Cognitive Media Processing #8

Nobuaki Minematsu

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 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)

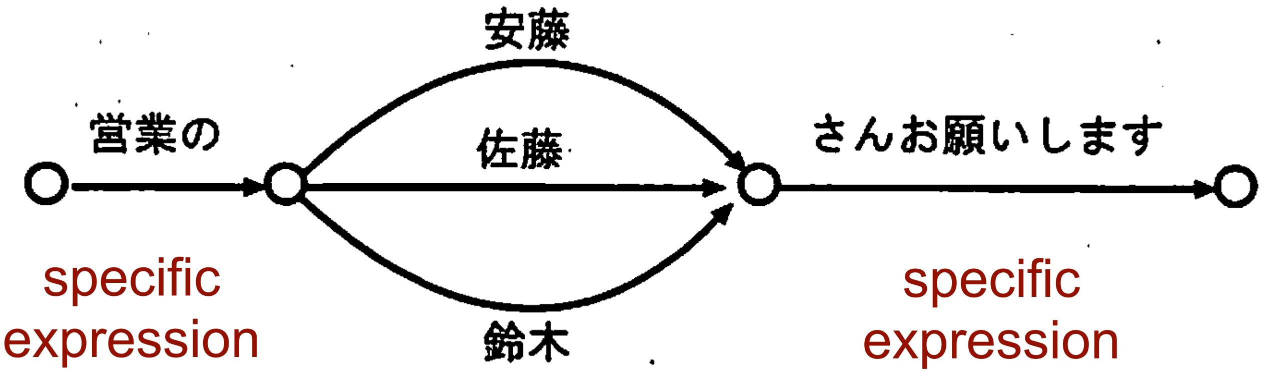
Continuous speech (connected word) recognition

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)

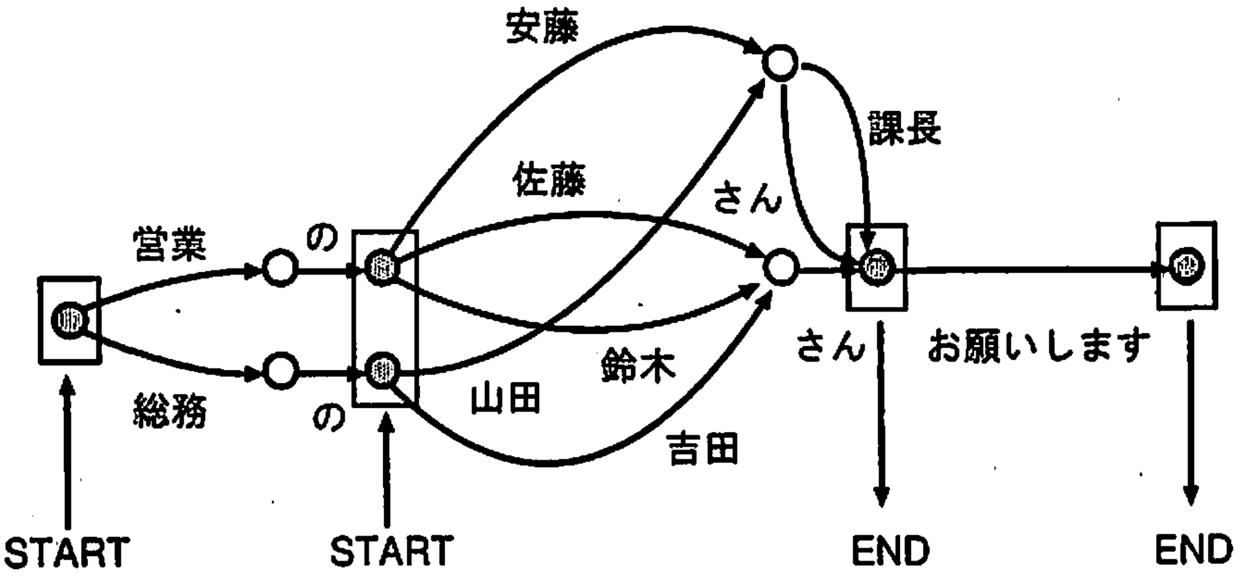
Representation of syntax (grammar)

