

Speech Processing

Chung-Hsien Wu

Professor

Department of Computer Science and Information Engineering

National Cheng Kung University

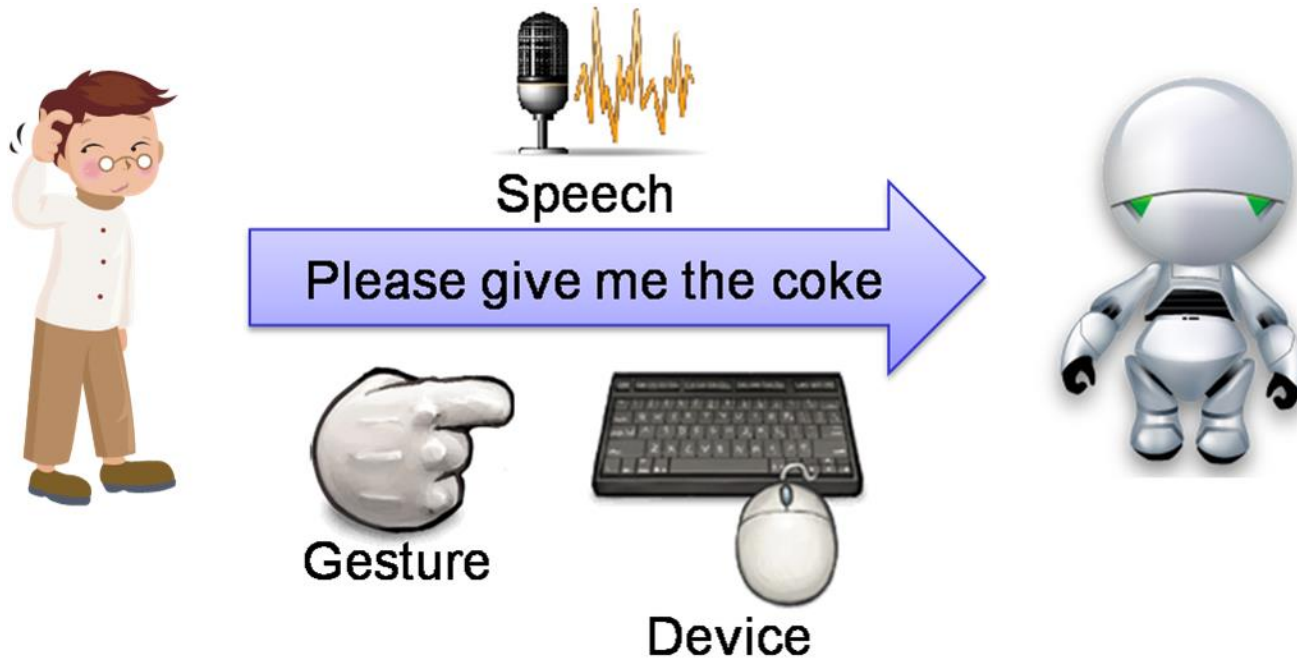
chunghsienwu@gmail.com

Outline

- Introduction
- Speech Production
- Speech Signal Representation
- Speech Coding
- Speech Synthesis
- Automatic Speech Recognition

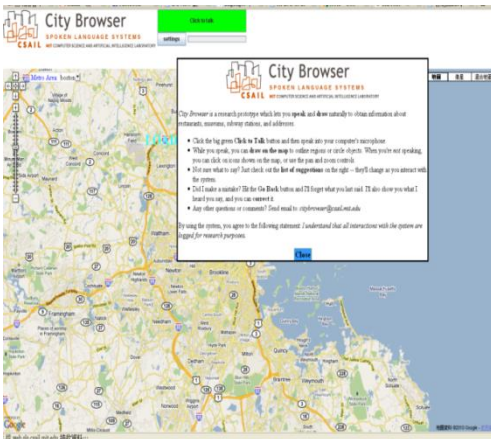
Introduction

- “Speech” is the most natural way for people to interact with others, including a machine.



Introduction

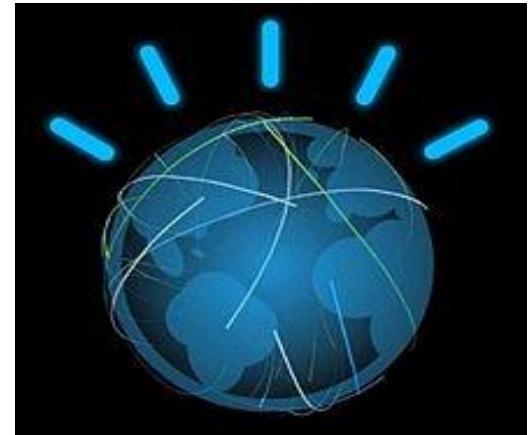
- In the past decade, spoken dialogue systems have been used in many domains for human-computer interaction, e.g., CityBrowser, iPhone 4S Siri, IBM DeepQA project (Waston)
- Examples



MIT CityBrowser

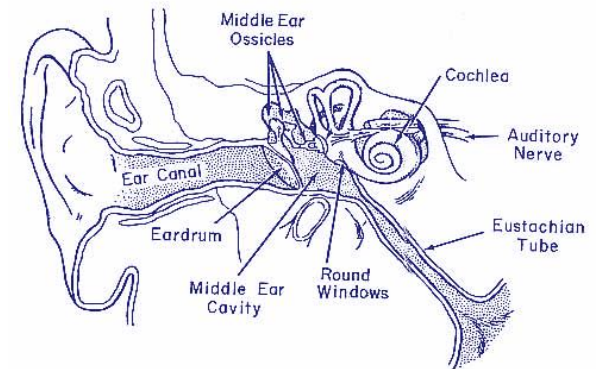
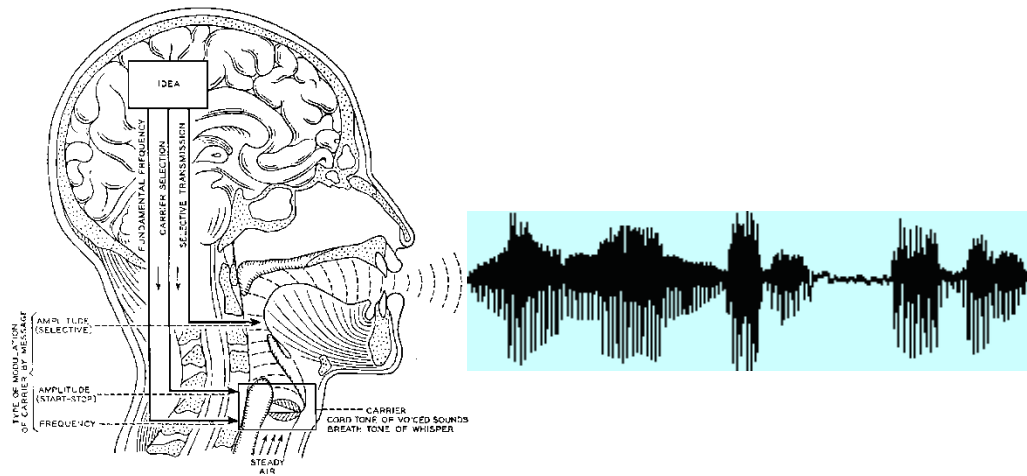


Apple
iPhone Siri

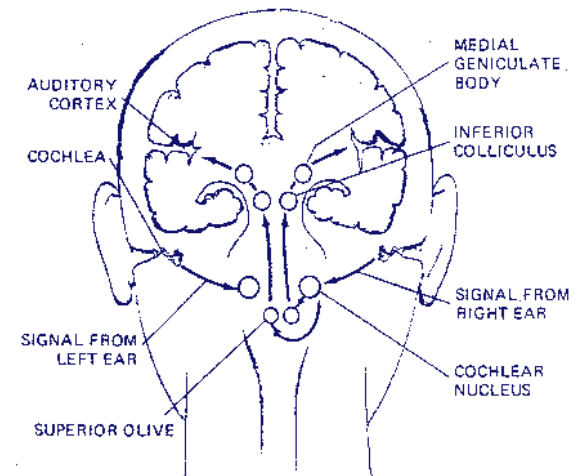


IBM Watson
wins Jeopardy

How do humans do it?



- Articulation produces sound waves which the ear conveys to the brain for processing

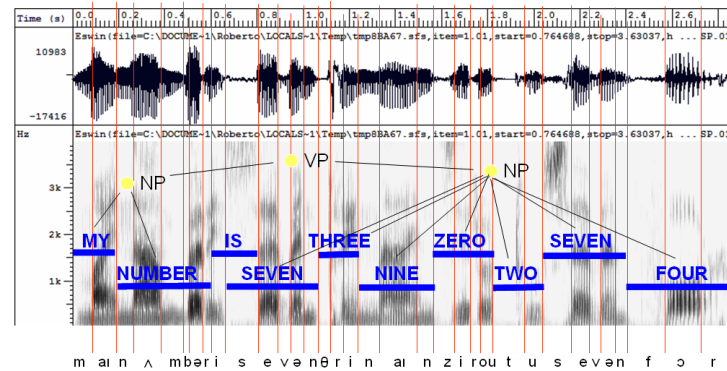
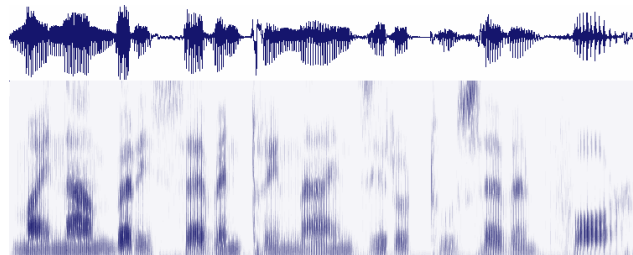


How might computers do it?



Acoustic waveform

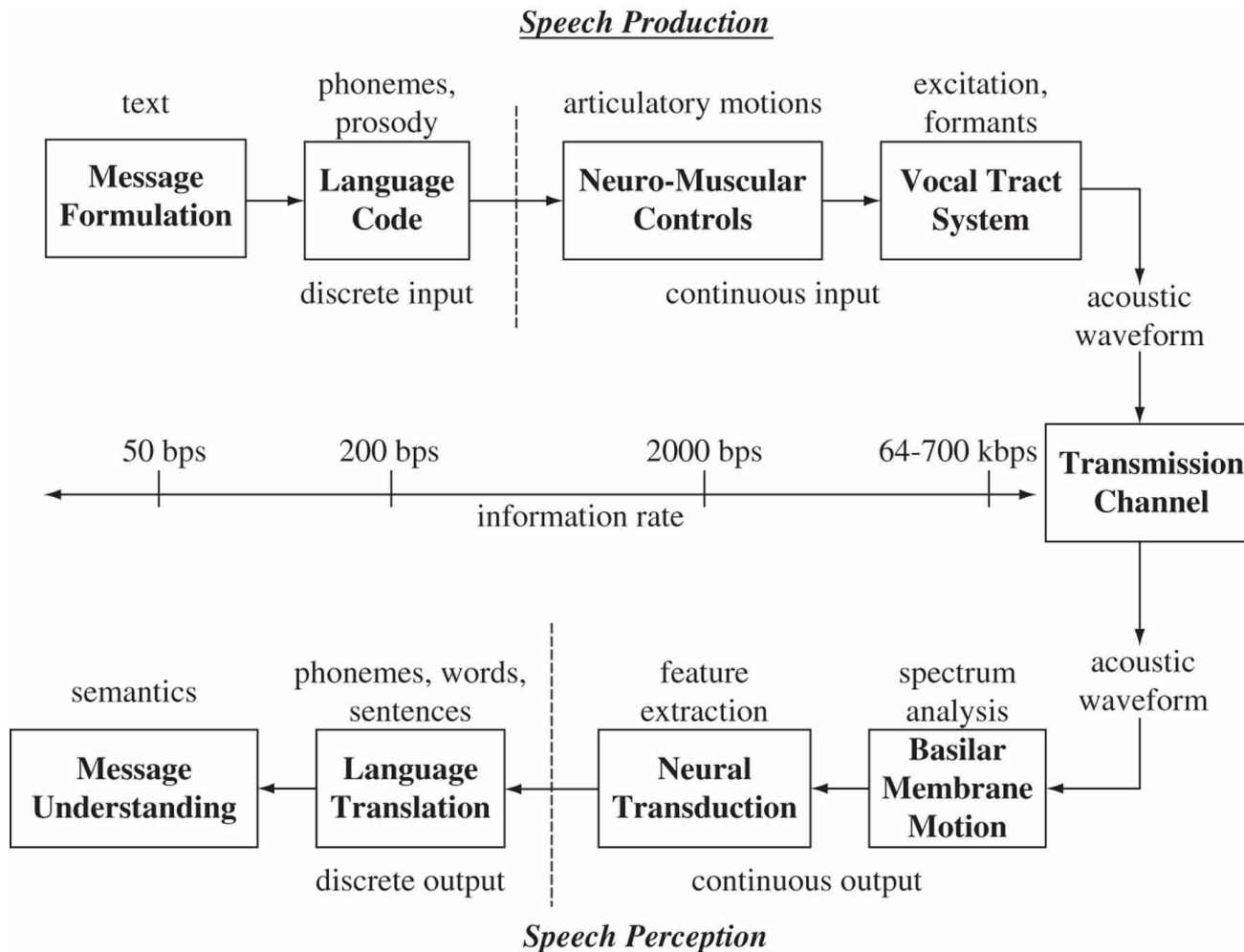
Acoustic signal



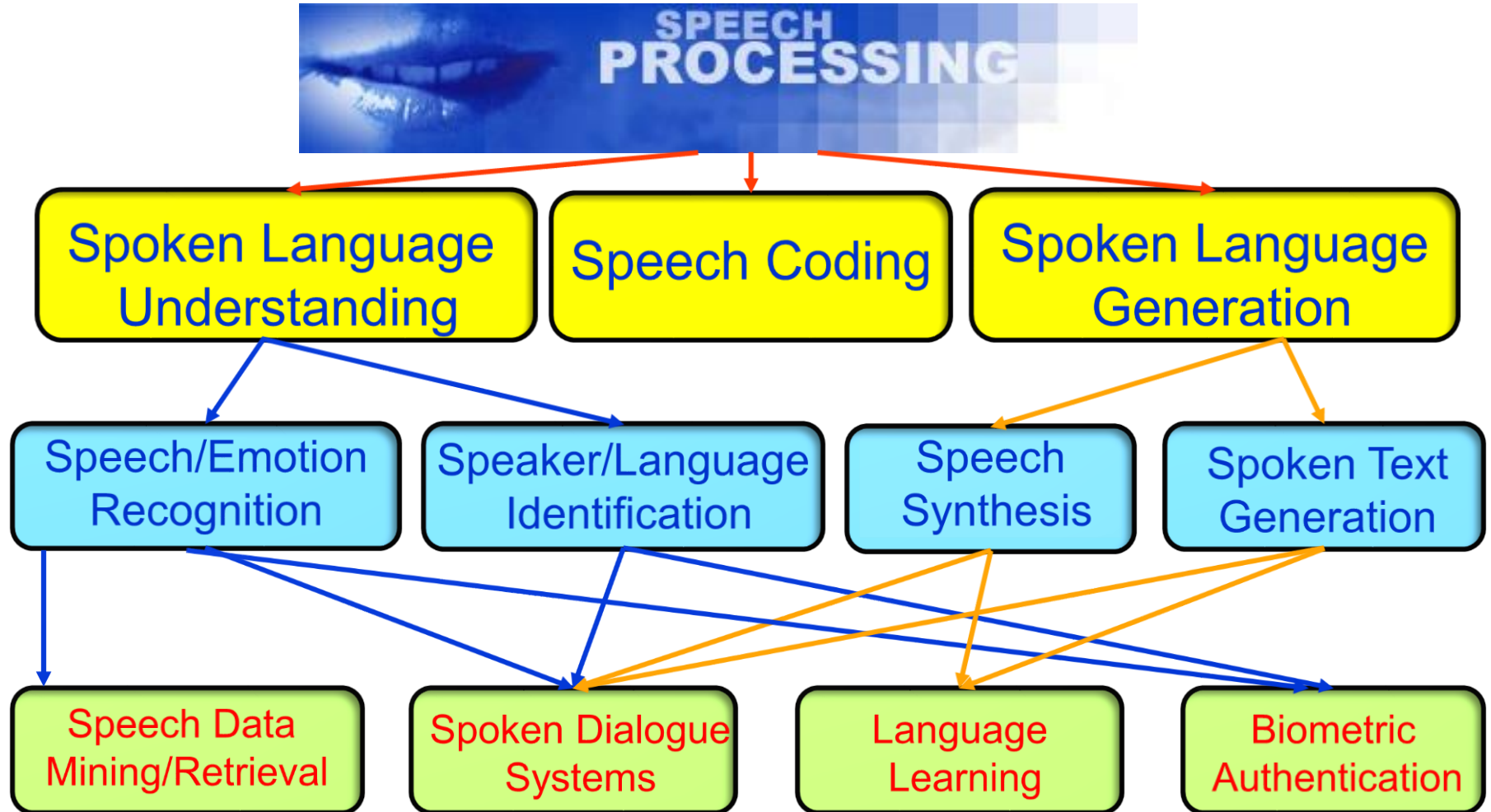
- Digitization
- Acoustic analysis of the speech signal
- Linguistic interpretation

Speech recognition

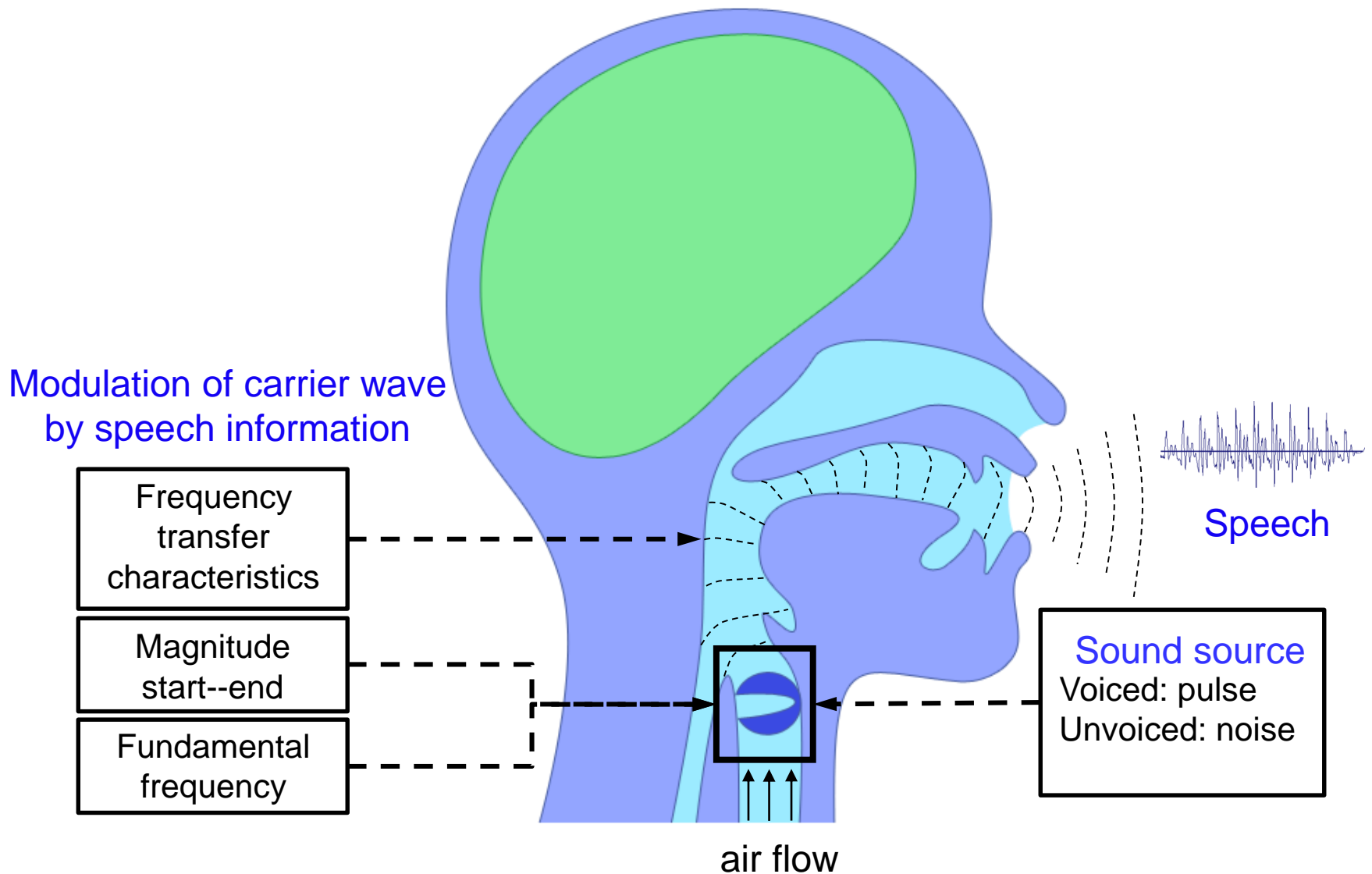
Speech Chain



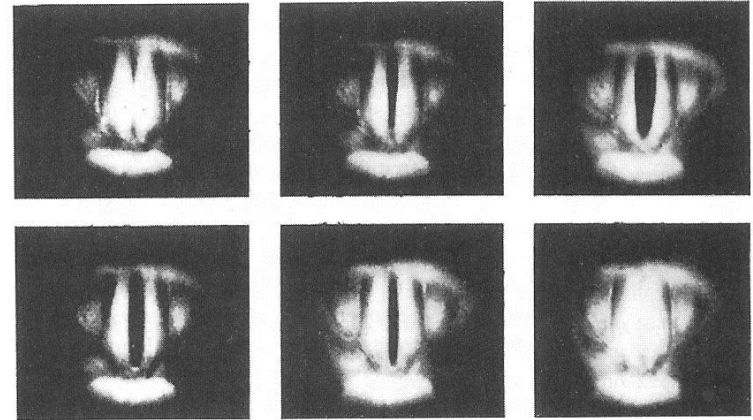
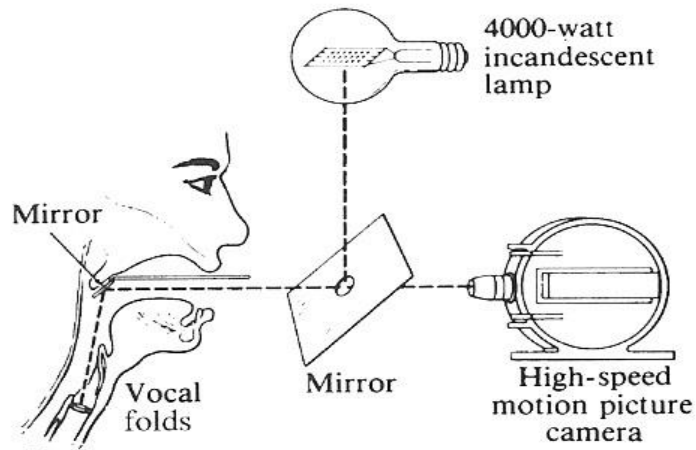
Hierarchy of Speech Processing



