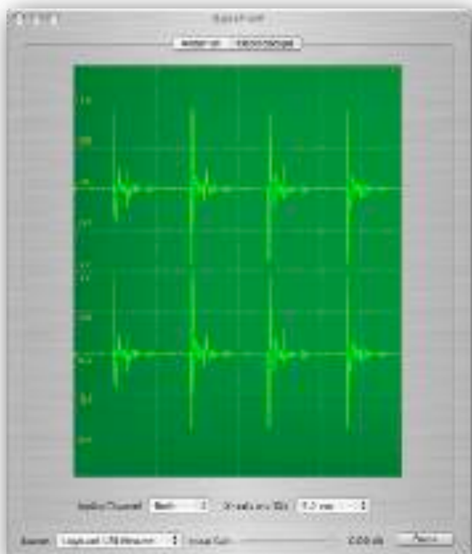
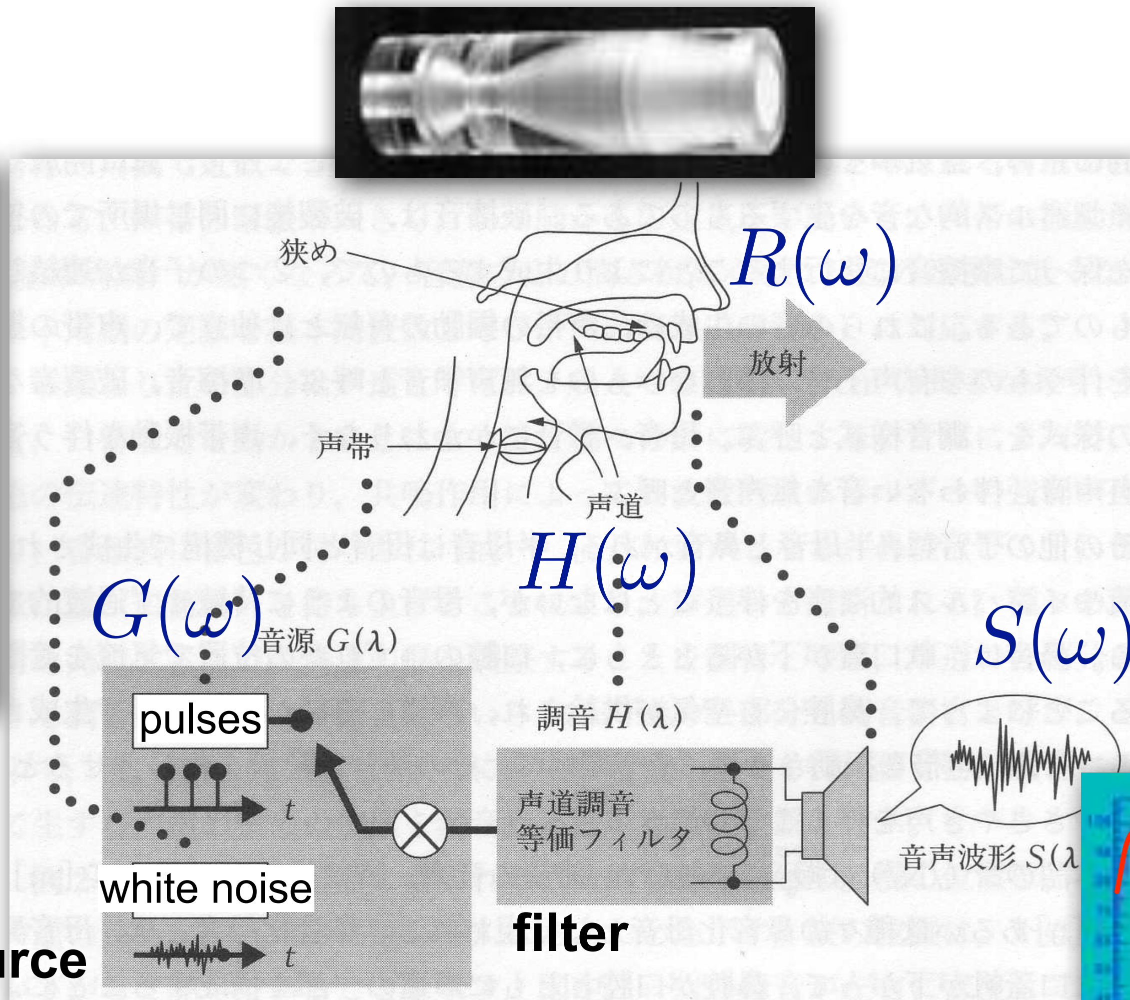
Cognitive Media Processing #7

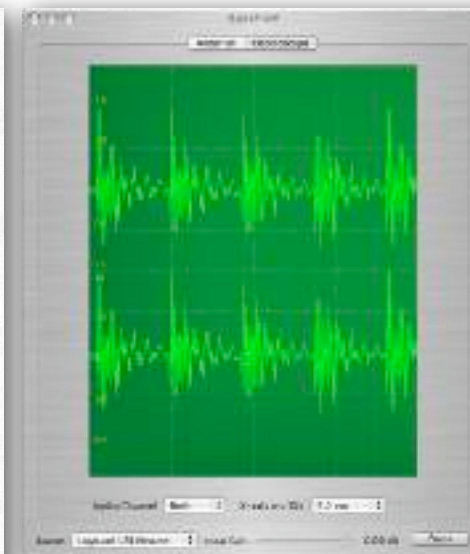
Nobuaki Minematsu

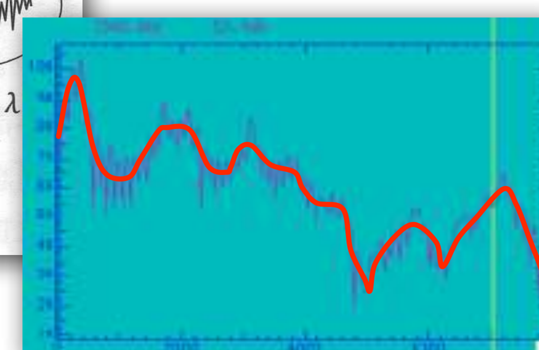
Modeling of speech production

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







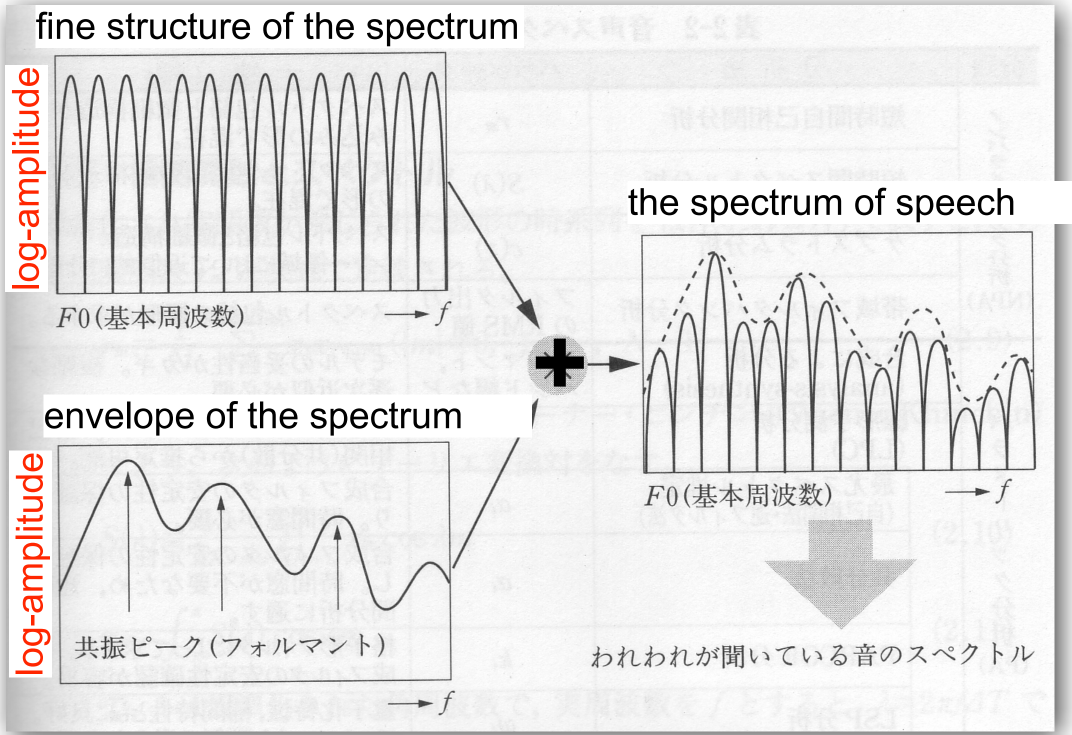


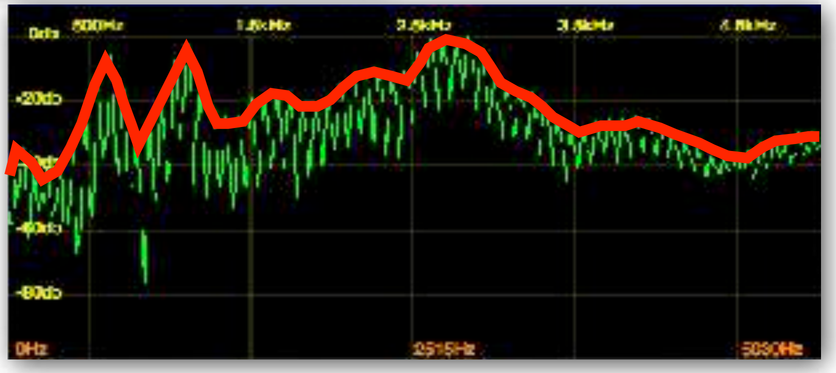
source

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

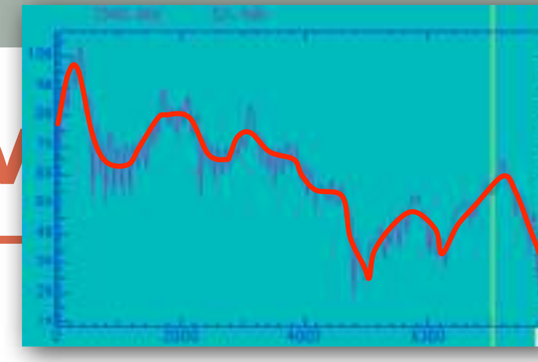
Modeling of vowel production

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

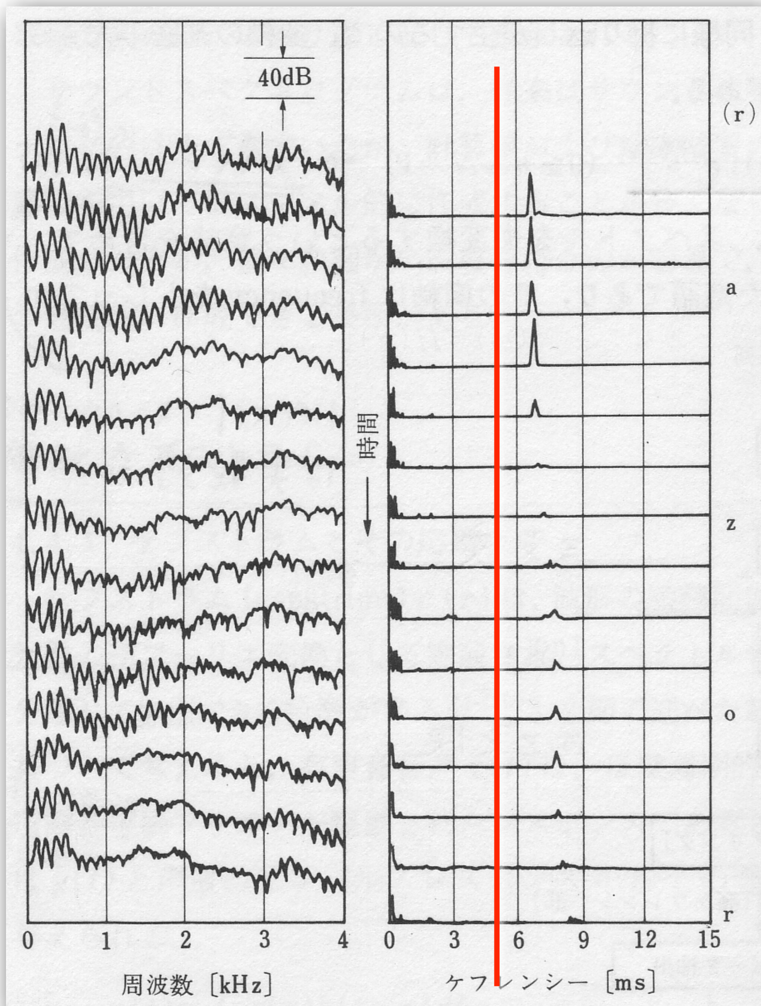


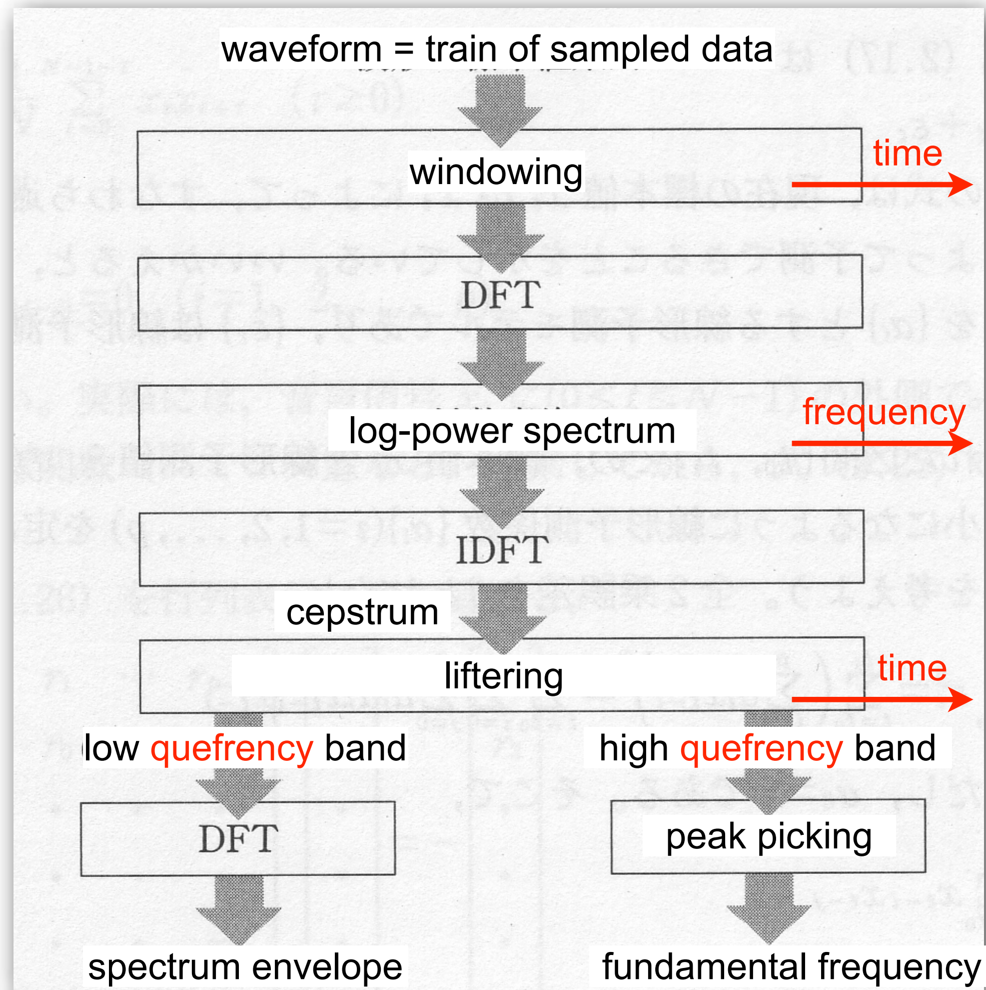


Extraction of spectrum env



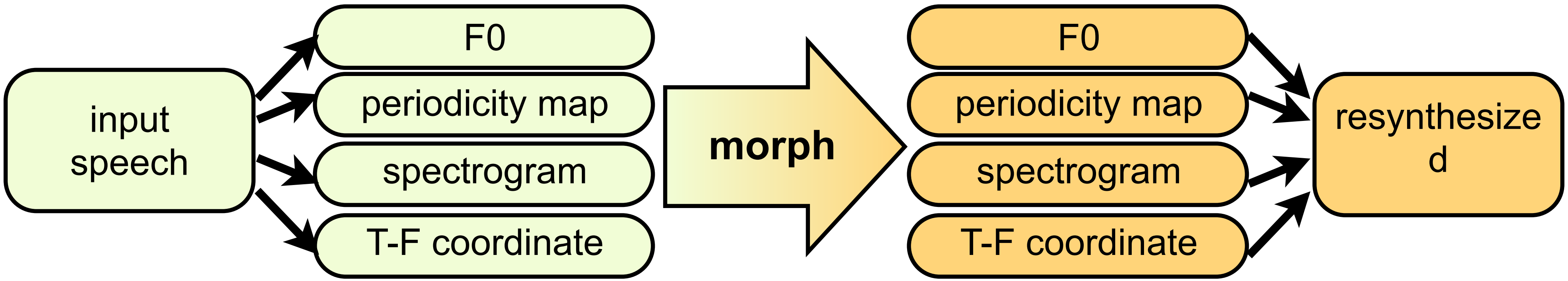
- Cepstrum method
  - Windowing + FFT + log-amplitude --> a spectrum with pitch harmonics
  - Smoothing (LPF) of the fine spectrum into its smoothed version





