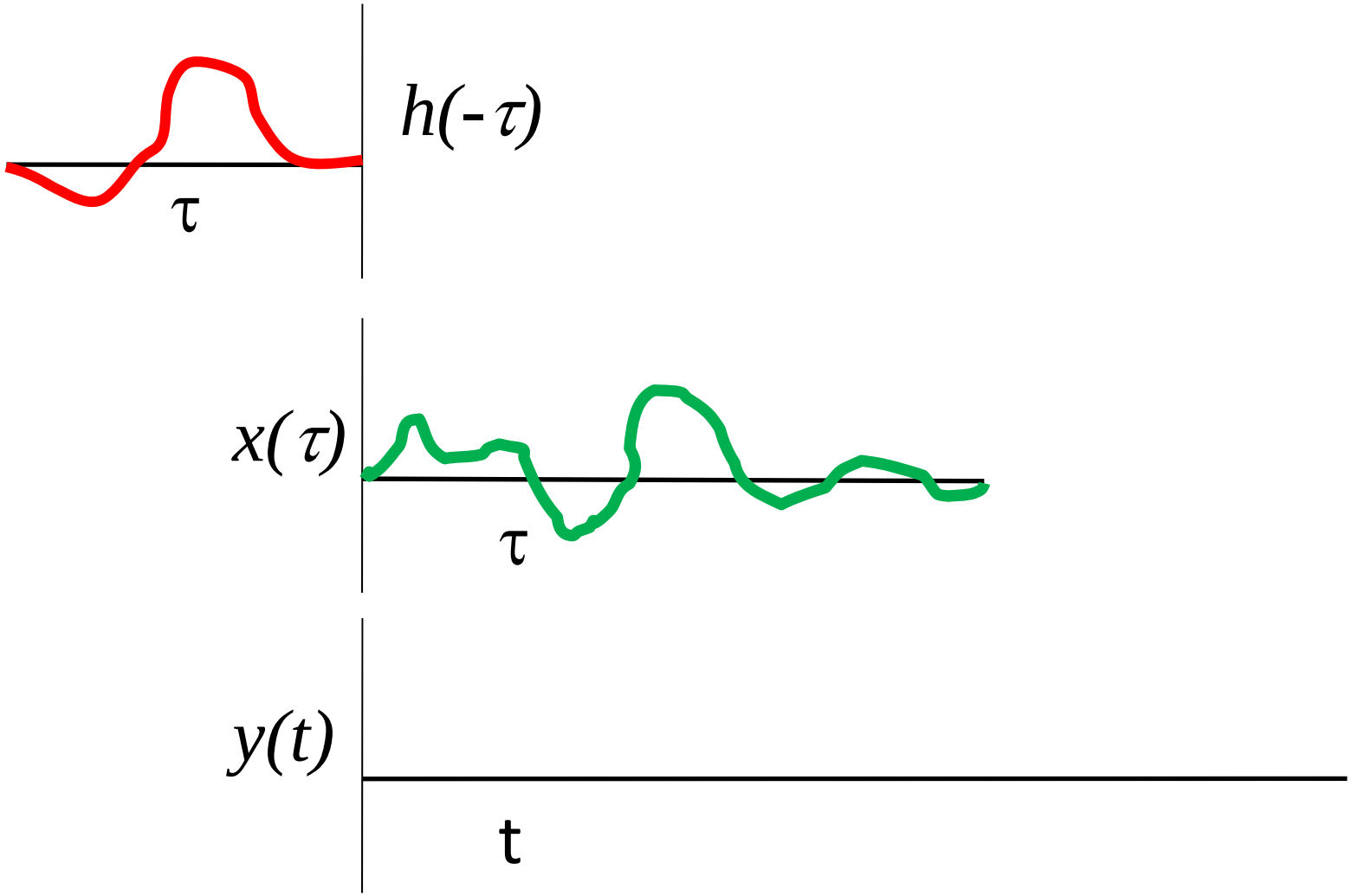
Operations involved

- **Folding**
- **Shifting**
- **Multiplication**
- **Summation**

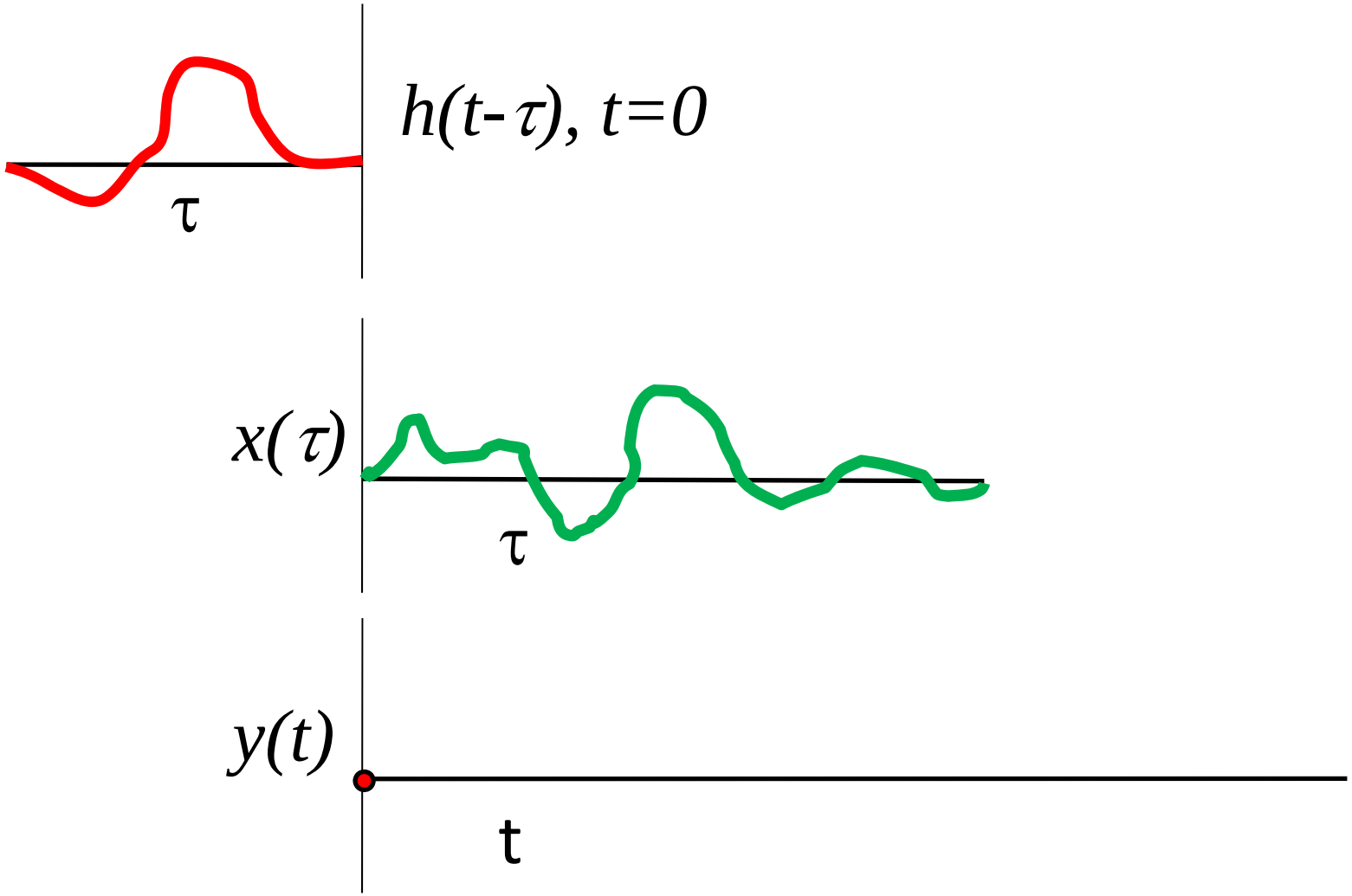
$$y(t) = \sum_{\tau=-\infty}^{\infty} x(\tau)h(t-\tau)$$



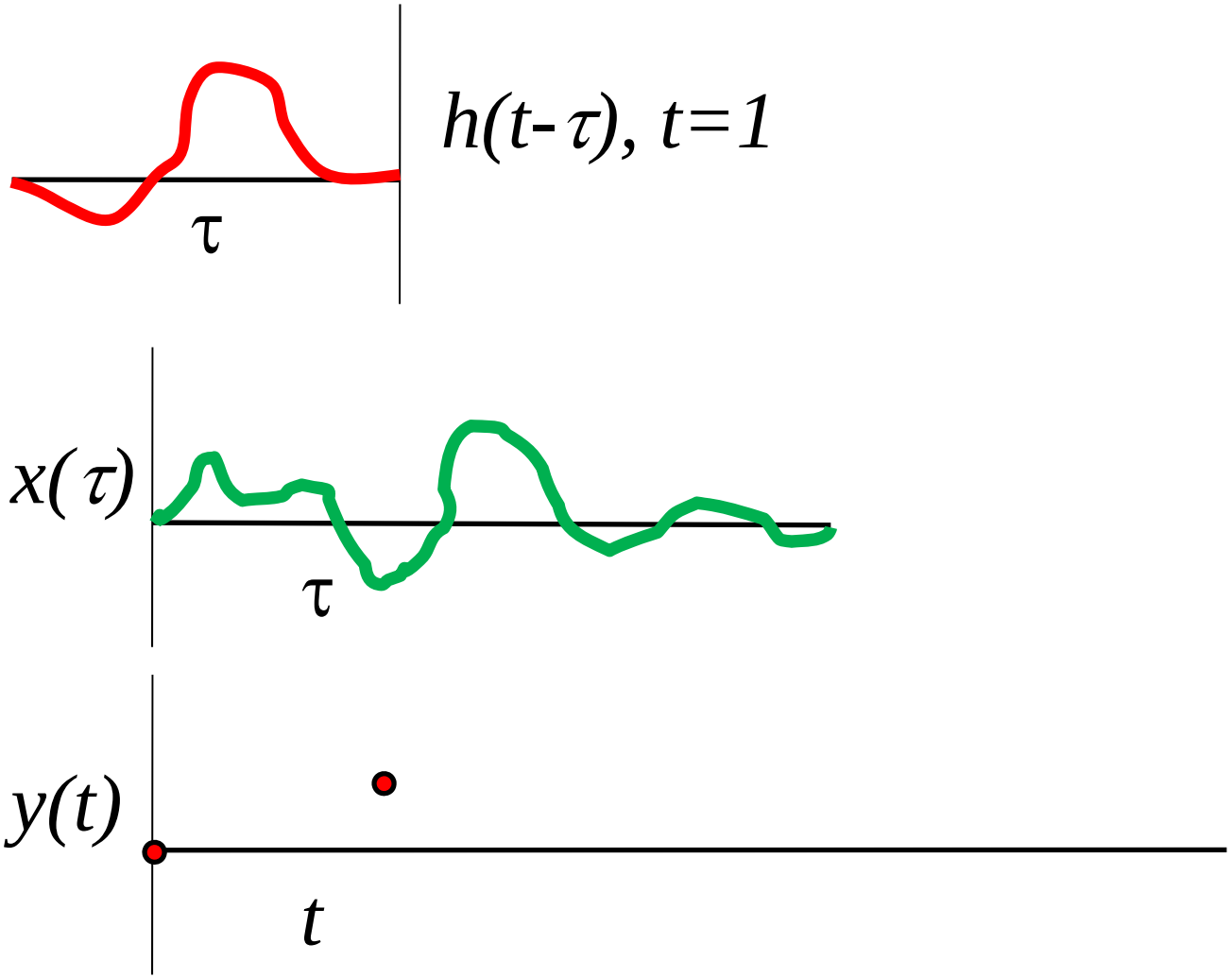
$$y(t) = \sum_{\tau=-\infty}^{\infty} x(\tau)h(t-\tau)$$



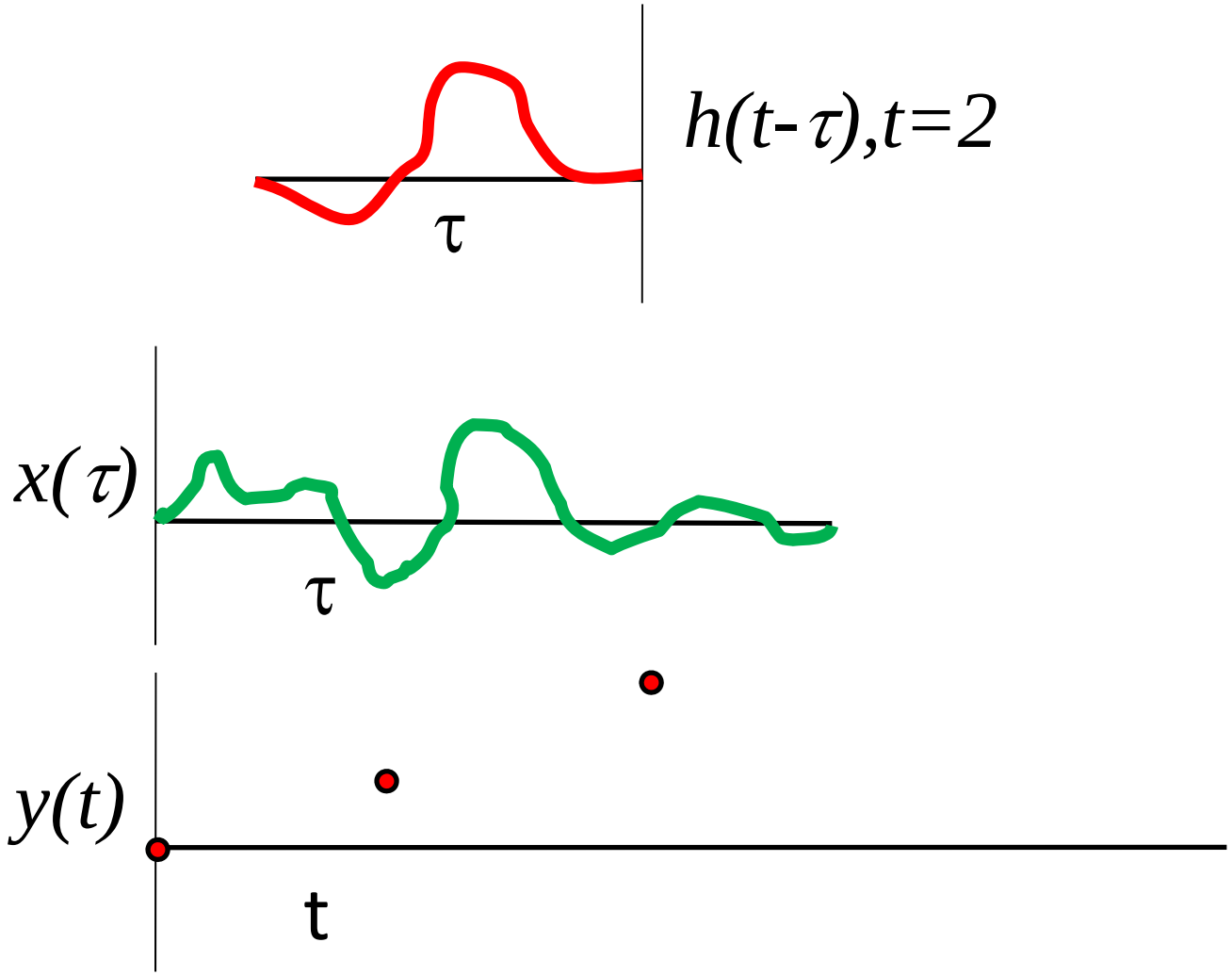
$$y(t) = \sum_{\tau=-\infty}^{\infty} x(\tau)h(t-\tau)$$



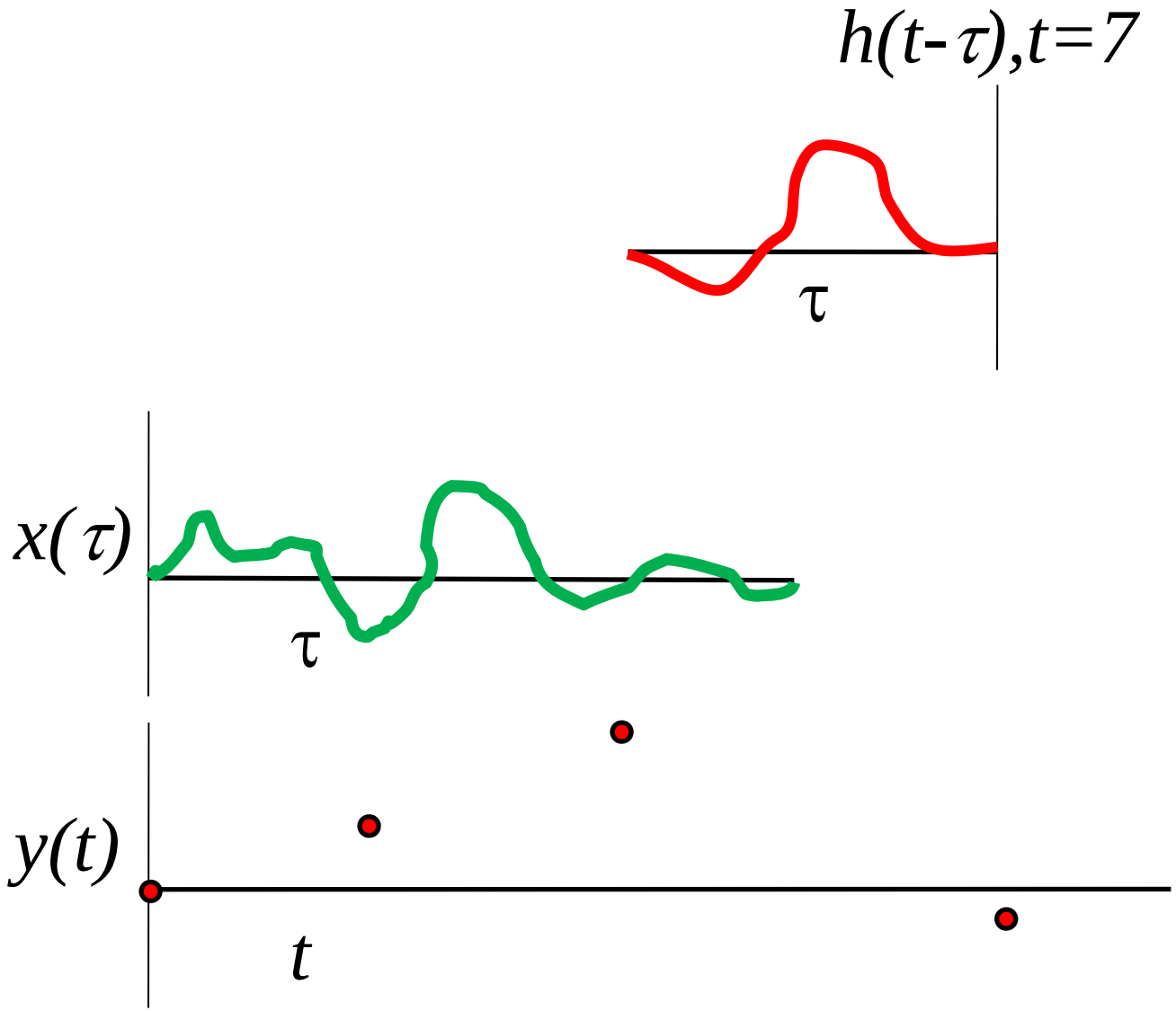
$$y(t) = \sum_{\tau=-\infty}^{\infty} x(\tau)h(t-\tau)$$



$$y(t) = \sum_{\tau=-\infty}^{\infty} x(\tau)h(t-\tau)$$

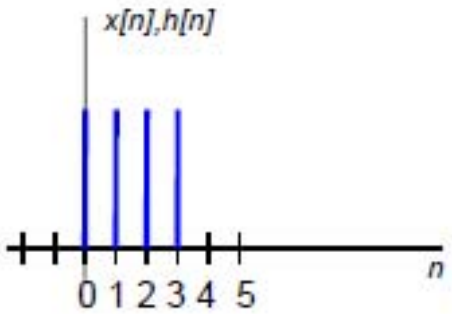


$$y(t) = \sum_{\tau=-\infty}^{\infty} x(\tau)h(t-\tau)$$



Convolution Example

$$x[n]=\begin{cases}1 & 0\le n\le 3\\ 0 & \text{otherwise}\end{cases} \qquad h[n]=\begin{cases}1 & 0\le n\le 3\\ 0 & \text{otherwise}\end{cases}$$

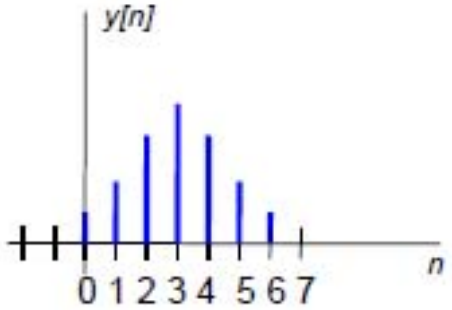


What is $y[n]$ for this system?

Solution :

$$y[n]=x[n]*h[n]=\sum_{m=-\infty}^{\infty}h[m]\,x[n-m]$$

$$=\begin{cases}\sum_{m=0}^{n}1\cdot 1=(n+1) & 0\le n\le 3\\ \sum_{m=n-3}^{3}1\cdot 1=(7-n) & 4\le n\le 6\\ 0 & n\le 0,\ n\ge 7\end{cases}$$



Circular convolution

• *Circular convolution of x(n) and h(n) is defined as the convolution of h(n) with a periodic signal $x_p(n)$:*

$$y_p[n] = x_p[n] * h[n]$$

$$y_p[n] = \sum_{n=0}^{N-1} h[n] x_p[m-n]_N$$

$$m = 0,1,.....N-1$$

where

$$x_p(n) = x(n \bmod N), \qquad -\infty < n < \infty$$

Correlation

Correlation is a mathematical operation that is very similar to convolution. Just as with convolution, correlation uses two signals to produce a third signal. This third signal is called the **cross-correlation** of the two input signals.

If a signal is correlated with *itself*, the resulting signal is instead called the **autocorrelation**.

Correlation

$$r_{xy}(l) = \sum_{n=-\infty}^{\infty} x(n)\,y(n-l) \qquad l = 0,\pm 1,\pm 2,\ldots$$

$$r_{xy}(l) = \sum_{n=-\infty}^{\infty} x(n+l)\,y(n) \qquad l = 0,\pm 1,\pm 2,\ldots$$

Where, $r_{xy}(l)$ is the correlation coefficients

Computation of correlation

$$r_{xy}(l) = \begin{cases} \sum_{n=l}^{M-1+l} x(n)y(n-l) & 0 \le l \le N-M \\ \sum_{n=l}^{N-1} x(n)y(n-l) & N-M \le l \le N-1 \end{cases}$$

```
FOR l=1 to lmax
{
  NL=M+1-l
IF(NL>N-1) NL=NL-1
R(L)=0.0
FOR(K=l TO NL
{
 R(l)=R(l)+X(K)*Y(K-l)
}
}
```

Convolution vs. correlation

- Convolution is the relationship between a system's input signal, output signal, and impulse response.
- Correlation is a way to detect a known waveform in a noisy background.
- The similar mathematics is only a convenient coincidence.

LTI Discrete-Time Systems

- Linearity (superposition):

$$\mathrm{T}\{ax_1[n]+bx_2[n]\} = a\mathrm{T}\{x_1[n]\}+b\mathrm{T}\{x_2[n]\}$$

- Time-Invariance (shift-invariance):

$$x_1[n] = x[n-n_d] \quad \Rightarrow \quad y_1[n] = y[n-n_d]$$

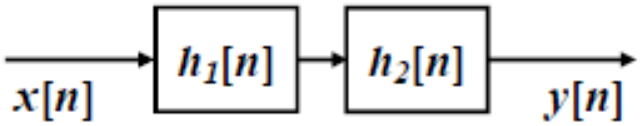
- LTI implies discrete convolution:

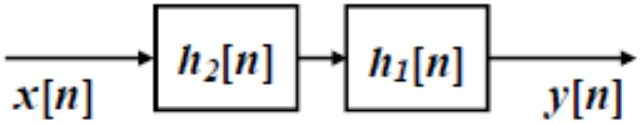
$$y[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k] = x[n]*h[n] = h[n]*x[n]$$

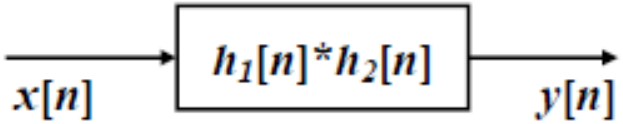
Linear Time-Invariant Systems

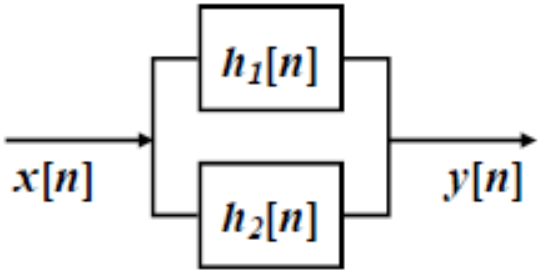
- Easiest to **understand**
- Easiest to **manipulate**
- **Powerful processing** capabilities
- Characterized completely by their response to unit sample, *h(n), via* ***convolution relationship***
- Basis for **linear filtering**
- Used as **models for speech production**

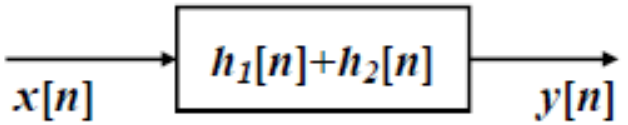
Equivalent LTI Systems





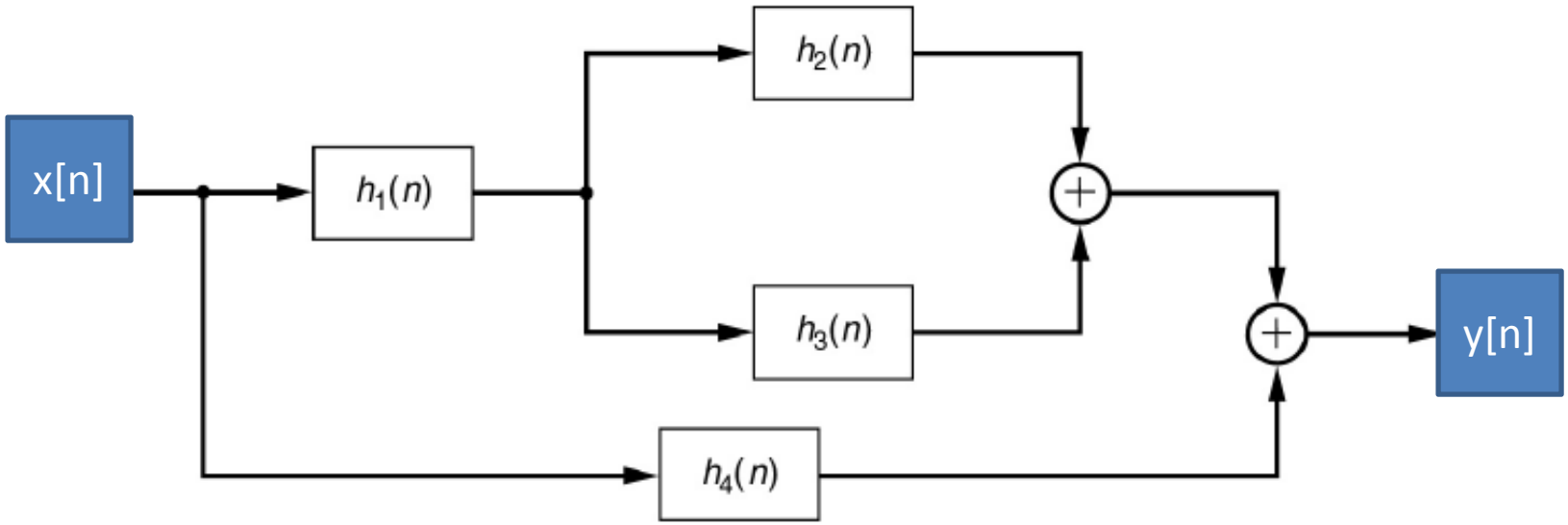






$$h_1[n]*h_2[n]= h_2[n]*h_1[n]$$

$$h_1[n]+h_2[n]= h_2[n]+h_1[n]$$



Find y[n]?