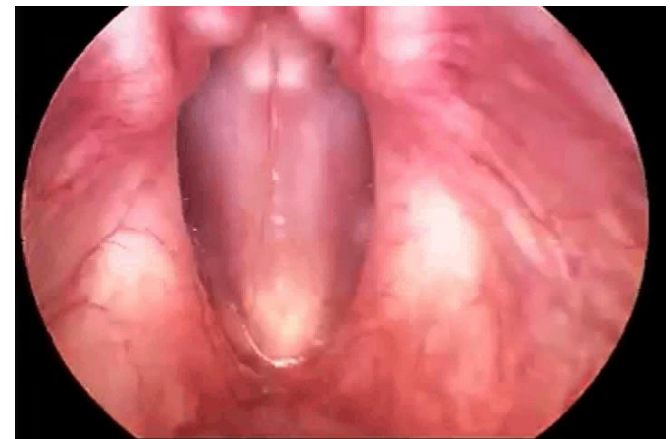
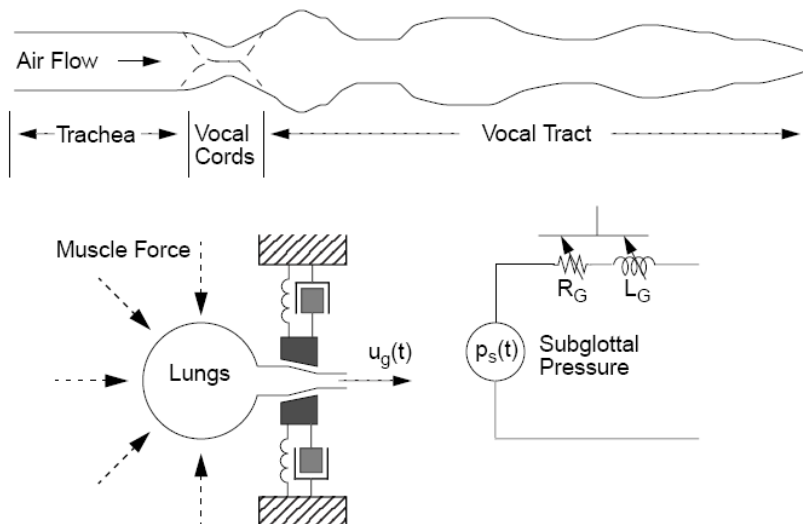
Part 1- Speech Production



Vocal Fold Vibrations



Excitation Model



Speech Production

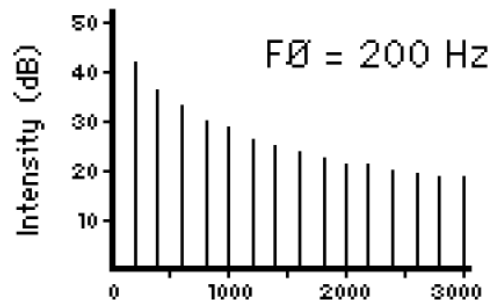
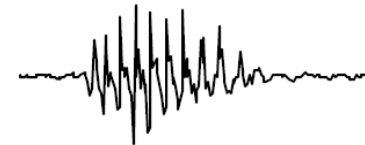
Glottal pulses



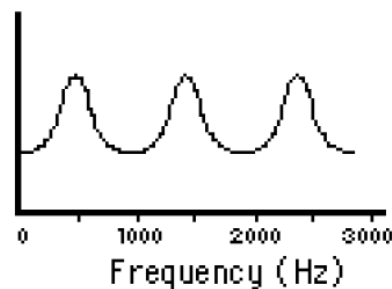
Vocal tract



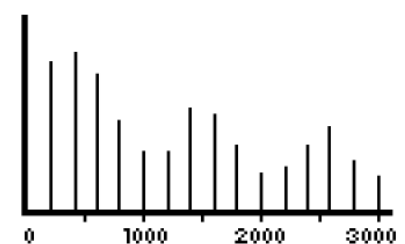
Speech signal



SOURCE SPECTRUM

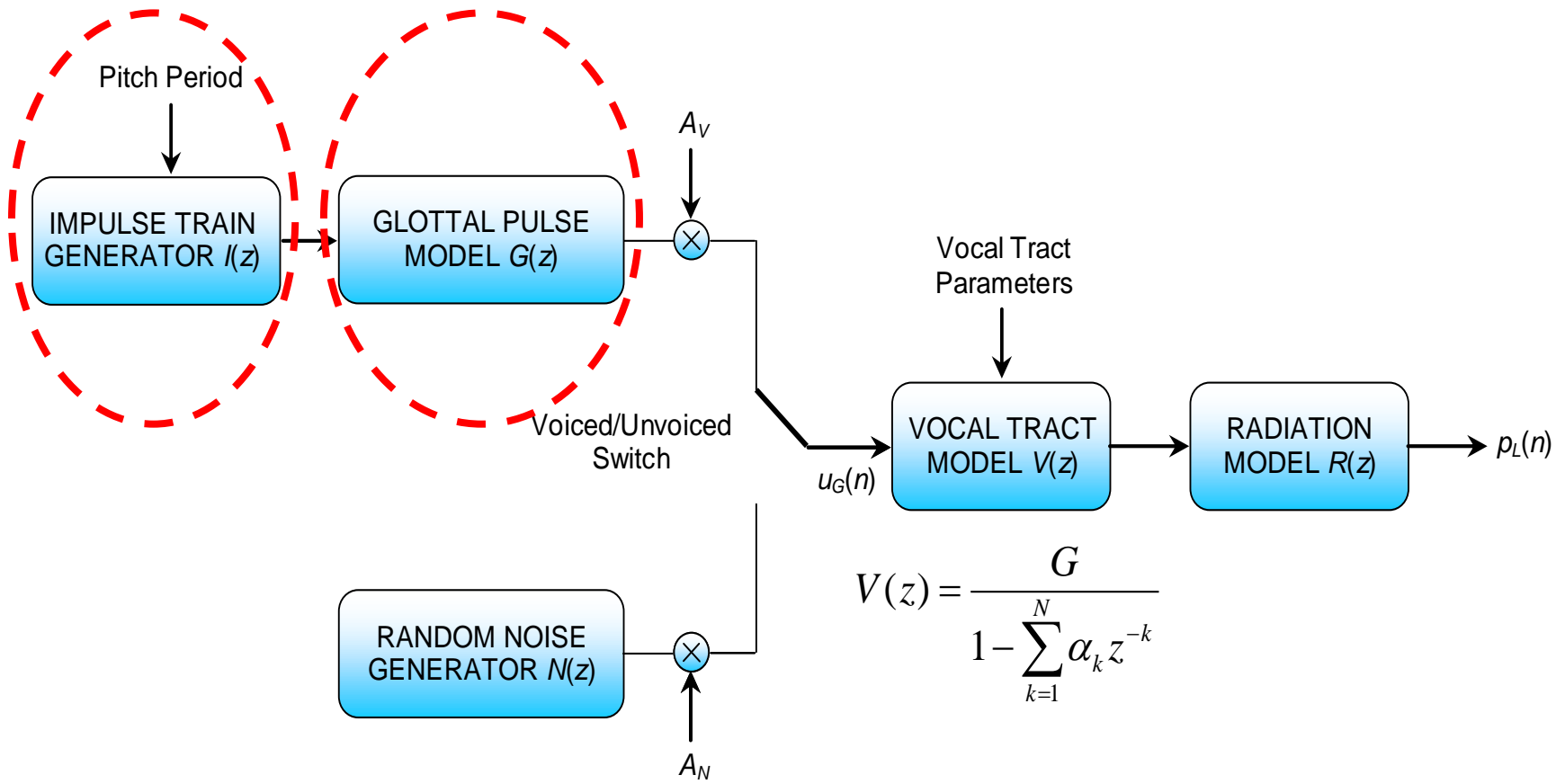


FILTER FUNCTION



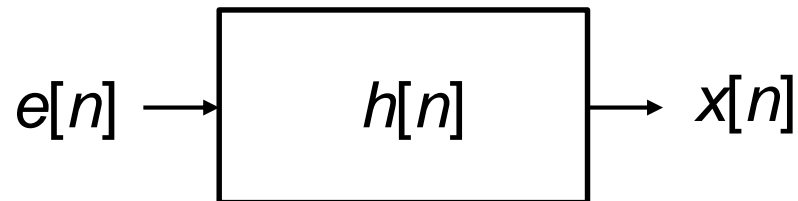
OUTPUT ENERGY SPECTRUM

Complete Digital Model



Part 2- Speech Signal Representation

- Several representations for speech signals are useful in speech coding, synthesis, and recognition.
- We describe methods to compute both the source or *excitation* $e[n]$ and the filter $h[n]$ from the speech signal $x[n]$.



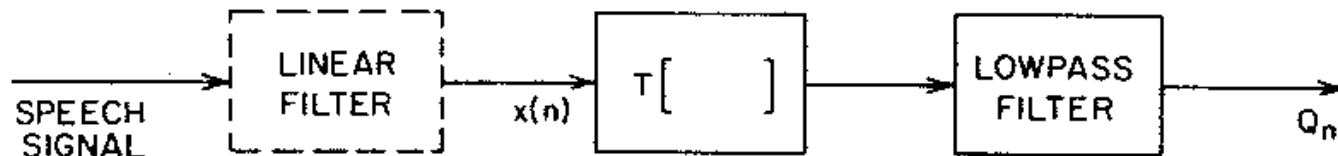
A. Time Domain Analysis

Assumption :

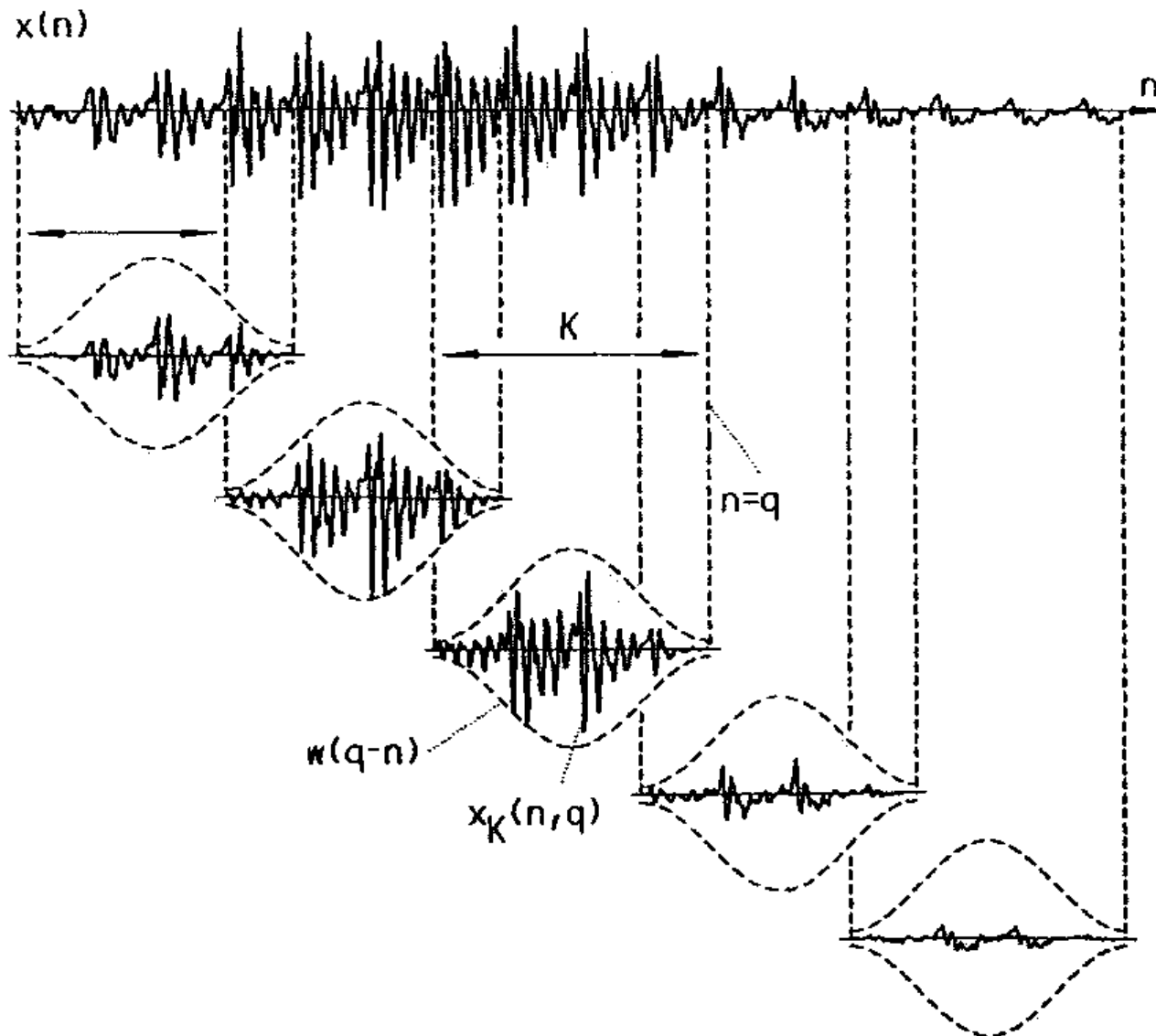
The properties of the speech signal change relatively slowly with time

→ **”Short-Time”** Processing

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)]w(n-m)$$

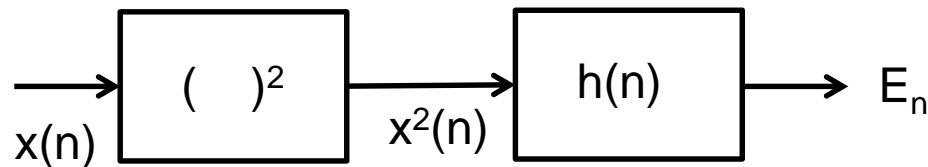


“Short-Time” Processing



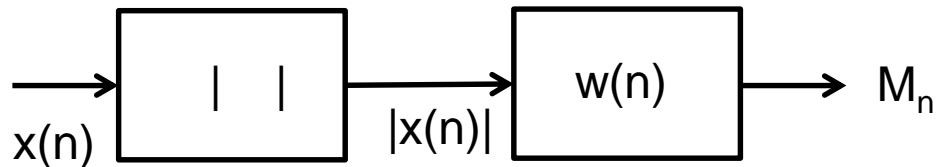
■ Short-Time Energy :

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2$$



■ Average Magnitude :

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)| w(n-m)$$

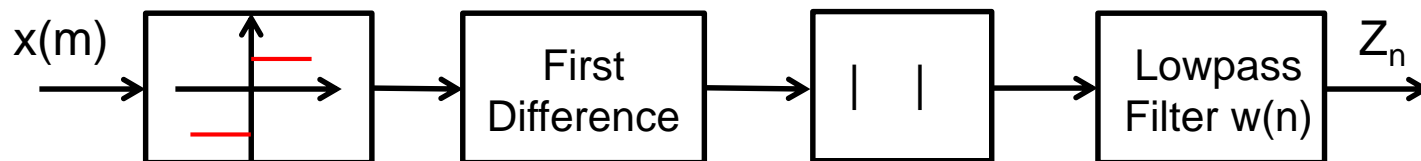
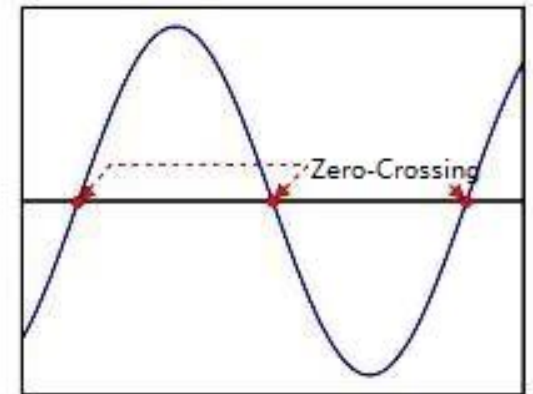


■ Average Zero-Crossing Rate (crossings/sample) :

Definition:
$$Z_n = \sum_{m=-\infty}^{\infty} |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]| w(n-m)$$

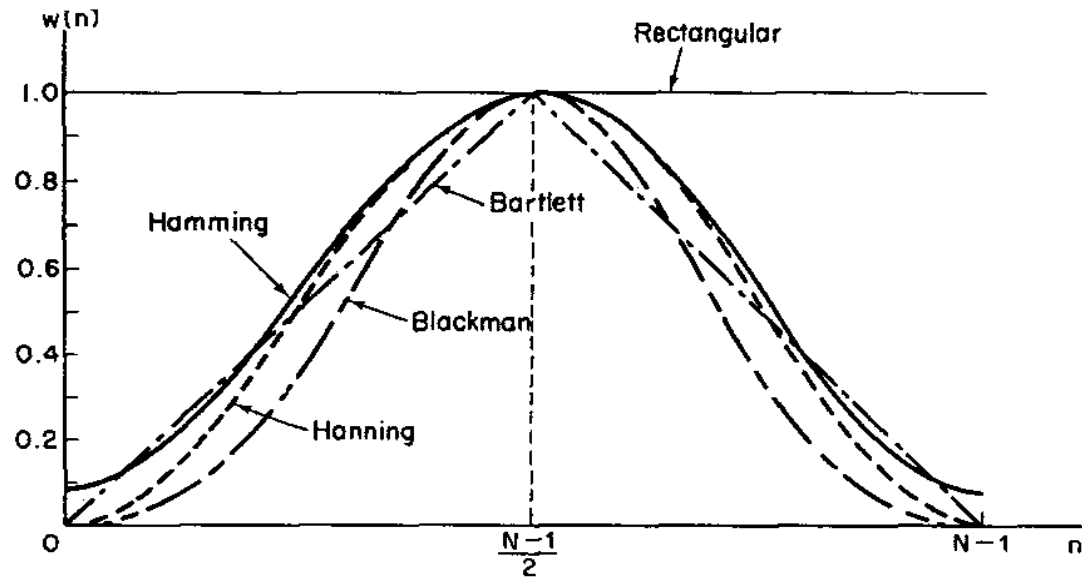
where $\text{sgn}(x[m]) = 1 \quad x[m] \geq 0$
 $\quad \quad \quad = -1 \quad x[m] < 0$

and $w(n) = \frac{1}{2N} \quad 0 \leq n \leq N-1$
 $\quad \quad = 0 \quad \text{otherwise}$





Windows:



■ Rectangular window

$$h(n) = 1 \quad 0 \leq n \leq N-1$$
$$= 0 \quad \text{otherwise}$$

■ Hamming window

$$h(n) = 0.54 - 0.46 \cos(2\pi n(N-1)), \quad 0 \leq n \leq N-1$$
$$= 0 \quad \text{otherwise}$$

■ Window Duration :

- N is too small (1 pitch period)

 - E_n will fluctuate

- N is too large (on the order of several pitch periods)

 - E_n will change very slowly

■ Suitable practical choice for N :