Advanced technology for analysis

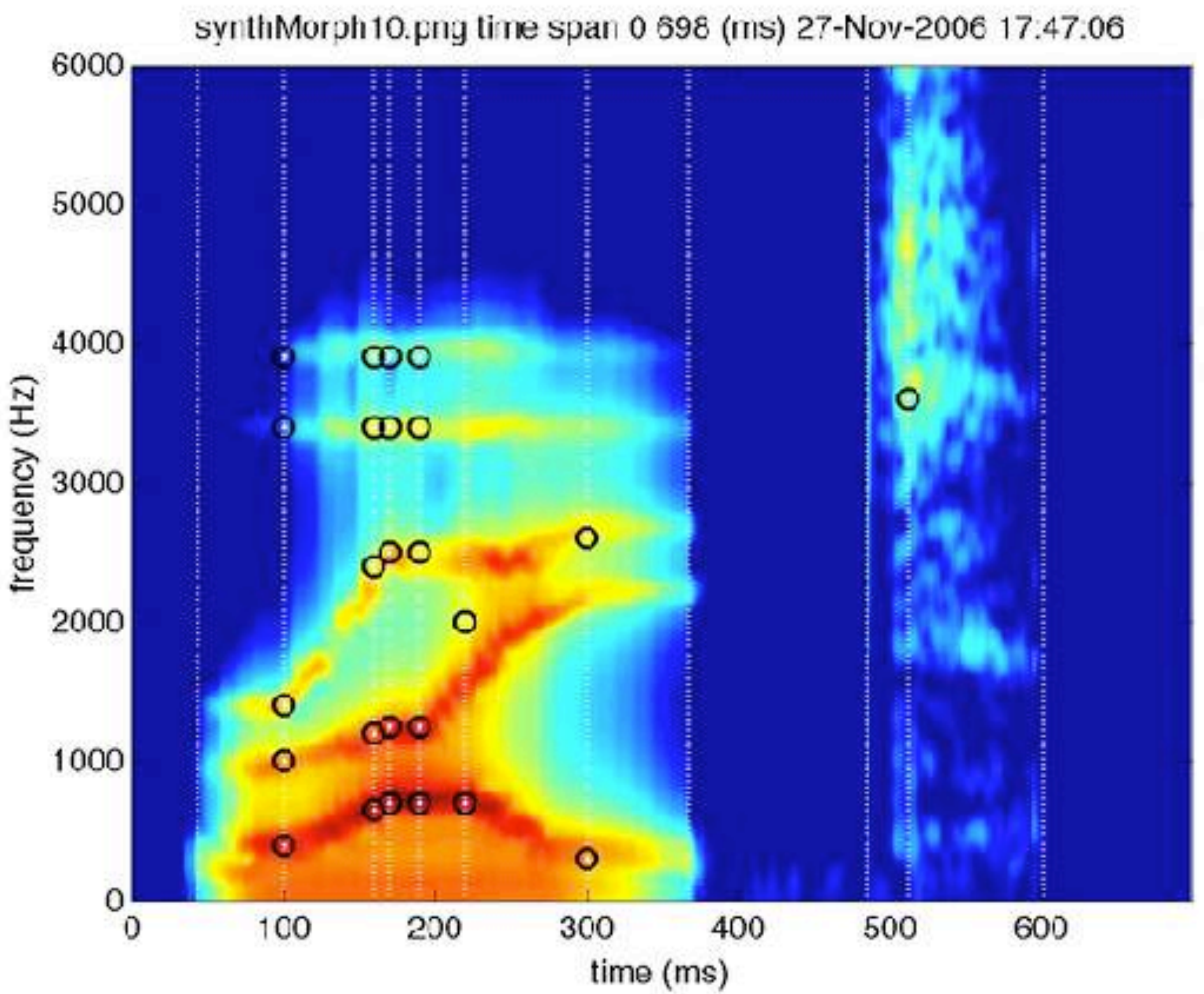
- STRAIGHT [Kawahara'06]
  - High-quality analysis-resynthesis tool
    - Decomposition of speech into
      - Fundamental frequency, spectrographic representations of power, and that of periodicity
    - High-quality speech morphing tool



  - 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

- R to L morphing bet. r/l-ight generated by Klatt synthesizer [Kubo+'98]



Spectrum reading

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

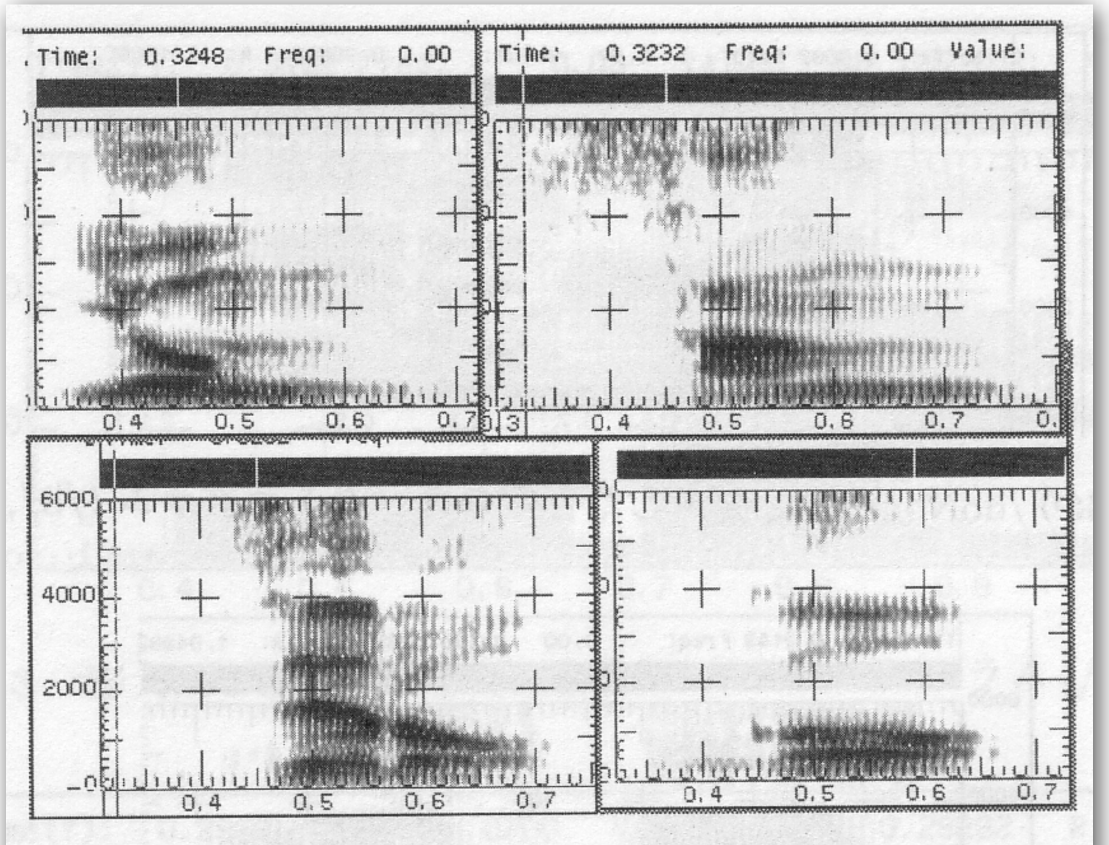


図 2.19 数字音声のスペクトログラム

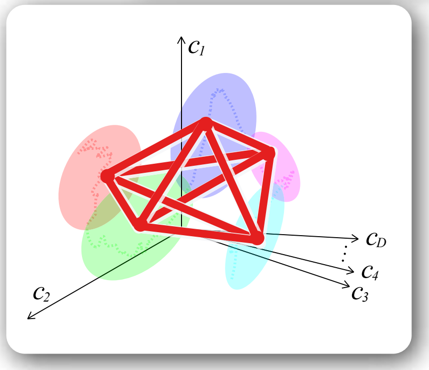
  - This is the task that is done by a speech recognizer.

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



AI Forum

- A joint forum between UToky

Ece Kamar

Deputy Lab Director, Microsoft

› Bio

⌄ Keynote Abstract

Phase Transition in AI

of

AI technologies in the way we live and work.

specialized training in medicine. Examples will be shown of how the general intelligence of GPT-4 can be used, with implications for the current and future practice of medicine.

Logical and expressive

- Logical information and expressive information
  - Factors (bases) to describe expressive information
    - Facial expressions (as example)
      - 6 factors of surprise, fear, dislike, anger, happiness, and sorrow
      - A still debatable problem in psychology
    - Theory of mind [D. Premack et. al.'78]
      - The ability to attribute mental states to oneself and others and to understand that others have different mental states than one's own.
        - Different individuals have different minds.
        - Those who can't have theory of mind have difficulty in understanding this fact.
      - One of the theories that explains the cause of autism (自閉症) [S. Baron-Cohen 91]
        - Difficulty in reading the mind of others and understanding that everybody has one's own mind.
        - Difficulty in reading the facial expressions.
        - Abnormality in information processing in the "old" brain.

a, a, a, a, a, a,
etc

Chat-GPT-4 can pass the test of the theory of mind?

Chat-GPT-4 have emotional intelligence?

Speech Communication Tech.

- Speech recognition -

Nobuaki Minematsu

Today's menu

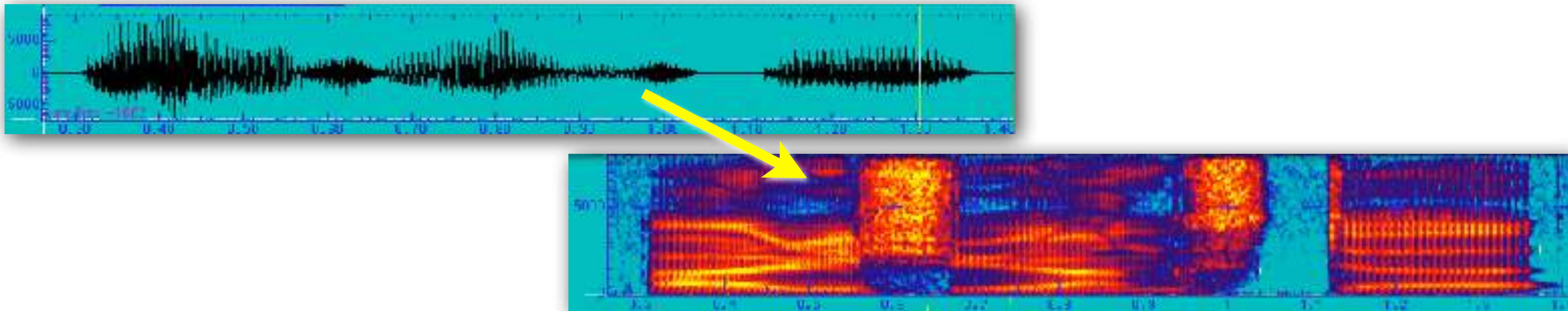
- Fundamentals of statistical speech recognition
- Acoustic models (HMM) for speech recognition
- From word-based HMMs to phoneme-based HMMs
- From GMM-HMM to DNN-HMM

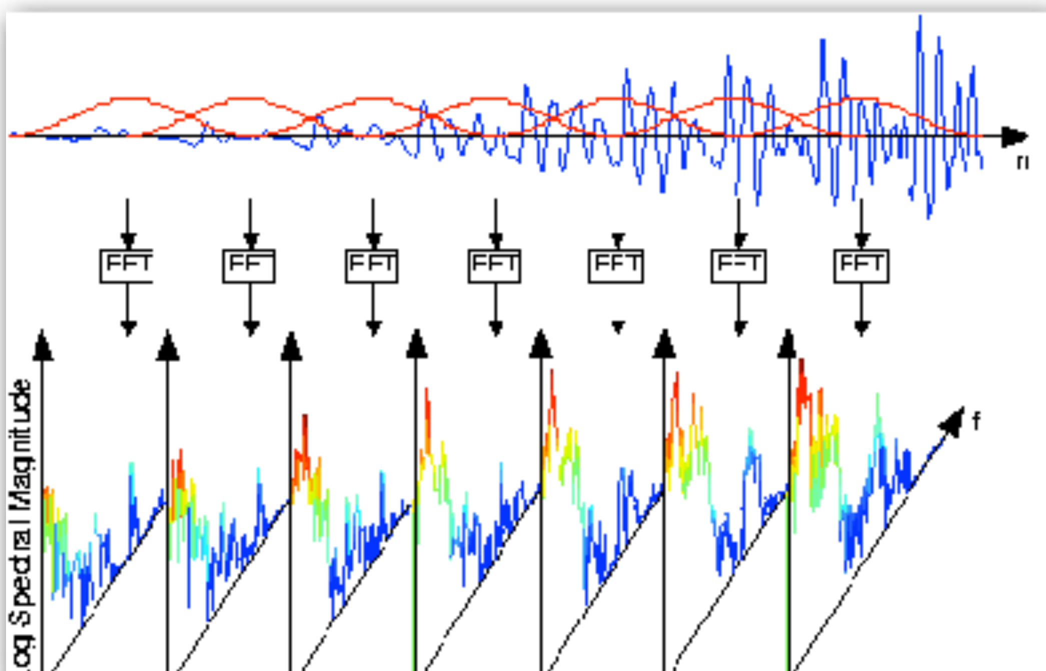
- Speech recognition using network grammars
- Speech recognition using N-grams
- Speech recognition using NN-based language models

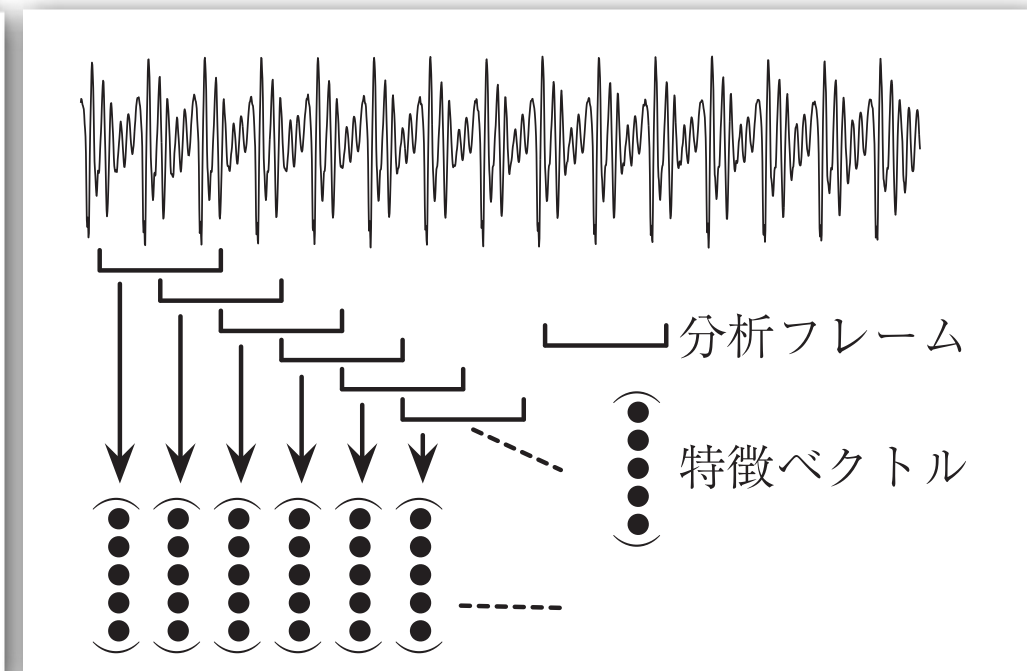
module-based ASR

- Module-based ASR to one-package (E2E) ASR (next week)

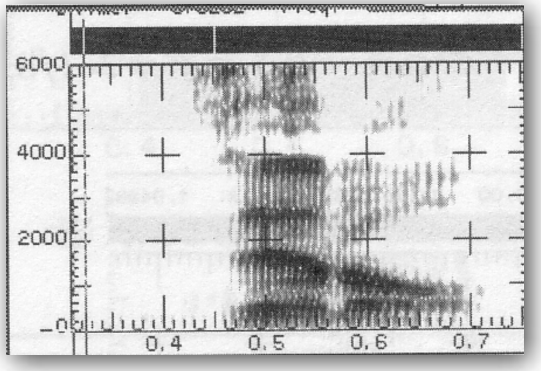
Waveforms --> spectrums --> sequence of feature vectors



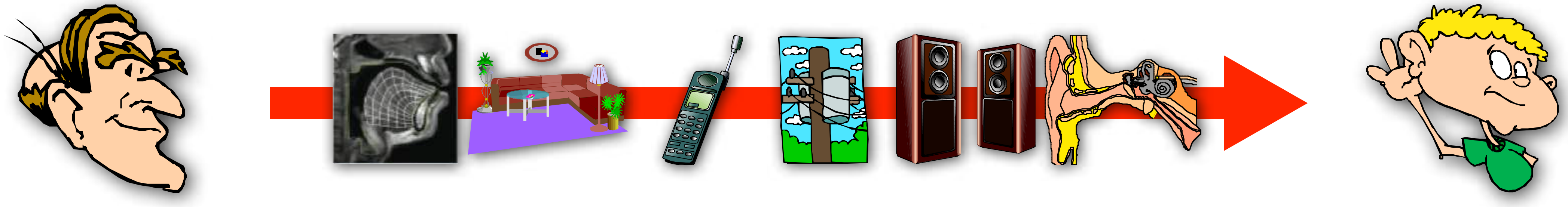




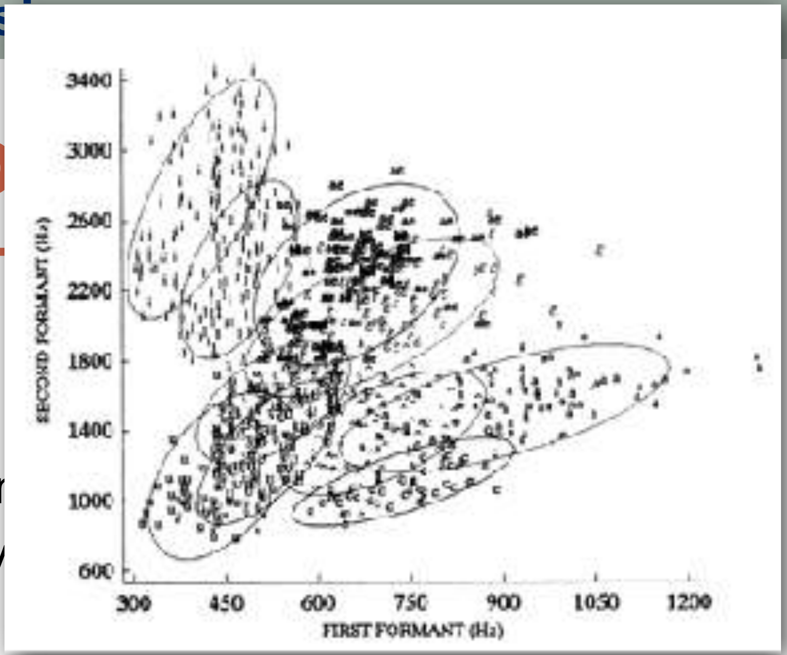
Difficulty of ASR



- Task of Automatic Speech Recognition (ASR)
  - Automatic identification of what is said by any speaker
    - Input: spectrum (feature vector) 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 represent word sequences in naturally and spontaneously generated speech?
    - How to treat ungrammatical utterances, word fragments, filled pauses, etc ?
  - ASR machines do not understand the content of what is spoken.

