Audio & Speech Technology

[6] Speech Synthesis & Recognition

Phoneme

Phoneme = Smallest segment of speech sound

Vowels | Consonants

EXAMPLE

IPA	Example
i:	see,heat
a:	arm,father
u:	blue,foot
ɔ:	call,horse
ɜ:	turn,learn
ɪ	ship,hit
æ	hat,black
ʊ	put,could
ʌ	cup,duck

IPA	Example	IPA	Example
b	book	l	leg,little
d	day,did	z	zoo,lazy
f	find,if	ʒ	pleasure
g	give,flag	ʃ	she,crash
h	how	tʃ	check
j	yes,yellow	n	no,ten
k	cat,back	ŋ	sing

IPA = International Phonetic Alphabet

Thai Phonemes

SHORT VOWELS	
IPA	ตัวอักษร
a	◌ะ, ◌ั◌
e	เ◌ะ, เ◌็◌
ɛ	แ◌ะ, แ◌็◌
i	◌ิ, ◌ิ◌
o	โ◌ะ, ◌◌
ɔ	เ◌าะ, ◌็อ◌
u	◌ุ, ◌ุ◌
ɯ	◌ึ, ◌ึ◌
ɤ	เ◌อะ

(เสียงสระสั้น)

LONG VOWELS	
IPA	ตัวอักษร
a:	◌า, ◌า◌
e:	เ◌, เ◌◌
ɛ:	แ◌, แ◌◌
i:	◌ี, ◌ี◌
o:	โ◌, โ◌◌
ɔ:	◌อ, ◌อ◌
u:	◌ู, ◌ู,
ɯ:	◌ือ, ◌ื◌
ɤ:	เ◌อ, เ◌ิ◌

(เสียงสระยาว)

DIPHTHONGS	
IPA	ตัวอักษร
iaʔ, iəʔ	เ◌ียะ
ia, iə	เ◌ีย, เ◌ีย◌
uaʔ, uəʔ	◌ัวะ
ua, uə	◌ัว, ◌ว◌
ɯaʔ, ɯəʔ	เ◌ือะ
ɯa, ɯə	เ◌ือ, เ◌ือ◌

(เสียงสระควบ)

See more at http://en.wikipedia.org/wiki/International_Phonetic_Alphabet_chart_for_English_dialects

Thai Phonemes

CONSONANTS			
IPA	ตัวอักษร	IPA	ตัวอักษร
b	บ	p	ป
d	ฎ,ด	p^h	ผ,พ,ภ
f	ฝ,ฟ	r	ร,หร
h	ห,ฮ	s	ซ,ศ,ษ,ส
j	ญ,ย,อย,หย	t	ฏ,ต
k	ก	t^h	ฐ,ฑ,ฒ,ถ,ท,ธ
k^h (/x/)	ข,ฃ,ค,ฅ,ฆ	tɕ	จ
l	ล,ฬ,หล	tɕ^h	ฉ, ช, ฌ
m	ม,หม	w	ว,หว
n	ณ,น,หน	ʔ	อ, ะ
ŋ	ง,หง		

TONES	
IPA	วรรณยุกต์
a	เสียงสามัญ
à	เสียงเอก
â	เสียงโท
á	เสียงตรี
ǎ	เสียงจัตวา

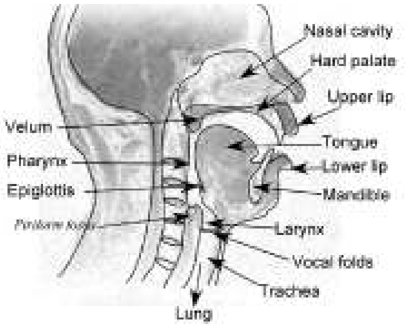
Ex

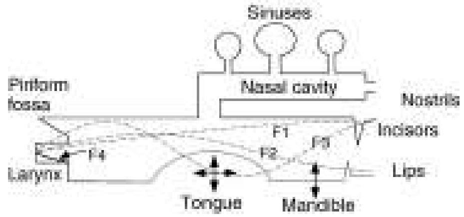
เงินมาก

⬇

ŋɤ̂:bma:k

Speech Production





Voiced Speech

Unvoiced Speech

Excitation Signal → Vocal Tract Filter → Lips Radiation Filter → Speech Signal

http://ars.els-cdn.com/content/image/1-s2.0-S016763930700177X-gr1.jpg

Voiced Speech

- Vocal cords vibrate
- Excitation is periodic signal
- Characterized by Low Freq.

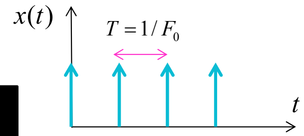
Time Domain

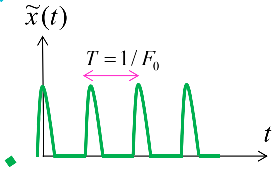
$$\tilde{x}(t) = A \cdot g(t) * x(t)$$

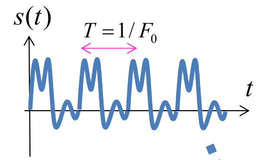
$$s(t) = l(t) * v(t) * \tilde{x}(t)$$

Impulse Train Generator → $x(t)$ → Glottal Pulse Model G(f)

Gain A → × → $\tilde{x}(t)$ → Vocal Tract Filter V(f) → Lips Radiation Filter L(f) → Voiced Speech $s(t)$







Voiced Speech

- Vocal cords vibrate
- Excitation is periodic signal
- Characterized by Low Freq.

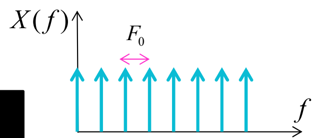
Frequency Domain

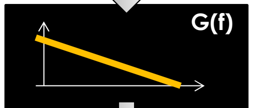
$$\tilde{X}(f) = AG(f)X(f)$$

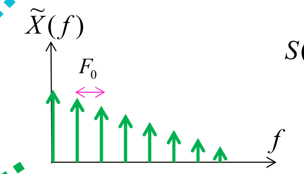
$$S(f) = L(f)V(f)\tilde{X}(f)$$

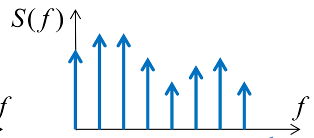
Impulse Train Generator → $x(t)$ → G(f)

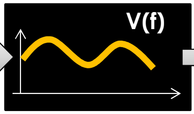
Gain A → × → $\tilde{x}(t)$ → V(f) → L(f) → Voiced Speech $s(t)$