100-200 for 10KHz sampling rate

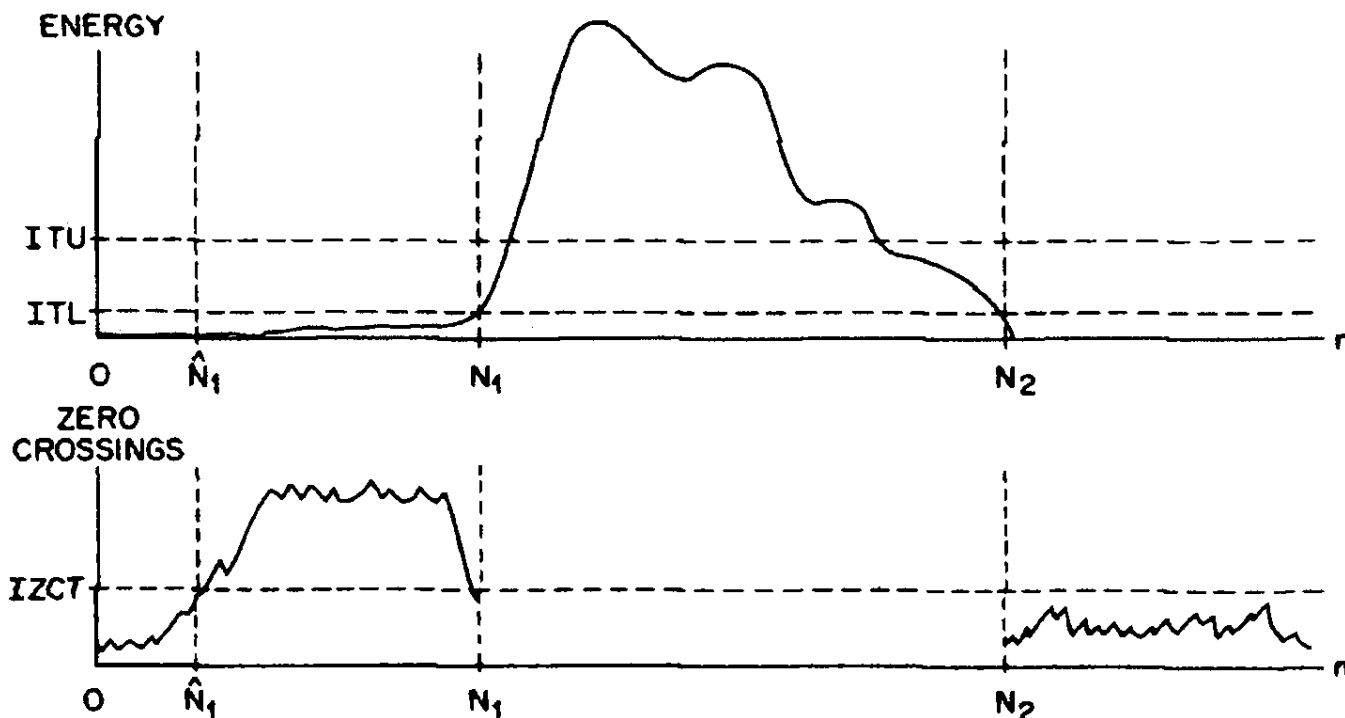
i.e. 10-20ms duration

End-Point Detection Using Energy and Zero-Crossings

Algorithm:

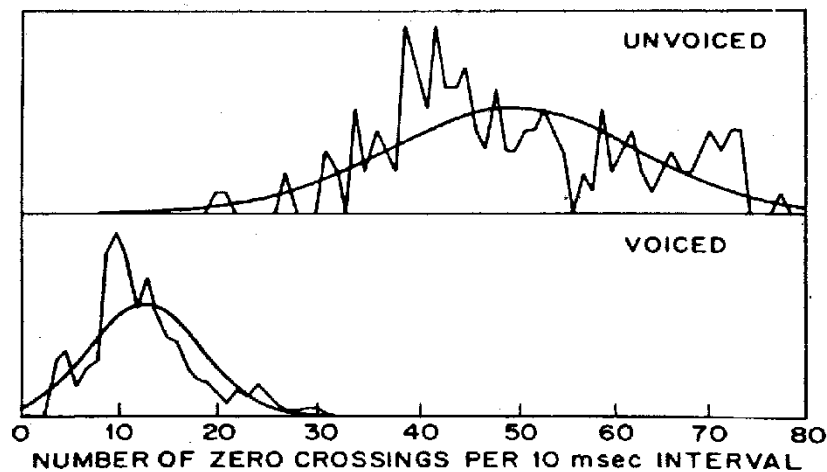
1. It is assumed that the first 100ms of the interval contains no speech
→ Compute ITU, ITL, and IZCT
2. Find the interval, in which the average magnitude exceeds a conservative threshold, ITU
3. Working backward from the point at which the average magnitude first exceeded the threshold ITU, the point N1(N2) where the average magnitude first falls below a lower threshold ITL is tentatively selected as the beginning (end) point.
4. Move backwards from N1 comparing the zero-crossing rate to a threshold (IZCT). If the zero-crossing rate exceeds the threshold 3 or more times, the beginning point N1 is moved back to the first point at which the ZC threshold was exceeded. Otherwise, N1 is defined as the beginning.
5. A similar procedure is followed at the end.

Operation



Distribution

---Zero crossing rate

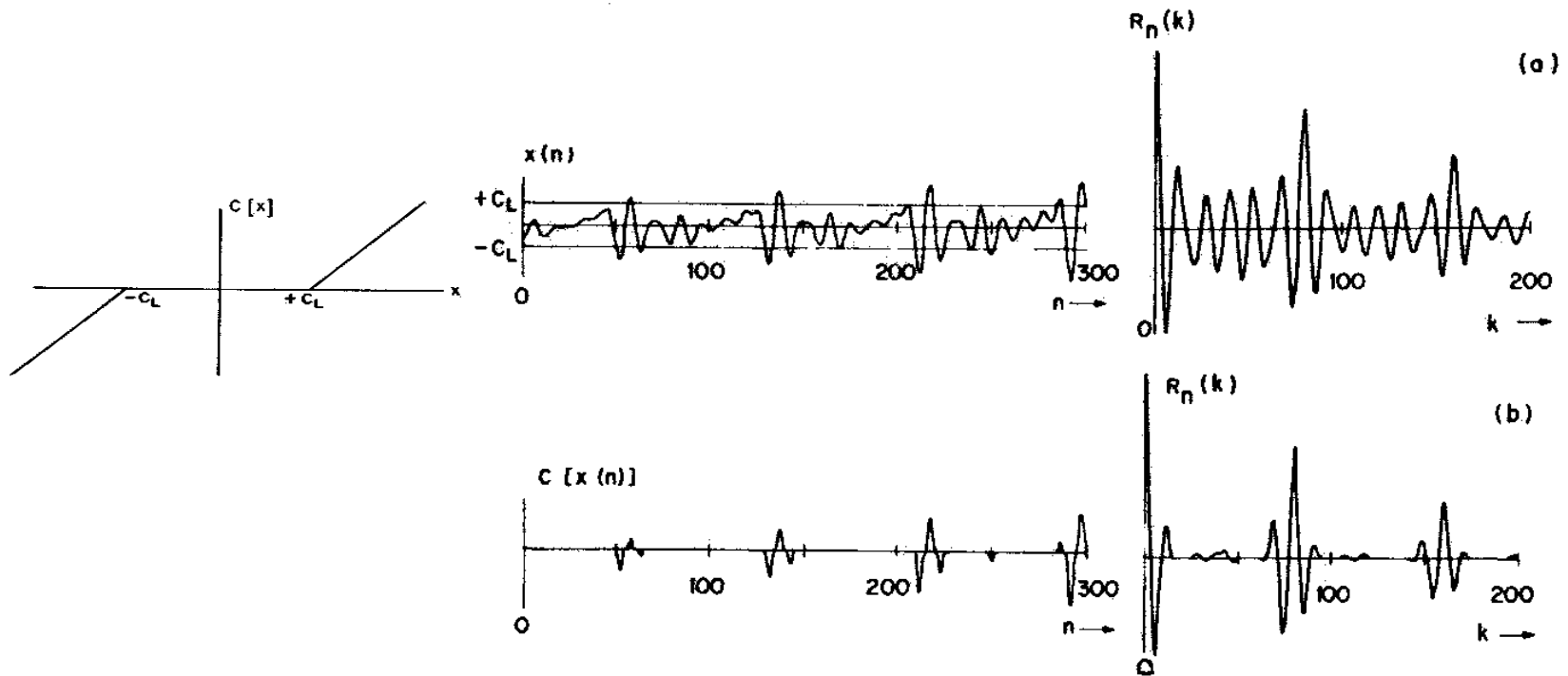


Pitch Detection

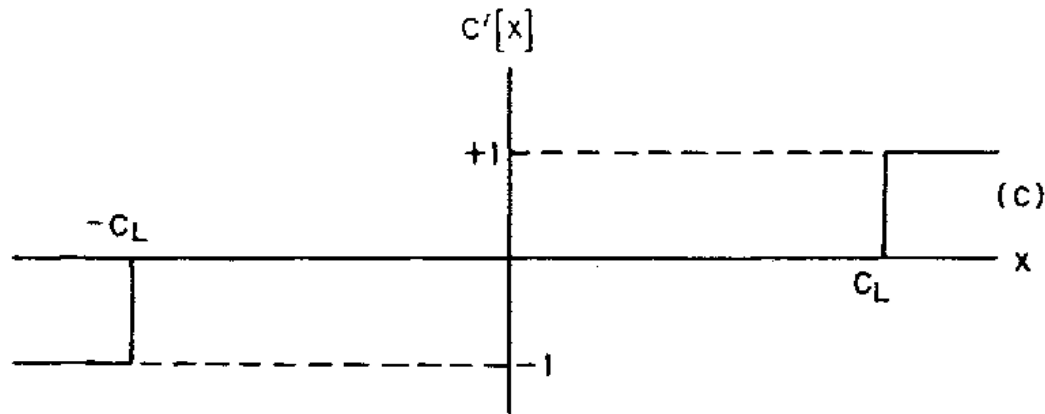
■ Short-Time Autocorrelation Function

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)w_1(n-m)x(m+k)w_2(n-(m+k))$$

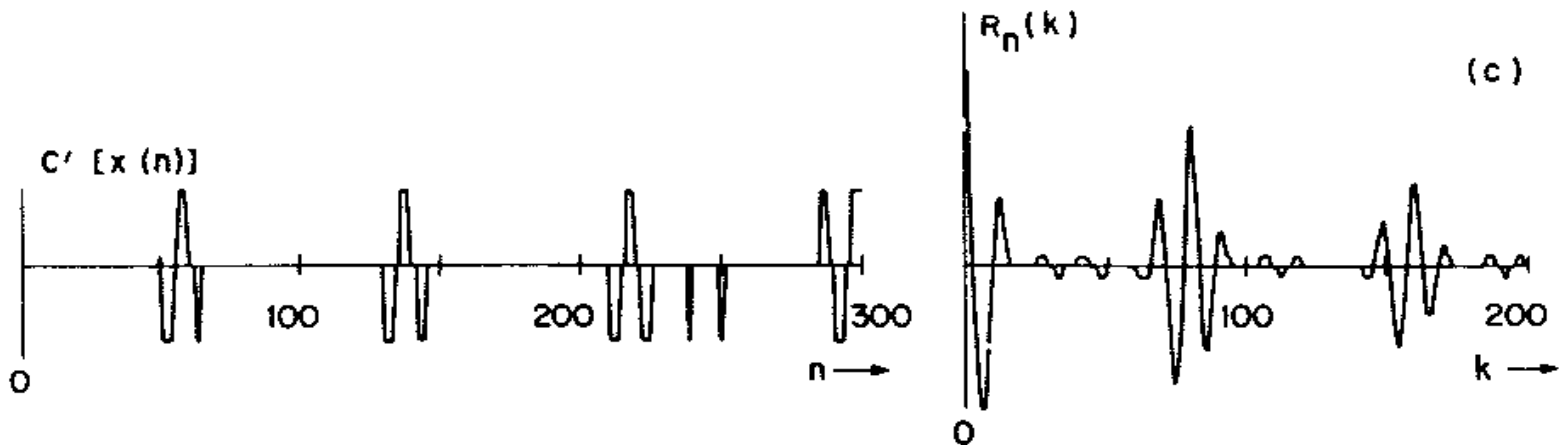
■ Pitch Detection using Center Clipping



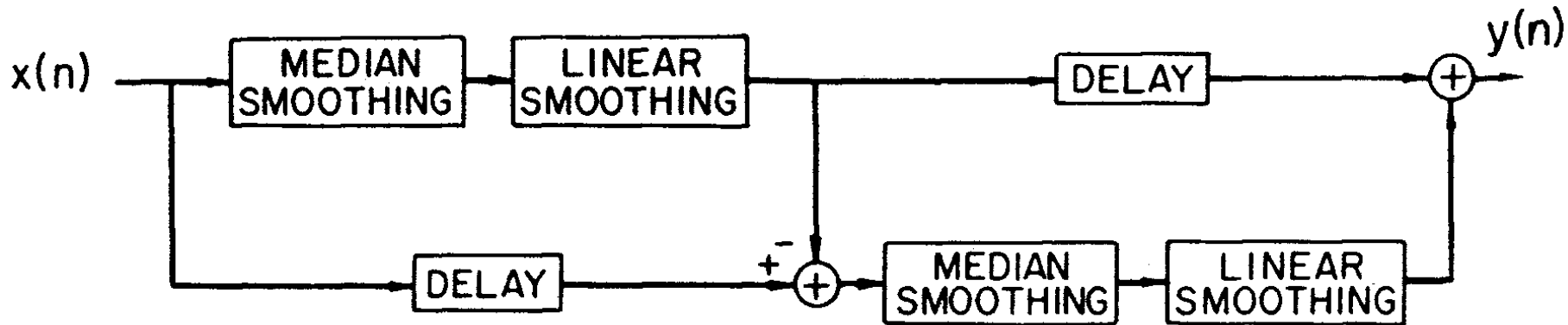
3-Level Center Clipping



$$\begin{aligned}
 y(n+m)y(n+m+k) &= 0 && \text{if } y(n+m) = 0 \text{ or } y(n+m+k) = 0 \\
 &= +1 && \text{if } y(n+m) = y(n+m+k) \\
 &= -1 && \text{if } v(n+m) \neq 0 \text{ and } v(n+m+k) \neq 0
 \end{aligned}$$



Median Smoothing



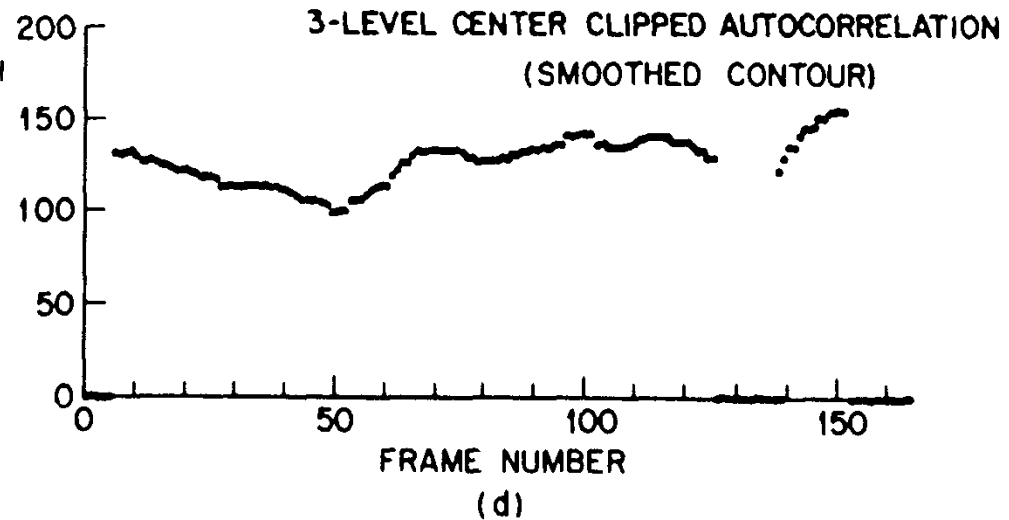
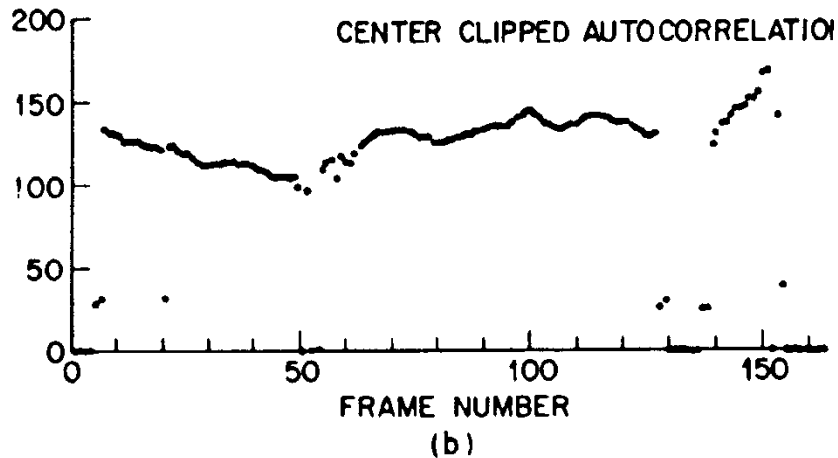
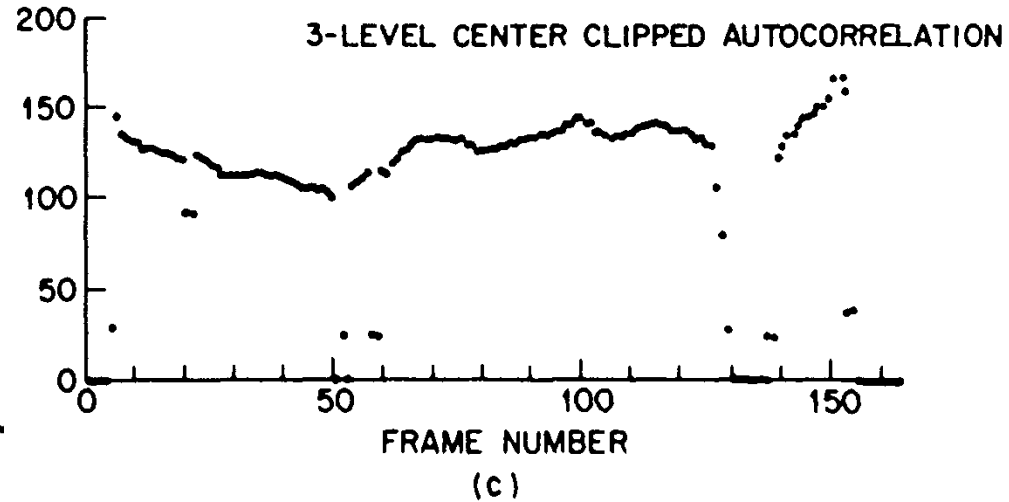
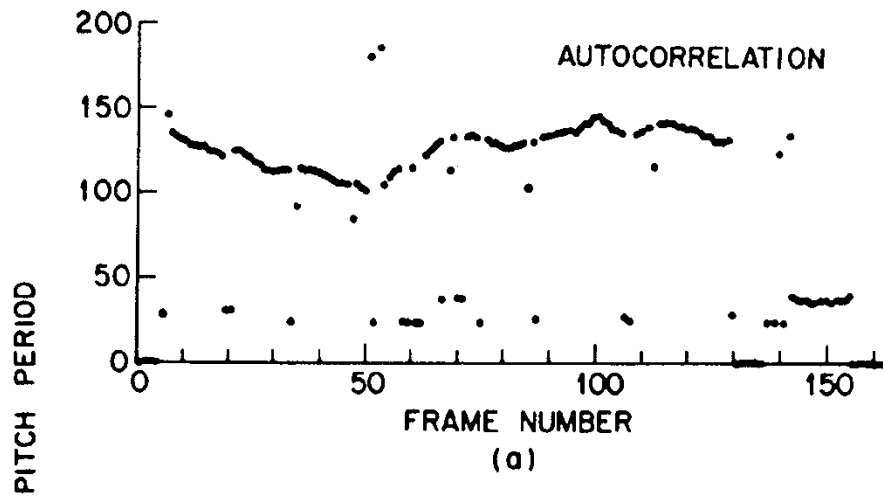
■ Median Smoother : Output = median of the input numbers
e.g. $M(23,40,58,70,120) = 58$

■ Linear Smoother (hanning filter) :

$$\begin{array}{ll} h(n) = 1/4 & n = 0 \\ & = 1/2 \quad n = 1 \\ & = 1/4 \quad n = 2 \end{array}$$