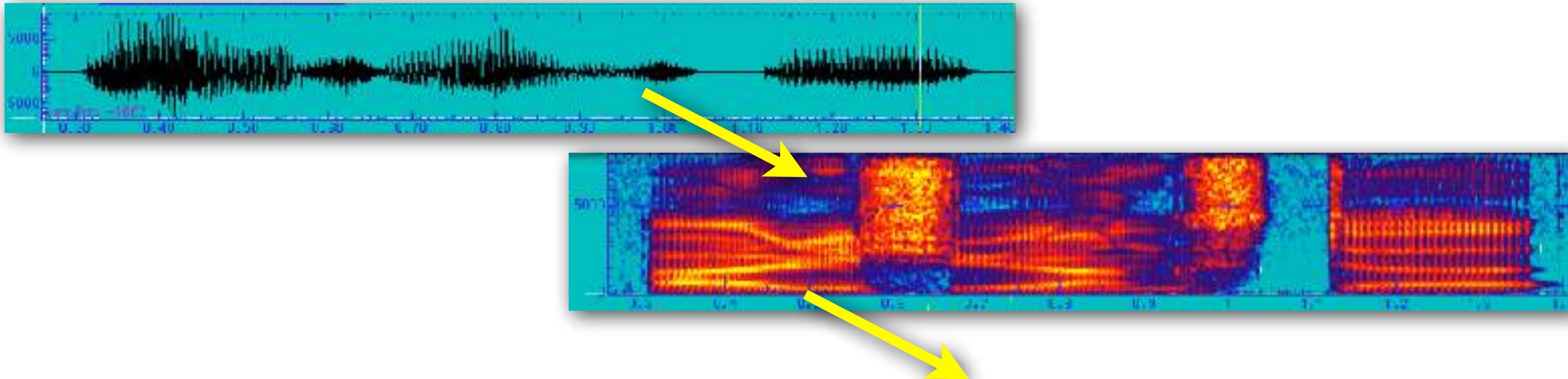


How to make a difficult pro



- Statistical framework of ASR
  - Solution of argmax_{w} P(w|o)
    - P(w): prior knowledge of what kind of words or phoner
    - P(w|o): conditional probability of word observation, giv
      - (specific) o --> w1, w2, w3, ...? o --> p1, p2, p3, ...?
      - Data collection is very difficult to characterize or estimate 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 read aloud w or p !!
    - P(o|w): acoustic model, P(w): linguistic model
      - Two separate modules + the other one that searches for the word sequence that maximizes P(w,o)

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



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

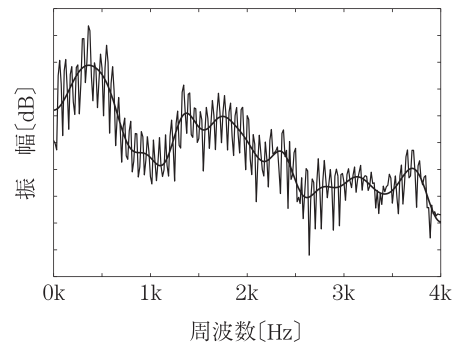
$$\arg\max_{w} P(w_1, w_2, \ldots, w_N | o_1, \ldots, o_t, \ldots, o_T) =$$

$$\arg\max_{w} P(o_1, \ldots, o_t, \ldots, o_T | w_1, w_2, \ldots, w_N) P(w_1, w_2, \ldots, w_N)$$

o : cepstrum vector

Cep. distortion and DTW

- Cepstrum vector = spectrum envelope

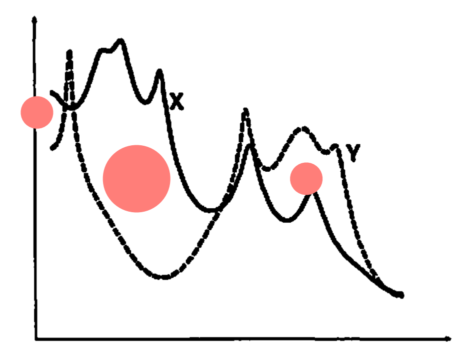


$$\longrightarrow \vec{c}$$

- 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.

$$D_n = \left(\log|S_n'| - \overline{\log|S_n|}\right) - \left(\log|T_n'| - \overline{\log|T_n|}\right)$$

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

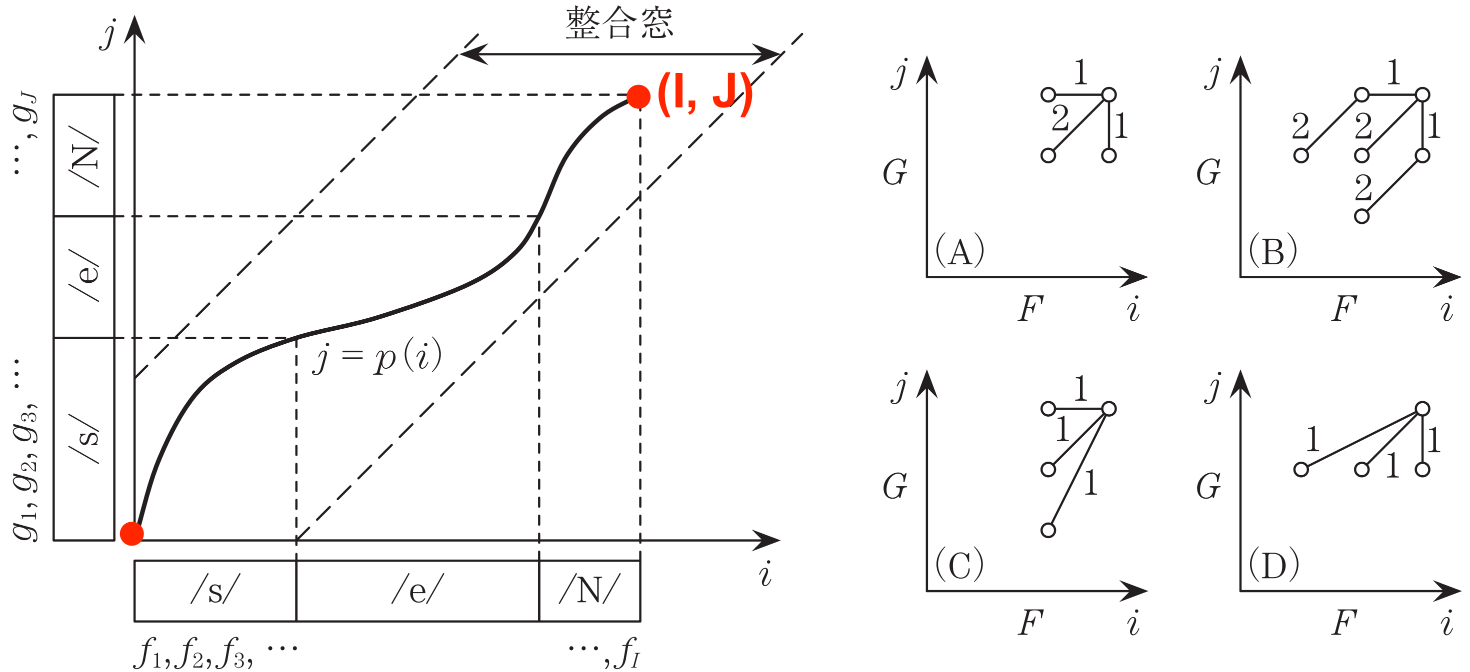


Cep. distortion and DTW

- 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.

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

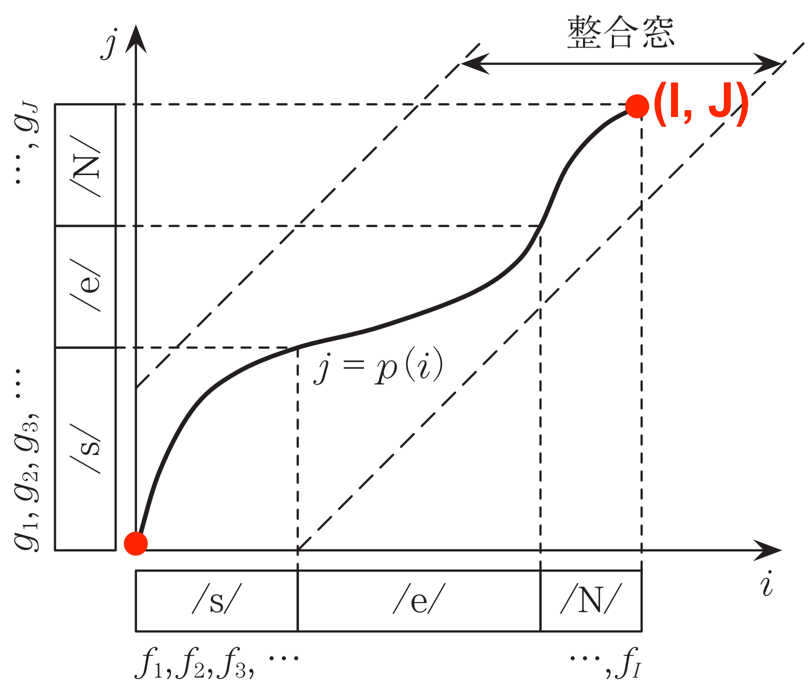
- Local distance : $d(f_i, g_j)$ = Euclid distance of the corresponding two cepstrum vectors.

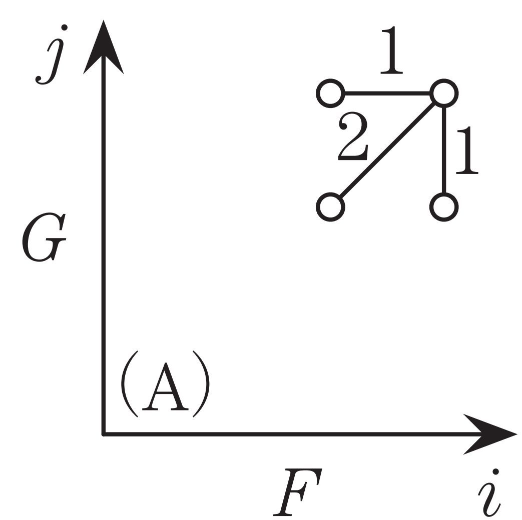


Cep. distortion and DTW

- Total distance accumulated up to point (i,j) = D(i,j)
  - d(i,j) = local 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 GMM-HMM to DNN-HMM

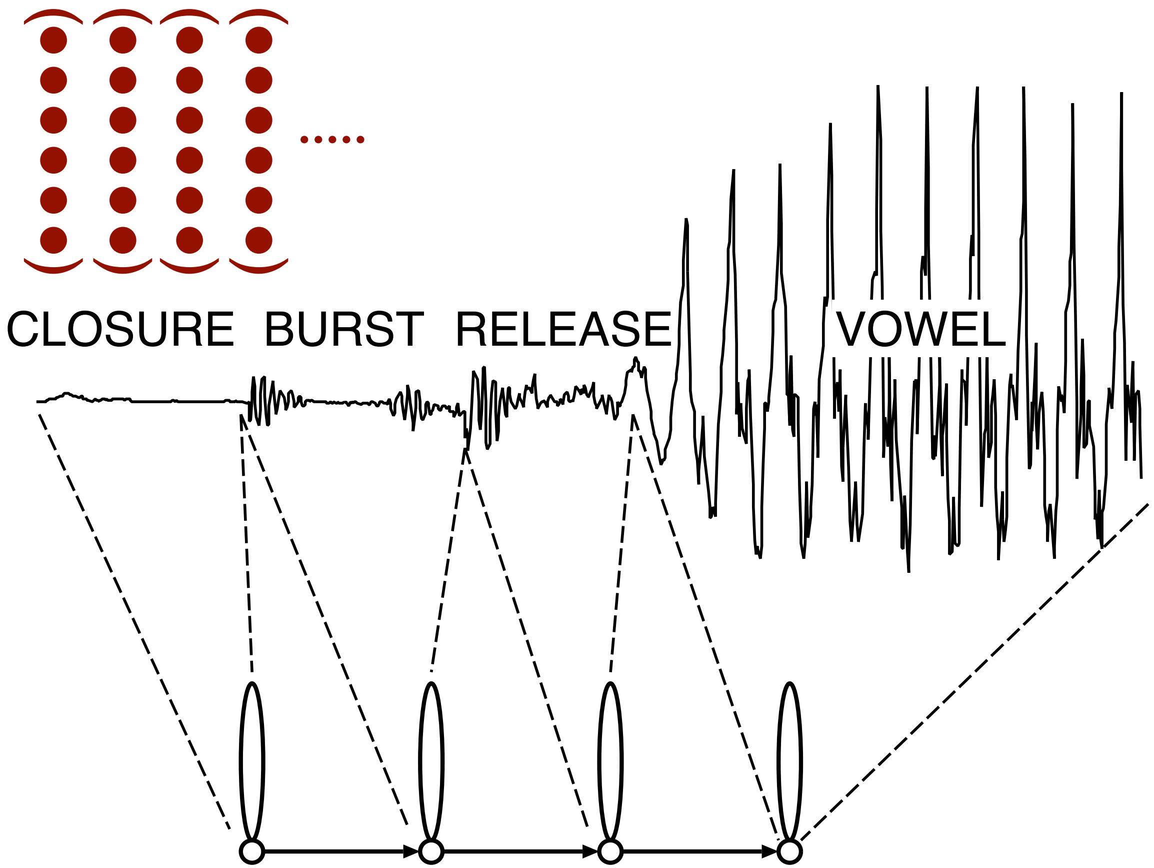
- Speech recognition using network grammars
- Speech recognition using N-grams
- Speech recognition using NN-based language models

- Module-based ASR to one-package (E2E) ASR (next week)

How to make a difficult problem tractable?