- 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。



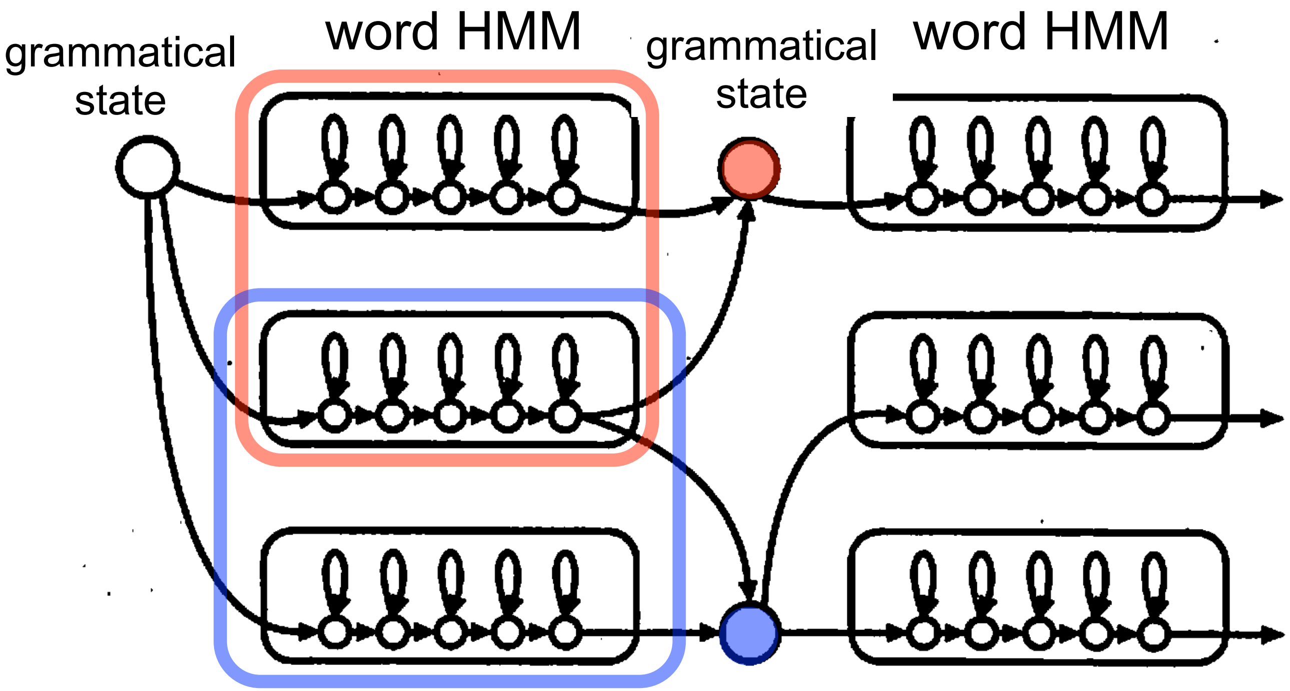
variable

Network grammar with a finite set of states



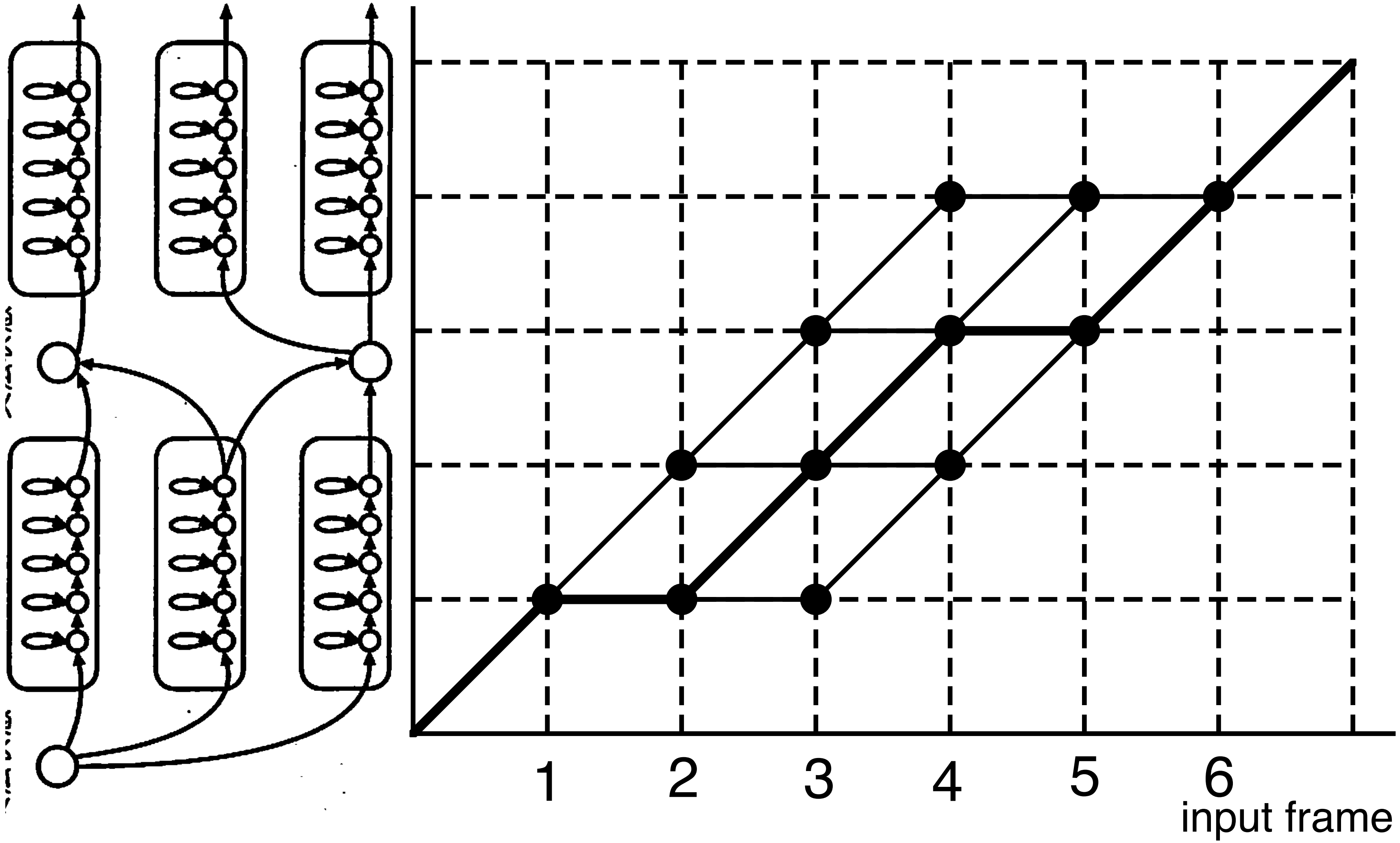
A sentence is accepted if it starts at one of the initial states and ends at one of the final states.

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 isolated words

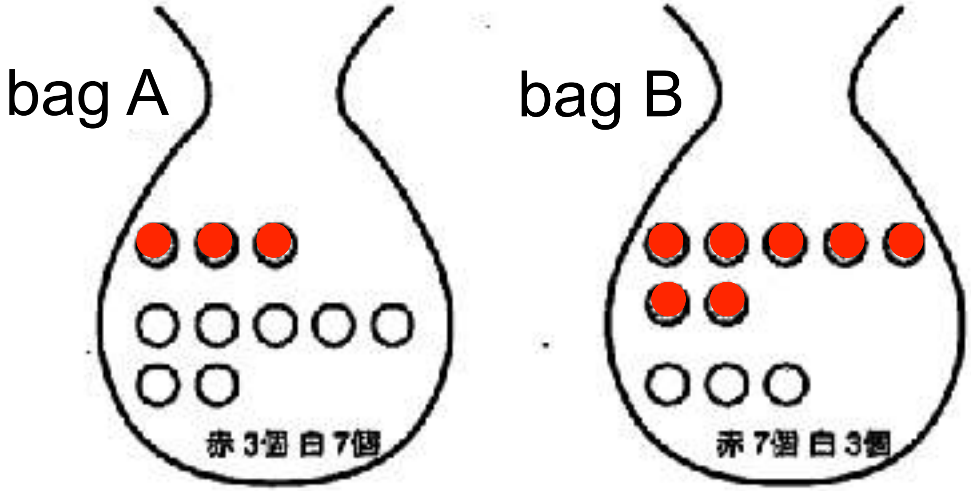


Search for the maximum likelihood path

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：$\frac{3}{10}\times\frac{7}{10}\times\frac{3}{10}=0.063$ 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(袋A | ●○●) = 0.063 × 0.75 = 0.04725
P(袋B | ●○●) = 0.147 × 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

$$P(x_1,\cdots,x_n) = \underbrace{P(x_n|\overset{\text{n-1 words}}{\underline{x_1,\cdots,x_{n-1}}})} \quad P(x_1,\cdots,x_{n-1})$$

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

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

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

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 ....

  $P(\text{speech} \mid \text{lecture}, \text{on})$

  $= C(\text{lecture}, \text{on}, \text{speech}) / C(\text{lecture}, \text{on})$

  $P(\text{recognition} \mid \text{on}, \text{speech})$

  $= C(\text{on}, \text{speech}, \text{recognition}) / C(\text{on}, \text{speech})$

  $P(w3 \mid w1, w2)$

  $= C(w1, w2, w3) / C(w1, w2)$

- Typical problems of calculating N-gram prob

  $C(w1, w2, w3) = 0$ --> N-gram prob. = 0 ??