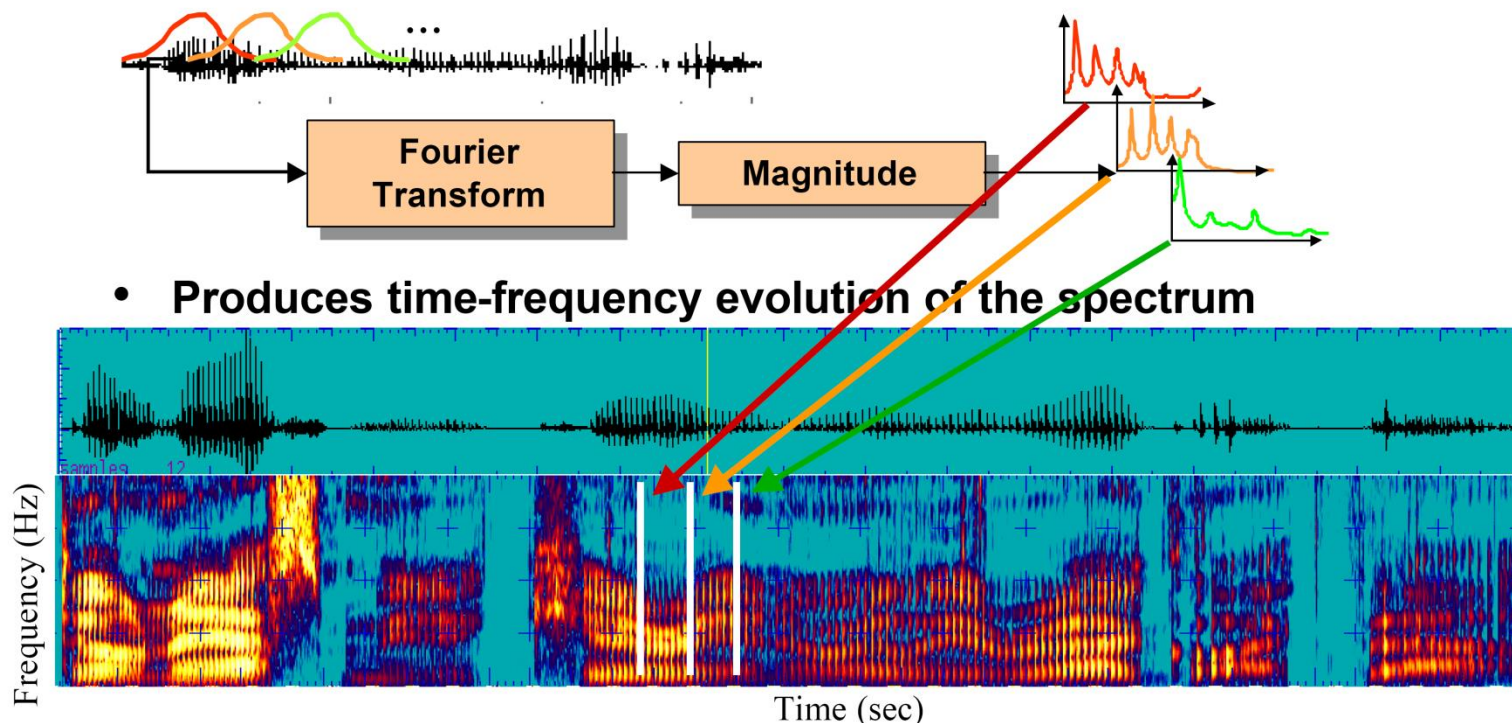
Example

WE WERE AWAY A YEAR AGO



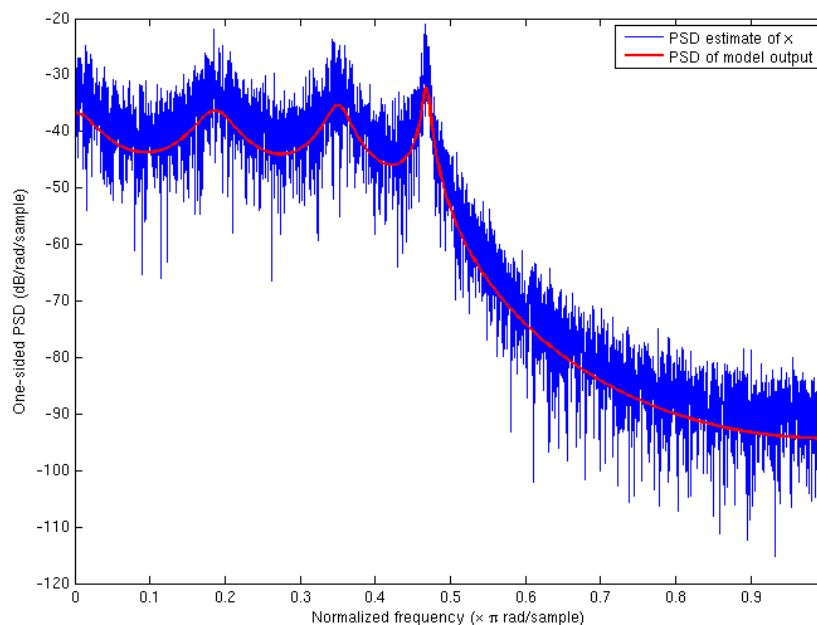
B. Frequency Domain Analysis

- Speech is a continuous evolution of the vocal tract
 - Need to extract time series of spectra
 - Use a sliding window - 16 ms window, 8 ms shift



C. Linear Predictive Coding

- A very powerful method for speech analysis
- Also known as **LPC analysis** or **auto-regressive (AR) modeling**
- Widely used because it is fast and simple
- An effective way of estimating the main parameters of speech signals



Linear Predictive Coding

- An all-pole filter with a sufficient number of poles is a good approximation for speech signals.
- We could model the filter $H(z)$ in the digital model as:

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

where p is the order of the LPC analysis. The **inverse filter** $A(z)$ is defined as

$$A(z) = 1 - \sum_{k=1}^p a_k z^{-k}$$

Taking inverse z-transforms in Eq. (1) results in:

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

Linear Predictive Coding

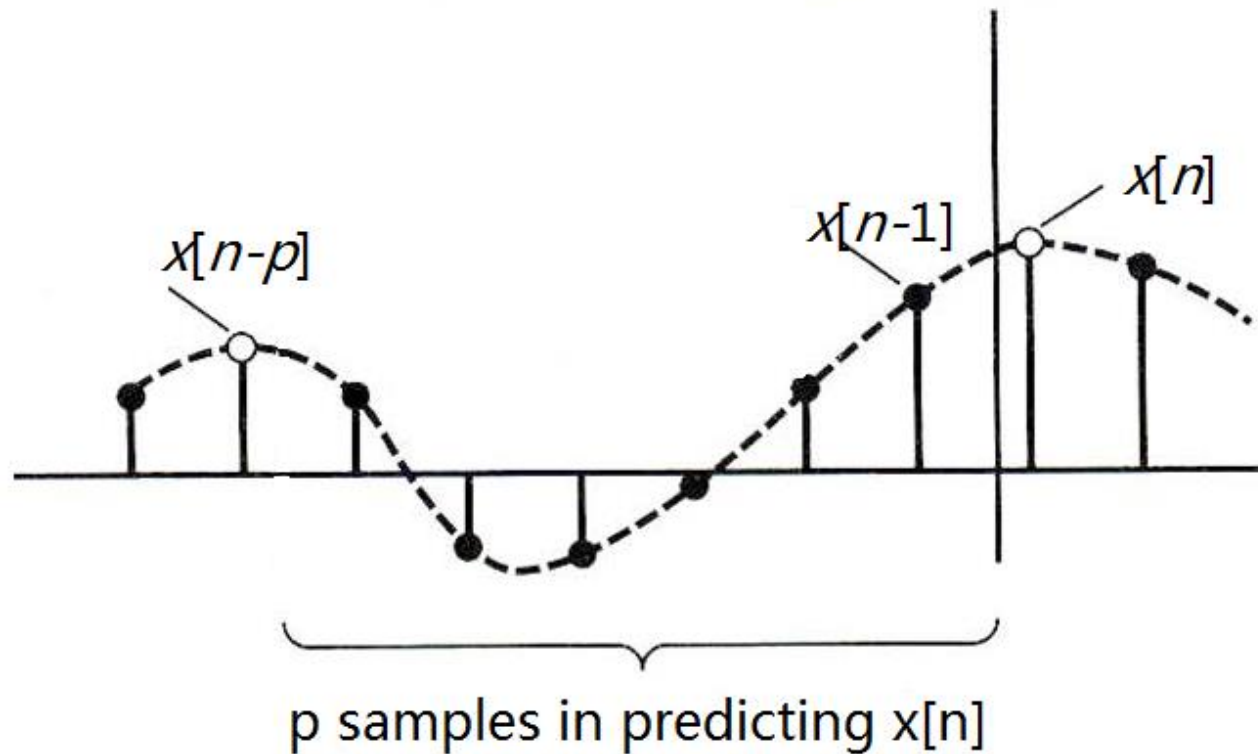
- Linear predictive coding gets its name from the fact that it predicts the current sample as a linear combination of its past p samples:

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

- The prediction error when using this approximation is (Yule-Walker Equation):

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

Linear Predictive Coding



Linear Predictive Coding

The Orthogonality Principle

- To estimate the predictor coefficients from a set of speech samples, we use the **short-time analysis** technique.
- Define $x_m[n]$ as a segment of speech selected in the vicinity of sample m :

$$x_m[n] = x[m + n]$$

- Define the short-term prediction error for that segment as

$$E_m = \sum_n e_m^2[n] = \sum_n (x_m[n] - \hat{x}_m[n])^2 = \sum_n \left(x_m[n] - \sum_{j=1}^p a_j x_m[n-j] \right)^2$$

Linear Predictive Coding

The Orthogonality Principle

- Given a signal $x_m[n]$, we estimate its corresponding LPC coefficients as those that minimize the total prediction error E_m .
- Taking the derivative of E_m with respect to a_i and equating to 0, we obtain:

$$\langle e_m, x_m^i \rangle = \sum_n e_m[n] x_m[n-i] = 0 \quad 1 \leq i \leq p$$

where we have defined e_m and x_m as vectors of samples, and their inner product has to be 0.

Solution of the LPC Equations

Autocorrelation Method

- In the autocorrelation method, we assume that $x_m[n]$ is 0 outside the interval $0 \leq n < N$:

$$x_m[n] = x[m+n]w[n]$$

with $w[n]$ being a window which is 0 outside the interval $0 \leq n < N$

- The corresponding prediction error $e_m[n]$ is non-zero over the interval $0 \leq n < N+p$, and the total prediction error :

$$E_m[n] = \sum_{n=0}^{N+p-1} e_m^2[n]$$

Solution of the LPC Equations

Autocorrelation Method

■ Solve the following matrix equation

$$\sum_{j=1}^p a_j R_m[|i - j|] = R_m[i]$$

$$\begin{pmatrix} R_m[0] & R_m[1] & R_m[2] & \dots & R_m[p-1] \\ R_m[1] & R_m[0] & R_m[1] & \dots & R_m[p-2] \\ R_m[2] & R_m[1] & R_m[0] & \dots & R_m[p-3] \\ \dots & \dots & \dots & \dots & \dots \\ R_m[p-1] & R_m[p-2] & R_m[p-3] & \dots & R_m[0] \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \dots \\ a_p \end{pmatrix} = \begin{pmatrix} R_m[1] \\ R_m[2] \\ R_m[3] \\ \dots \\ R_m[p] \end{pmatrix}$$

Solution of the LPC Equations

Autocorrelation Method

■ The matrix in the previous equation is symmetric and all the elements in its diagonals are identical. (**Toeplitz**)

■ Durbin's recursion:

1. Initialization $E^0 = R[0]$

2. Iteration. For $i = 1, \dots, p$ do the following recursion

$$k_i = \left(R[i] - \sum_{j=1}^{i-1} a_j^{i-1} R[i-j] \right) / E^{i-1}$$

$$a_i^i = k_i$$

$$a_j^i = a_j^{i-1} - k_i a_{i-j}^{i-1}, \quad 1 \leq j \leq i$$

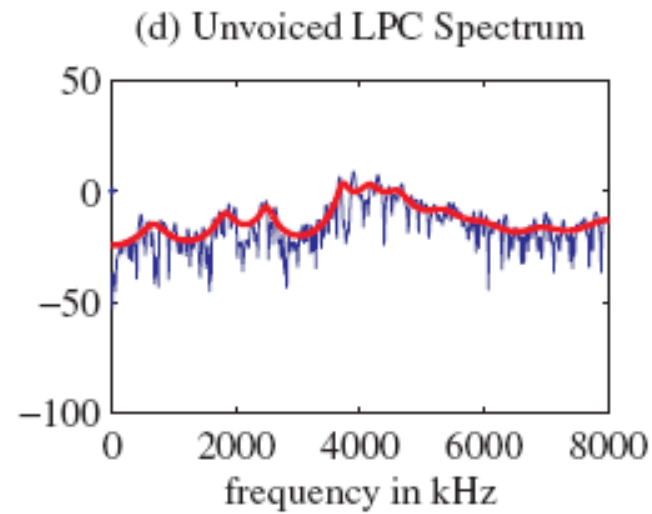
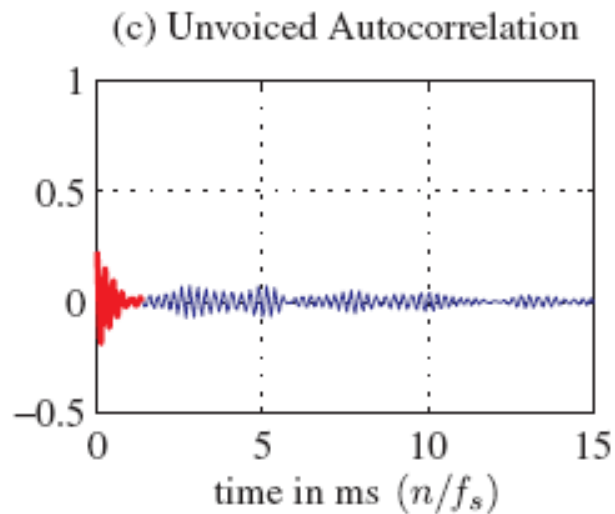
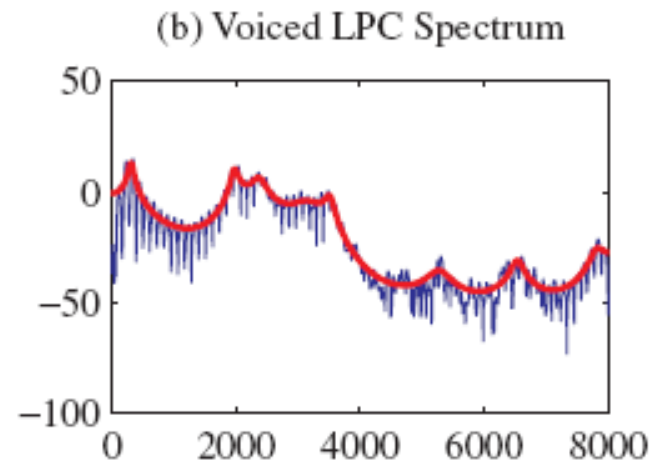
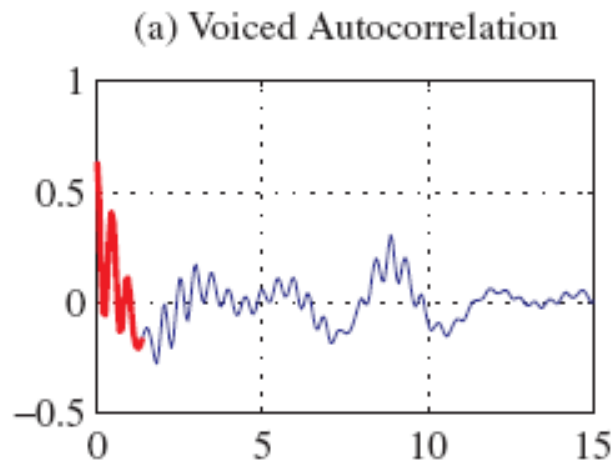
$$E^i = (1 - k_i^2) E^{i-1}$$

3. Final solution:

$$a_j = a_j^p \quad 1 \leq j \leq p$$

where k_j (reflection coefficients) are bounded between -1 and 1.

A Real Example of Autocorrelation & LPC



D. Homomorphic Analysis

- A homomorphic transformation $x[n]=D(x[n])$ is a transformation that converts a convolution

$$x[n] = e[n] * h[n]$$

into a sum

$$x[n] = \hat{e}[n] + h[n]$$

Homomorphic Analysis of Speech Signal

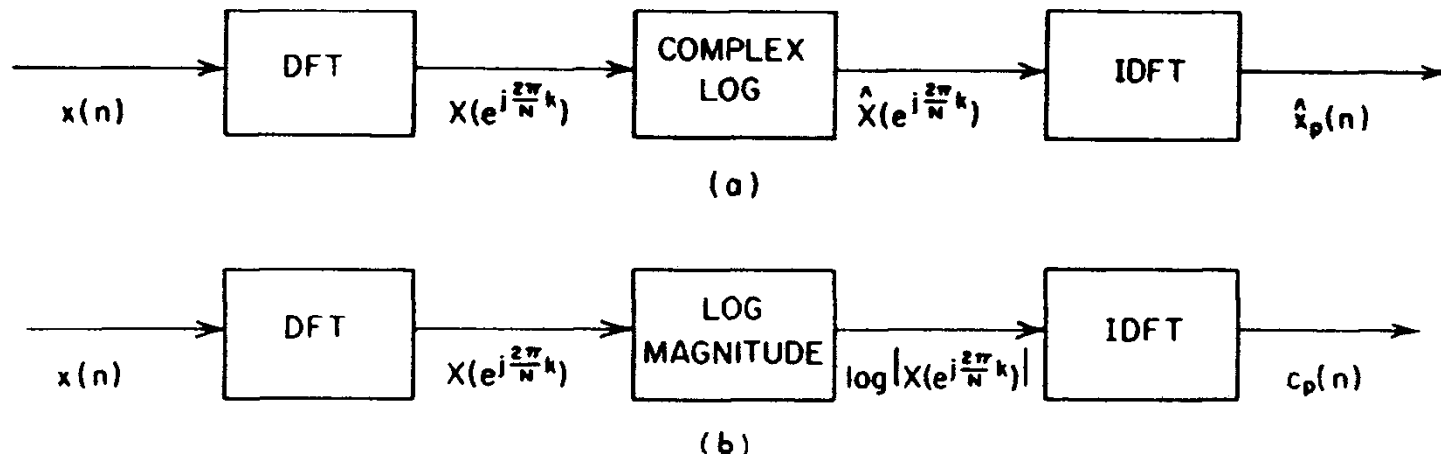
■ Voiced Speech :

$$s(n) = p(n) * g(n) * v(n) * r(n)$$

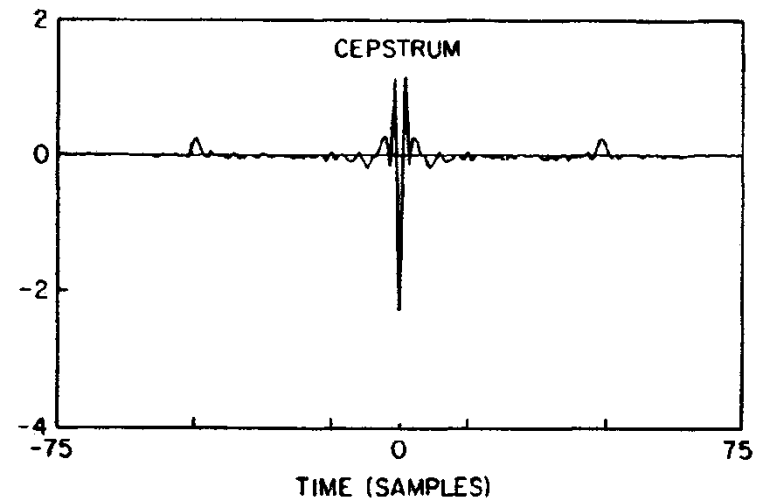
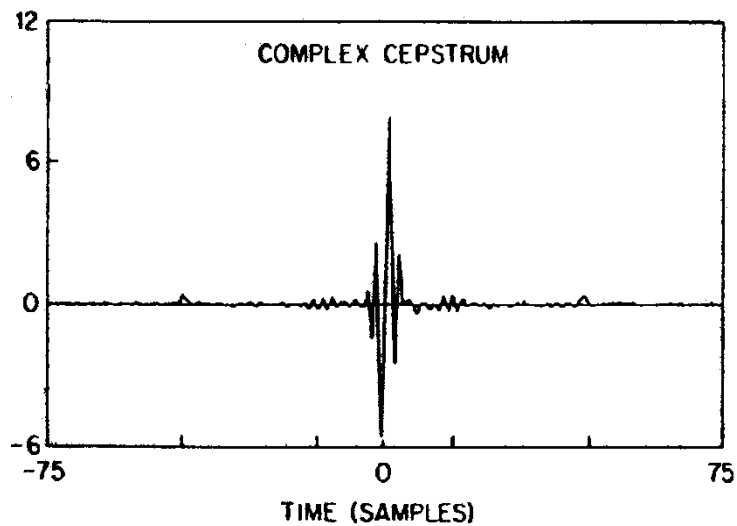
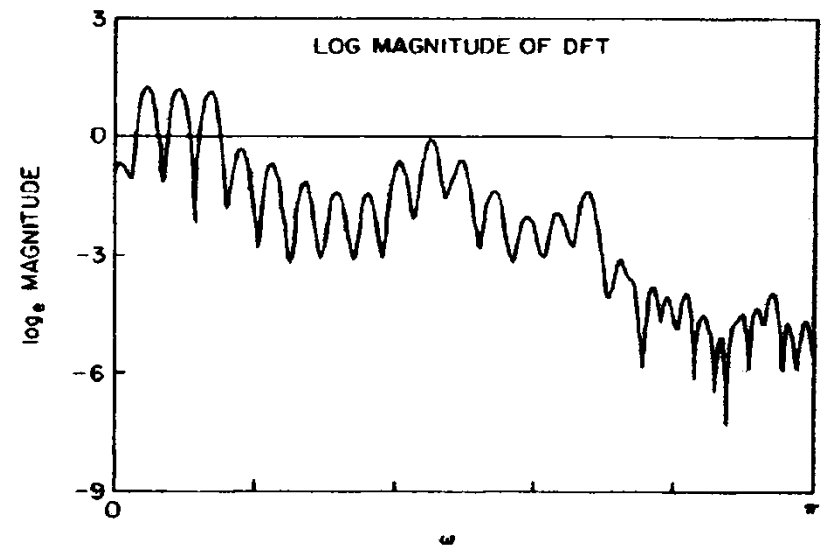
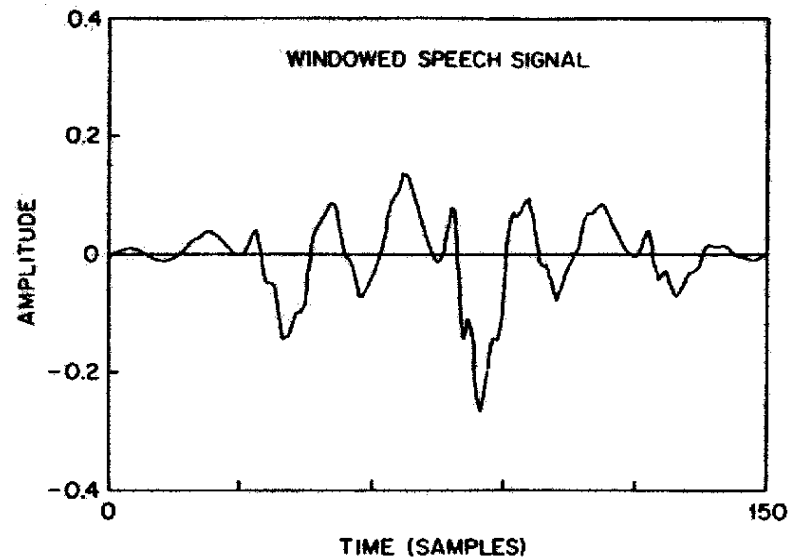
■ Unvoiced Speech :

$$s(n) = u(n) * v(n) * r(n)$$

→ Implementation :



Cepstrum

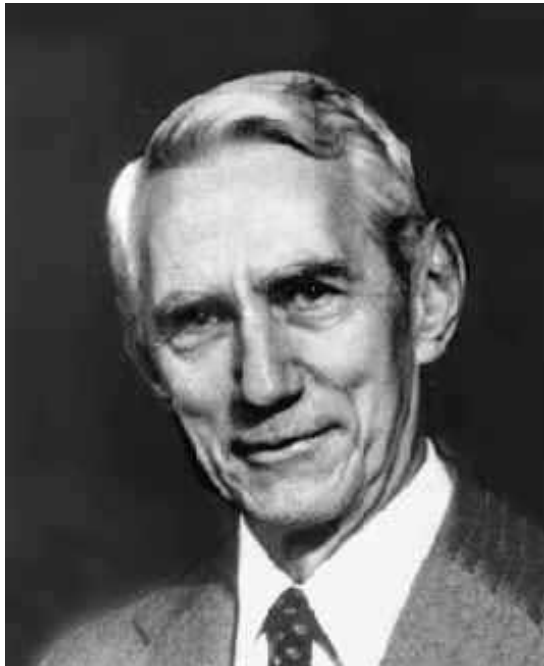


Part 3- Speech Coders

- Reduction in bit rate is the primary purpose of speech coding
 - Previous bit stream can be compressed to a lower rate by removing redundancy in the signal
 - *Lossless compression* : original signal can be recovered exactly
 - *Lossy compression* : original signal can not be recovered exactly

Claude Elwood Shannon

(April 30, 1916 – February 24, 2001) was an American mathematician, electronic engineer, and cryptographer known as "the father of information theory".