- 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 estimate 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 read aloud w or p !!
    - P(o|w): acoustic model, P(w): linguistic model
      - Two separate modules + the other one that searches for the word sequence that maximizes P(w,o)

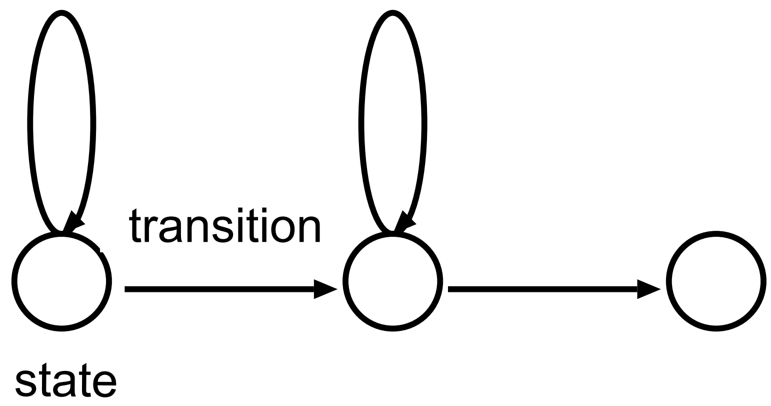
Hidden Markov Model as generative model



Probabilistic generative model

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

Hidden Markov Process



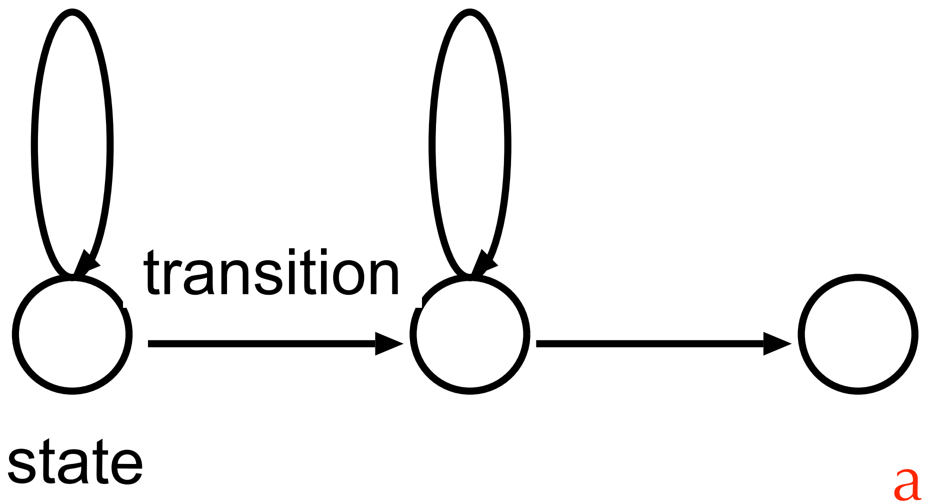
$$P(x_n \mid \underbrace{x_{n-1}, \cdots, x_1}_{\text{previous observations}}) = P(x_n \mid \underbrace{S_n}_{\text{current state}})$$

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.

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. : $P(o|s_t = i) = b_i(o) \; = \mathcal{N}(o; \mu_i, \Sigma_i)$

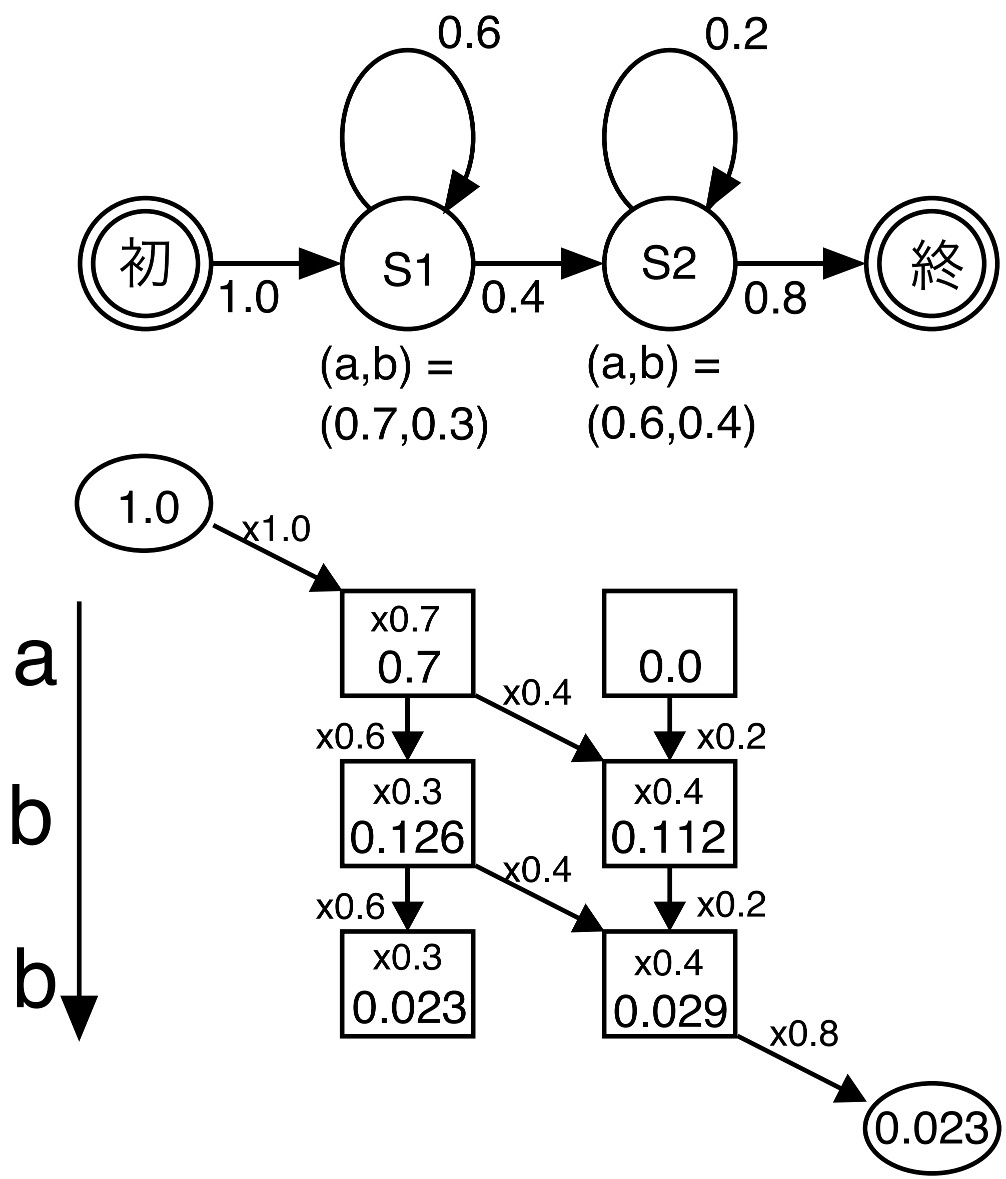
Forward prob.

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

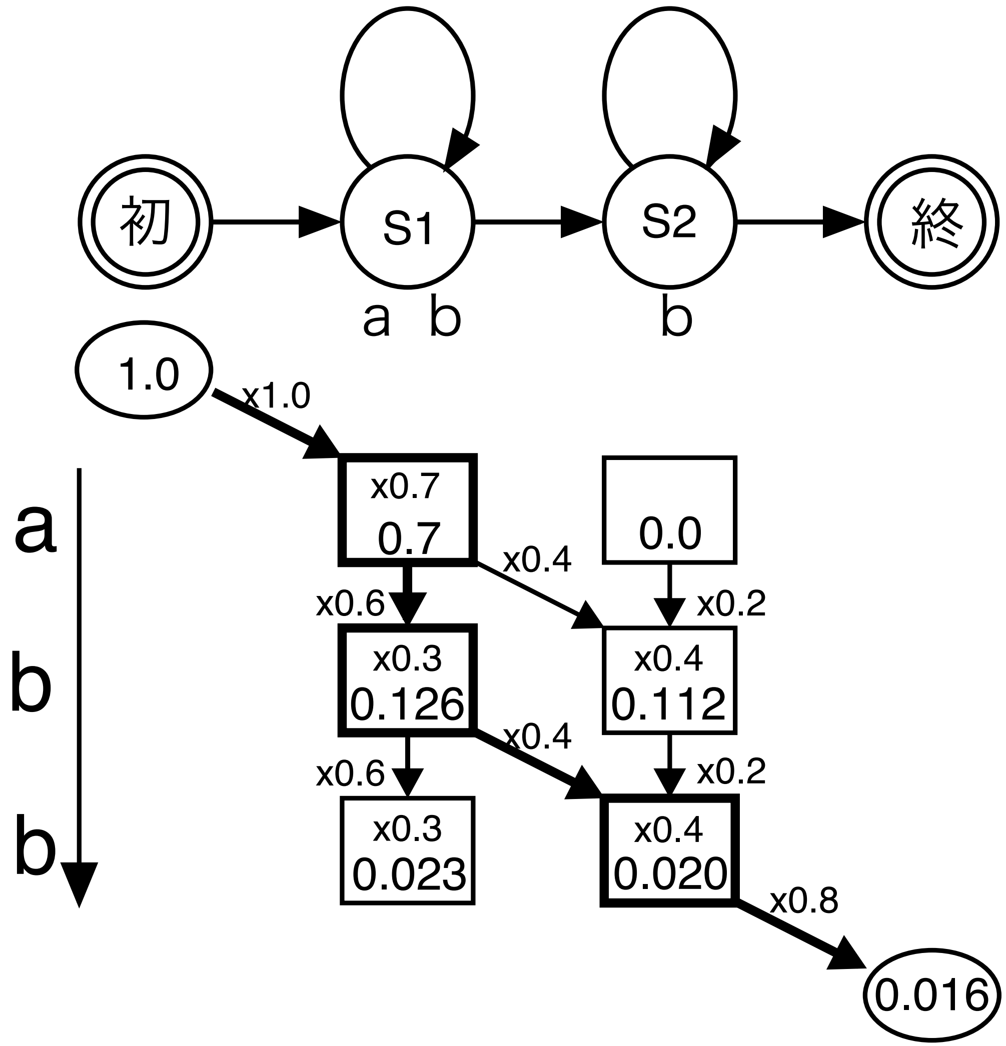
Backward prob.

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

Output probability of observation sequence (Trellis)

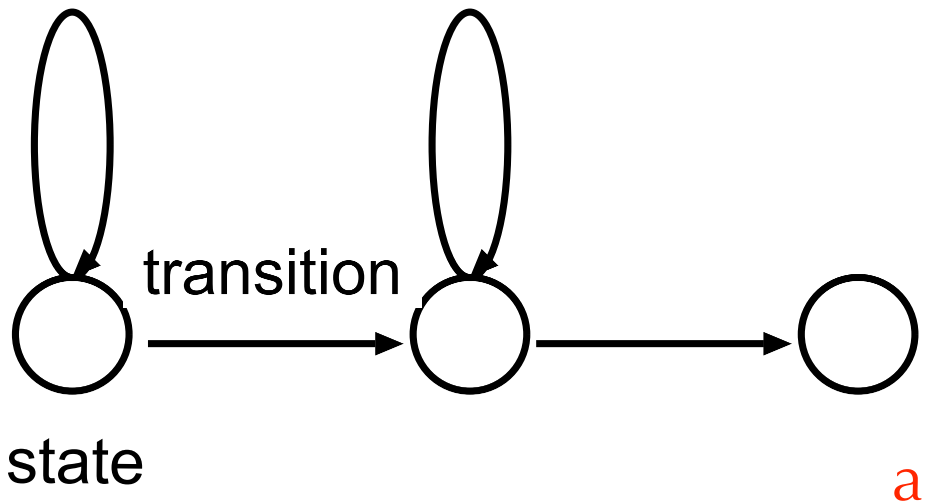


Output probability of observation sequence (Viterbi)



The maximum likelihood path is only adopted.

Parameters of HMM



single Gaussian or
a mixture of Gaussians

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

Forward prob.

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

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.

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

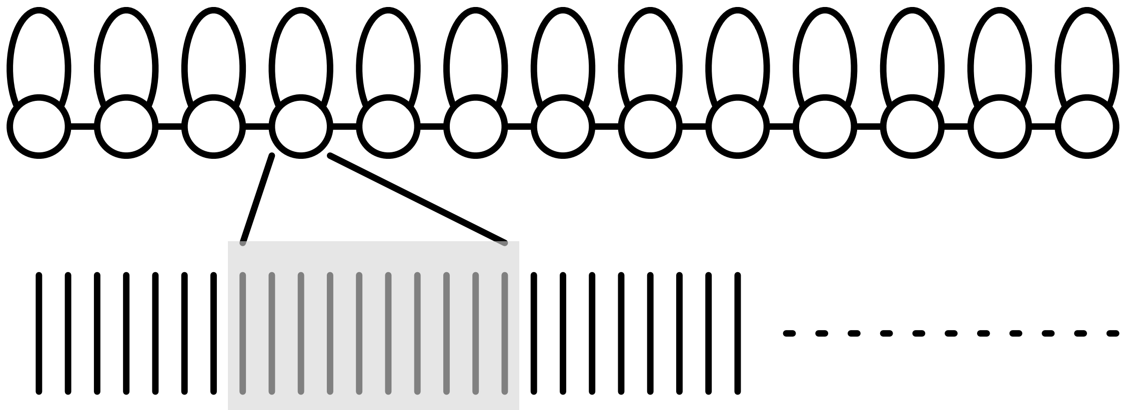
$\rightarrow \quad \alpha_j(t)\beta_j(t) = P(O, s(t) = j|M)$

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

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

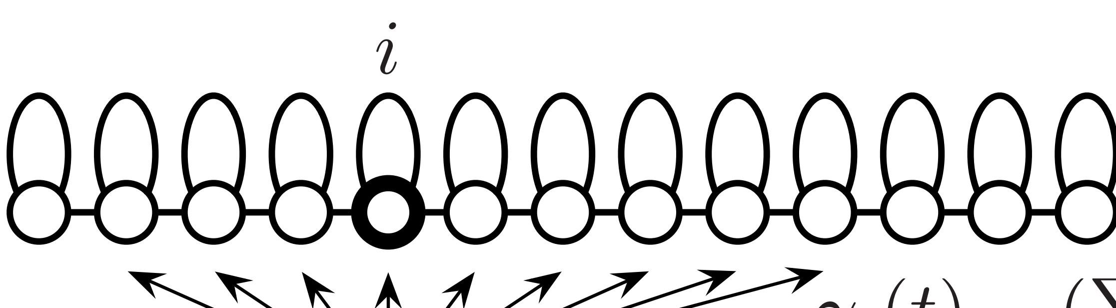
$\rightarrow \quad \hat{\mu}_j = \dfrac{\sum_t L_j(t) \cdot o_t}{\sum_t L_j(t)} = \dfrac{\sum_t \alpha_j(t)\beta_j(t) \cdot o_t}{\sum_t \alpha_j(t)\beta_j(t)} \qquad P(O|\hat{M}) \geq P(O|M)$

Estimation of HMM parameters

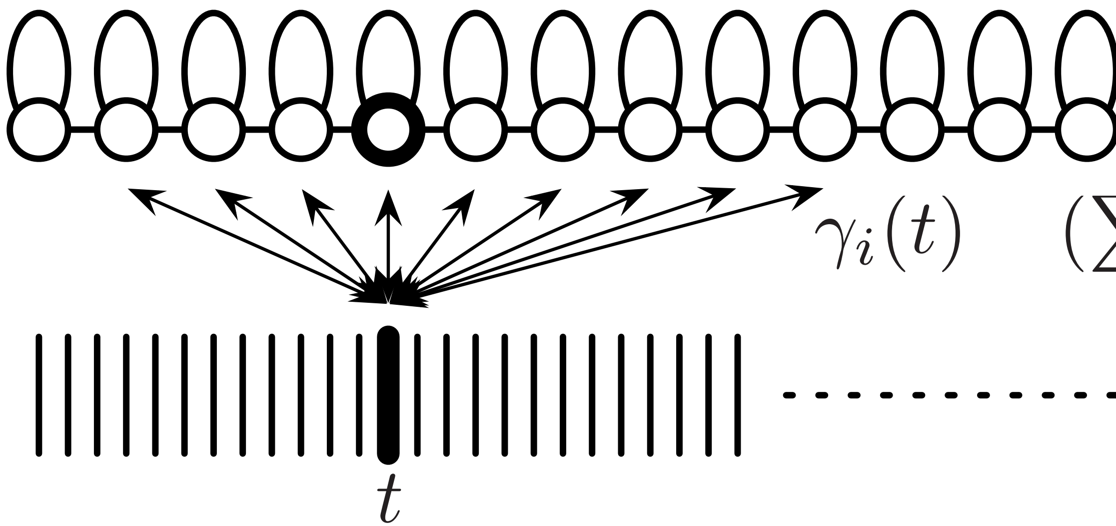


$$\mu = \frac{1}{T}\sum_t o_t = \frac{\sum_t \frac{1}{T} o_t}{\sum_t \frac{1}{T}}$$

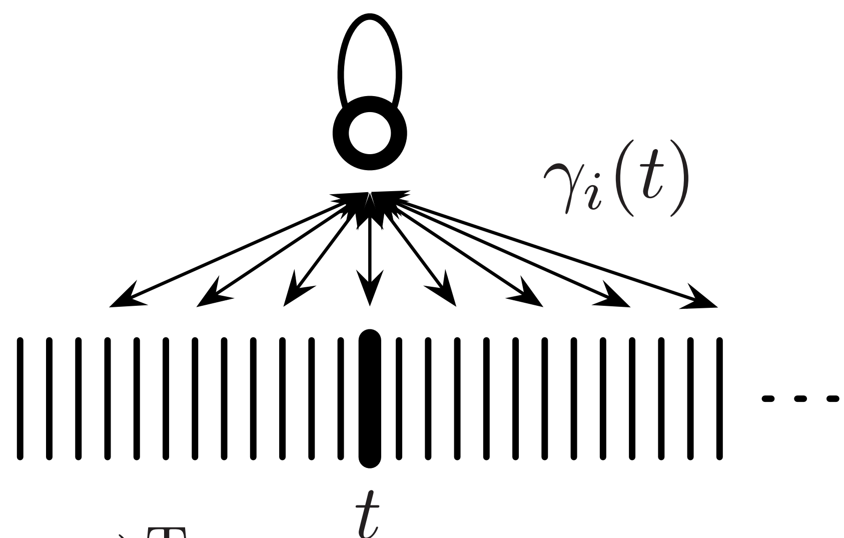
$$\Sigma = \frac{1}{T}\sum_t (o_t - \mu)(o_t - \mu)^{\mathrm{T}}$$



