Speech and Language Processing

Lecture 1

Speech recognition based on GMM, HMM, and N-gram

Information and Communications Engineering Course

Takahiro Shinozaki

Manabu Okumura

Lecture Plan (Shinozaki's part)

I gives the first 6 lectures about speech recognition. Through these lectures, the backbone of the latest speech recognition techniques is explained.

1. 10/19 (remote)
   Speech recognition based on GMM, HMM, and N-gram
2. 10/19 (remote)
   Maximum likelihood estimation and EM algorithm
3. 10/20 (remote)
   Bayesian network and Bayesian inference
4. 10/20 (remote)
   Variational inference and sampling
5. 10/22 (remote)
   Neural network based acoustic and language models
6. 10/22 (remote)
   Weighted finite state transducer (WFST) and speech decoding

Handouts

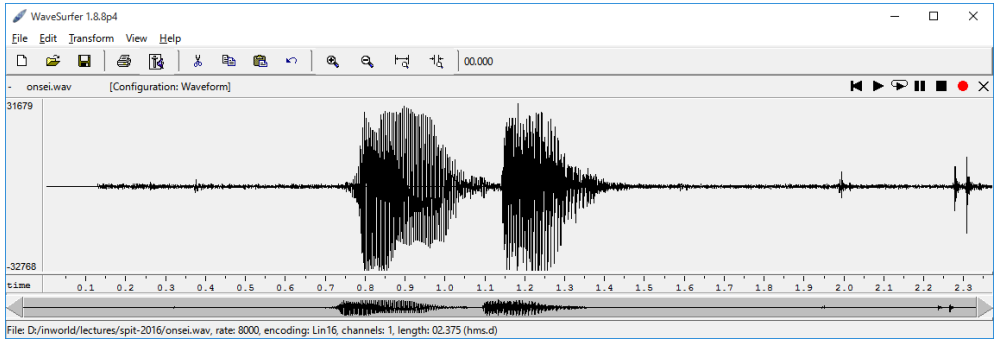
- All the materials are available at my home page:

http://www.ts.ip.titech.ac.jp/shinot/lectures/asrintro/

Perception of Speech Sound

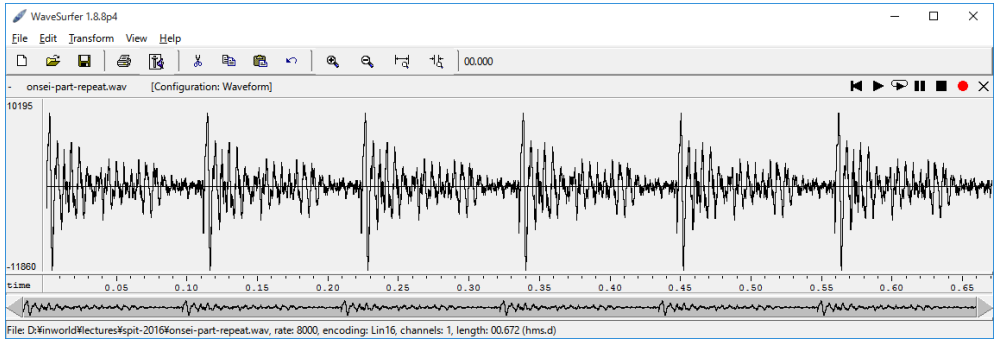
- Speech is a sound made by a physical process
- Frequency and time pattern are important

Speech sound pronunciating “Onsei”

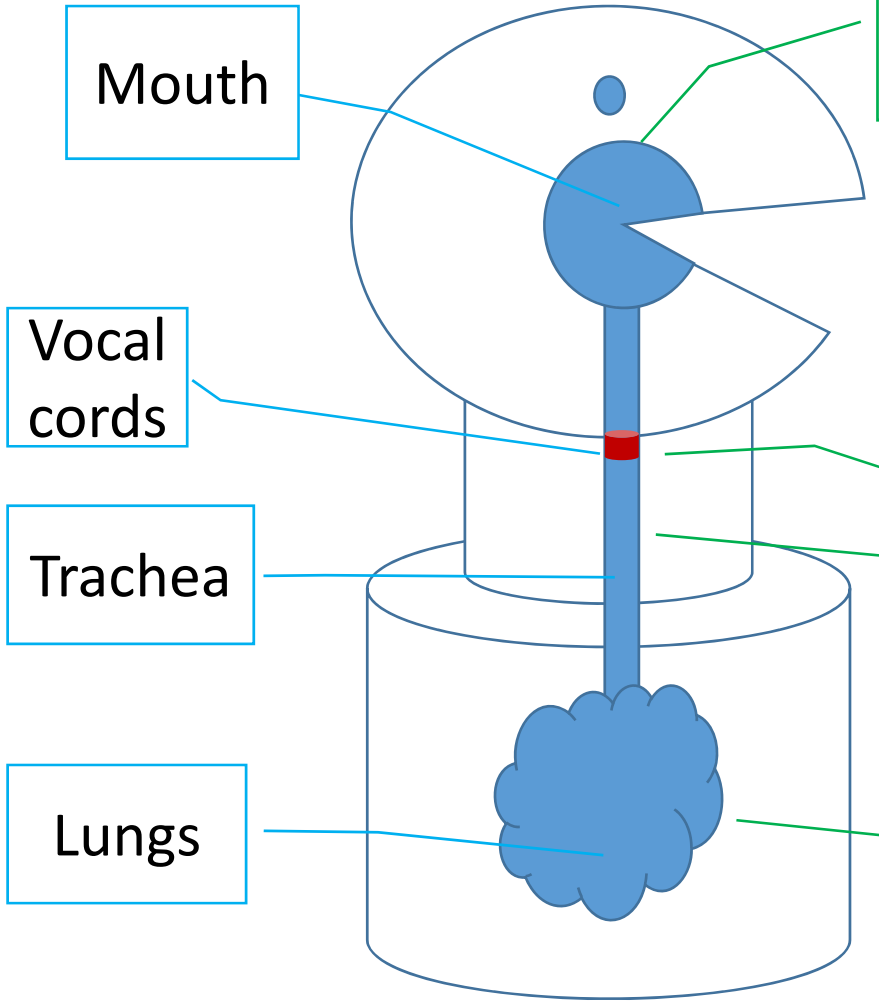


Sound is available in web version

Cut the region near “n” and repeat it 6 times.



Utterance Generation



Mouth

Vocal cords

Trachea

Lungs

Transfer characteristics H(ω) is modulated by moving the mouth

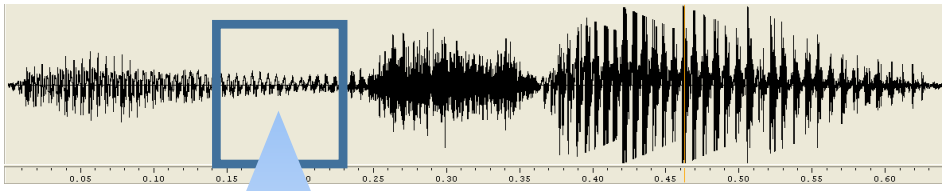
Speech sound G(ω)･H(ω)

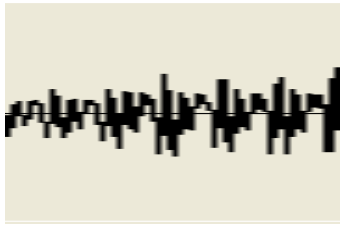
The membrane vibrate and generates beep sound G(ω)