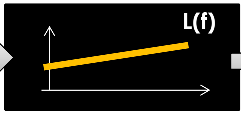






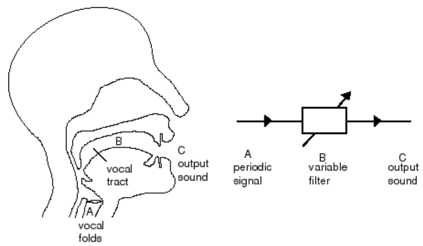


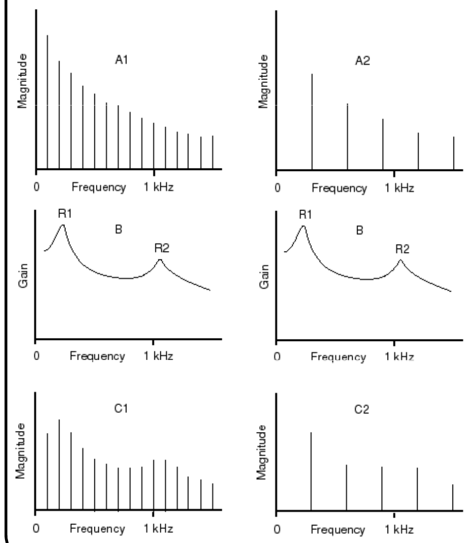




Voiced Speech

- Different pitches
- Same phoneme





Fundamental Frequency

- Male : 50-200 Hz
- Female : 150-300 Hz
- Child : 200-400 Hz

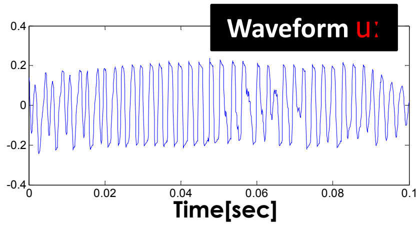
http://www.animations.physics.unsw.edu.au/jw/SPEECH_HELIUM/sourcefilter.gif

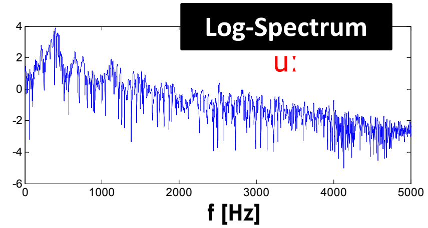
Voiced Speech

- All vowel in English = Voiced speech.
- Voiced consonants

IPA	Example
b	book
d	day,did
g	give,flag
v	vanilla
z	zoo,lazy
ð	then
ʒ	pleasure
dʒ	jump,gin

IPA	Example
l	leg,little
m	man
n	no,ten
ŋ	sing
r	red,try
w	window
j	yes,yellow



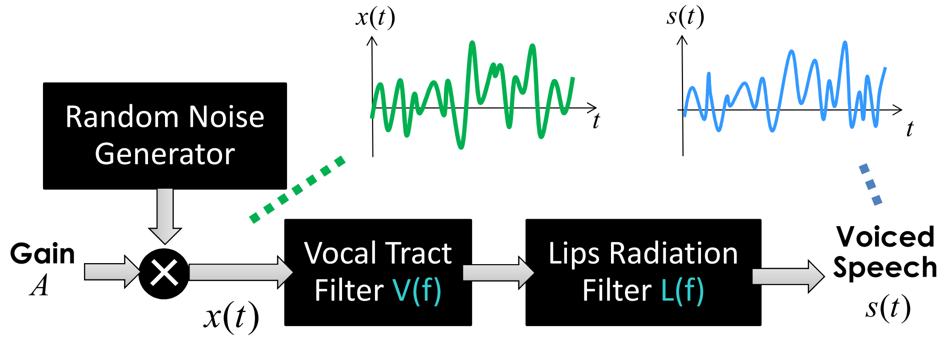


Unvoiced Speech

- Vocal cords dose not vibrate
- Excitation is random noise
- Characterized by High Freq.

Time Domain

$$s(t) = l(t) * v(t) * x(t)$$

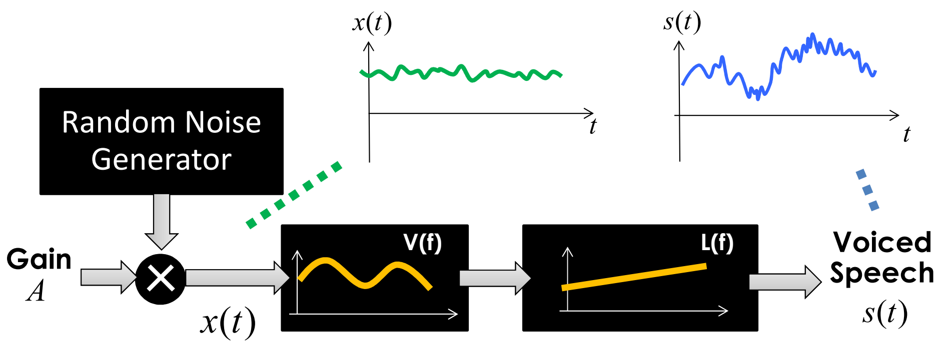


Unvoiced Speech

- Vocal cords dose not vibrate
- Excitation is random noise
- Characterized by High Freq.

Frequency Domain

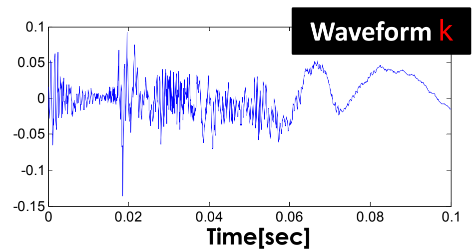
$$S(f) = L(f)V(f)X(f)$$

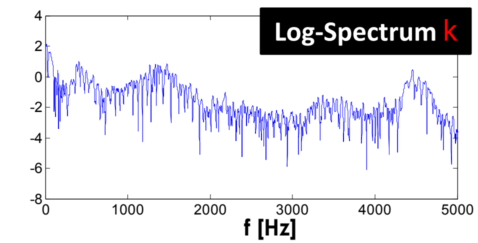


Unvoiced Speech

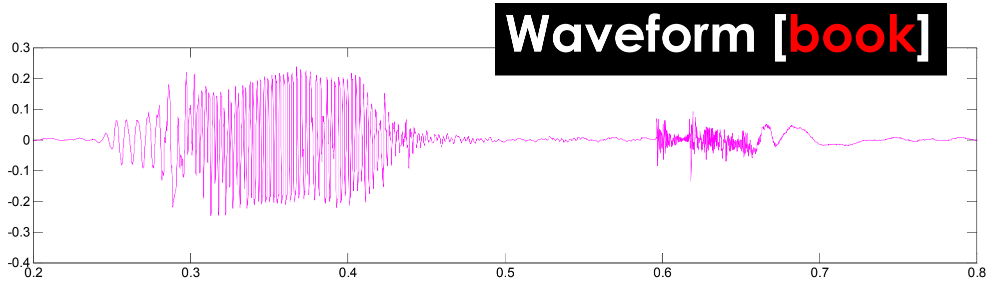
- Unvoiced consonants

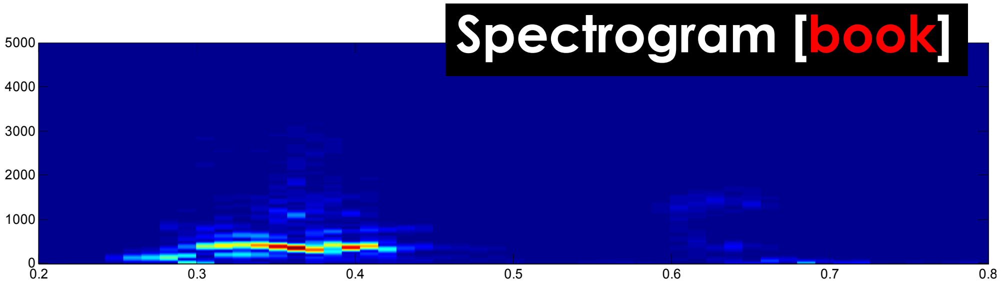
IPA	Example
p	pet
f	find,if
θ	thirty,both
t	ten
s	sir
ʃ	she,crash
tʃ	check
k	king,cat