Direct Form I

$$y[n] = -\sum_{k=1}^{N} a_k y[n-k] + \sum_{k=0}^{M} b_k x[n-k]$$

- Transfer function of recursive LTI system

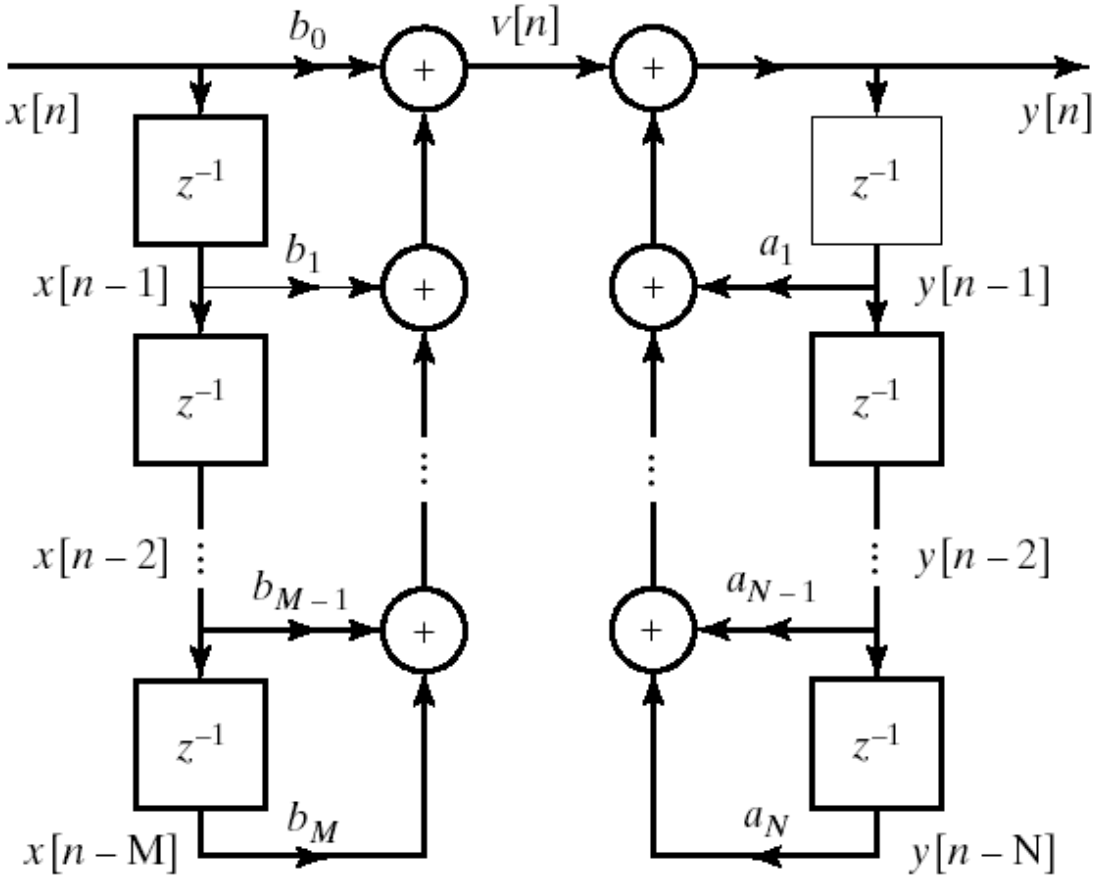
$$v[n] = \sum_{k=0}^{M} b_k x[n-k]$$

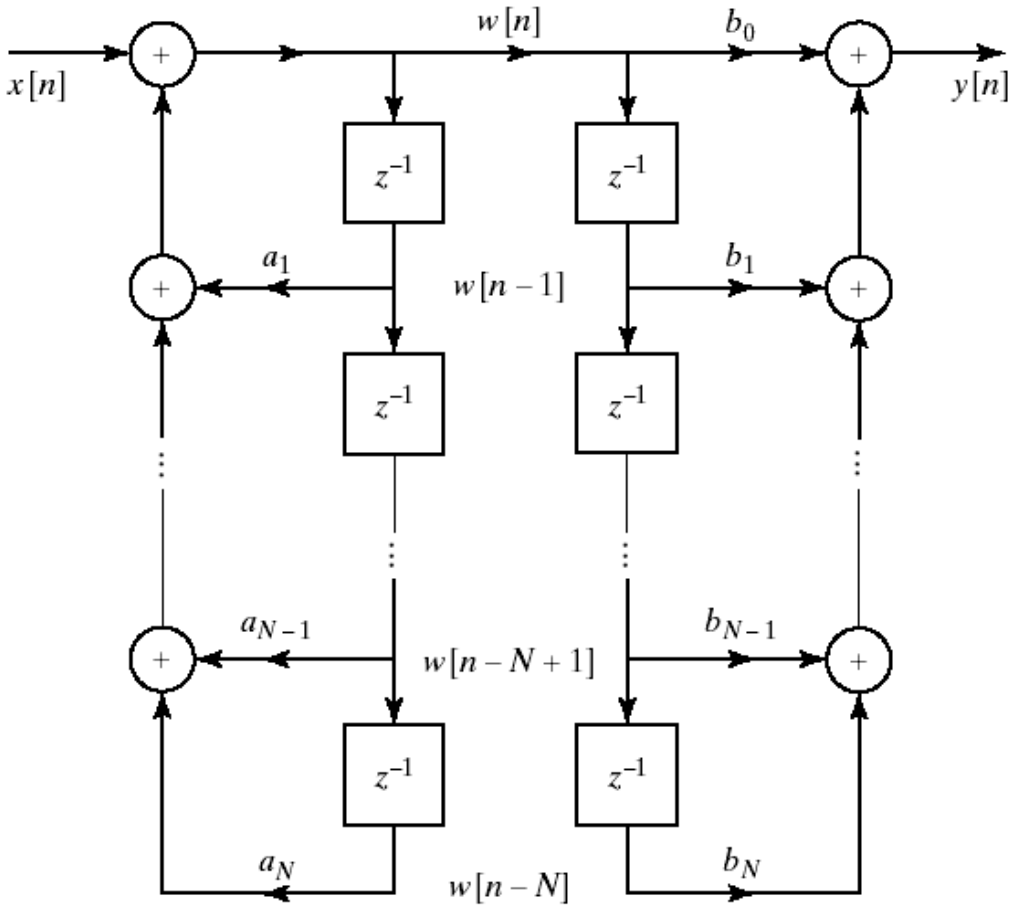
$$y[n] = -\sum_{k=1}^{N} a_k y[n-k] + v[n]$$

$$w[n] = -\sum_{k=1}^{N} a_k w[n-k] + x(n)$$

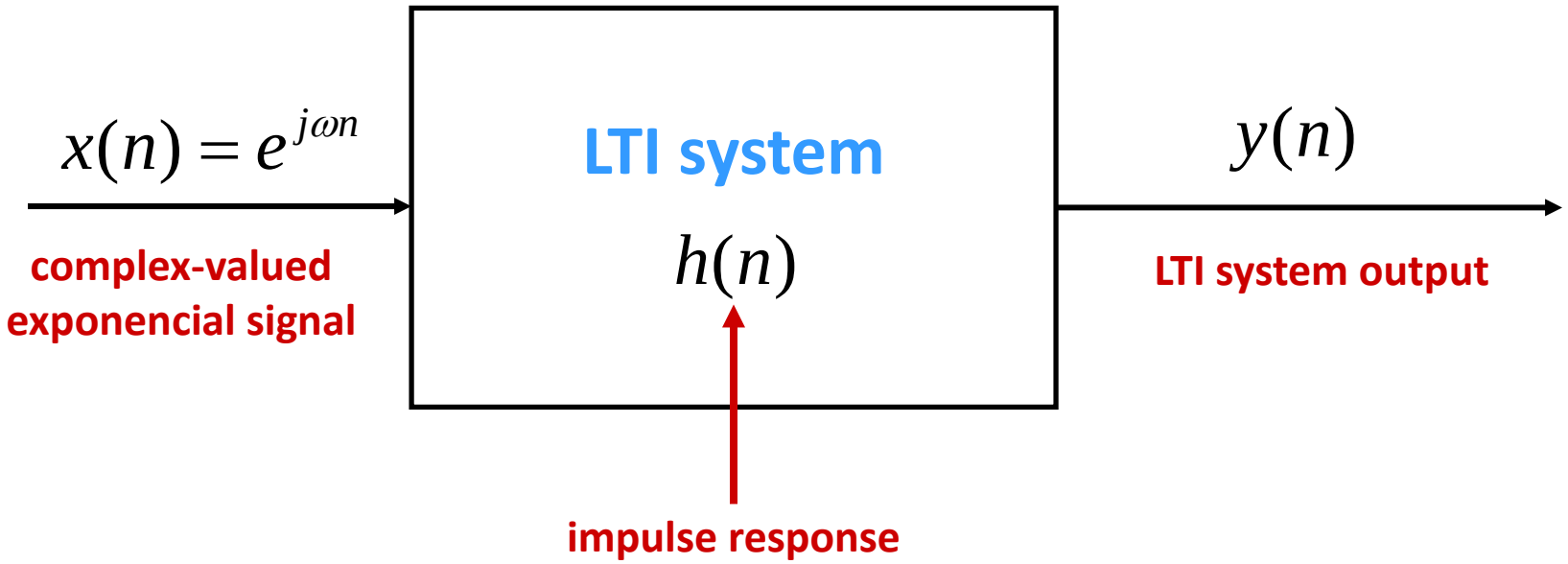
$$y[n] = \sum_{k=0}^{M} b_k w[n-k]$$

Direct Form I




x[n]
w[n]
b0
y[n]
z−1
z−1
a1
w[n − 1]
b1
z−1
z−1
aN−1
w[n − N + 1]
bN−1
z−1
z−1
aN
w[n − N]
bN

Frequency-Domain Representation of Discrete Signals and LTI Systems



$$y(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k)$$

LTI system output:

$$y(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) = \sum_{k=-\infty}^{\infty} h(k)e^{j\omega(n-k)} =$$

$$= \sum_{k=-\infty}^{\infty} h(k)e^{-j\omega k}e^{j\omega n} = e^{j\omega n}\sum_{k=-\infty}^{\infty} h(k)e^{-j\omega k}$$

$$y(n) = e^{j\omega n}H(e^{j\omega})$$

Frequency response: $H(e^{j\omega}) = \sum_{k=-\infty}^{\infty} h(k)e^{-j\omega k}$

$$H(e^{j\omega}) = \left|H(e^{j\omega})\right| e^{j\phi(\omega)}$$

$$H(e^{j\omega}) = \mathrm{Re}\left[H(e^{j\omega})\right] + j\,\mathrm{Im}\left[H(e^{j\omega})\right]$$

$$H(e^{j\omega}) = \sum_{k=-\infty}^{\infty} h(k)\cos\omega k + j\left[-\sum_{k=-\infty}^{\infty} h(k)\sin\omega k\right]$$

$$\mathrm{Re}\left[H(e^{j\omega})\right] = \sum_{k=-\infty}^{\infty} h(k)\cos\omega k$$

$$\mathrm{Im}\left[H(e^{j\omega})\right] = -\sum_{k=-\infty}^{\infty} h(k)\sin\omega k$$

Magnitude response:

$$\left|H(e^{j\omega})\right| = \sqrt{\mathrm{Re}\left[H(e^{j\omega})\right]^2 + \mathrm{Im}\left[H(e^{j\omega})\right]^2}$$

Phase response:

$$\phi(\omega) = \arg\left[H(e^{j\omega})\right] = arctg\,\frac{\mathrm{Im}\left[H(e^{j\omega})\right]}{\mathrm{Re}\left[H(e^{j\omega})\right]}$$

Group delay function:

$$\tau(\omega) = -\frac{d\phi(\omega)}{d\omega}$$

Comments on symmetry properties

For LTI systems with real-valued impulse response, the magnitude response, phase responses, the real component of and the imaginary component of $H(e^{j\omega})$ possess these symmetry properties:

The real component: ***even function*** of ω periodic with period 2π

$$\mathrm{Re}\left[H(e^{-j\omega})\right] = \mathrm{Re}\left[H(e^{j\omega})\right]$$

The imaginary component: ***odd function*** of ω periodic with period 2π

$$\mathrm{Im}\left[H(e^{-j\omega})\right] = -\mathrm{Im}\left[H(e^{j\omega})\right]$$

The magnitude response: even function of ω periodic with period 2π

$$\left|H(e^{j\omega})\right| = \left|H(e^{-j\omega})\right|$$

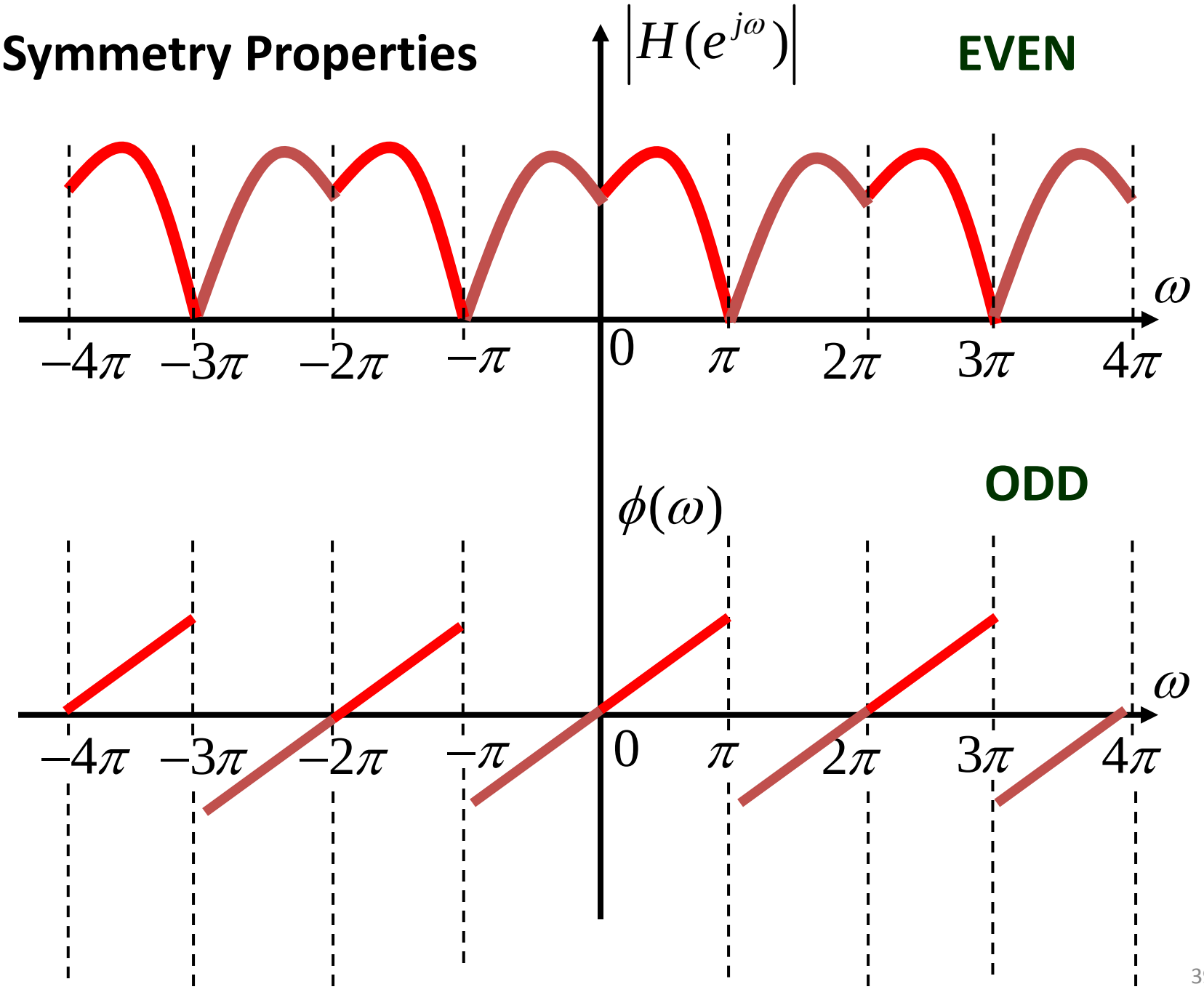
The phase response: odd function of ω periodic with period 2π

$$\arg\left[H(e^{-j\omega})\right] = -\arg\left[H(e^{j\omega})\right]$$

Consequence:

If we known $\left|H(e^{j\omega})\right|$ and $\phi(\omega)$ for $0 \le \omega \le \pi$, we can describe these functions (i.e. also $H(e^{j\omega})$) for all values of ω

Symmetry Properties



Normalized Frequency

It is often desirable to express the frequency response of an LTI system in terms of units of frequency that involve sampling interval *T*. In this case, the expressions:

$$H(e^{j\omega}) = \sum_{k=-\infty}^{\infty} h(k)e^{-j\omega k} \qquad h(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(e^{j\omega})e^{j\omega n}d\omega$$

are modified to the form:

$$H(e^{j\omega T}) = \sum_{k=-\infty}^{\infty} h(kT)e^{-j\omega kT}$$

$$h(nT) = \frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} H(e^{j\omega T})e^{j\omega nT}d\omega$$

$H(e^{j\omega T})$ is periodic with period $2\pi/T = 2\pi F$, where F is sampling frequency. $F/2 \rightarrow \pi$

Solution: **normalized frequency approach**:

Example:

$$F = 100kHz \qquad F/2 = 50kHz \qquad 50kHz \rightarrow \pi$$

$$f_1 = 20kHz \qquad \omega_1 = \frac{20x10^3}{50x10^3}\pi = \frac{2\pi}{5} = 0.4\pi$$

$$f_2 = 25kHz \qquad \omega_2 = \frac{25x10^3}{50x10^3}\pi = \frac{\pi}{2} = 0.5\pi$$

Discrete Time Fourier Transform (DTFT)

Discrete Time Fourier Transform

- Continuous time Fourier transform, when the signal is sampled:

$$x_s(t) \leftrightarrow \sum_{n=-\infty}^{\infty} x(nT)e^{-jn\omega T}$$

- Assuming $x(nT) = x[n]$ $\Omega = \omega T$
- Discrete-Time Fourier Transform (DTFT):

$$X(\Omega) = X(e^{j\omega}) = X(z)\,|_{z=e^{j\omega}} = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n}$$

$$x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}d\omega$$

❑ **DTFT is periodic in frequency with period of 2π**

$$e^{j\Omega} = e^{j(\Omega+2\pi)} = e^{j\Omega}e^{j2\pi} = e^{j\Omega}$$

Example

Impulse $x[n] = \delta[n], \quad X(e^{j\omega}) = 1$

Delayed impulse $x[n] = \delta[n - n_0], \quad X(e^{j\omega}) = e^{-j\omega n_0}$

Step function $x[n] = u[n], \quad X(e^{j\omega}) = \dfrac{1}{1 - e^{-j\omega}}$

Rectangular window $x[n] = u[n] - u[n - N], \; X(e^{j\omega}) = \dfrac{1 - e^{-j\omega N}}{1 - e^{-j\omega}}$

Exponential $x[n] = a^n \; u[n], \quad X(e^{j\omega}) = \dfrac{1}{1 - ae^{-j\omega}}, \; a < 1$

Table 6-1. Properties of the Fourier Transform

Property	Sequence	Fourier transform
	$x[n]$	$X(\Omega)$
	$x_1[n]$	$X_1(\Omega)$
	$x_2[n]$	$X_2(\Omega)$
Periodicity	$x[n]$	$X(\Omega + 2\pi) = X(\Omega)$
Linearity	$a_1 x_1[n] + a_2 x_2[n]$	$a_1 X_1(\Omega) + a_2 X_2(\Omega)$
Time shifting	$x[n - n_0]$	$e^{-j\Omega n_0} X(\Omega)$
Frequency shifting	$e^{j\Omega_0 n} x[n]$	$X(\Omega - \Omega_0)$
Conjugation	$x^*[n]$	$X^*(-\Omega)$
Time reversal	$x[-n]$	$X(-\Omega)$
Time scaling	$x_{(m)}[n] = \begin{cases} x[n/m] & \text{if } n = km \\ 0 & \text{if } n \neq km \end{cases}$	$X(m\Omega)$
Frequency differentiation	$nx[n]$	$j\dfrac{dX(\Omega)}{d\Omega}$
First difference	$x[n] - x[n-1]$	$(1 - e^{-j\Omega})X(\Omega)$
Accumulation	$\sum_{k=-\infty}^{n} x[k]$	$\pi X(0)\delta(\Omega) + \dfrac{1}{1 - e^{-j\Omega}} X(\Omega)$
		$\lvert\Omega\rvert \leq \pi$
Convolution	$x_1[n] * x_2[n]$	$X_1(\Omega)X_2(\Omega)$
Multiplication	$x_1[n]x_2[n]$	$\dfrac{1}{2\pi} X_1(\Omega) \otimes X_2(\Omega)$
Real sequence	$x[n] = x_e[n] + x_o[n]$	$X(\Omega) = A(\Omega) + jB(\Omega)$
		$X(-\Omega) = X^*(\Omega)$
Even component	$x_e[n]$	$\mathrm{Re}\{X(\Omega)\} = A(\Omega)$
Odd component	$x_o[n]$	$j\ \mathrm{Im}\{X(\Omega)\} = jB(\Omega)$

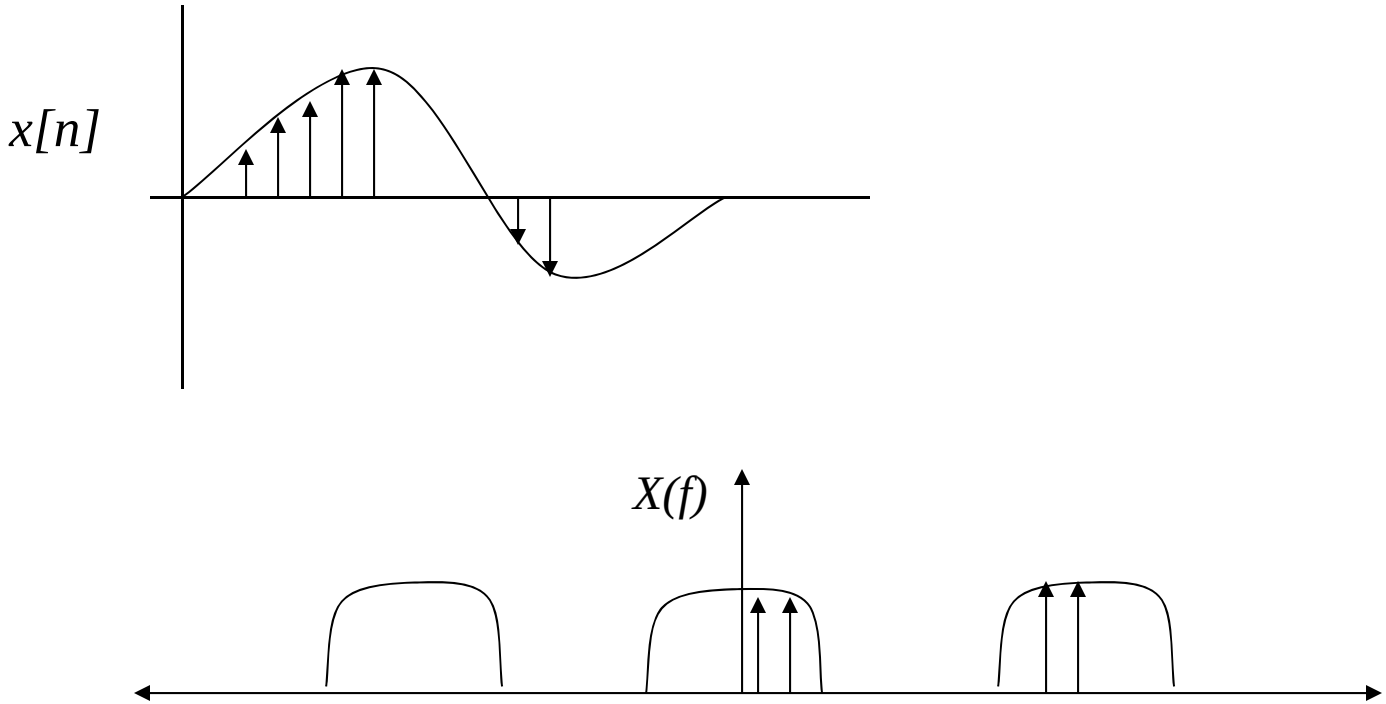
Parseval's relations

$$\sum_{n=-\infty}^{\infty} x_1[n]x_2[n] = \frac{1}{2\pi}\int_{2\pi} X_1(\Omega)X_2(-\Omega)\,d\Omega$$

$$\sum_{n=-\infty}^{\infty} |x[n]|^2 = \frac{1}{2\pi}\int_{2\pi} |X(\Omega)|^2\,d\Omega$$

DISCRETE FOURIER TRANSFORM (DFT)

x[n] is the discrete signal FT is $X_s(f)$



Frequency domain sampling of the FT→DFT

DFT $$X[k] = \sum_{n=0}^{N-1} x[n] W_N^{kn}$$ k=0,1,2,3…….N-1

IDFT $$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] W_N^{-kn}$$ n=0,1,2,3…….N-1

Where W_N *is define as* $W_N = e^{-j2\pi/N}$

DFT is the set of N sample {X[k]} of the Fourier transform X(ω) for a finite –duration sequence {x[n]} of length L<=N. the sampling of X(ω) occurs at the N equally spaced frequencies $\omega_k = 2\pi k/N$, k=0,1,2,3… N-1

$$e^{-j\theta} = \cos\theta - j\sin\theta$$

$$X[k] = \sum_{n=0}^{N-1} x[n] W_N^{kn} = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N}$$

$$= \sum_{n=0}^{N-1} x[n][\cos(2\pi kn/N) - j\sin(2\pi kn/N)$$

$$X[k] = X_{real}[k] + j\, X_{imag}[k]$$

$$|x[k]| = X_{power}[k] = \sqrt{X_{real}^2[k] + X_{img}^2[k]}$$

$$X_\theta[k] = \tan^{-1}\left[\frac{X_{img}[k]}{X_{real}[k]}\right]$$

$$f(k) = kf_s/N$$

Properties of DET

Periodicity: *if x[n] and X[k] are an N-point DFT then*

$$x[n+N]=x[n] \quad \text{for all } n$$

$$X[k+N]=X[k] \quad \text{for all } k$$

Linearity: if

$$x_1(n) \xleftrightarrow[N]{DFT} X_1(n) \qquad x_2(n) \xleftrightarrow[N]{DFT} X_2(n)$$

$$a_1x_1(n)+a_2x_2(n) \xleftrightarrow[N]{DFT} a_1X_1(n)+a_2X_2(n)$$

Symmetry:

DFT Shifting theorem:

$$X_{shifted}(l) = e^{-j2\pi ln/N} X(l)$$

x[n] is shifted to the right by l sample

DFT Leakage:

DFT Resolution, Zero stuffing:

DFT Magnitudes:

When a real input signal contains a sine wave component of peak amplitude A_0 with an integral number of cycles over N input samples the output magnitude of the DFT for that particular sine wave is M_r where

$$M_r = A_0 * N/2$$

If the DFT input is a complex sinusoid of magnitude A_0 with an integral number of cycles over N input samples the output magnitude of the DFT is M_c where

$$M_c = A_0 N$$

Fast Fourier Transform

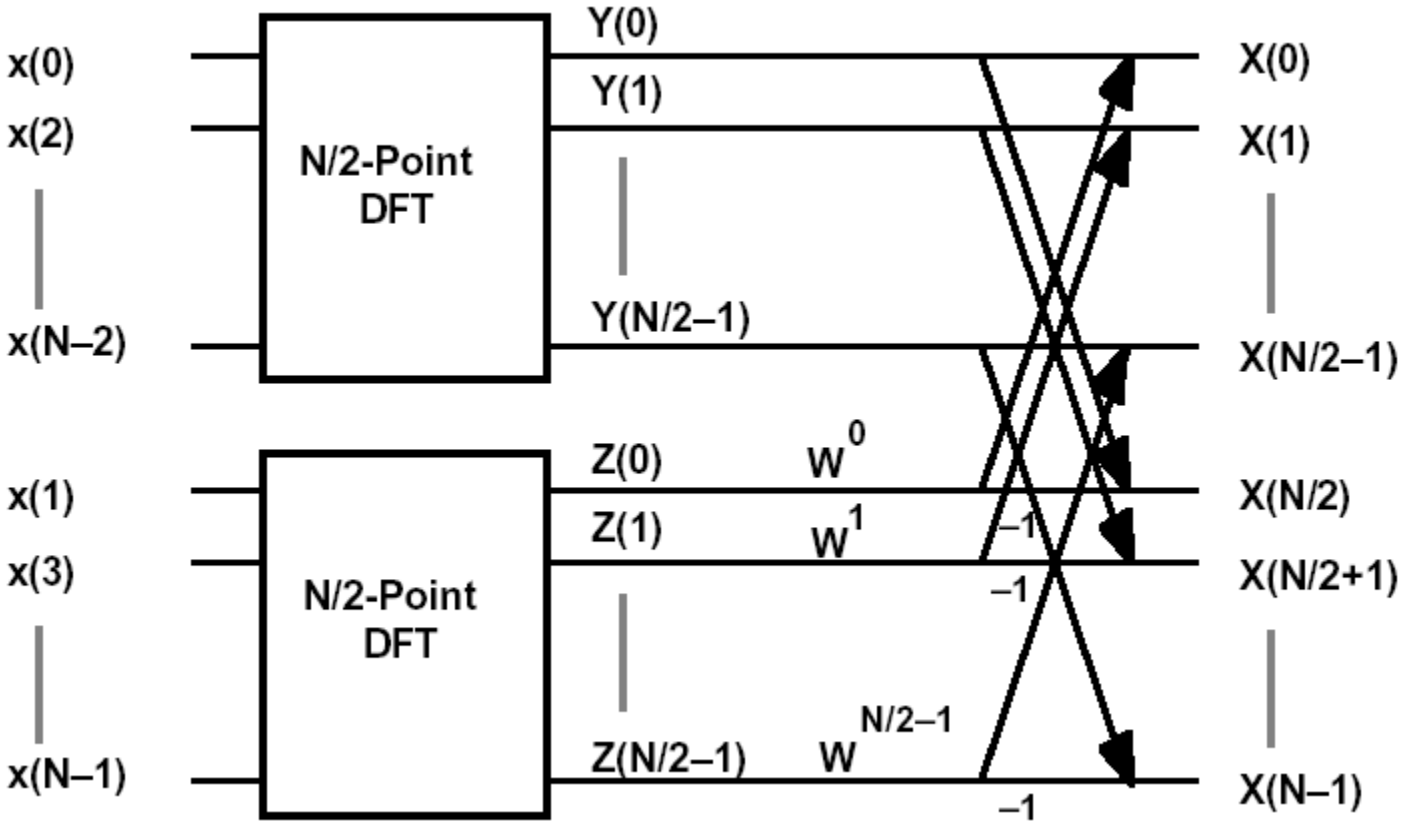
let the computation of $N=2^v$ point DFT , split the N point data sequence into two N/2 point data sequence f1(n),f2(n) corresponding the even-number and odd-numberd samples of x(n)

f1(n)=x(2n), f2(n)=x(2n+1)

Thus f1(n) and f2(n) are obtained by decimating x(n) by a factor of 2 and hence the resulting FFT algorithm is called *Decimating in time algorithm:*

$$X(k)=\sum_{n=0}^{N-1}x(n)W_N^{kn}$$

$$=\sum_{n-even}x(n)W_N^{kn}+\sum_{n-odd}x(n)W_N^{kn}$$


x(0)
x(2)
x(N–2)
N/2-Point DFT
Y(0)
Y(1)
Y(N/2–1)
x(1)
x(3)
x(N–1)
N/2-Point DFT
Z(0)
Z(1)
Z(N/2–1)
W^0
W^1
W^N/2–1
–1
–1
–1
X(0)
X(1)
X(N/2–1)
X(N/2)
X(N/2+1)
X(N–1)

Decimation in Frequency

$$X(k) = \sum_{n=0}^{N/2-1} x(n) W_N^{kn} + \sum_{n=N/2}^{N-1} x(n) W_N^{kn}$$

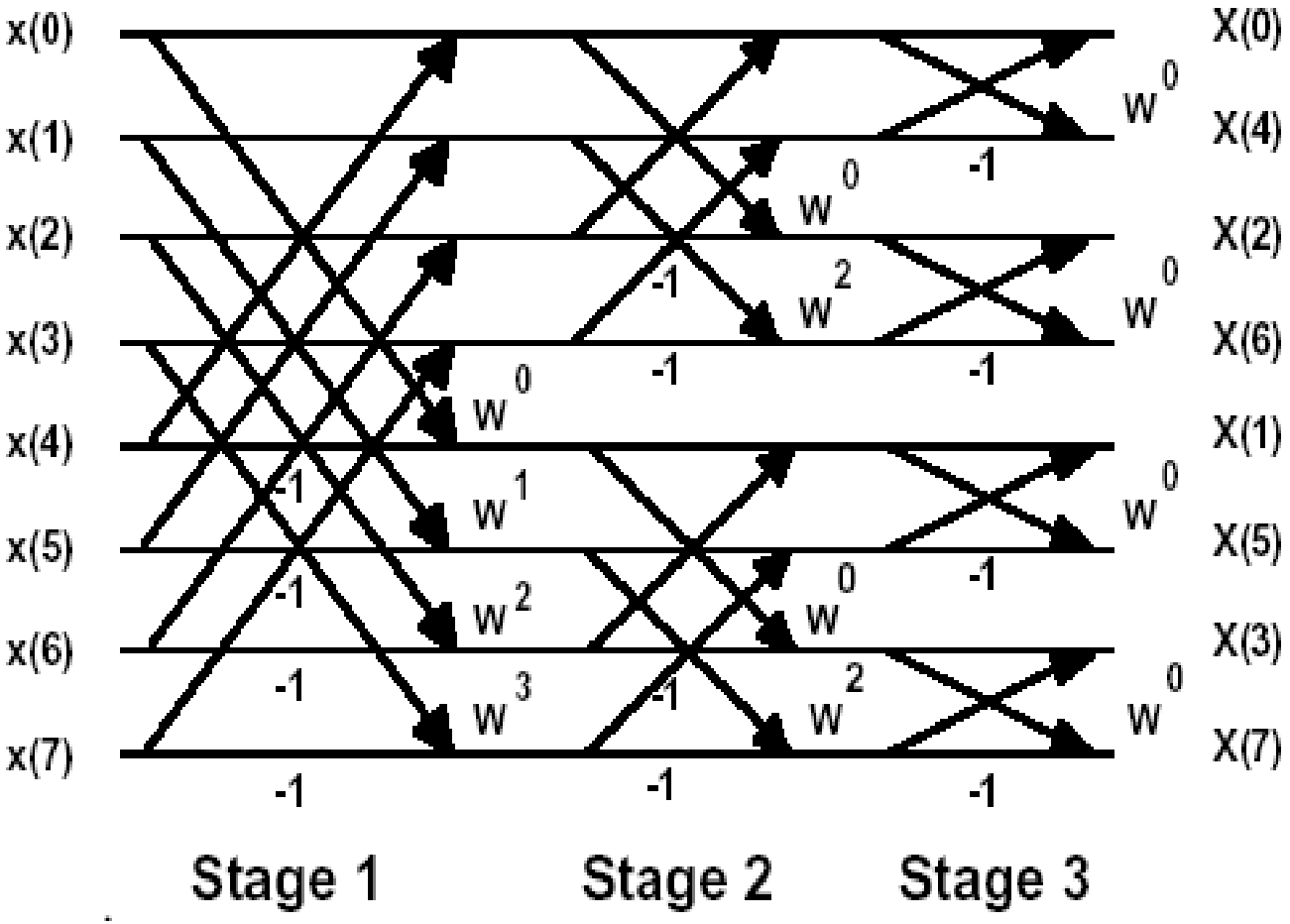
$$= \sum_{n=0}^{N/2-1} x(n) W_N^{kn} + W_N^{kN/2} \sum_{n=0}^{N/2-1} x(n+\frac{N}{2}) W_N^{kn}$$

