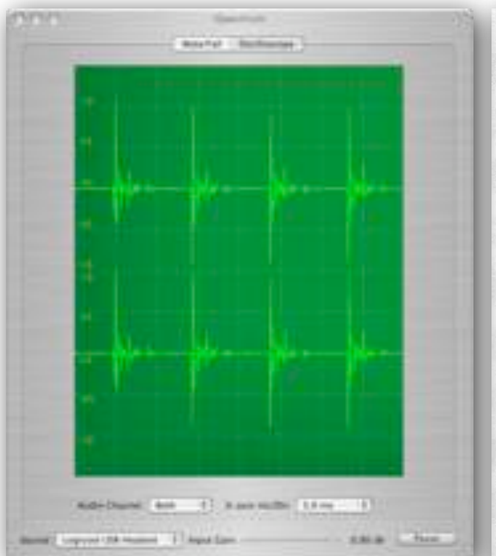
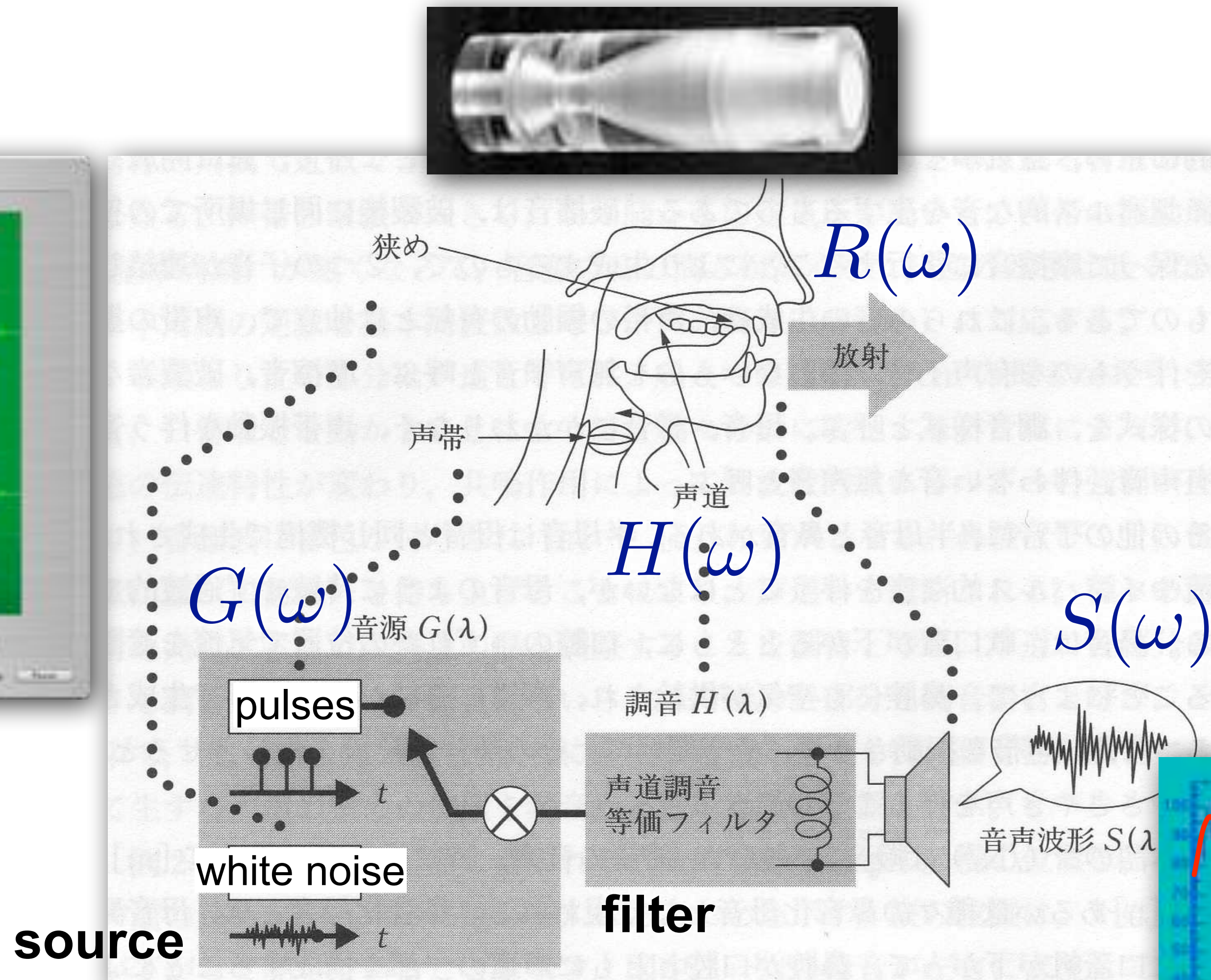
Cognitive Media Processing #7

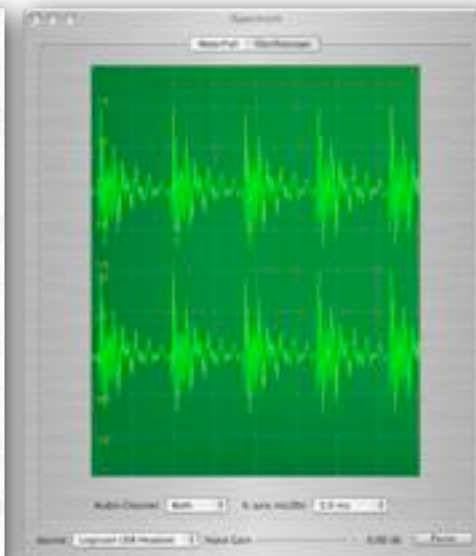
Nobuaki Minematsu

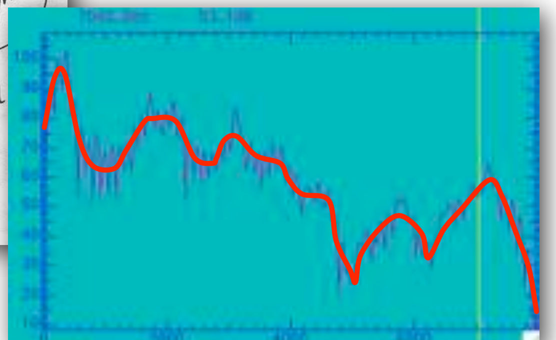
Modeling of speech production

- Mathematical modeling of speech production -- source & filter model --
  - Linear independence between source and filter





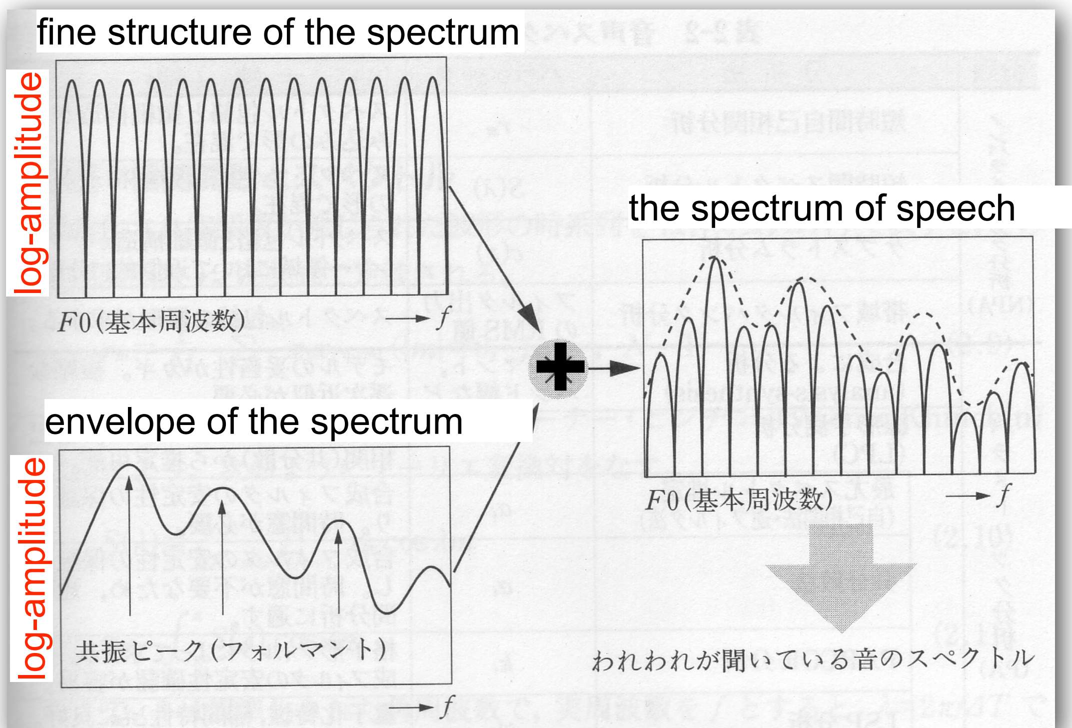


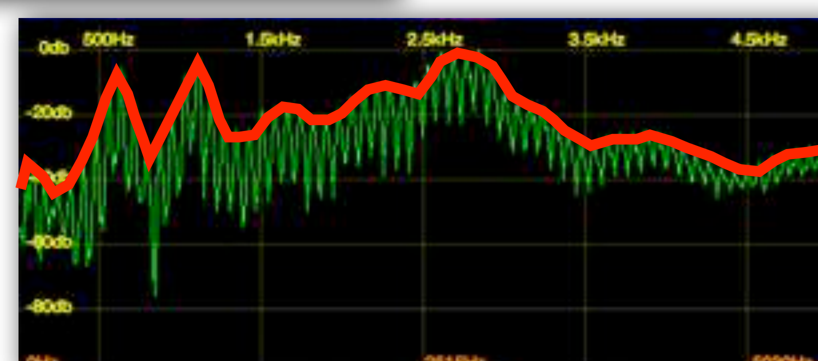


$$S(\omega) = G(\omega)H(\omega)R(\omega)$$

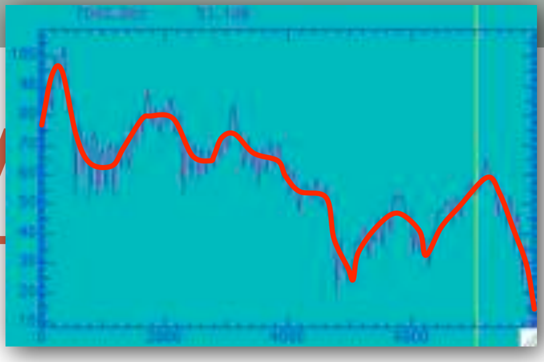
Modeling of vowel production

- Mathematical modeling of speech production -- source & filter model --
  - Separation between the spectrums of source and filter

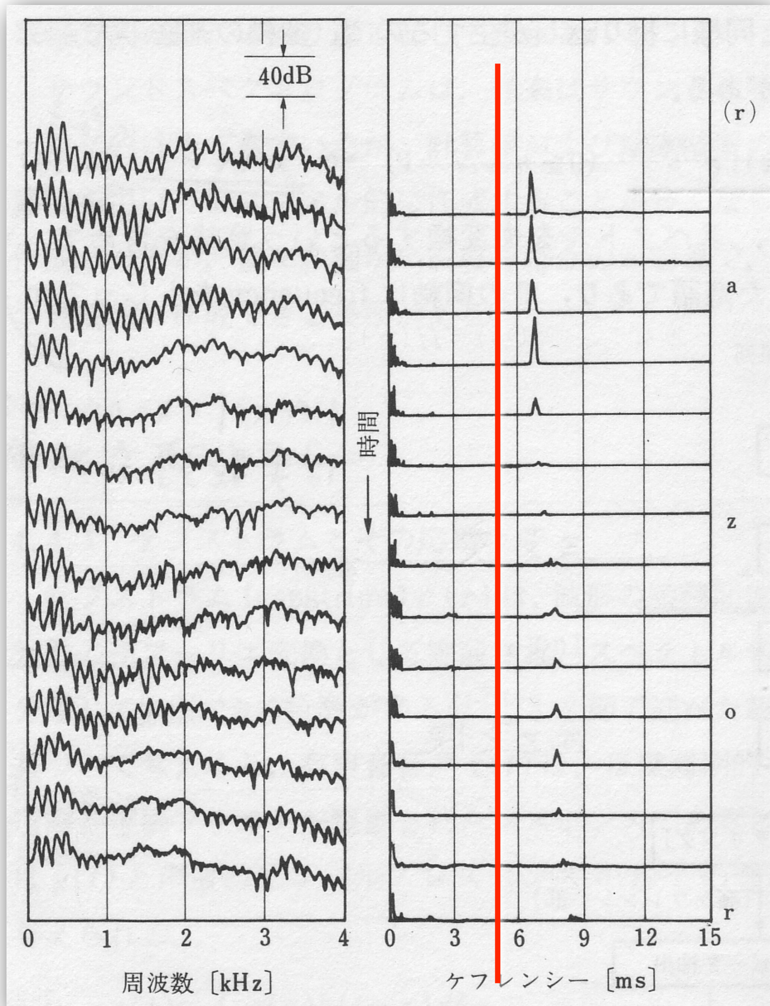


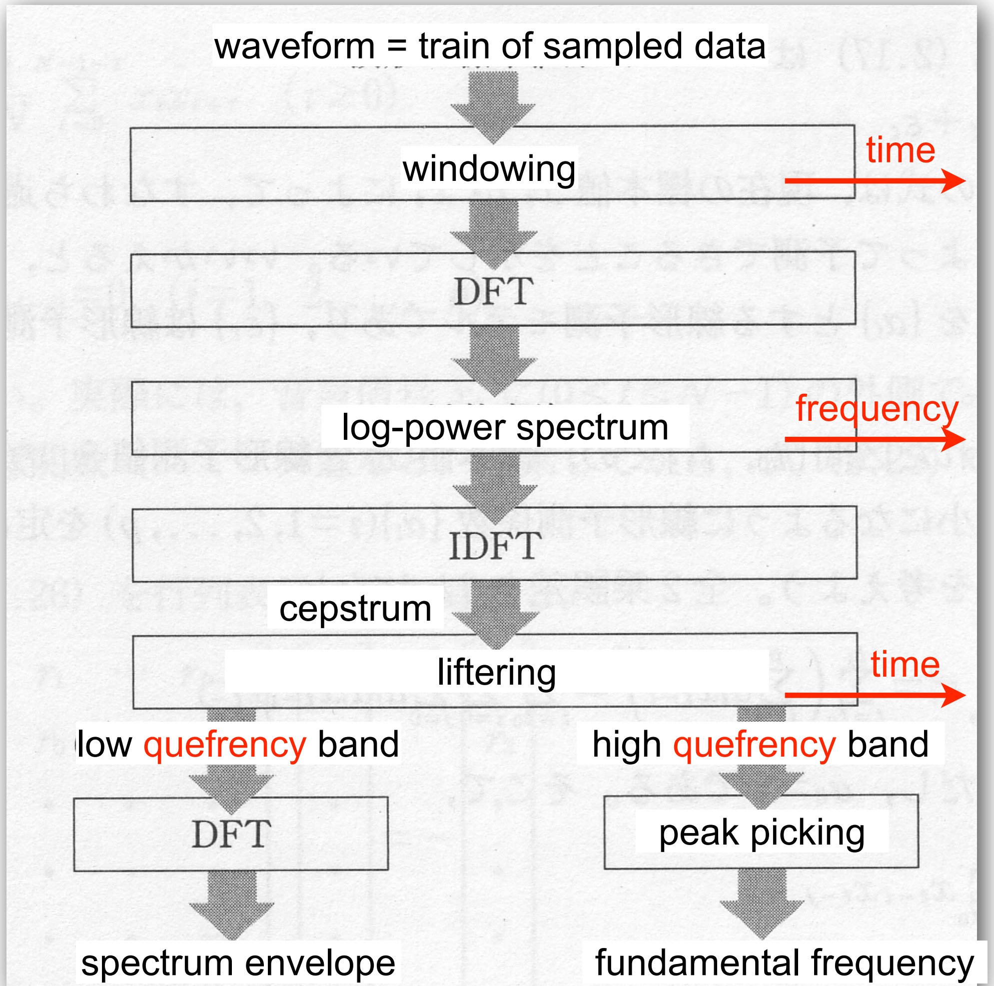


Extraction of spectrum env



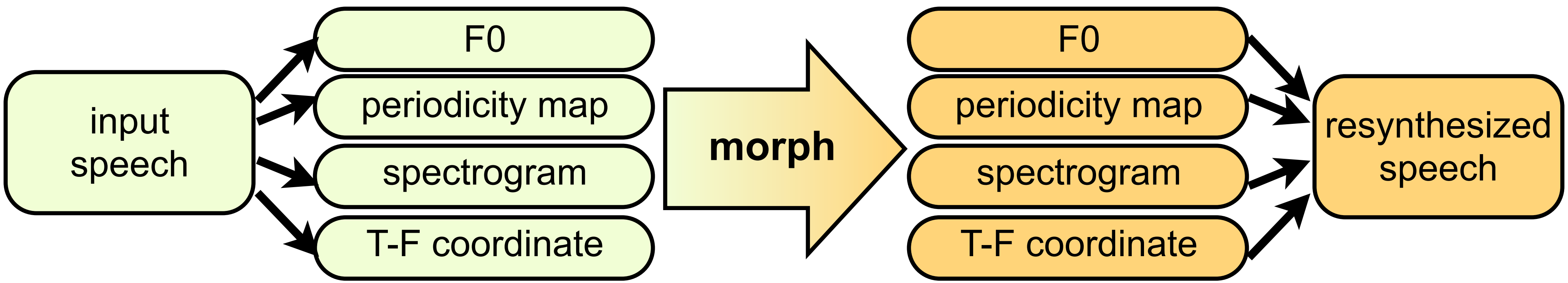
- Cepstrum method
  - Windowing + FFT + log-amplitude --> a spectrum with pitch harmonics
  - Smoothing (LPF) of the fine spectrum into its smoothed version