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



$\gamma_i(t) \quad (\sum_i \gamma_i(t) \equiv 1.0)$



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

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

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

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

Today's menu

- Fundamentals of statistical speech recognition
- Acoustic models (HMM) for speech recognition
- From word-based HMMs to phoneme-based HMMs
- From GMM-HMM to DNN-HMM

- Speech recognition using network grammars
- Speech recognition using N-grams
- Speech recognition using NN-based language models

- Module-based ASR to one-package (E2E) ASR (next week)

Phonemes

The minimum units of spoken language

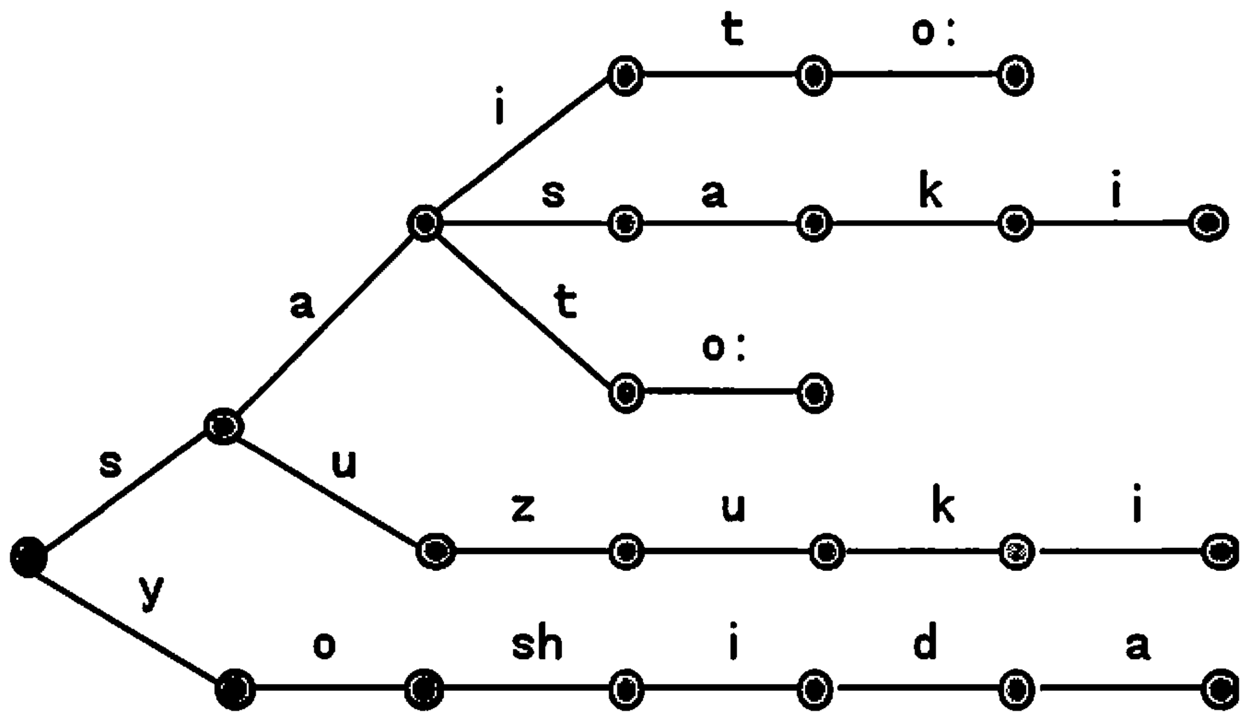
Vowels	short vowels	a, i, u, e, o
	long vowels	a:, i:, u:, e:, o:
Consonants	plosives	b, d, g, p, t, k
	fricatives	s, sh, z, j, f, h
	affricates	ch, ts
	拗音:	ky, py, ..
	semi-vowels	r, w, y
	nasals	m, n, N

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

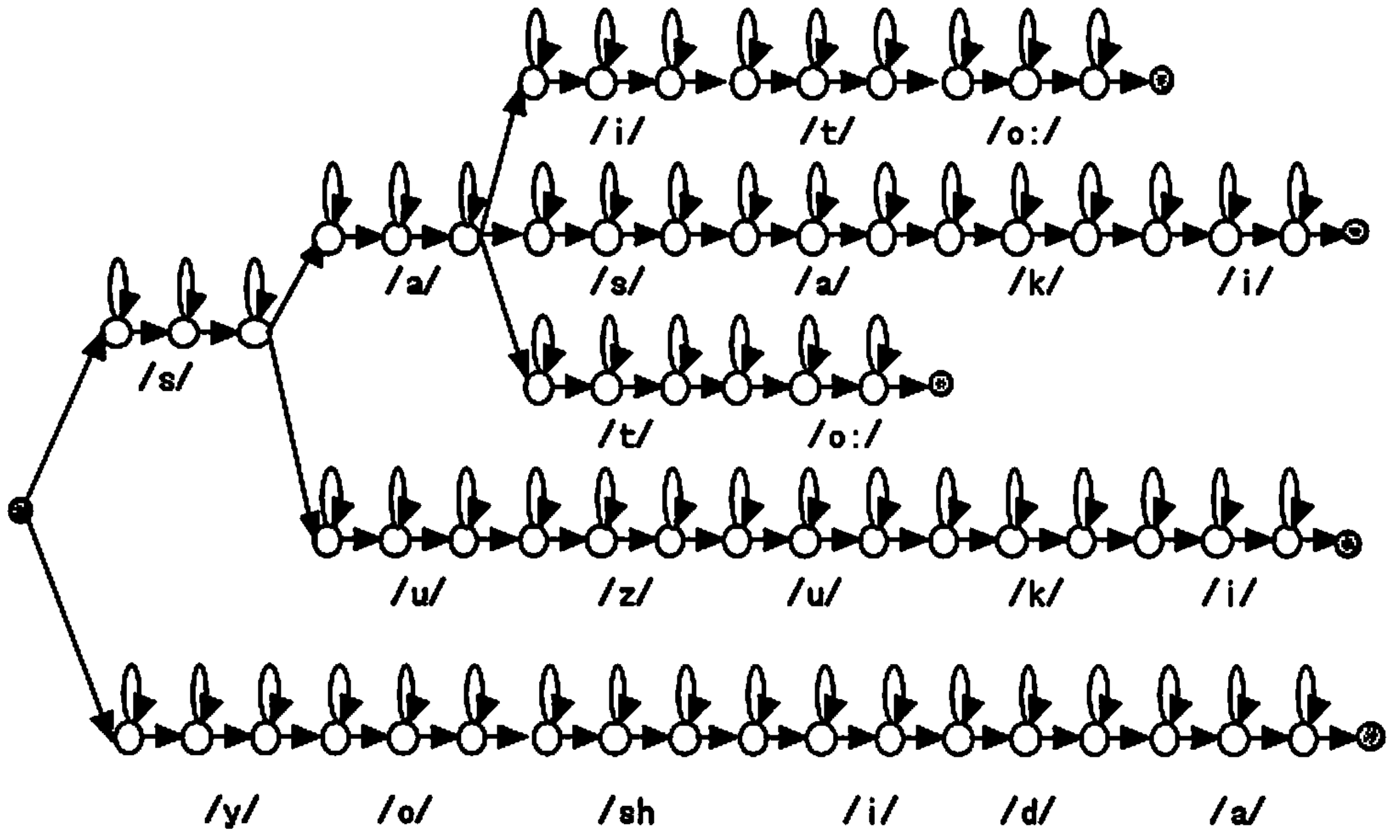
Tree lexicon (compact representation of the words)



The following words are stored as a tree.

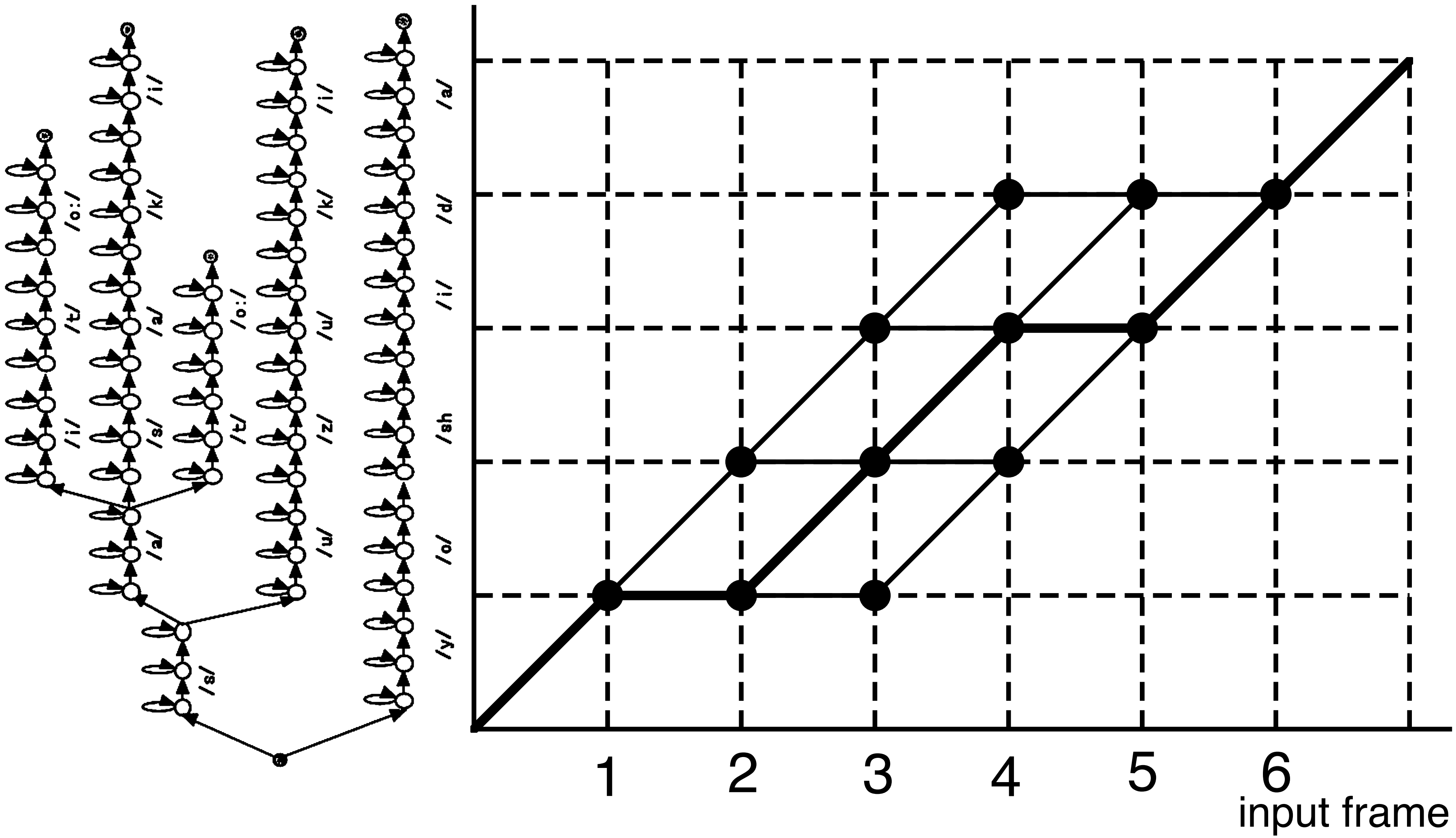
saito: (斉藤), sasaki (佐々木), sato: (佐藤)
suzuki (鈴木) , yoshida (吉田)

Tree-based lexicon using phoneme HMMs



Generation of state-based network containing all the candidate words

Recognition of names



Search for the maximum likelihood path

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

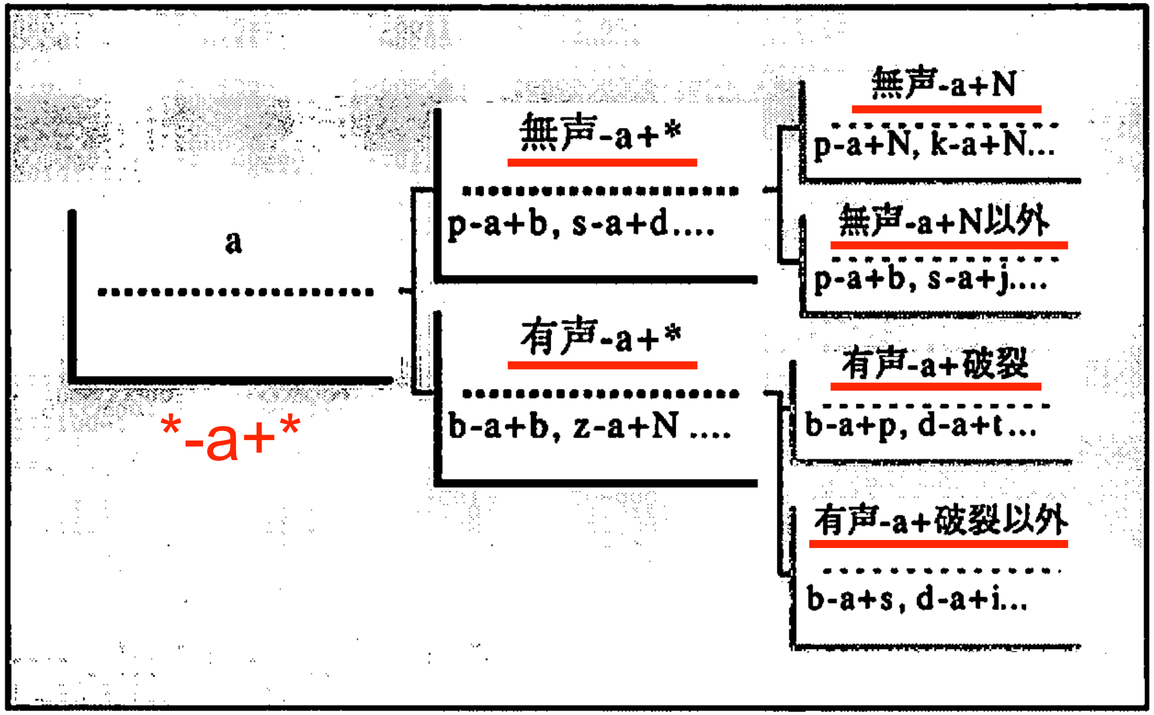
trihphone

50 vs. 50*50*50 = 125,000

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 $\approx$ 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 effect is easy to model.
	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 GMM-HMM to DNN-HMM

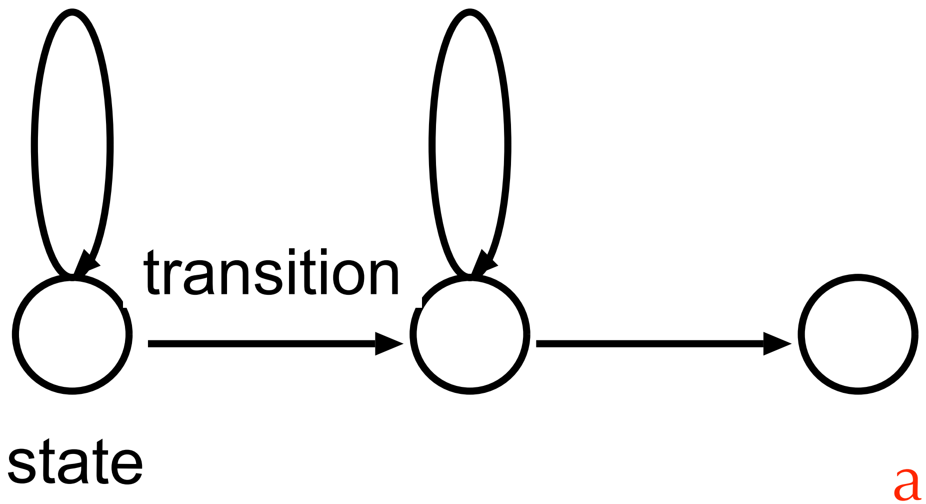
- Speech recognition using network grammars
- Speech recognition using N-grams
- Speech recognition using NN-based language models

- Module-based ASR to one-package (E2E) ASR (next week)

How to make a difficult problem tractable?

- 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 estimate 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 read aloud w or p !!
    - P(o|w): acoustic model, P(w): linguistic model
      - Two separate modules + the other one that searches for the word sequence that maximizes P(w,o)

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.

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

Backward prob.