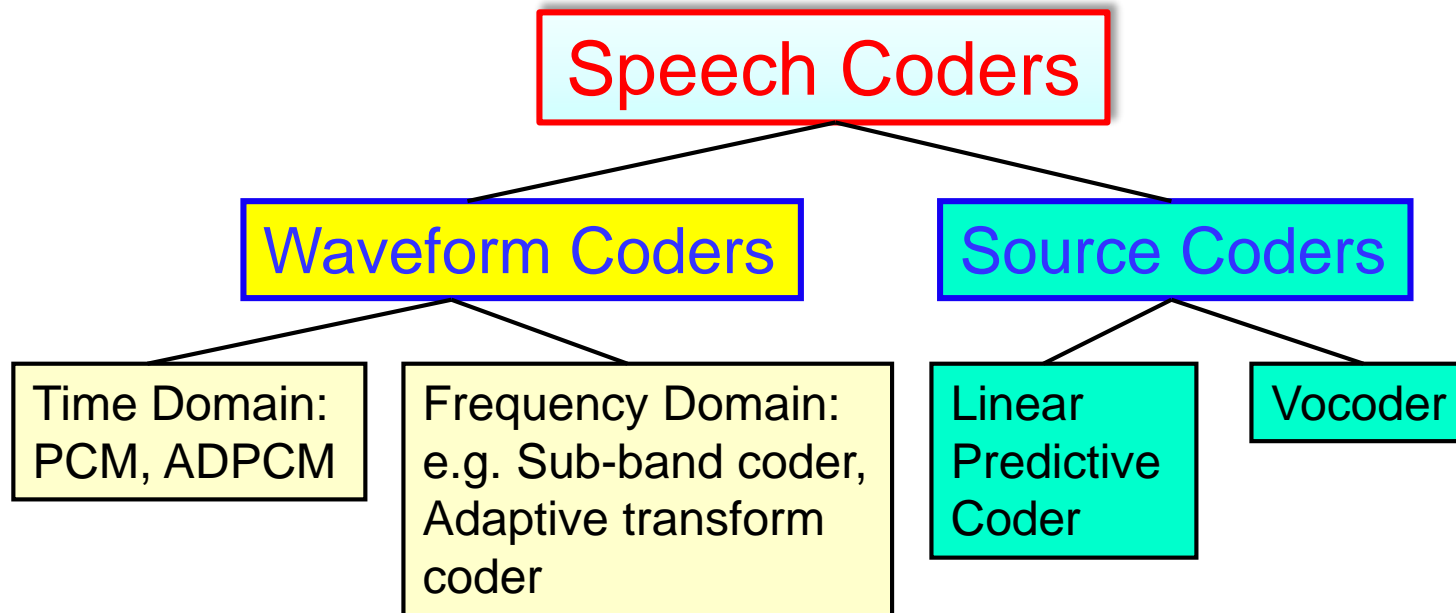


Harry Nyquist

(February 7, 1889 – April 4, 1976) was an important contributor to communication theory.



Taxonomy of Speech Coders



Measure of Quality

- Bit rate and quality are intimately related
 - lower the bit rate, lower the quality
- While the bit rate is inherently a number, the most widely used measure of quality is the *Mean Opinion Score* (MOS)

Excellent	Good	Fair	Poor	Bad
5	4	3	2	1

- Another measure of quality is the *signal-to-noise* (SNR), defined as the ratio between the signal's energy & noise's energy in terms of dB :

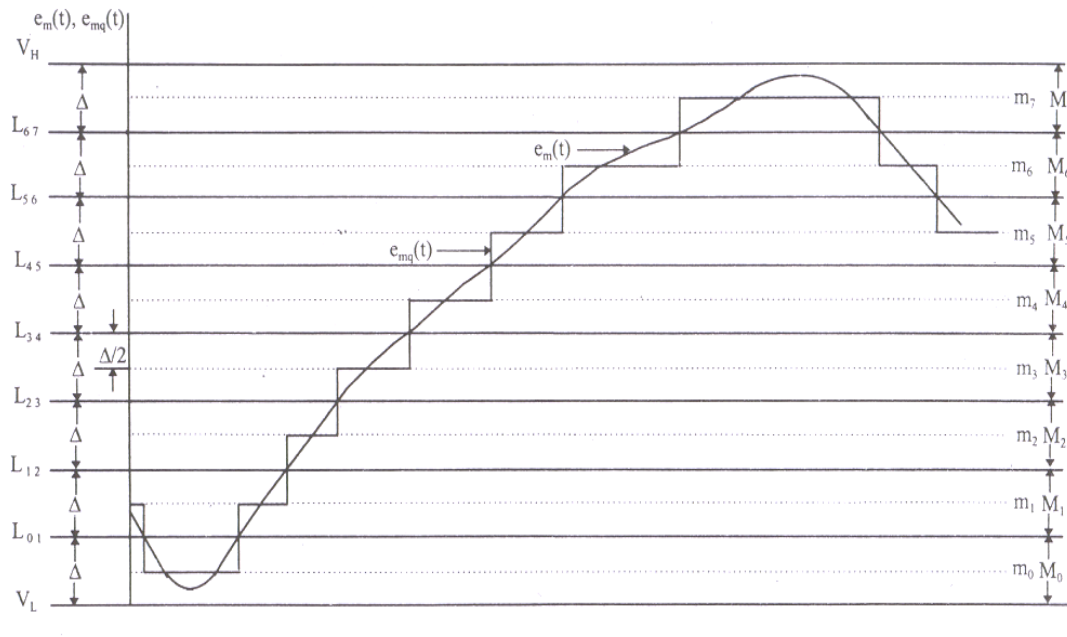
$$SNR = \frac{\sigma_x^2}{\sigma_e^2} = \frac{E\{x^2[n]\}}{E\{e^2[n]\}}$$

A. Scalar Waveform Coders

- Several waveform coding techniques
 - Linear PCM
 - μ -law
 - A-law PCM
 - APCM
 - DPCM
 - DM

Linear Pulse Code Modulation(PCM)

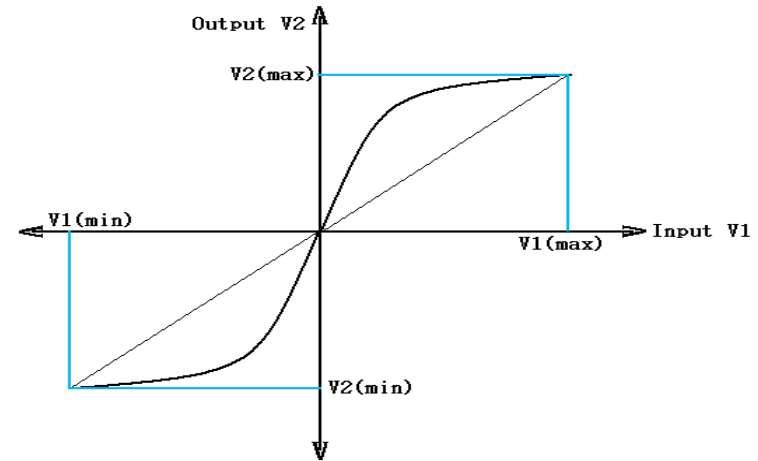
- Analog-to digital converters perform both sampling and quantization simultaneously
- With B bits, it is possible to represent 2^B separate quantization levels.
- The output of the quantizer $x[n]$ is given by $x[n] = Q\{x[n]\}$



μ -law and A-law PCM

■ μ -law

$$y[n] = X_{\max} \frac{\log[1 + \mu \frac{|x[n]|}{X_{\max}}]}{\log[1 + \mu]} \text{sign}\{x[n]\}$$



■ A-law

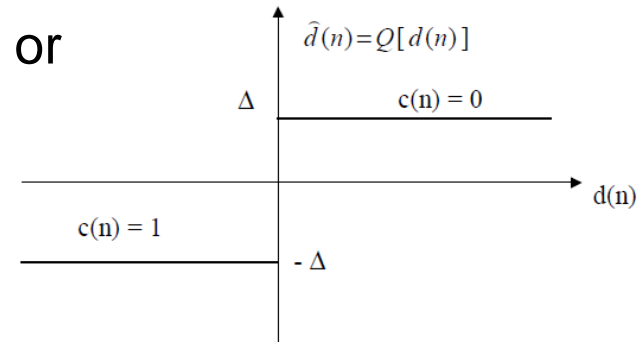
$$y[n] = \begin{cases} \frac{A}{1 + \log A} \frac{x[n]}{X_{\max}}, & 0 \leq \frac{|x[n]|}{X_{\max}} \leq \frac{1}{A} \\ X_{\max} \frac{1 + \log[\frac{A|x[n]|}{X_{\max}}]}{1 + \log A} \text{sign}\{x[n]\}, & \frac{1}{A} < \frac{|x[n]|}{X_{\max}} \leq 1 \end{cases}$$

■ Typical values in practice: $\mu = 255$, $A = 87.6$

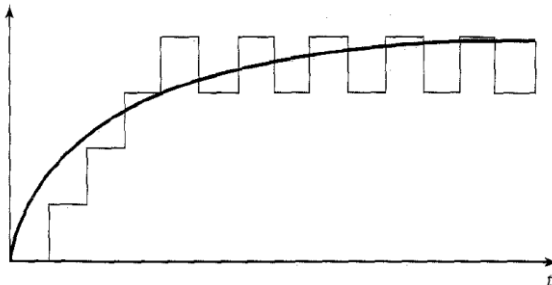
Delta Modulation(DM)

- DM is a 1-bit DPCM, which predicts the current sample to be the same as the past sample
- Transmit whether the current sample is above or below the previous sample $x[n] = x[n-1]$

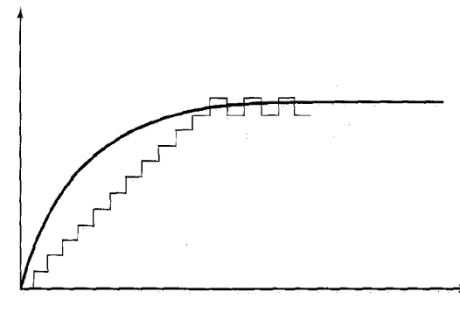
$$d[n] = \begin{cases} \Delta & x[n] > x[n-1] \\ -\Delta & x[n] \leq x[n-1] \end{cases} \quad \Delta \text{ is the step size}$$



- If Δ is too small, the reconstructed signal will not increase as fast as the original signal → **slope overload distortion**
- When the slope is small, the step size Δ also determines the peak error → **granular noise**



Large Δ and Granular noise

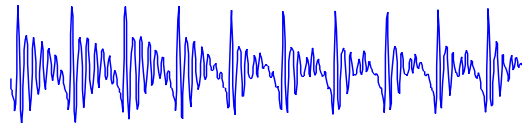


Small Δ and slope overload distortion

B. Vocoder

Voice Coder/Decoder

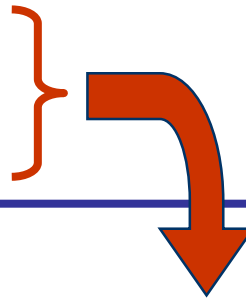
Encoder



Original Speech

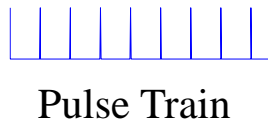
Analysis:

- Voiced/Unvoiced decision
- Pitch Period (voiced only)
- Signal power (Gain)



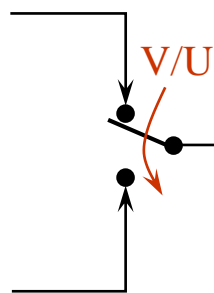
Decoder

Pitch
Period



Pulse Train

Signal Power

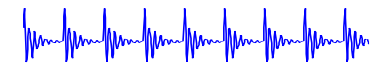


G

Vocal Tract
Model

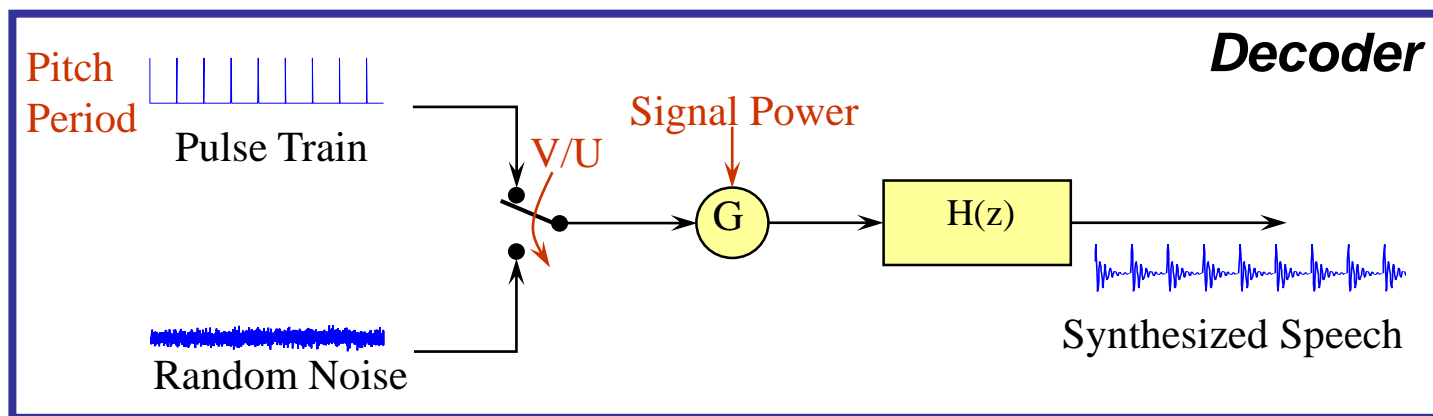


Random Noise



Synthesized Speech

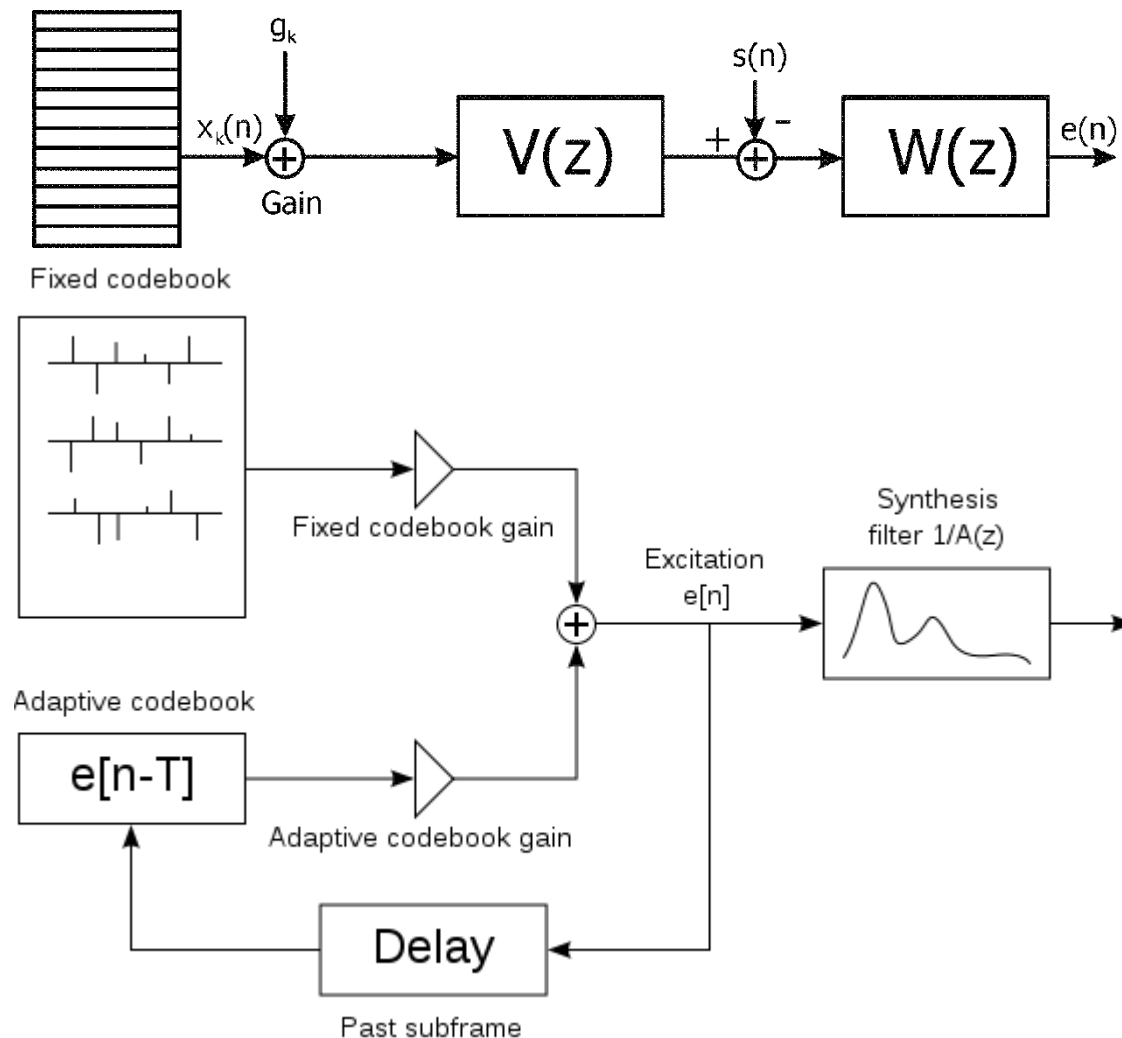
LPC Vocoder



- Filter $h_m[n]$ for frame m changes at regular intervals. If this filter is represented with linear predictive coding, it is called an LPC vocoder
- In addition to transmitting the gain and LPC coefficients, the encoder has to determine whether the frame is voiced or unvoiced, as well as the pitch period P for voiced frames

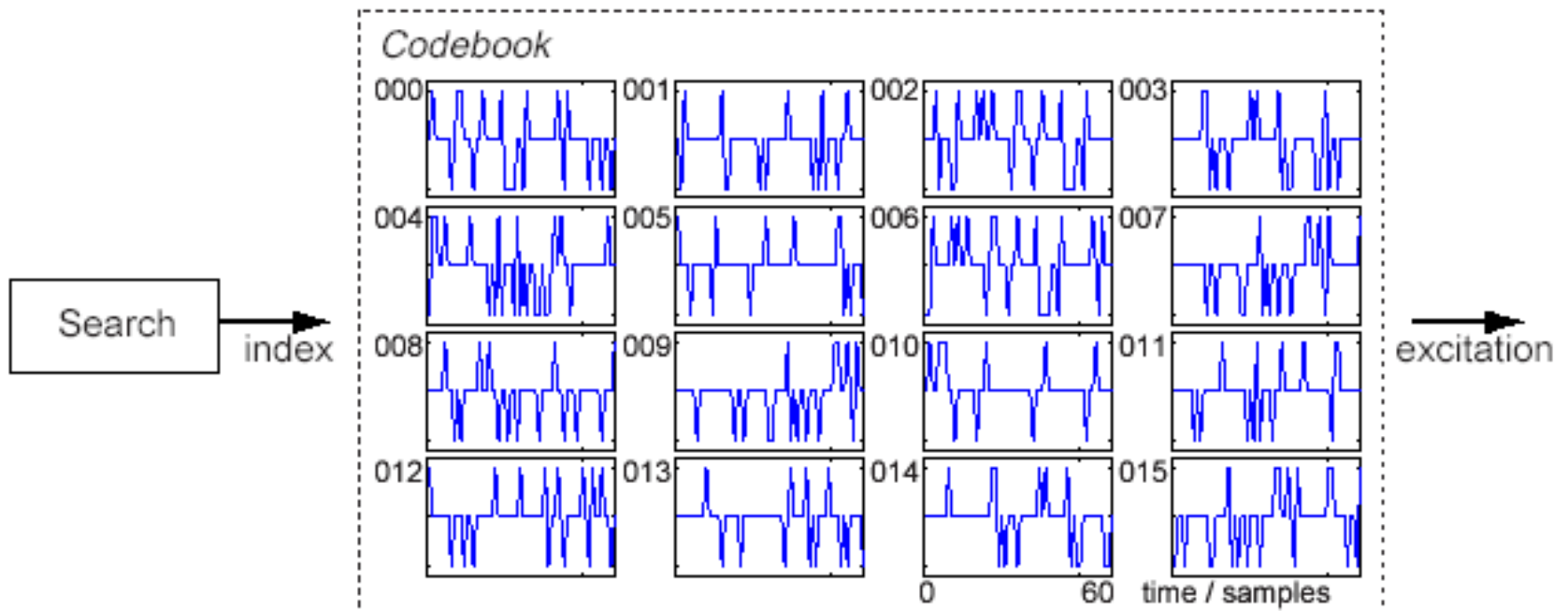
Code Excited Linear Prediction(CELP)