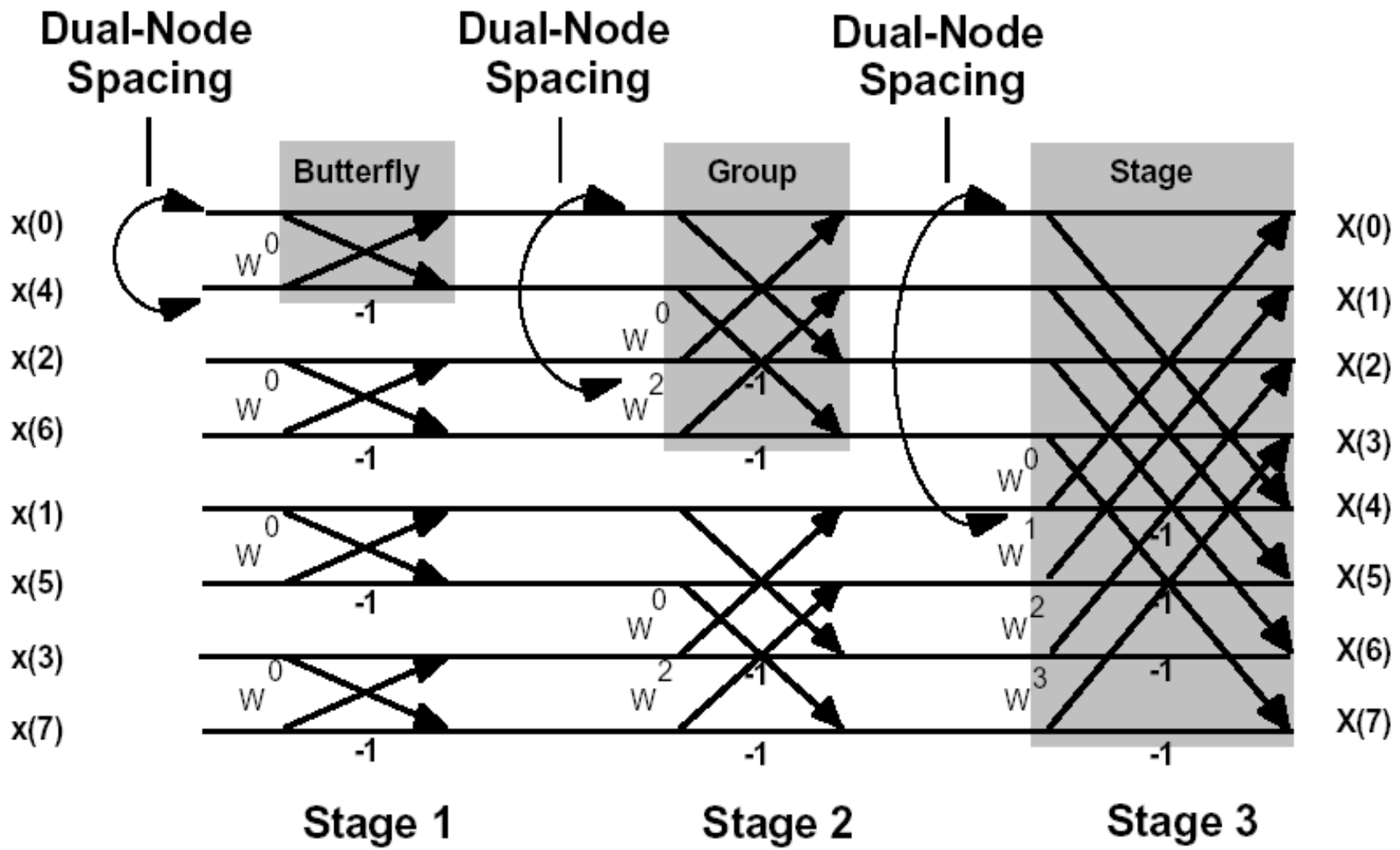
Now $W_N^{kN/2} = (-1)^k$

$$X(k) = \sum_{n=0}^{N/2-1} \left[x(n) + (-1)^k x(n+N/2) \right] W_N^{kN/2}$$

$$X(2k) = \sum_{n=0}^{N/2-1} \left[x(n) + x(n+N/2) \right] W_N^{kN/2}$$

$$X(2k+1) = \sum_{n=0}^{N/2-1} \left[x(n) - x(n+N/2) \right] W_N^{kN/2}$$

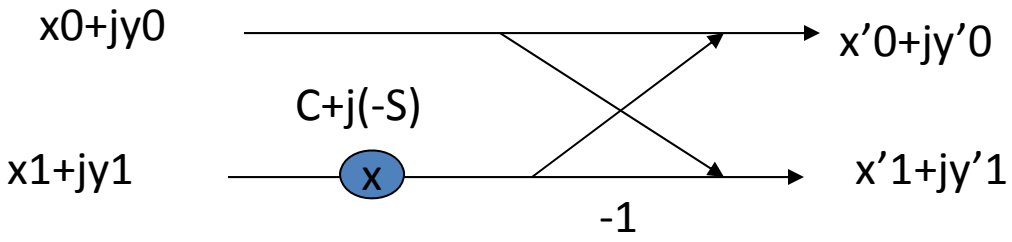

x(0)
x(1)
x(2)
x(3)
x(4)
x(5)
x(6)
x(7)
X(0)
X(4)
X(2)
X(6)
X(1)
X(5)
X(3)
X(7)
W 0
W 0
W 0
W 0
W 1
W 2
W 3
W 0
W 2
W 0
-1
Stage 1
Stage 2
Stage 3

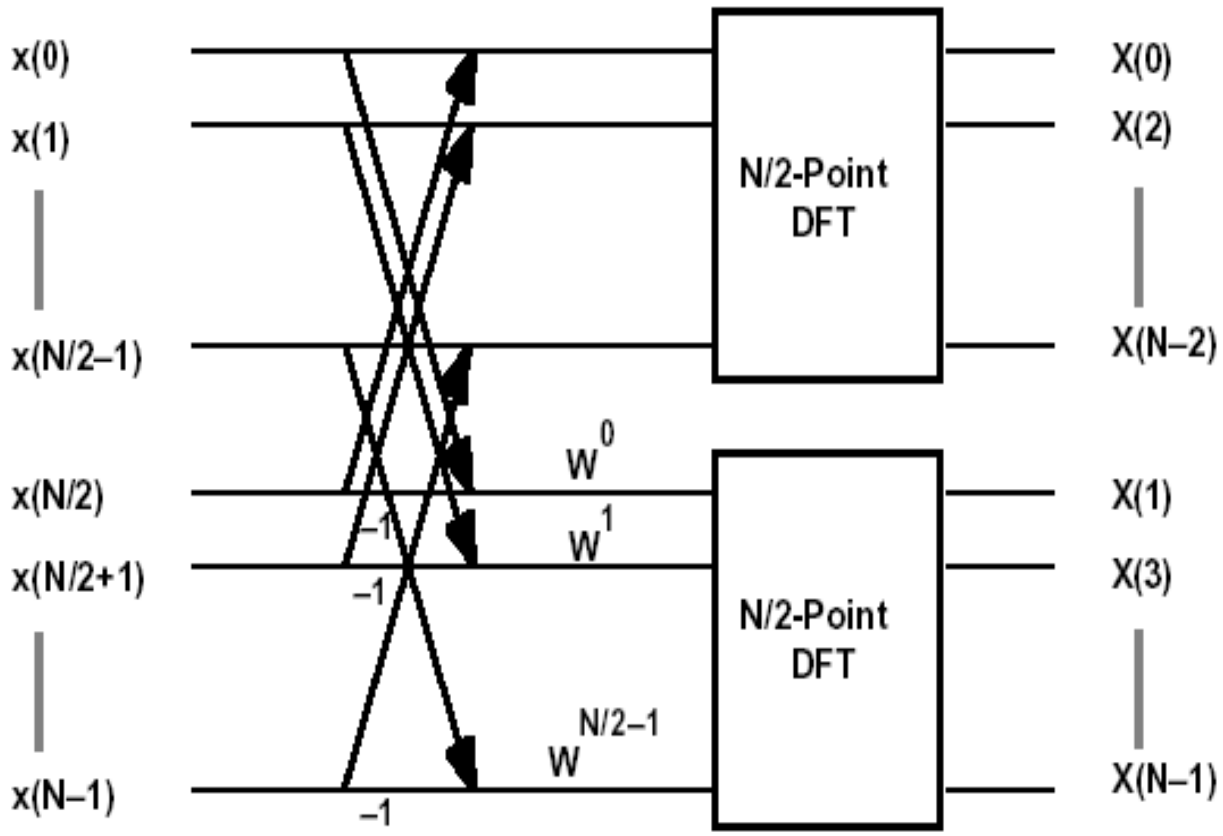

Dual-Node Spacing
Dual-Node Spacing
Dual-Node Spacing
Butterfly
Group
Stage
x(0)
x(4)
x(2)
x(6)
x(1)
x(5)
x(3)
x(7)
X(0)
X(1)
X(2)
X(3)
X(4)
X(5)
X(6)
X(7)
W 0
W 0
W 0
W 0
W 0
W 2
W 0
W 2
W 1
W 2
W 3
-1
Stage 1
Stage 2
Stage 3

The characteristics of an N-point DIT FFT with bit-reversed inputs are summarized below.

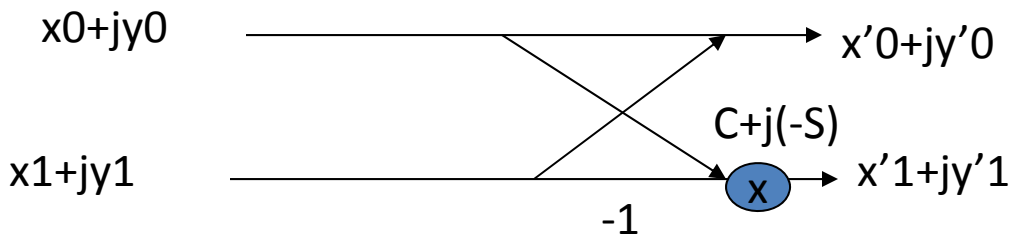
	Stage 1	*Stage 2*	*Stage 3*	*Stage Log_2N*
Number of Groups	N/2	N/4	N/8	1
Butterflies per Group	1	2	4	N/2
Dual-Node Spacing	1	2	4	N/2
Twiddle Factor Exponents	(N/2)k, k=0	(N/4)k, k=0, 1	(N/8)k, k=0, 1, 2, 3	k, k=0 to N/2–1

$$W_N = e^{-j2\pi/N} = \cos(2\pi/N) - j\sin(2\pi/N)$$

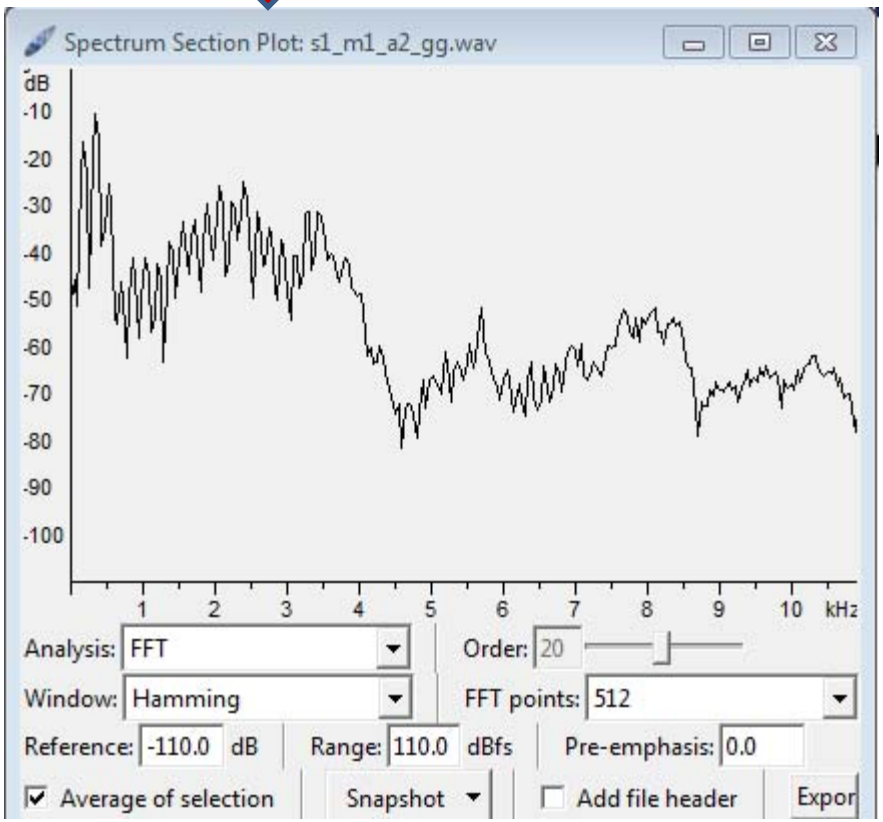



x(0)
x(1)
x(N/2–1)
x(N/2)
x(N/2+1)
x(N–1)
N/2-Point DFT
N/2-Point DFT
X(0)
X(2)
X(N–2)
X(1)
X(3)
X(N–1)
W 0
W 1
W N/2–1
–1
–1
–1

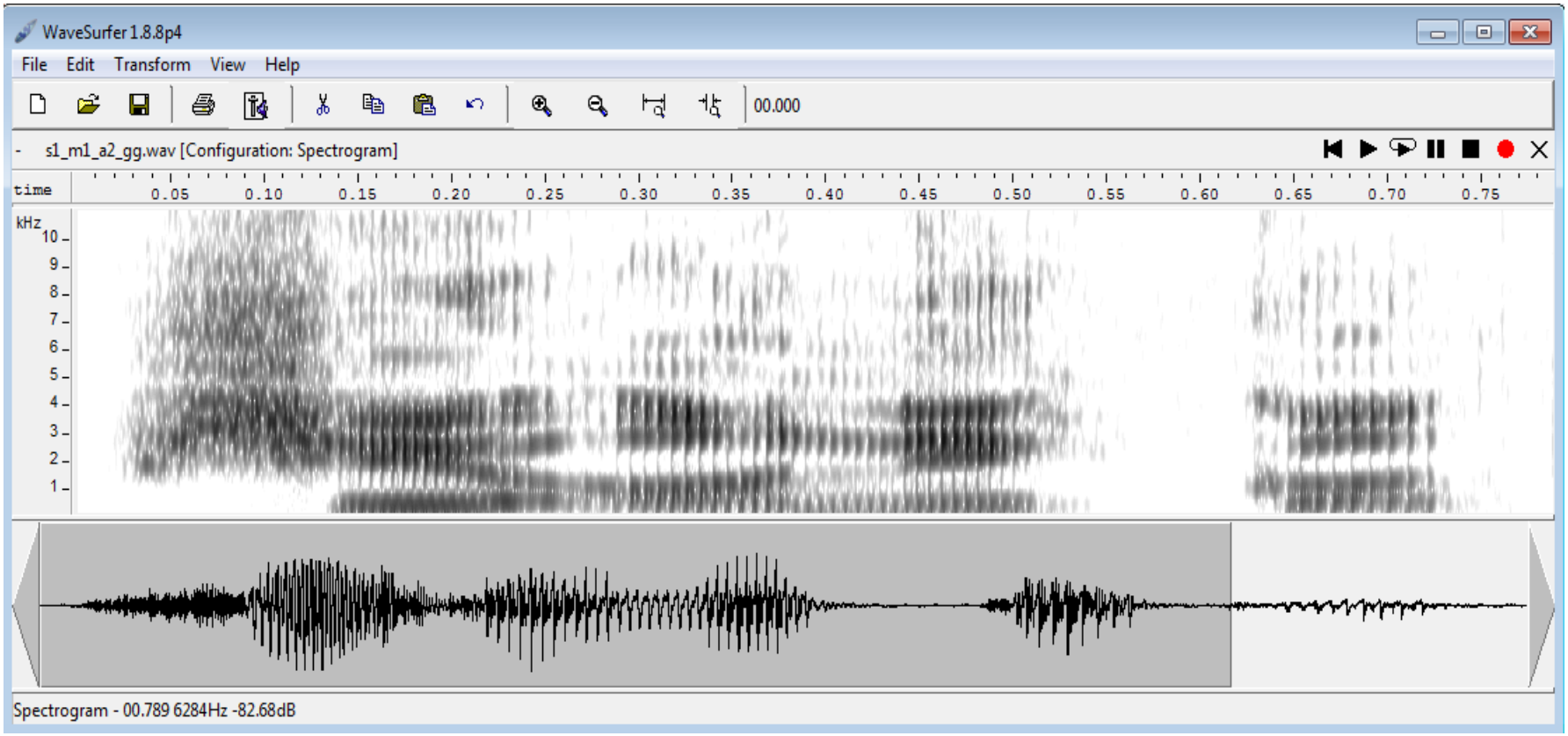
	Stage 1	*Stage 2*	*Stage 3*	*Stage Log_2N*
Number of Groups	1	2	4	N/2
Butterflies per Group	N/2	N/4	N/8	1
Dual-Node Spacing	N/2	N/4	N/8	1
Twiddle Factor Exponents	n, n=0 to N/2–1	2n, n=0 to N/4–1	4n, n=0 to N/8–1	(N/2)n, n=0



Frequency spectrum



Spectrogram



Digital Filter

Digital Filter Implementation

- input and output satisfy linear difference equation of the form:

$$y[n]-\sum_{k=1}^{N} a_k y[n-k]=\sum_{r=0}^{M} b_r x[n-r]$$

- evaluating z-transforms of both sides gives:

$$Y(z)-\sum_{k=1}^{N} a_k z^{-k} Y(z)=\sum_{r=0}^{M} b_r z^{-r} X(z)$$

$$Y(z)\left(1-\sum_{k=1}^{N} a_k z^{-k}\right)=X(z)\sum_{r=0}^{M} b_r z^{-r}$$

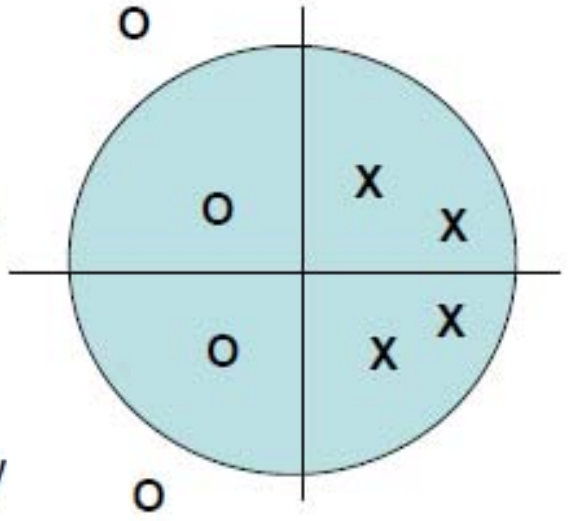
$$H(z)=\frac{Y(z)}{X(z)}=\frac{\sum_{r=0}^{M} b_r z^{-r}}{1-\sum_{k=1}^{N} a_k z^{-k}}$$

canonic form showing poles and zeros

Digital Filters

- $H(z)$ is a rational function in z^{-1}

$$H(z) = \frac{A\prod_{r=1}^{M}(1-c_r z^{-1})}{\prod_{k=1}^{N}(1-d_k z^{-1})} \Rightarrow M \text{ zeros, } N \text{ poles}$$

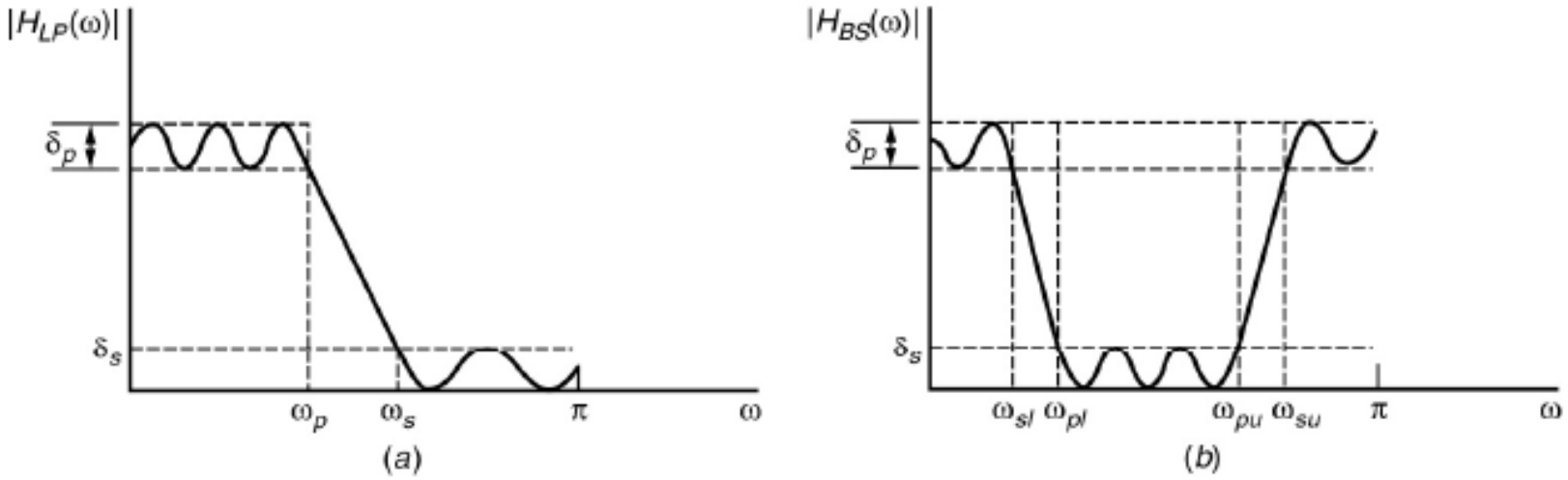


- converges for $|z| > R_1$, with $R_1 < 1$ for stability

$\Rightarrow$

all poles of $H(z)$ inside the unit circle for a stable, causal system

Equiripple Design Specifications



ω_p = normalized edge of passband frequency

ω_s = normalized edge of stopband frequency

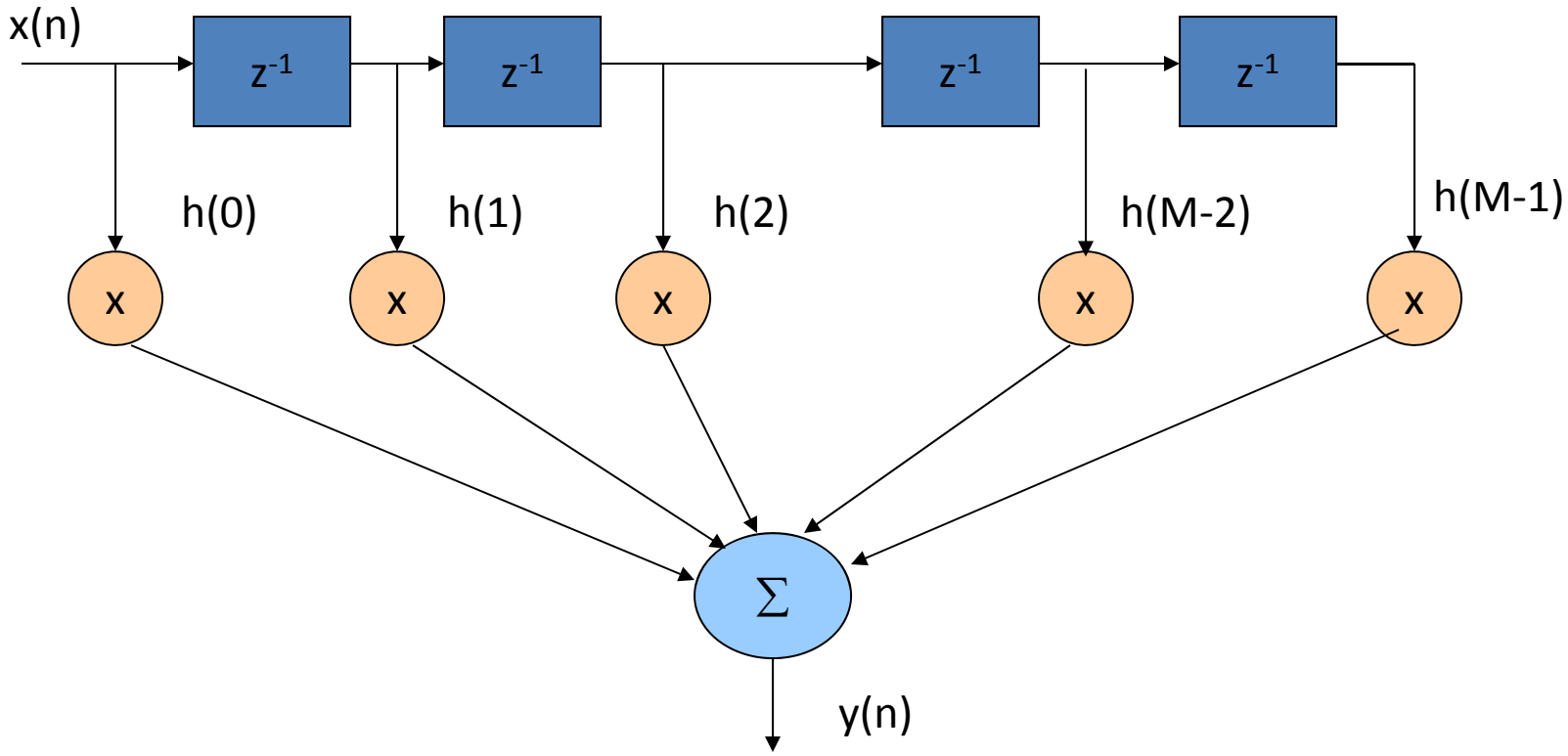
δ_p = peak ripple in passband

δ_s = peak ripple in stopband

$\Delta\omega = \omega_s - \omega_p$ = normalized transition bandwidth

A finite impulse response(FIR) filter is a discrete linear time-invariant system whose output is based on the weighted summation of a finite number of past input.

$$y(n) = \sum_{k=0}^{M-1} b_k \, x(n-k)$$



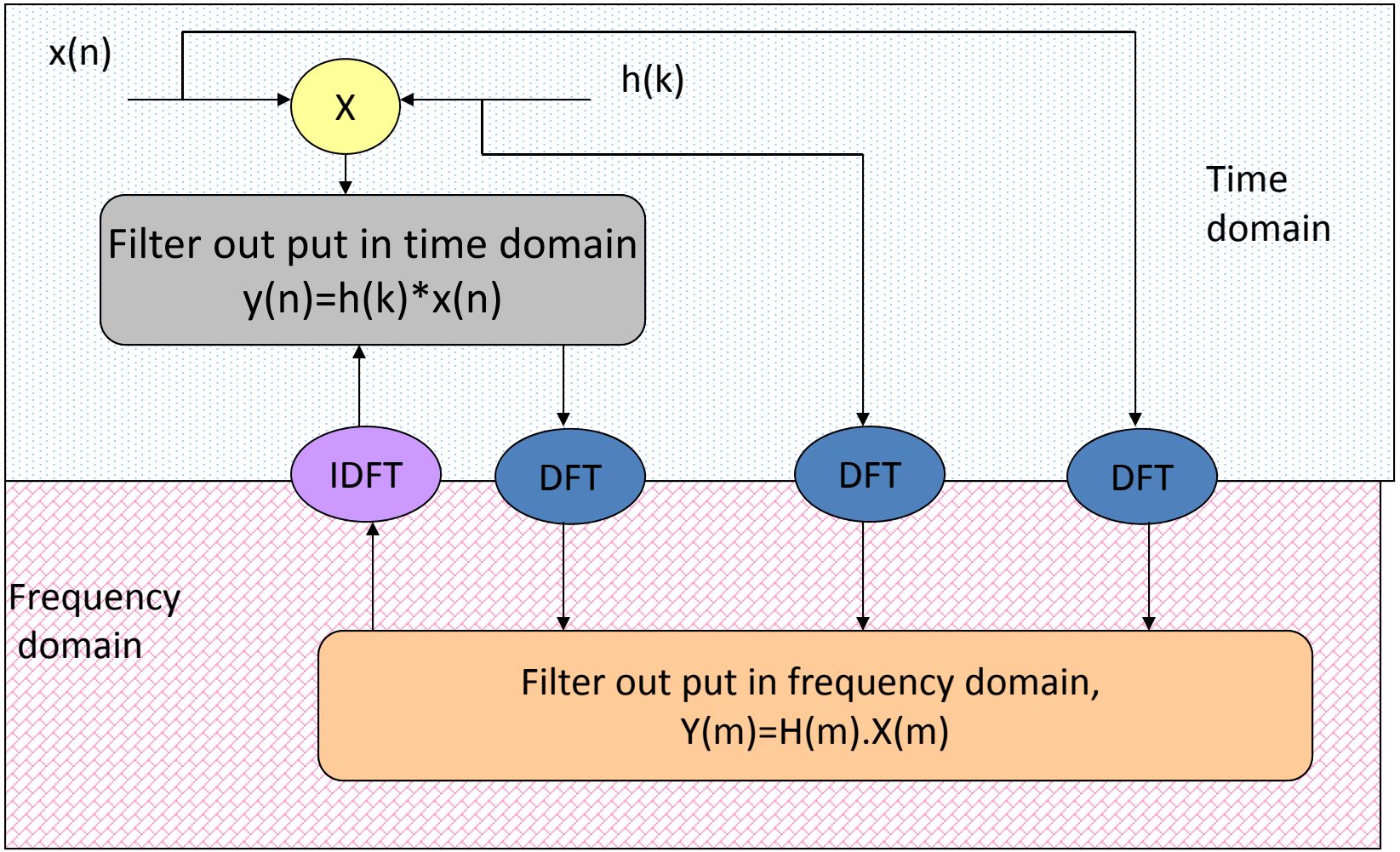
$$y(n) = h(0)x(n) + h(1)x(n-1) + \ldots\ldots + h(M-2)x(n-M+2) + h(M-1)x(n-M+1)$$

$$\Rightarrow y(n) = \sum_{k=0}^{M-1} h(k)x(n-k)$$

DFT

$$y(n) = h(k) * x(n) \Leftrightarrow H(m).X(m)$$

IDFT



Algebraic determination of time-domain coefficients of low pass filter

1. Develop an expression for the discrete frequency response H(m)
2. Apply that expression to the inverse DFT equation to get the time domain h(k)
3. Evaluate that h(k) expression as a function of time

Let $H_d(w)$ is the frequency response of a low-pass filter of time response $h_d(n)$

$$H_d(\omega) = \sum_{n=0}^{\infty} h_d(n) e^{-j\omega n}$$

$$h_d(n) = \int_{-\pi}^{\pi} H_d(\omega) e^{-j\omega n} d\omega$$

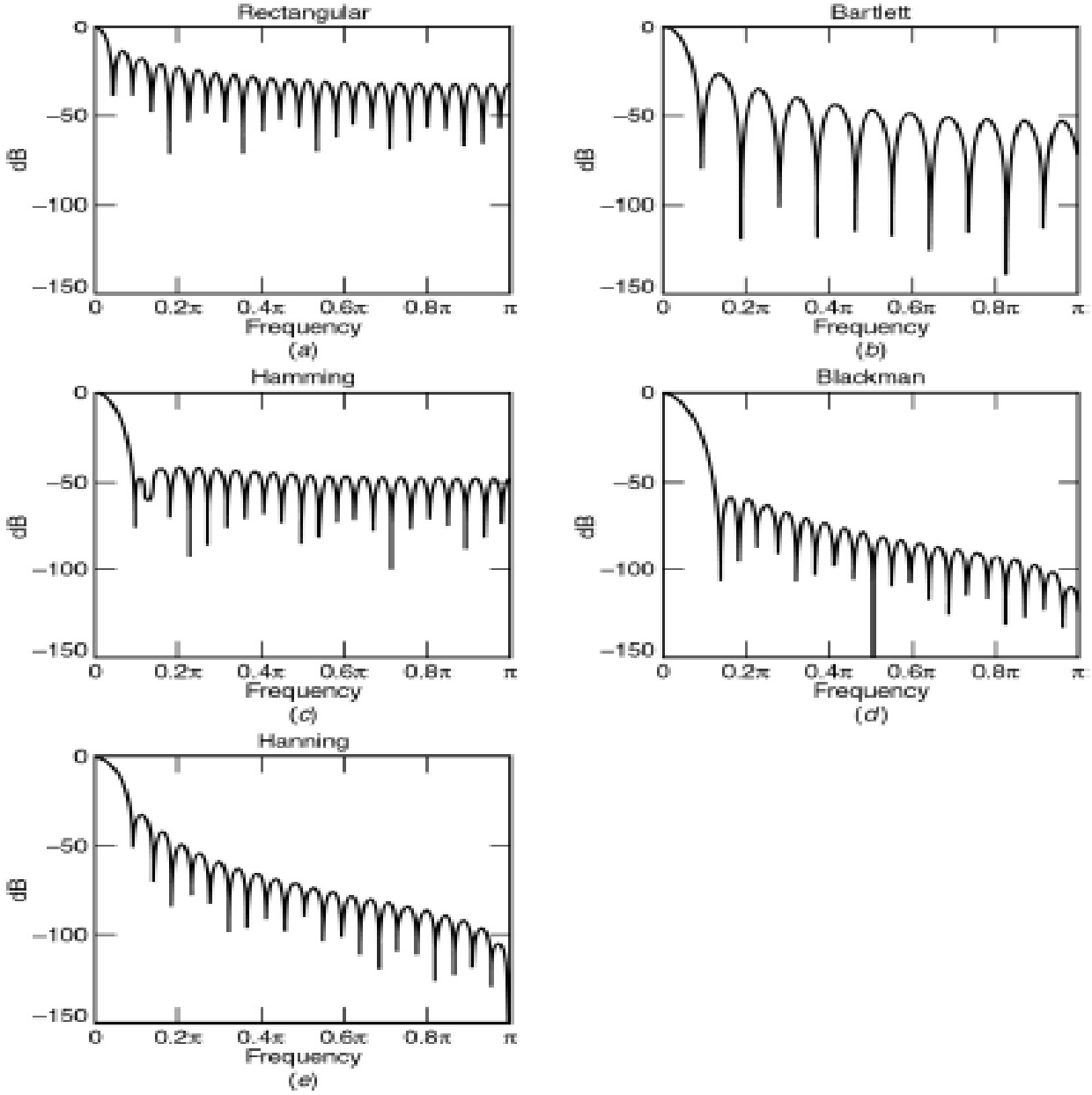
The unit sample response obtained from the above equation is infinite in duration and must be truncated at some point say n=M-1 for a FIR filter of length M this truncation is equivalent to multiplying $h_d(n)$ by a window function w(n).