The air flows in the pipe

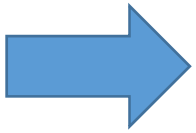
Makes compressed air

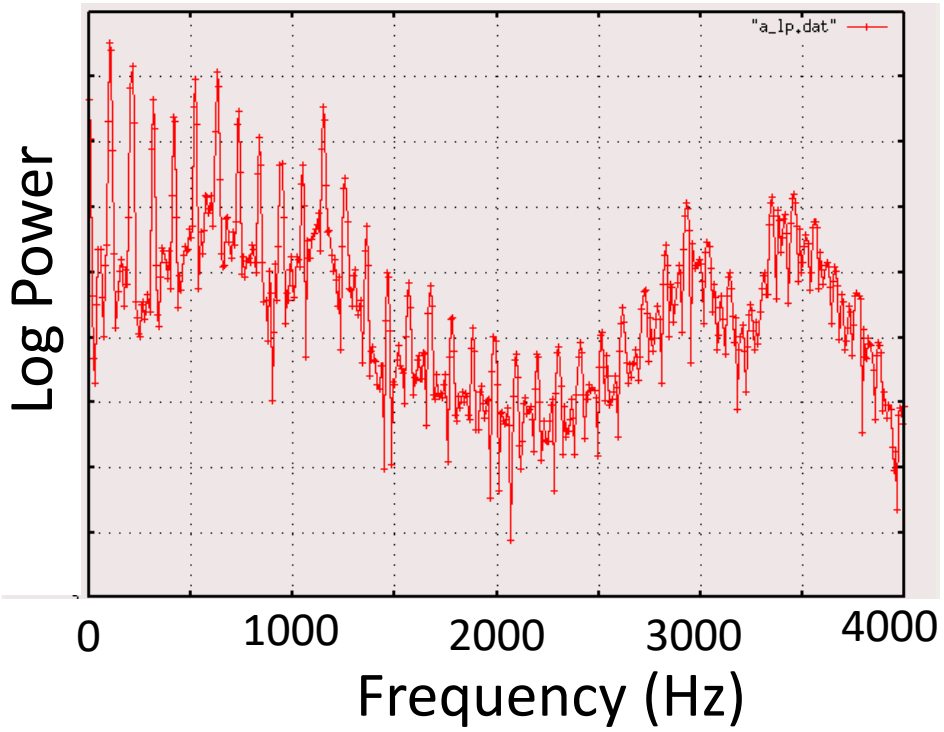
Spectral Analysis



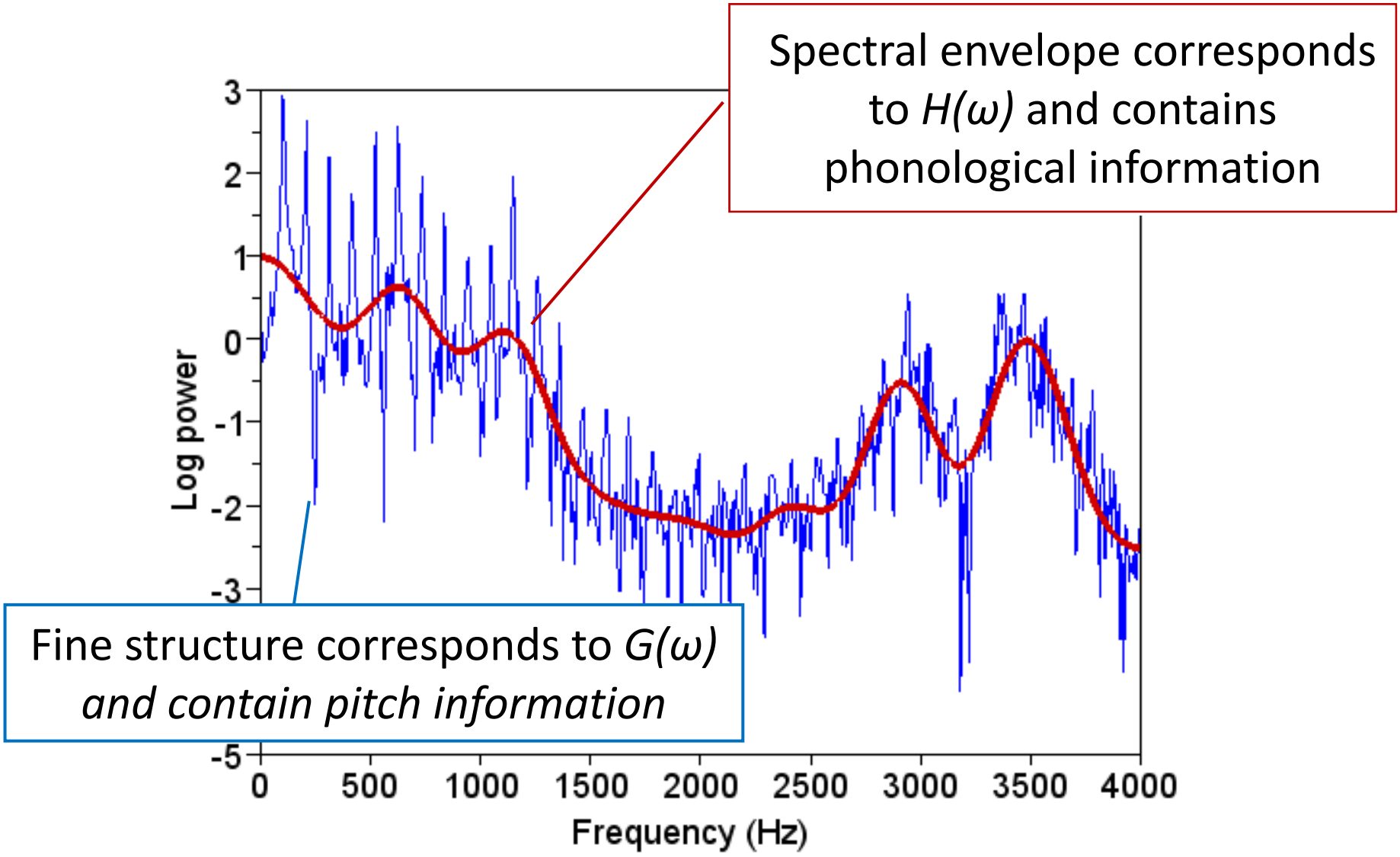


Fourier transform

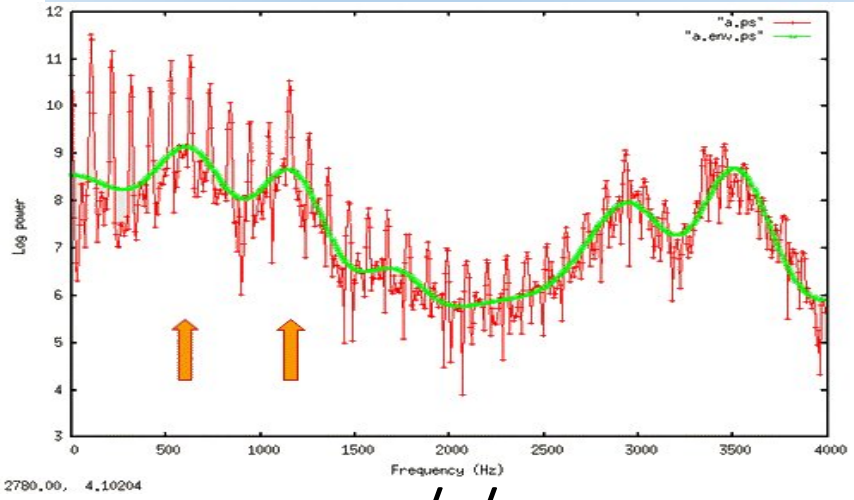




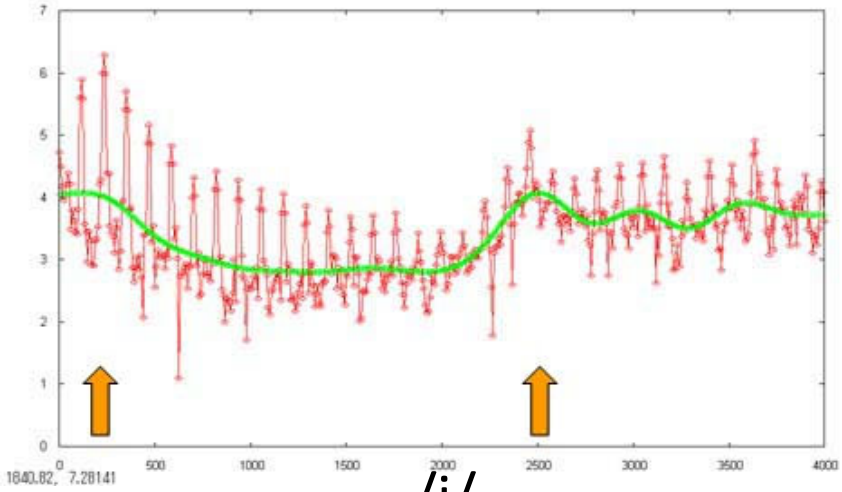
Spectral Envelope



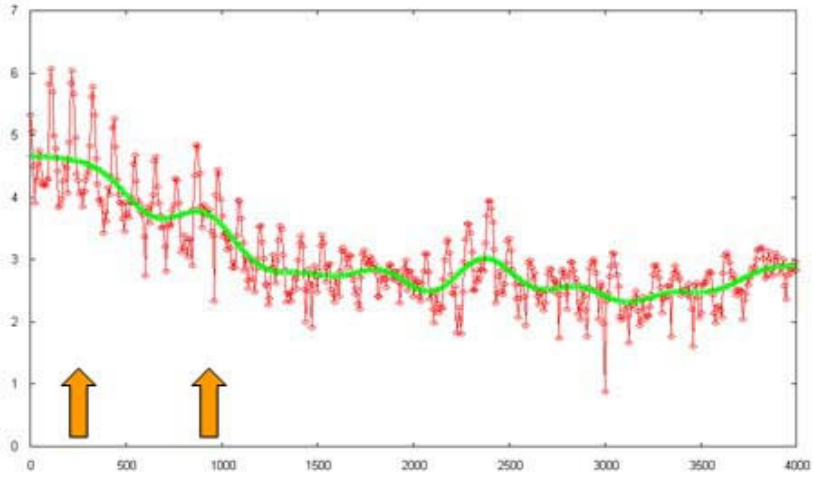
Vowels and Spectral Envelopes



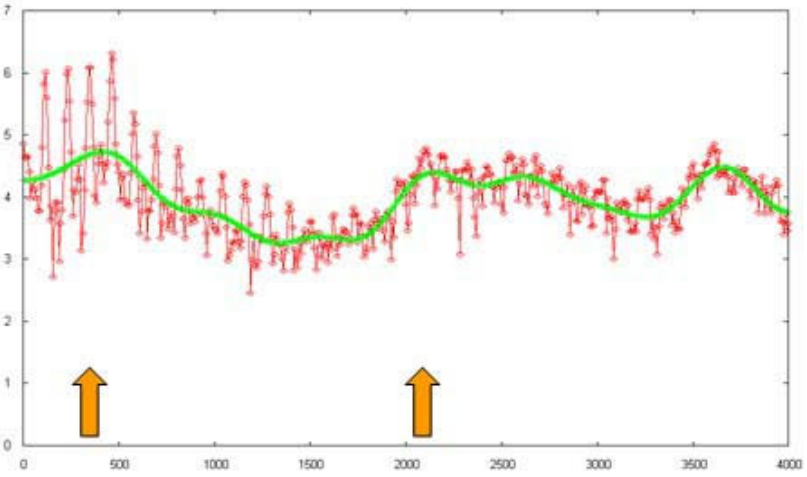
/a/



/i/



/u/



/e/

Experiment: Replacing the Sound Source *G*

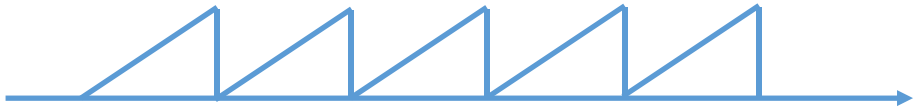