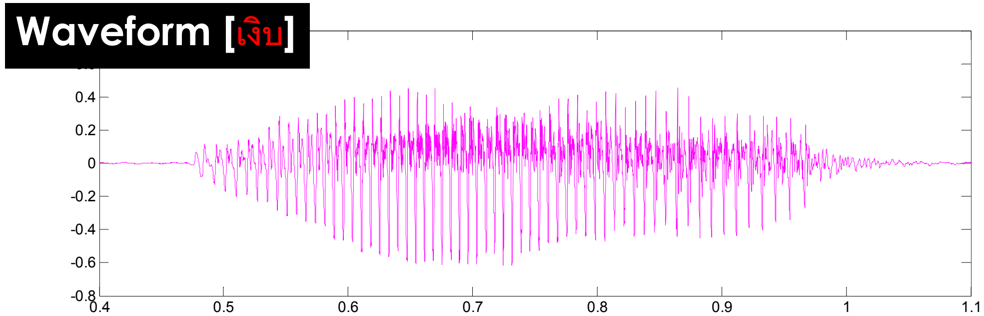


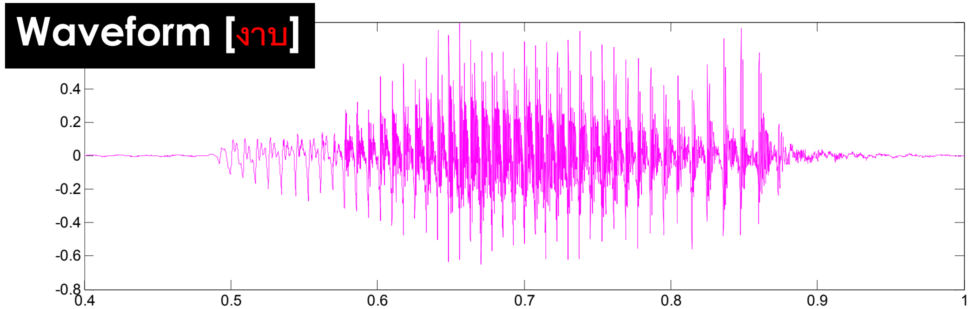


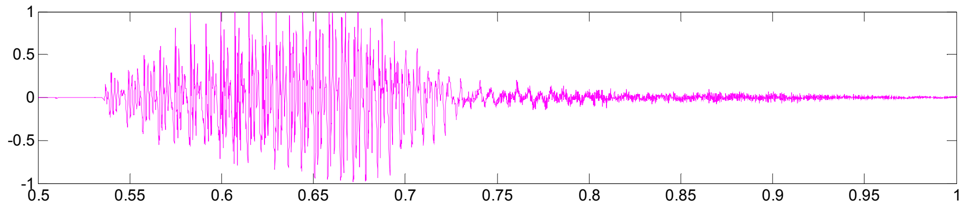
Voiced VS Unvoiced

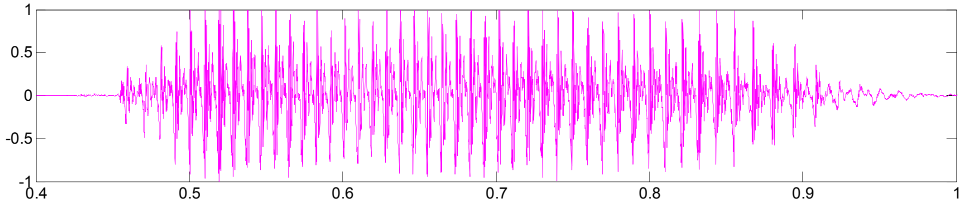


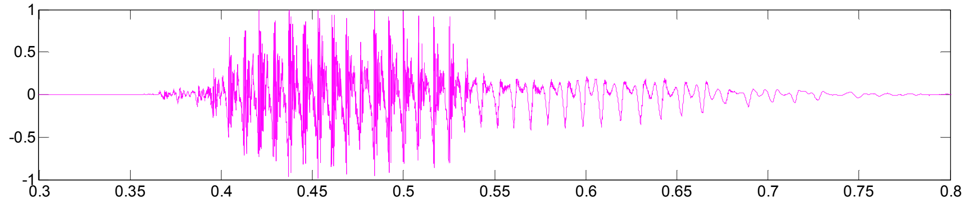


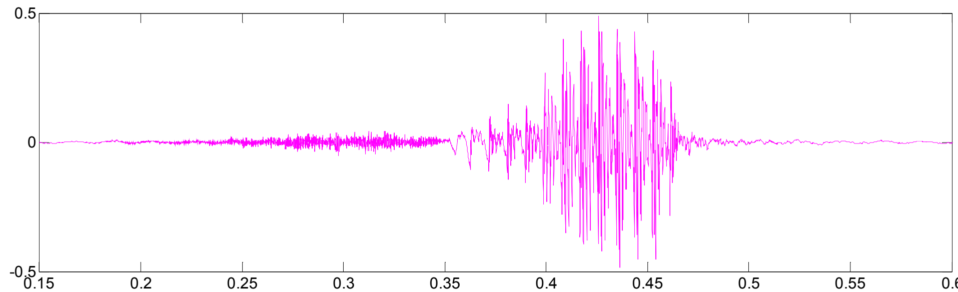


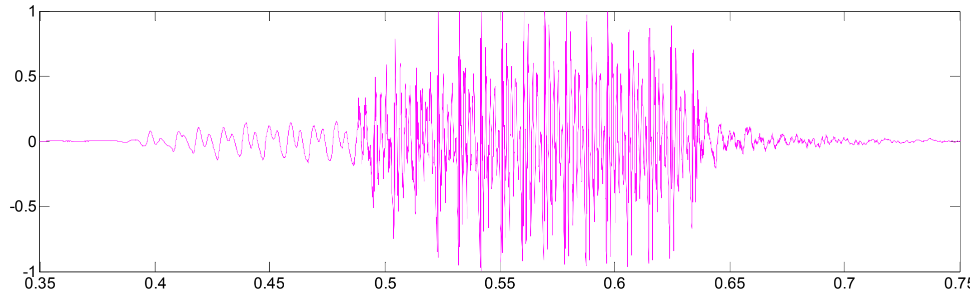


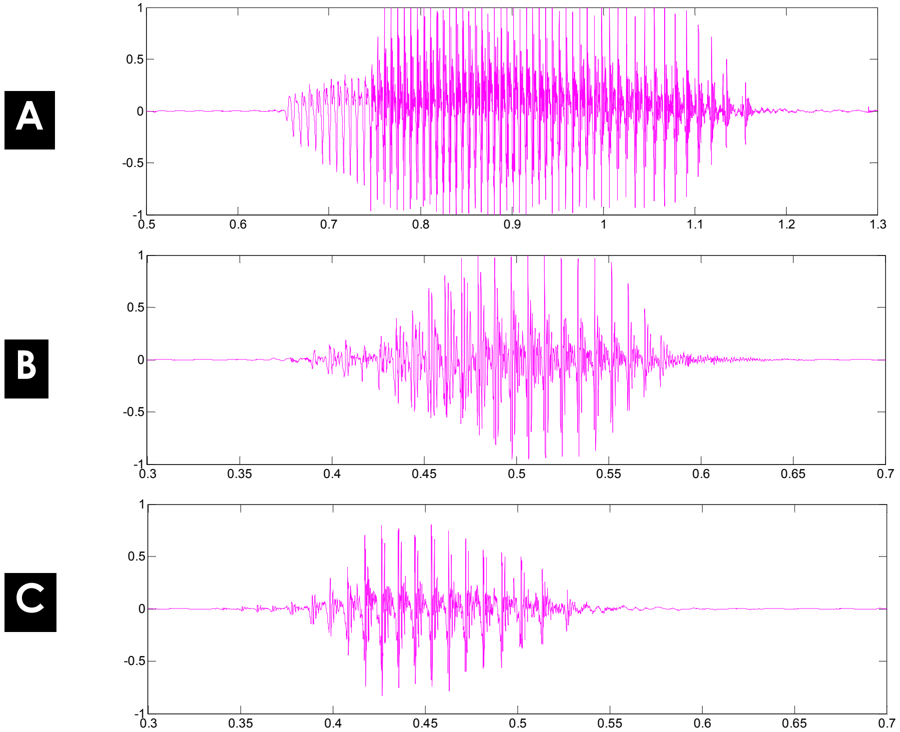






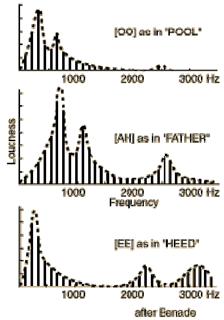






Formant

- **Vocal tract** = Resonance Tube
- **Formant** = Resonant Frequency of Vocal Tract
- Vocal tract has a fixed characteristic in the order of 10 ms
- 3-4 formants present below 4kHz of speech

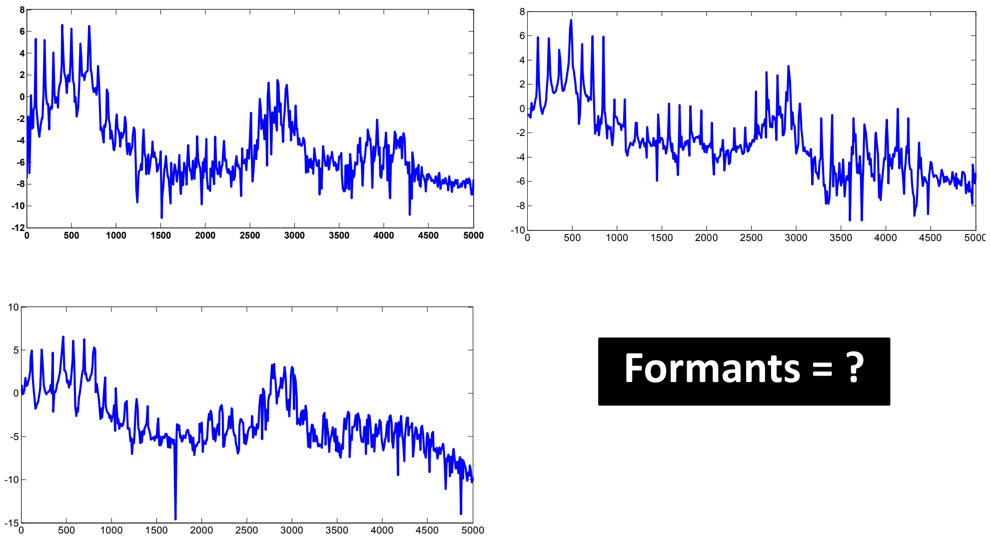


Phonetic Symbol	Example Word	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)
/ow/	bought	570	840	2410
/oo/	boot	300	870	2240
/u/	foot	440	1020	2240
/a/	hot	730	1090	2440
/uh/	but	520	1190	2390
/er/	bird	490	1350	1690
/ae/	bat	660	1720	2410
/e/	bet	530	1840	2480
/i/	bit	390	1990	2550
/iy/	beet	270	2290	3010

http://hyperphysics.phy-astr.gsu.edu/hbase/music/imgmus/vow5.gif
http://cnx.org/content/m15459/latest/sub_formants-voweltable.png