Analysis-by-synthesis principle

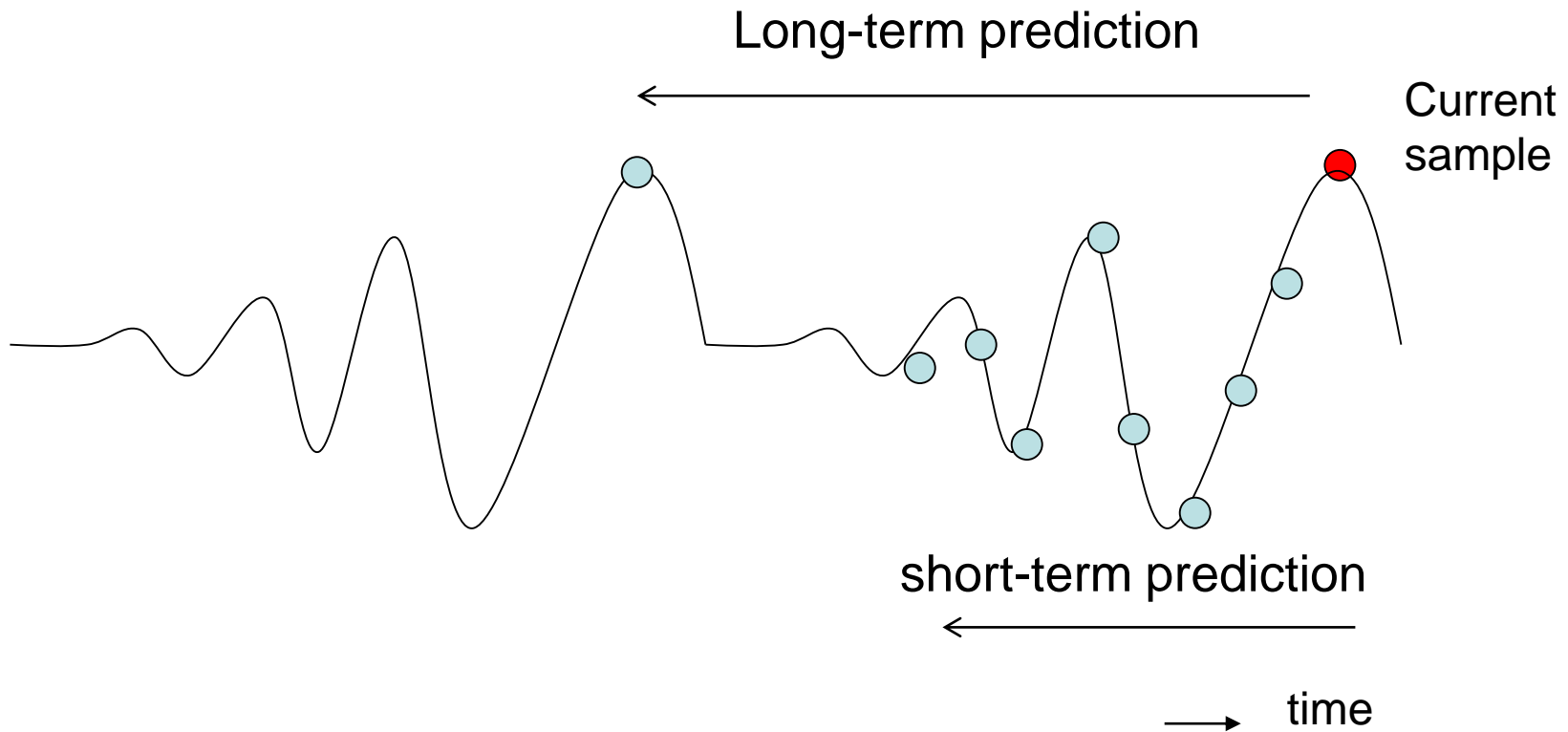


CELP

- Represent excitation with codebook
 - e.g. 512 sparse excitation vectors
 - linear search for minimum weighted error?



Prediction



- Short-term - resonance of vocal tract
- Long-term - periodicity of voiced speech (vocal cord vibration)

Part 4- Speech Synthesis

- Four different families of speech generation approaches
 - Limited-domain waveform concatenation
 - Concatenative synthesis with no waveform modification
 - Concatenative systems with waveform modification
 - Rule-based systems

A. Concatenative Speech Synthesis

- A speech segment is synthesized by simply playing back a waveform with matching phoneme string
- An utterance is synthesized by concatenating together several speech fragments
- Speech segments are greatly affected by coarticulation
- There can be spectral or prosodic discontinuities:
 - Spectral discontinuities occur when the formants at the concatenation point do not match
 - Prosodic discontinuities occur when the pitch at the concatenation point does not match

Choice of Unit

- The unit should lead to low concatenation distortion
- The unit should lead to low prosodic distortion
- The unit should be generalizable
- The unit should be trainable

Unit length	Unit type	#Units	Quality
<div>Short</div> <div>↓</div> <div>Long</div>	Phoneme	42	<div>Low</div> <div>↓</div> <div>High</div>
	Diphone	~1500	
	Triphone	~30K	
	Demisyllable	~2000	
	Syllable	~15K	
	Word	100K–1.5M	
	Phrase	∞	
	Sentence	∞	

Optimal Unit String

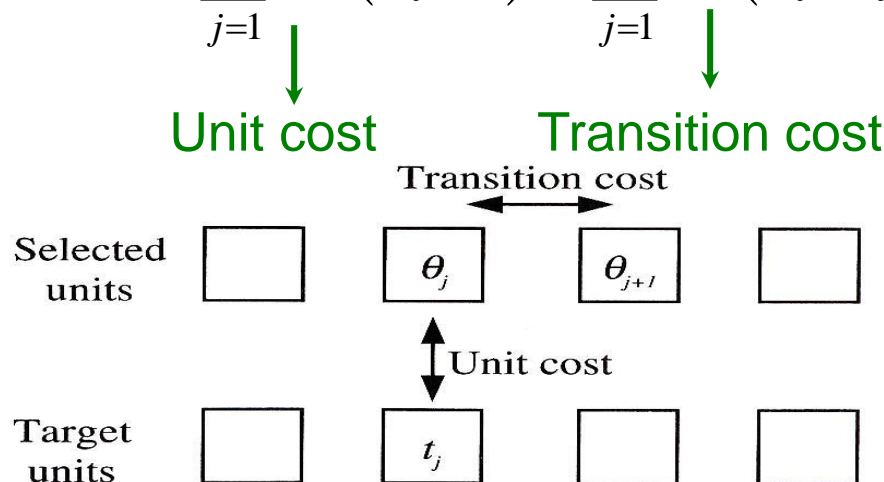
The Decoding Process

- Choose the optimal string of units for a given phonetic string that best matches the desired prosody
- The quality of a unit string is typically dominated by spectral and pitch discontinuities at unit boundaries.
- Discontinuities can occur because of
 - Differences in phonetic contexts
 - Incorrect segmentation
 - Acoustic variability
 - Different prosody

Objective Function

- Cost function between the segment concatenation Θ and the target T

$$C(\Theta, T) = \sum_{j=1}^N C_u(\theta_j, T) + \sum_{j=1}^{N-1} C_t(\theta_j, \theta_{j+1})$$



- The optimal speech segment sequence of units $\hat{\Theta}$ can be found as the one that minimizes the overall cost

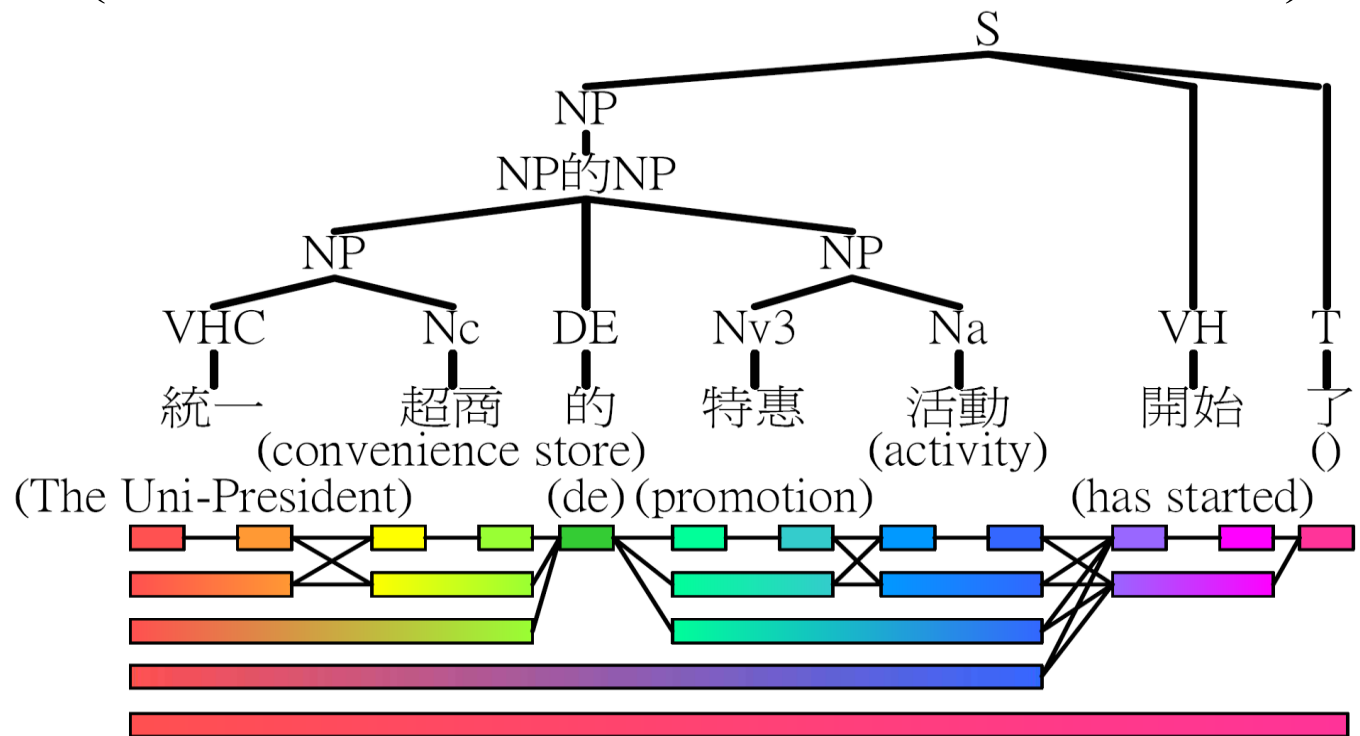
$$\hat{\Theta} = \arg \min_{\Theta} C(\Theta, T)$$

Variable-Length Unit Selection

- Minimize the **substitution (unit)** and **concatenation (transition)** cost

$$\hat{u}_{1:N} = \arg \min_{u'_{1:N}} \left(C_{Sub}(u_1, u'_1) + C_{Con}(u'_1, u'_2) + C_{Sub}(u_2, u'_2) + C_{Con}(u'_2, u'_3) + \dots + C_{Con}(u'_{N-1}, u'_N) + C_{Sub}(u_N, u'_N) \right)$$

Syntactic
Structure





Search
Lattice

Demo

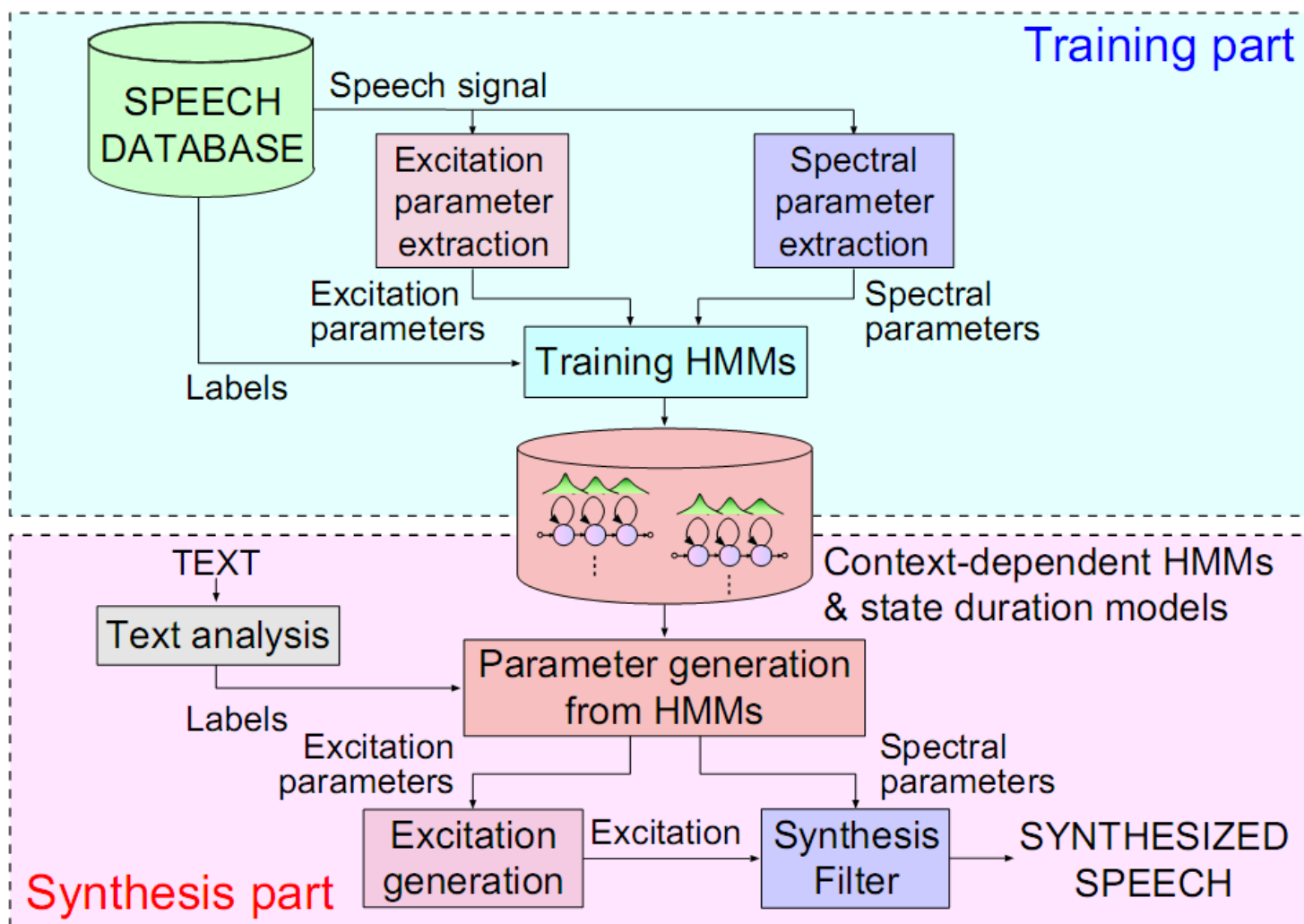
Unit Selection

● 訊飛語音合成技術--中國安徽科技大學 

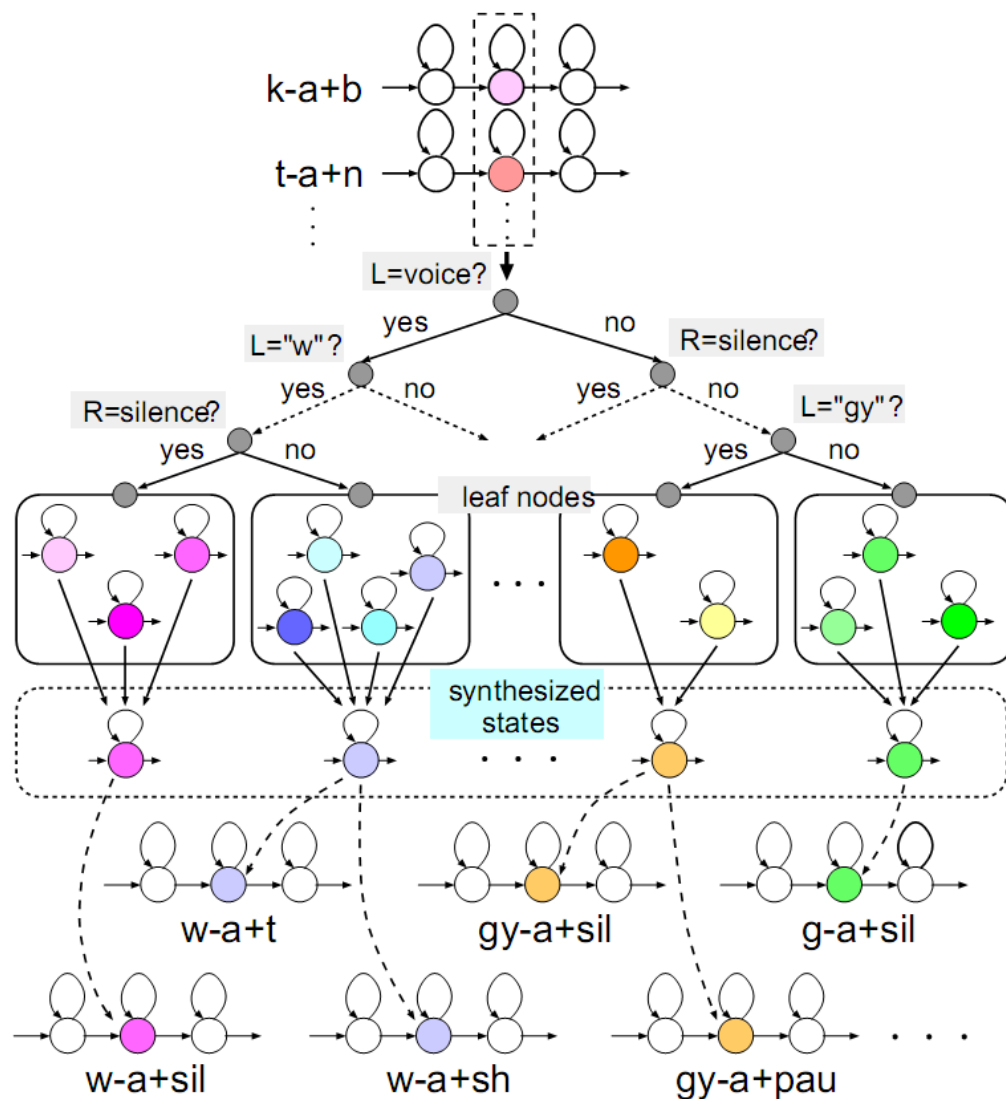
● Proposed approach

- 中秋闔家團圓，邊賞月邊烤肉已經變成台灣的習俗。 
- 藝術家在這裡專心創作，期望把台灣藝術推向國際。 

B. HMM-Based Speech Synthesis



Decision tree-based state clustering



Speech Parameter Generation

- For given HMM λ , determine a speech parameter vector sequence $\mathbf{O} = [\mathbf{o}_1^T, \mathbf{o}_2^T, \dots, \mathbf{o}_T^T]^T$, which maximizes





$$P(\mathbf{O} | \lambda) = \sum_{\mathbf{Q}} P(\mathbf{O} | \mathbf{Q}, \lambda) P(\mathbf{Q} | \lambda)$$
$$\approx \max_{\mathbf{Q}} P(\mathbf{O} | \mathbf{Q}, \lambda) P(\mathbf{Q} | \lambda)$$



$$\mathbf{Q}_{\max} = \arg \max_{\mathbf{Q}} P(\mathbf{Q} | \lambda)$$

$$\mathbf{O}_{\max} = \arg \max_{\mathbf{O}} P(\mathbf{O} | \mathbf{Q}_{\max}, \lambda)$$