1. Record a voice

2. Analyze the voice, and decompose it to the sound source ***G*** and the transmission characteristics ***H***

X(ω) （original voice）	=	G(ω)	×	H(ω)

4. Replace the sound source ***G*** with another one ***G'***

E.g：sawtooth wave



5. Compute G'(ω) × H(ω) and re-generate waveform

G'(ω)	×	H(ω)	=	X'(ω) （synthesized sound）

Synthesized Voice Changing *G*

Sawtooth G’ (100Hz)  $G'(\omega)H(\omega)$ 

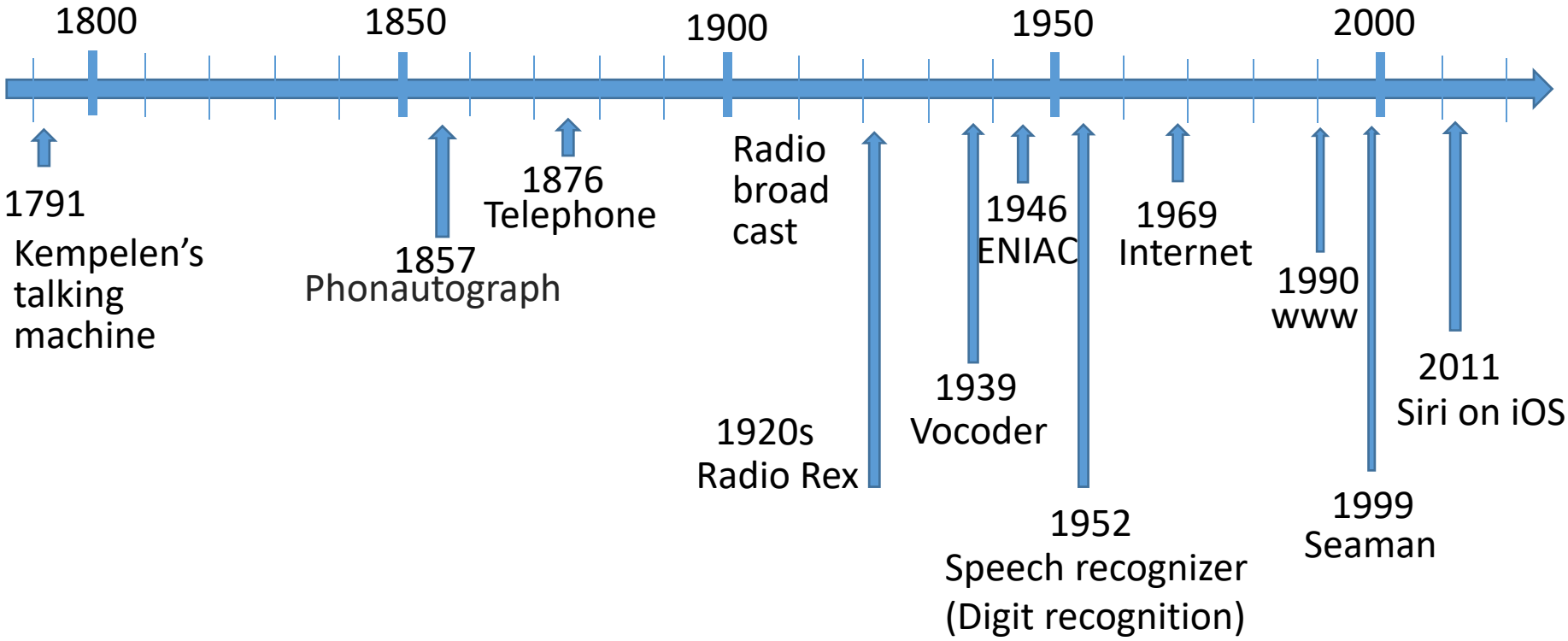
Sawtooth G’ (300Hz)  $G'(\omega)H(\omega)$ 