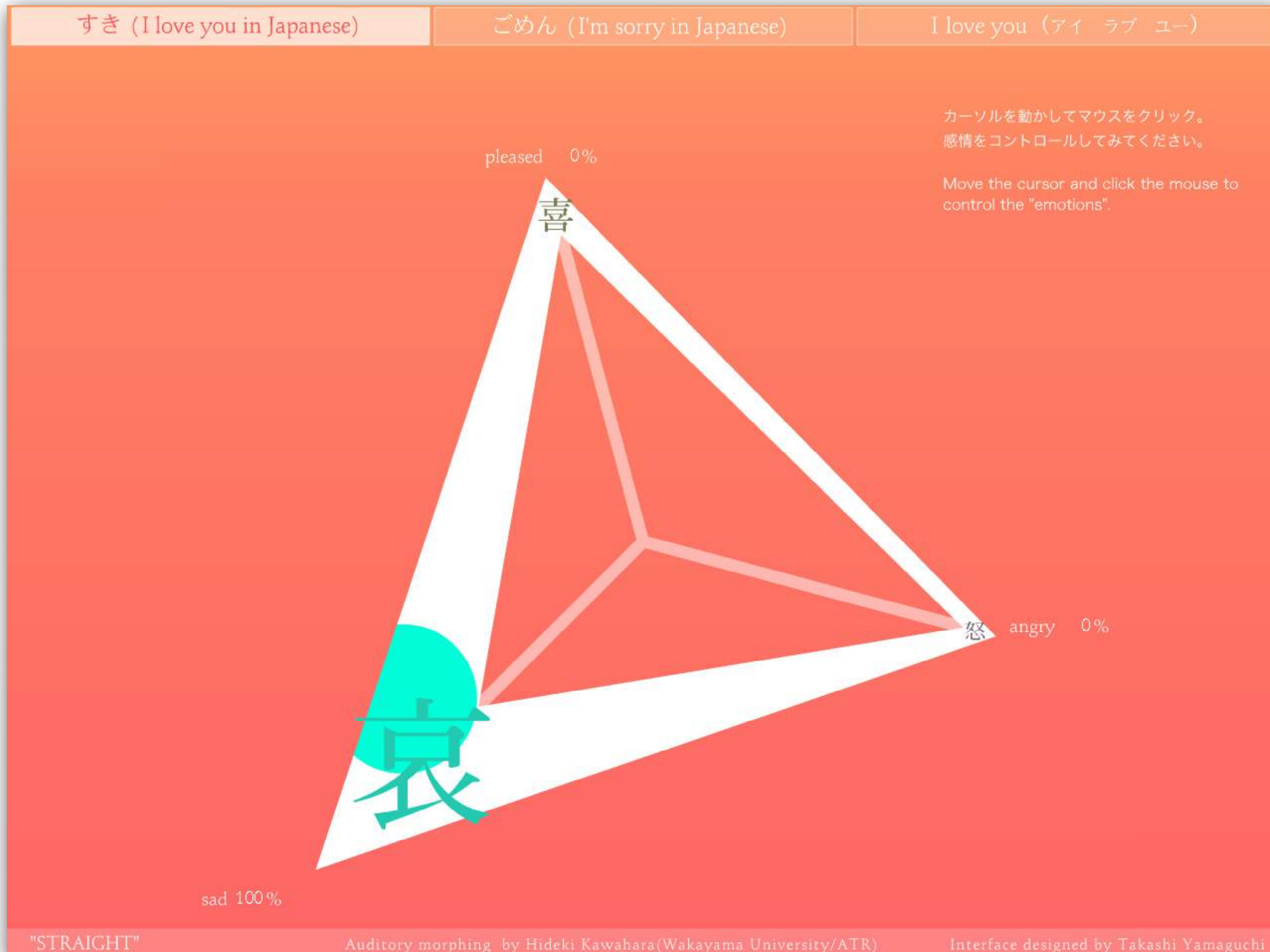
Advanced technology for analysis

- STRAIGHT [Kawahara'06]
  - High-quality analysis-resynthesis tool
    - Decomposition of speech into
      - Fundamental frequency, spectrographic representations of power, and that of periodicity
    - High-quality speech morphing tool



  - Spectrographic representation of power
    - F0 adaptive complementary set of windows and spline based optimal smoothing
  - Instantaneous frequency based F0 extraction
    - With correlation-based F0 extraction integrated
  - Spectrographic representation of periodicity
    - Harmonic analysis based method

Examples of speech morphing



Spectrum reading

- What are these?
  - Hint : they are numbers.

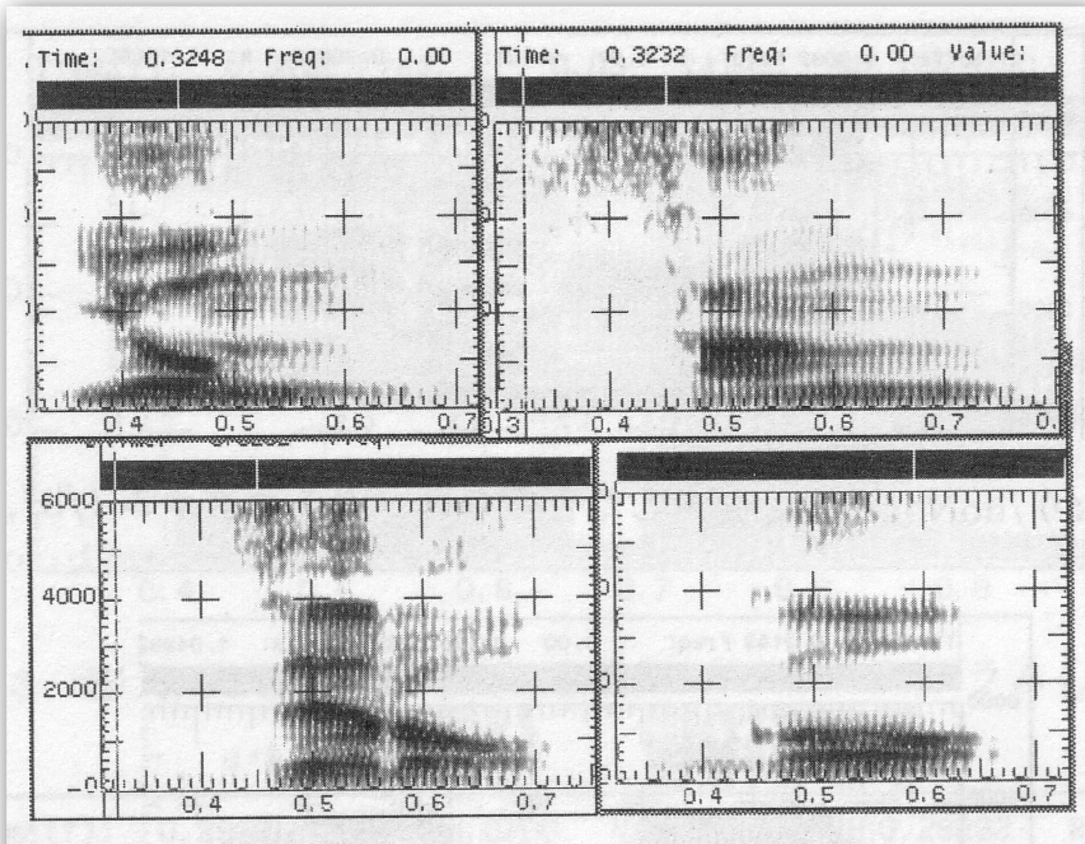


図 2.19 数字音声のスペクトログラム

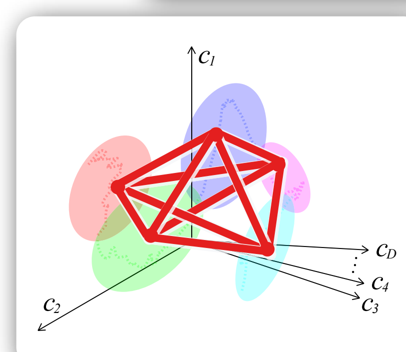
- This is the task that is done by a speech recognizer.

Title of each lecture

- Theme-1
  - ~~Multimedia information and humans~~
  - ~~Multimedia information and interaction between humans and machines~~
  - ~~Multimedia information used in expressive and emotional processing~~
  - ~~A wonder of sensation - synesthesia -~~

- Theme-2
  - ~~Speech communication technology - articulatory & acoustic phonetics -~~
  - ~~Speech communication technology - speech analysis -~~
  - Speech communication technology - speech recognition -
  - Speech communication technology - speech synthesis -

- Theme-3
  - A new framework for “human-like” speech machines #1
  - A new framework for “human-like” speech machines #2
  - A new framework for “human-like” speech machines #3
  - A new framework for “human-like” speech machines #4



Speech Communication Tech.

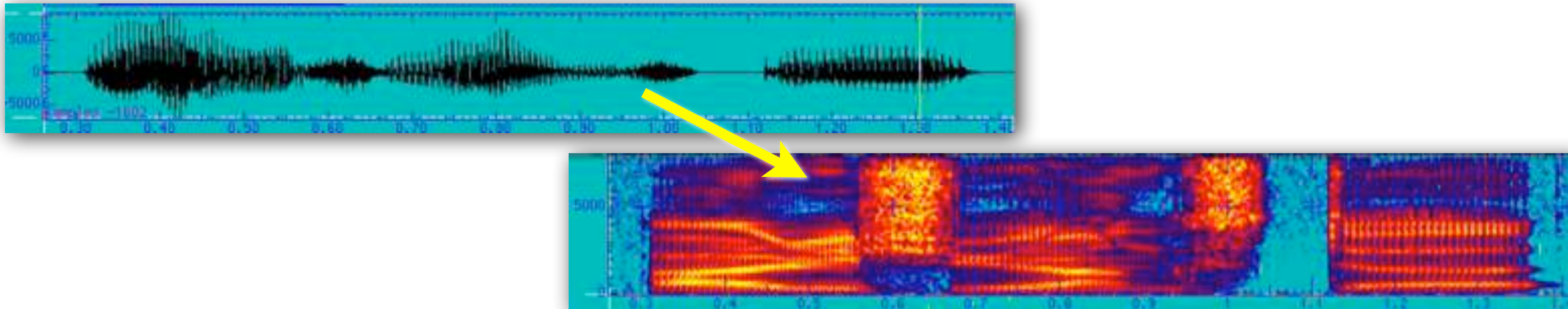
- Speech recognition -

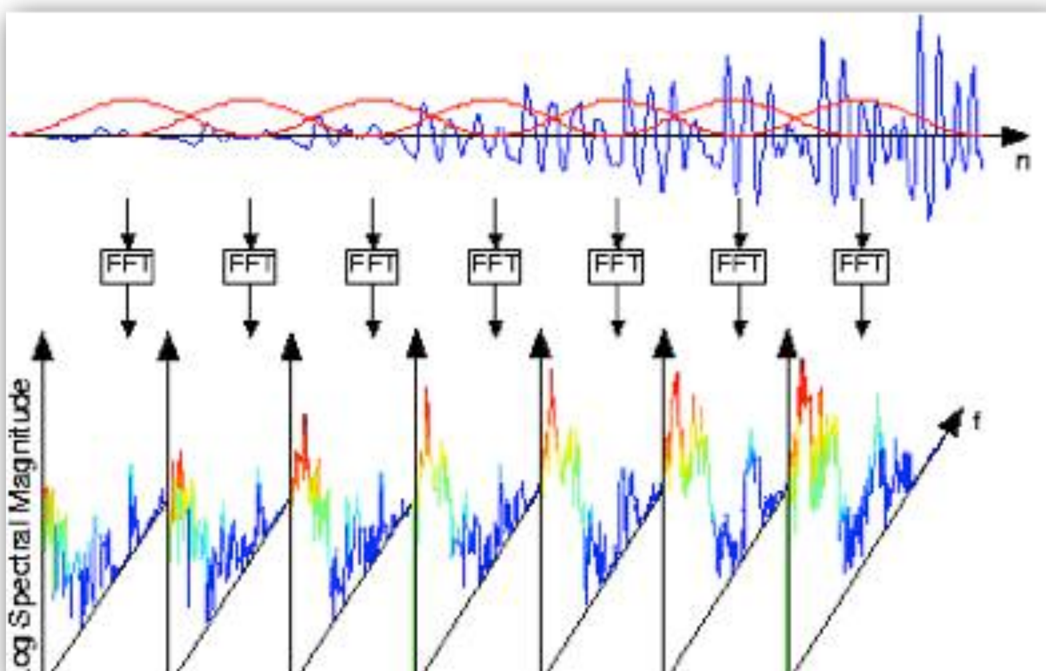
Nobuaki Minematsu

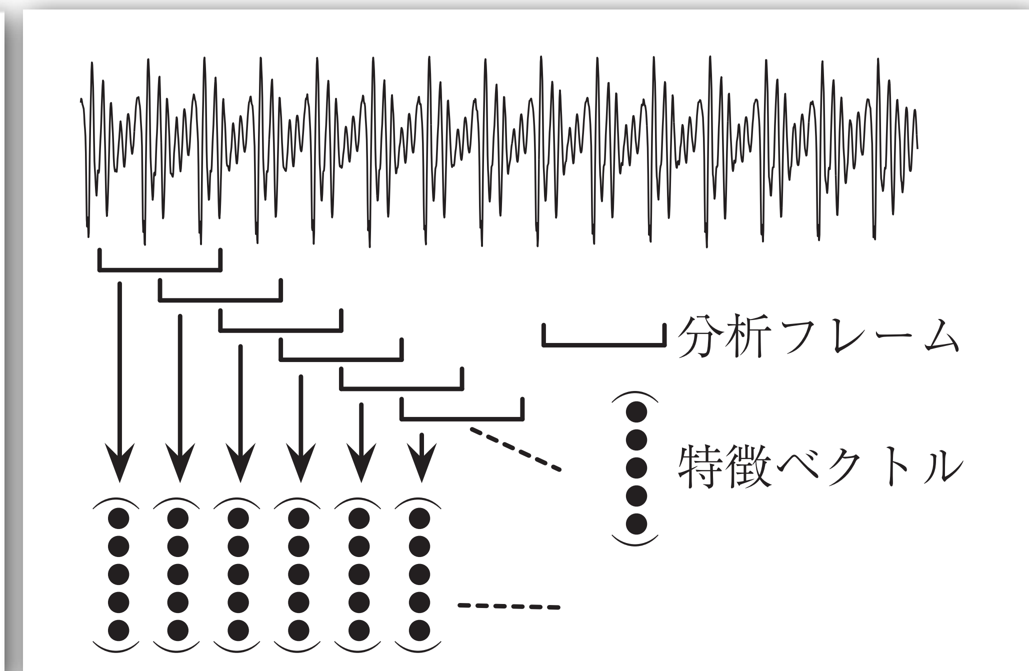
Today's menu

- Fundamentals of statistical speech recognition
- Acoustic models (HMM) for speech recognition
- From word-based HMMs to phoneme-based HMMs
- From HMM-GMM to HMM-DNN
- Speech recognition using network grammars
- Speech recognition using N-grams
- Speech recognition using NN-based language models

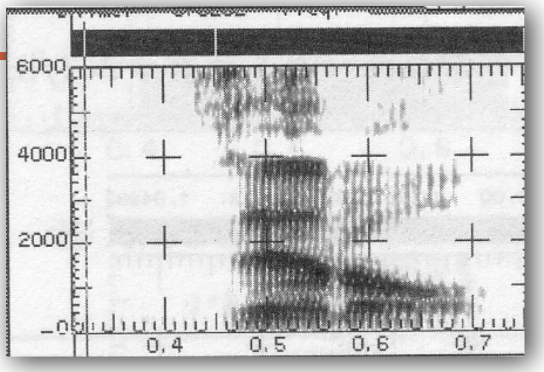
Waveforms --> spectrums --> sequence of feature vectors







Difficulty of ASR

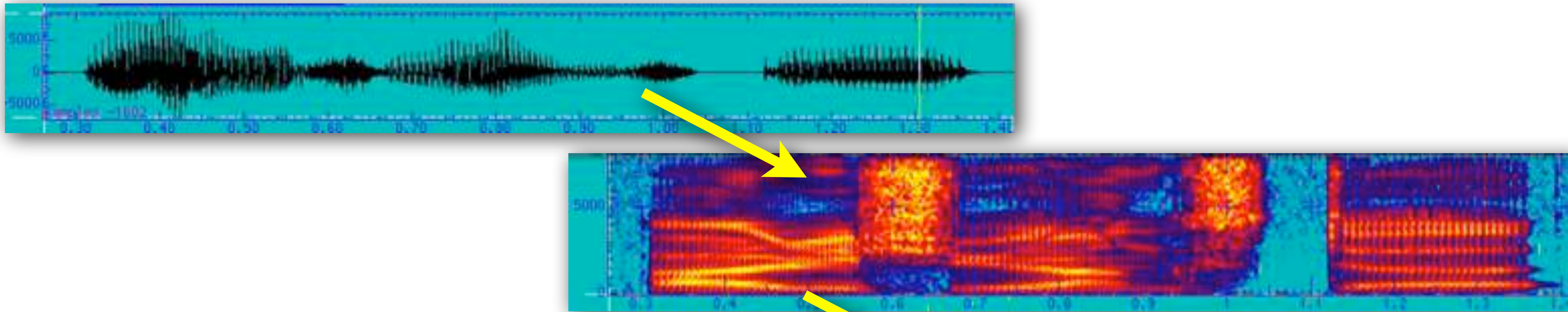


- Task of Automatic Speech Recognition (ASR)
  - Automatic identification of what is said by any speaker
    - Input: spectrum sequence
    - Output: word sequence
- Acoustic difficulty of ASR
  - A large acoustic diversity of one and the same linguistic content, e.g. word
    - Factors of the diversity: speaker identity, age, gender, speaking style, channel, line, etc.
    - Not explicitly represented in the written form of language.
- Linguistic difficulty of ASR
  - We're not speaking like the written form of language.
    - How to characterize word sequences in naturally and spontaneously generated speech?
    - How to treat ungrammatical utterances, word fragments, filled pauses, etc ?