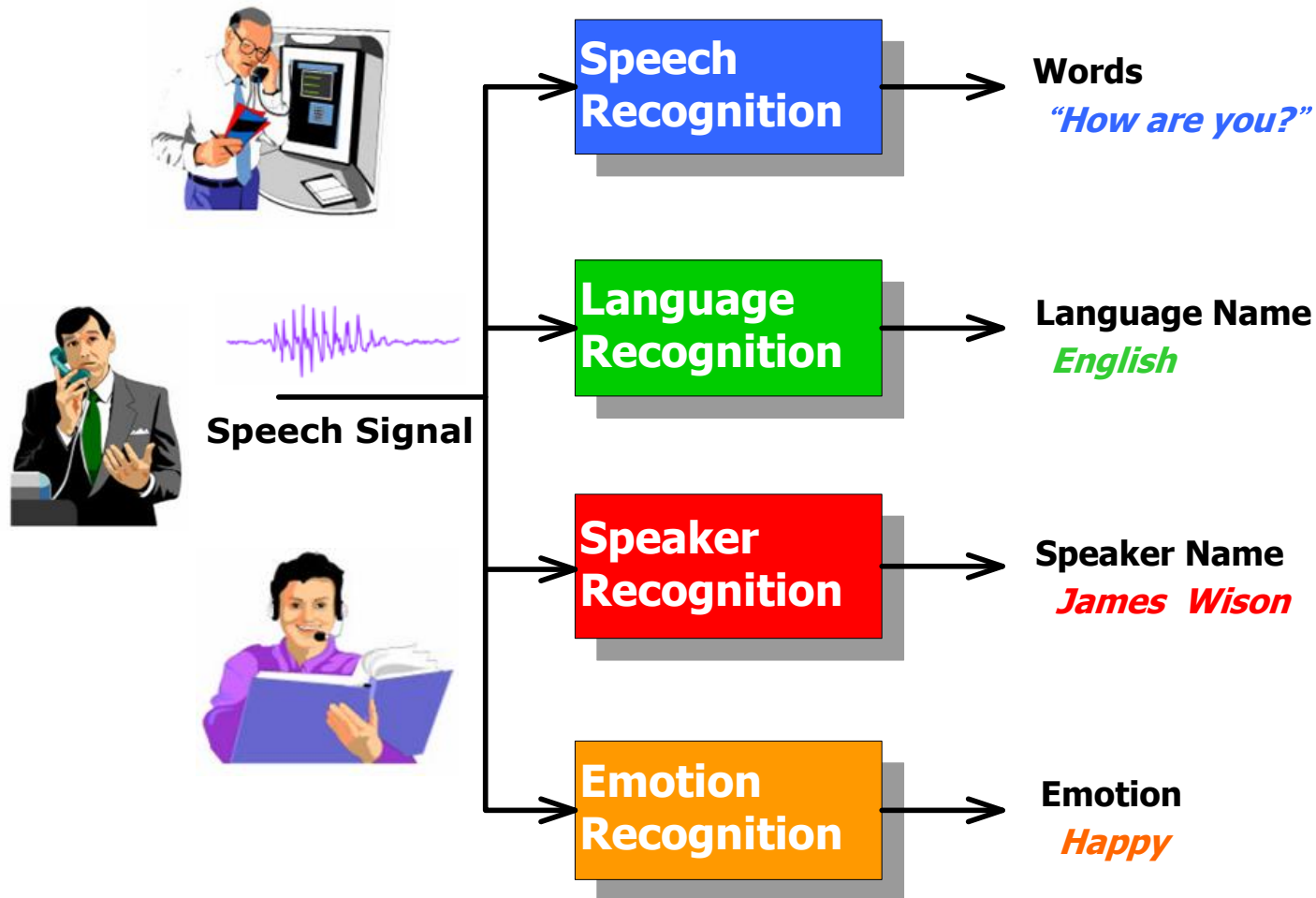
Speaker Morphing --- Demo

- 放眼整片海洋，章魚的智慧可是數一數二，科學家發現，章魚能夠獨自解決複雜的問題，像是以觸角來扭開罐頭，具有所謂的概念智力。
- Original voices: Female  Male 
- Female to male: 
- Male to female: 

Voice Morphing

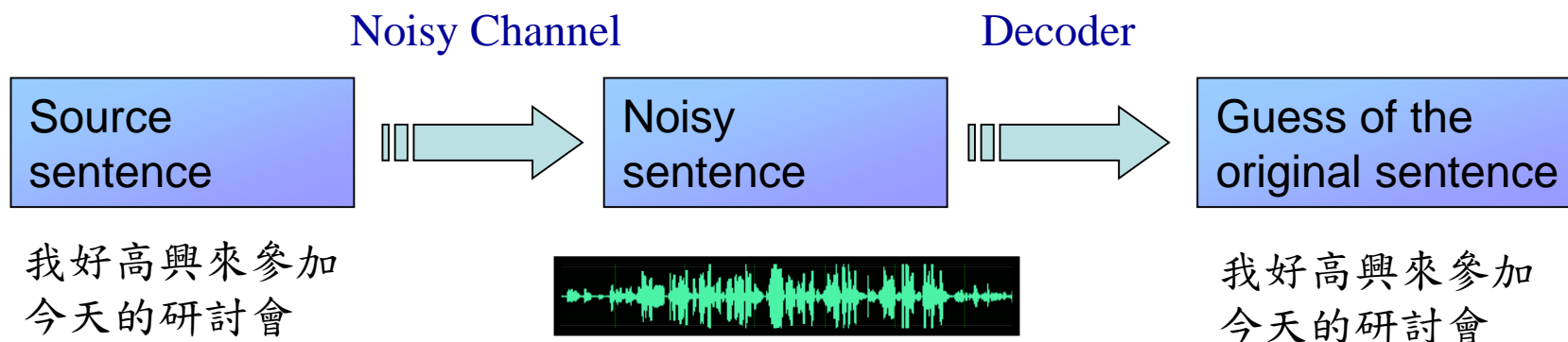
■ Demo

Part 5- Automatic Speech Recognition

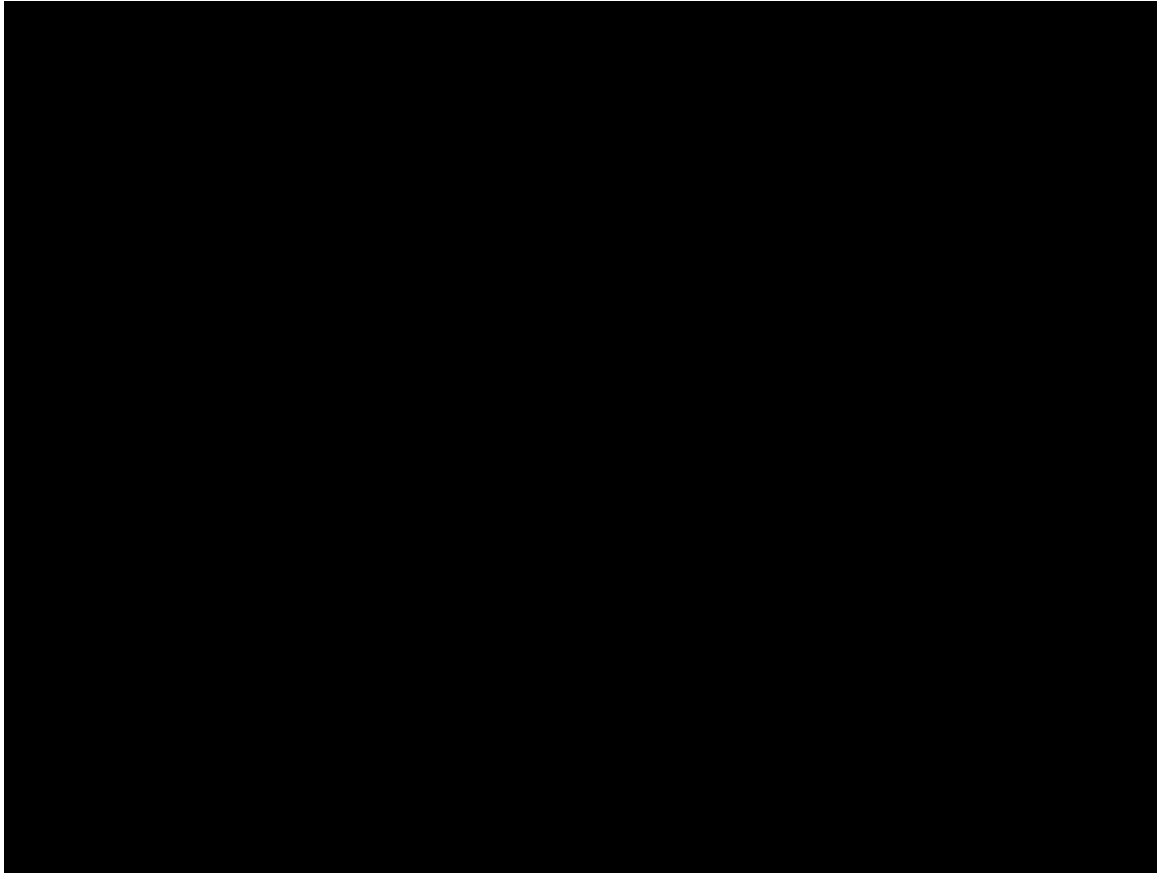


The Noisy Channel Model

- **Automatic speech recognition** (ASR) is a process by which an acoustic speech signal is converted into a set of words [Rabiner et al., 1993]
- **The noisy channel model** [Lee et al., 1996]
 - Acoustic input considered a noisy version of a source sentence



McGurk Effect



The Noisy Channel Model

- *What is the most likely sentence out of all sentences in the language given some acoustic input O ?*
- Treat acoustic input O as sequence of individual observations
 - $O = o_1, o_2, o_3, \dots, o_t$
- Define a sentence as a sequence of words:
 - $W = w_1, w_2, w_3, \dots, w_n$

$$\hat{W} = \arg \max_{W \in L} P(W | O)$$

$$\hat{W} = \arg \max_{W \in L} \frac{P(O | W)P(W)}{P(O)}$$

$$\hat{W} = \arg \max_{W \in L} P(O | W)P(W)$$

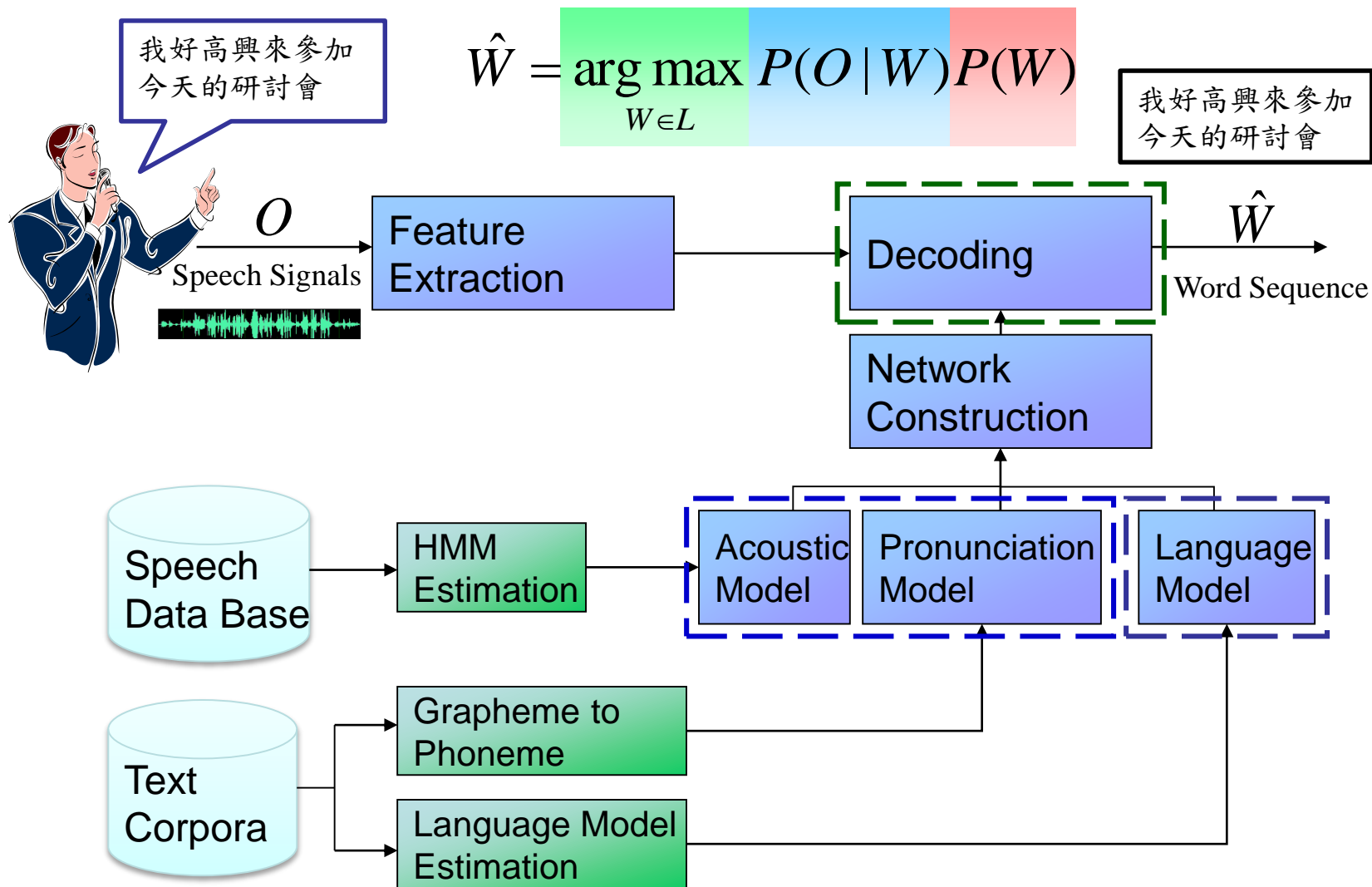


Bayes rule

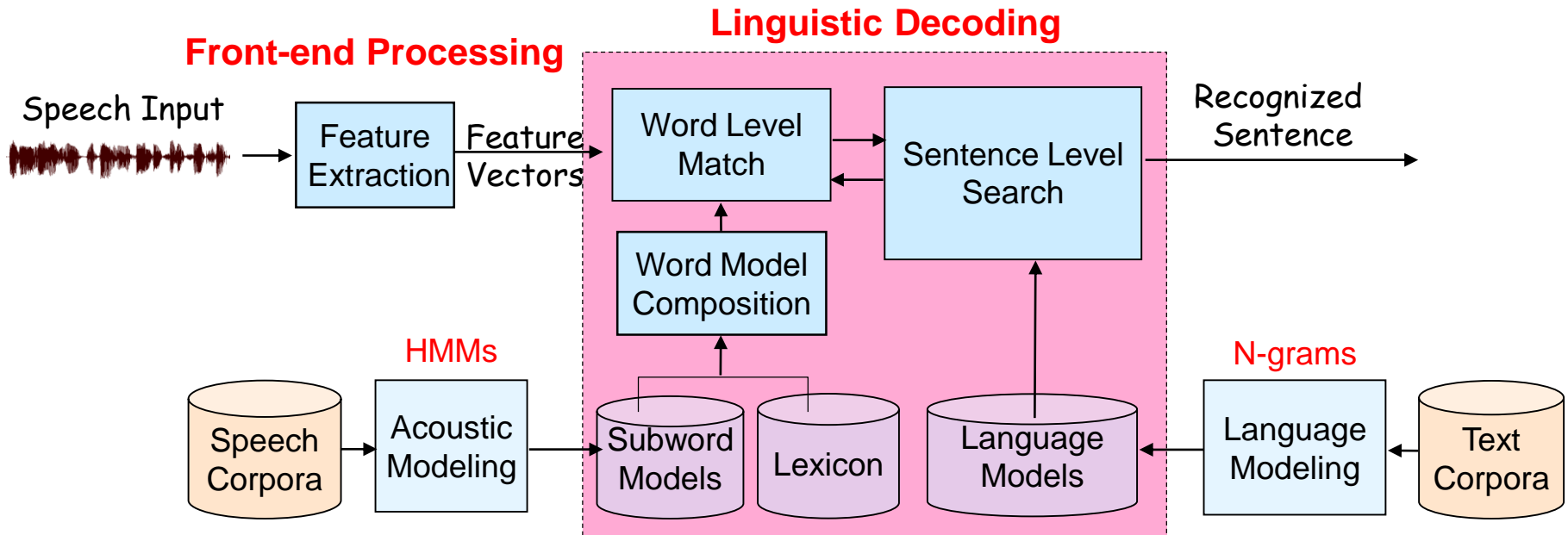


Golden rule

Speech Recognition Architecture



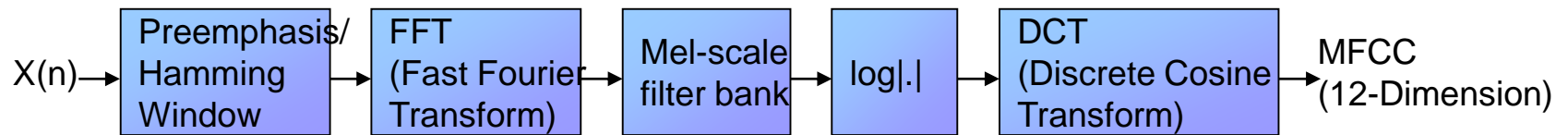
Large Vocabulary Continuous Speech Recognition



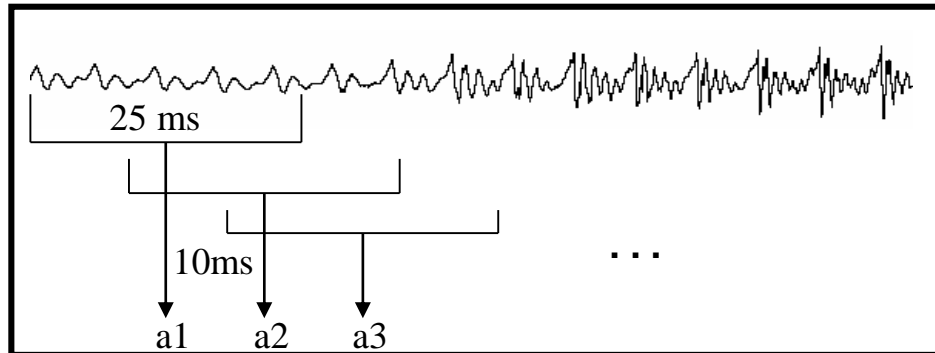
- Front-end Processing (前端處理) is a spectral analysis (頻譜分析) that derives feature vectors to capture salient spectral characteristics of speech input
- Linguistic decoding (語言解碼) combines word-level matching and sentence-level search to perform an inverse operation to decode the message from the speech waveform

Feature Extraction

- The *Mel-Frequency Cepstrum Coefficients* (MFCC) is a popular choice [Paliwal, 1992]



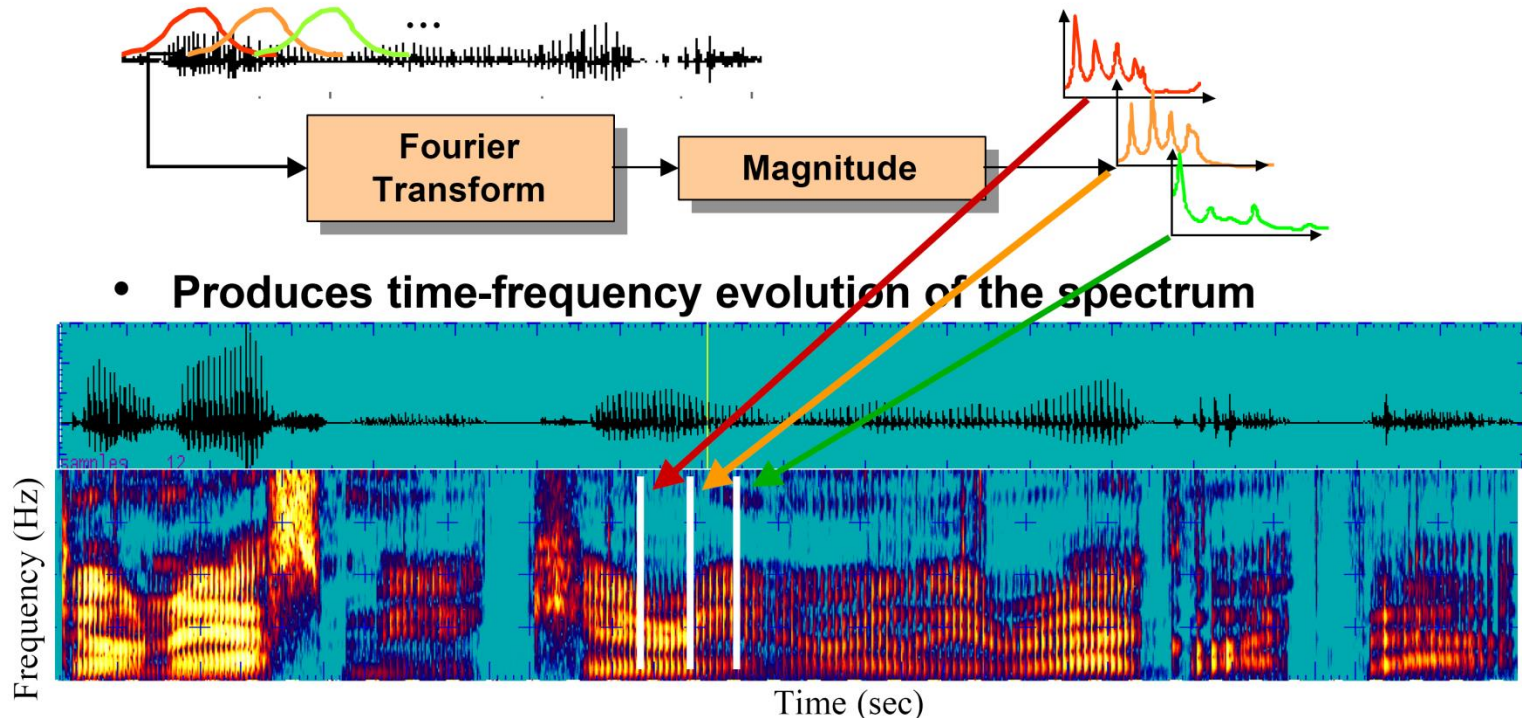
- Frame size : 25ms / Frame rate : 10ms



- 39 feature per 10ms frame
 - Absolute : Log Frame Energy (1) and MFCCs (12)
 - Delta : First-order derivatives of the 13 absolute coefficients
 - Delta-Delta : Second-order derivatives of the 13 absolute coefficients