$$h(n) = h_d(n) w(n)$$

Window Function for FIR Filter Design

Name of Window	Window function
Bartlett(triangular)	$1 - \frac{2\left\lvert n - \frac{M-1}{2}\right\rvert}{M-1}$
Blackman	$0.42 - 0.5\cos\frac{2\pi n}{M-1} + 0.08\cos\frac{4\pi n}{M-1}$
Hamming	$0.54 - 0.46\cos\frac{2\pi n}{M-1}$
Hanning	$\frac{1}{2}\left(1 - \cos\frac{2\pi n}{M-1}\right)$
Kaiser	$\frac{I_0\left[\alpha\sqrt{(M-1)^2 - \left(n - \frac{M-1}{2}\right)^2}\right]}{I_0\left[\alpha\left(\frac{M-1}{2}\right)\right]}$

Common Windows (Frequency)



Example

Let the frequency response of a low-pass filter as

$$H_d(\omega) = \begin{cases} 1e^{-j\omega(M-1)/2} & 0 \le |\omega| \le \omega_c \\ 0 & \text{otherwise} \end{cases}$$

A delay of (M-1)/2 unit is incorporated into H(w) in anticipation of forcing the filter to be of length M

$$h_d(n) = \frac{1}{2\pi}\int_{-\omega_c}^{\omega_c} e^{jw(n-\frac{M-1}{2})} d\omega$$

$$h(n) = \frac{\sin \omega_c (n - \frac{M-1}{2})}{\pi\left(n - \frac{M-1}{2}\right)} \quad 0 \le n \le M-1, n \ne \frac{M-1}{2}$$

$$h\left(\frac{M-1}{2}\right) = \omega_c / \pi$$

Type of Window	Approximate Transition width of main Lobe	Peak Sidelobe
Rectangular	$4\pi/M$	-13
Bartlett	$8\pi/M$	-27
Hanning	$8\pi/M$	-32
Hamming	$8\pi/M$	-43
Blackman	$12\pi/M$	-58

Band-pass FIR filter design:

$$h_{bp}(k) = h_{lp}(k) \cdot s_{shift}(k)$$

High-pass FIR filter design:

$$h_{bp}(k) = h_{lp}(k) \cdot s_{shift}(k)$$

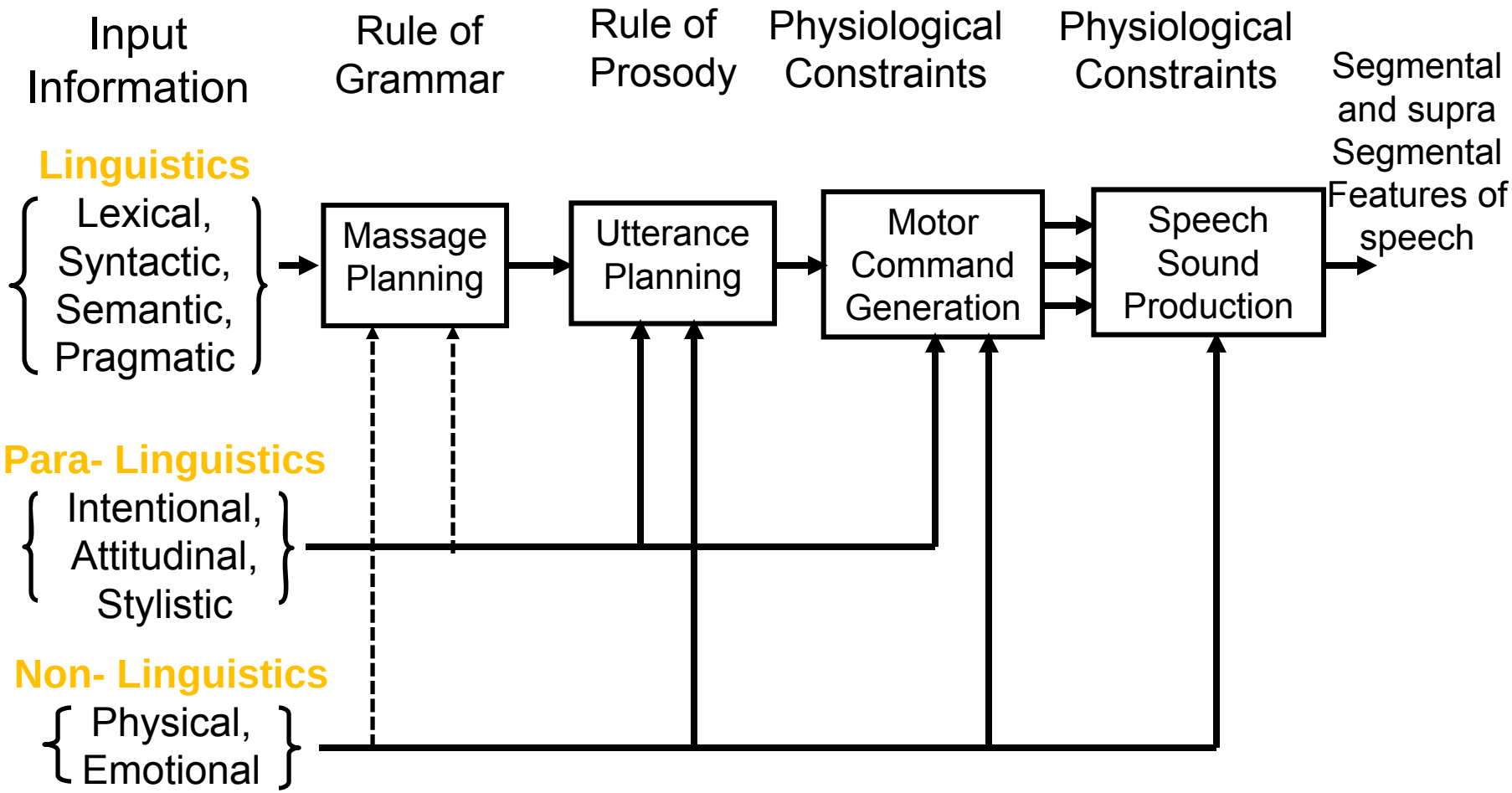
IIR Design Methods

Impulse invariant transformation – match the analog impulse response by sampling; resulting frequency response is aliased version of analog frequency response

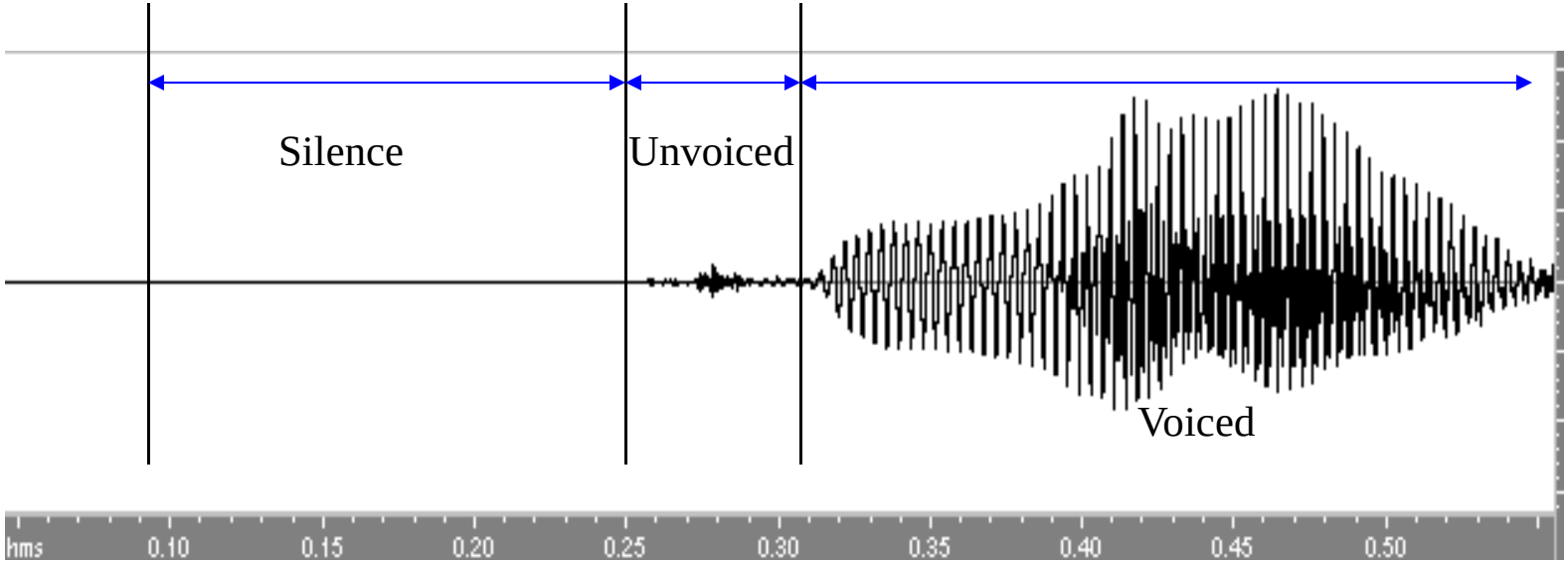
Bilinear transformation – use a transformation to map an analog filter to a digital filter by warping the analog frequency scale (0 to infinity) to the digital frequency scale (0 to π); use frequency pre-warping to preserve critical frequencies of transformation (i.e., filter cutoff frequencies)

Human Speech Production and Source Filter model

Information manifestation in the segmental and suprasegmental features of speech



@ Fujisaki

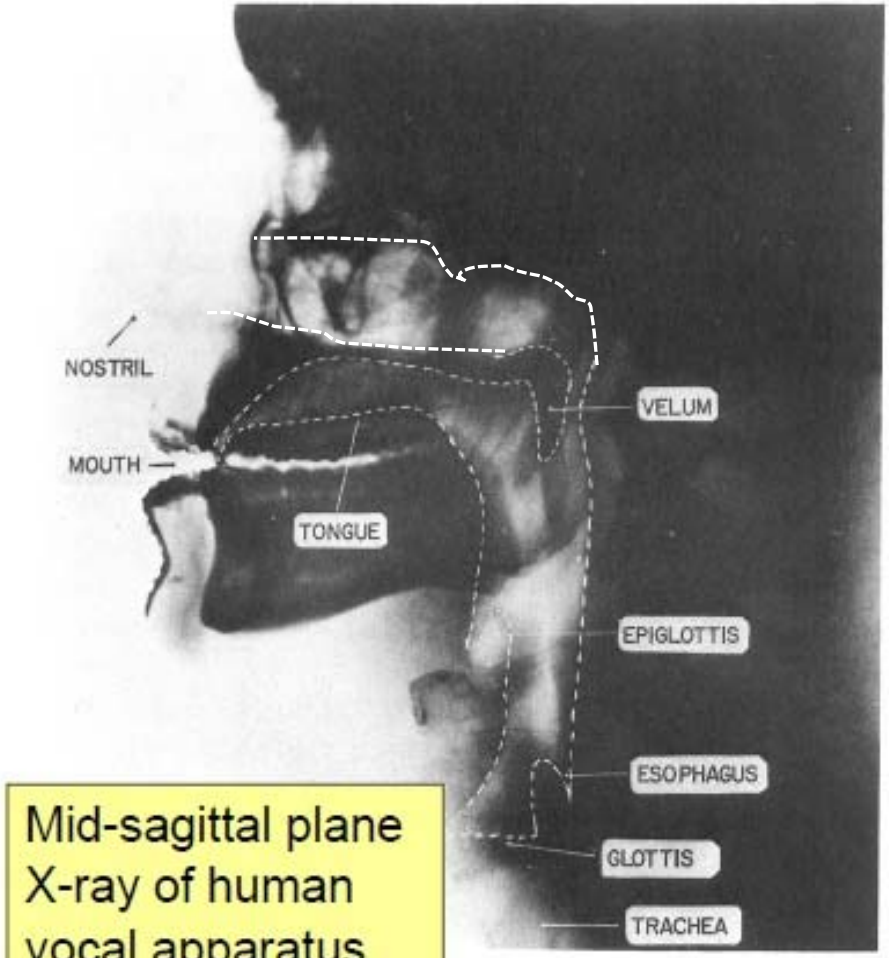

Silence
Unvoiced
Voiced
hms
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50

Basics Definitions

Speech is composed of a sequence of sounds
Sounds/Phonemes serve as a symbolic representation of information to be shared between humans (or humans and machines)

Arrangement of sounds is governed by rules of language (constraints on sound sequences, word sequences, etc)

Linguistics is the study of the rules of language
Phonetics is the study of the sounds of speech



Mid-sagittal plane X-ray of human vocal apparatus

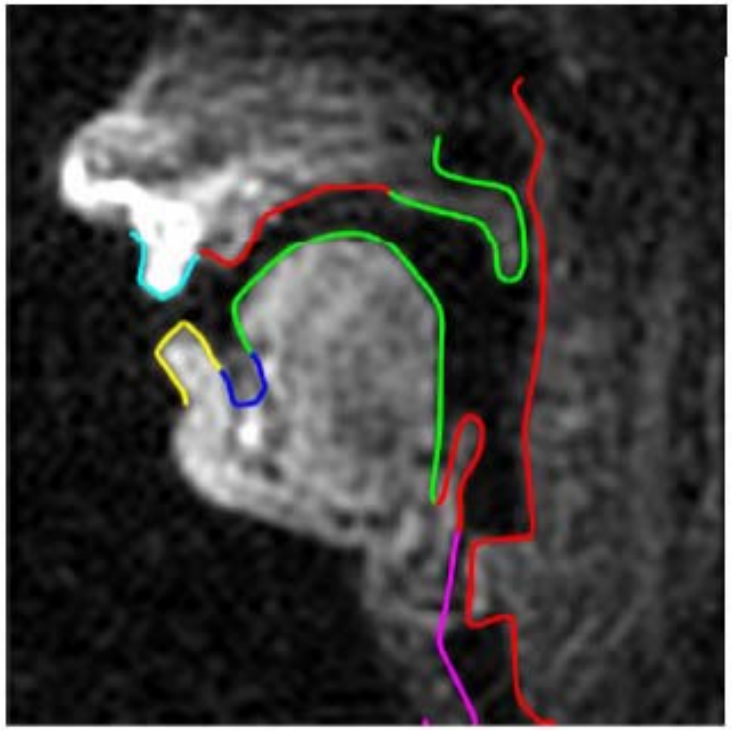
Vocal tract —dotted lines in figure; begins at the glottis (the vocal cords) and ends at the lips

- Consists of the pharynx (the connection from the esophagus to the mouth) and the mouth itself (the oral cavity)
- Average male vocal tract length is 17.5 cm
- Cross sectional area, determined by positions of the tongue, lips, jaw and velum, varies from zero (complete closure) to 20 sq cm

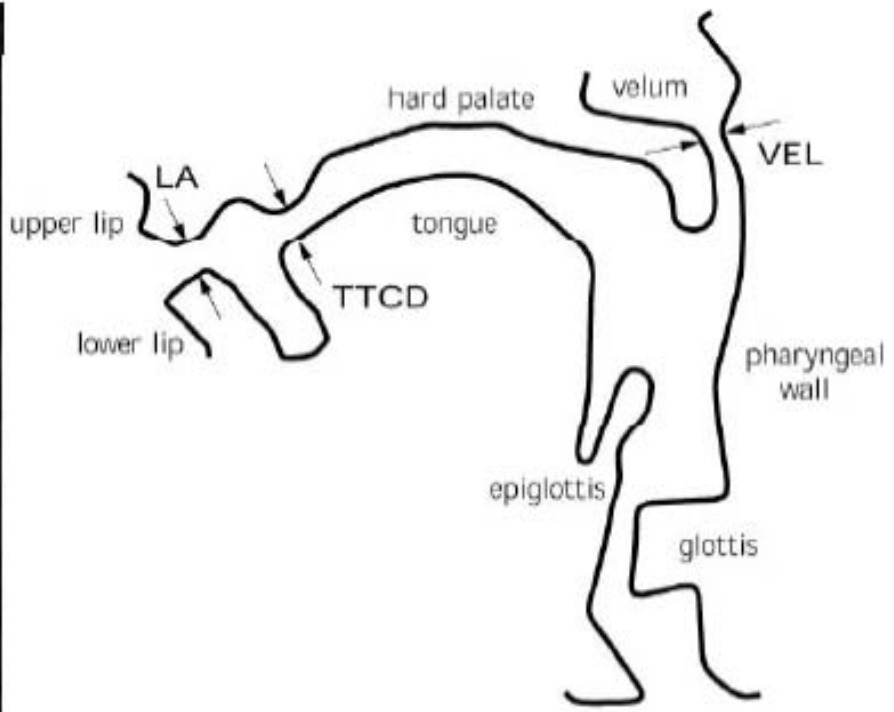
Nasal tract — begins at the velum and ends at the nostrils

Velum —a trapdoor-like mechanism at the back of the mouth cavity; lowers to couple the nasal tract to the vocal tract to produce the nasal sounds like /m/ (mom), /n/ (night), /ng/ (sing)

MRI of Speech Human Production system (Prof. Shri Narayanan, USC)

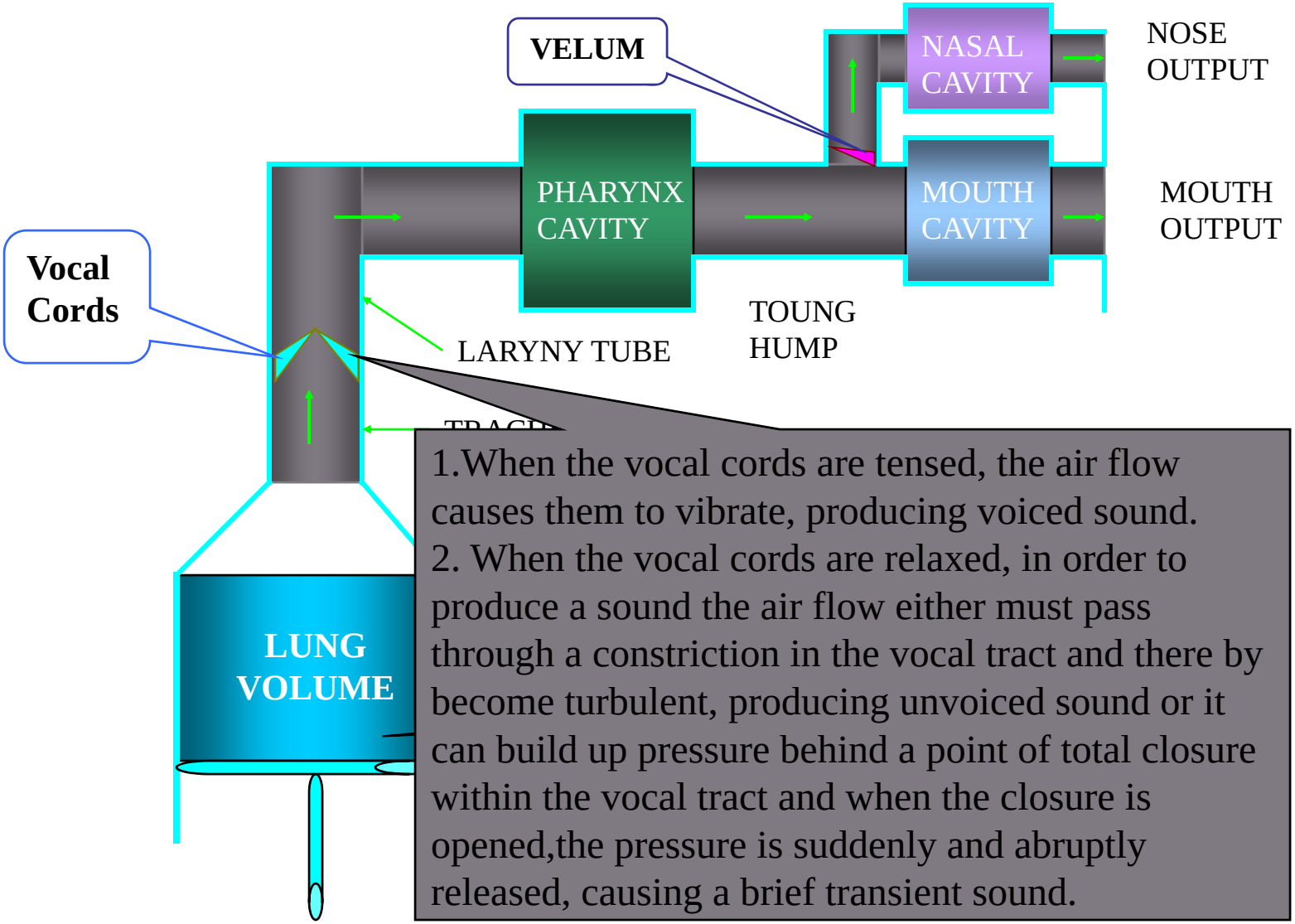


(a)

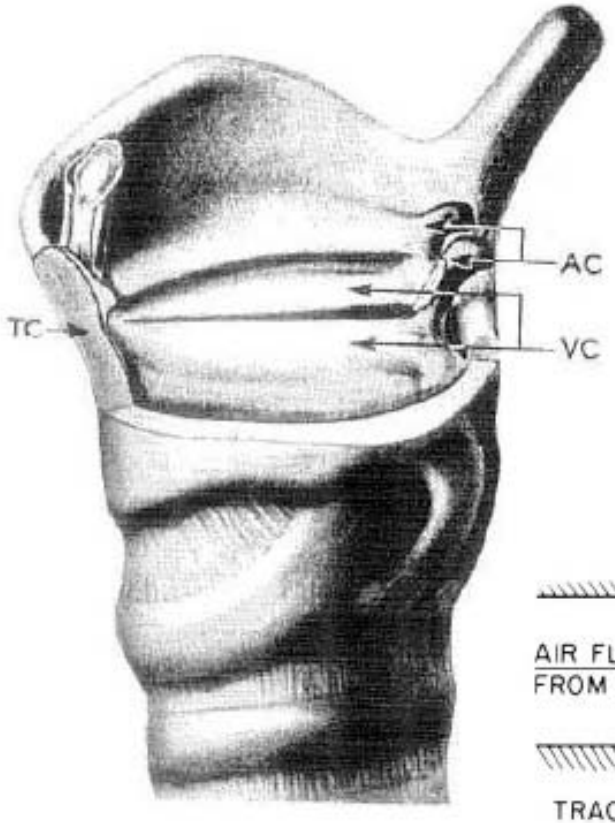


(b)

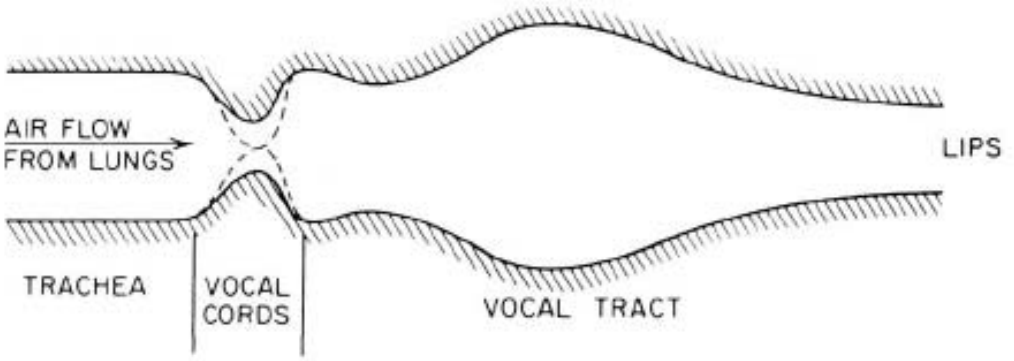
Schematic representation of the physiological mechanism of speech production



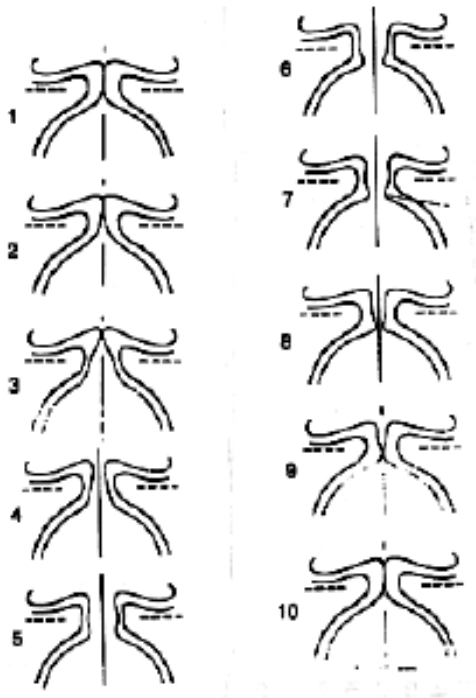
Vocal Cords



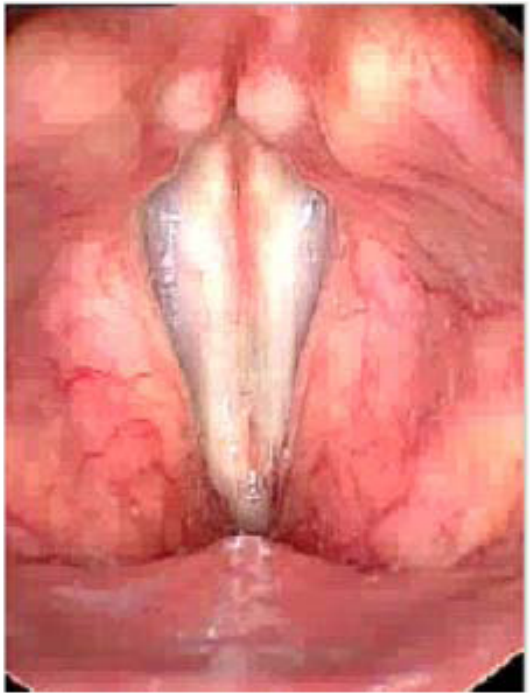
The vocal cords (folds) form a relaxation oscillator. Air pressure builds up and blows them apart. Air flows through the orifice and pressure drops allowing the vocal cords to close. Then the cycle is repeated.