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

GMM-HMM to DNN-HMM

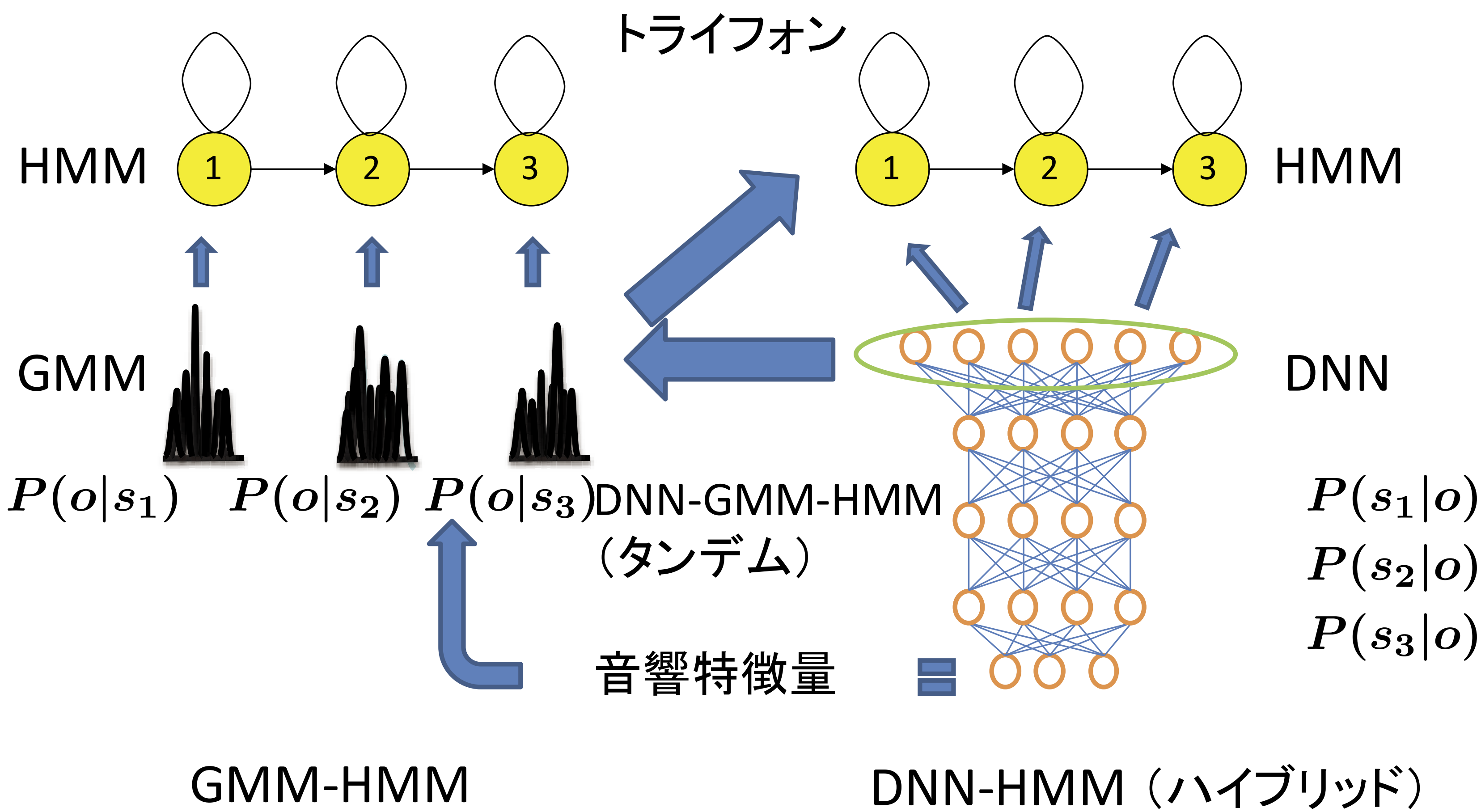
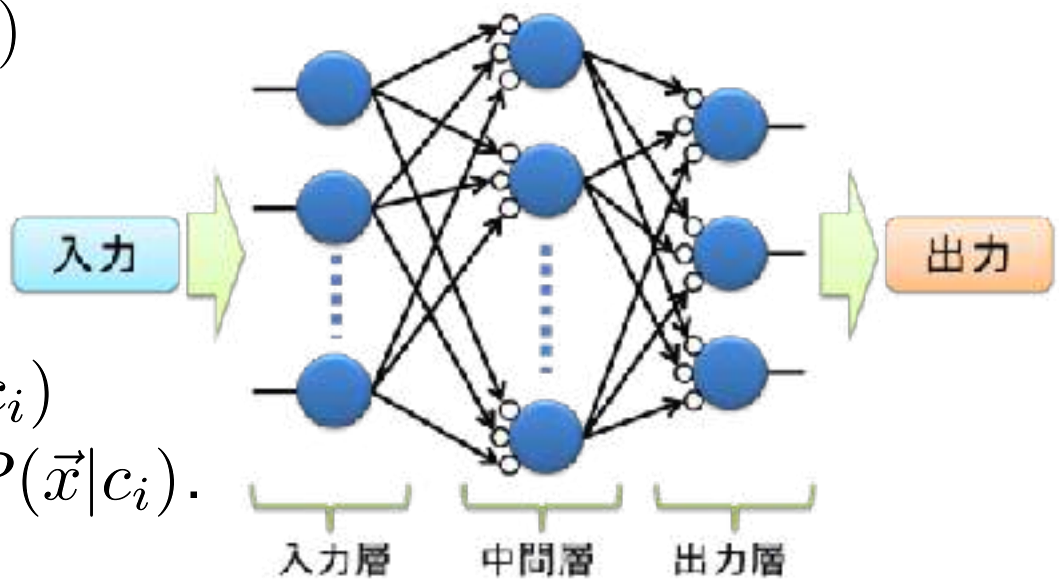


図 2　GMM-HMM と DNN-HMM

DNN as phoneme posterior calculator

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



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

GMM-HMM to DNN-HMM

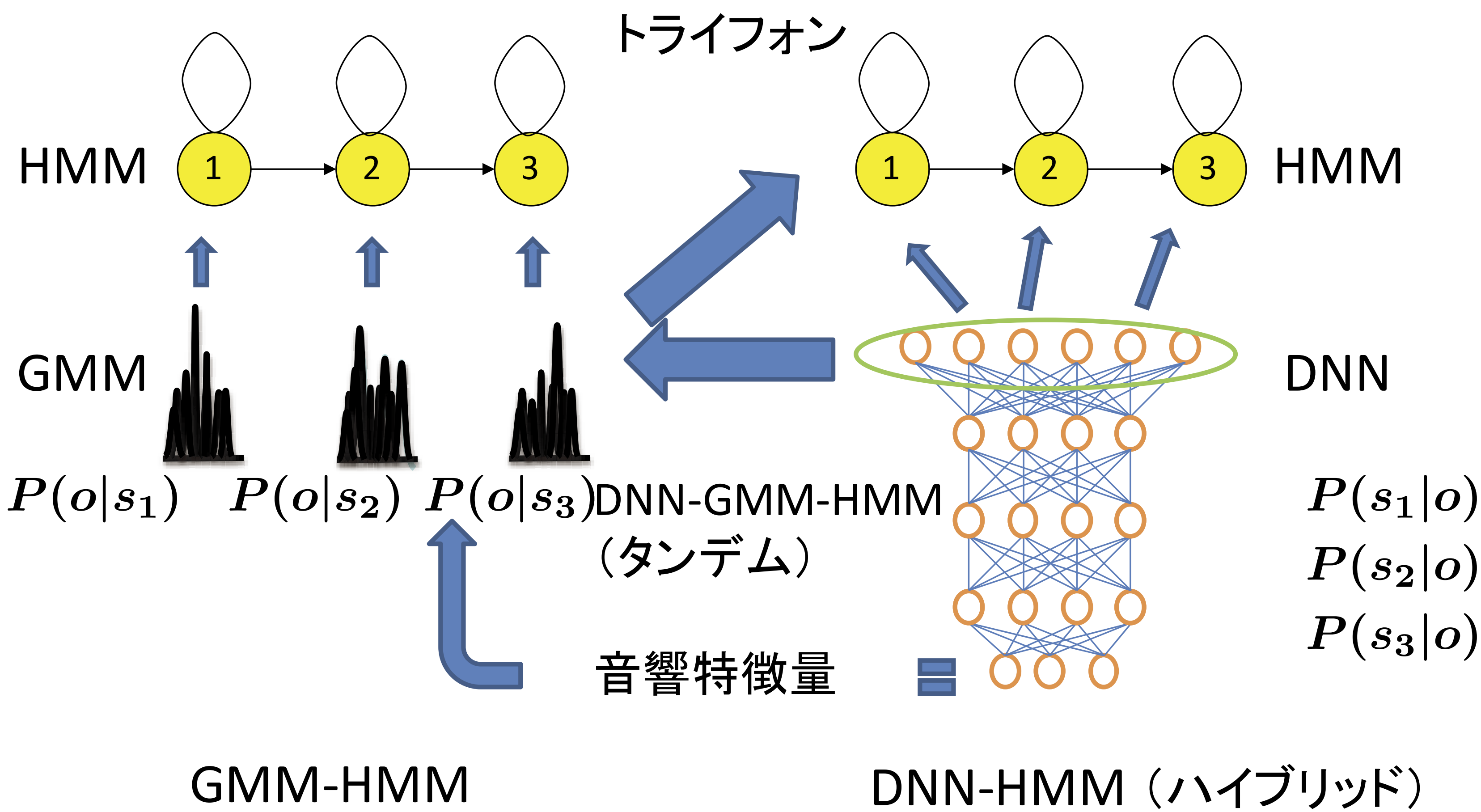
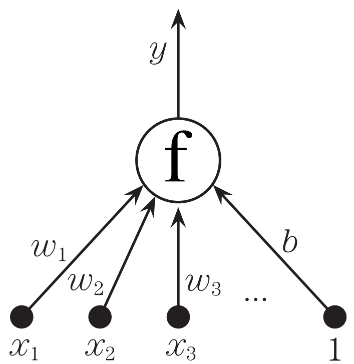


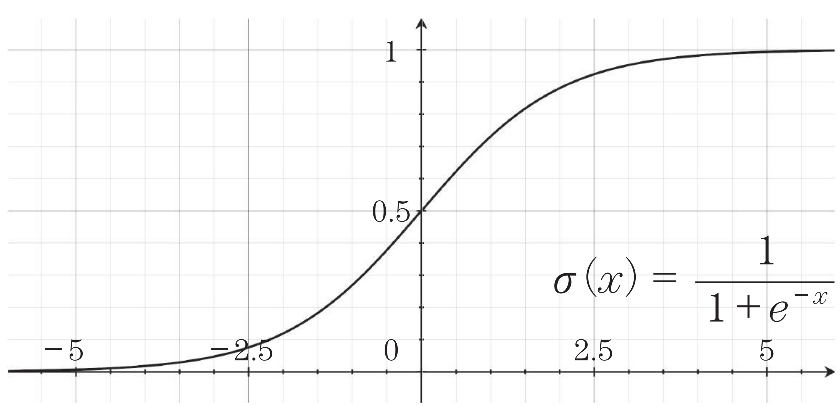
図 2　GMM-HMM と DNN-HMM

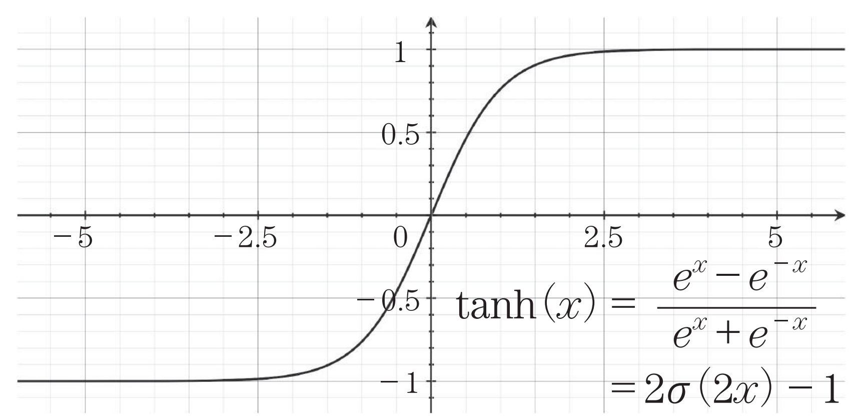
Artificial Neural Network

- A model of a single neuron
  - Linear transform + non-linear normalization



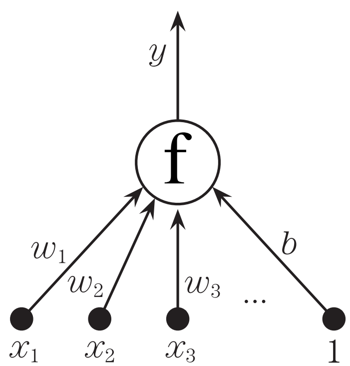
$$y = f\left(\sum_i w_i x_i + b\right) = f(u), \qquad \left(u = \sum_i w_i x_i + b\right)$$

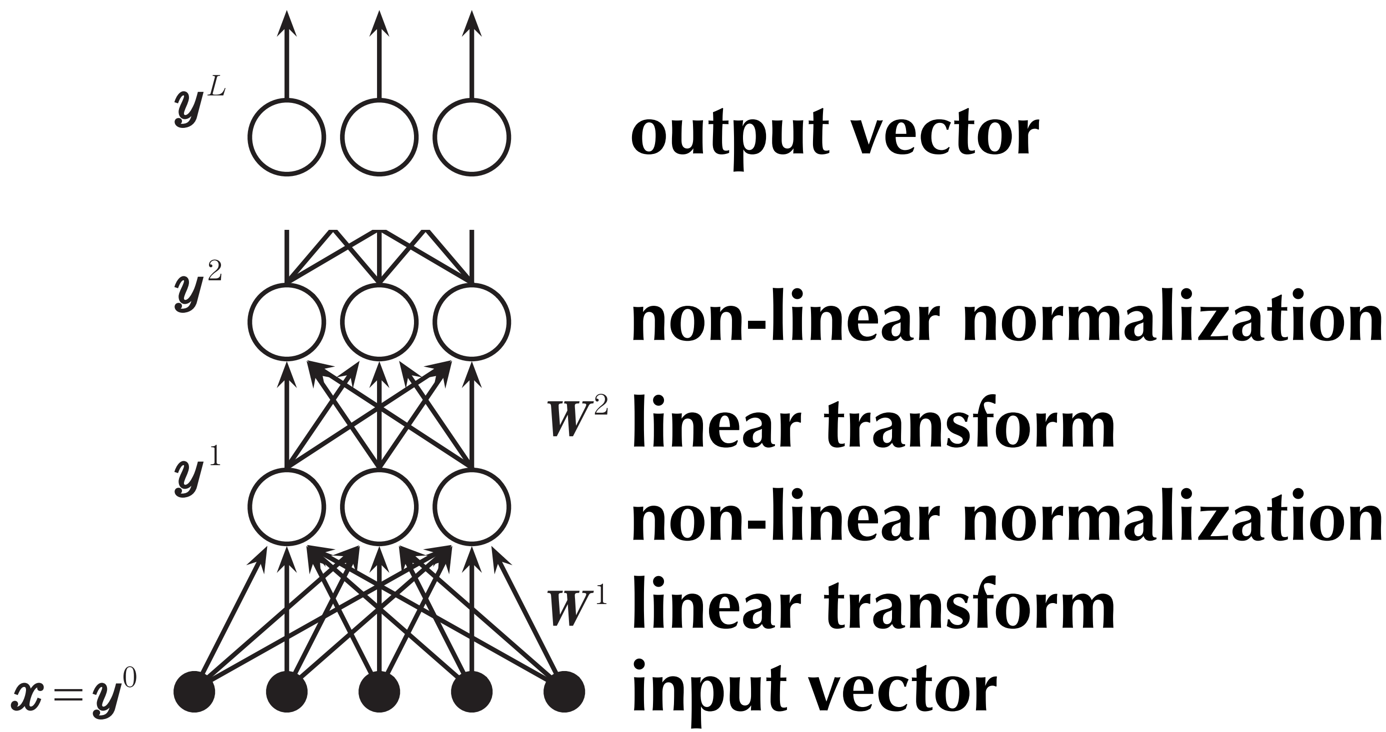




Artificial Neural Network

- A model of a network of neurons
  - Linear transform + non-linear normalization

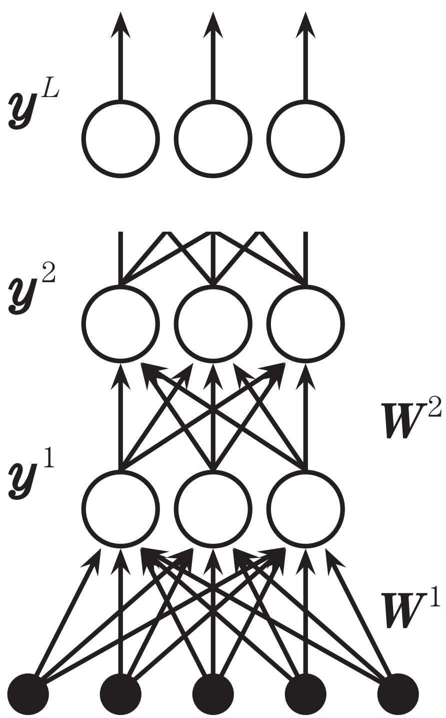




$$\boldsymbol{y}^l = \boldsymbol{f}\left(\boldsymbol{W}^l \boldsymbol{y}^{l-1} + \boldsymbol{b}^l\right) = \boldsymbol{f}\left(\boldsymbol{u}^l\right)$$

Artificial Neural Network

- How to train the network so that it can classify the input vector.
  - A classifier is trained so that it can output the posterior probability of P(c | x).
    - The dimension of the output vector = #classes
  - Training data = pairs of data and class
    - Output vector = (0, 0, 0, ..., 0, 1, 0, 0 0)
  - How to train the classifier so that it can output a probability distribution.
    - The final non-linear transform functions as normalizer for probability distribution