A well-known strategy for diversity

- Statistical framework of ASR
  - Solution of argmax_{w} P(w|o)
    - P(w): prior knowledge of what kind of words or phonemes are likely to be observed.
    - P(w|o): conditional probability of word observation, given acoustic observation of o.
      - (specific) o --> w1, w2, w3, ...? o --> p1, p2, p3, ...?
      - Data collection is very difficult to characterize or formulate P(w|o) directly.
  - Use of the Bayesian rule
    - $$P(w|o) = \frac{P(w,o)}{P(o)} = \frac{P(o|w)P(w)}{\sum_w P(o,w)} = \frac{P(o|w)P(w)}{\sum_w P(o|w)P(w)}$$
    - The denominator is independent of w.
    - Maximization of P(w|o) in terms of w is equal to that of P(o|w)P(w) (=P(o,w))
  - Solution of argmax_{w} P(o|w) P(w)
    - P(w): can be estimated from a large text corpus.
    - P(o|w): conditional probability of acoustic observation, given intended content of w.
      - (specific) w --> o1, o2, o3, ...? p --> o1, o2, o3, ...?
      - This data collection is possible enough by asking many speakers to say w or p !!
    - P(o|w): acoustic model, P(w): linguistic model
      - Separate two models and a program that can search for the word sequence that maximizes P(o,w)

Waveforms --> spectrums --> sequence of feature vectors



$$\boldsymbol{o_1, o_2, o_3, \ldots, o_t, \ldots, o_T}$$

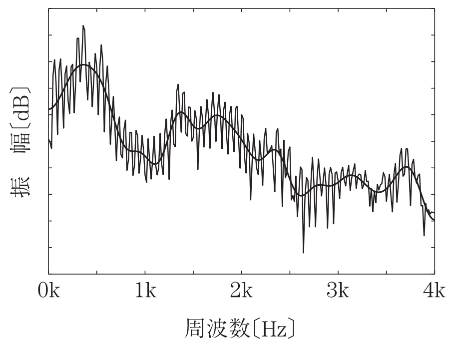
$$\arg\max_w P(w_1, w_2, \ldots, w_N | o_1, \ldots, o_t, \ldots, o_T) =$$

$$\arg\max_w P(o_1, \ldots, o_t, \ldots, o_T | w_1, w_2, \ldots, w_N) P(w_1, w_2, \ldots, w_N)$$

o : cepstrum vector

Cep. distortion and DTW

- Cepstrum vector = spectrum envelope

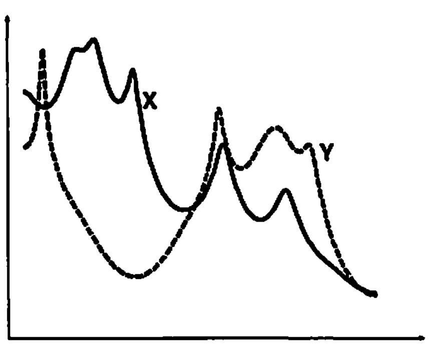


$$\longrightarrow \vec{c}$$

- 2 cepstrum vectors always satisfy the following equation.
  - log|Sn|, log|Tn|: 2 spectrums
  - log|S'n|, log|T'n|: 2 spectrum envelopes that are characterized by M cepstrums.
  - Euclid distance of cepstrums has a clear physical meaning.

$$D_n = \left(\log|S'_n| - \overline{\log|S_n|}\right) - \left(\log|T'_n| - \overline{\log|T_n|}\right)$$

$$2\sum_{k=1}^{M}\left(c_k^S - c_k^T\right)^2 = \frac{1}{N}\sum_{i=0}^{N-1} D_n^2$$

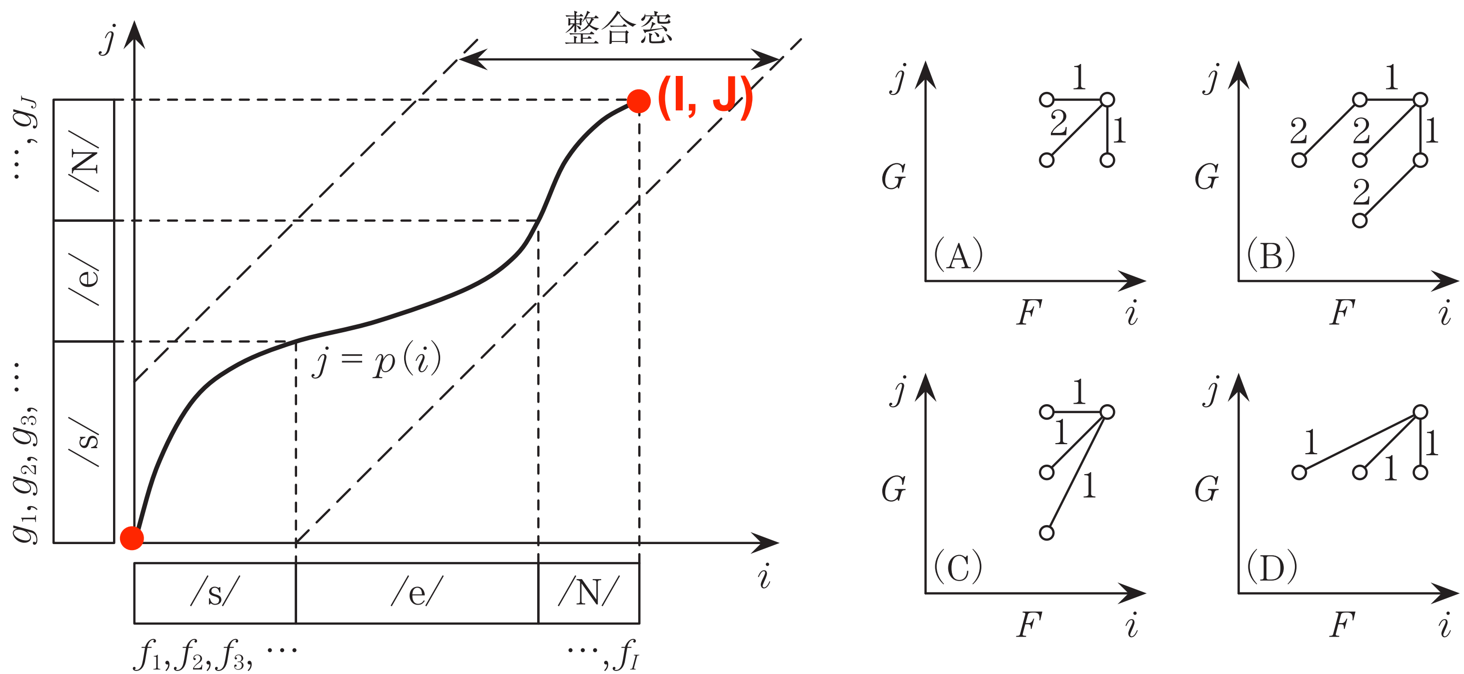


Cep. distortion and DTW

- Dynamic Time Warping
  - Temporal alignment between two utterances of the same content
  - Temporal alignment between two utterances of different contents
  - Finding the best path that minimizes the accumulated distortion along that path.

$$\min_{p}\left[\frac{1}{Z}\sum_{i=1}^{I} d(f_i, g_{p(i)})\right]$$

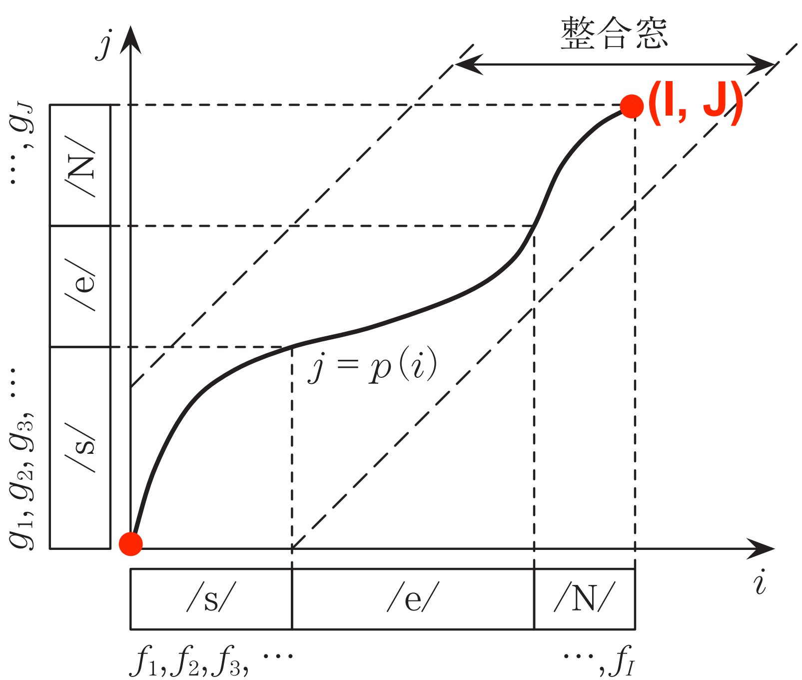
- Local distortion: $d(f_i, g_j)$ = Euclid distance of the corresponding two cepstrum vectors.

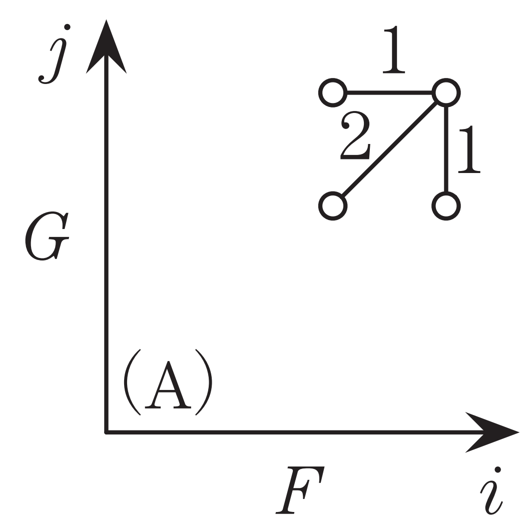


Cep. distortion and DTW

- Total distortion accumulated up to point (i,j) = D(i,j)
  - d(i,j) = local distortion (distance) between f_i and g_j.

$$D(i,j) = \min \begin{bmatrix} D(i,j-1) + d(i,j) \\ D(i-1,j-1) + 2d(i,j) \\ D(i-1,j) + d(i,j) \end{bmatrix} \rightarrow \min_p \left[\frac{1}{Z} \sum d(i,p(i)) \right] = \frac{1}{I+J-1} D(I,J)$$





Today's menu

- Fundamentals of statistical speech recognition
- Acoustic models (HMM) for speech recognition
- From word-based HMMs to phoneme-based HMMs
- From HMM-GMM to HMM-DNN
- Speech recognition using network grammars
- Speech recognition using N-grams
- Speech recognition using NN-based language models

A well-known strategy for diversity

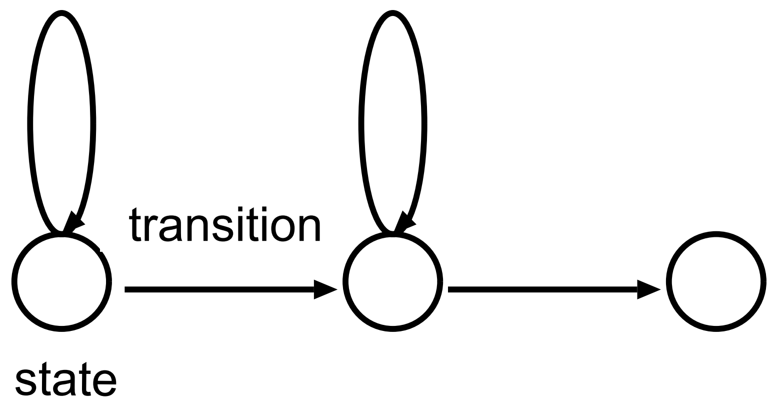
- Statistical framework of ASR
  - Solution of argmax_{w} P(w|o)
    - P(w): prior knowledge of what kind of words or phonemes are likely to be observed.
    - P(w|o): conditional probability of word observation, given acoustic observation of o.
      - (specific) o --> w1, w2, w3, ...?   o --> p1, p2, p3, ...?
      - Data collection is very difficult to characterize or formulate P(w|o) directly.
  - Use of the Bayesian rule
    - $$P(w|o) = \frac{P(w,o)}{P(o)} = \frac{P(o|w)P(w)}{\sum_w P(o,w)} = \frac{P(o|w)P(w)}{\sum_w P(o|w)P(w)}$$
    - The denominator is independent of w.
    - Maximization of P(w|o) in terms of w is equal to that of P(o|w)P(w) (=P(o,w))
  - Solution of argmax_{w} P(o|w) P(w)
    - P(w): can be estimated from a large text corpus.
    - P(o|w): conditional probability of acoustic observation, given intended content of w.
      - (specific) w --> o1, o2, o3, ...? p --> o1, o2, o3, ...?
      - This data collection is possible enough by asking many speakers to say w or p !!
    - P(o|w): acoustic model, P(w): linguistic model
      - Separate two models and a program that can search for the word sequence that maximizes P(o,w)

Markov Process

$$P(x_n|x_{n-1},\cdots,x_1) = P(x_n|x_{n-1})$$

- Signal at t = n depends only on the previous signal (t=n-1).
- If signal at t = n-1 is known, signals at t < n-1 have no effect on the next signal at t = n.

Hidden Markov Process



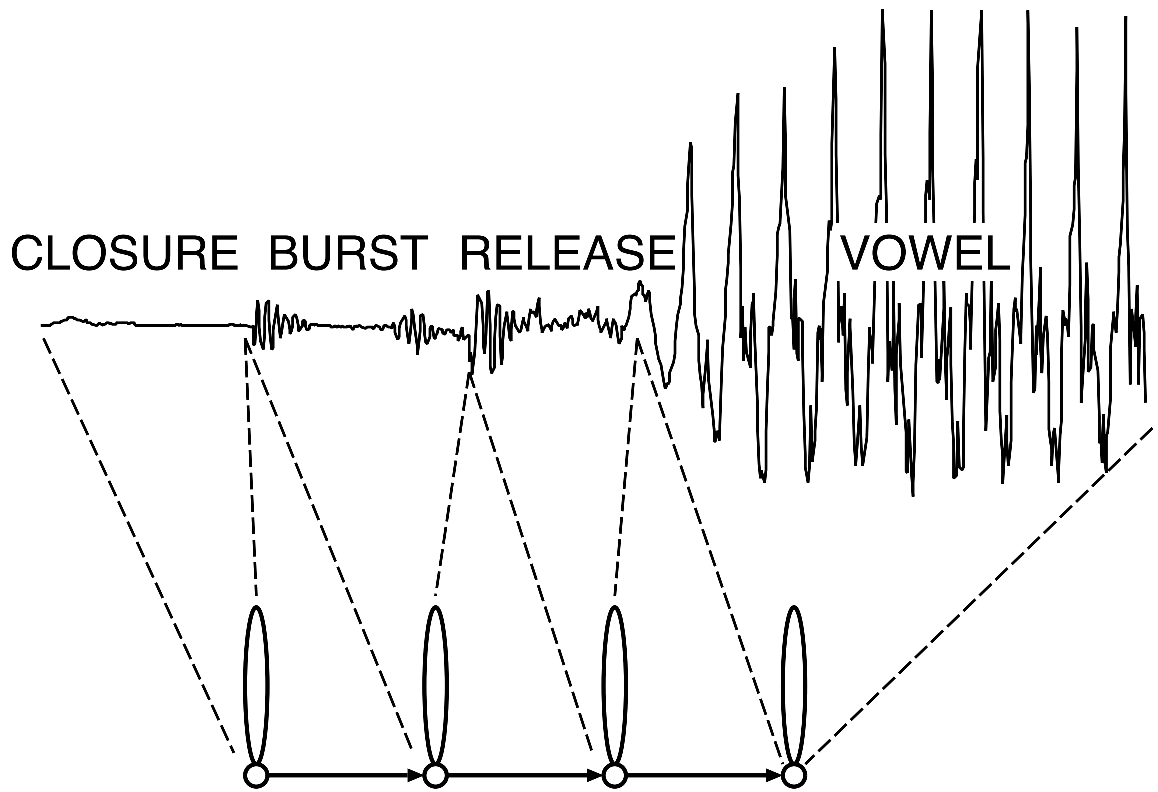
$$P(x_n | \underbrace{x_{n-1}, \cdots, x_1}_{\text{previous observations}}) = P(x_n | \underbrace{S_n}_{\text{current state}})$$

Observation sequence : $x_1, x_2, \cdots, x_n, \cdots$

(Hidden) state sequence : $S_1, S_2, \cdots, S_n, \cdots$

- Previous observations cannot determine the current state uniquely.
- Signals (features) are observed but states are hidden.

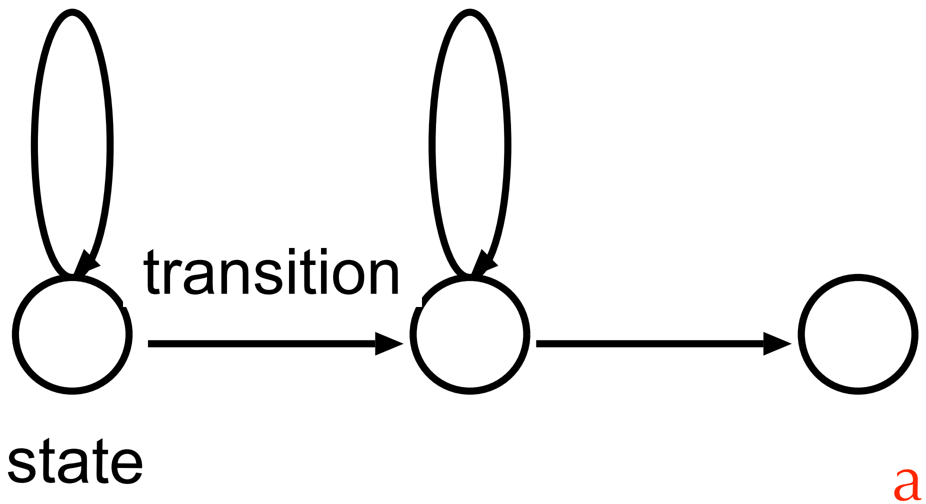
HMM as generative model



Probabilistic generative model

State transition is modeled as transition probability.
Output features are modeled as output probability.