Vocal Cord Views and Operation

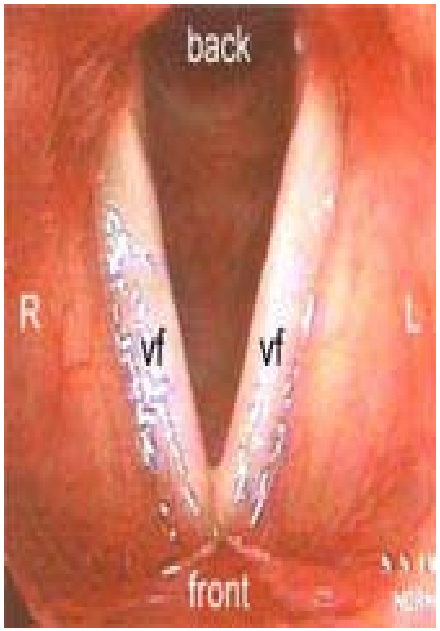


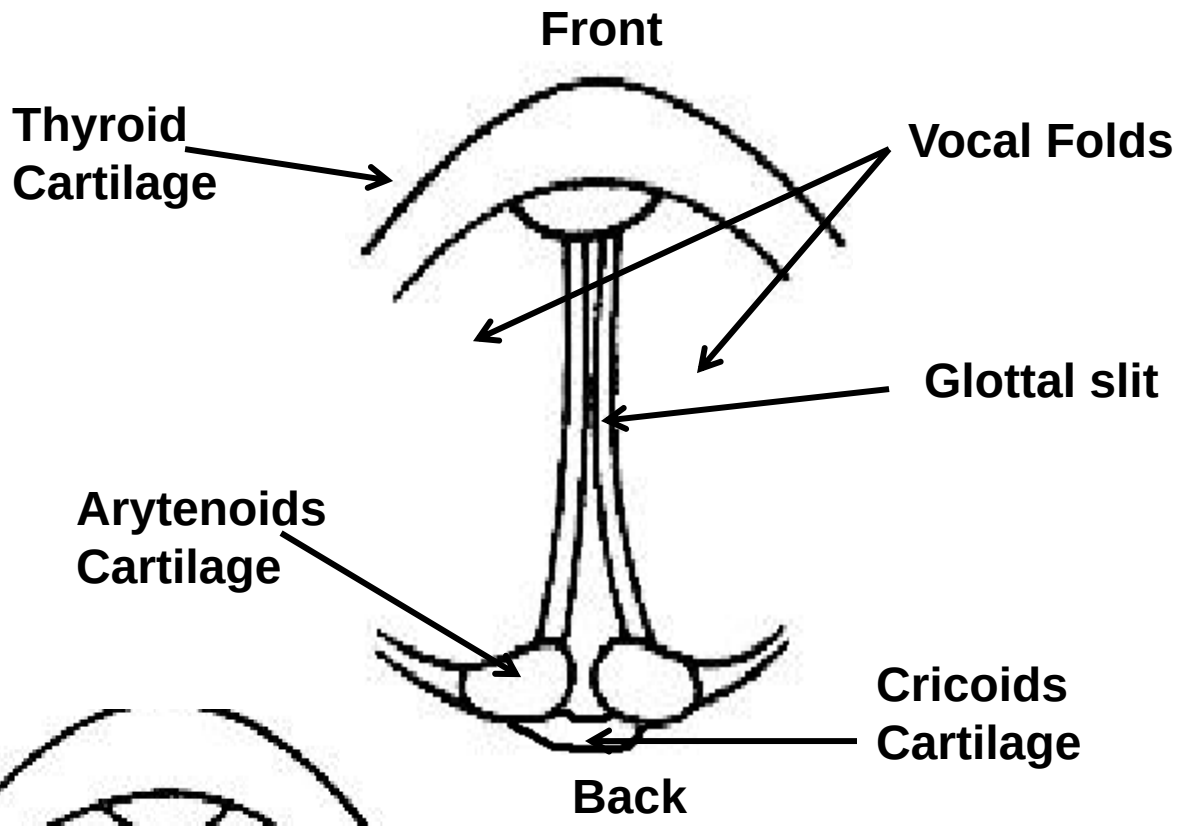
Bernoulli Oscillation

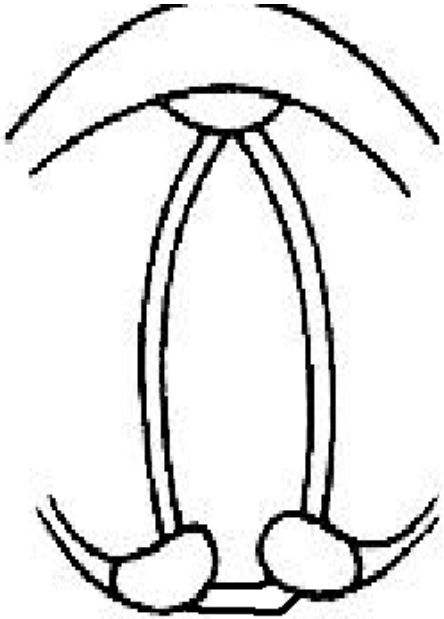


Tensed Vocal Cords – Ready to Vibrate

Vocal Cords – Open for Breathing

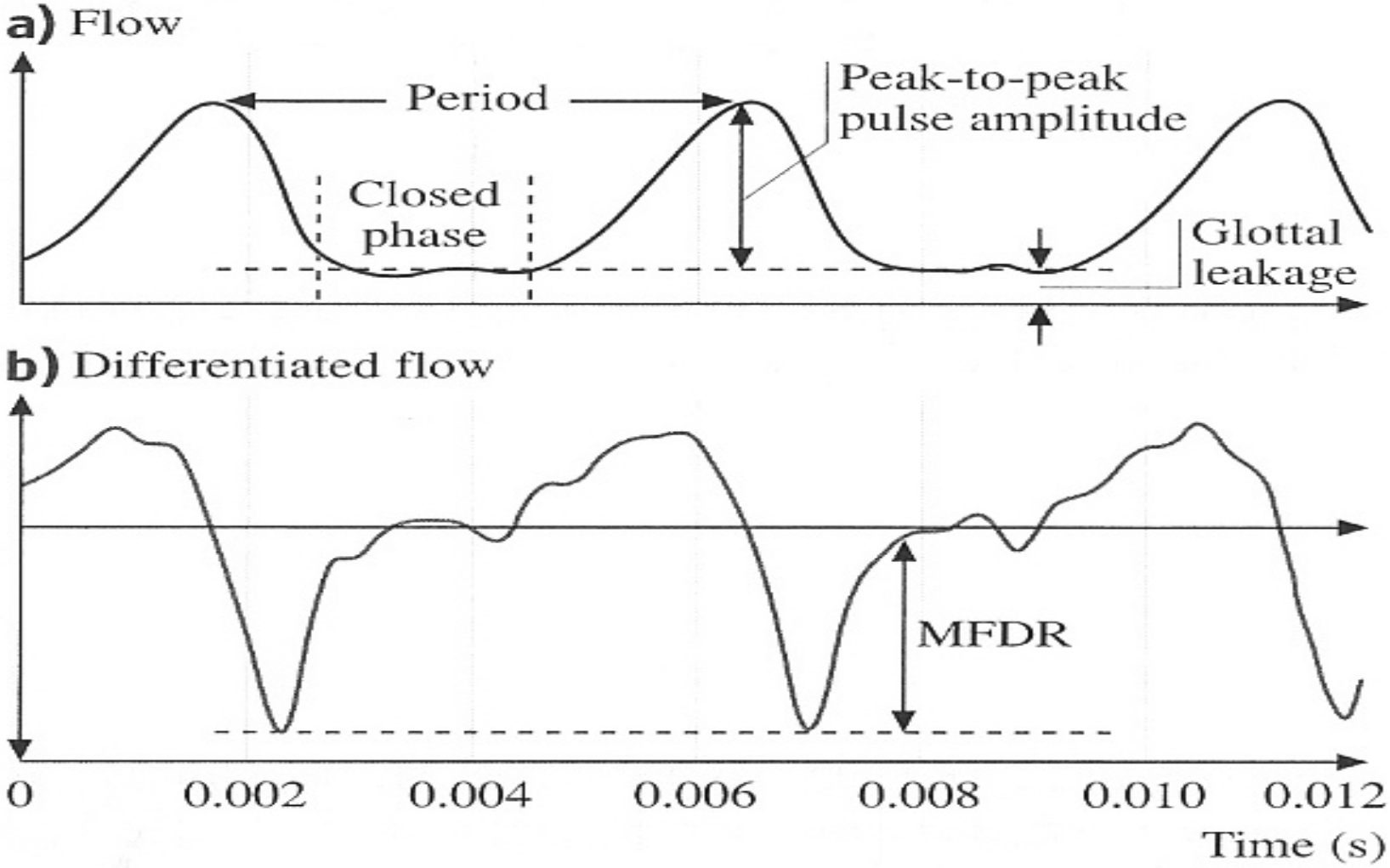






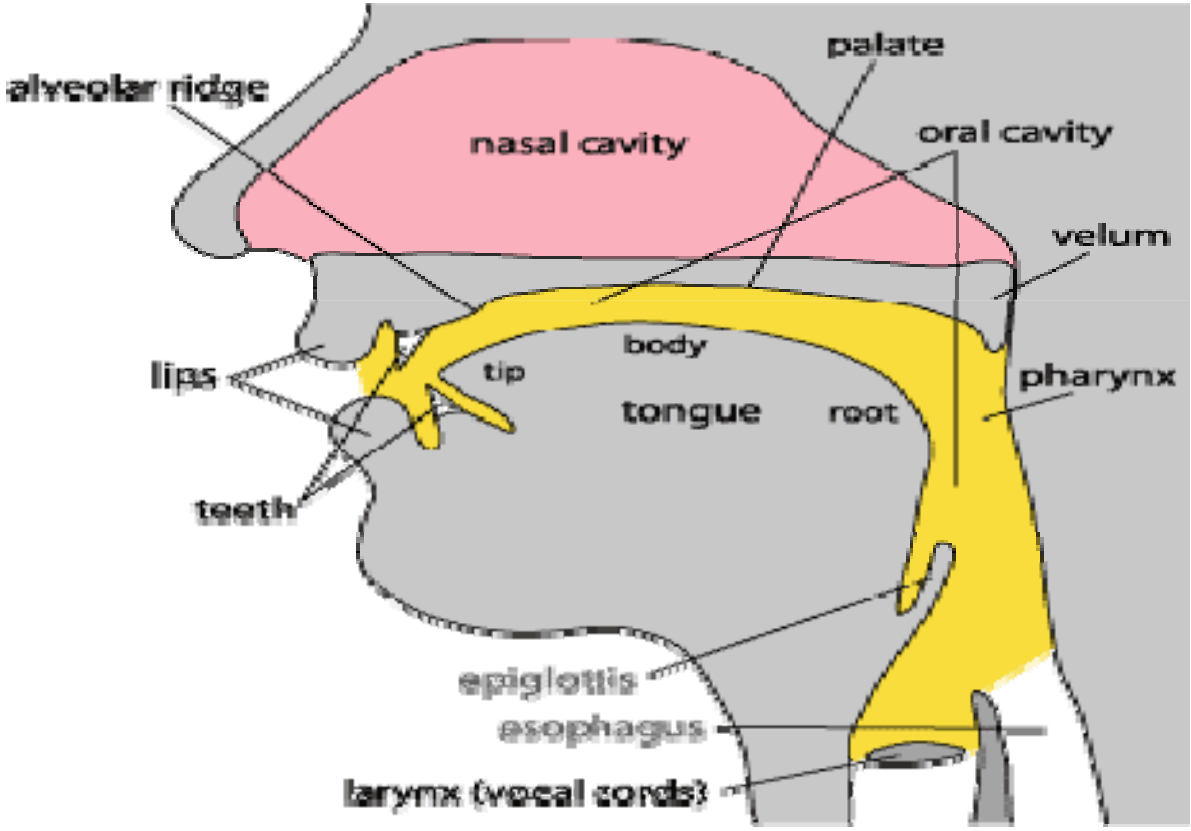
- Breathing
- Voiced
- Unvoiced

Glottal Flow

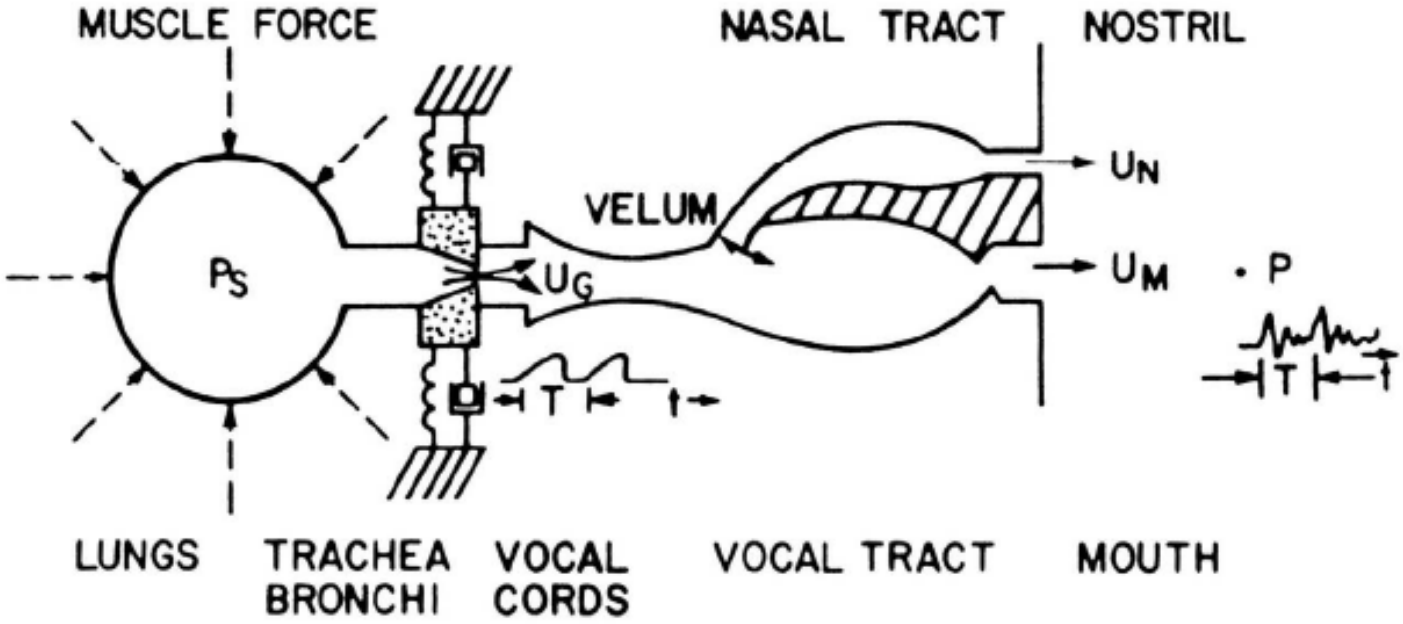


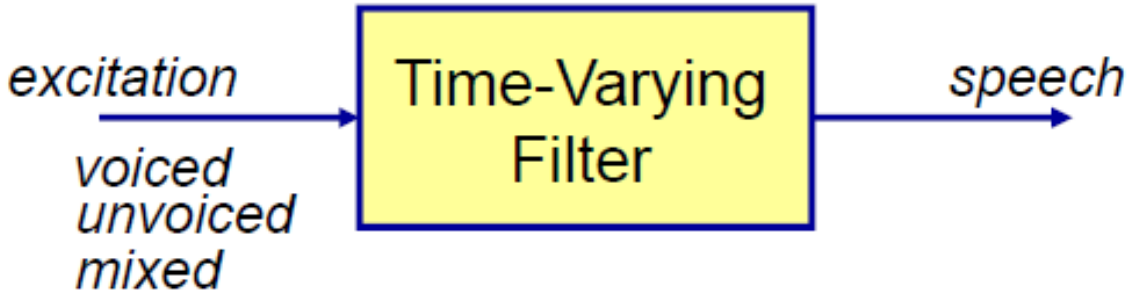
The Vocal Tract

- **The shape of the vocal tract transforms raw sound from the vocal folds into recognizable sounds.**

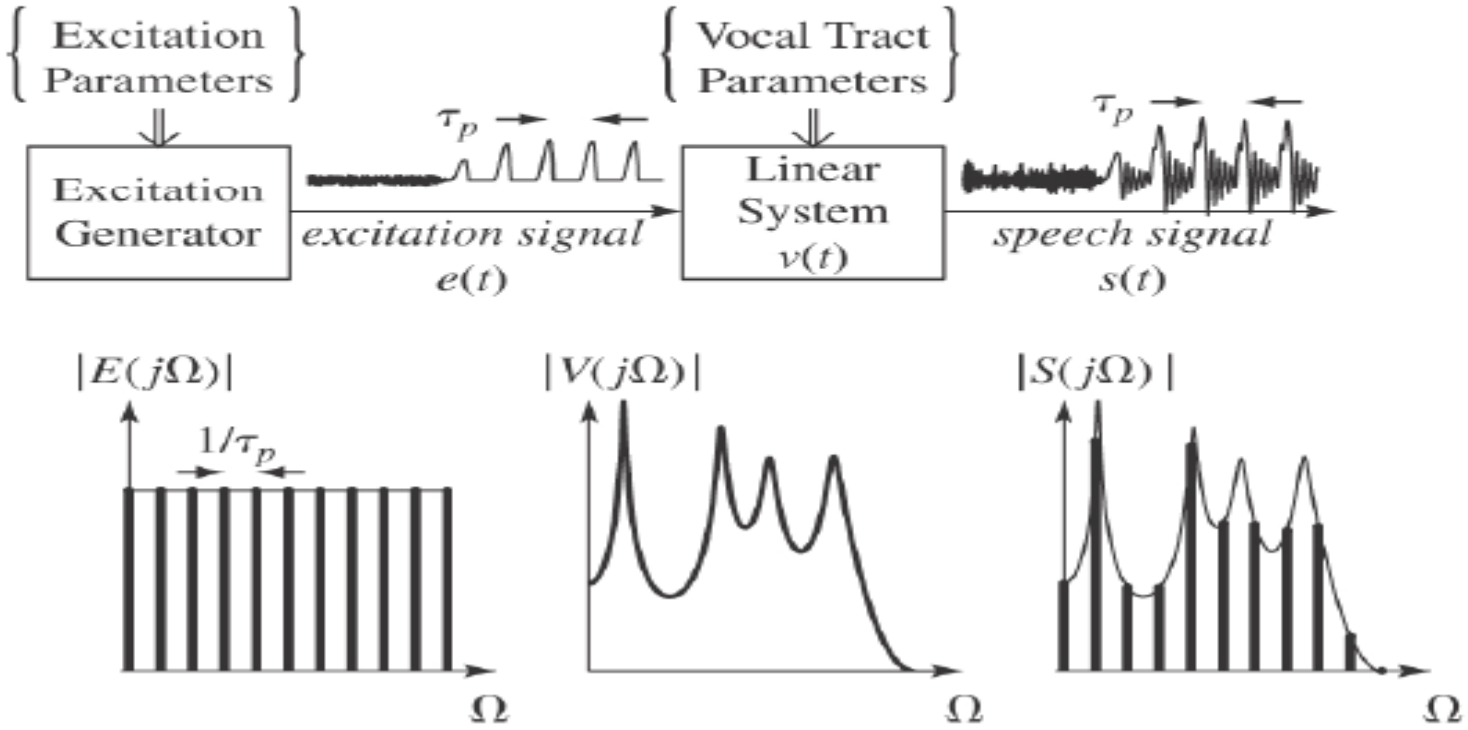


Abstractions of Physical Model





Source-System Model of Speech Production



Women and Men

- The acoustics of male and female vowels differ reliably along two different dimensions:
    1. Sound **Source**
    2. Sound **Filter**
- Source--F_0: Depends on length of vocal folds

  Shorter in women $\Rightarrow$ higher average F_0

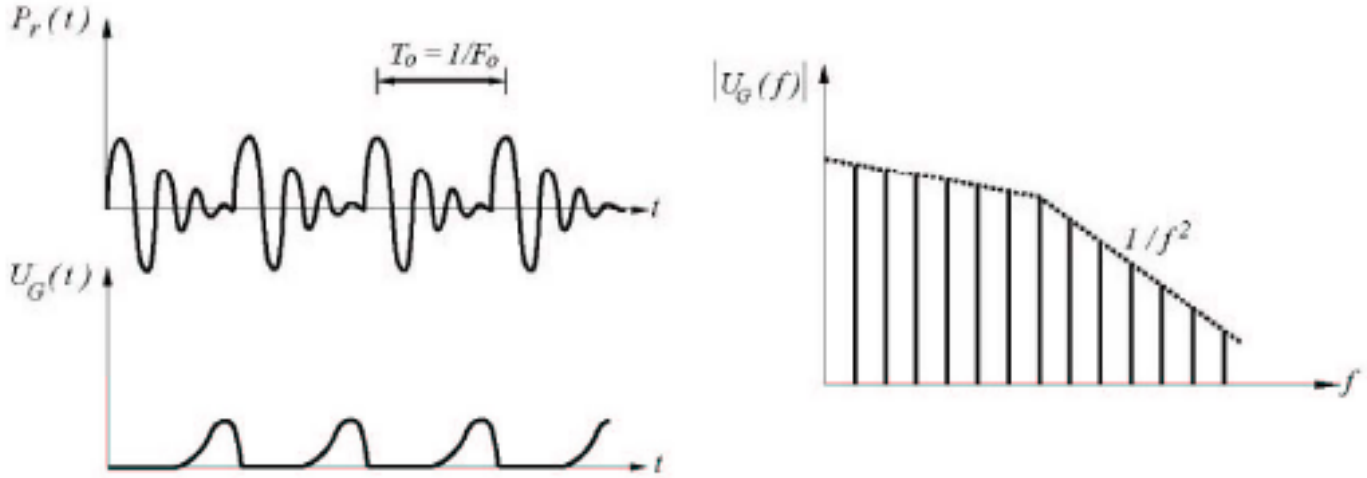
  Longer in men $\Rightarrow$ lower average F_0
- **Filter--Formants**: Depend on length of vocal tract

  shorter in women $\Rightarrow$ higher formant frequencies

  longer in men $\Rightarrow$ lower formant frequencies

Sound Source for Voiced Sounds

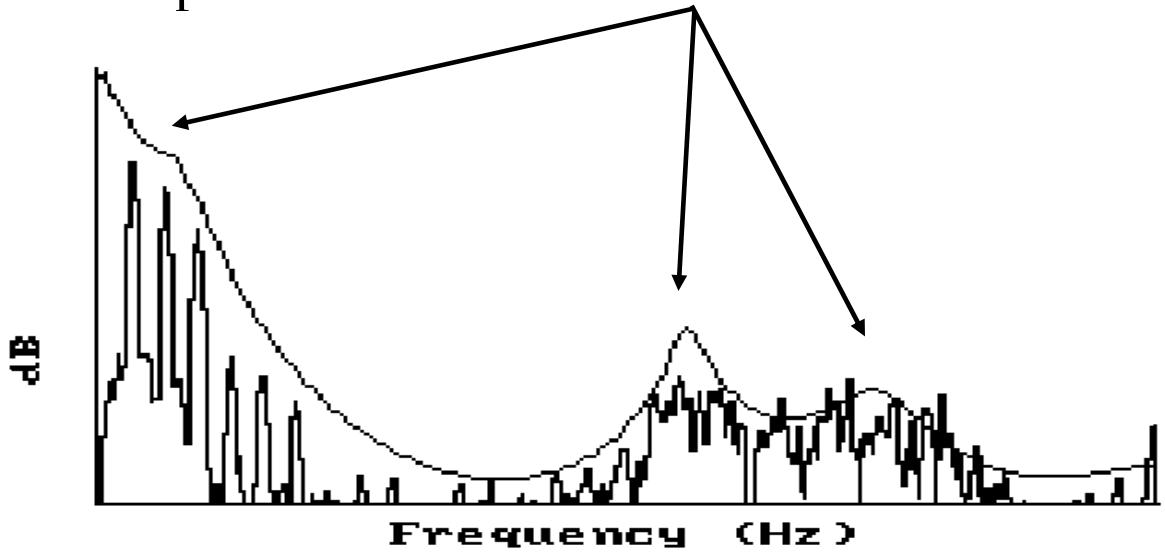
Modelled as a volume velocity source at glottis, $U_G(j\Omega)$



	F_0 ave (Hz)	F_0 min (Hz)	F_0 max (Hz)
Men	125	80	200
Women	225	150	350
Children	300	200	500

What is Formant??

➢To identify dissimilar sounds i.e., vowels, the ears are more sensitive to peaks in the signal spectrum. These resonant peaks in the spectrum are called formants.



Spectrographic view of vowel /i/

➢Formants are the characteristics partial that identify vowels to the listeners.

➢Formant with lowest frequency is called F1, the second F2 & the third F3. F1 & F2 are enough to disambiguate the vowel.

Sound Source for Unvoiced Sounds

- Turbulence noise is produced at a constriction in the vocal tract
  - Aspiration noise is produced at glottis
  - Frication noise is produced above the glottis

Articulatory and Acoustic phonetics

Manners and Place of Articulation

Place of articulation: During the articulation the airstreams through the vocal tract must be obstructed in some way. The place where the obstruction takes place is called the place of articulation

Manner of articulation: Manner of articulation is concerned with airflow ; the paths it take and the degree to which it is impeded by vocal tract constrictions.

The consonants are classified depending on the place of obstruction and manner of articulation.

/k/ Velar Un-aspirated unvoiced stop

Vowel sound specified in terms of the position of the tongue and the position of the lips.

/i/ High front Un-rounded

Manners of Articulation due to State of the Glottis

If the glottis are closed then it is voiced and if opened then it is unvoiced or voiceless.

Place of articulation

a. **Bilabial:** Bilabial sounds are produced when the two lips make the constriction

b. **Labiodentals:** These sounds are produced by contacting lower lip with the upper teeth.

c. **Dental:** Dental sounds are produced by the constriction of tip or blade of the tongue with the upper teeth.

d. **Alveolar:** The sound made by the tip or the blade of the tongue in contact against the alveolar ridge, which is the bony prominence immediately behind the upper teeth.

e. **Post alveolar:** The sound, which is articulated by the tip or the blade of the tongue with the back area of the alveolar ridge.

f. **Retroflex** : Retroflex sounds are made when the tip of the tongue curled back in the direction of the front part of the hard palate- in other words just behind the alveolar ridge. Depending on how far the tongue curls back, retroflexed could be apico-postalveolar or apico-palatal.

g. **Palatal:** This sound is produced when the constriction is made by the front part of the tongue with the hard palate.

h. **Velar:** It refers to a sound made by the back of the tongue against the soft palate.

i. **Uvular:** This sound is produced when the back of the tongue touches the uvula.

j. **Pharyngeal:** It refers to a sound produced in the pharynx, the tubular cavity, which constitutes the throat above the larynx.

k. **Glottal:** These are the sounds, which made in the larynx due to the closure or narrowing of the glottis.

Manner of articulation

a) Plosive, or stop
b) Nasal stop
c) Fricative
d) Affricate
e) Lateral
f) Approximant
g) Trill:
h) Flap and Tap

1. Voiced
2. Unvoiced
3. Aspiration

Classification of sound in linguistically distinct speech (phonemes)

- Vowels: a) Oral vowels b) nasal vowels

- Dipthongs: Dipthongs is a gliding monosyllabic speech sound that start at or near the articulatory position for one vowel and moves to or toward the position for another

- Semivowels: Semivowels are vowel like nature. They are generally characterized by gluding transition in vocal tract area function between adjacent phonemes.

- Consonant:
  a. Nasal consonants. b. unvoiced fricatives.
  c. Voiced fricative d. voiced and unvoiced stop/ **Plosive**

CONSONANTS (PULMONIC)

© 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	B			r					R		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	’	Examples:
ǀ	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
ǃ	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative

w Voiced labial-velar approximant

ɥ Voiced labial-palatal approximant

ʜ Voiceless epiglottal fricative

ʢ Voiced epiglottal fricative

ʡ Epiglottal plosive

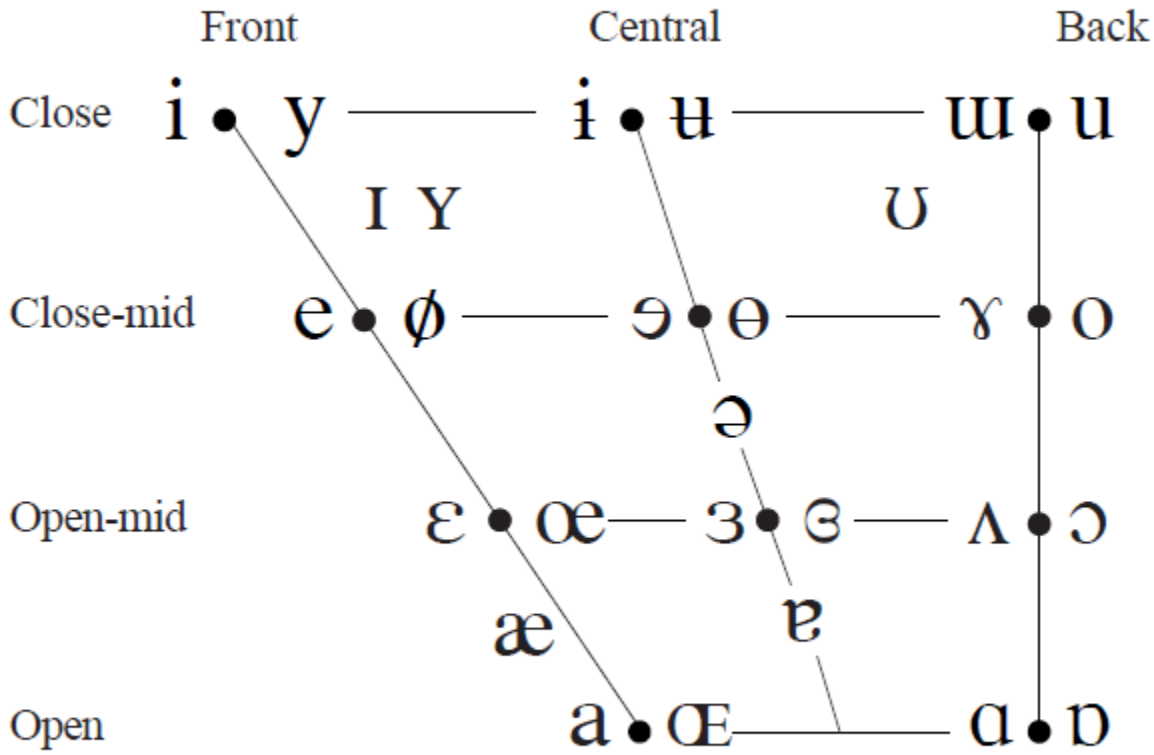
ɕ ʑ Alveolo-palatal fricatives

ɺ Voiced alveolar lateral flap

ɧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.

VOWELS

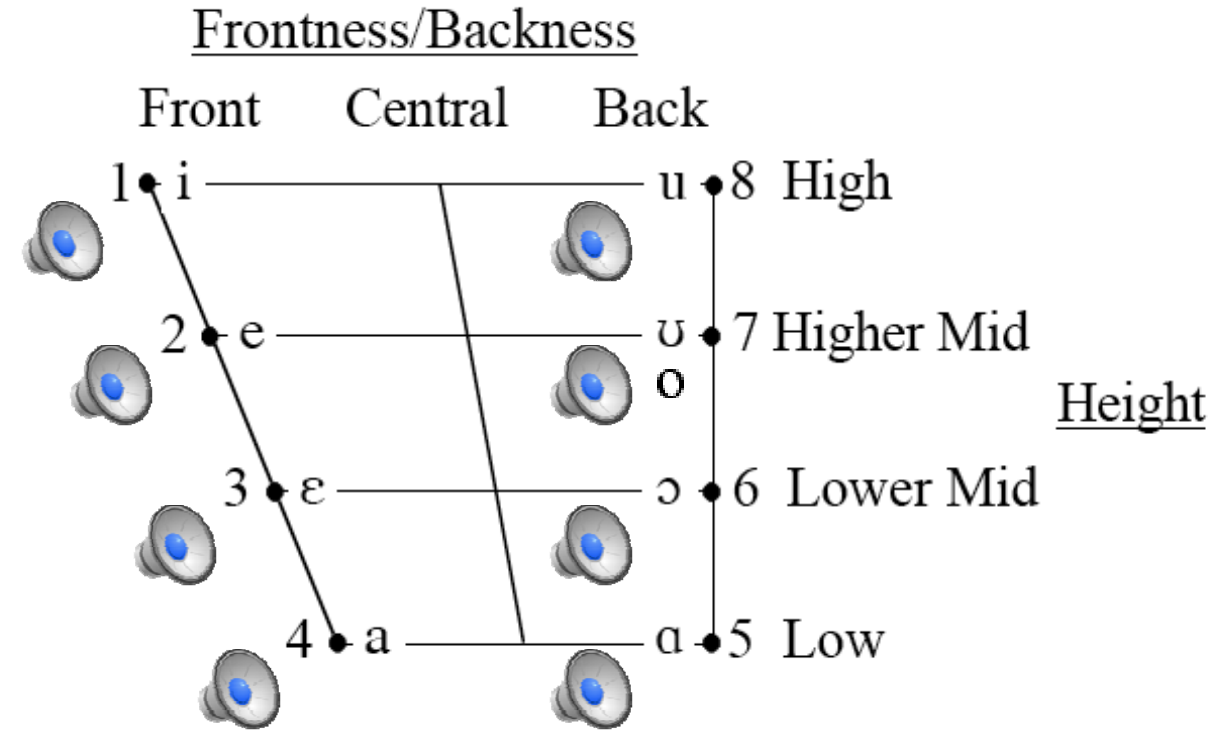


Where symbols appear in pairs, the one to the right represents a rounded vowel.

The Vowel Space

Cardinal Vowels recorded by Jones in 1965 when he was 75.

(Audio clips from: http://www.let.uu.nl/~audiufon/)



Primary cardinal vowels with rising intonation

Different Vowels, Different Formants

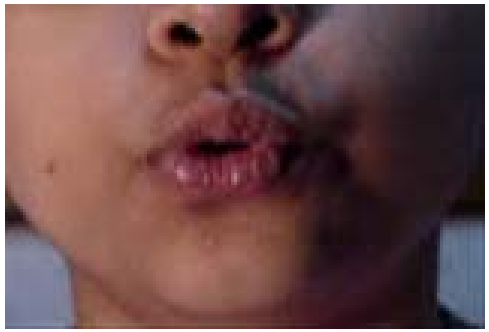
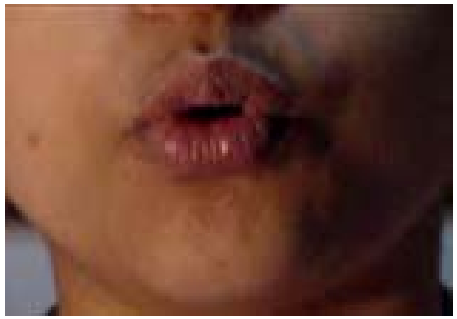
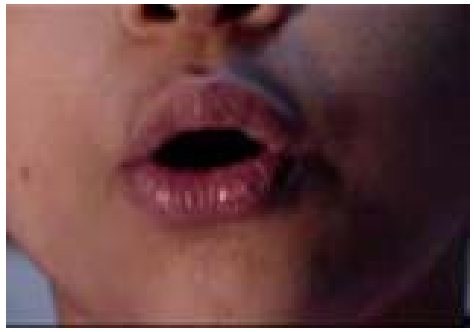
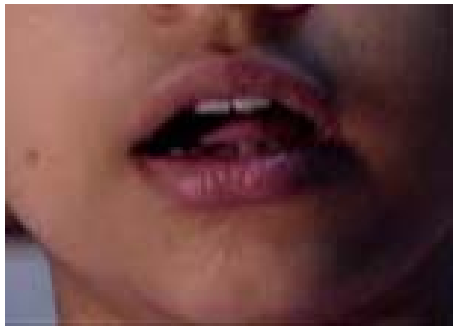
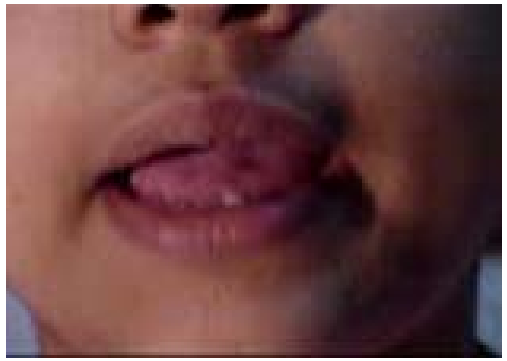
- The formant frequencies of [ə] resemble the resonant frequencies of a tube that is open at one end.
- For the average man (ref: Peter Ladefoged):
  - F1 = 500 Hz
  - F2 = 1500 Hz
  - F3 = 2500 Hz
- However, we can change the shape of the vocal tract to get different resonant frequencies.
- Vowels may be defined in terms of their characteristic resonant frequencies (**formants**).