Sawtooth G’ (500Hz)  $G'(\omega)H(\omega)$ 

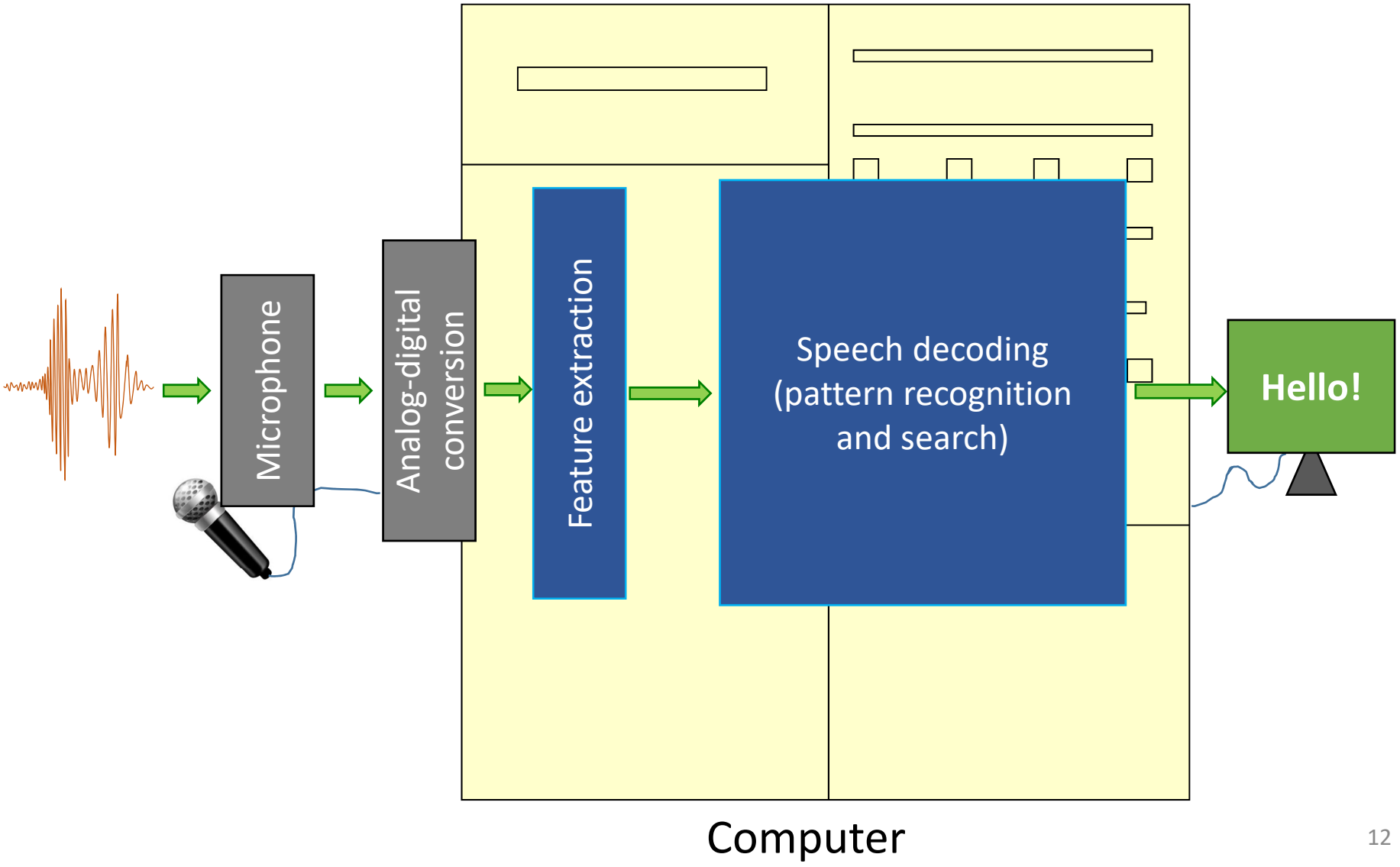
Music G’ (Acoustic11*)  $G'(\omega)H(\omega)$ 

*The music is from: http://maoudamashii.jokersounds.com/

History of Speech Technology

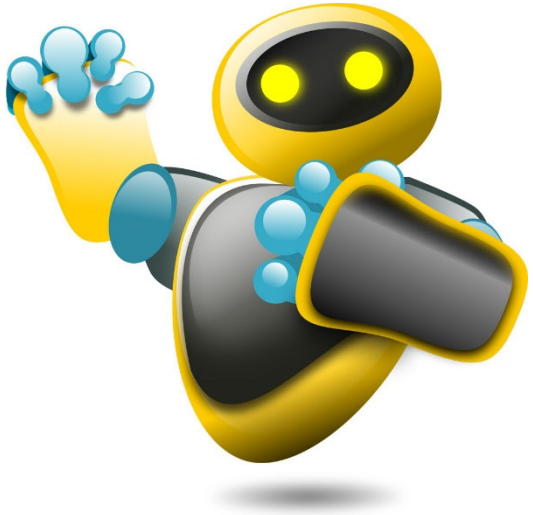


Organization of a Speech Recognition System



Applications of Speech Recognition

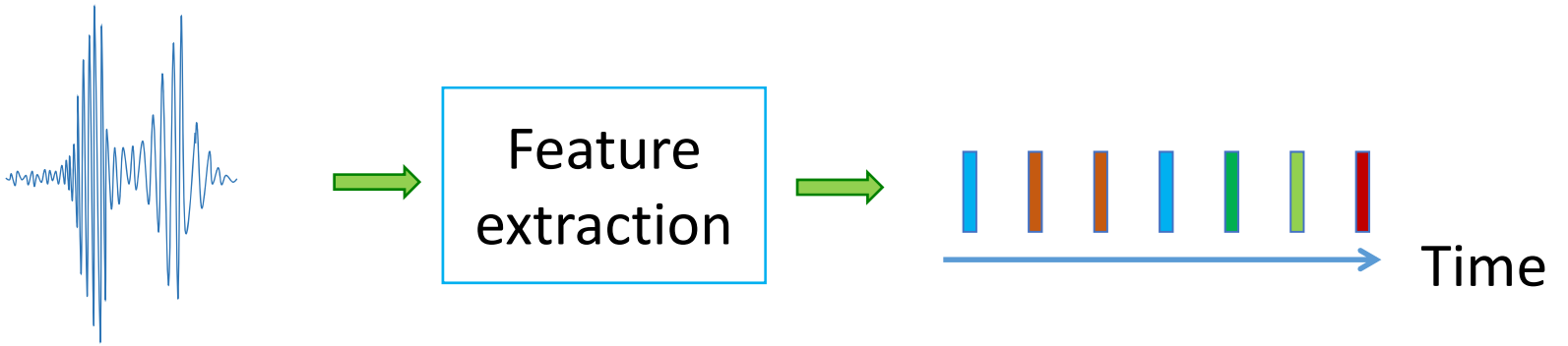
- Smartphone
  - Voice assistance
  - Speech-to-speech translation
- Judge
  - Speech retrieval system to support citizen judge
- Television
  - Automatic captioning system
- Car navigation
  - Voice commands
- Toy robots
  - Speech conversation



Feature Extraction

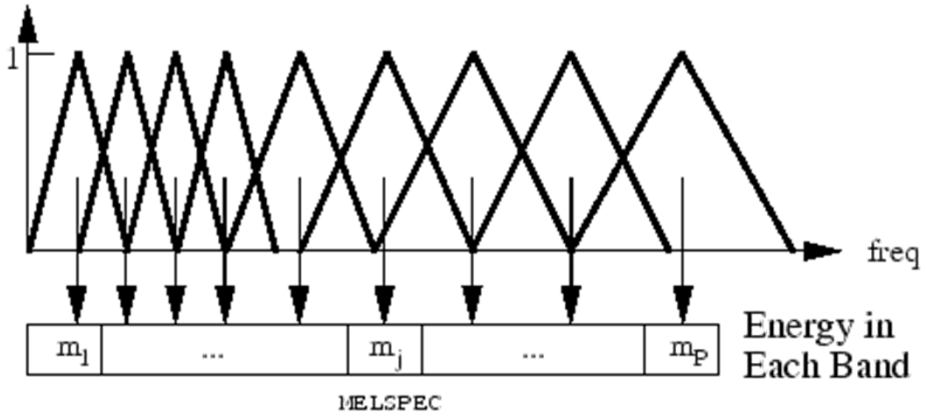
Extract useful information from the input signal in a convenient form for pattern recognition