Formant

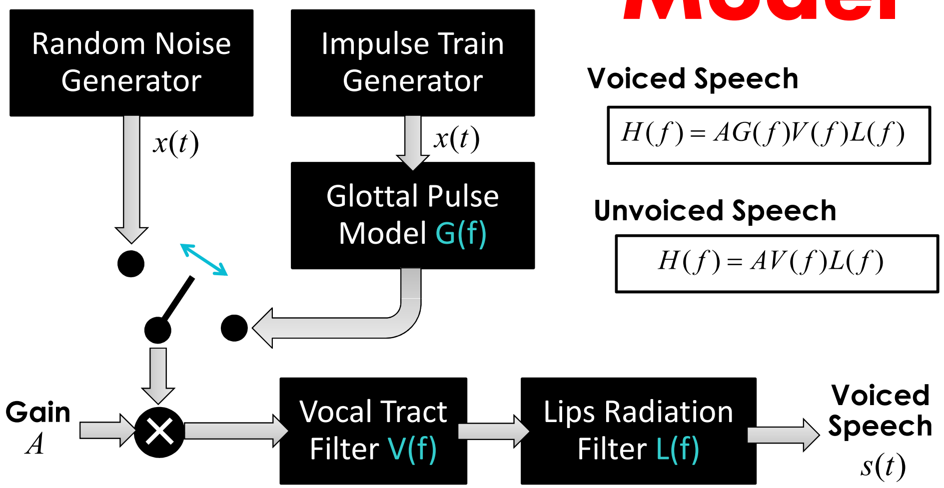
โอวววว์



Formants = ?

Speech Production Model

$$S(f) = H(f)X(f)$$



Voiced Speech

$$H(f) = AG(f)V(f)L(f)$$

Unvoiced Speech

$$H(f) = AV(f)L(f)$$

Speech Production Model

Frequency Domain

$$S(f) = H(f)X(f)$$

Time Domain

$$s[n] = \sum_{k=1}^{P} a_i s[n-k] + K \cdot x[n]$$

$K = Volumn\ Control$

$a_i = Filter\ Coefficients$

Depend on phonemes and speakers
Can be estimated by using LPC

$P \in \{10, 12, 14\}$

$s[n] = Speech\ Signal$

$x[n] = Excitation$

Impulse Train for voiced speech [Varying Pitch]

Random noise for unvoiced Speech

Linear Predictive Coding

Current sample — Previous samples

$$s[n] = \sum_{k=1}^{P} a_k s[n-k] + \varepsilon[n]$$

LPC Coefficients — Prediction Error

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} R[0] & R[1] & R[2] & \cdots & R[P-1] \\ R[1] & R[0] & R[1] & \cdots & R[P-2] \\ R[2] & R[1] & R[0] & \cdots & R[P-3] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R[P-1] & R[P-2] & R[P-3] & \cdots & R[0] \end{bmatrix}^{-1} \begin{bmatrix} R[1] \\ R[2] \\ R[3] \\ \vdots \\ R[P] \end{bmatrix}$$