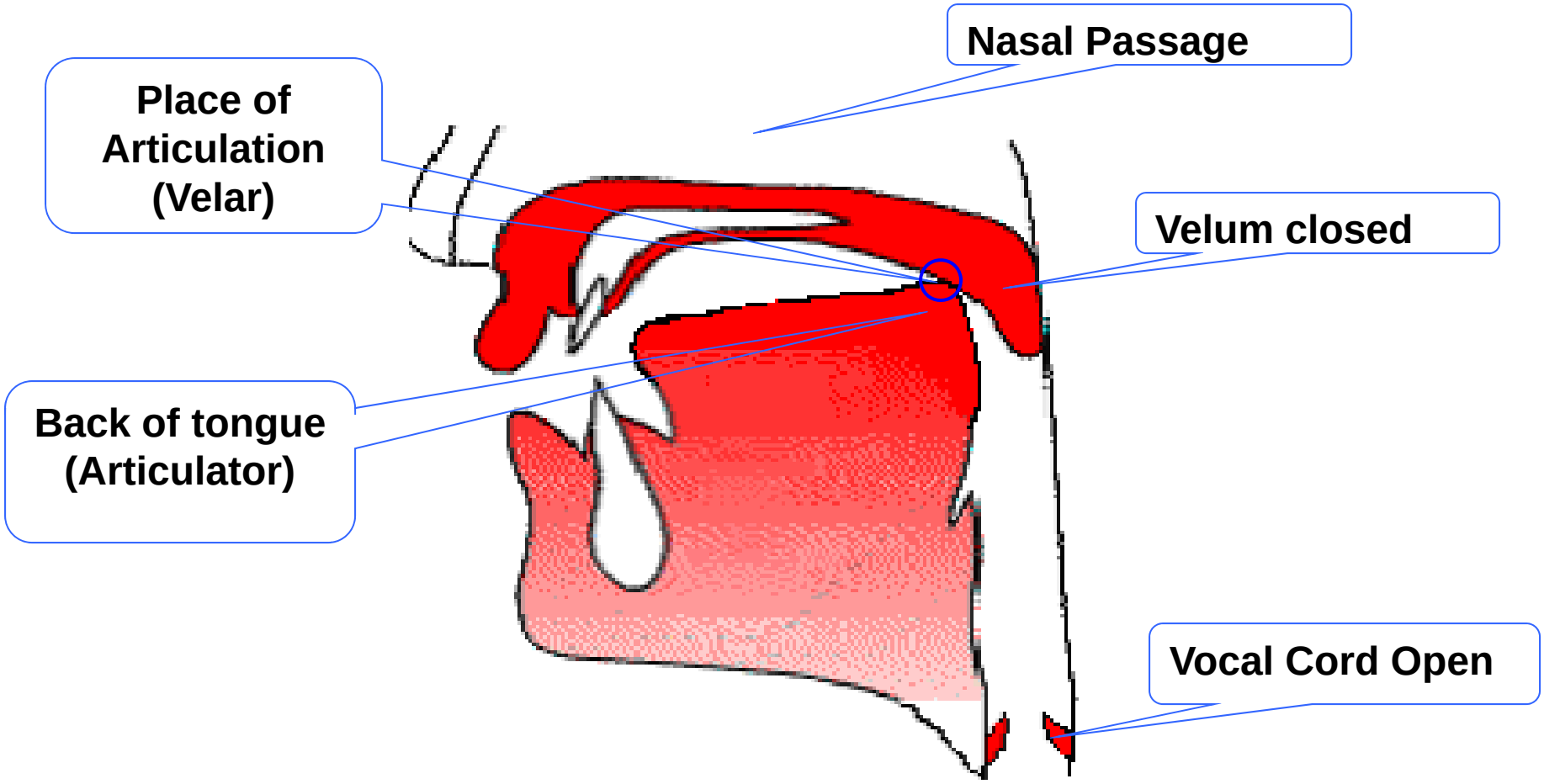
Articulatory description of Vowels

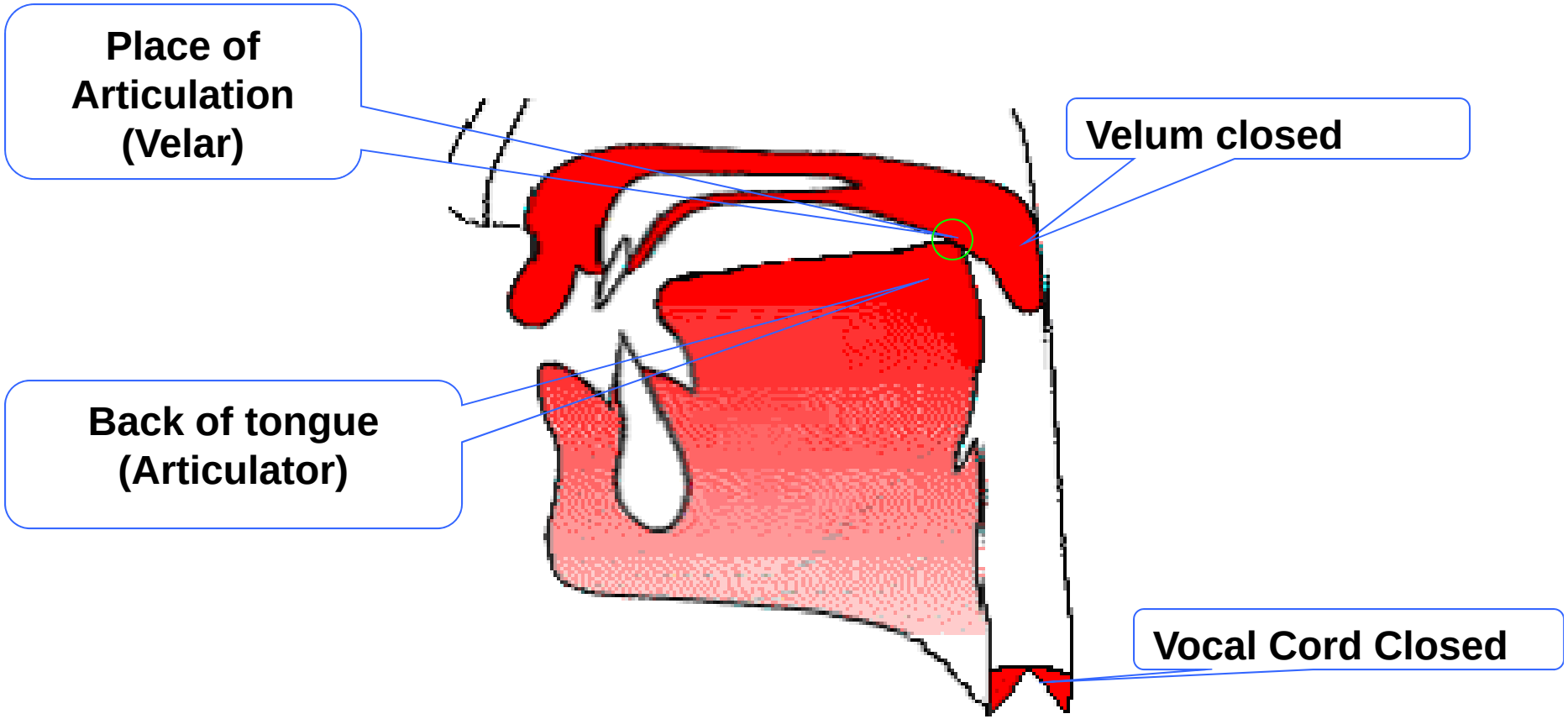
Vowels have traditionally been described according to following pseudo-articulatory parameters:

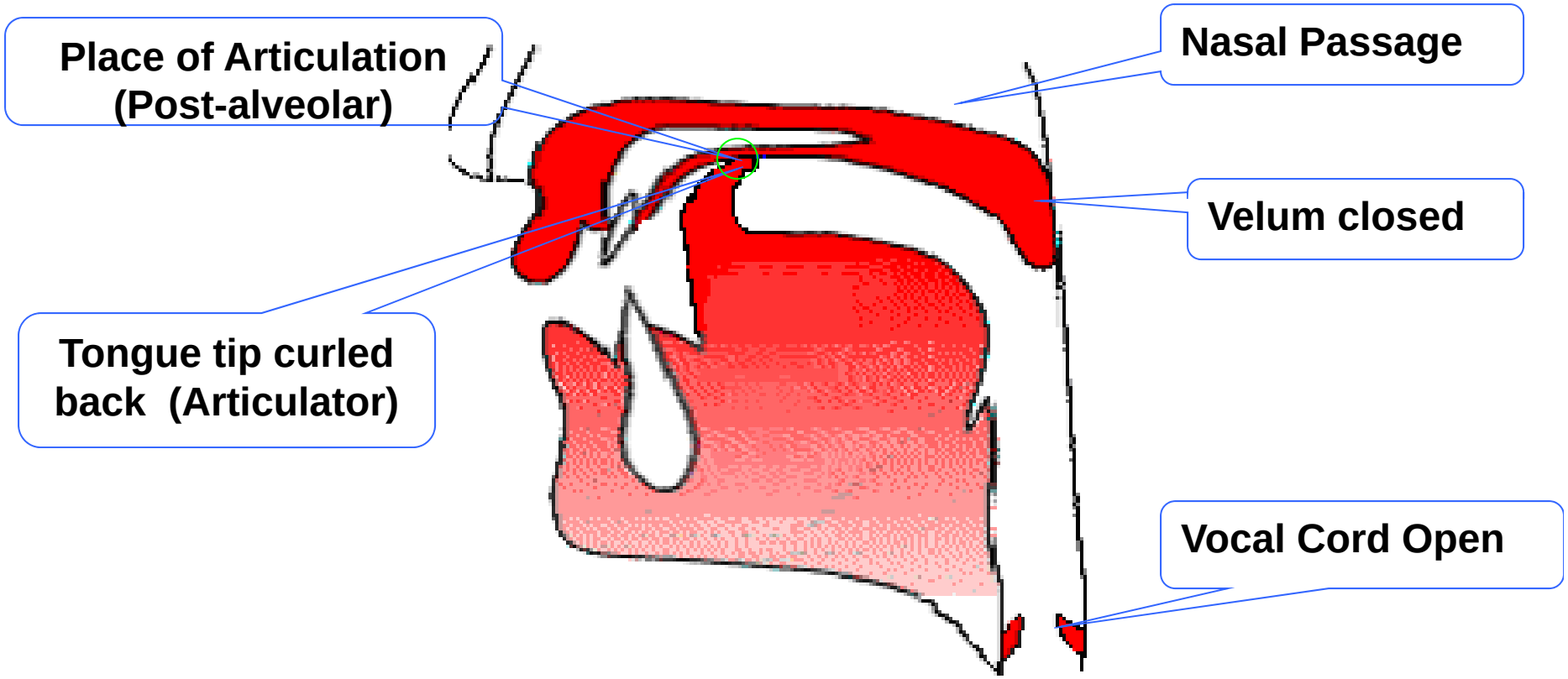
1. Height (of tongue) (F1)
2. Front/Back (of tongue)(F2)
3. Rounding (of lips)

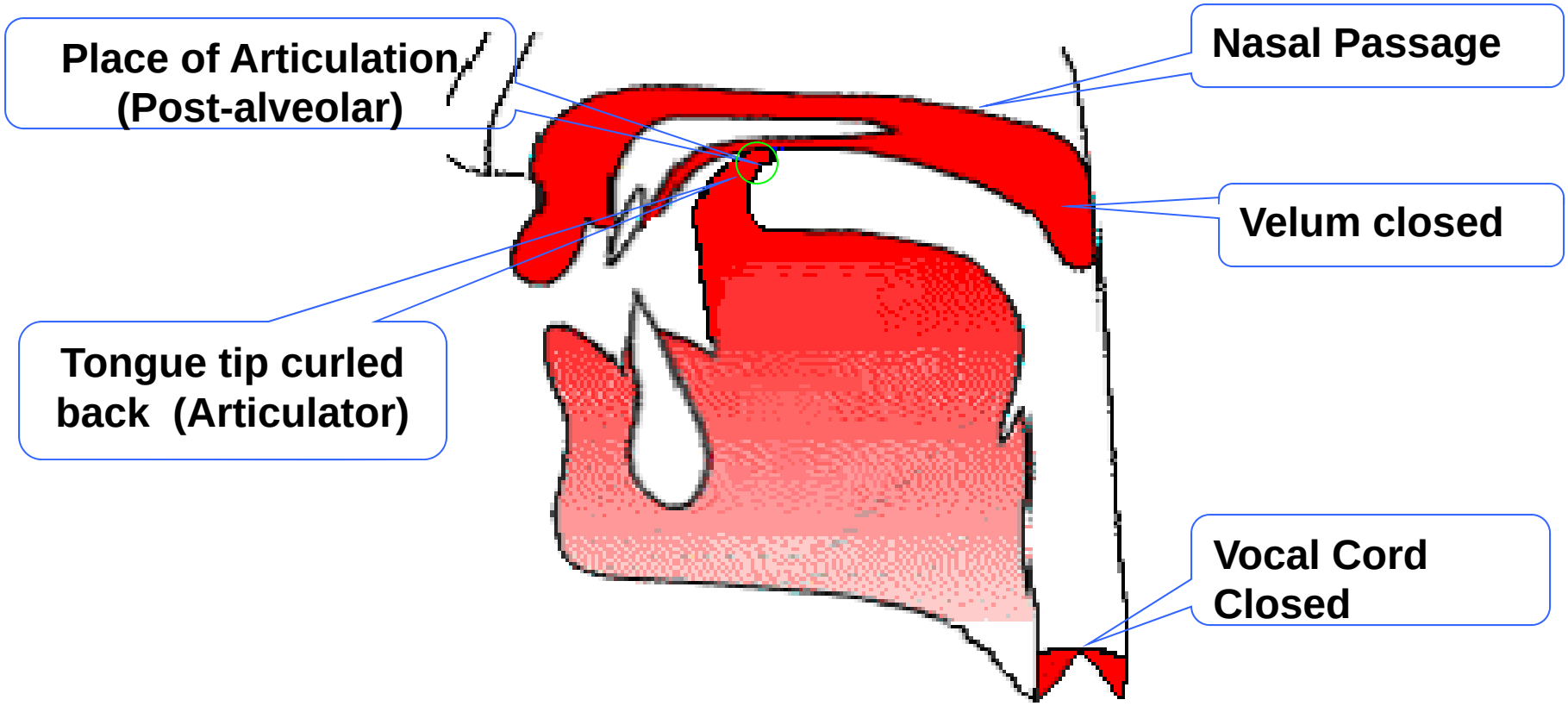
Lip rounding

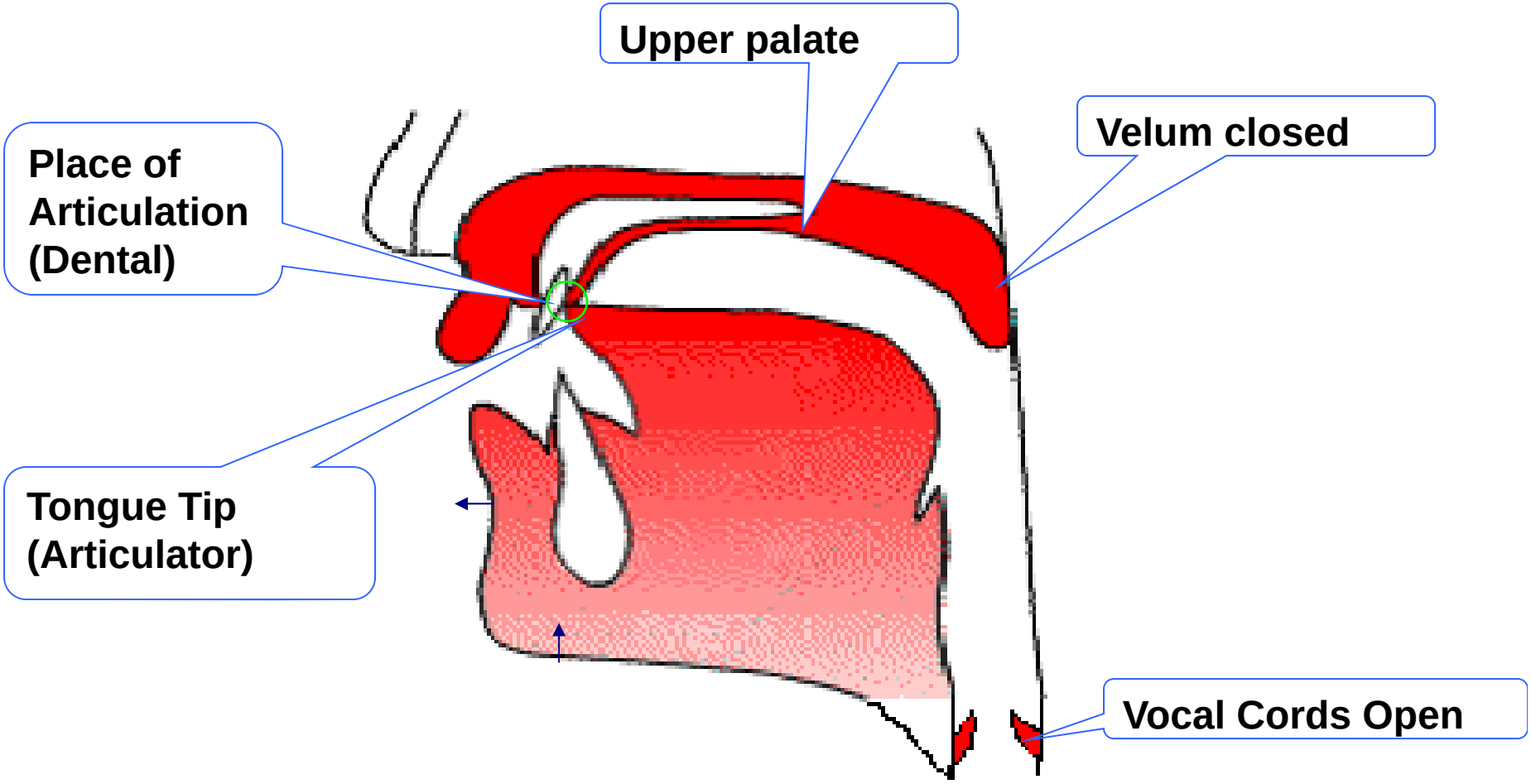






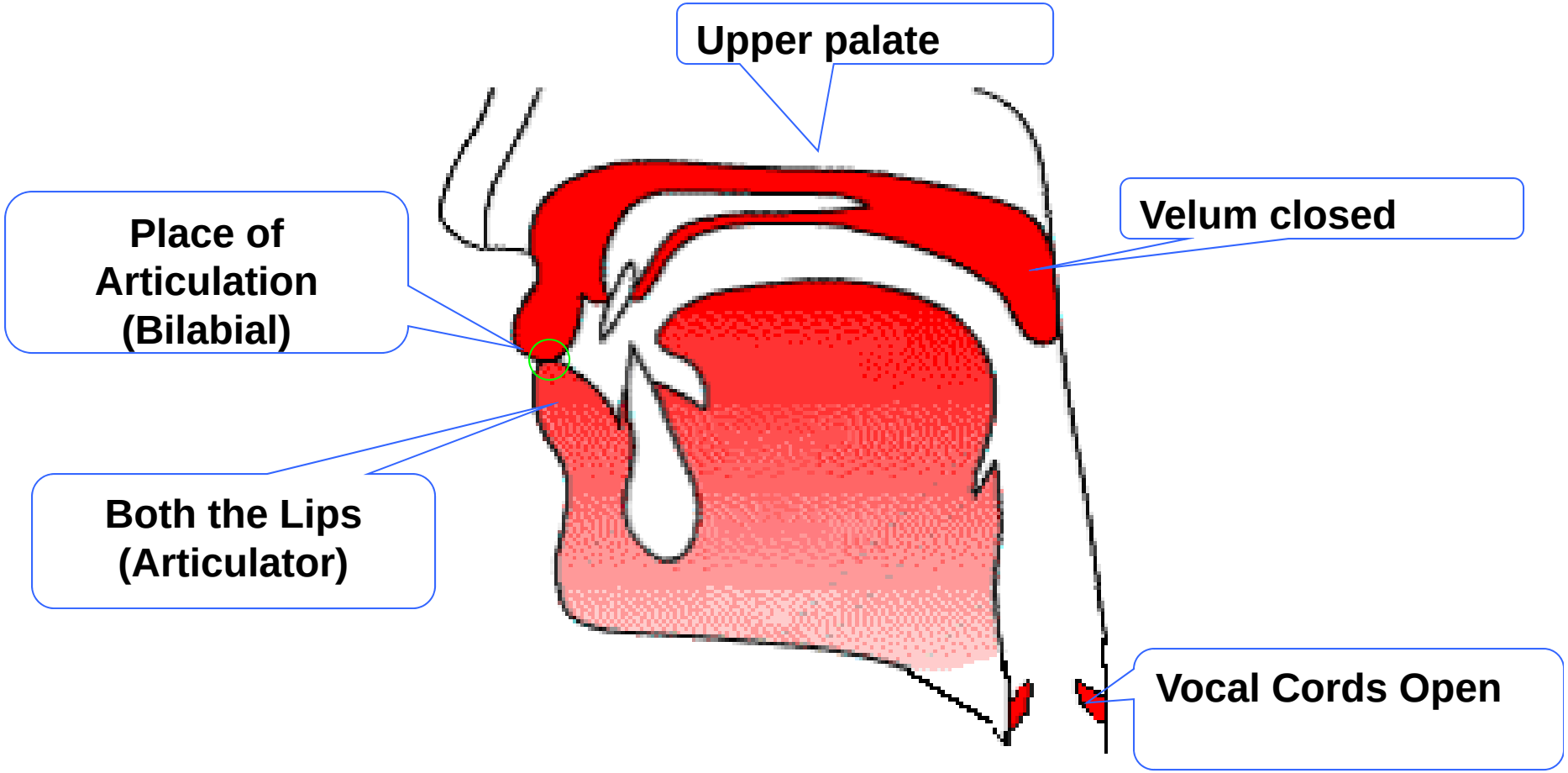

Nasal Passage
Place of Articulation (Velar)
Velum closed
Back of tongue (Articulator)
Vocal Cord Open

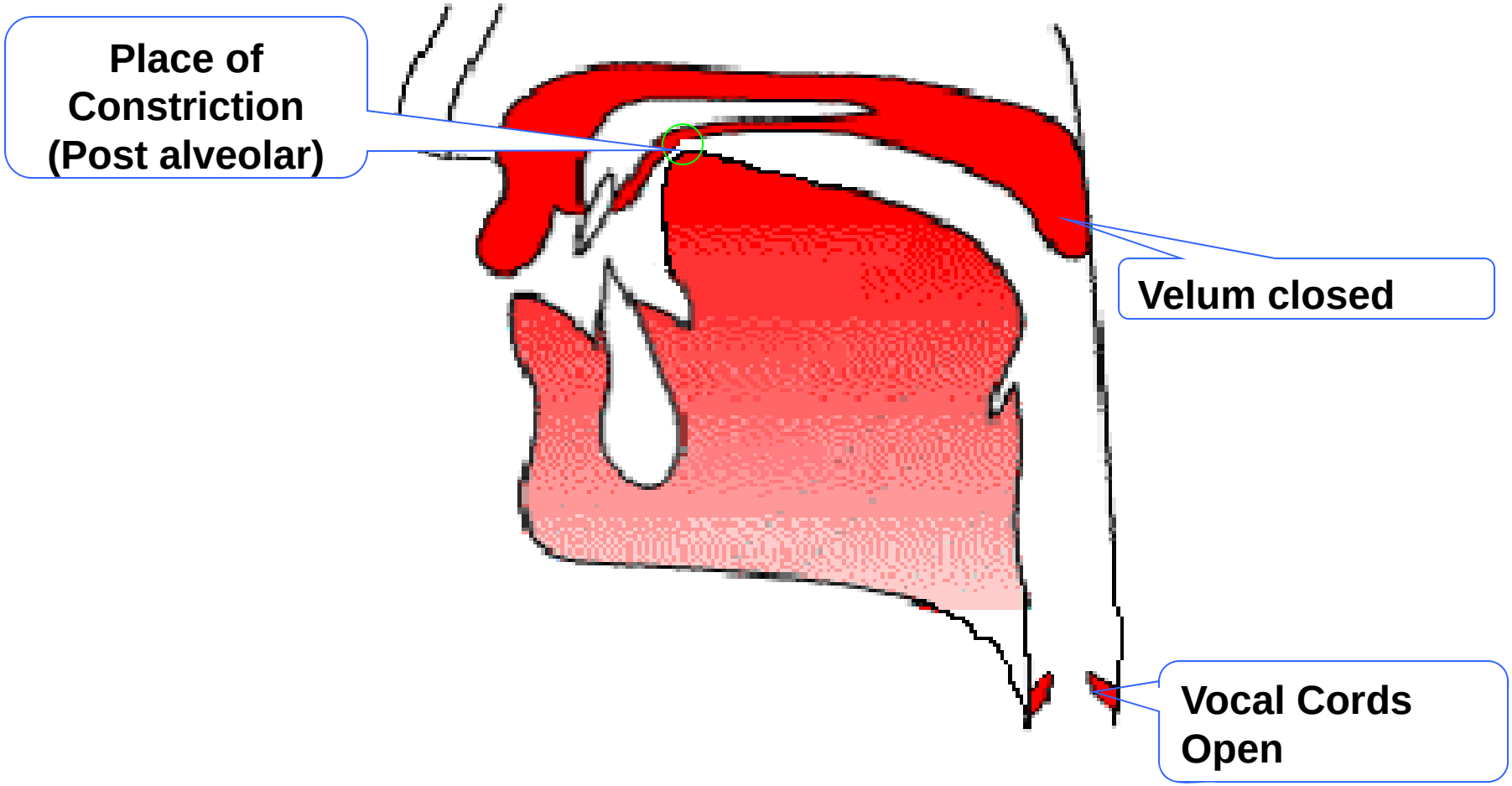

Place of Articulation (Velar)
Velum closed
Back of tongue (Articulator)
Vocal Cord Closed

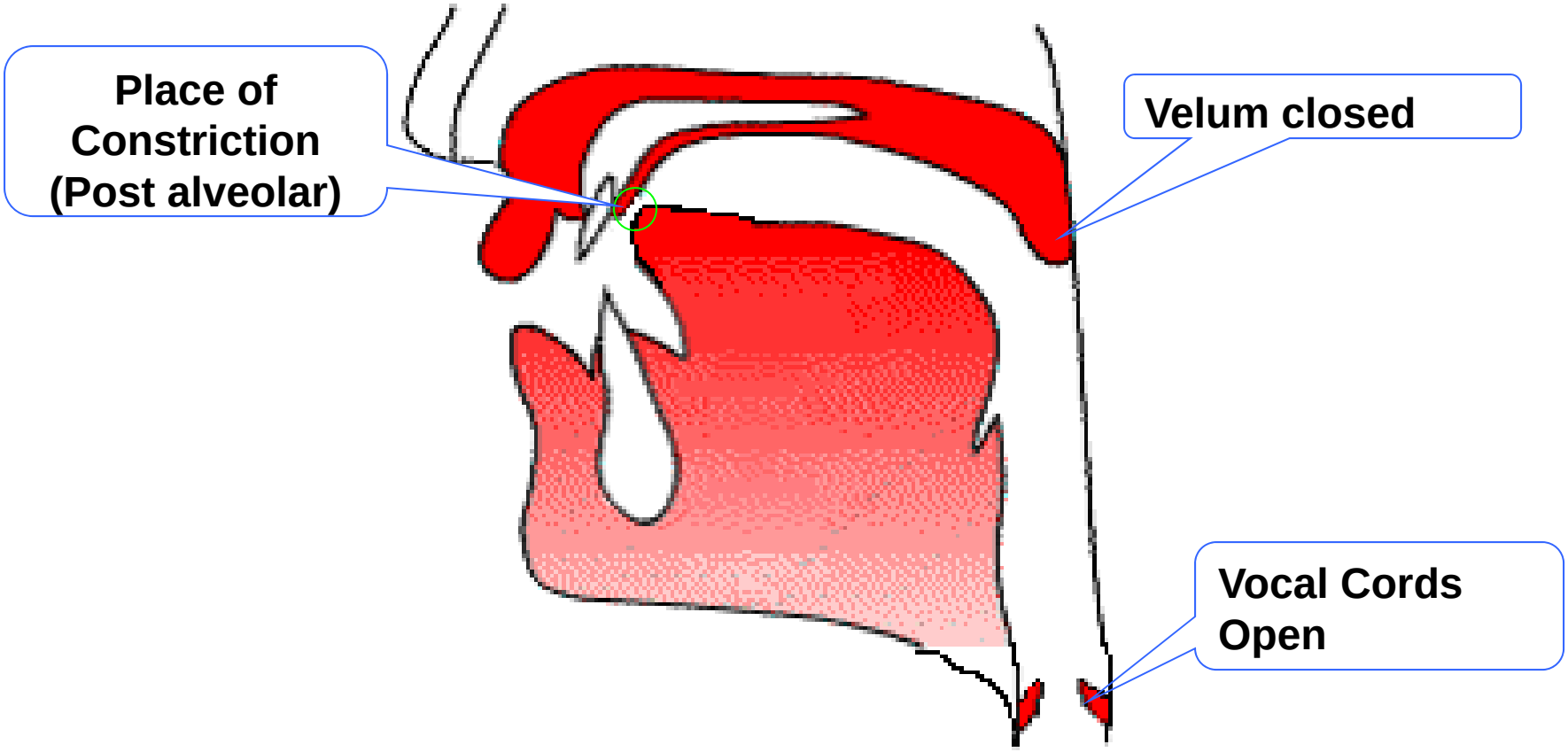

Place of Articulation
(Post-alveolar)
Nasal Passage
Velum closed
Tongue tip curled
back (Articulator)
Vocal Cord Open

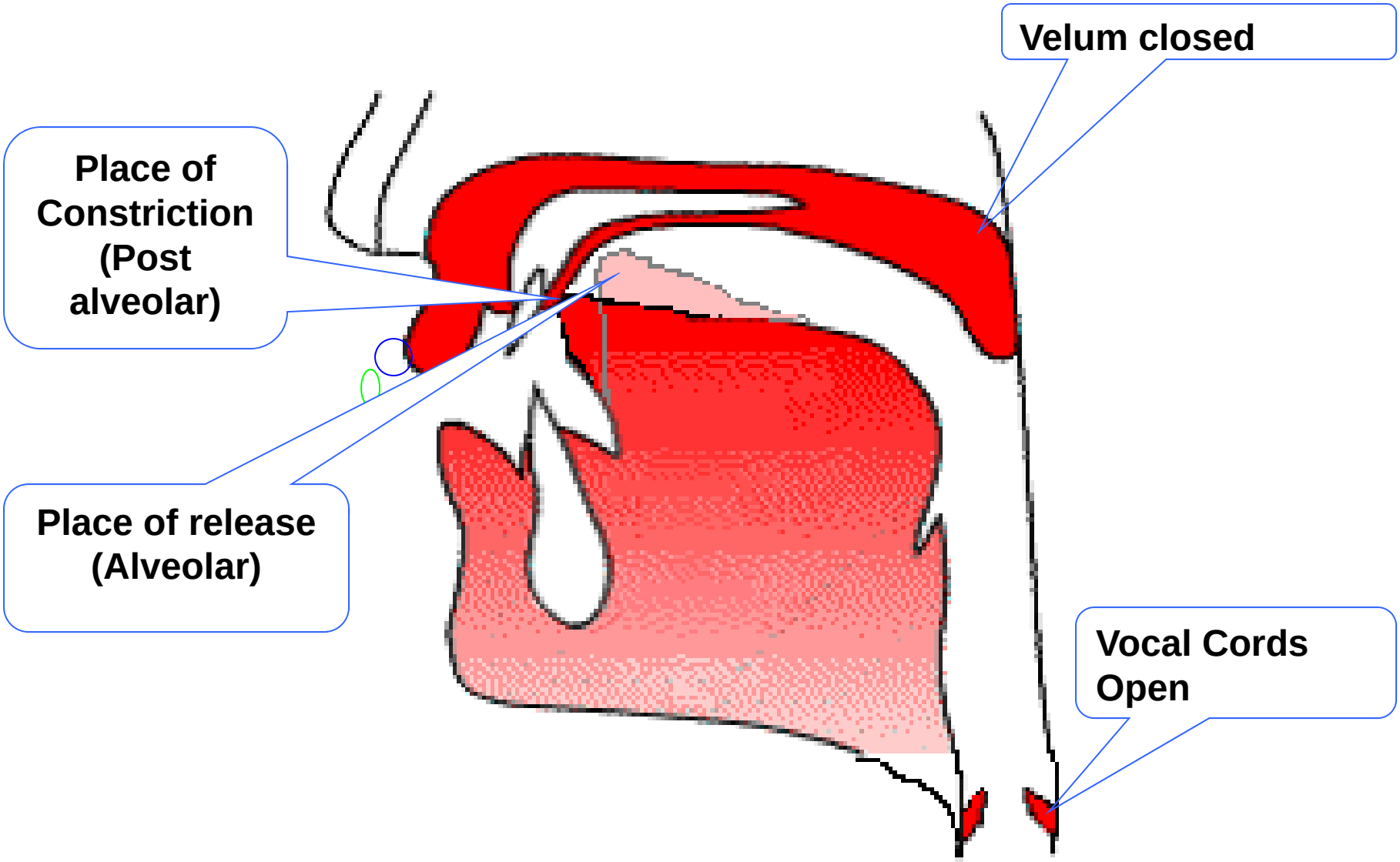

Place of Articulation (Post-alveolar)
Nasal Passage
Velum closed
Tongue tip curled back (Articulator)
Vocal Cord Closed

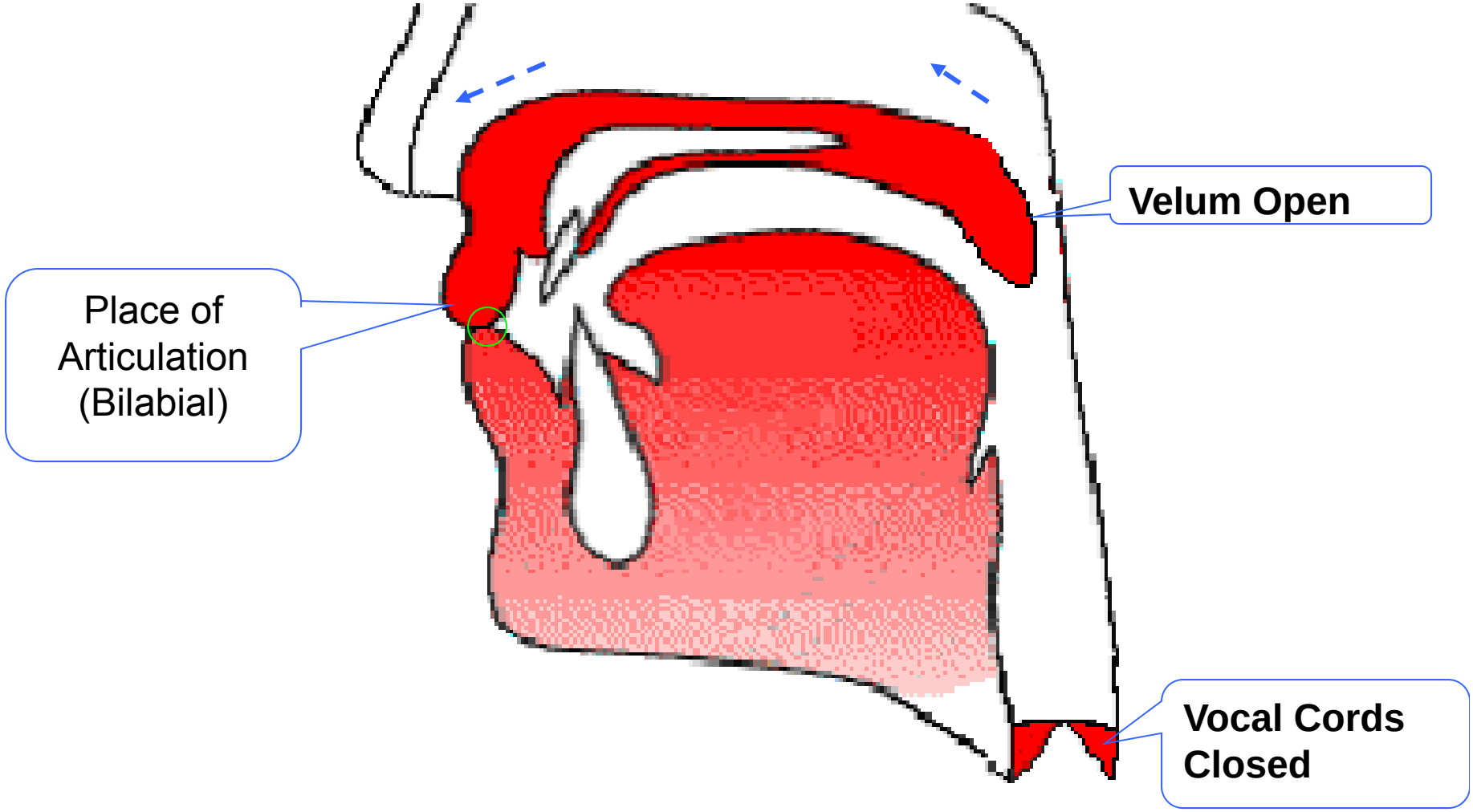

Upper palate
Velum closed
Place of Articulation (Dental)
Tongue Tip (Articulator)
Vocal Cords Open