Parameters of HMM



single Gaussian or
a mixture of Gaussians

- Transition prob. : $P(s_{t+1}|s_t = i) = \{a_{1i}, a_{2i}, ..., a_{ji}, ..., a_{Si}\}$
- Output prob. : $P(o|s_t = i) = b_i(o) \; = \mathcal{N}(o; \mu_i, \Sigma_i)$

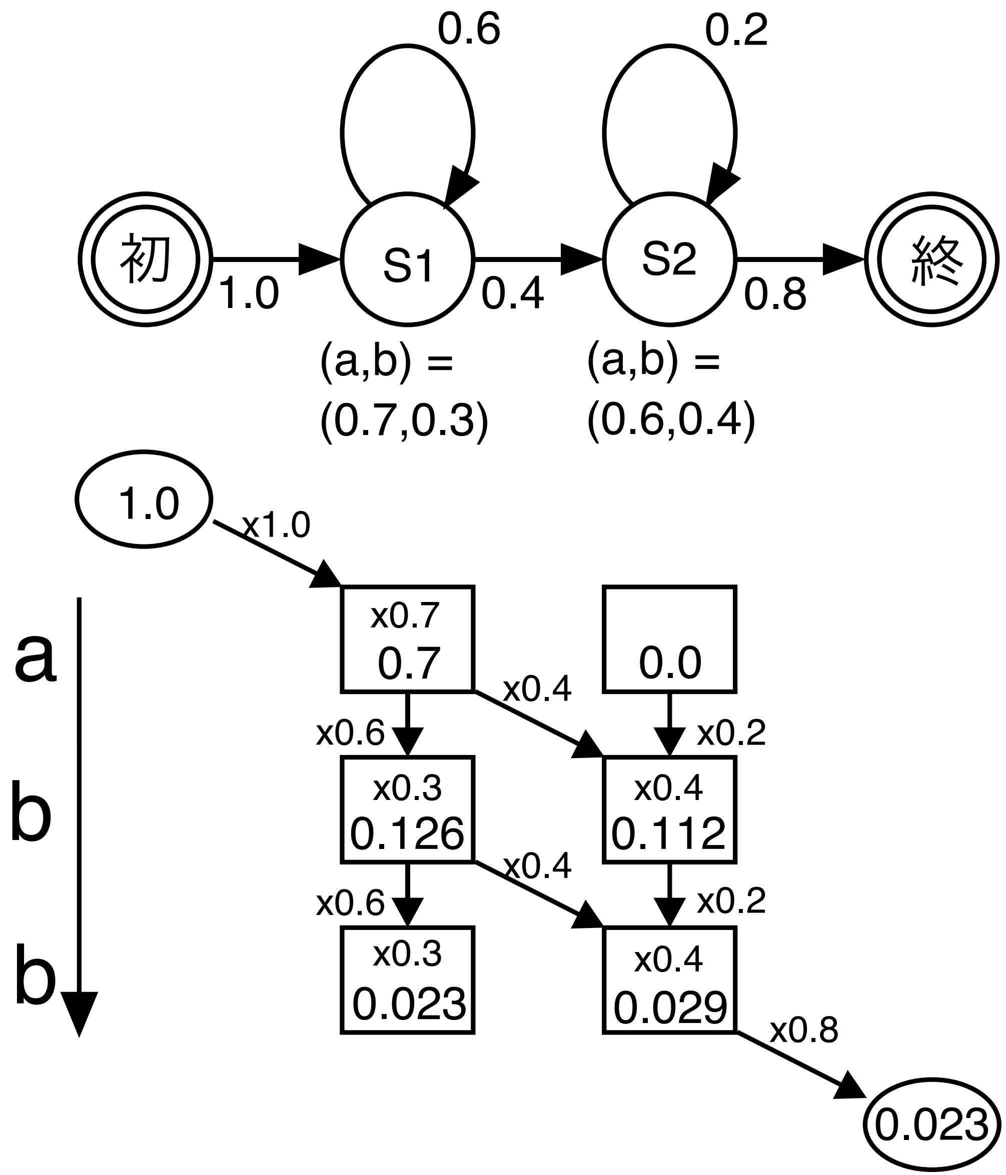
Forward prob.

$$\alpha_j(t) = P(o_1, \cdots, o_t, s(t) = j|M) \qquad = \sum_i \alpha_i(t-1)a_{ij}b_j(o_t)$$

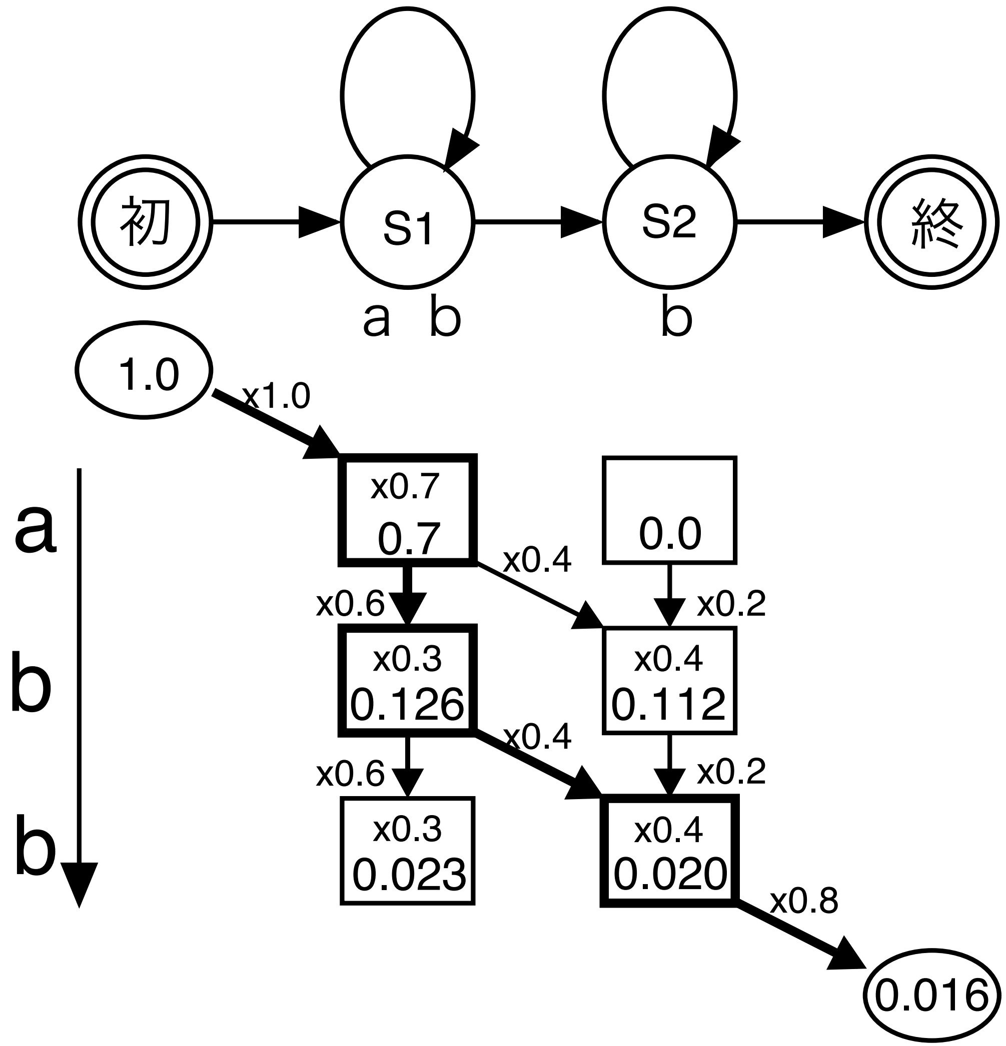
Backward prob.

$$\beta_j(t) = P(o_{t+1}, \cdots, o_T|s(t) = j, M) \qquad = \sum_i a_{ji}b_i(o_{t+1})\beta_i(t+1)$$

Output probability of observation sequence (Trellis)

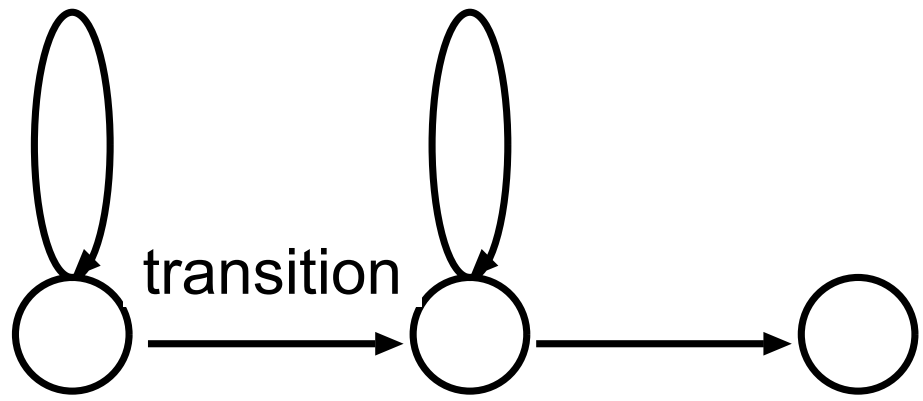


Output probability of observation sequence (Viterbi)



The maximum likelihood path is only adopted.

Parameters of HMM



single Gaussian or
a mixture of Gaussians

- Transition prob. : $P(s_{t+1}|s_t = i) = \{a_{1i}, a_{2i}, ..., a_{ji}, ..., a_{Si}\}$
- Output prob. : $P(o|s_t = i) = b_i(o) \; = \mathcal{N}(o; \mu_i, \Sigma_i)$

Forward prob.

$$\alpha_j(t) = P(o_1, \cdots, o_t, s(t) = j|M) \qquad = \sum_i \alpha_i(t-1) a_{ij} b_j(o_t)$$

Backward prob.

$$\beta_j(t) = P(o_{t+1}, \cdots, o_T|s(t) = j, M) \qquad = \sum_i a_{ji} b_i(o_{t+1}) \beta_i(t+1)$$

Estimation of HMM parameters

Estimation is done iteratively by updating old parameters.

- Forward prob.

$$\alpha_j(t) = P(o_1, \cdots, o_t, s(t) = j|M) \qquad = \sum_i \alpha_i(t-1)a_{ij}b_j(o_t)$$

- Backward prob.

$$\beta_j(t) = P(o_{t+1}, \cdots, o_T|s(t) = j, M) \quad = \sum_i a_{ji}b_i(o_{t+1})\beta_i(t+1)$$

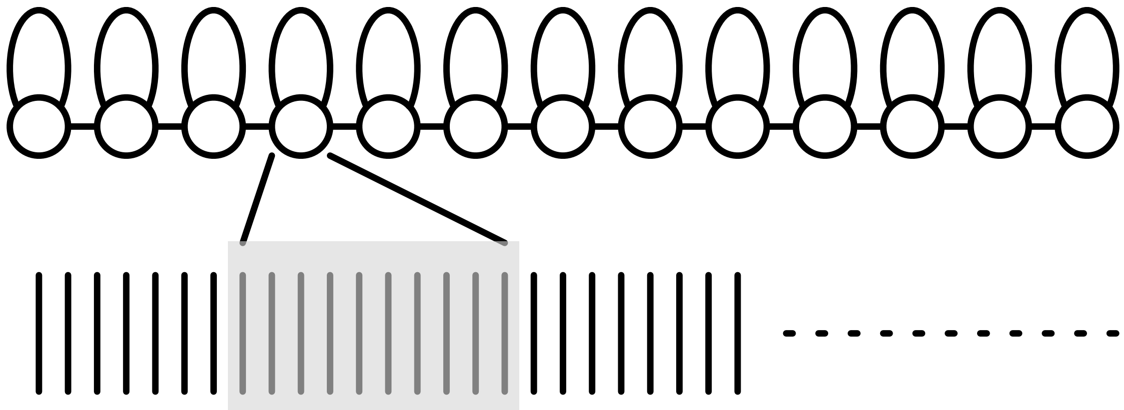
$\rightarrow \quad \alpha_j(t)\beta_j(t) = P(O, s(t) = j|M)$

$\rightarrow \quad P(s(t) = j|O, M) = \dfrac{\alpha_j(t)\beta_j(t)}{P(O|M)} = \dfrac{\alpha_j(t)\beta_j(t)}{\alpha_N(T)} = L_j(t)$

$\rightarrow$ Represents how strongly o_t is associated with state j.

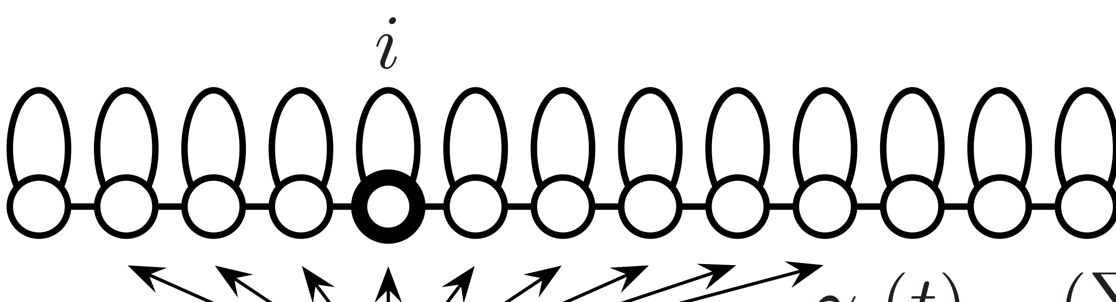
$\rightarrow \quad \hat{\mu}_j = \dfrac{\sum_t L_j(t)\cdot o_t}{\sum_t L_j(t)} = \dfrac{\sum_t \alpha_j(t)\beta_j(t)\cdot o_t}{\sum_t \alpha_j(t)\beta_j(t)}$ $\qquad P(O|\hat{M}) \geq P(O|M)$

Estimation of HMM parameters

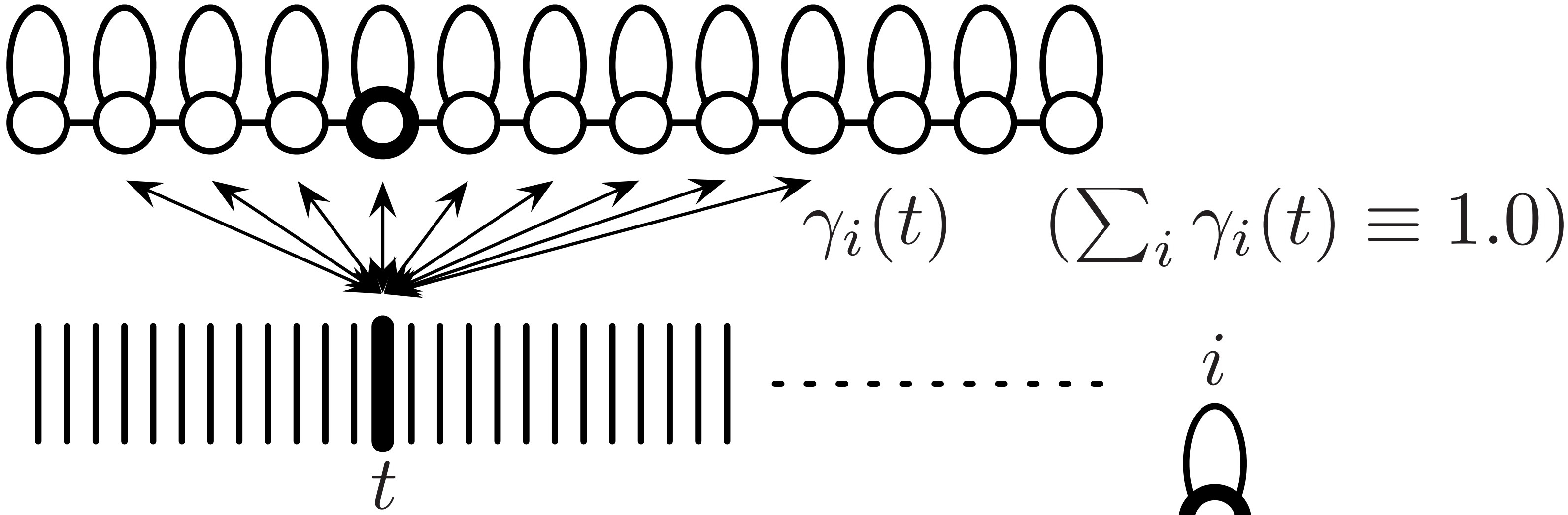


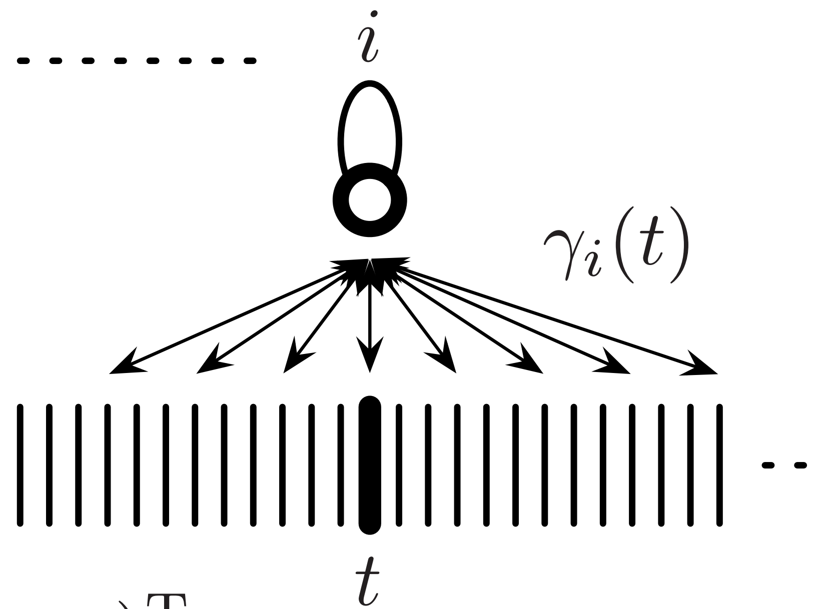
$$\mu = \frac{1}{T}\sum_t o_t = \frac{\sum_t \frac{1}{T} o_t}{\sum_t \frac{1}{T}}$$

$$\Sigma = \frac{1}{T}\sum_t (o_t - \mu)(o_t - \mu)^{\mathrm{T}}$$



$\alpha_i(t) \quad (\sum \alpha_i(t) = 1.0)$





$$\hat{\mu}_i = \frac{\sum_t \gamma_i(t) o_t}{\sum_t \gamma_i(t)}$$

$$\hat{\Sigma}_i = \frac{\sum_t \gamma_i(t)(o_t - \mu)(o_t - \mu)^{\mathrm{T}}}{\sum_t \gamma_i(t)}$$

$$P(O|\hat{M}) \geq P(O|M)$$

Estimation of HMM parameters

Estimation is done iteratively by updating old parameters.