- Help improving pattern recognition performance
- Reduce unnecessary memory and processing costs

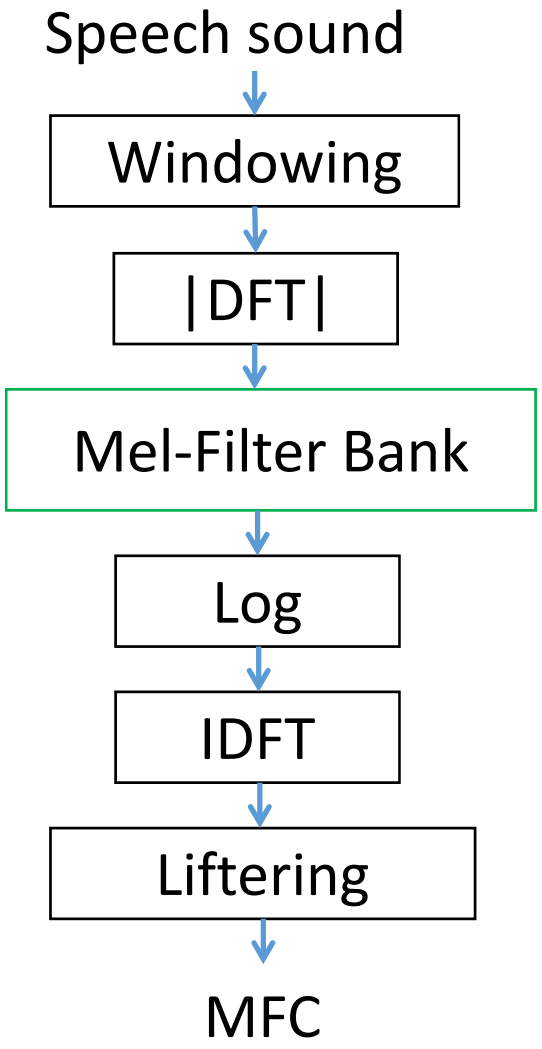


Mel-Frequency Cepstrum (MFC)

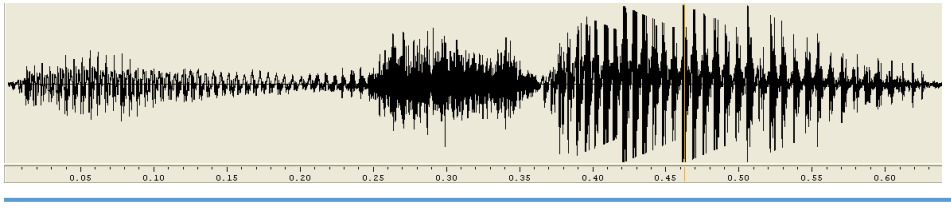
- Widely used features for speech recognition
- Emulate perceptual scale of pitches by using Mel-scale filter bank



Mel-Scale Filter Bank



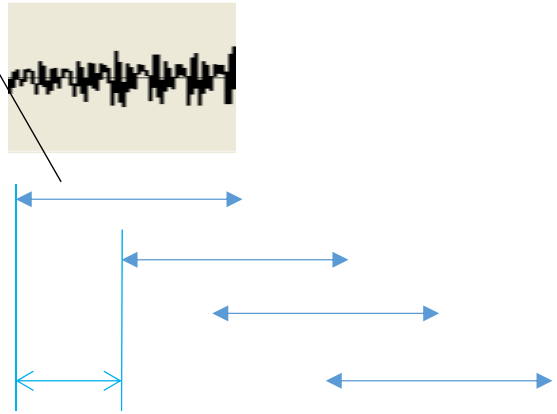
Typical Feature Extraction Process



16kHz sampling
16bit quantization

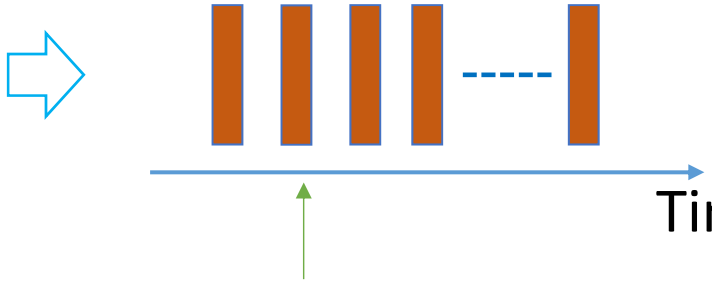
Time

Window width: 32ms (=512samples/16kHz)



shift: 10ms

Feature sequence
Sequence of real valued vectors
Rate=100Hz



Time

A vector is called a
“frame”

Speech Decoding

***O**: Input acoustic features (or a feature sequence)*
***W** : A symbol (or a symbol sequence) to recognize*
e.g. phone, word, word sequence, etc.

Statistical Speech Recognition

- Use probability distribution to model speech sounds

$$\hat{W} = \underset{W}{\arg\max}\, P(W \mid O)$$

Speech recognizer

Speech model

Acoustic Model and Language Model

- By using the Bayes' theorem, the probability is decomposed to two parts
- P(*O*) is independent of the maximization of *W*, and can be ignored

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

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

$$= \arg\max_{W} P(O \mid W)P(W)$$

Speech recognizer	Acoustic model (AM)	Language model (LM)

Problem Settings

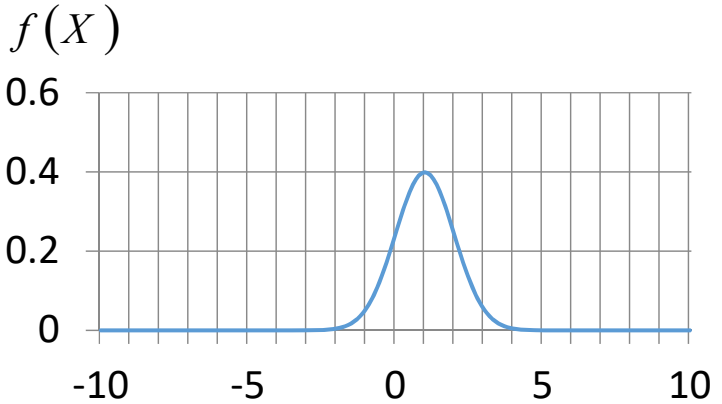
- Frame-wise vowel recognition
  - ***O***: feature vector of a single frame
  - ***W***: One of the vowels (For Japanese: a,i,u,e,o)
- Isolated phone recognition
  - ***O***: A sequence of feature vectors of a segment of phone utterance
  - ***W***: One of the phones
- Isolated word recognition
  - ***O***: A sequence of feature vectors of a segment of word utterance
  - ***W***: One of the words in a vocabulary
- Continuous word recognition
  - ***O***: A sequence of feature vectors of an utterance
  - ***W***: Sequence of words