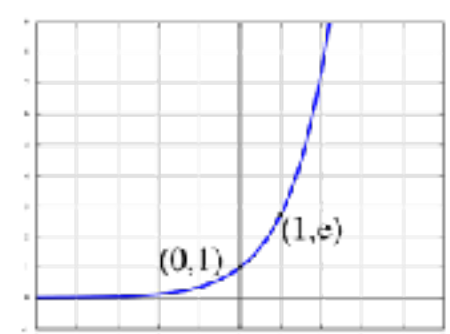
output vector
softmax

non-linear normalization
linear transform
non-linear normalization
linear transform
input vector

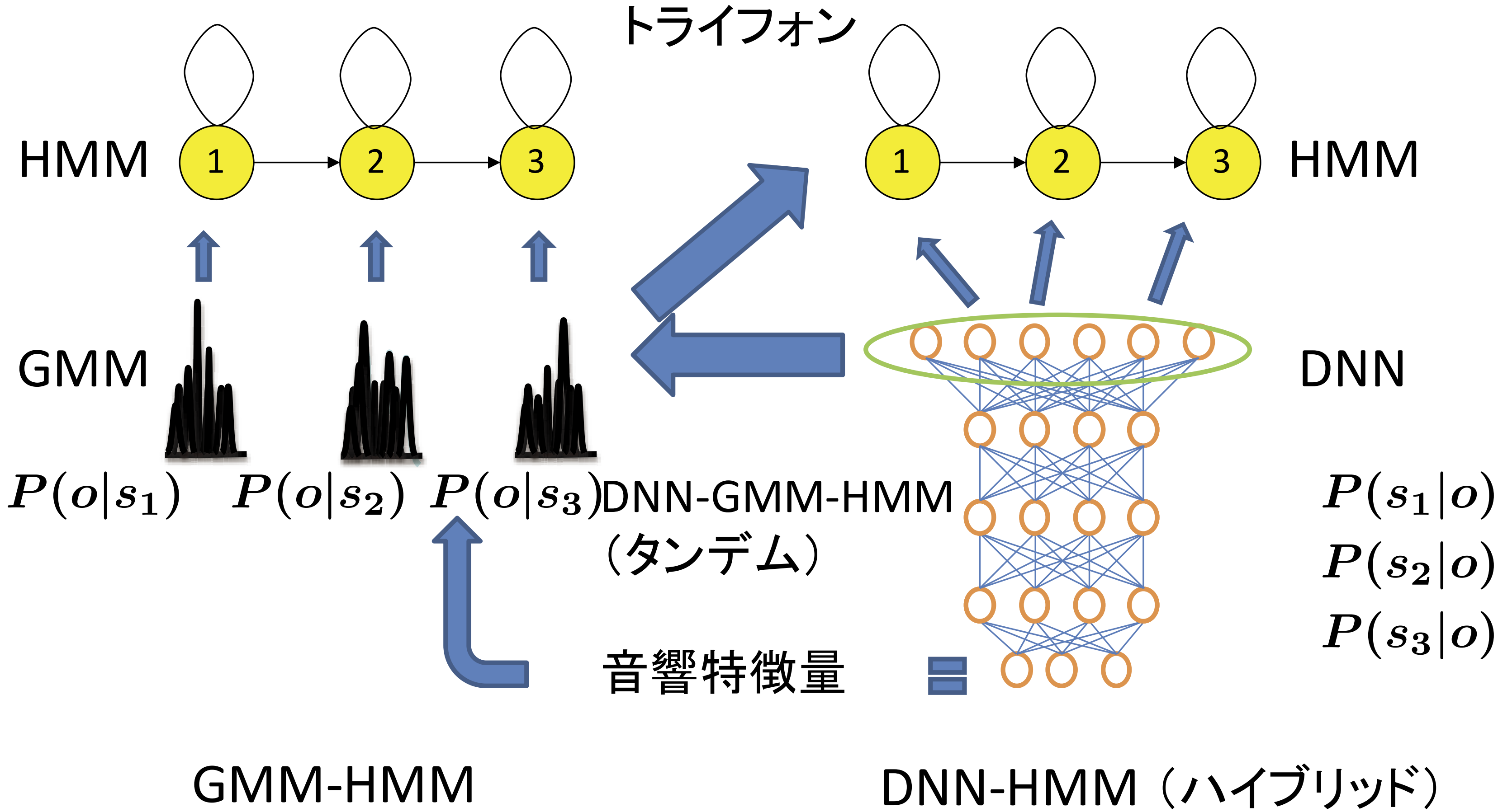
$$\boldsymbol{y}^l = \boldsymbol{f}\left(\boldsymbol{W}^l \boldsymbol{y}^{l-1} + \boldsymbol{b}^l\right) = \boldsymbol{f}\left(\boldsymbol{u}^l\right)$$

$$\boldsymbol{u}^L = \boldsymbol{W}^L \boldsymbol{y}^{L-1} + \boldsymbol{b}^L$$

$$P(C_j|\boldsymbol{o}_t) = \frac{\exp(u_j^L)}{\sum_k \exp(u_k^L)}$$



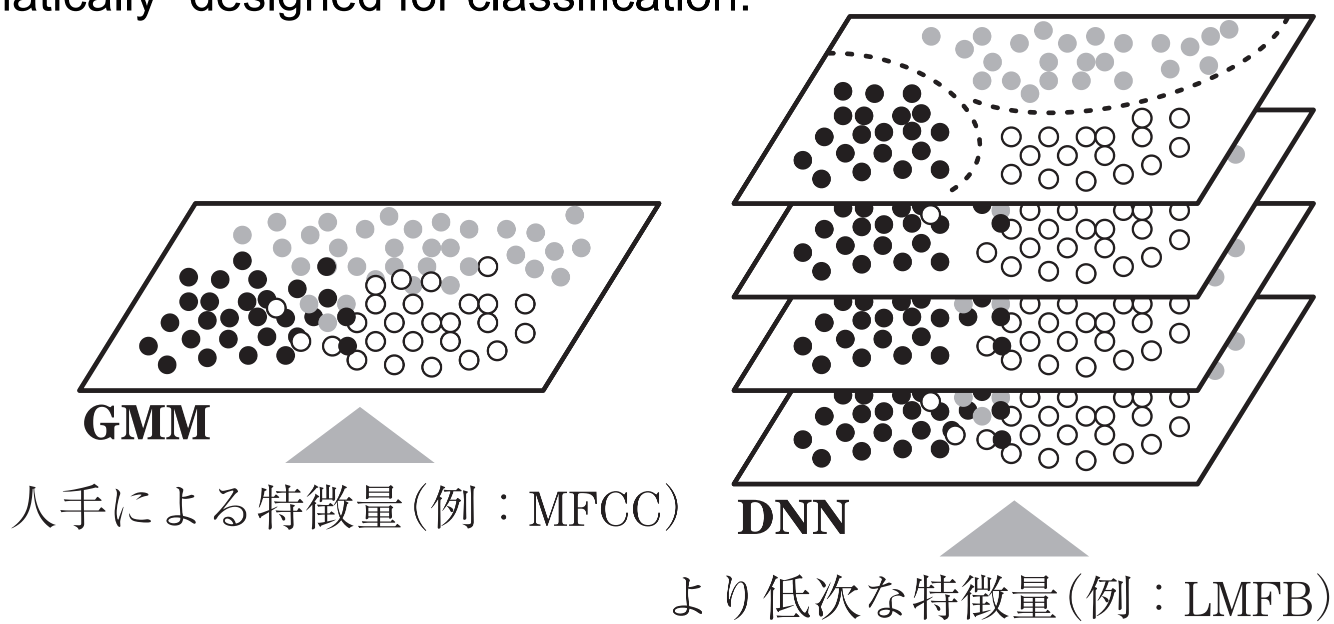
GMM-HMM to DNN-HMM



- How to obtain the HMM state for each frame in the training data?
  - DNN-HMM trains GMM-HMM internally at first.
  - (Forced) alignment between GMM-HMM and training data is done.
  - Then, the state for each frame is fixed and labeled.

Why GMM-HMM < DNN-HMM?

- GMM = Generative model, DNN = Discriminative model
  - Generative model has to characterize the probability distribution of manually-crafted features such as cepstrum coefficients, given classes (= P(o | c))
  - Discriminative model has to characterize the probability distribution of classes, given acoustic observations (= P(c | o))
    - o → linear transform + non-linear normalization → o'
    - o' → linear transform + non-linear normalization → o''
    - Multiple "feature" transformations are trained (designed) so that better features are "automatically" designed for classification.



Today's menu

- Fundamentals of statistical speech recognition
- Acoustic models (HMM) for speech recognition
- From word-based HMMs to phoneme-based HMMs
- From GMM-HMM to DNN-HMM

- Speech recognition using network grammars
- Speech recognition using N-grams
- Speech recognition using NN-based language models

- Module-based ASR to one-package (E2E) ASR (next week)

How to make a difficult problem tractable?

- 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 estimate 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 read aloud w or p !!
    - P(o|w): acoustic model, P(w): linguistic model
      - Two separate modules + the other one that searches for the word sequence that maximizes P(w,o)

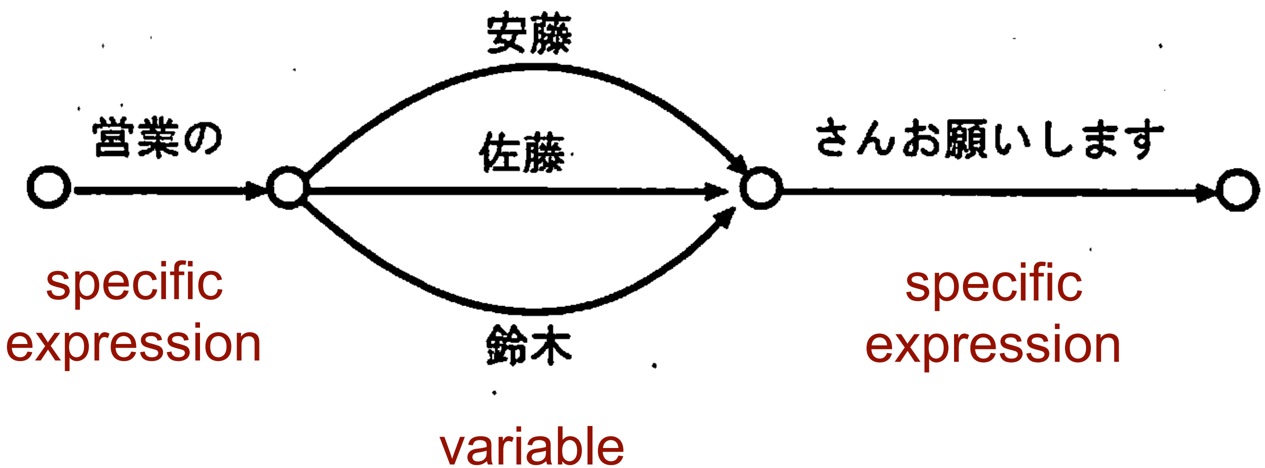
Continuous speech (connected word) recognition

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

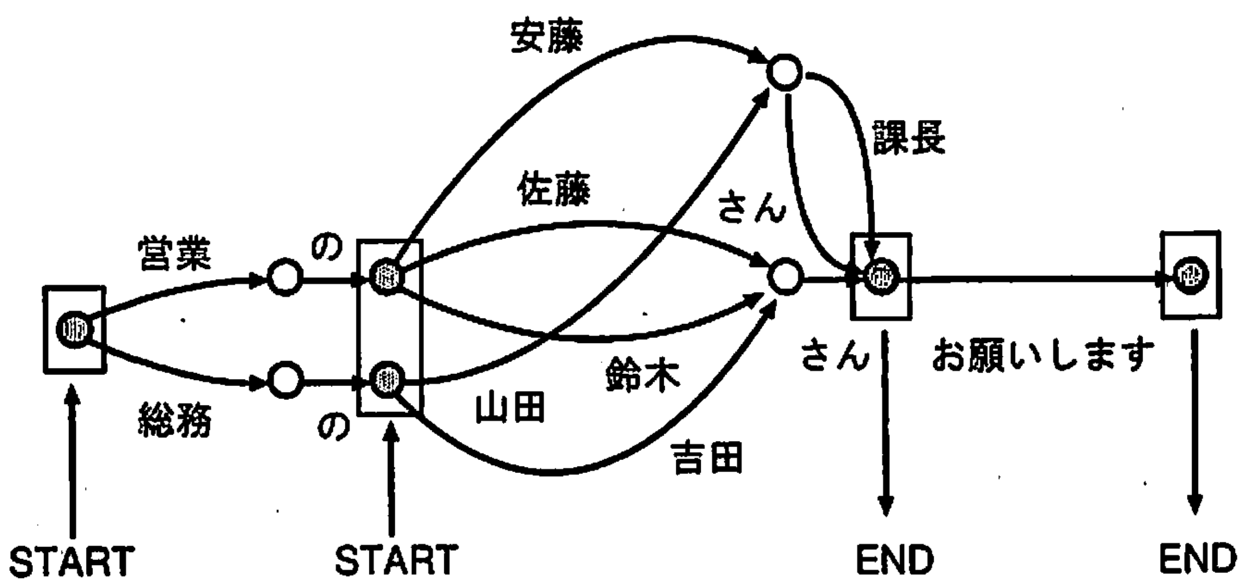
- Constraints on words and their sequences
  - ∗ Vocabulary: a set of candidate words
  - ∗ Syntax: how words are arranged linearly.
  - ∗ Semantics: can be represented by word order??
- Examples of unaccepted sentences
  - ∗ 私/は/マッキンポッシュ/を/使う。(lexical error)
  - ∗ 私/マッキントッシュ/は/使う/を。(syntax error)
 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