Linear Predictive Coding

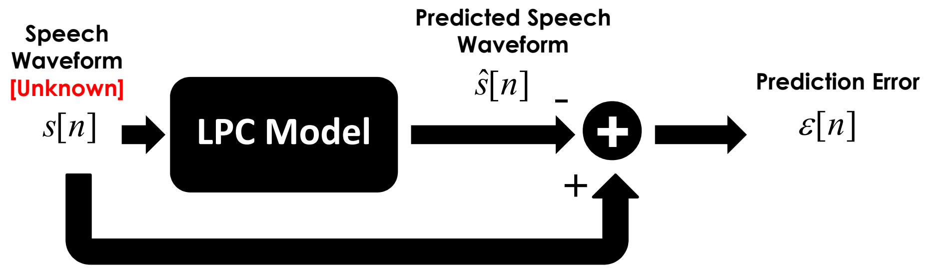
$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} R[0] & R[1] & R[2] & \cdots & R[P-1] \\ R[1] & R[0] & R[1] & \cdots & R[P-2] \\ R[2] & R[1] & R[0] & \cdots & R[P-3] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R[P-1] & R[P-2] & R[P-3] & \cdots & R[0] \end{bmatrix}^{-1} \begin{bmatrix} R[1] \\ R[2] \\ R[3] \\ \vdots \\ R[P] \end{bmatrix}$$

Speech Waveform [Phoneme] → **Training LPC Model** → **LPC Coefficients** $\{a_i\}$

Speech Production Model

$$s[n] = \sum_{k=1}^{P} a_i s[n-k] + K \cdot x[n]$$

Linear Predictive Coding



LPC Model

$$s[n] = \sum_{k=1}^{P} a_k s[n-k] + \varepsilon[n]$$

Speech Production Model

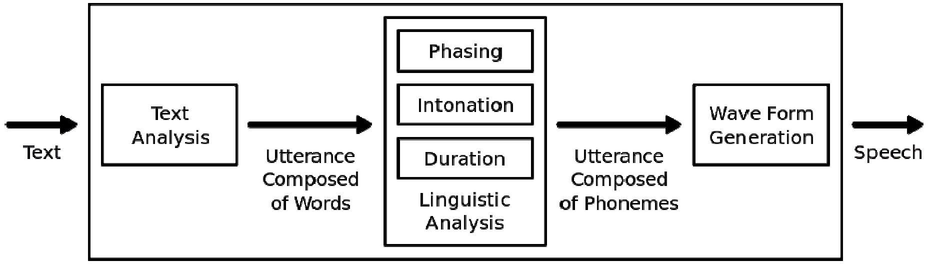
$$s[n] = \sum_{k=1}^{P} a_i s[n-k] + K \cdot x[n]$$

Excitation can be estimated from prediction error

- **Voiced/Unvoiced Classes**
- **Pitches**

$$\varepsilon[n] = K \cdot x[n]$$

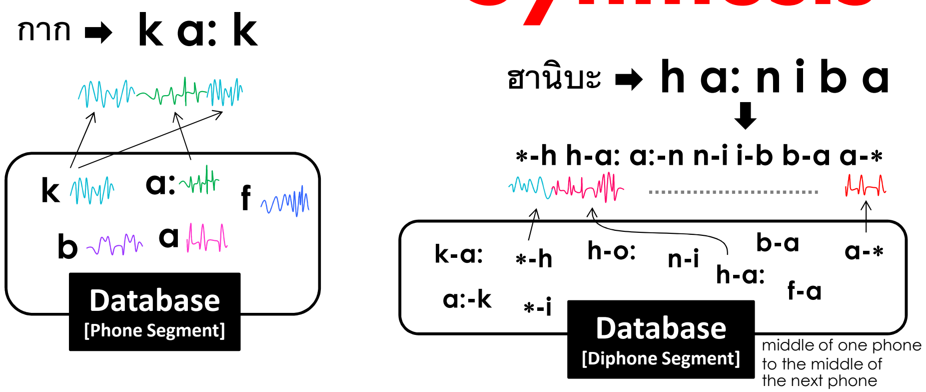
Speech Synthesis



TTS : Text-to-Speech

http://en.wikipedia.org/wiki/Speech_synthesis

Concatenative Synthesis



- Concatenate segments of recorded speech
- Produce natural synthesized speech
- Large databases of segmented recorded speech (phones, diphones, words, phrases, sentences)
- Unit selection synthesis, Diphone synthesis, Domain-specific synthesis

Time & Pitch Manipulation

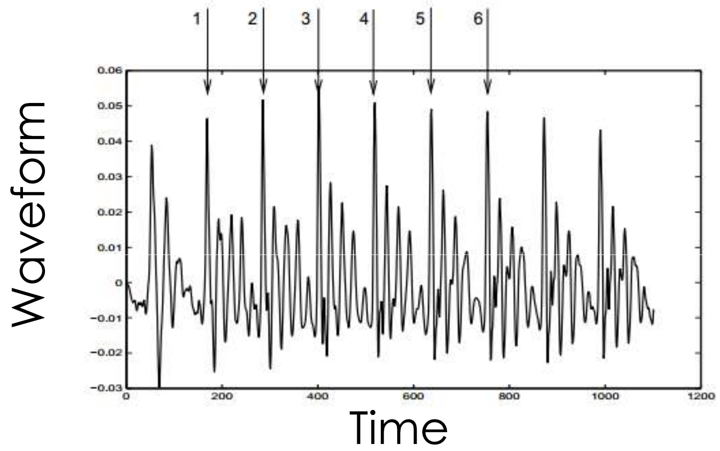
- **Time Stretching =** Change duration of sound (Shorten/Lengthen) without affecting its pitches
- **Pitch Shifting/Scaling =** Change pitch of sound (Lower/Higher) without affecting its duration

Techniques

- Resampling
- Phase Vocoder
- PSOLA (Pitch Synchronous Overlap and Add)

PSOLA Pitch Synchronous Overlap and Add

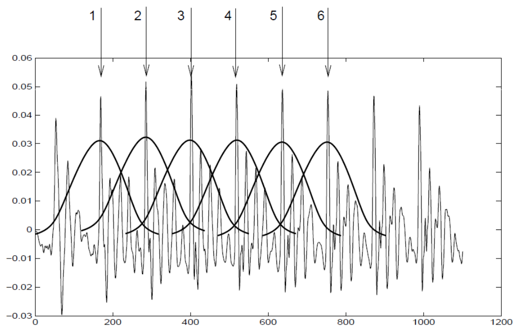
Epoch (Pitch Mark) = Single instant in each pitch period that serves as an "anchor"

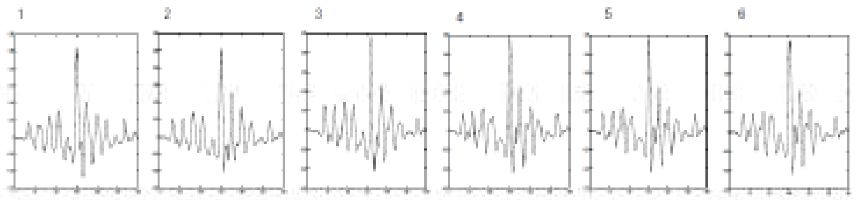


Paul Taylor, Text-to-Speech Synthesis

PSOLA Pitch Synchronous Overlap and Add

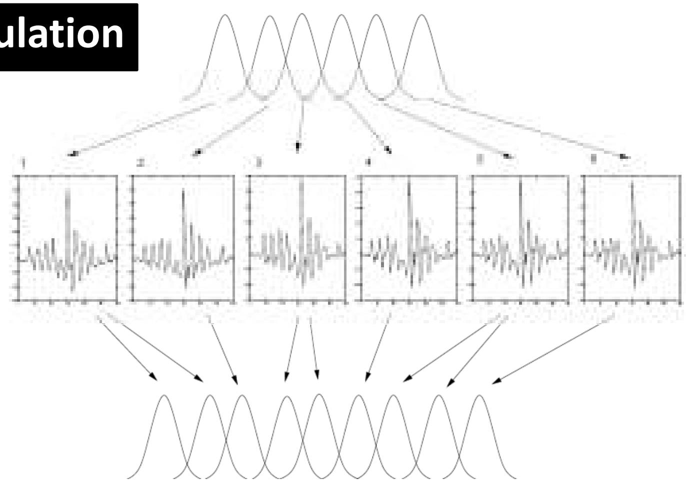
Time segmentation using overlapping window with center at epoch