Upper palate
Velum closed
Place of Articulation (Bilabial)
Both the Lips (Articulator)
Vocal Cords Open

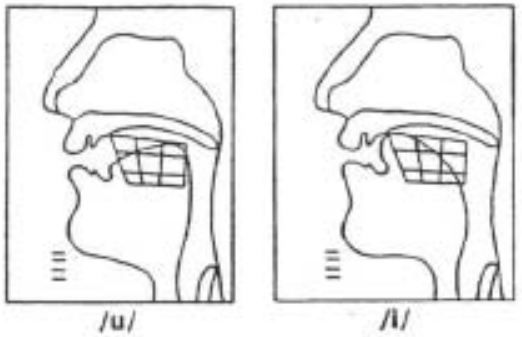

Place of Constriction (Post alveolar)
Velum closed
Vocal Cords Open


Place of Constriction (Post alveolar)
Velum closed
Vocal Cords Open


Velum closed
Place of Constriction (Post alveolar)
Place of release (Alveolar)
Vocal Cords Open

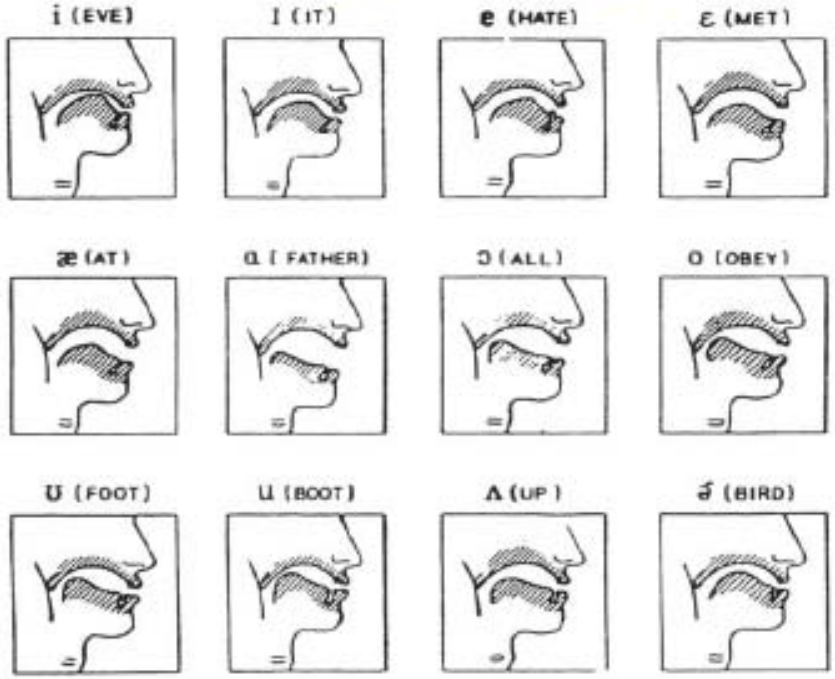

Velum Open
Place of Articulation (Bilabial)
Vocal Cords Closed

Vowel Articulatory Shapes

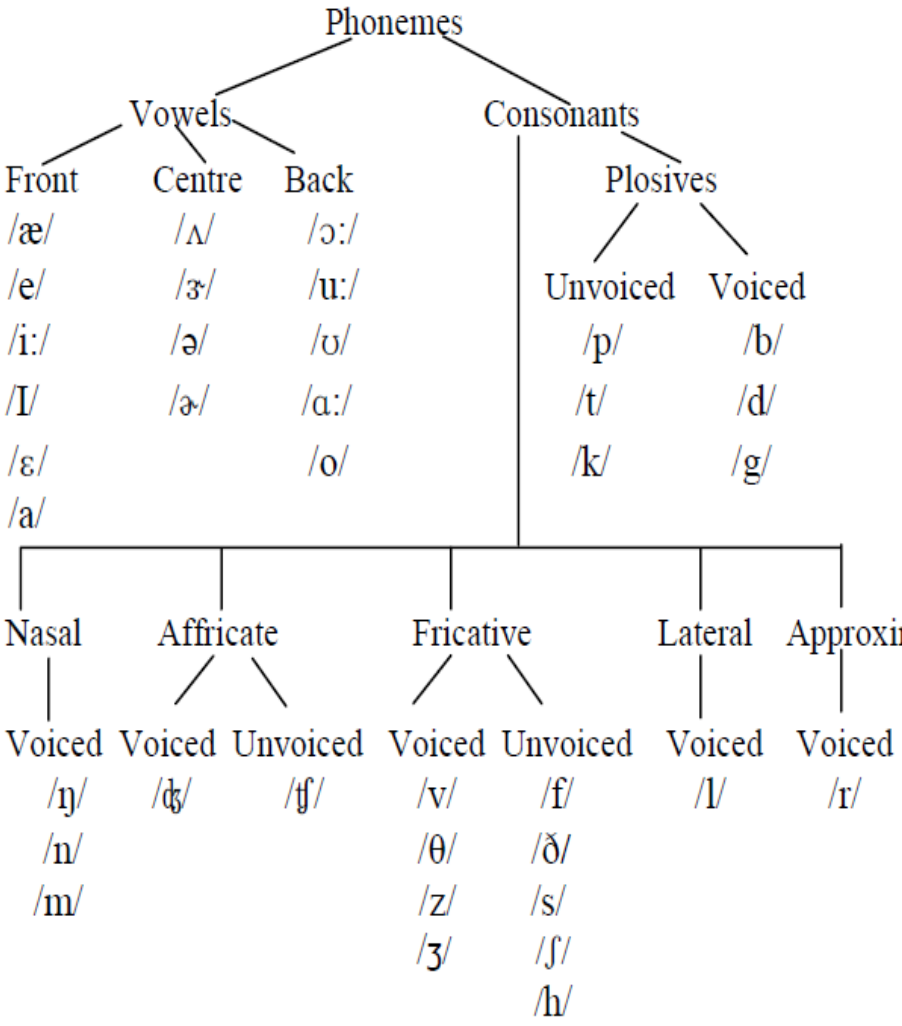


TONGUE POSITION

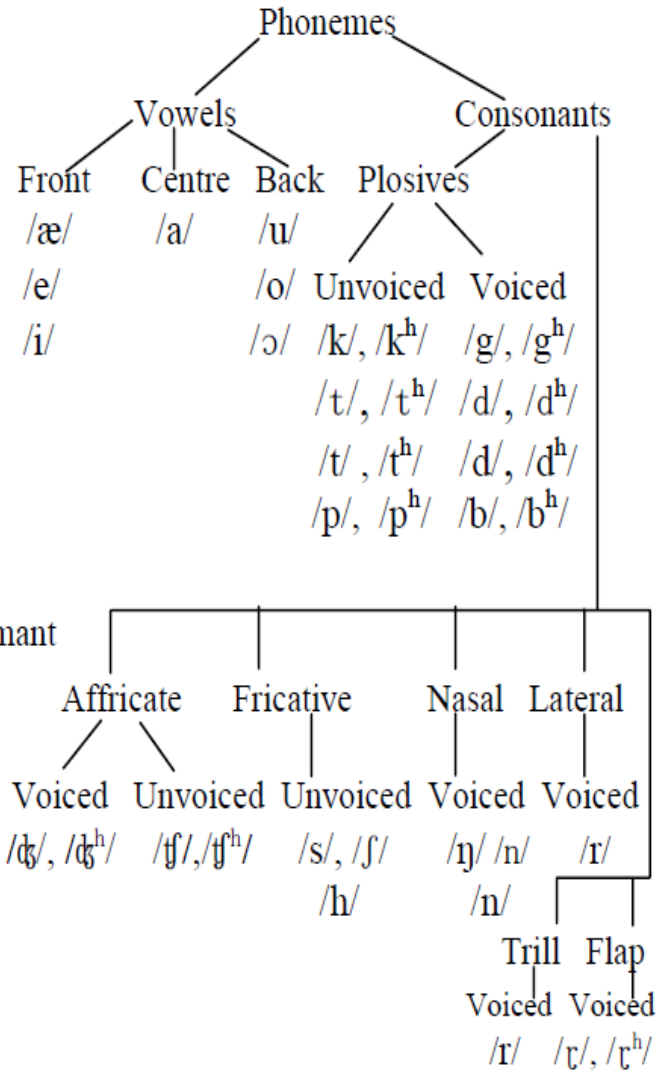
TONGUE HEIGHT	FRONT	BACK
HIGH	1. i	
MID	2. I	.7 u
	3. ɛ	.6 U
LOW	4. ae	.5 a



- tongue hump position (front, mid, back)
- tongue hump height (high, mid, low)
- /IY/, /IH/, /AE/, /EH/ => front => high resonances
- /AA/, /AH/, /AO/ => mid => energy balance
- /UH/, /UW/, /OW/ => back => low frequency resonances



English



Bengal

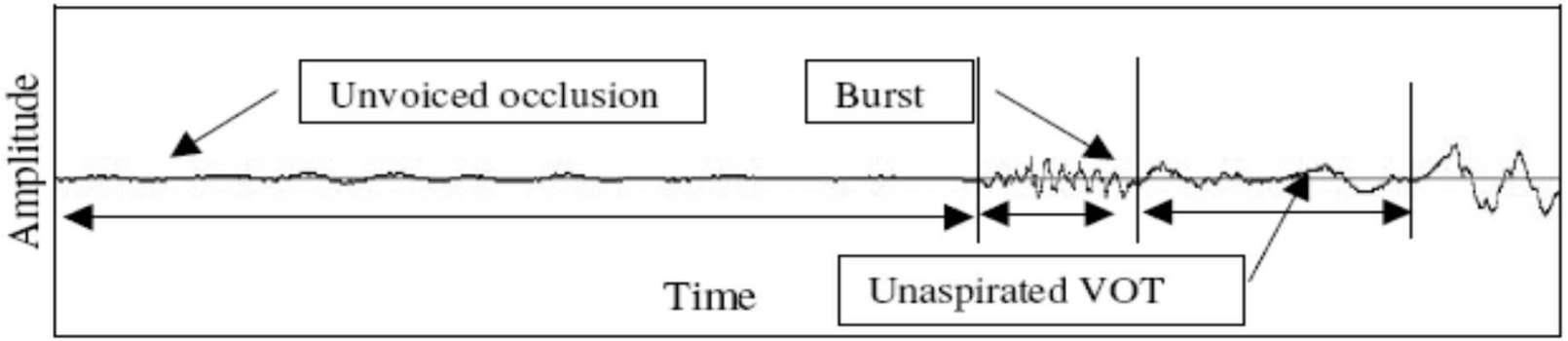


Figure 4.5 Example segment of unaspirated unvoiced stop /k/

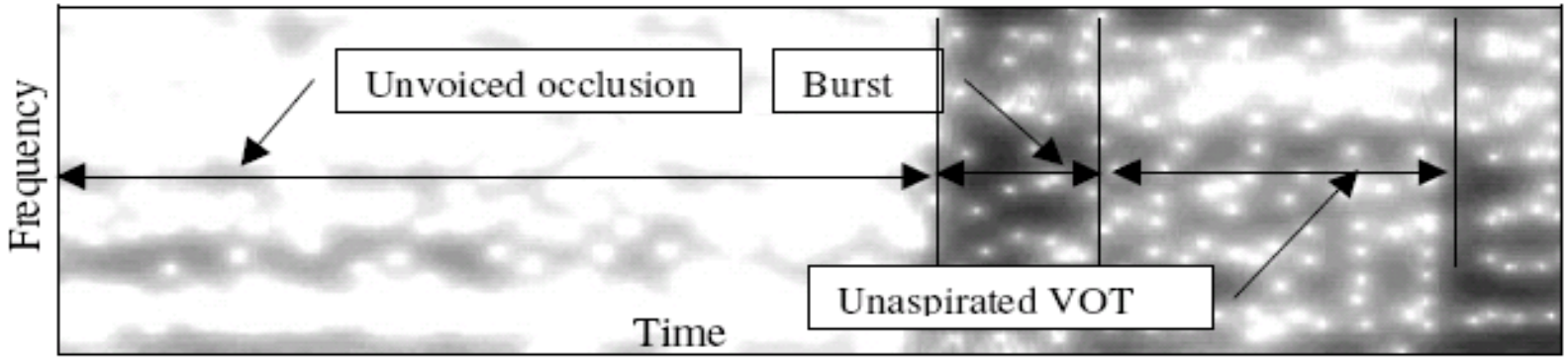


Figure 4.5a Spectrogram of the example unaspirated unvoiced stop /k/

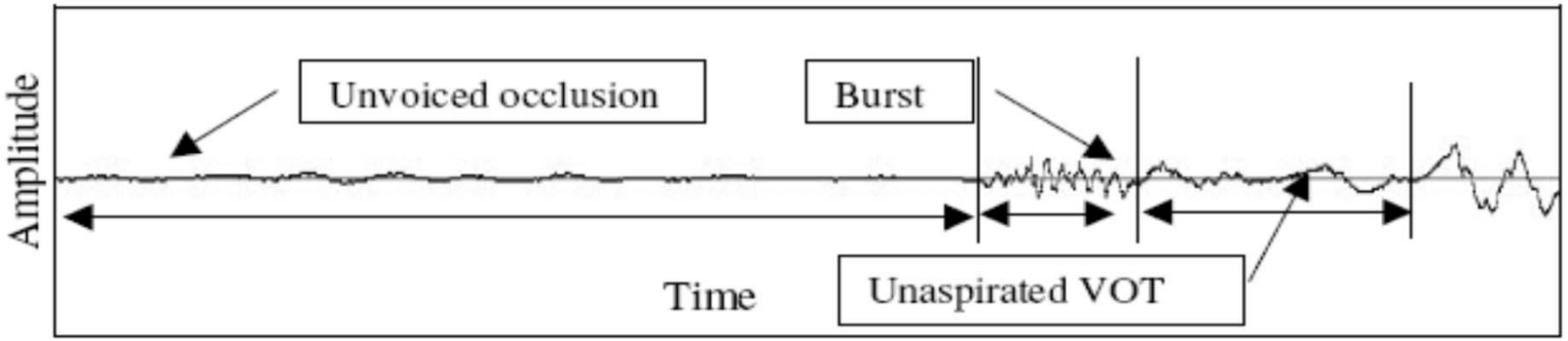


Figure 4.5 Example segment of unaspirated unvoiced stop /k/

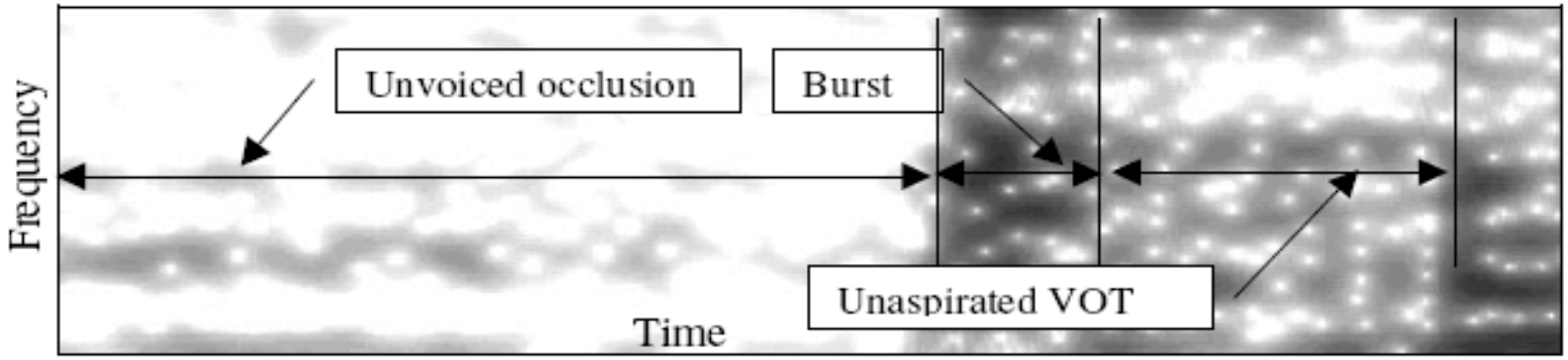


Figure 4.5a Spectrogram of the example unaspirated unvoiced stop /k/

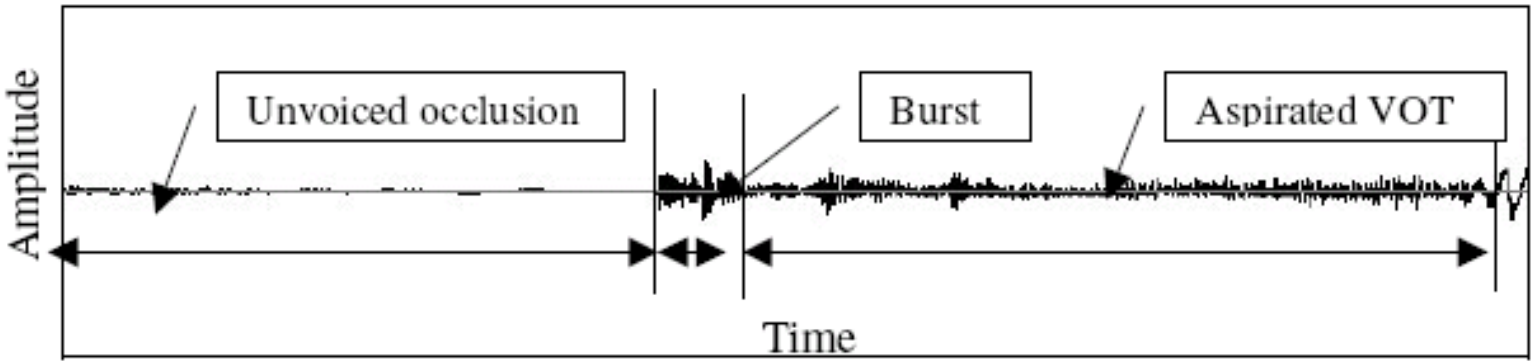


Figure 4.6 Example segment of an aspirated unvoiced stop /k^h/

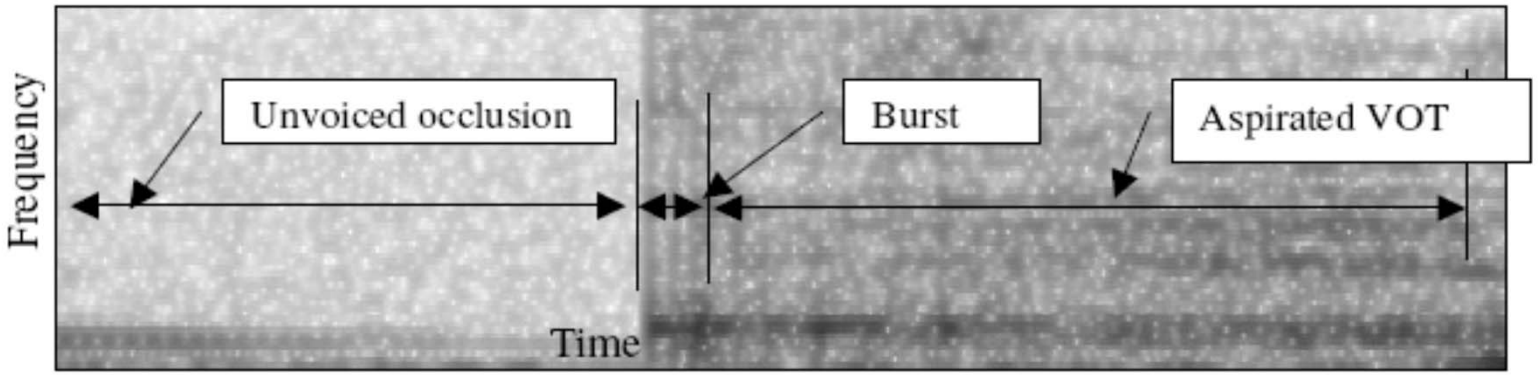


Figure 4.6a Spectrogram example segment of aspirated unvoiced stop /k^h/

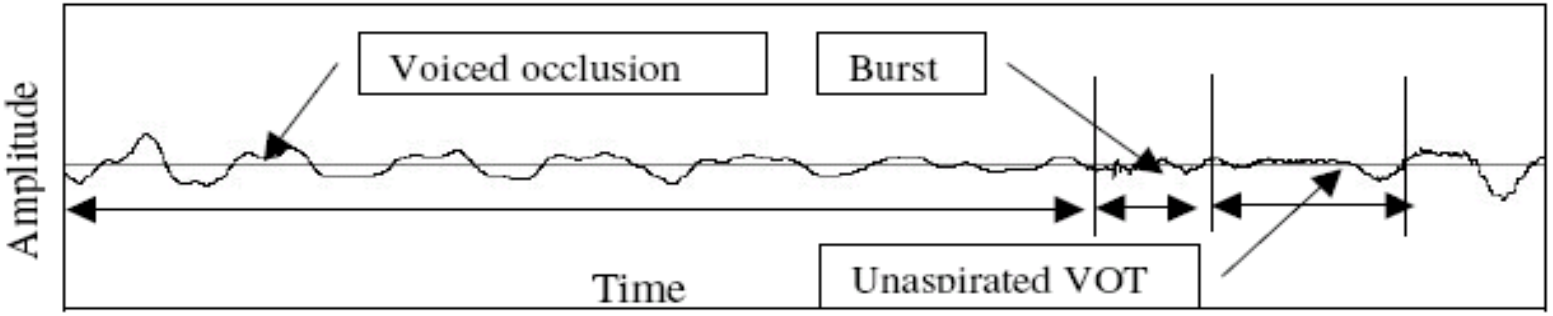


Figure 4.7 Example segment of unaspirated voiced stop /g/

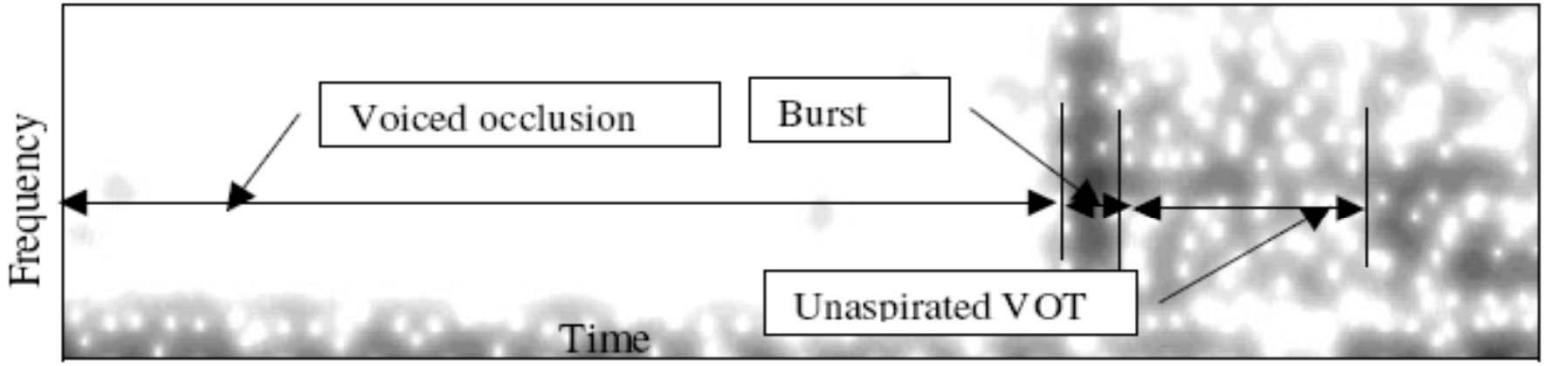


Figure 4.7a Spectrogram segment of unaspirated voiced stop /g/

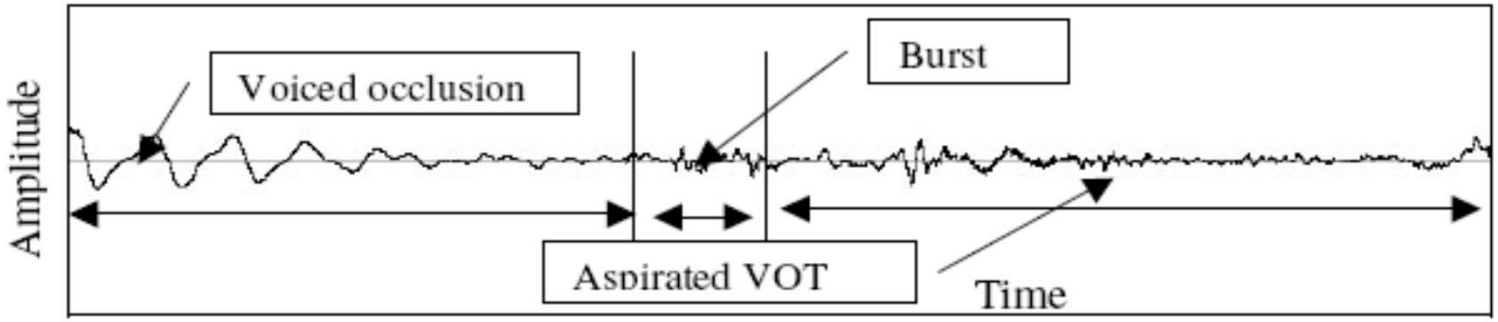


Figure 4.8 Example segment of aspirated voiced stop /g^h/

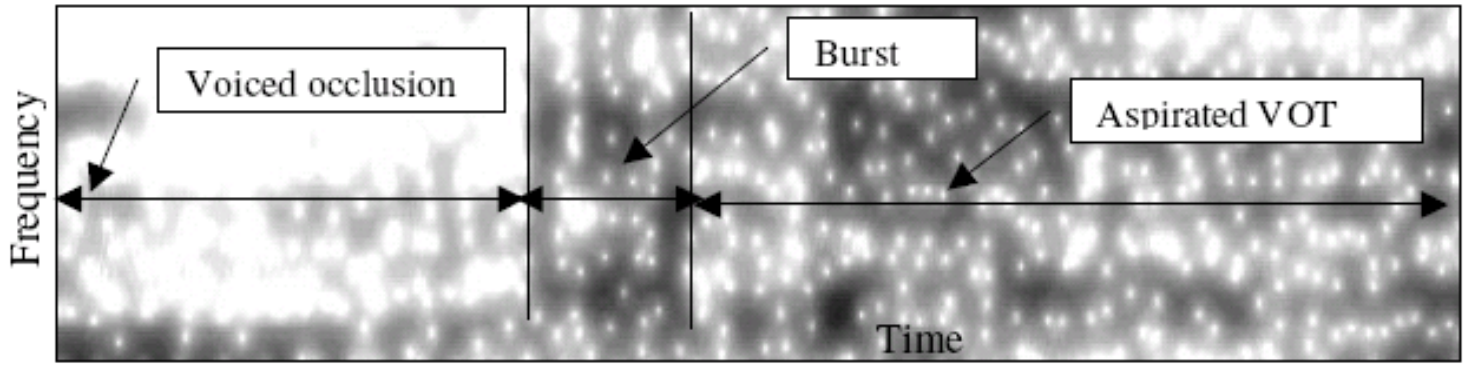


Figure 4.8a Spectrogram example segment of aspirated voiced stop /g^h/

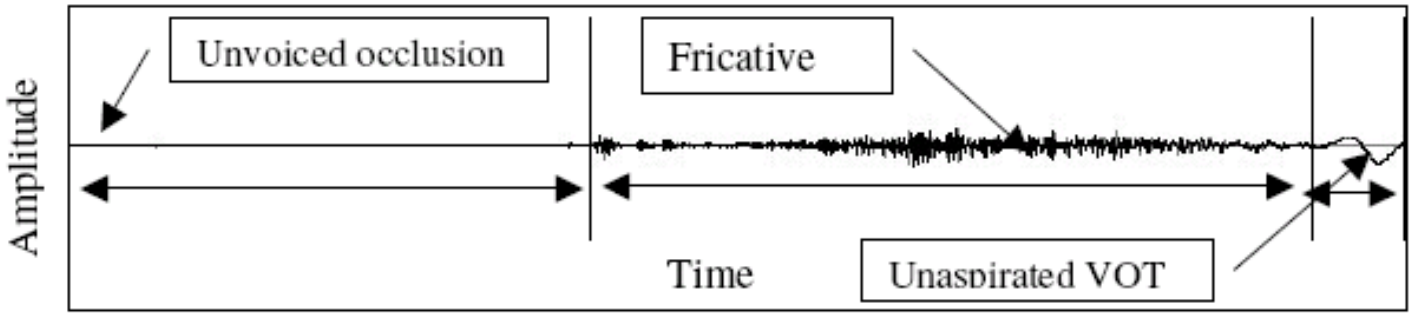


Figure 4.9 Example segment of unaspirated unvoiced affricates / ʧ/

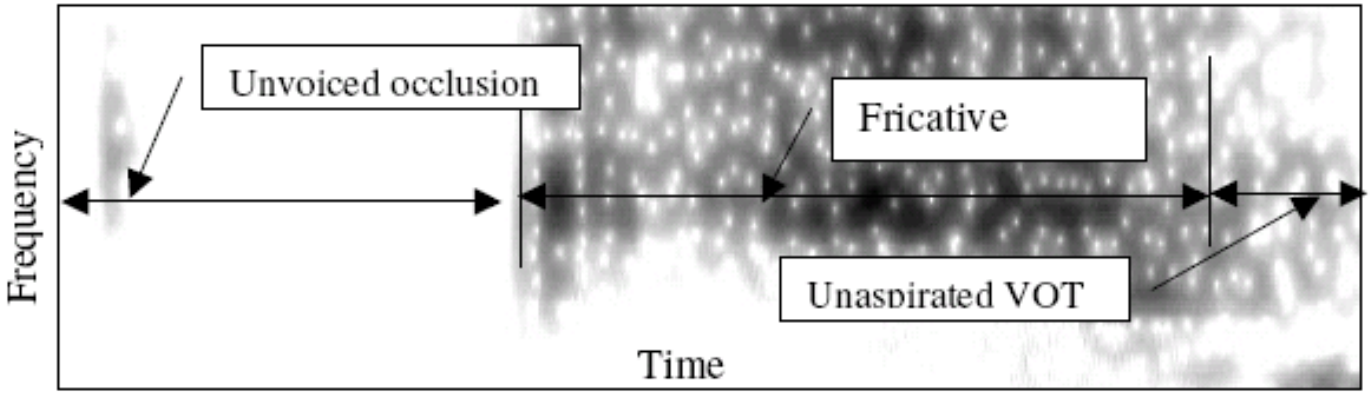


Figure 4.9a Spectrogram of the example segment of unaspirated unvoiced affricates /ʧ/