Gaussian Distribution

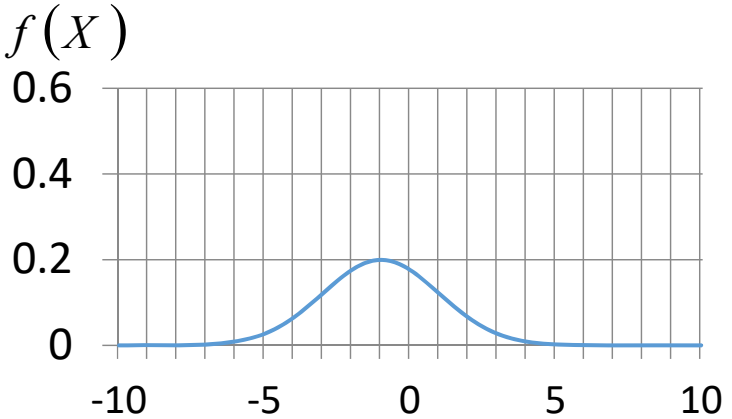
- Defined by two parameters mean μ and standard deviation σ (σ^2 is variance)

$$N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

It satisfies: $0 < N(x \mid \mu, \sigma^2), \quad \int_{-\infty}^{\infty} N(x \mid \mu, \sigma^2) dx = 1$



$\mu = 1, \quad \sigma = 1$



$\mu = -1, \quad \sigma = 2$

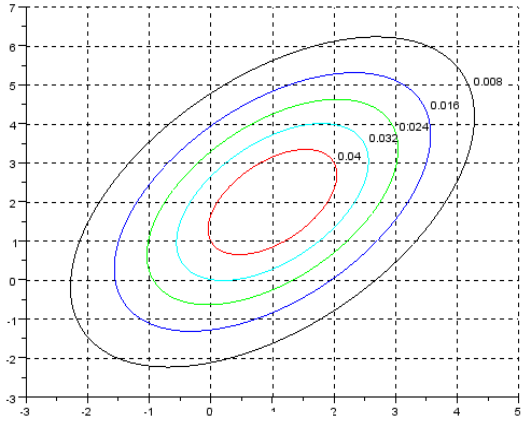
Multivariate Gaussian Distribution

- For D-dimensional vector **x**, it is defined using a mean vector **μ** and a covariance matrix **Σ**:

$$N(\mathbf{x}\,|\,\boldsymbol{\mu},\mathbf{S}) = \frac{1}{\sqrt{(2\pi)^D\,|\mathbf{S}|}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

|S| denotes determinant of S

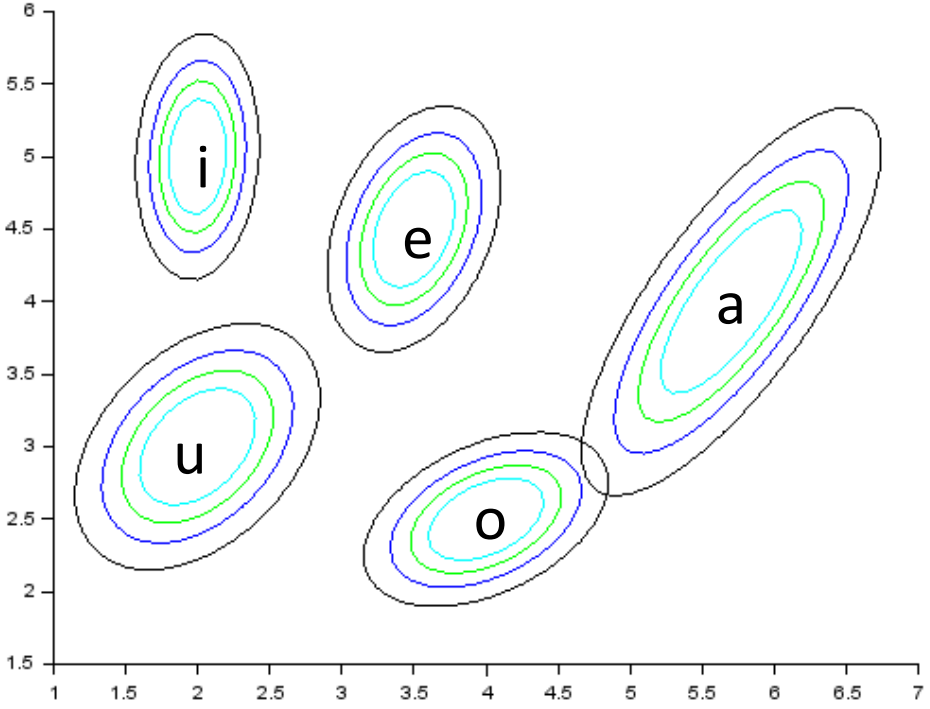
Contour plot of an example of a two dimensional Gaussian distribution



Gaussian Distribution based AM

- Fit a Gaussian distribution for each vowel

$$P_w(O) = N(x \mid \mu_w, \sigma_w^2) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left\{-\frac{1}{2\sigma_w^2}(x-\mu_w)^2\right\}$$



How to fit the distributions?

We will consider this problem later in the lecture of maximum likelihood estimation

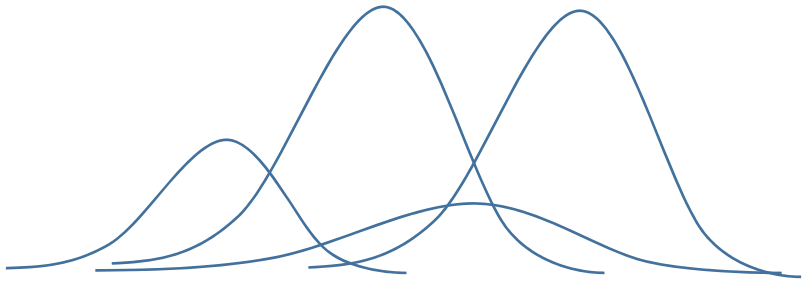
Gaussian Mixture Model (GMM)

- By mixing multiple Gaussian distributions, a complex distribution can be expressed
  → Useful to improve recognition performance

$$GMM(X) = \sum_i w_i N_i(X \mid \mu_i, S_i) \qquad \sum_{m=1}^{M} w_k = 1.0$$

w_i : Mixture weight

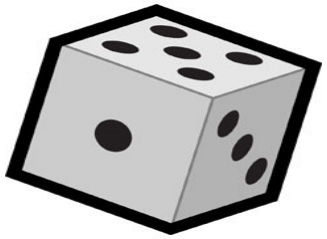
N_i : Component Gaussian distribution with mean μ_i and covariance S_i



Categorical Distribution

- The distribution is represented by a table
- The probability distribution of a skewed die is an example of categorical distribution

Vowel	a	i	u	e	o
Probability	0.3	0.1	0.2	0.1	0.3



1-of-K Representation

- The same probability as the table description can be represented as an equation by using 1-of-K representation

Value W	1-of-K representation $W=(w_1, w_2, w_3, w_4, w_5)$	Probability $\boldsymbol{\rho}=(\rho_1, \rho_2, \rho_3, \rho_4, \rho_5)$
1 （a）	1,0,0,0,0	$Pr(W=1)=\rho_1=0.3$
2 （i）	0,1,0,0,0	$Pr(W=2)=\rho_2=0.1$
3 （u）	0,0,1,0,0	$Pr(W=3)=\rho_3=0.2$
4 （e）	0,0,0,1,0	$Pr(W=4)=\rho_4=0.1$
5 （o）	0,0,0,0,1	$Pr(W=5)=\rho_5=0.3$

$$p(W)=\prod_{k=1}^{K}\rho_k^{\,w_k}$$

Example of Frame-wise Vowel Recognition

- Gaussian distribution based acoustic model
- Categorical distribution based language model