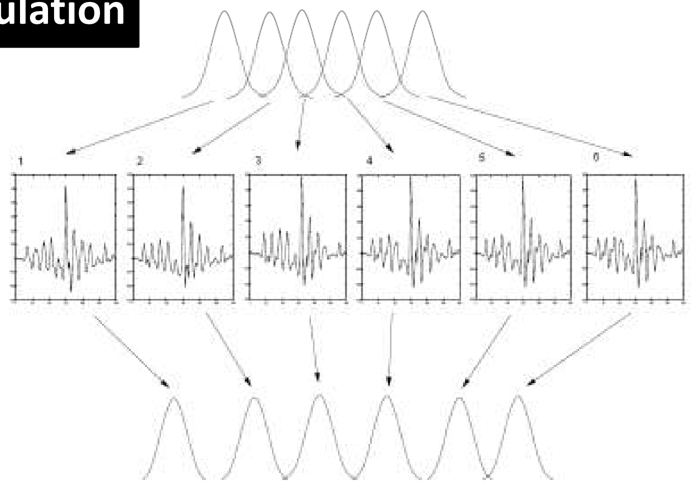
PSOLA Pitch Synchronous Overlap and Add

Time Manipulation

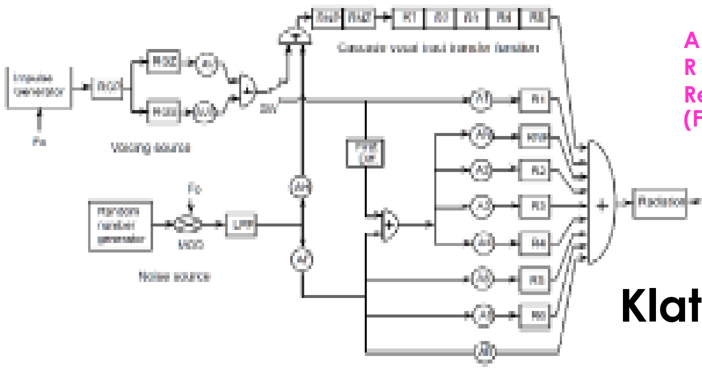


PSOLA Pitch Synchronous Overlap and Add

Pitch Manipulation



Formant Synthesis



A = Gain
R = Filter for Specific Resonance Frequency (Formant)

Klatt synthesiser

- Create a waveform using acoustic model and does not use human speech samples at runtime
- Smaller program, Higher synthesizing speed, avoiding the acoustic glitches
- Artificial, robotic-sounding speech

Formant Synthesis

Single Formant Synthesis Filter

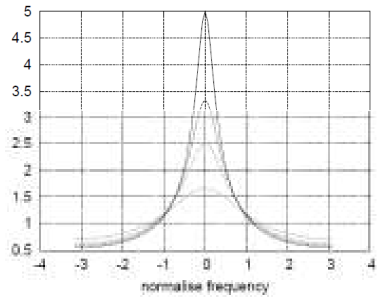
$$y[n] = s[n] + 2e^{-\pi B/F_S}\cos\left(2\pi\frac{F_R}{F_S}\right)y[n-1] - e^{-2\pi B/F_S}\,y[n-2]$$

F_R = Resonance Frequency (Formant)

B = Bandwidth of Resonance Filter

F_S = Sampling Frequency

	/a/	/e/	/i/	/o/	/u/	BW
F1	750	469	281	468	312	90
F2	1187	2031	2281	781	1219	110
F3	2595	2687	3187	2656	2469	170
F4	3781	3375	3781	3281	3406	250
F5	4200	4200	4200	4200	4200	300



Paul Taylor, Text-to-Speech Synthesis
Alain de Cheveigné, FORMANT BANDWIDTH AFFECTS THE IDENTIFICATION OF COMPETING VOWELS

HMM-based Synthesis

We know phone sequence and duration for each phone. The phone sequence tells us which models to use in which order, but not which states to use, or which observations to generate. A duration tells us how many observations that should be generated, but again not which states to generate from.

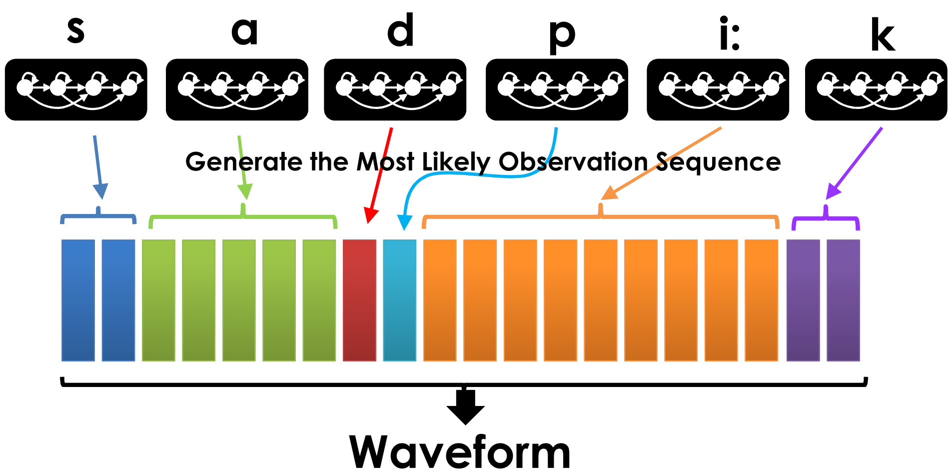
Generate the most likely sequence of observations from the sequence of models

Each state will generate its mean observation

HMM-based Synthesis

สัตว์ปีก → s a d p i: k

Duration 2 5 1 1 9 2 Frames



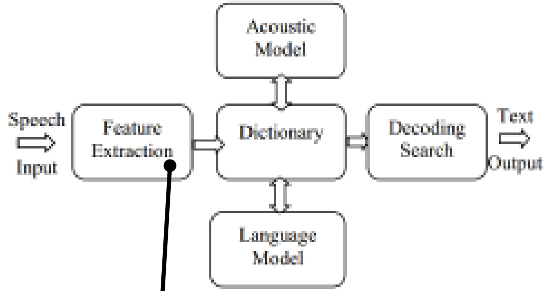
Speech Recognition

Speaker Model	Speaker Dependence [used by a single speaker] Speaker Independence [used by any speaker]
Vocabulary Size	Small Vocabulary [10 words] Medium Vocabulary [100 words] Large Vocabulary [1000 words] Very Large Vocabulary [10000 words] Out-of-Vocabulary [10000 words]
Speech Utterance	Isolated Words [single word] Connected Words / Discontinuous Speech [full sentence separated by silence] Continuous Speech [Naturally spoken sentences] Spontaneous Speech [Include mispronunciations, false-starts, and nonwords]