- Forward prob.

$$\alpha_j(t) = P(o_1, \cdots, o_t, s(t) = j|M) \qquad = \sum_i \alpha_i(t-1) a_{ij} b_j(o_t)$$

- Backward prob.

$$\beta_j(t) = P(o_{t+1}, \cdots, o_T|s(t) = j, M) \quad = \sum_i a_{ji} b_i(o_{t+1}) \beta_i(t+1)$$

$\rightarrow \quad \alpha_j(t)\beta_j(t) = P(O, s(t) = j|M)$

$\rightarrow \quad P(s(t) = j|O, M) = \dfrac{\alpha_j(t)\beta_j(t)}{P(O|M)} = \dfrac{\alpha_j(t)\beta_j(t)}{\alpha_N(T)} = L_j(t)$

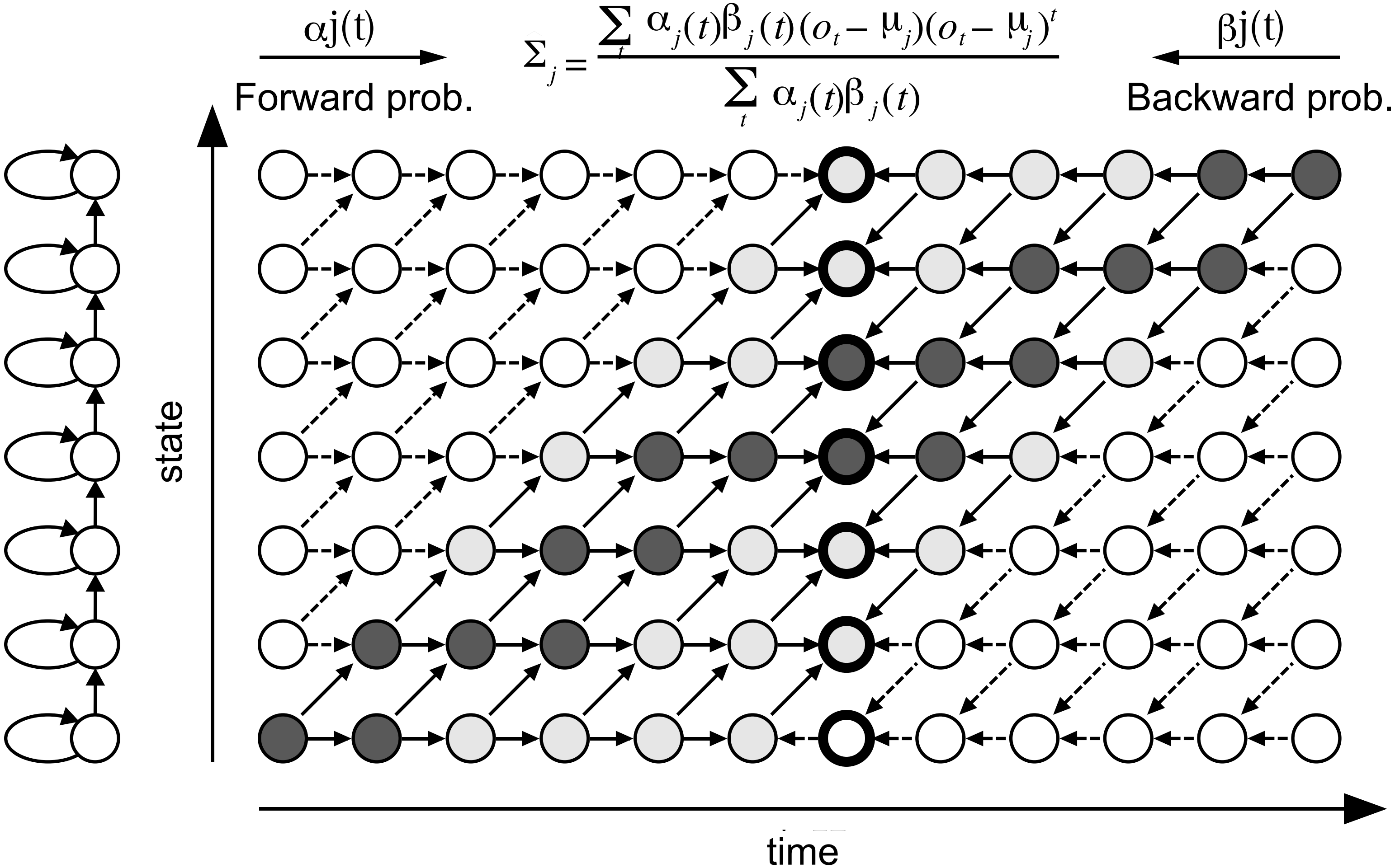
$\rightarrow$ Represents how strongly o_t is associated with state j.

$\rightarrow \quad \hat{\mu}_j = \dfrac{\sum_t L_j(t) \cdot o_t}{\sum_t L_j(t)} = \dfrac{\sum_t \alpha_j(t)\beta_j(t) \cdot o_t}{\sum_t \alpha_j(t)\beta_j(t)}$ $\qquad P(O|\hat{M}) \geq P(O|M)$

Estimation of HMM parameters

$$\mu_j = \frac{\sum_t \alpha_j(t)\beta_j(t)\, o_t}{\sum_t \alpha_j(t)\beta_j(t)}$$

$$\Sigma_j = \frac{\sum_t \alpha_j(t)\beta_j(t)(o_t - \mu_j)(o_t - \mu_j)^t}{\sum_t \alpha_j(t)\beta_j(t)}$$



Estimation of HMM parameters

- When the number of training data is 1,

$$\hat{\mu}_j = \frac{\sum_t L_j(t) \cdot o_t}{\sum_t L_j(t)}, \quad \hat{\Sigma}_j = \frac{\sum_t L_j(t) \cdot (o_t - \mu_j)(o_t - \mu_j)^t}{\sum_t L_j(t)}$$

- When the number of training data is R (>1),

$$\hat{\mu}_j = \frac{\sum_r \left[\sum_t L_j^r(t) \cdot o_t^r\right]}{\sum_r \left[\sum_t L_j^r(t)\right]} = \frac{\sum_r \frac{1}{P^r}\left[\sum_t \alpha_j^r(t)\beta_j^r(t) \cdot o_t^r\right]}{\sum_r \frac{1}{P^r}\left[\sum_t \alpha_j^r(t)\beta_j^r(t)\right]}$$

$$\hat{\Sigma}_j = \frac{\sum_r \left[\sum_t L_j^r(t) \cdot (o_t^r - \mu_j)(o_t^r - \mu_j)^t\right]}{\sum_r \left[\sum_t L_j^r(t)\right]} = \cdots$$

#speakers = several thousands

Recognition of isolated words

$$\arg\max_W P(W|O) = \arg\max_W P(O|W)P(W) = \arg\max_W P(O|W)$$

if prior probability of W is evenly distributed.

$$\arg\max_M P(O|M) = \arg\max_M \left\{\sum_X P(O,X|M)\right\}$$

$$\downarrow \quad (X = \text{path})$$

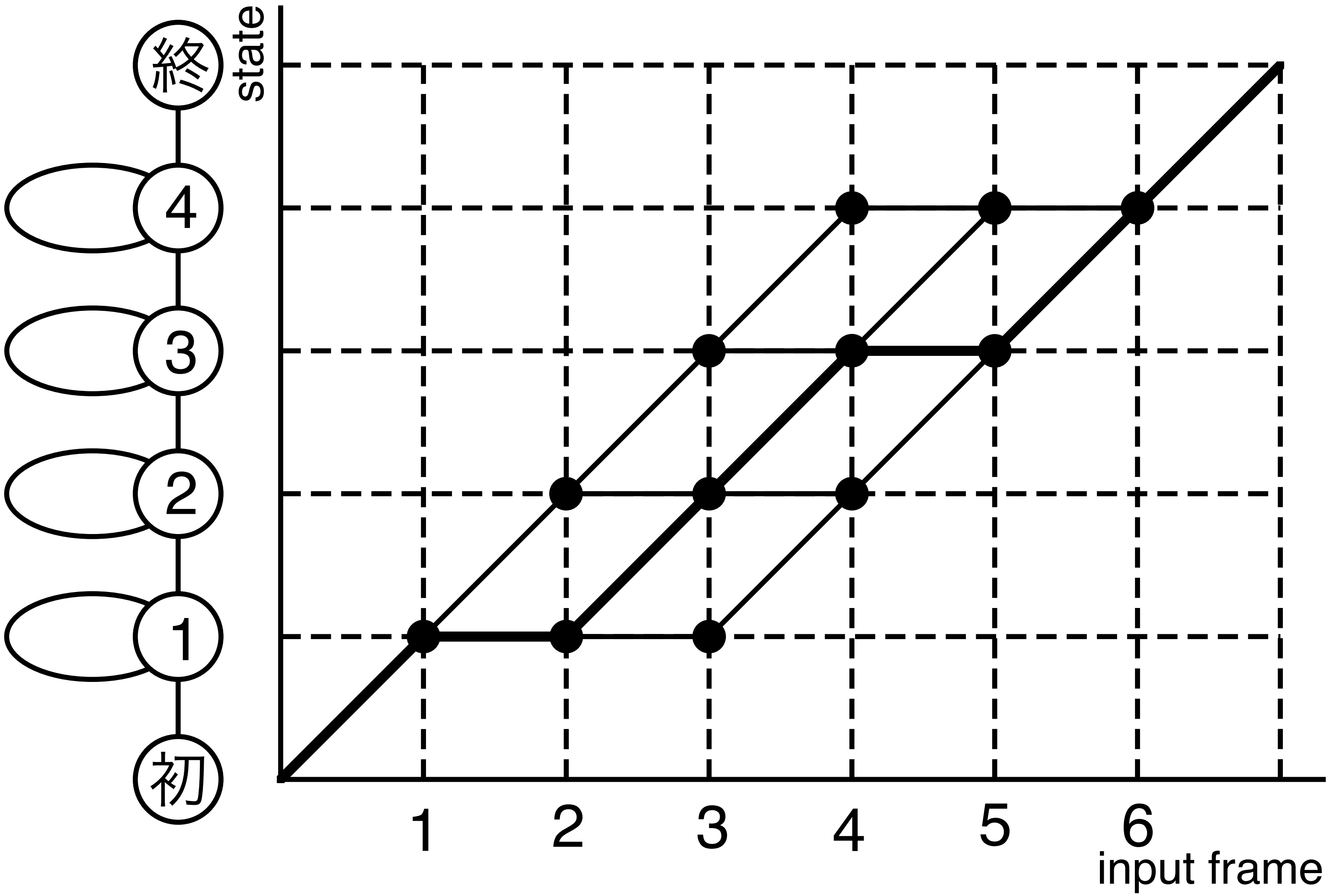
$$\arg\max_M \hat{P}(O|M) = \arg\max_M \left\{\max_X P(O,X|M)\right\}$$

$$\alpha_j(t) = \sum_i \alpha_i(t-1)a_{ij}b_j(o_t), \quad (\alpha_N(T) \equiv P(O|M))$$

$$\downarrow$$

$$\phi_j(t) = \max_i \phi_i(t-1)a_{ij}b_j(o_t), \quad (\phi_N(T) \equiv \hat{P}(O|M))$$

Recognition of isolated words



Search for the maximum likelihood path

Today's menu

- Fundamentals of statistical speech recognition
- Acoustic models (HMM) for speech recognition
- From word-based HMMs to phoneme-based HMMs
- From HMM-GMM to HMM-DNN
- Speech recognition using network grammars
- Speech recognition using N-grams
- Speech recognition using NN-based language models

Phonemes

The minimum units of spoken language

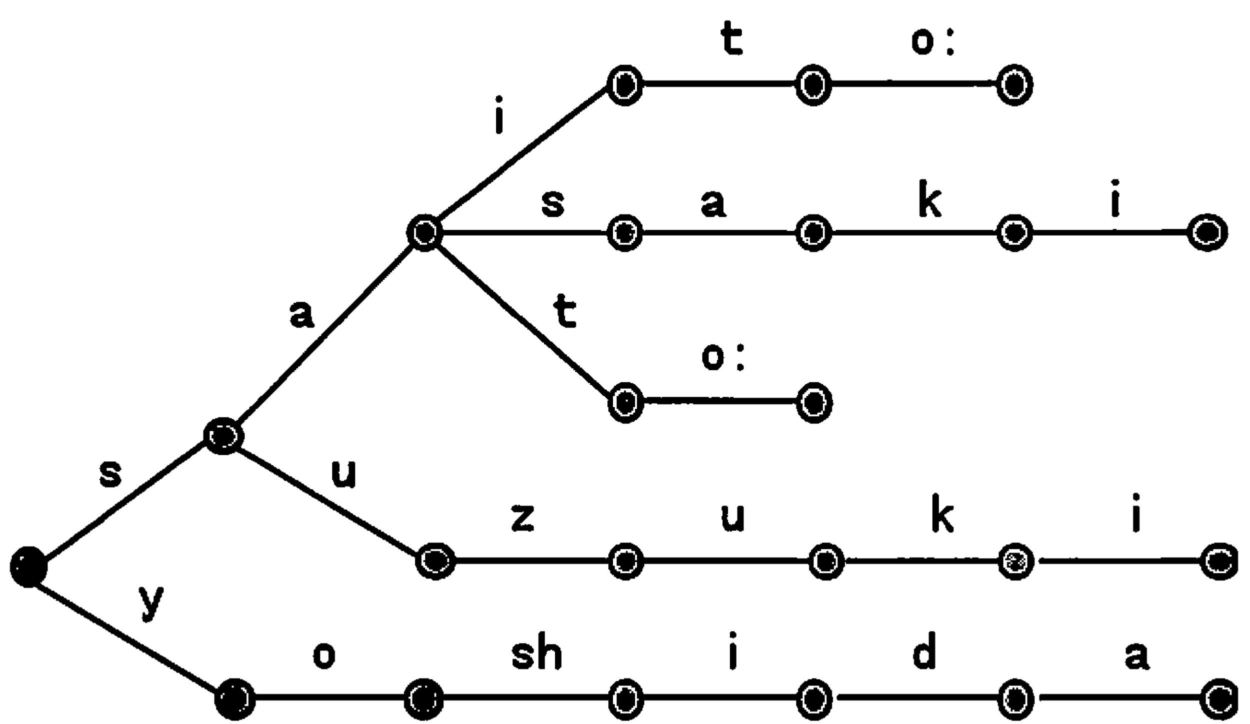
Vowels	short vowels	a, i, u, e, o
	long vowels	a:, i:, u:, e:, o:
Consonants	plosives	b, d, g, p, t, k
	fricatives	s, sh, z, j, f, h
	affricates	ch, ts
	拗音:	ky, py, ..
	semi-vowels	r, w, y
	nasals	m, n, N

Word lexicon (word dictionary)

Examples required for automated call centers

鈴木	s u z u k i
佐藤	s a t o:
吉田	y o sh i d a
さん	s a N
総務	s o: m u
営業	e: gy o:
課長	k a ch o:
の	n o
お願いします	o n e g a i sh i m a s u

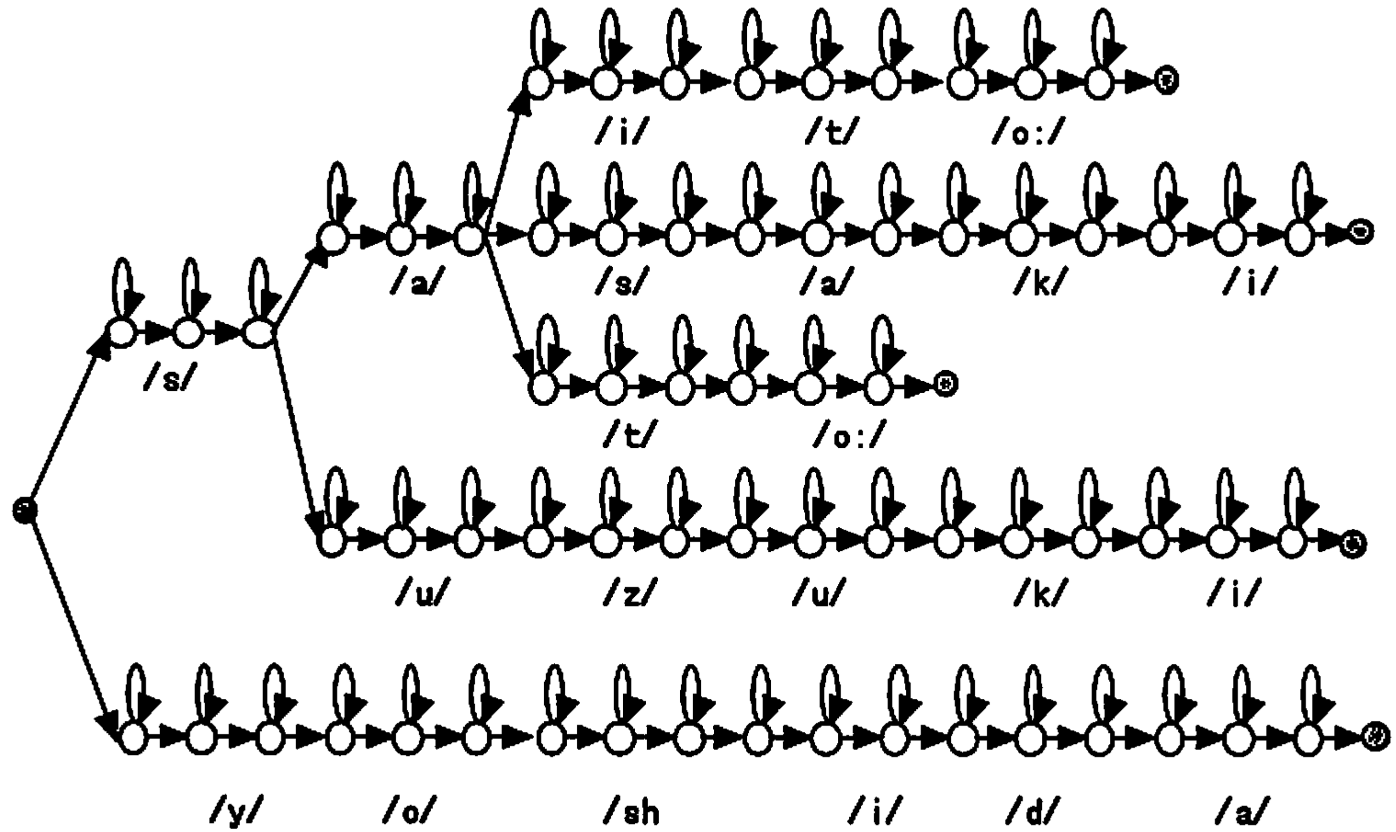
Tree lexicon (compact representation of the words)