Feature Extraction

- Speech is a continuous evolution of the vocal tract
 - Need to extract time series of spectra
 - Use a sliding window - 16 ms window, 8 ms shift



Selection of Fundamental Speech Units

- Phoneme-like units (PLUs)
 - The simplest set of fundamental speech units
 - /s/ /ɛ/ /g/ /m/ /ə / /n/ /t/
- Units other than Phones
 - Syllables, /seg/ /men/ /t/
 - Demisyllables, /sɛ/ /ɛ g/ /mə / /ə n/ /t/
- Units with Linguistic Context Dependency
 - triphones

Acoustic Model

■ Provide $P(O|Q) = P(\text{features}|\text{phone})$

■ Modeling Units [Bahl et al., 1986]

- Context-independent : Phoneme
- Context-dependent : Diphone, Triphone, Quinphone
 - $p_L\text{-}p\text{-}p_R$: left-right context triphone

■ Typical acoustic model [Juang et al., 1986]

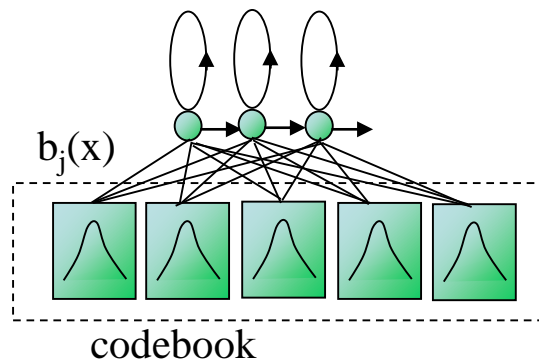
– Continuous-density Hidden Markov Model

$$\lambda = (A, B, \pi)$$

– Distribution : Gaussian Mixture

$$b_j(x_j) = \sum_{k=1}^K c_{jk} N(x_t; \mu_{jk}, \Sigma_{jk})$$

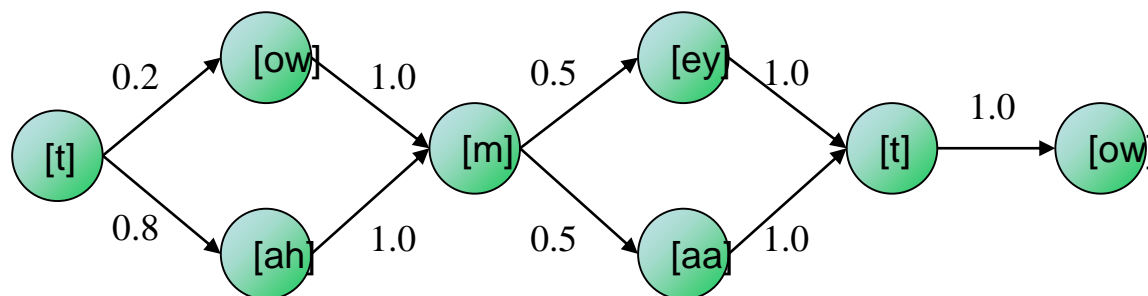
– HMM Topology : 3-state left-to-right model for each phone, 1-state for silence or pause



Pronunciation Model

- Provide $P(Q|W) = P(\text{phone}|\text{word})$
- Word Lexicon [Hazen et al., 2002]
 - Map legal phone sequences into words according to phonotactic rules
 - G2P (Grapheme to phoneme) : Generate a word lexicon automatically
 - Several word may have multiple pronunciations
- Example

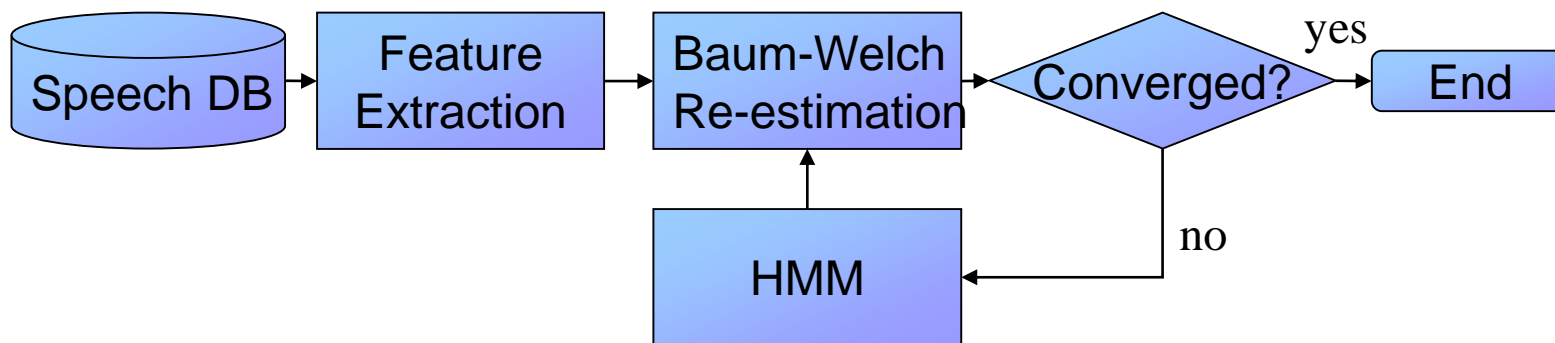
- Tomato



- $P([\text{towmeytow}]|\text{tomato}) = P([\text{towmaatow}]|\text{tomato}) = 0.1$
- $P([\text{tahmeytow}]|\text{tomato}) = P([\text{tahmaatow}]|\text{tomato}) = 0.4$

Training

■ Training process [Lee et al., 1996]

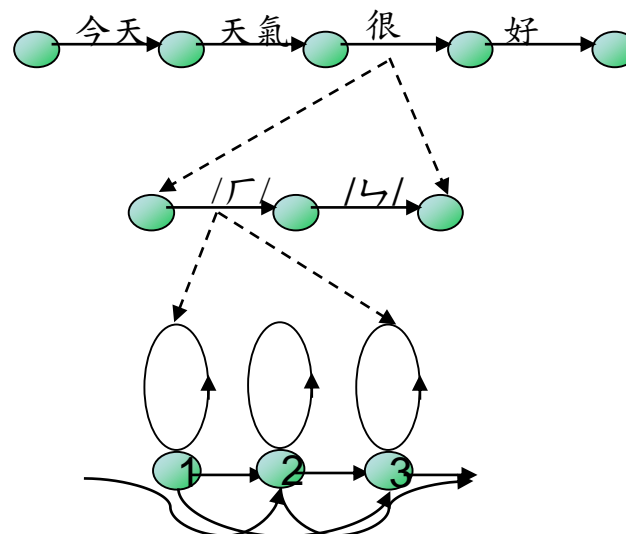


■ Network for training

Sentence HMM 今天天氣很好

Word HMM 很

Phone HMM /ɿ/



Language Model

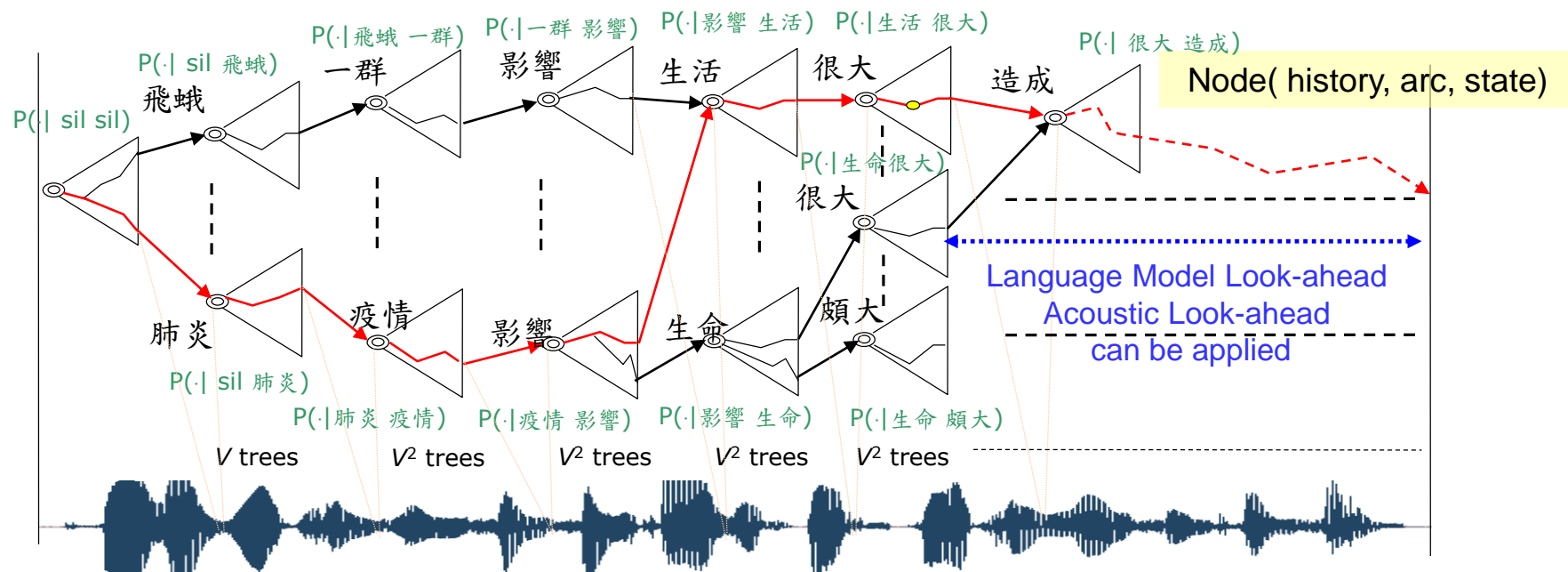
- Provide $P(W)$; the probability of the sentence [Beaujard et al., 1999]
 - We saw this was also used in the decoding process as the probability of transitioning from one word to another.
 - Word sequence : $W = w_1, w_2, w_3, \dots, w_n$

$$P(w_1 \cdots w_n) = \prod_{i=1}^n P(w_i \mid w_1 \cdots w_{i-1})$$

- The problem is that we cannot reliably estimate the conditional word probabilities, $P(w_i \mid w_1 \cdots w_{i-1})$ for all words and all sequence lengths in a given language
- n-gram Language Model
 - n-gram language models use the previous n-1 words to represent the history
$$P(w_i \mid w_1 \cdots w_{i-1}) = P(w_i \mid w_{i-(n-1)} \cdots w_{i-1})$$
 - Bi-grams are easily incorporated in a viterbi search

Language Model

■ Tree-Copy Search with Bigram/Trigram Language Modeling

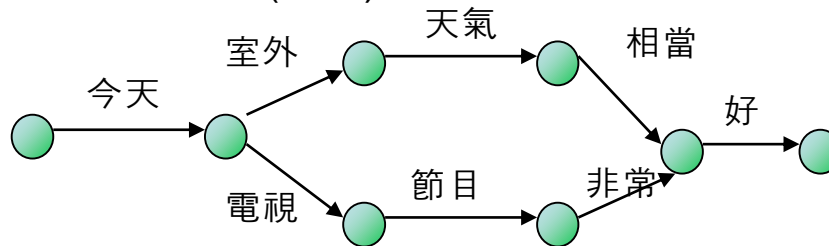


- The pronunciation lexicon (發音詞典) is structured as a tree
- n-gram language modeling (n-連語言模型), a word's occurrence is dependent on the previous n-1 words
- Search through all possible tree copies to find a best sequence of word hypotheses

Language Model

■ Example

- Finite State Network (FSN)



- Context Free Grammar (CFG)

$S \rightarrow VP\ NP$
 $NP \rightarrow D\ N$
 $VP \rightarrow V\ NP$ and $VP \rightarrow V$
 $D \rightarrow [the]$
 $N \rightarrow [dog];[cat]$

- Bigram

$P(電視|今天)=0.2$ $P(室外|今天)=0.5$
 $P(天氣|室外)=1.0$ $P(節目|電視)=0.5$
...

Context Free Grammar

- The structure of an English sentence can be best represented in a **hierarchical tree**.

- For example:

The dog chased a cat into the garden

GRAMMAR: a set of PS rules

$S \rightarrow NP VP$

$NP \rightarrow D N$

$VP \rightarrow V NP \mid V NP PP$

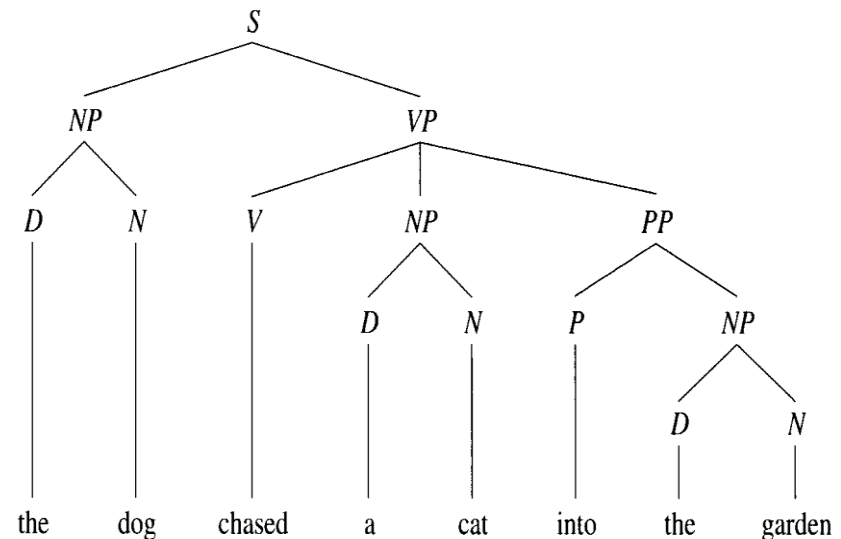
$PP \rightarrow P NP$

$D \rightarrow \text{the} \mid \text{a}$

$N \rightarrow \text{dog} \mid \text{cat} \mid \text{garden}$

$V \rightarrow \text{chased} \mid \text{saw}$

$P \rightarrow \text{into}$

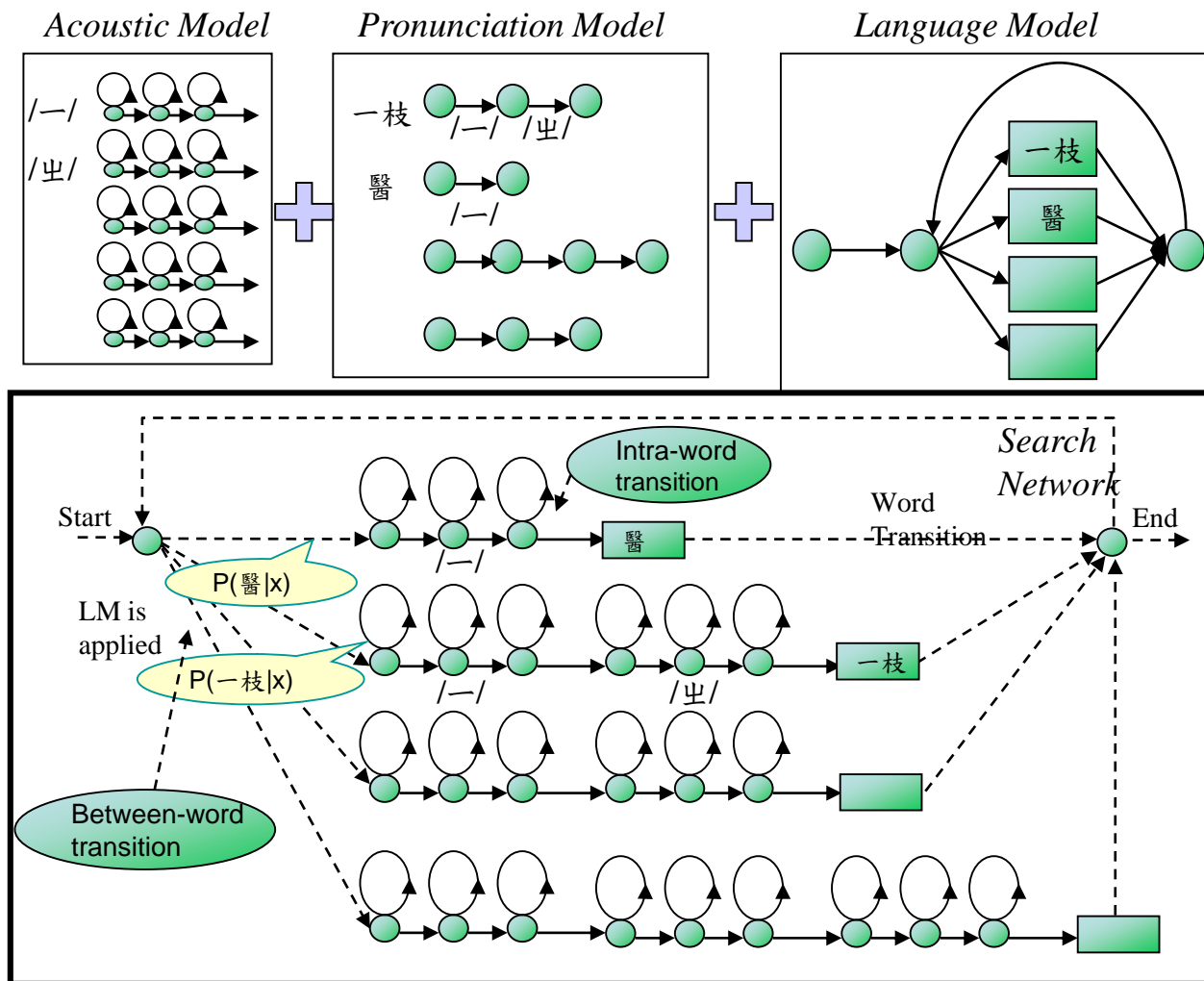


An equivalent expression:

$[_s[_{NP}[_D \text{the}][_N \text{dog}]][_{VP}[_V \text{chased}][_{NP}[_D \text{a}][_N \text{cat}]][_{PP}[_P \text{into}][_{NP}[_D \text{the}][_N \text{garden}]]]]]$

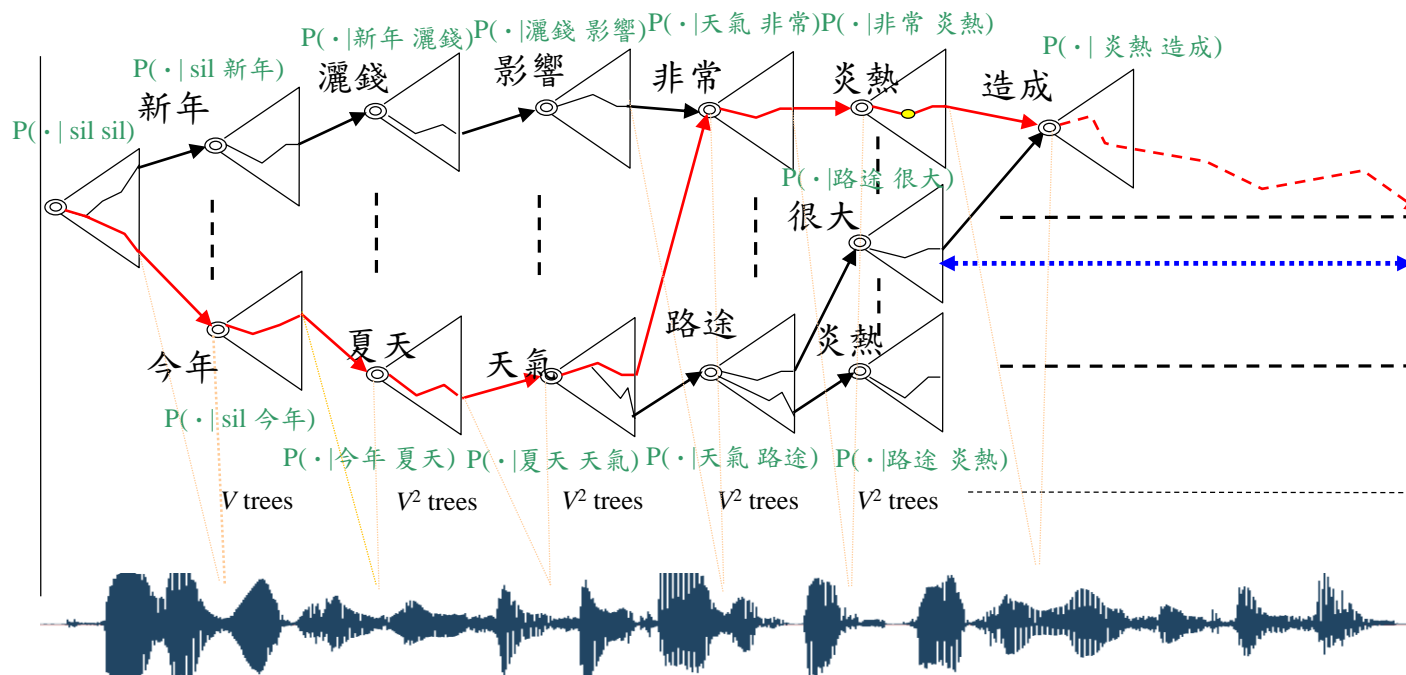
Network Construction

- Expanding every word to state level, we get a search network
[Demuynck et al., 1997]



Decoding

- Find $\hat{W} = \arg \max_{W \in L} P(W | O)$
- Viterbi Search** : Dynamic Programming



Multimedia Systems and Applications

Project -Speech Processing

1. Sample speech signal via sound card (Pronounce 多媒體系統與應用 in Chinese using continuous speech)
2. Speech signal analysis in Time Domain
 1. Waveform
 2. End point detection
 3. Energy contour
 4. Pitch contour
 5. Zero-crossing rate contour
3. Present the methods you used and the results
4. Submission due: **Two weeks after the project is assigned.**