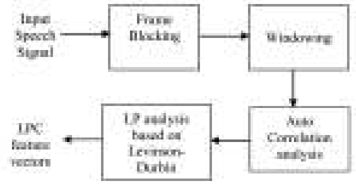
Growth of ASR

Year	Progress of ASR System
1952	Digit Recognizer
1976	1000 word connected recognizer with constrained grammar
1980	1000 word LSM recognizer (separate words w/o grammar)
1988	Phonetic typewriter
1993	Read texts (WSJ news)
1998	Broadcast news, telephone conversations
1998	Speech retrieval from broadcast news
2002	Rich transcription of meetings, Very Large Vocabulary, Limited Tasks, Controlled Environment
2004	Finnish online dictation, almost unlimited vocabulary based on morphemes
2006	Machine translation of broadcast speech
2008	Very Large Vocabulary, Limited Tasks, Arbitrary Environment
2009	Quick adaptation of synthesized voice by speech recognition (in a project where TKK participates in)
2011	Unlimited Vocabulary, Unlimited Tasks, Many Languages, Multilingual Systems for Multimodal Speech Enabled Devices
Future Direction	Real time recognition with 100% accuracy, all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics or accent.

Speech Recognition



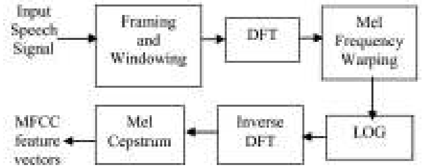
LPC Features



Speech Features

- Easy to measure extracted speech features
- It should not be susceptible to mimicry
- It should show little fluctuation from one speaking environment to another
- It should be stable over time
- It should occur frequently and naturally in speech

MFCC Features



Speech Recognition

Approaches

- Hidden Markov Models (HMM)
  - Monophone Models
  - Triphone Models
- Neural Networks (NN)
- Dynamic Time Warping (DTW)

HMM Speech Recognition

$$\hat{W} = \arg\max_{W} P(W \mid X) = \arg\max_{W} \frac{P(X \mid W) P(W)}{P(X)}$$

X = Observation Sequence ที่มี

W = Symbol Sequence ที่เป็นไปได้

Sequence of phoneme (word, phrase, sentences)

$\hat{W}$ = Symbol Sequence ที่ Recognized ได้ สำหรับ Observation Sequence X

$P(X|W)$ = Prob. ของการเกิด X จากคำ/วลี/ประโยค W

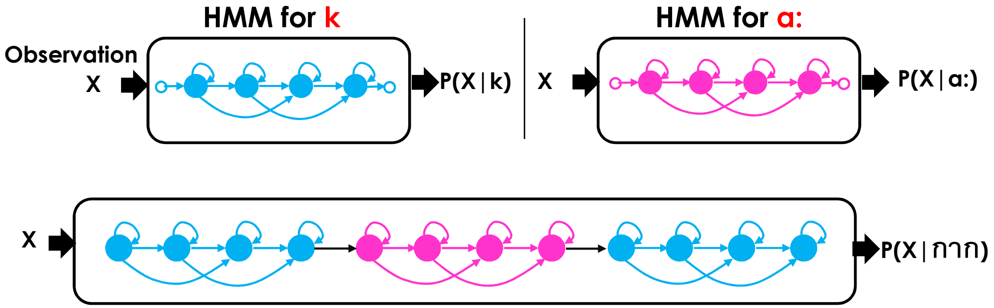
หาได้จาก Acoustic Model

$P(W)$ = Prob. ของการเกิด คำ/วลี/ประโยค W

หาได้จาก Language Model

Acoustic Model

- Statistical model that describe the probability of acoustic observation sequence

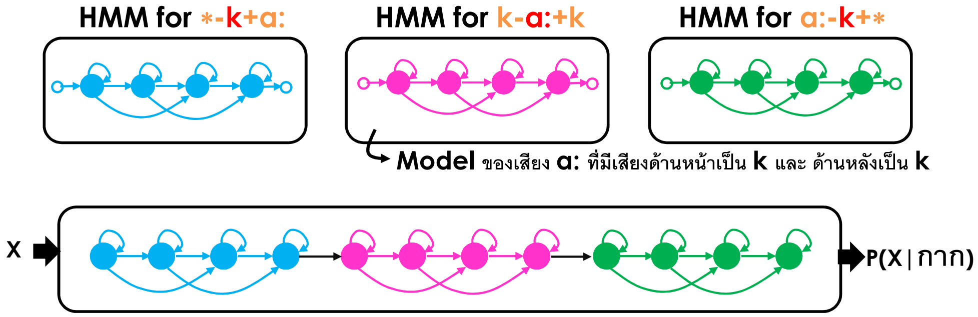


Composed HMM for word กาก [ka:k]

Monophone Model [Context Independent]

Acoustic Model

- Statistical model that describe the probability of acoustic observation sequence



Composed HMM for word กาก [ka:k]

Triphone Model [Context Dependent]

Acoustic Model

Monophone Model [Context Independent]

แม่งง = m, ɛ:, ŋ, o, ŋ

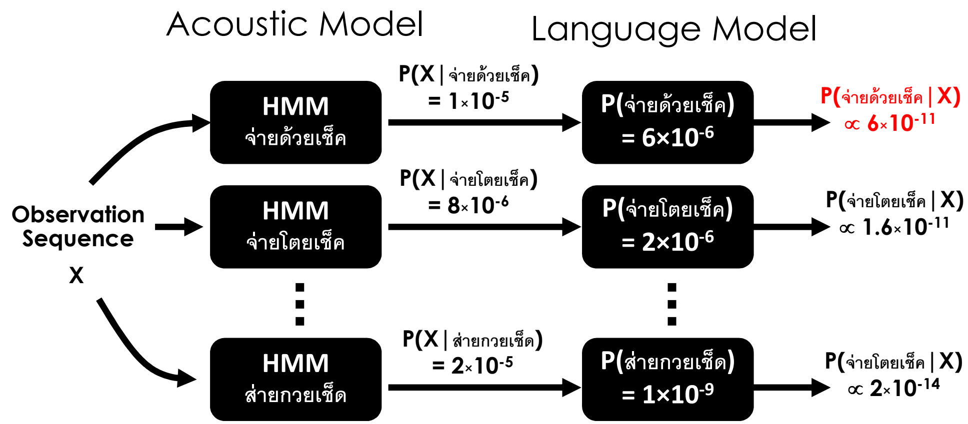
Triphone Model [Context Dependent]

Word Internal แม่งง = *-m+ɛ:, m-ɛ:+*, *-ŋ+o, ŋ-o+ŋ, o-ŋ+*

Cross-Word แม่งง = *-m+ɛ:, m-ɛ:+ŋ, ɛ:-ŋ+o, ŋ-o+ŋ, o-ŋ+*

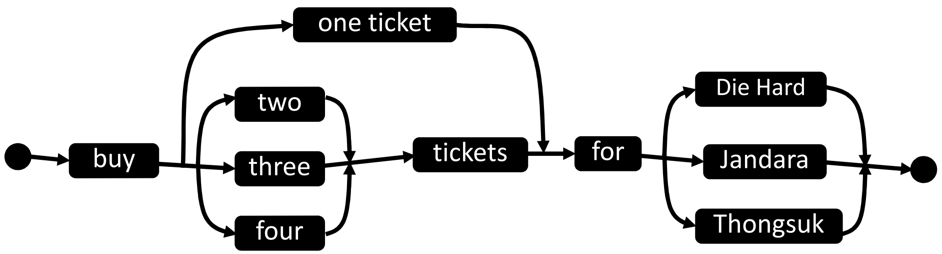
Language Model

- Statistical Model that describe the probability of the symbol sequence (word/ phrase/ sentence)