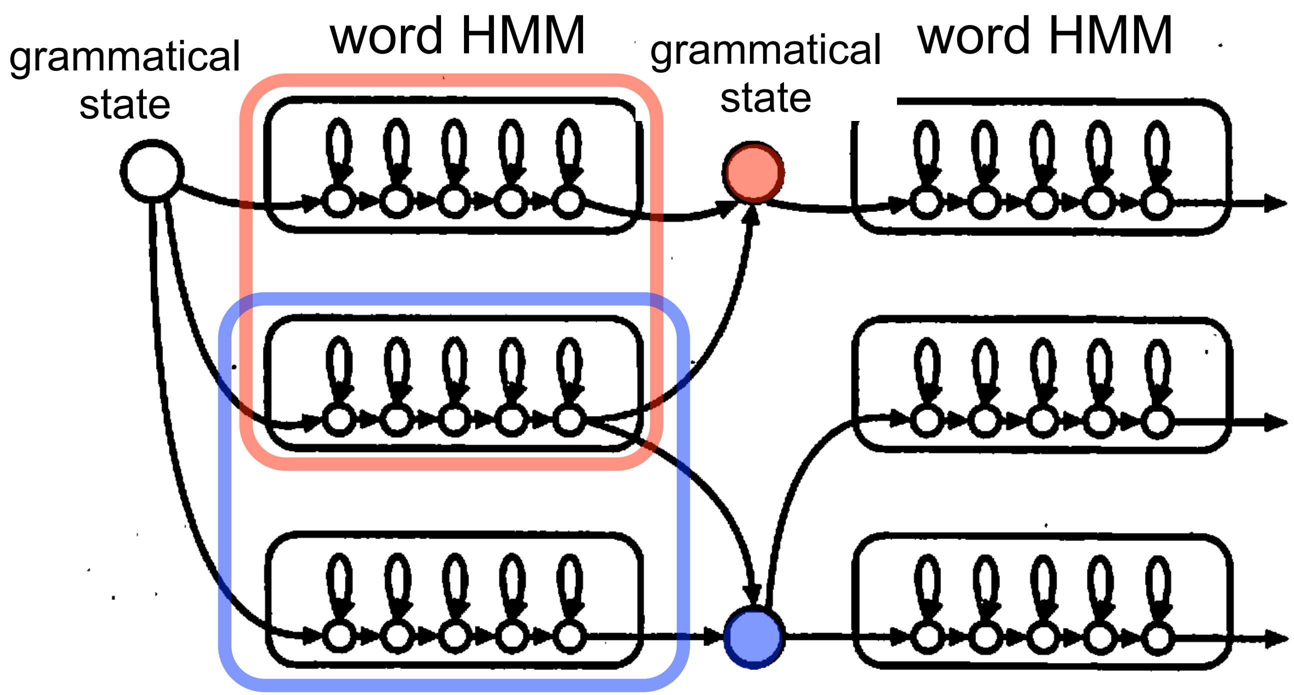


Network grammar with a finite set of states



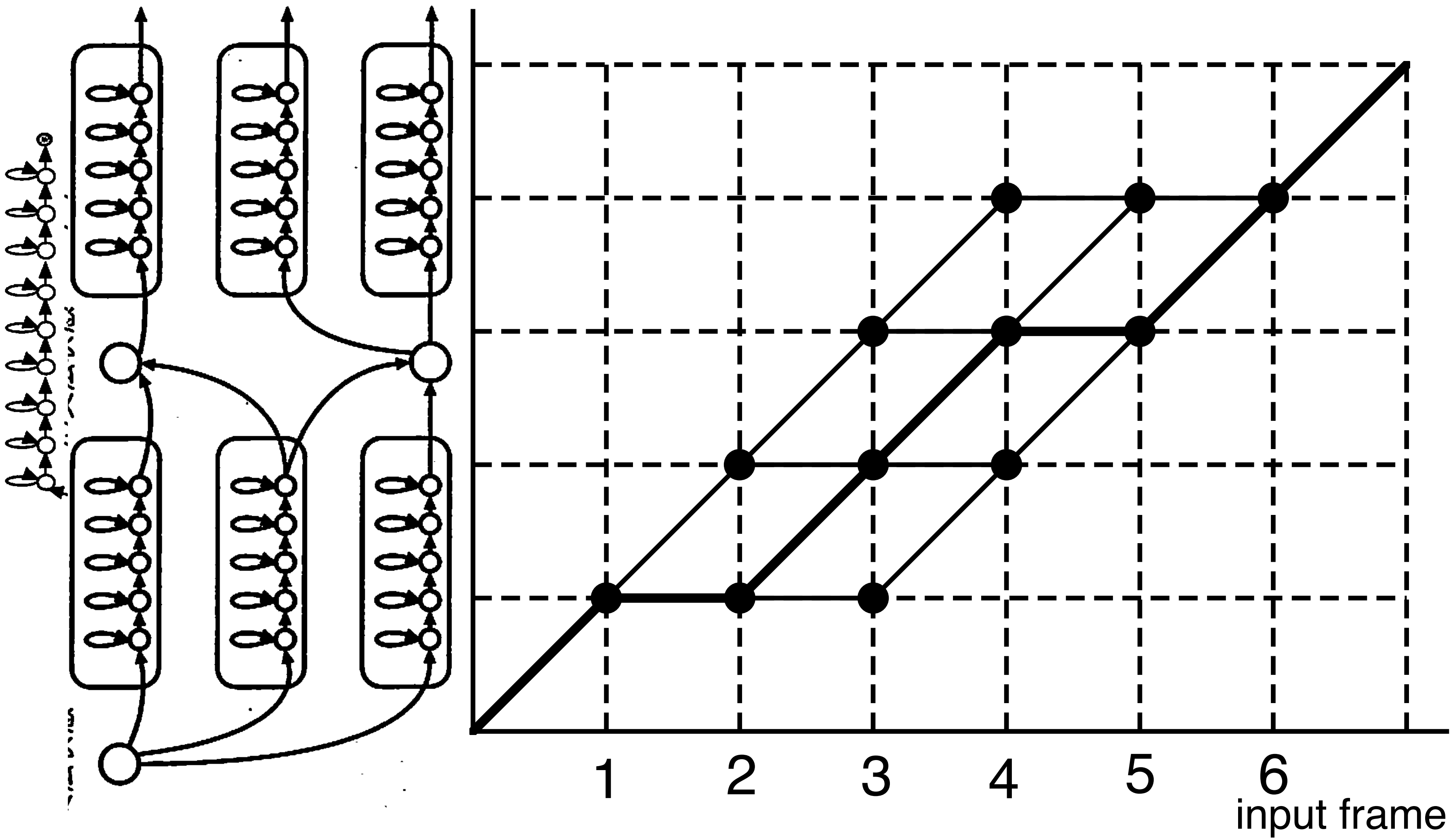
A sentence is accepted if it starts at one of the initial states and ends at one of the final states.

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.

Recognition of names

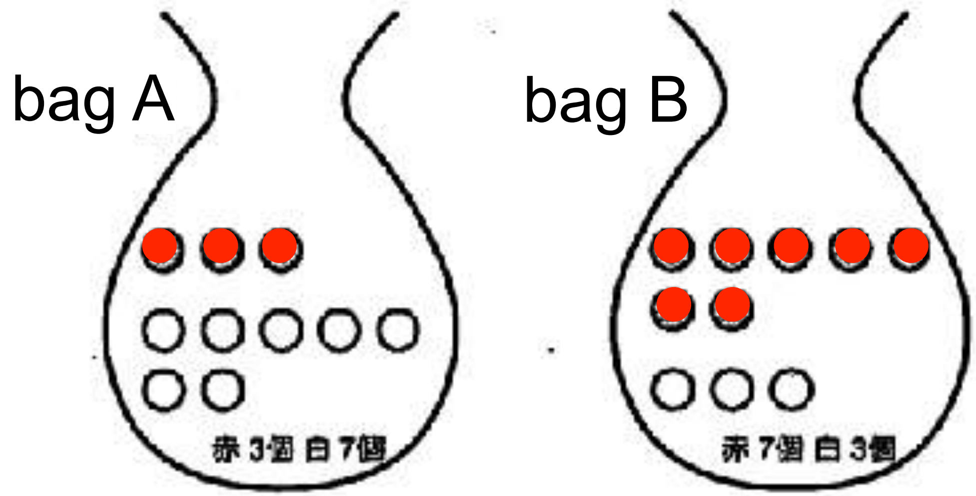


Search for the maximum likelihood path

How to make a difficult problem tractable?

- 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 estimate 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 read aloud w or p !!
    - P(o|w): acoustic model, P(w): linguistic model
      - Two separate modules + the other one that searches for the word sequence that maximizes P(w,o)

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 more likely 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

$$P(\text{袋A} \mid ●○●) = 0.063 \times 0.75 = 0.04725$$
$$P(\text{袋B} \mid ●○●) = 0.147 \times 0.25 = 0.03675$$

The bag used is likely 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

n-1 words

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

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

N-1 words

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

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

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

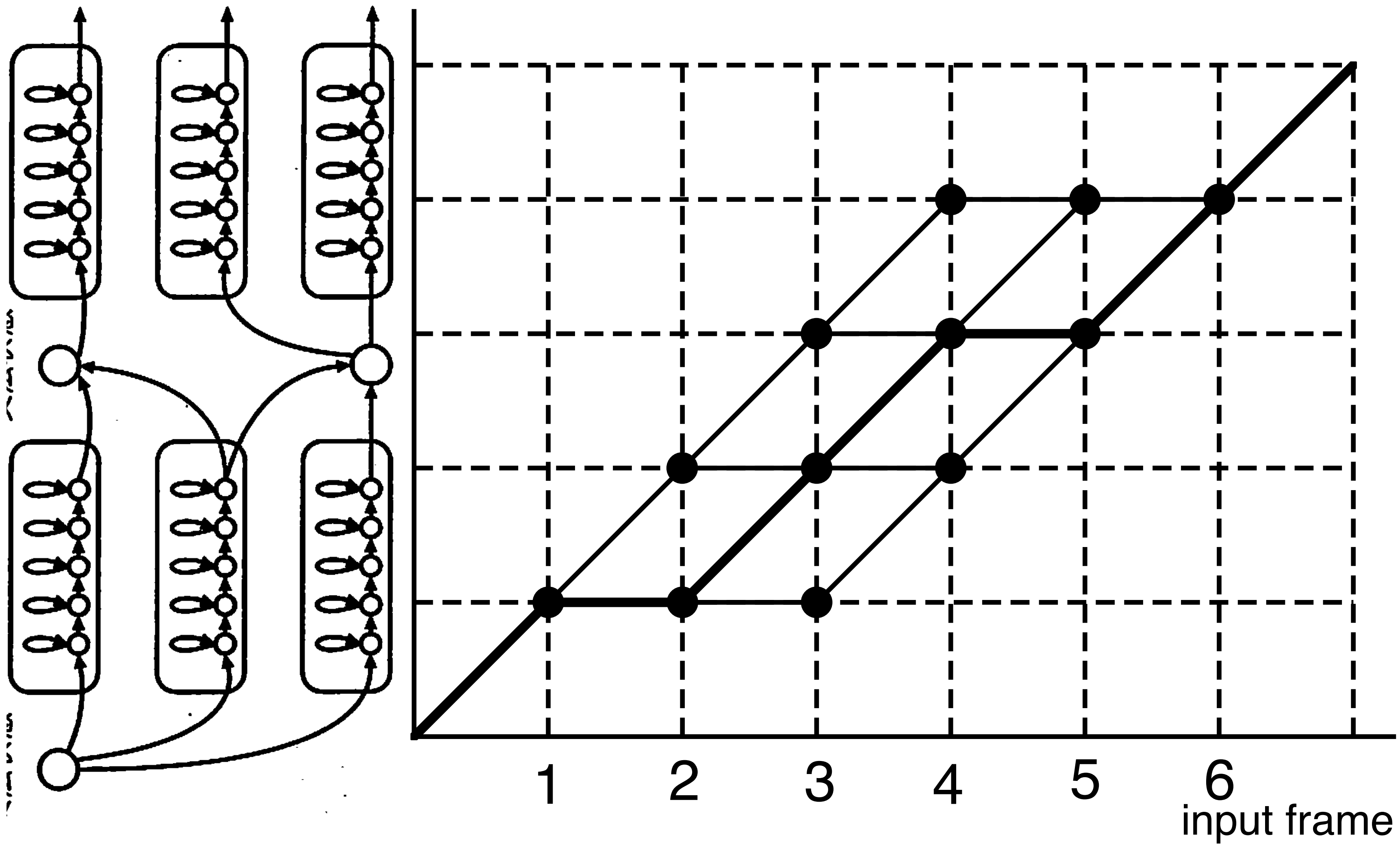
I'm giving a lecture on speech recognition technology to university students.

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 | w2) or β x P(w3) are substituted as P (w3 | w1, w2).
  Context dependencies are ignored to some degree.

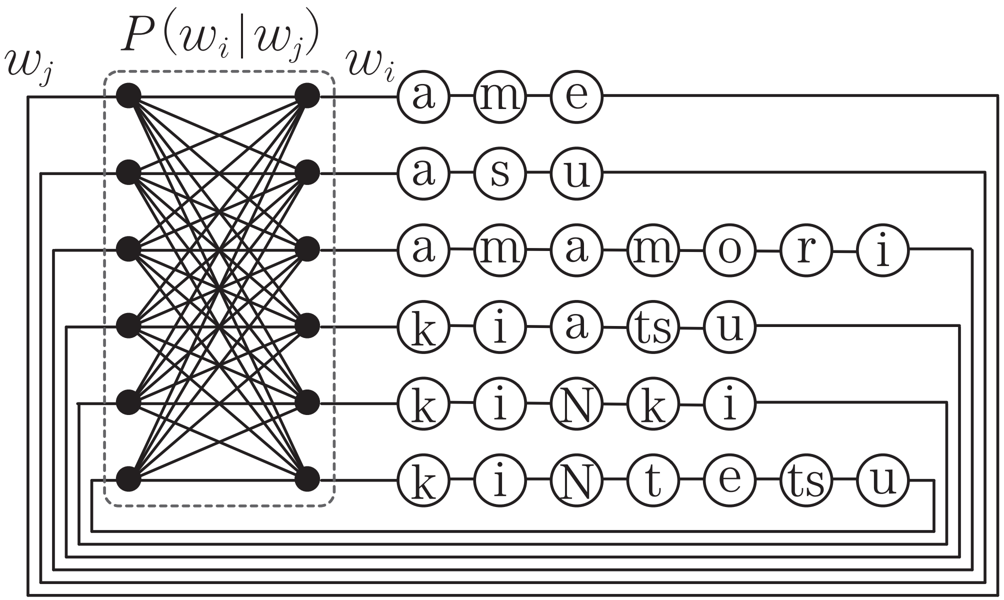
Recognition of isolated words

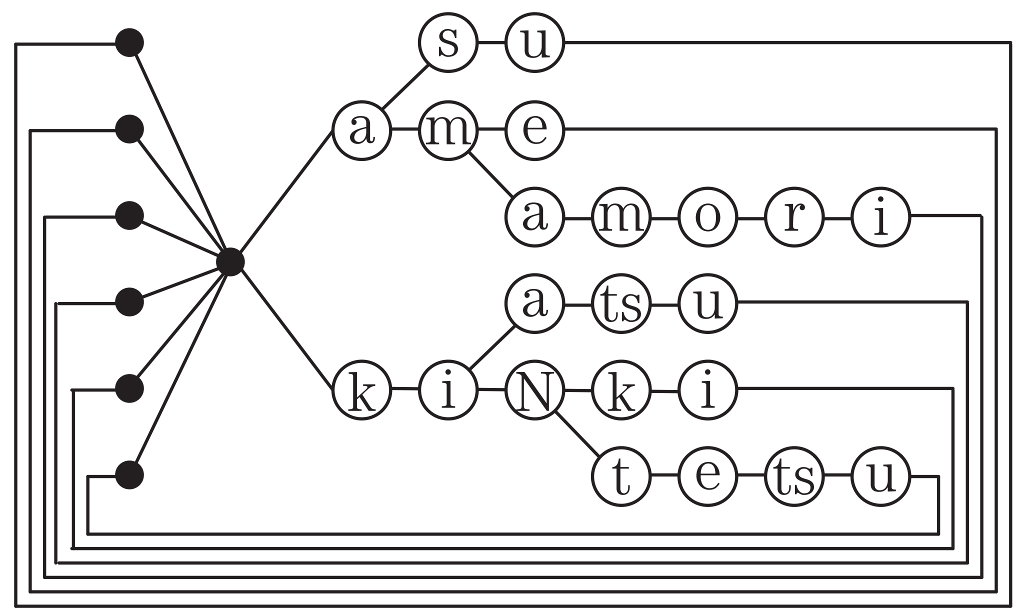


Search for the maximum likelihood path

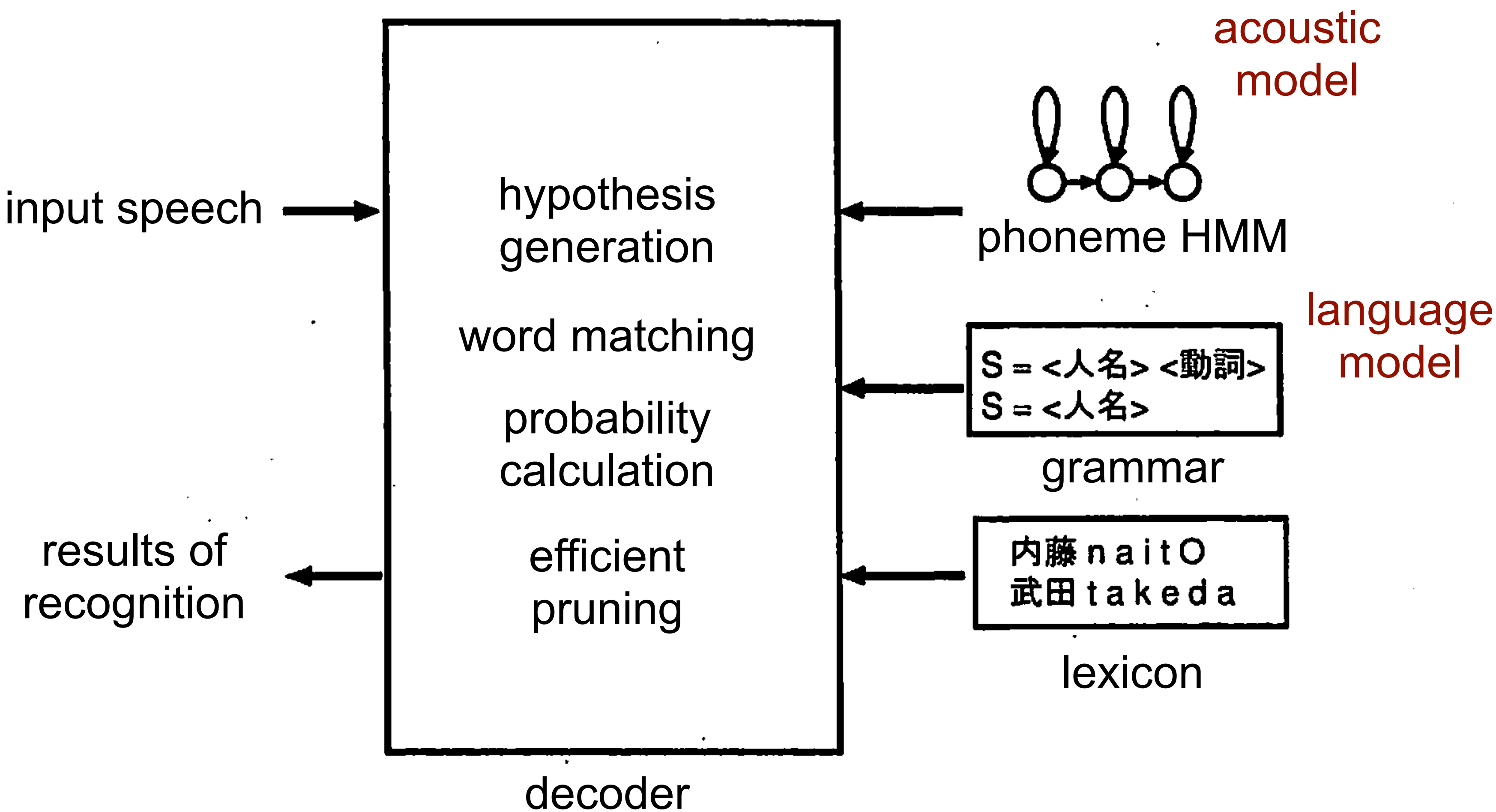
2-gram as network grammar

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





Development of a speech recognition system



Module-based ASR

Today's menu

- Fundamentals of statistical speech recognition
- Acoustic models (HMM) for speech recognition
- From word-based HMMs to phoneme-based HMMs
- From GMM-HMM to DNN-HMM

- Speech recognition using network grammars
- Speech recognition using N-grams
- Speech recognition using NN-based language models

- Module-based ASR to one-package (E2E) ASR (next week)

Recommended books