$$\hat{W} = \underset{W \in \{a,i,u,e,o\}}{\arg\max}\ P(W \mid O)$$

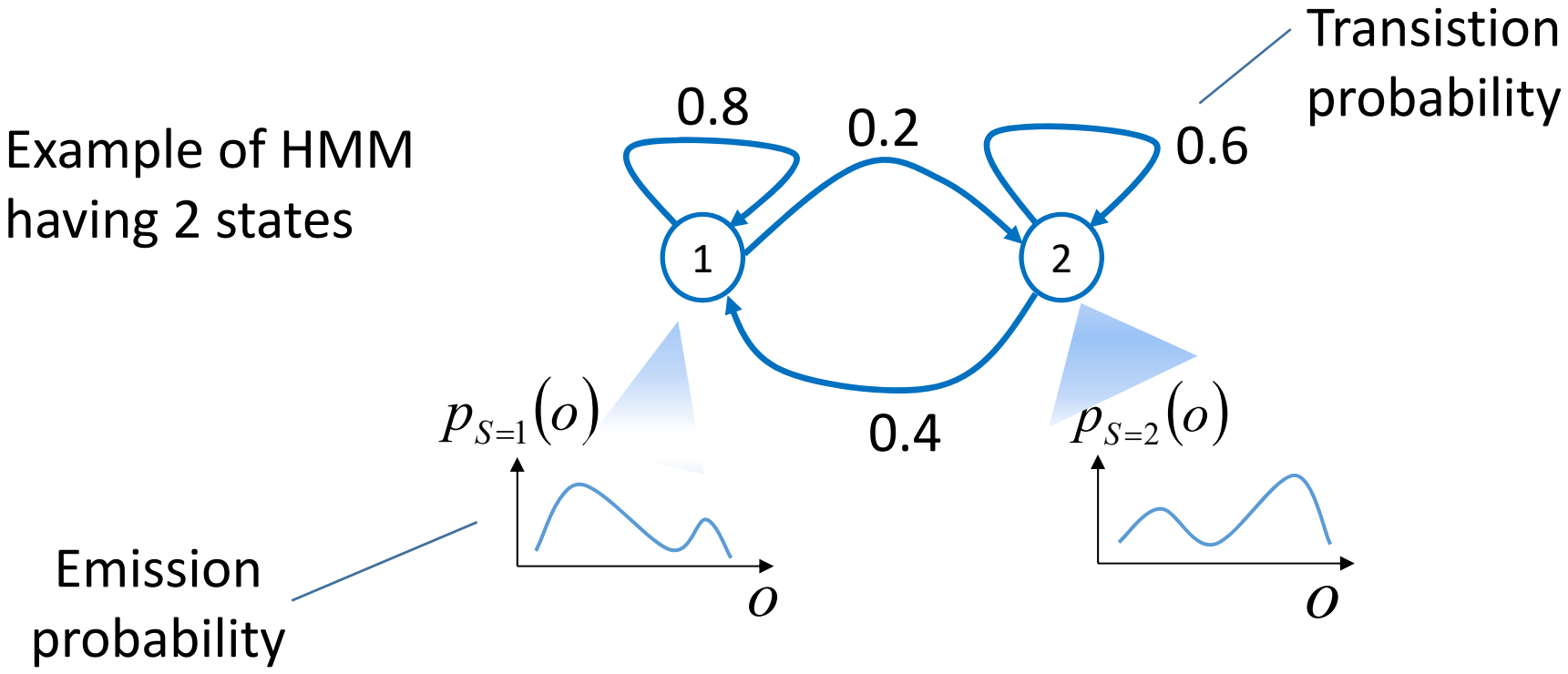
$$= \underset{W \in \{a,i,u,e,o\}}{\arg\max}\ N(O \mid \mu_W, \Sigma_W) \prod_k \rho_k^{w_k}$$

Modeled by a set of Gaussian distributions prepared for each W

Modeled by a categorical distribution

Hidden Markov Model (HMM)

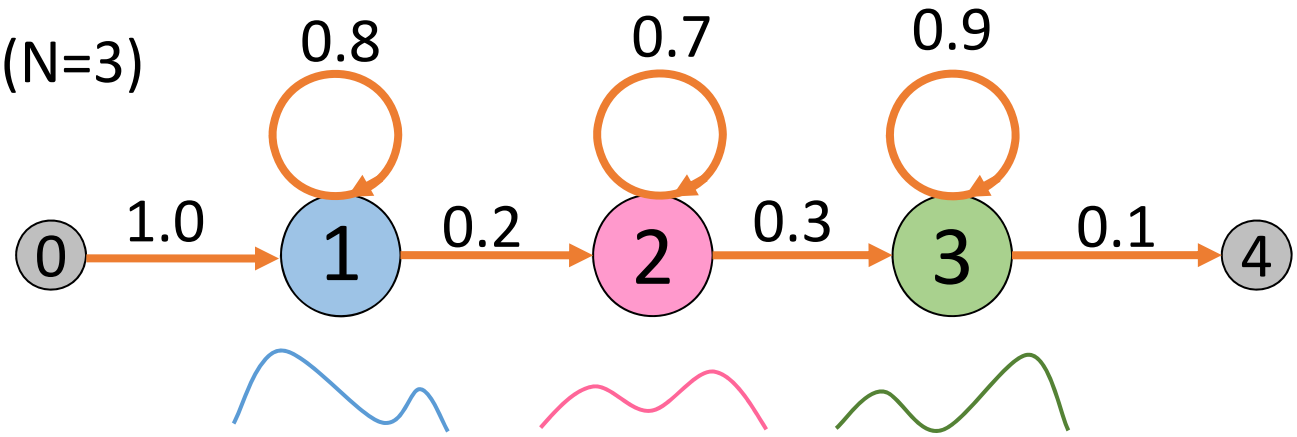
- A probabilistic model for sequential data
- Defined by a set of states, state transition probabilities, and state emission probabilities



HMM based Acoustic Model

- Popular HMM design for acoustic modeling
  - Left-to-right transition structure
  - Non-emitting start and end states

Example of 3 state (N=3) left-to-right HMM



For a feature sequence $O=<o_1, o_2,\ldots,o_T>$ with length T,

$$P(O)=\sum_{S\in SS}\left\{\prod_{t=1}^{T}P(s_t\mid s_{t-1})P(o_t\mid s_t)\right\}P(s_{Fin}\mid s_T),\quad s_0=0,\quad s_{Fin}=N+1,\quad s_t\in\{1,2,\cdots,N\}$$