Language Model

Finite State Networks



Assume that words are Independent

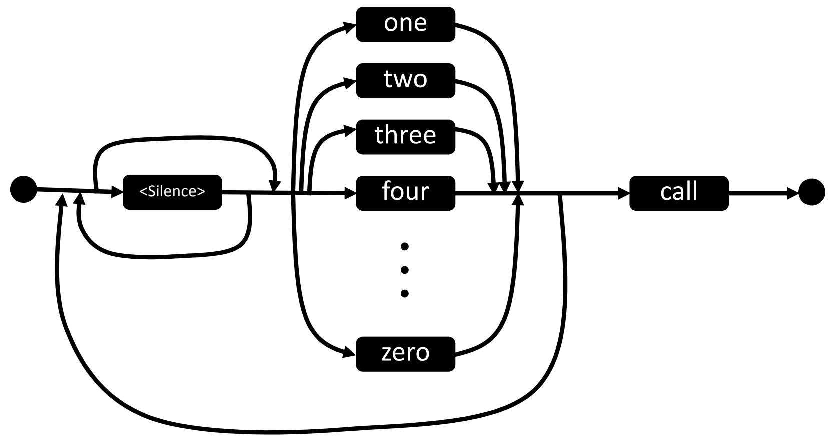
P("buy two tickets for Jandara") = P(two)P(Jandara)

P("buy one ticket for Die Hard") = P(one ticket)P(Die Hard)

P("buy one Die Hard") = 0

Language Model

Finite State Networks



Language Model

n-gram Language Model

$$P(w_1, w_2, \ldots, w_L) = \prod_{i=1}^{L} P(w_i | w_1, w_2, \ldots, w_{i-1})$$

$$\approx \prod_{i=1}^{L} P(w_i | w_{i-(n-1)}, w_{i-(n-2)}, \ldots, w_{i-1})$$

Unigram

$$P(w_1, \ldots, w_L) \approx \prod_{i=1}^{L} P(w_i)$$

Bigram

$$P(w_1, \ldots, w_L) \approx \prod_{i=1}^{L} P(w_i | w_{i-1})$$

Trigram

$$P(w_1, \ldots, w_L) \approx \prod_{i=1}^{L} P(w_i | w_{i-1}, w_{i-2})$$

Language Model

P(หมีเป็นสัตว์กินเบียร์) = P(หมี)×P(เป็น|หมี)×P(สัตว์|เป็น,หมี)
×P(กิน|สัตว์,เป็น,หมี)×P(เบียร์|กิน,สัตว์,เป็น,หมี)

Unigram Approximation

P(หมีเป็นสัตว์กินเบียร์) ≈ P(หมี)×P(เป็น)×P(สัตว์)×P(กิน)×P(เบียร์)

Bigram Approximation

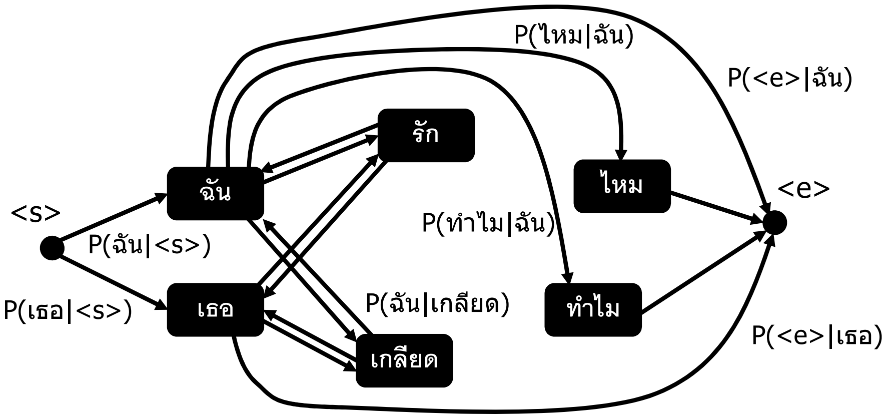
P(หมีเป็นสัตว์กินเบียร์) ≈ P(หมี|<start>)×P(เป็น|หมี)×P(สัตว์|เป็น)
×P(กิน|สัตว์)×P(เบียร์|กิน)×P(<end>|เบียร์)

Trigram Approximation

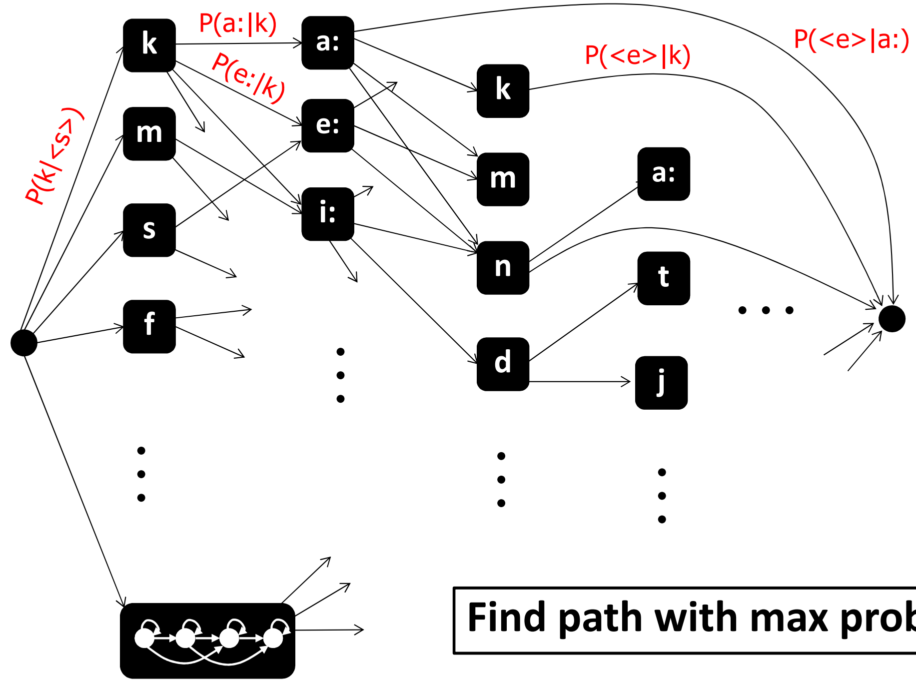
P(หมีเป็นสัตว์กินเบียร์) ≈ P(หมี|<start>,<start>)×P(เป็น|หมี,<start>)×P(สัตว์|เป็น,หมี)
×P(กิน|สัตว์,เป็น)×P(เบียร์|กิน,สัตว์)×P(<end>|เบียร์,กิน)

Language Model

Bigram Approximation



P(เธอเกลียดฉัน|X) ∝ P(X|เธอเกลียดฉัน)P(เธอ|<s>)P(เกลียด|เธอ)P(ฉัน|เกลียด) P(<e>|ฉัน)



Find path with max prob.