The following words are stored as a tree.

saito: (斉藤), sasaki (佐々木), sato: (佐藤)
suzuki (鈴木) , yoshida (吉田)

Tree-based lexicon using phoneme HMMs



Generation of state-based network containing all the candidate words

Coarticulation and context-dependent phone models

Acoustic features of a specific kind of phone depends on its phonemic context.

model of /k/ = *-k+* = a-k+i a-k+e a-k+a a-k+u a-k+o ····· e-k+o i-k+o

monophone

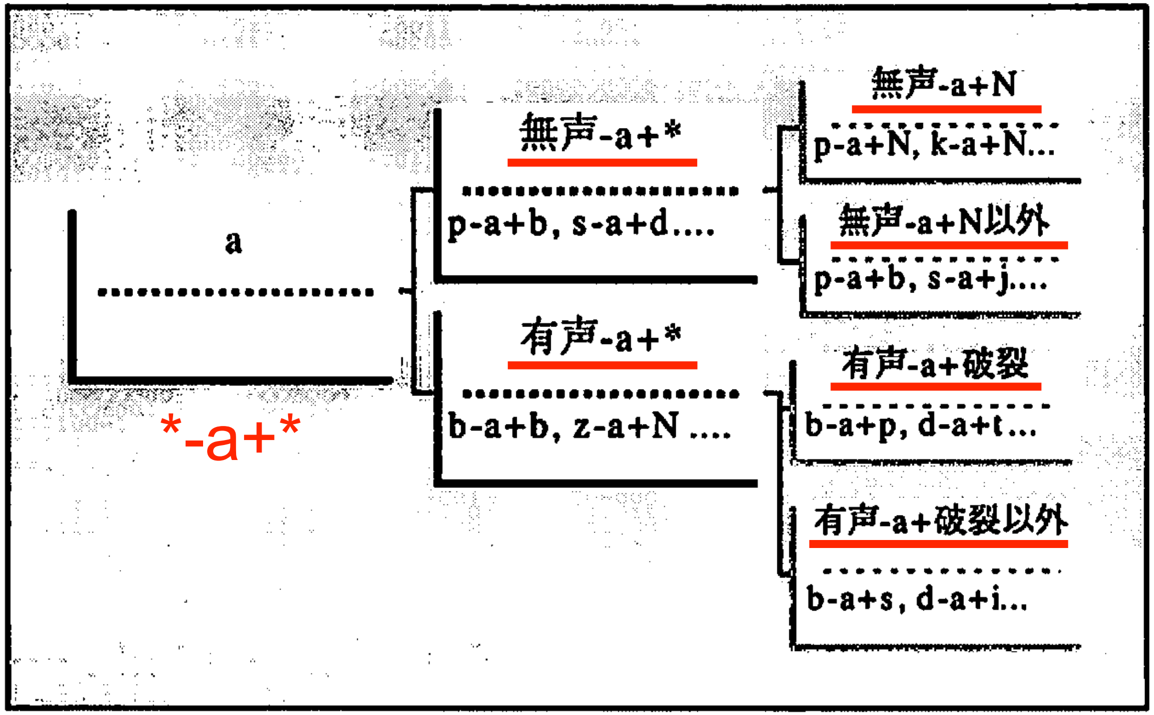
model of /k/ preceded by /a/ and succeeded by /i/ = a-k+i

trihphone

A phoneme is defined by referring to the left and the right context (phoneme)

Clustering of phonemic contexts

Number of logically defined trihphones = N x N x N (N ≈ 40)
Clustering of the contexts to reduce #triphones.



Context clustering is done based on phonetic attributes of the left and the right phonemes.

Unit of acoustic modeling

word model	merit:	Within-word coarticulation is easy to be modeled.
	demerit:	For new words, actual utterances are needed. #models will be easily increased.
	use:	Small vocabulary speech recognition systems
phoneme model	merit:	Easy to add new words to the system.
	demerit:	Long coarticulation effect is ignored. Every word has to be represented as phonemic string.
	use:	Large vocabulary speech recognition systems

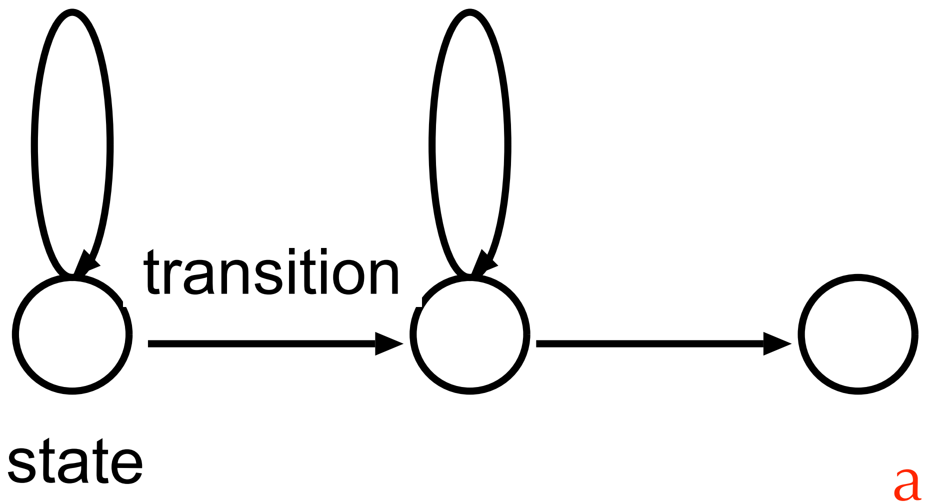
Today's menu

- Fundamentals of statistical speech recognition
- Acoustic models (HMM) for speech recognition
- From word-based HMMs to phoneme-based HMMs
- From HMM-GMM to HMM-DNN
- Speech recognition using network grammars
- Speech recognition using N-grams
- Speech recognition using NN-based language models

A well-known strategy for diversity

- Statistical framework of ASR
  - Solution of argmax_{w} P(w|o)
    - P(w): prior knowledge of what kind of words or phonemes are likely to be observed.
    - P(w|o): conditional probability of word observation, given acoustic observation of o.
      - (specific) o --> w1, w2, w3, ...?   o --> p1, p2, p3, ...?
      - Data collection is very difficult to characterize or formulate P(w|o) directly.
  - Use of the Bayesian rule
    - $$P(w|o) = \frac{P(w,o)}{P(o)} = \frac{P(o|w)P(w)}{\sum_w P(o,w)} = \frac{P(o|w)P(w)}{\sum_w P(o|w)P(w)}$$
    - The denominator is independent of w.
    - Maximization of P(w|o) in terms of w is equal to that of P(o|w)P(w) (=P(o,w))
  - Solution of argmax_{w} P(o|w) P(w)
    - P(w): can be estimated from a large text corpus.
    - P(o|w): conditional probability of acoustic observation, given intended content of w.
      - (specific) w --> o1, o2, o3, ...? p --> o1, o2, o3, ...?
      - This data collection is possible enough by asking many speakers to say w or p !!
    - P(o|w): acoustic model, P(w): linguistic model
      - Separate two models and a program that can search for the word sequence that maximizes P(o,w)

Parameters of HMM



single Gaussian or
a mixture of Gaussians

- Transition prob. : $P(s_{t+1}|s_t = i) = \{a_{1i}, a_{2i}, \ldots, a_{ji}, \ldots, a_{Si}\}$
- Output prob. : $P(o|s_t = i) = b_i(o) \;= \mathcal{N}(o; \mu_i, \Sigma_i)$

Forward prob.

$$\alpha_j(t) = P(o_1, \cdots, o_t, s(t) = j|M) \quad = \sum_i \alpha_i(t-1)a_{ij}b_j(o_t)$$

Backward prob.

$$\beta_j(t) = P(o_{t+1}, \cdots, o_T|s(t) = j, M) \quad = \sum_i a_{ji}b_i(o_{t+1})\beta_i(t+1)$$

GMM-HMM to DNN-HMM

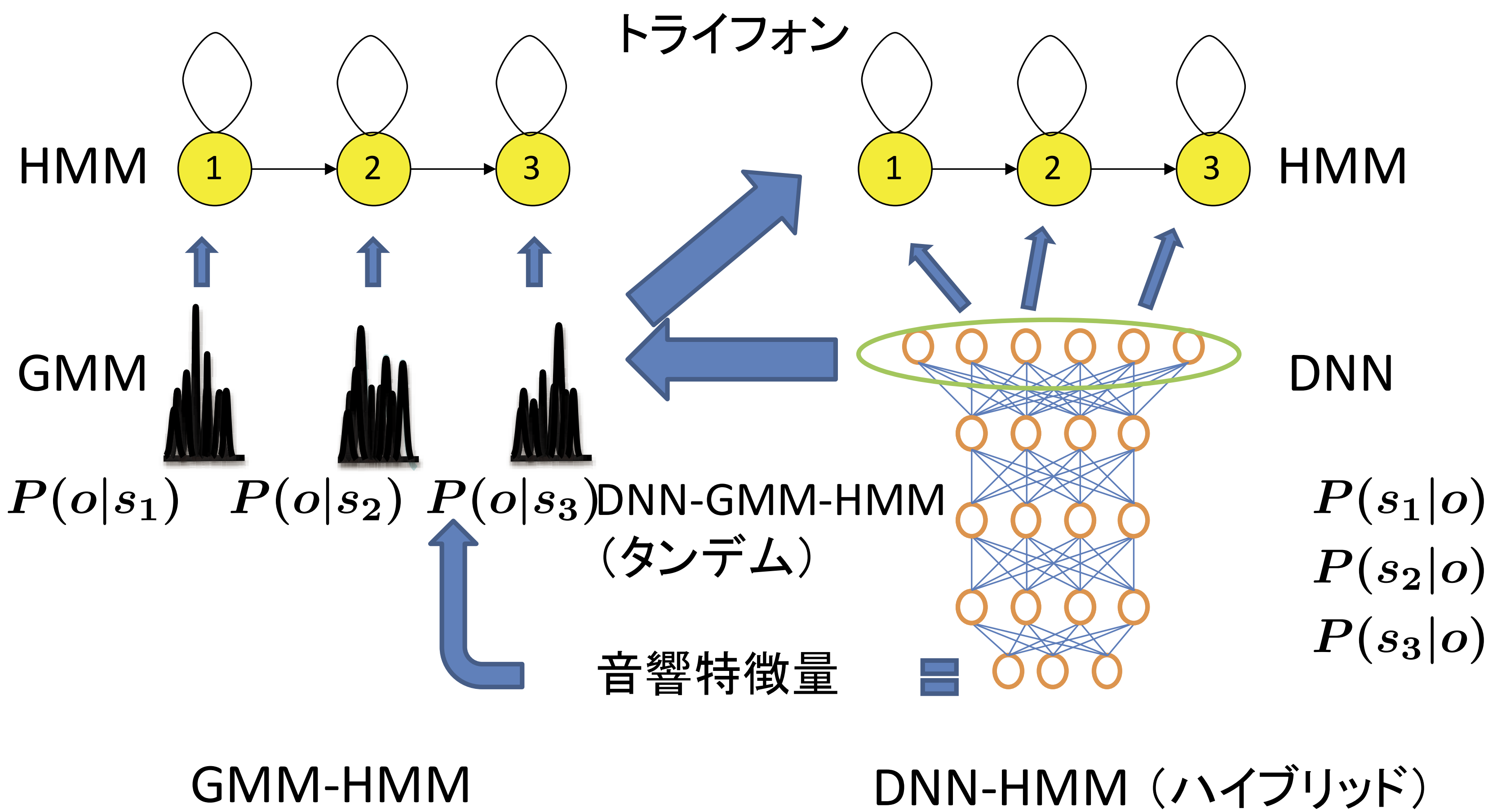
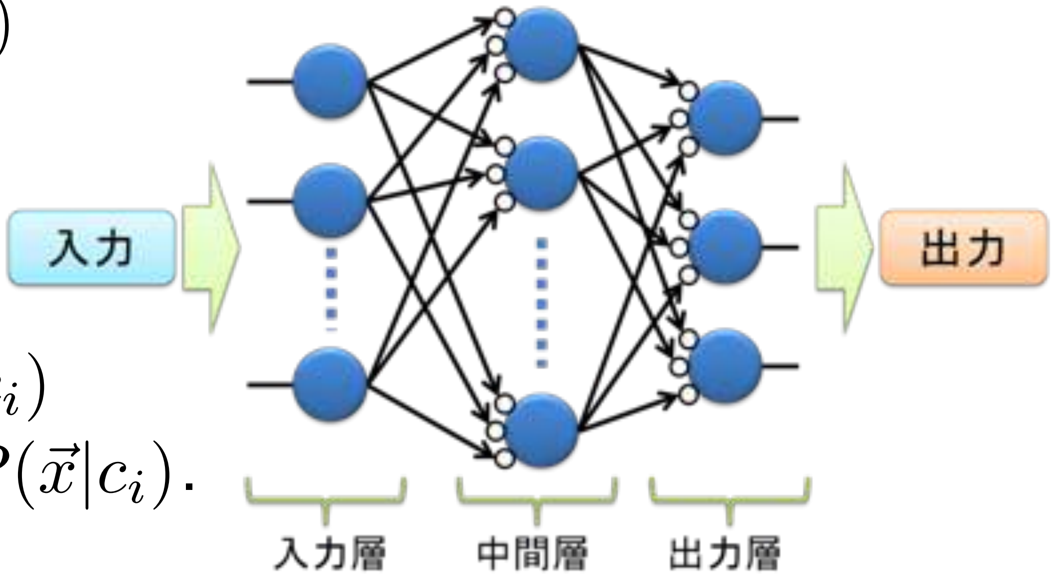


図 2　GMM-HMM と DNN-HMM

DNN as phoneme posterior calculator

- cepstrum feature $\vec{x} \longrightarrow P(c_i|\vec{x})$



- HMM-GMM is a model of $P(\vec{x}|c_i)$
  $P(c_i|\vec{x})$ has to be changed to $P(\vec{x}|c_i)$.

- The Bayesian rule, again.

$$P(\vec{x}|c_i) = \frac{P(c_i|\vec{x})P(\vec{x})}{P(c_i)}$$

Which is better, $P(\vec{x}|c_i)$ calculated by HMM-GMM or $P(\vec{x}|c_i)$ calculated by HMM-DNN with the Bayesian rule?