$S=<0, s_1, s_2,\ldots,s_T, N+1>$, SS is a set of all possible state sequences

Example of Isolated Phone Recognition

- HMM based acoustic model
- Categorical distribution based language model

$$\hat{W} = \arg\max_{W \in \{phones\}} P(W \mid O)$$

$$= \arg\max_{W \in \{phones\}} HMM(O \mid \lambda_W) \prod_k \rho_k^{w_k}$$

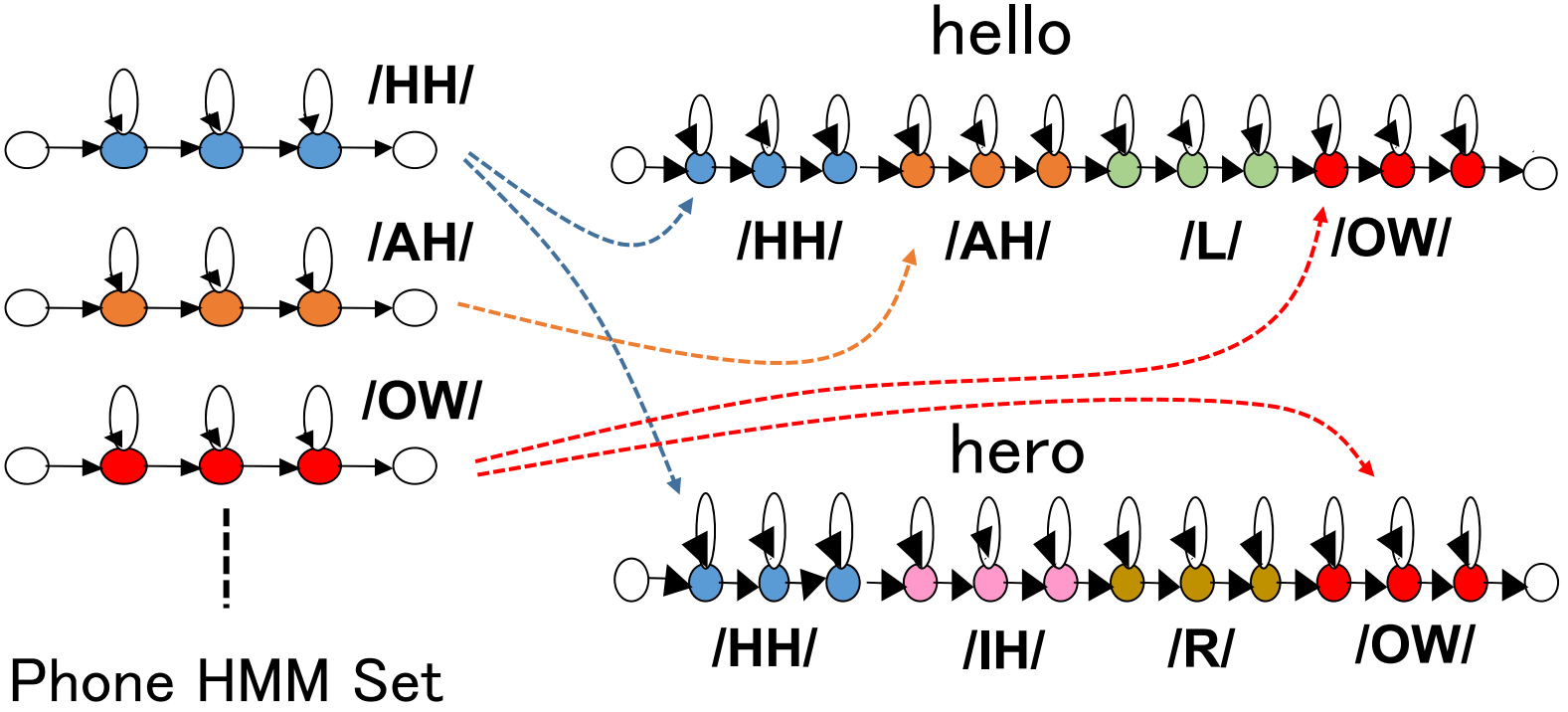
All parameters of an HMM: Transition probabilities and emission probabilities

Modeled by a set of HMMs prepared for each W

Modeled by a categorical distribution

Phone based Word Modeling

- When # of words is large, preparing an HMM for each word is difficult since # of parameters increases
- Phone based modeling composes arbitral word models by concatenating phone models



Example of Isolated Word Recognition

- HMM based acoustic model
  - A word is directly modeled by an HMM, OR
  - The phone based word modeling strategy can be used
- Categorical distribution based language model

$$\hat{W} = \arg\max_{W \in \{words\}} P(W \mid O)$$

$$= \arg\max_{W \in \{words\}} HMM(O \mid \lambda_W) \prod_k \rho_k^{w_k}$$

Modeled by a set of HMMs prepared for each W

Modeled by a categorical distribution

N-gram Model

- Assumes that the appearance of a word in an utterance depends on at max N-1 preceding words as context
- Represented by a set of categorical distributions prepared for each context

$$P(w_1 w_2 w_3 \cdots w_T)$$
$$= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)P(w_4 \mid w_1 w_2 w_3)\cdots P(w_T \mid w_1 w_2 \cdots w_{T-1})$$

N-gram approximation
Ignore history (or context)
older than N-1 words

When N=2

$$\approx P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2)P(w_4 \mid w_3)\cdots P(w_T \mid w_{T-1})$$

Unigram

- N=1
  - Do not consider the history at all
  - Same as the product of individual word probabilities

$$P(w_1 w_2 w_3 \cdots w_T)$$
$$\approx P(w_1)P(w_2)P(w_3)P(w_4)\cdots P(w_T)$$
$$= \prod_{t=1}^{T} P(w_t)$$

Example：

P(“Today is a sunny day”) =
P(“today”)P(“is”)P(“a”)P(“sunny”)P(“day”)

Bi-gram

- N=2
  - Consider only a previous word as the history

$$P(w_1 w_2 w_3 \cdots w_T)$$
$$\approx P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2)P(w_4 \mid w_3)\cdots P(w_T \mid w_{T-1})$$
$$= P(w_1)\prod_{t=2}^{T} P(w_t \mid w_{t-1})$$

Example：

P(“Today is a sunny day”) =

P(today)P(is|today)P(a|is)P(sunny|a)P(day|sunny)

Tri-gram

- N=3
  - Consider two previous words as the history
  - Popular in speech recognition

$$P(w_1 w_2 w_3 \cdots w_T)$$
$$\approx P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)P(w_4 \mid w_2 w_3)\cdots P(w_T \mid w_{T-2} w_{T-1})$$
$$= P(w_1)P(w_2 \mid w_1)\prod_{t=3}^{T} P(w_t \mid w_{t-2}, w_{t-1})$$

Example：

P(“Today is a sunny day”) =

P(today)P(is|today)P(a|today, is)P(sunny|is, a)P(day|a, sunny)

Example of Continuous Word Recognition

- HMM based acoustic model
  - The phone based modeling approach can be applied to make utterance HMM
- N-gram based language model

$$\hat{W} = \underset{W \in \{utterances\}}{\arg\max} P(W \mid O)$$

$$= \underset{W \in \{utterances\}}{\arg\max} HMM(O \mid \lambda_W) Ngram(W)$$

Utterance HMM

Modeled by an N-gram, e.g. Tri-gram

Problem: How to Perform argmax?

- For continuous word recognition, # utterance is huge
  - E.g. If the vocabulary size *V* is 10000, and the utterance length *L* is 10, # utterances is $10000^{10} = 10^{40}$
- Enumerating all the utterances is impossible!

$$\hat{W} = \underset{W \in \{utterances\}}{\arg\max}\ HMM(O \mid \lambda_W) Ngram(W)$$

How to do this?

We will consider this problem later in the lecture of WFST

Exercise 1.1

Suppose W is a vowel and O is a MFCC feature vector. Suppose that $P_{AM}(O|W)$ is an acoustic model and $P_{LM}(W)$ is a language model. Obtain a vowel $\hat{W}$ that maximizes $P(W|O)$ when the acoustic and language model log likelihoods are given as in the following table.

$$\hat{W} = \arg\max_{W \in \{a,i,u,e,o\}} \{P(W \mid O)\}$$

Vowel V	a	i	u	e	o
log(P(O|V))	-13.4	-10.5	-30.1	-15.2	-17.0
log(P(V))	-1.61	-2.30	-1.61	-1.39	-1.39

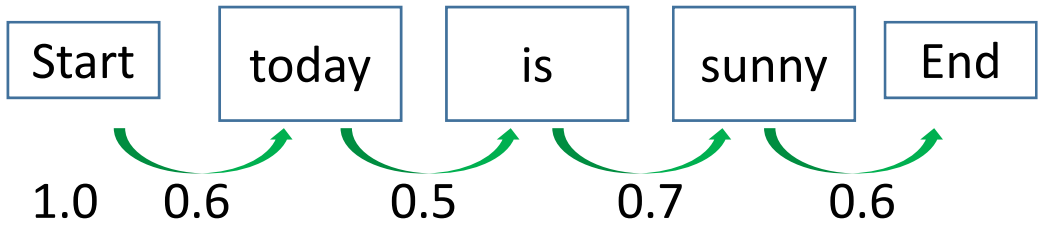
Exercise 1.2

- The following table defines a Bi-gram P(Word|Context)

C＼W	today	is	sunny	End
Start	0.6	0.1	0.2	0.1
today	0.1	0.5	0.3	0.1
is	0.1	0.1	0.7	0.1
sunny	0.1	0.1	0.2	0.6

*P(Start)=1.0

Example ：



1.0x0.6x0.5x0.7x0.6=0.126

Exercise 1.2 (Cont.)

- Based on the bigram definition of the previous slide, compute the probability of the following sentences

1) P(“Start today sunny today sunny End”)

= 

2) P(“Start today today sunny sunny End”)

=