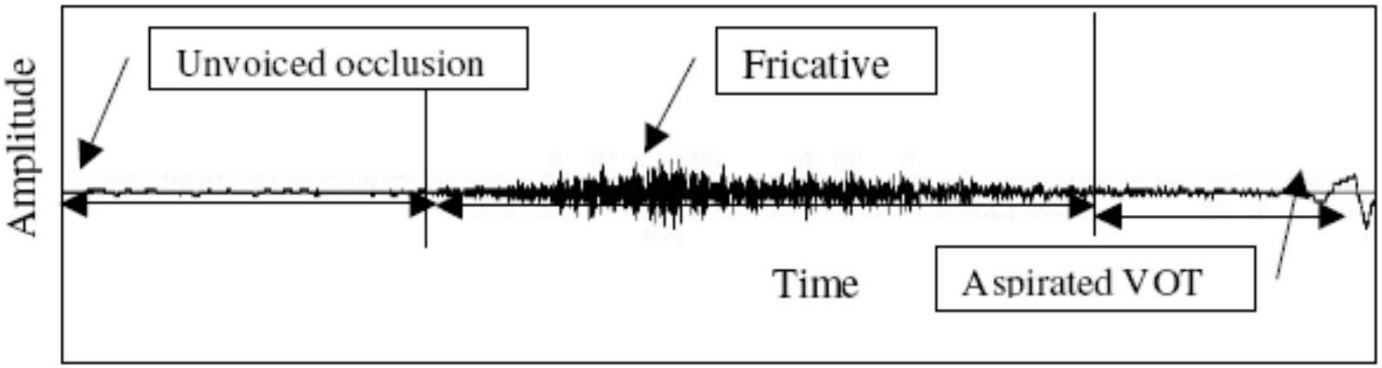


Figure 4.10 Example segment of aspirated unvoiced affricates / tʃʰ/

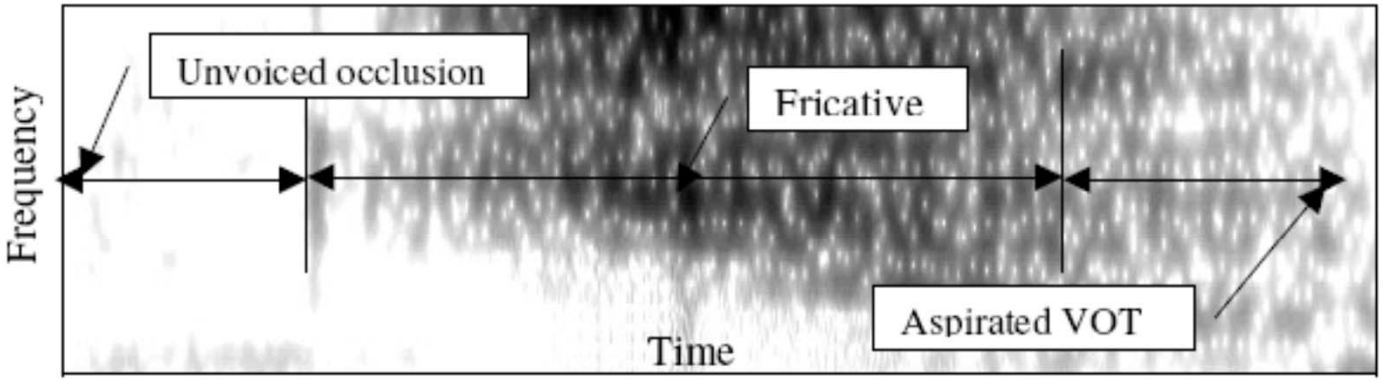


Figure 4.10a Spectrogram of the example segment of aspirated unvoiced affricates /tʃʰ/

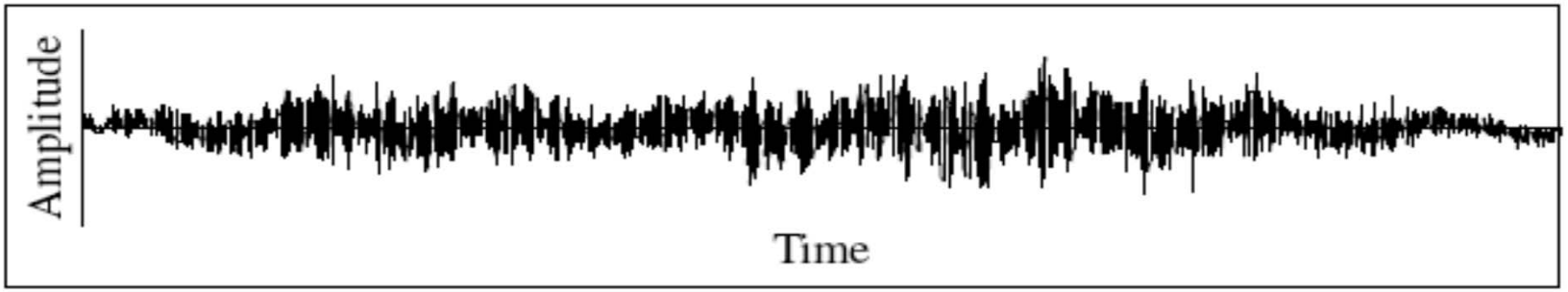


Figure 4.4 An example of sibilant sound segment /s/

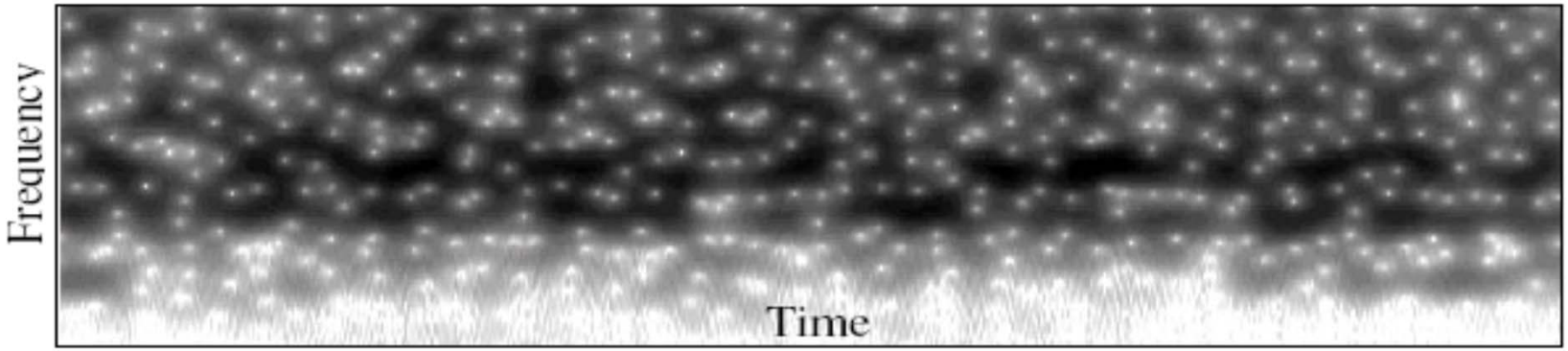


Figure 4.4a Spectrogram of the example sibilant sound segment /s/

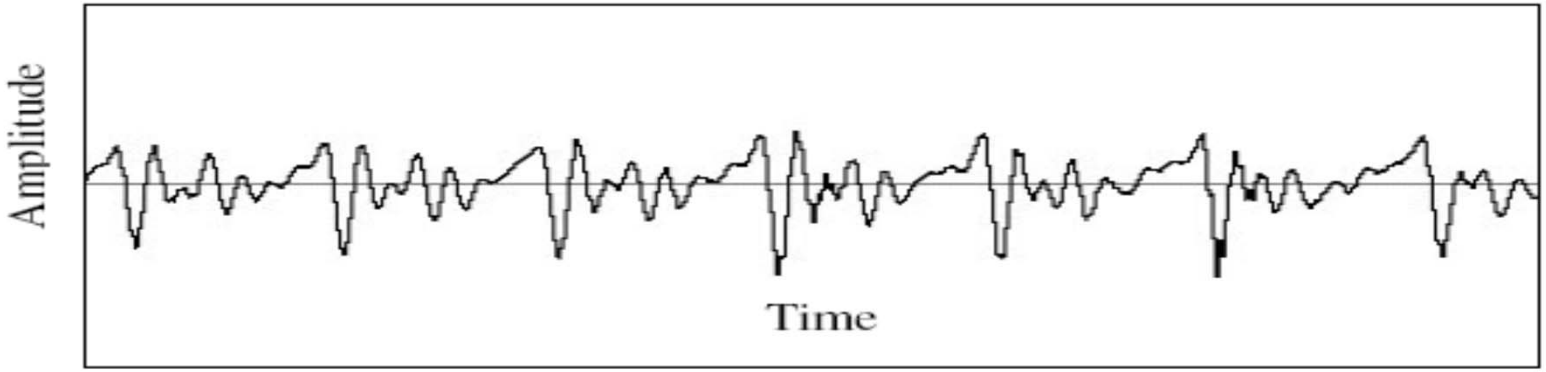


Figure 4.2 Segment of a voiced sound /ɔ/

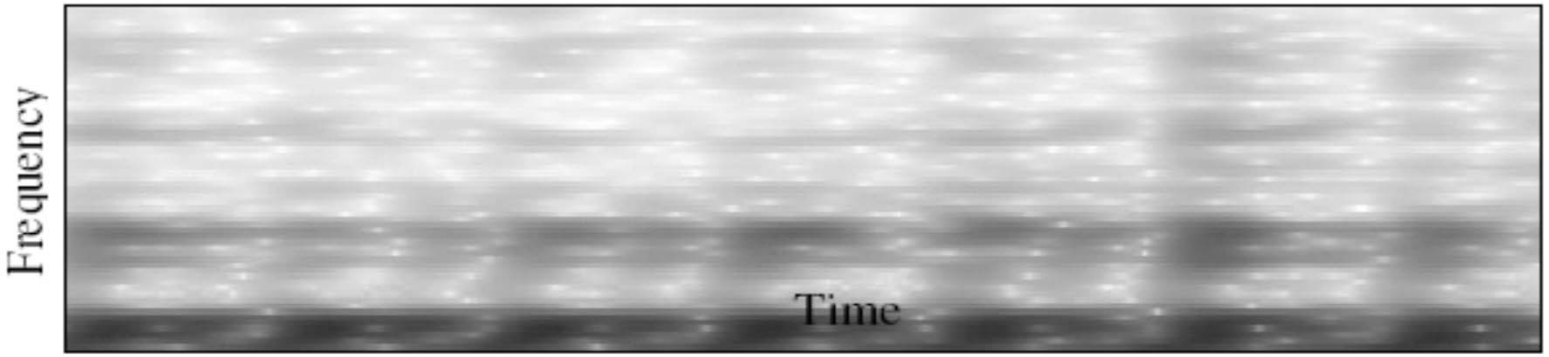
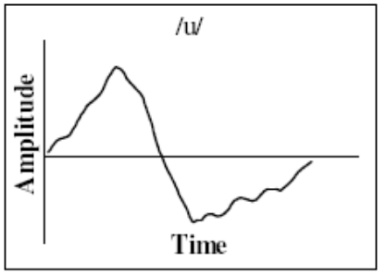
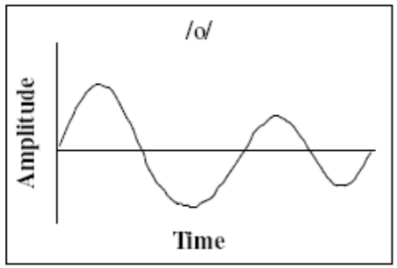
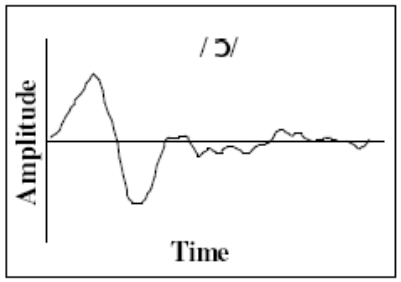


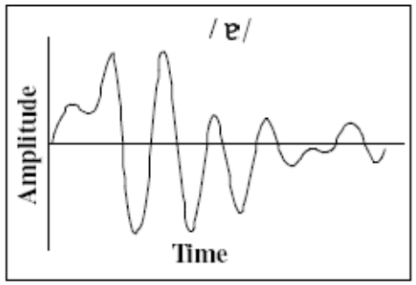
Figure 4.3 Spectrogram of the voice sound /ɔ/

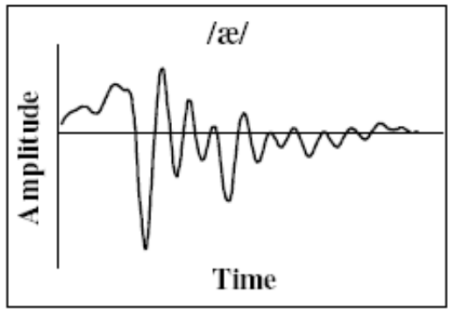
Time Domain Shape

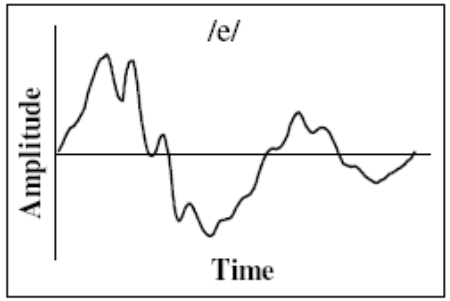


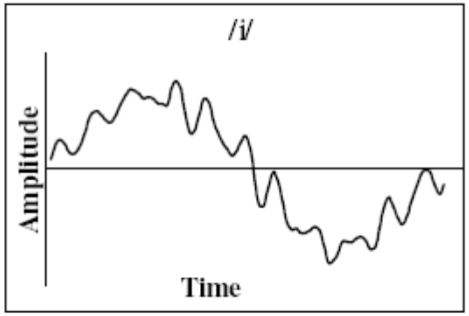










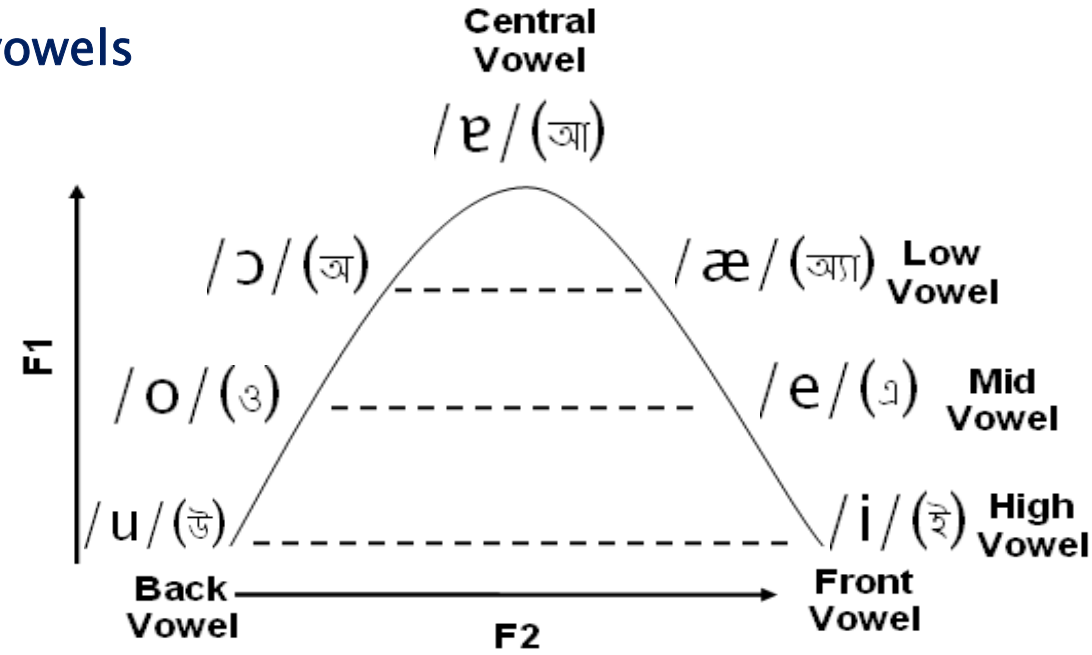


Classification of vowels

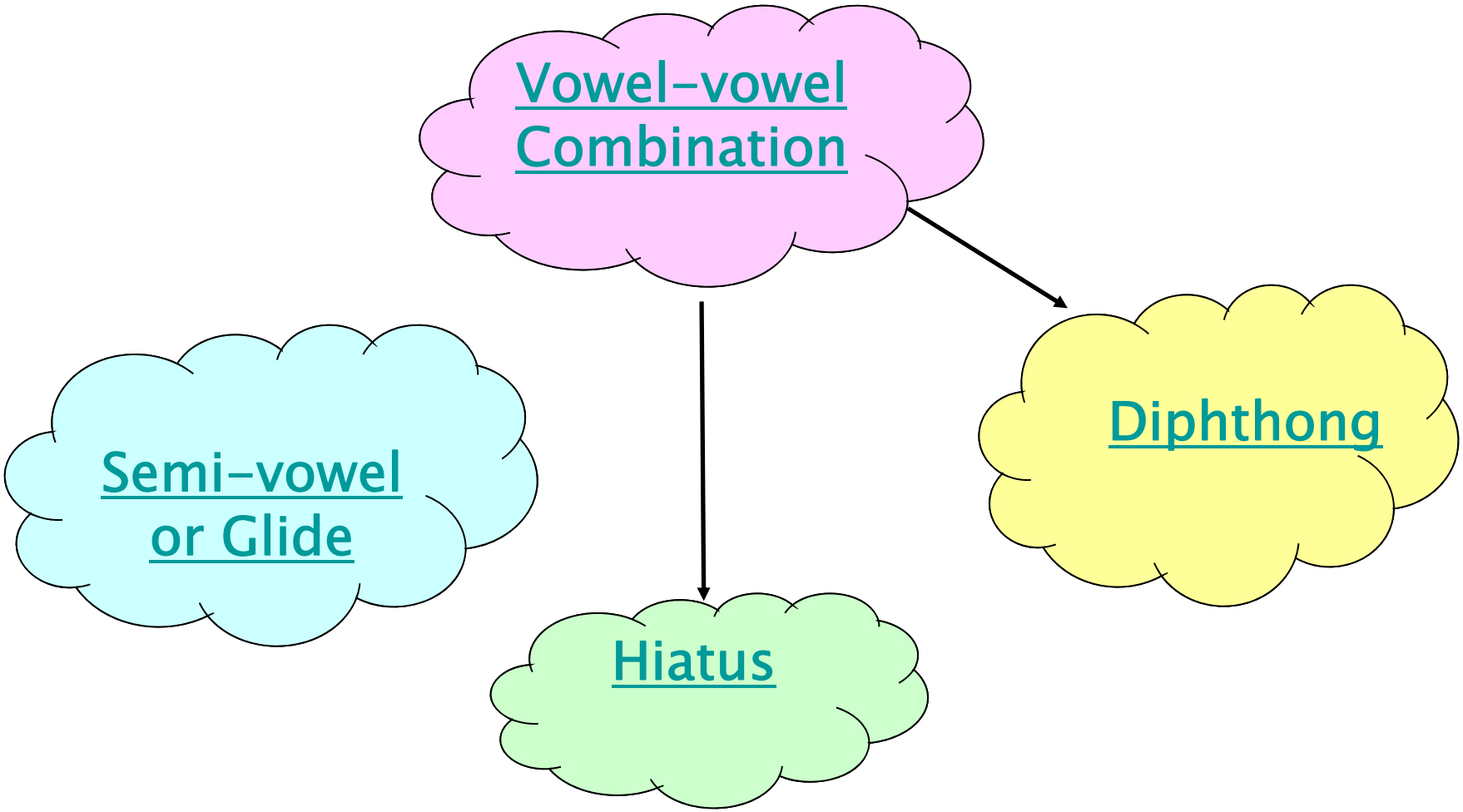
F1 & F2 are primarily determined by the position of tongue. F1 has a higher frequency when the tongue is lowered and F2 has a higher frequency when the tongue is forwarded.

Vowels are classified according to the height and position of the tongue inside the mouth.

Bangla vowels



Position of Bangla Vowels in Cardinal Vowel Diagram


Vowel-vowel Combination
Diphthong
Semi-vowel or Glide
Hiatus

Vowel-Vowel Combination

A. In continuous speech two vowels can come together in two different situations.

B. They may be in a single word.

C. They may be part of two adjacent words i.e., one word ends with a vowel and the next word starts with a vowel.

D. If the two vowels are within a single word, they may either be in two distinct syllables, or may merge into one syllable.

Examples : /bʰulei/ (ভুলেই) /pɐik/ (পাইক)


Diphthong

- A **diphthong** is a monosyllabic vowel combination involving a quick but smooth movement from one vowel to another, often interpreted by listeners as a single vowel sound or phoneme.

- It is a sequence of two different or same vowels that are part of a single syllable. Usually one of the vowels is stronger than the other.

.

- Examples:

Bangla Word :

/ʧɐi/ (চই)

Bangla Word :

/bʰulei/ (ভুলেই)

Hiatus

➢When two vowels coming together without any contraction or elision are pronounced separately as distinct from Diphthongs they are termed as **hiatus**.

➢Hiatus may be of two types:

1) **Internal Hiatus→** which occurs within a word.

Example: Bangla Word :

/pɐik/ (পাইক)

2) **External Hiatus** → which refers to the break between two successive words. In this situation the first word ends with a vowel and the second word starts with a vowel.

Example:

Bangla Sentence :

/ɐmi iliʃ kʰɐbo/ (আমি ইলিশ খাব)

Semi-vowel or Glide

➢**Semi-vowel** refers to a sound functioning as a consonant but lacking the PHONETIC characteristics normally associated with consonants.

➢Its QUALITY is phonetically that of a vowel; though its DURATION is much less than that typical of vowel.

➢ Examples:

Bangla Word :

/ɔjon/ (অয়ন)

Bangla Word :

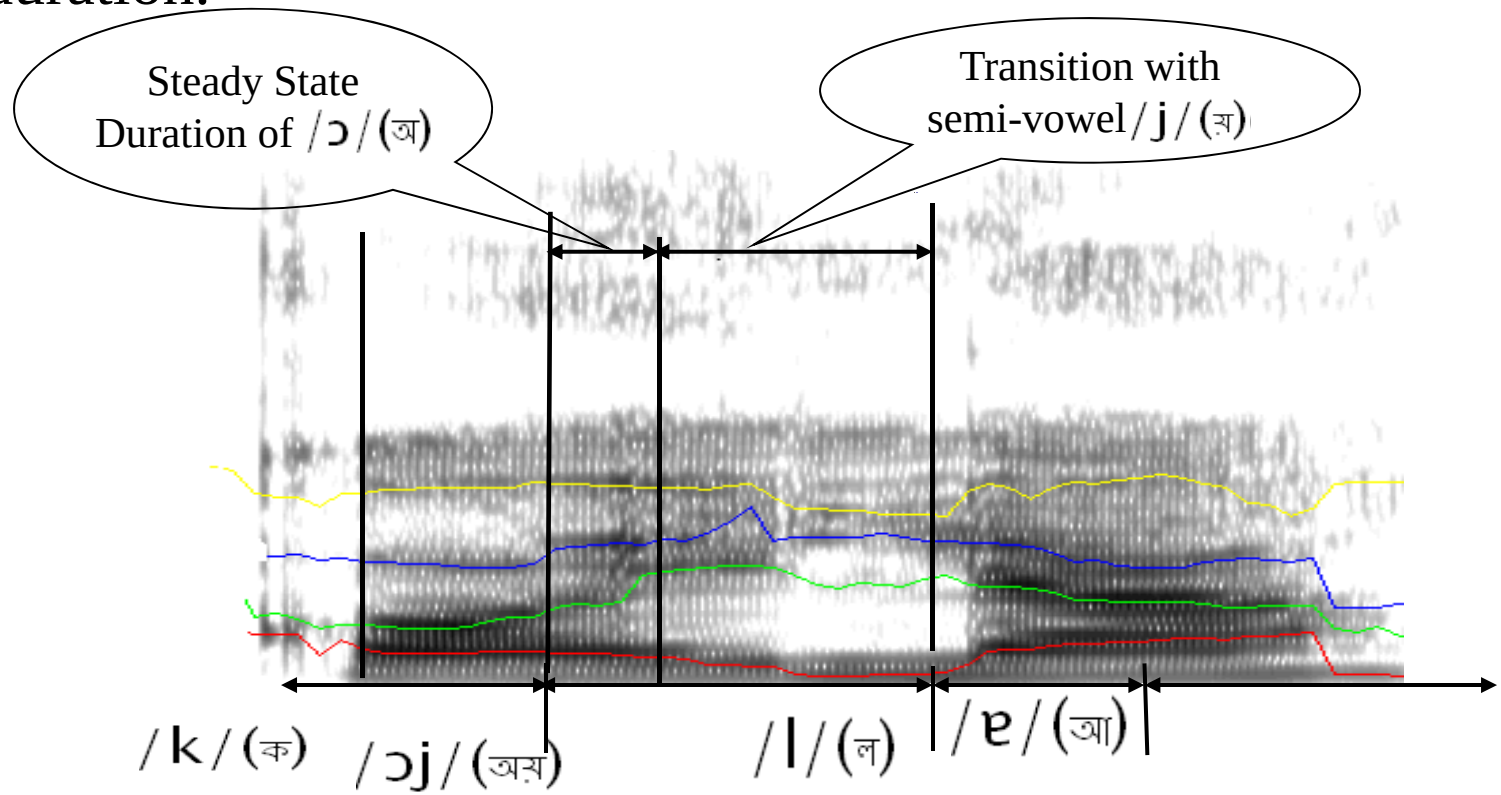
/meje/ (মেয়ে)

Bangla Word :

/hɐvɐ/ (হাওয়া)

Semi-vowel after a vowel

Vowel-semivowel combination (V-j) consists of transitional duration with semivowel along with the preceding vowel's steady state duration.

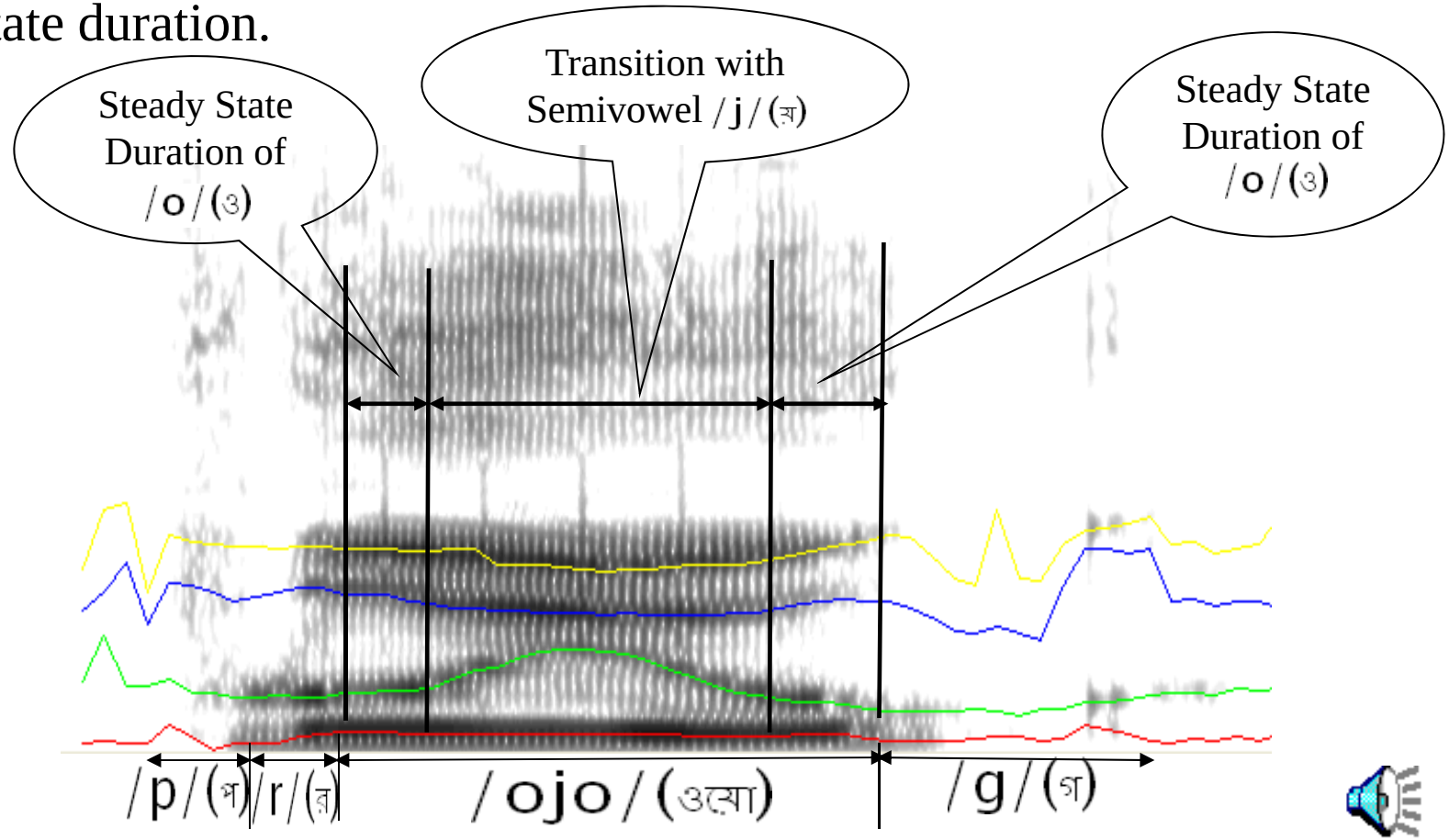


Spectrographic View of Bangla Word /kɔjlɐ/ (কয়লা)
with V–j combination /ɔj/ (অয়)

Play

Semi-vowel in between two vowels

➢Vowel-semivowel-vowel combination consists of transitional duration with semivowel along with the preceding and succeeding vowels' steady state duration.



Spectrographic View of Bangla Word (/projog/(প্রয়োগ)) with V-j-V combination

Play /ojo/(ওয়ো)

English Speech Sounds

A Condensed List of Phonetic Symbols for American English

Phoneme	ARPAbet	Example	Phoneme	ARAPAbet	Example
/i/	IY	beat	/ŋ/	NX	sing
/ɪ/	IH	bit	/p/	P	pet
/e/ (eʸ)	EY	bait	/t/	T	ten
/ɛ/	EH	bet	/k/	K	kit
/æ/	AE	bat	/b/	B	bet
/ɑ/	AA	Bob	/d/	D	debt
/ʌ/	AH	but	/g/	G	get
/ɔ/	AO	bought	/h/	HH	hat
/o/ (oʷ)	OW	boat	/f/	F	fat
/ʊ/	UH	book	/θ/	TH	thing
/u/	UW	boot	/s/	S	sat
/ə/	AX	about	/š/	SH	shut
/ɨ/	IX	roses	/v/	V	vat
/ɝ/	ER	bird	/ð/	DH	that
/ɚ/	AXR	butter	/z/	Z	zoo
/ɑʷ/	AW	down	/ž/	ZH	azure
/ɑʸ/	AY	buy	/č/	CH	church
/ɔʸ/	OY	boy	/ǰ/	JH	judge
/y/	Y	you	/ʍ/	WH	which
/w/	W	wit	/ l̥ /	EL	battle
/r/	R	rent	/ m̥ /	EM	bottom
/l/	L	let	/ n̥ /	EN	button
/m/	M	met	/T/	DX	batter
/n/	N	net	/ʔ/	Q	(glottal stop)

ARPABET representation

- ***48 sounds***
  - **18 vowels/diphthongs**
  - **4 vowel-like consonants**
  - **21 standard consonants**
  - **4 syllabic sounds**
  - **1 glottal stop**

Consonents		Manner of Articulation				
S/N	Place of Articulation		Unvoiced		Voiced	
			Un-Aspirated	Aspirated	Un-Aspirated	Aspirated
1	Velar	Stop	/k/	/kʰ/	/g/	/gʰ/
2	Post-alveolar (Retroflex)		/ʈ/	/ʈʰ/	/ɖ/	/ɖʰ/
3	Dental		/t/	/tʰ/	/d/	/dʰ/
4	Bilabial		/p/	/pʰ/	/b/	/bʰ/
5	Alveolar -Post alveolar	Affricate	/ʧ/	/ʧʰ/	/ʤ/	/ʤʰ/
6	Alveolar	Fricative	/s/			
7	Post alveolar		/ʃ/			
8	Glottal		/h/		//	
9	Velar	Nasal Murmur			/ŋ/	
10	Palatal				/ɳ/	
11	Dental				/n/	
12	Bilabial				/m/	

S/N	Place of Articulation	Manner of Articulation				
			Unvoiced		Voiced	
			Un-Aspirated	Aspirated	Un-Aspirated	Aspirated
13	Dental	Lateral			/l/	
14	Alveolar	Trill			/r/	
15	Post alveolar	Retroflex Flap			/ɽ/	/ɽh/
16	Palatal	Approximant			/j/	
17	Bilabial				/w/	
Vowel						
1	Back vowel	Close, Rounded			/u/	
2	Back vowel	Close-mid, Rounded			/o/	
3	Back vowel	Open, Rounded			/ɔ/	
4	Front vowel	Open, Unrounded			/a/	
5	Front vowel	Open-mid, Unrounded			/æ/	
6	Front vowel	Close-mid, Unrounded			/e/	
7	Front vowel	Close, Unrounded			/i/	

TUTORIAL

1.Write the place and manner of articulation of the following phoneme

/k/, /g/,/u/,/g^h/, /ɽ/, /ʃ/

2. Write out the phonetic transcription for the following words:

/she/, /phonetic/, /marks/, /speech/,

How many syllable is present in each of the above word.

3. Draw Schematic representation of the physiological mechanism of speech production system and explain how the a voiced sound is produce.

4. A voiced operated lift operation is designee using the following words

a. stop, b. up, c. down d. floor e. first f. second g. third h. fourth and i. ground.

Figure 1 shows wideband spectrograms of one version of each of these words. Using your knowledge of acoustic phonetics, determine which wideband spectrogram corresponds to which word.

5. The following waveform is for the utterance /kolkata/ and the waveform samples are at a sampling rate of FS =22050 Hz. Segment the waveform into regions of "Voiced Speech (V)" and "Non-Voiced Speech (N)".

6. Which formant frequency is related to tongue height and which formant related to tougue position

7. Why the child speech has high F0 and formant compare to a adult