GMM-HMM to DNN-HMM

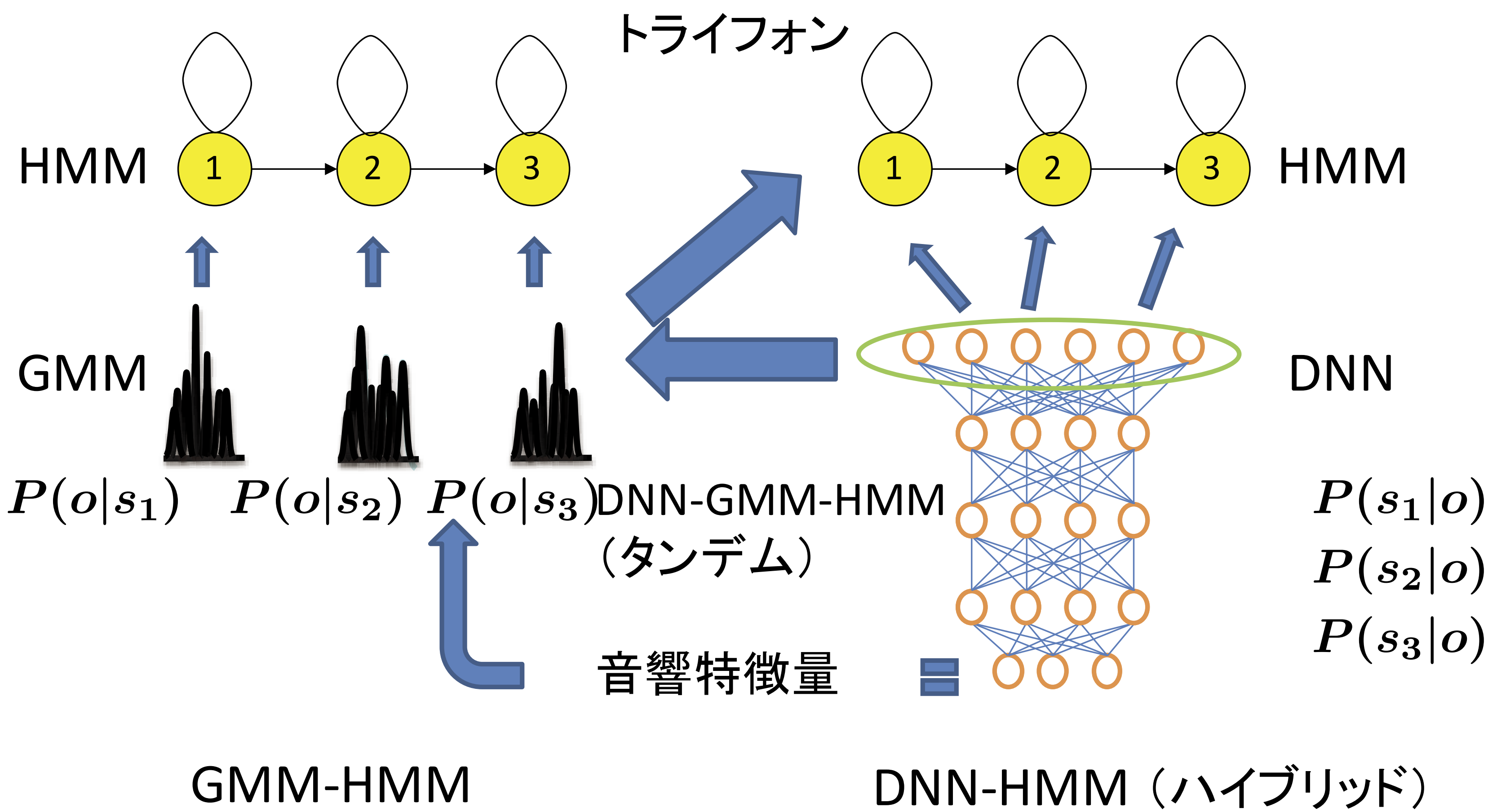


図 2　GMM-HMM と DNN-HMM

Today's menu

- Fundamentals of statistical speech recognition
- Acoustic models (HMM) for speech recognition
- From word-based HMMs to phoneme-based HMMs
- From HMM-GMM to HMM-DNN
- Speech recognition using network grammars
- Speech recognition using N-grams
- Speech recognition using NN-based language models

A well-known strategy for diversity

- Statistical framework of ASR
  - Solution of argmax_{w} P(w|o)
    - P(w): prior knowledge of what kind of words or phonemes are likely to be observed.
    - P(w|o): conditional probability of word observation, given acoustic observation of o.
      - (specific) o --> w1, w2, w3, ...?   o --> p1, p2, p3, ...?
      - Data collection is very difficult to characterize or formulate P(w|o) directly.
  - Use of the Bayesian rule
    - $$P(w|o) = \frac{P(w,o)}{P(o)} = \frac{P(o|w)P(w)}{\sum_w P(o,w)} = \frac{P(o|w)P(w)}{\sum_w P(o|w)P(w)}$$
    - The denominator is independent of w.
    - Maximization of P(w|o) in terms of w is equal to that of P(o|w)P(w) (=P(o,w))
  - Solution of argmax_{w} P(o|w) P(w)
    - P(w): can be estimated from a large text corpus.
    - P(o|w): conditional probability of acoustic observation, given intended content of w.
      - (specific) w --> o1, o2, o3, ...? p --> o1, o2, o3, ...?
      - This data collection is possible enough by asking many speakers to say w or p !!
    - P(o|w): acoustic model, P(w): linguistic model
      - Separate two models and a program that can search for the word sequence that maximizes P(o,w)

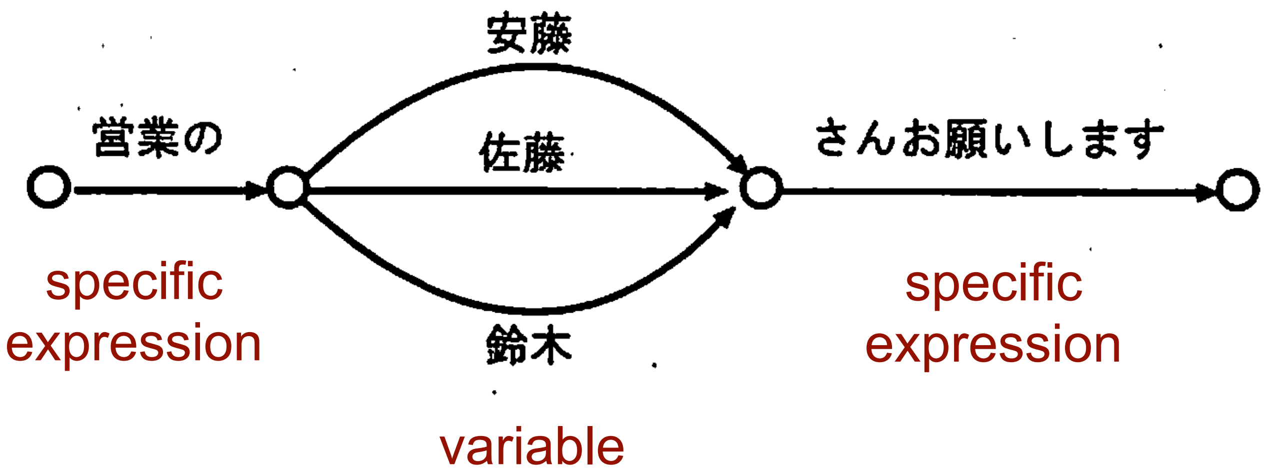
Continuous speech (connected word) recognition

Repetitive matching between an input utterance and word sequences that are allowed in a specific language

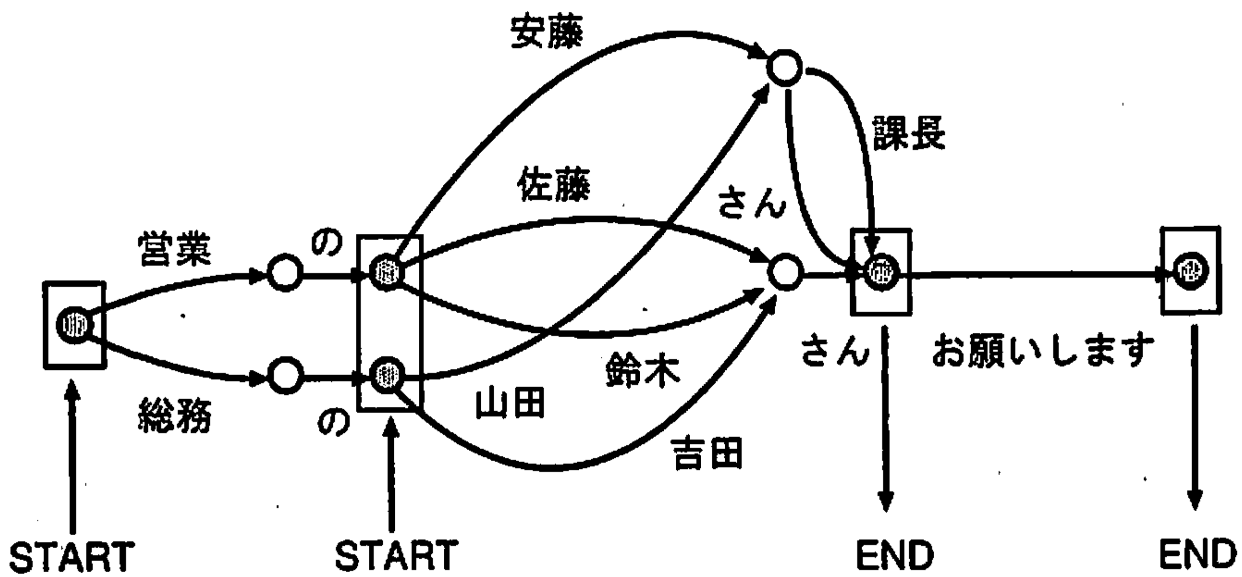
- Constraints on words and their sequences (ordering)
  - ∗ Vocabulary: a set of candidate words
  - ∗ Syntax: how words are arranged linearly.
  - ∗ Semantics: can be represented by word order??
- Examples of unaccepted sentences
  - ∗ 私/は/マッキンポッシュ/を/使う。(lexical error)
  - ∗ 私/マッキントッシュ/は/使う/を。(syntax error)
  - ∗ 私/は/マッキントッシュ/を/破る。(semantic error)

Representation of syntax (grammar)

- 営業の安藤さんお願いします。
- 営業の佐藤さんお願いします。
- 営業の鈴木さんお願いします。

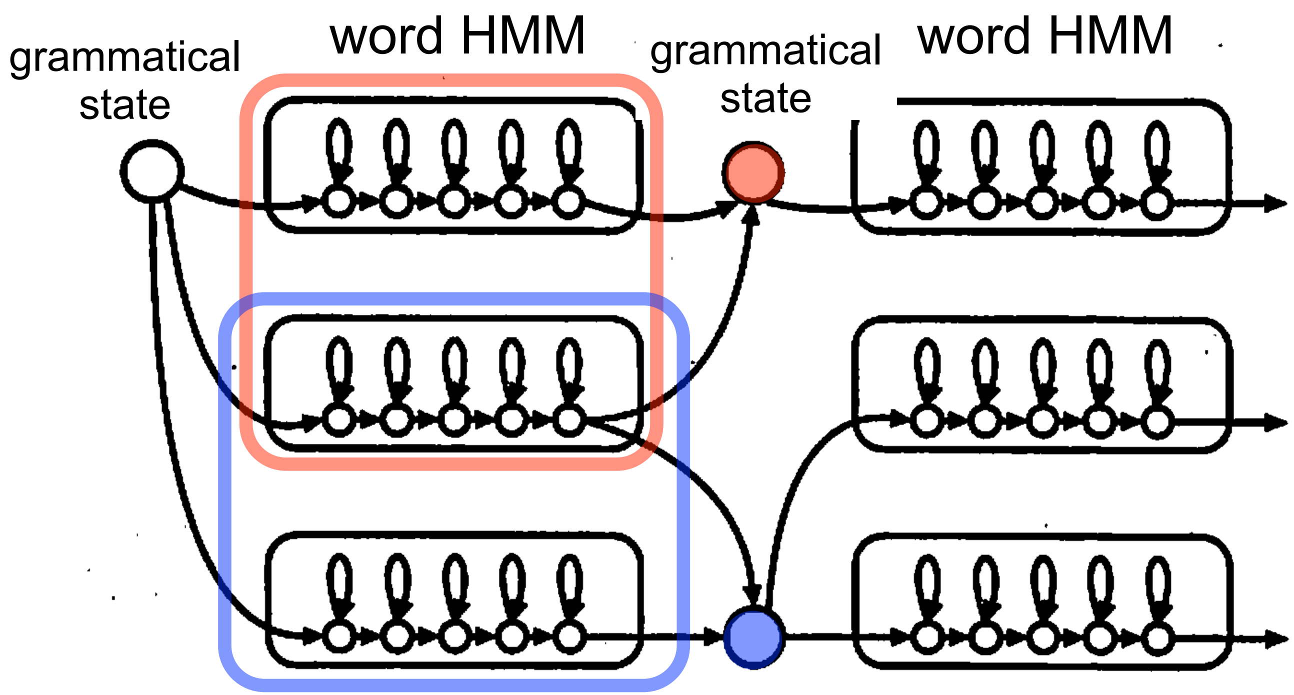


Network grammar with a finite set of states



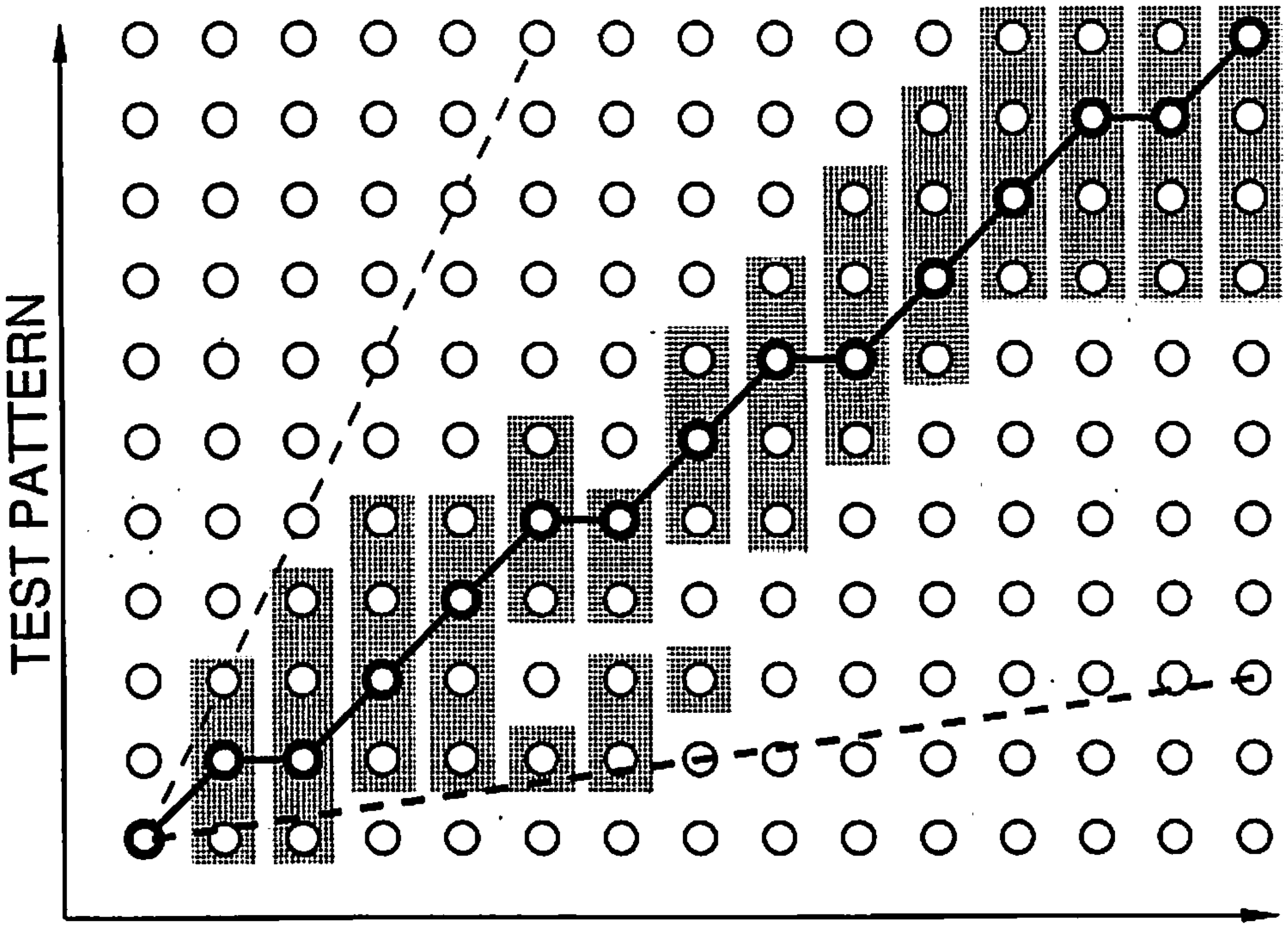
A sentence is accepted if it starts at one of the initial states and ends at one of the final states.

Speech recognition using a network grammar



When a grammatical state has more than one preceding words, the word of the maximum probability (or words with higher probabilities) is adopted and it will be connected to the following candidate words.

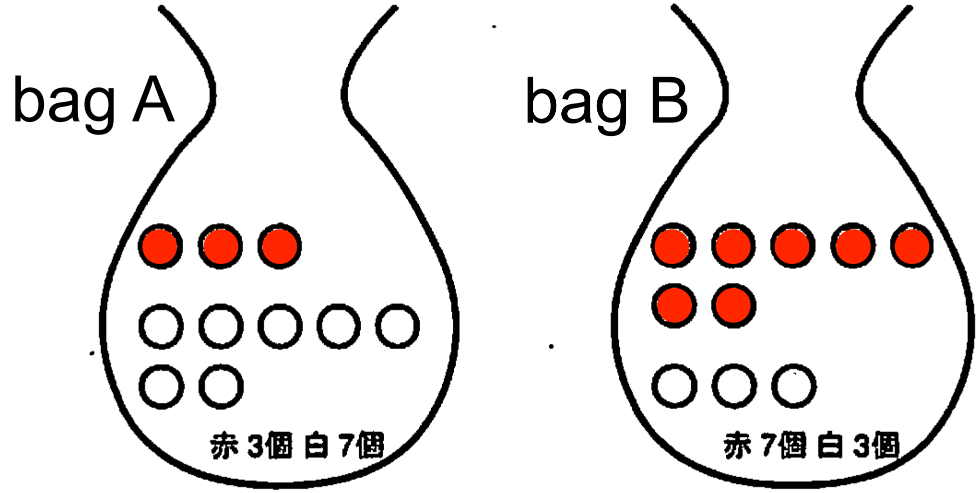
Viterbi search algorithm



A well-known strategy for diversity

- Statistical framework of ASR
  - Solution of argmax_{w} P(w|o)
    - P(w): prior knowledge of what kind of words or phonemes are likely to be observed.
    - P(w|o): conditional probability of word observation, given acoustic observation of o.
      - (specific) o --> w1, w2, w3, ...?   o --> p1, p2, p3, ...?
      - Data collection is very difficult to characterize or formulate P(w|o) directly.
  - Use of the Bayesian rule
    - $$P(w|o) = \frac{P(w,o)}{P(o)} = \frac{P(o|w)P(w)}{\sum_w P(o,w)} = \frac{P(o|w)P(w)}{\sum_w P(o|w)P(w)}$$
    - The denominator is independent of w.
    - Maximization of P(w|o) in terms of w is equal to that of P(o|w)P(w) (=P(o,w))
  - Solution of argmax_{w} P(o|w) P(w)
    - P(w): can be estimated from a large text corpus.
    - P(o|w): conditional probability of acoustic observation, given intended content of w.
      - (specific) w --> o1, o2, o3, ...? p --> o1, o2, o3, ...?
      - This data collection is possible enough by asking many speakers to say w or p !!
    - P(o|w): acoustic model, P(w): linguistic model
      - Separate two models and a program that can search for the word sequence that maximizes P(o,w)

Probabilistic decision



Observation: You pick a ball three times. The colors are ● ○ ●.

Probabilities of P(●○●| A) and P(●○●| B)

$$袋A : \frac{3}{10}\times\frac{7}{10}\times\frac{3}{10} = 0.063 \quad 袋B : \frac{7}{10}\times\frac{3}{10}\times\frac{7}{10} = 0.147$$

Decision: The bag used is supposed to be B.

Statistical framework of speech recognition

$$P(W|A) = \frac{P(A,W)}{P(A)} = \frac{P(A|W)P(W)}{P(A)} = \frac{P(A|W)P(W)}{\sum_W P(A|W)P(W)}$$

A = Acoustic, W = Word

- P(bag|●○●) --> P(bag=A|●○●) or P(bag=B|●○●)
- P(●○●|bag=A) : prob. of bag A's generating ●○●.
- P(bag) --> P(bag=A) or P(bag=B) Which bag is easier to be selected?

If we have three bags of type-A and one bag of type-B, then

$$P(袋A \mid ●○●) = 0.063 \times 0.75 = 0.04725$$

$$P(袋B \mid ●○●) = 0.147 \times 0.25 = 0.03675$$

The bag used is supposed to be A.

N-gram language model

The most widely-used implementation of P(w)

Only the previous N-1 words are used to predict the following word.
(N-1)-order Markov process

$$P(x_1,\cdots,x_n) = \underbrace{P(x_n|x_1,\cdots,x_{n-1})}_{\approx P(x_n|x_{n-N+1},\cdots,x_{n-1})} P(x_1,\cdots,x_{n-1})$$
$$\approx P(x_n|x_{n-N+1},\cdots,x_{n-1})P(x_1,\cdots,x_{n-1})$$
$$\approx \prod_{i=1}^{n} P(x_i|x_{n-N+1},\cdots,x_{i-1})$$

N-1 = 1 --> bi-gram
N-1 = 2 --> tri-gram

I'm giving a lecture on speech recognition technology to university students.

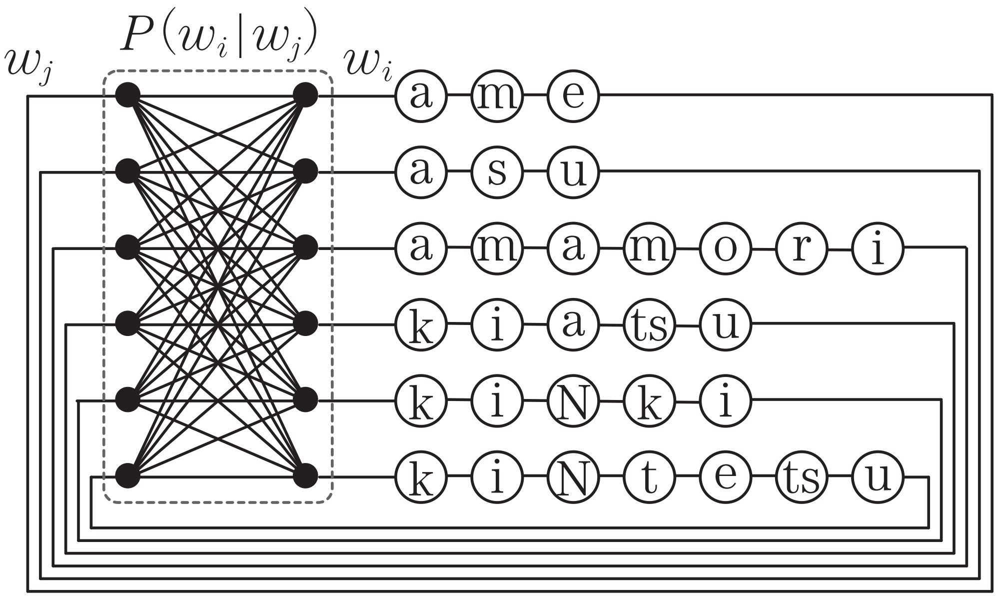
P(a | I'm, giving), P(lecture | giving, a), P(on | a, lecture),
P(speech | lecture, on), P(recognition | on, speech), ...

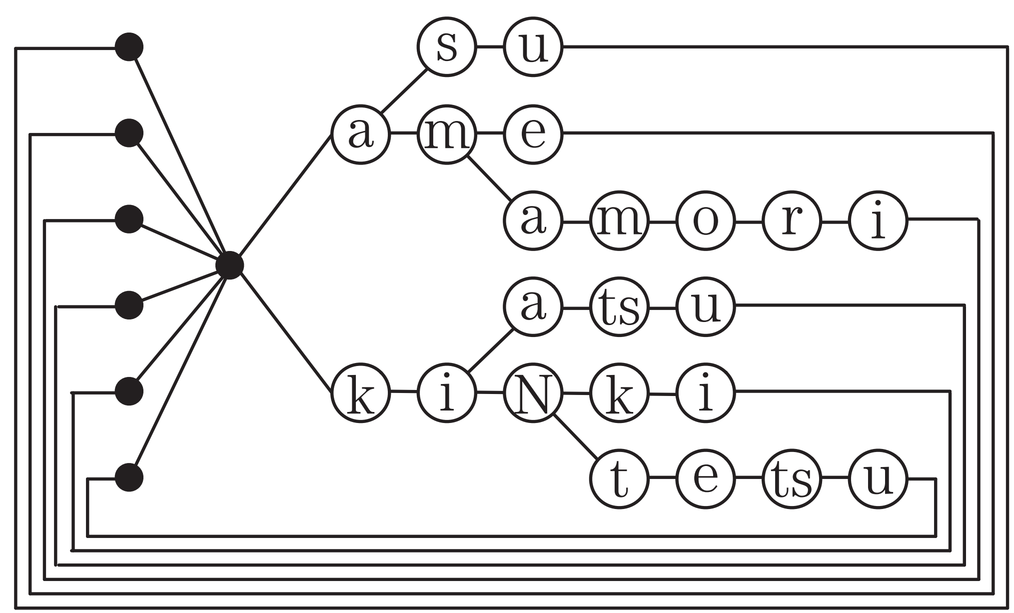
How to calculate N-gram prob.

- .... lecture on speech recognition ....
  P(speech | lecture, on)
  = C (lecture, on, speech) / C (lecture, on)
  P(recognition | on, speech)
  = C (on, speech, recognition) / C (on, speech)
  P(w3 | w1, w2)
  = C (w1, w2, w3) / C (w1, w2)
- Typical problems of calculating N-gram prob
  C (w1, w2, w3) = 0 --> N-gram prob. = 0 ??
  C (w1, w2) = 0 --> N-gram prob. = ???
  α x P(w3 | w1) or β x P(w3) are substituted as P (w3 | w1, w2).
  Context dependencies are ignored to some degree.

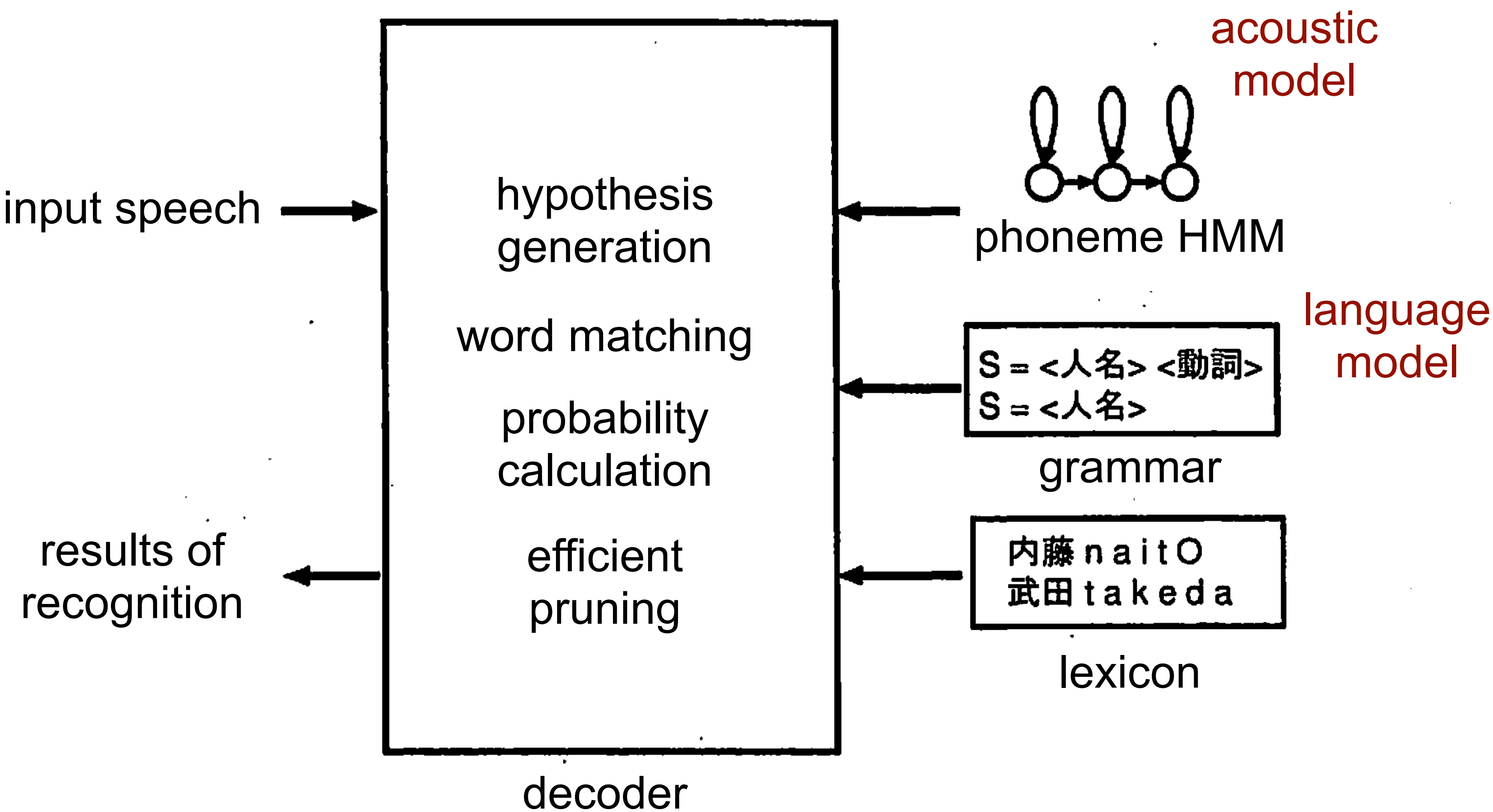
2-gram as network grammar

- 2-gram as network grammar and as tree-based network grammar





Development of a speech recognition system



ASR under various conditions

種々の条件下における認識結果例

- ❒ ***連続音韻認識結果(triphone の任意連結)***
  SIL Q be: kokudeoNobetonamukita: Nhe: einokokumiNnomewachimetakuSIL SIL Q ayagadoo: ojo: o: watsunerumadeiniwa SIL SIL tsukanarinosaigetsohichiootoshita

- ❒ ***連続音節認識結果(上記＋音節構造の知識導入)***
  SIL げいこくでおんおべとなむきたんへいのこくみんのめわちめたく SIL SIL っあやがどおじょおわつねるまでいにわ SIL SIL つかなりのさいげつおひちおおとした SIL

- ❒ ***連続単語認識結果(上記＋単語の知識導入，語彙数=20K)***
  1st pass 米穀 ネオン ベトナム 機関 平 残っ 区民 度目 月 目立っ 句 。 ？ カヤ 花道 王女 大和 詰める まで なり なさい えっ 消費 治療 落とし 他
  2nd pass 米穀 ネオン ベトナム 帰還 平 残っ 区民 度目 月 目立っ く 、 、 カヤ 門 王女 大和 詰める まで 庭 り なさい れ 曹 陽 治療 落とし た

ASR under various conditions

種々の条件下における認識結果例

- ***連続音韻認識結果(triphone の任意連結)***
  SIL Q be: kokudeoNobetonamukita: Nhe: einokokumiNnomewachimetakuSILSILQayagadoo: ojo: o: watsunerumadeiniwaSILSILtsukanarinosaigetsohichiiootoshita

- ***連続音節認識結果(上記＋音節構造の知識導入)***
  SILげいこくでおんおべとなむきたんへいのこくみんのめわちめたくSILSILっあやがどおじょおわつねるまでいにわSILSILつかなりのさいげつおひちおおとしたSIL

- ***連続単語認識結果(上記＋単語の知識導入，語彙数=20K)***
  1st pass 米穀 ネオン ベトナム 機関 平 残っ 区民 度目 月 目立っ 句 。 ？ カヤ 花道 王女 大和 詰める まで なり なさい えっ 消費 治療 落とし 他
  2nd pass 米穀 ネオン ベトナム 帰還 平 残っ 区民 度目 月 目立っ く 、 、 カヤ 門 王女 大和 詰める まで 庭 り なさい れ 曹 陽 治療 落とし た

- ***大語彙連続音声認識結果(上記＋単語間の連鎖知識導入)***
  1st pass 米国のベトナム帰還兵の国民の目が冷たく、彼らは同情を集めるまでには、かなりの歳月を必要 落とし た 。
  2nd pass 米国のベトナム帰還兵の国民の目は冷たく、彼らが同情を集めるまでには、かなりの歳月を必要 と し た 。

- ***正解文***
  米国でもベトナム帰還兵への国民の目は冷たく、彼らが同情を集めるまでには かなりの歳月を必要とした。

Today's menu

- Fundamentals of statistical speech recognition
- Acoustic models (HMM) for speech recognition
- From word-based HMMs to phoneme-based HMMs
- From HMM-GMM to HMM-DNN
- Speech recognition using network grammars
- Speech recognition using N-grams
- Speech recognition using NN-based language models

N-gram language model

The most widely-used implementation of P(w)

Only the previous N-1 words are used to predict the following word.
(N-1)-order Markov process

$$P(x_1,\cdots,x_n) = \underbrace{P(x_n|x_1,\cdots,x_{n-1})}_{\approx P(x_n|x_{n-N+1},\cdots,x_{n-1})} P(x_1,\cdots,x_{n-1})$$

$$\approx P(x_n|x_{n-N+1},\cdots,x_{n-1})P(x_1,\cdots,x_{n-1})$$

$$\approx \prod_{i=1}^{n} P(x_i|x_{n-N+1},\cdots,x_{i-1})$$

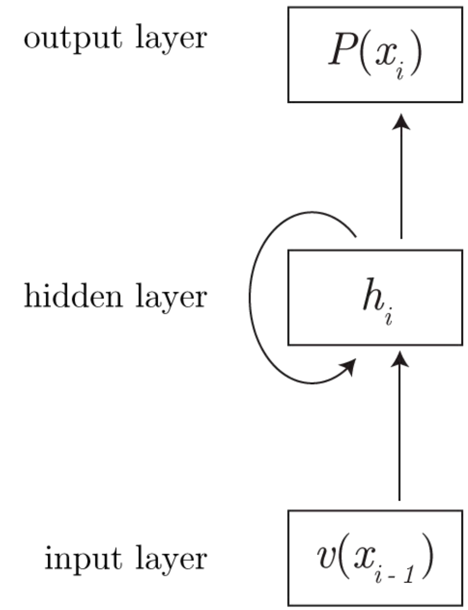
N-1 = 1 --> bi-gram
N-1 = 2 --> tri-gram

I'm giving a lecture on speech recognition technology to university students.

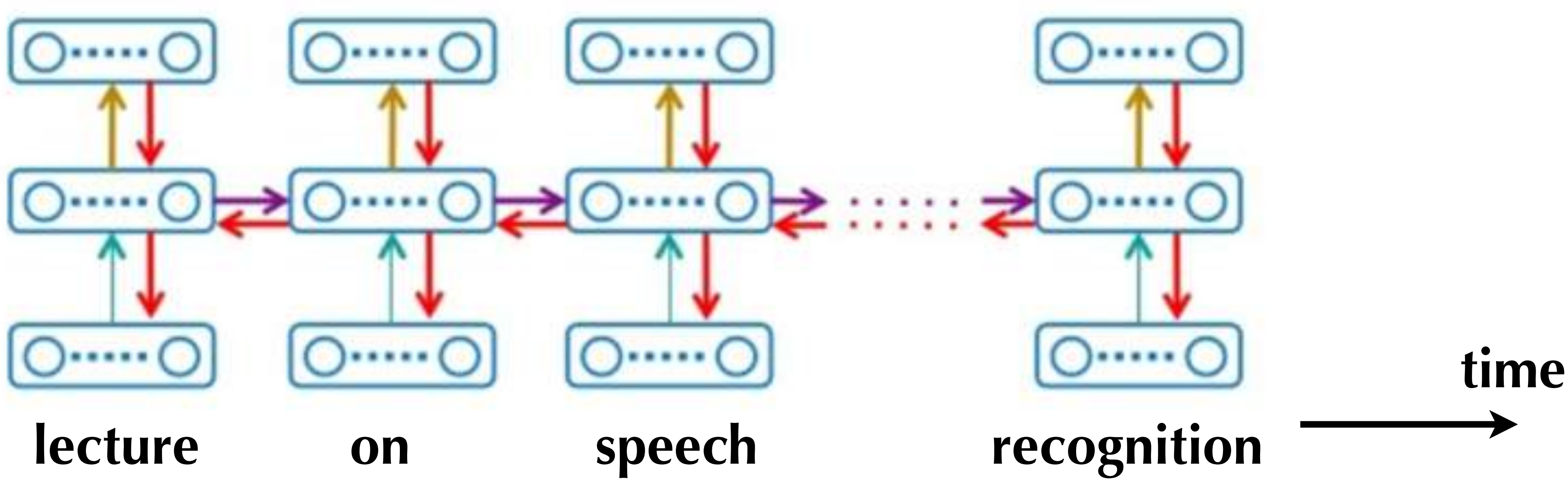
P(a | I'm, giving), P(lecture | giving, a), P(on | a, lecture),
P(speech | lecture, on), P(recognition | on, speech), ...

Recurrent NN-based LM

- v(x) = word features of related to x
  P(x) = probability of word x
  h = hidden layer



- $P(w_1), P(w_2), ..., P(w_N)$ at each time index



Recommended books