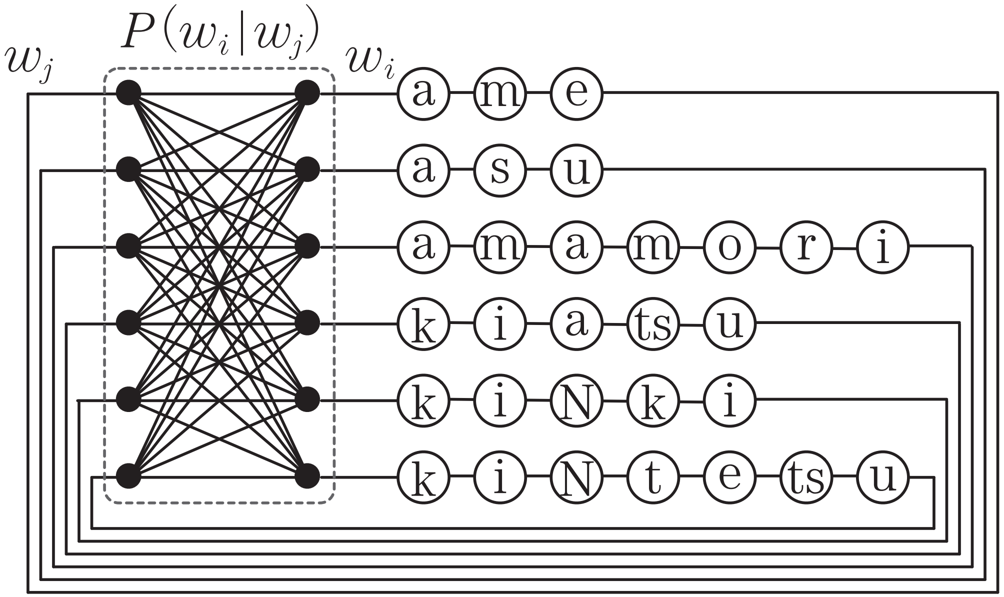
  $C(w1, w2) = 0$ --> N-gram prob. = ???

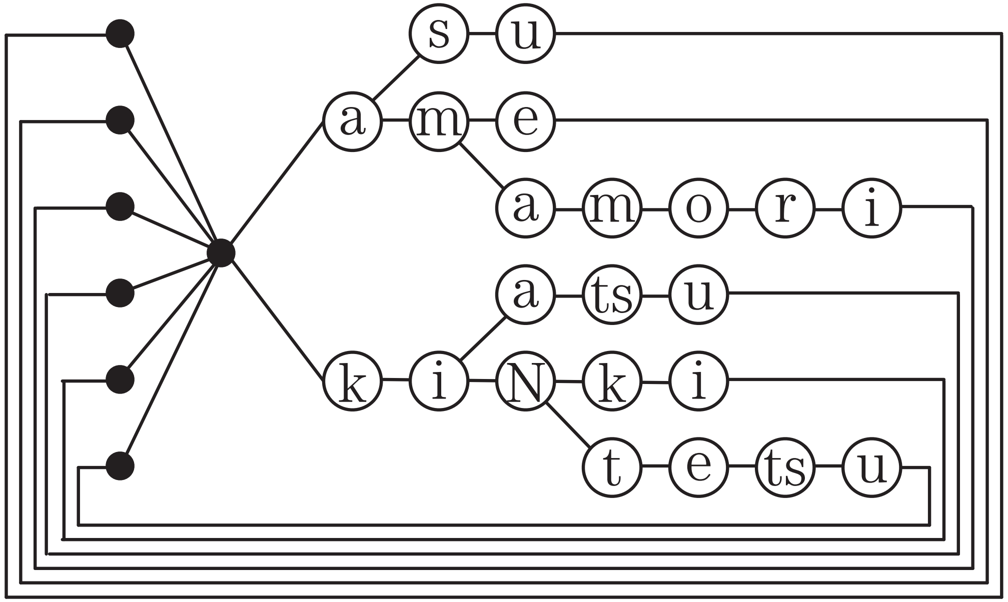
  $\alpha \times P(w3 \mid w1)$ or $\beta \times P(w3)$ are substituted as $P(w3 \mid w1, w2)$.

  Context dependencies are ignored to some degree.

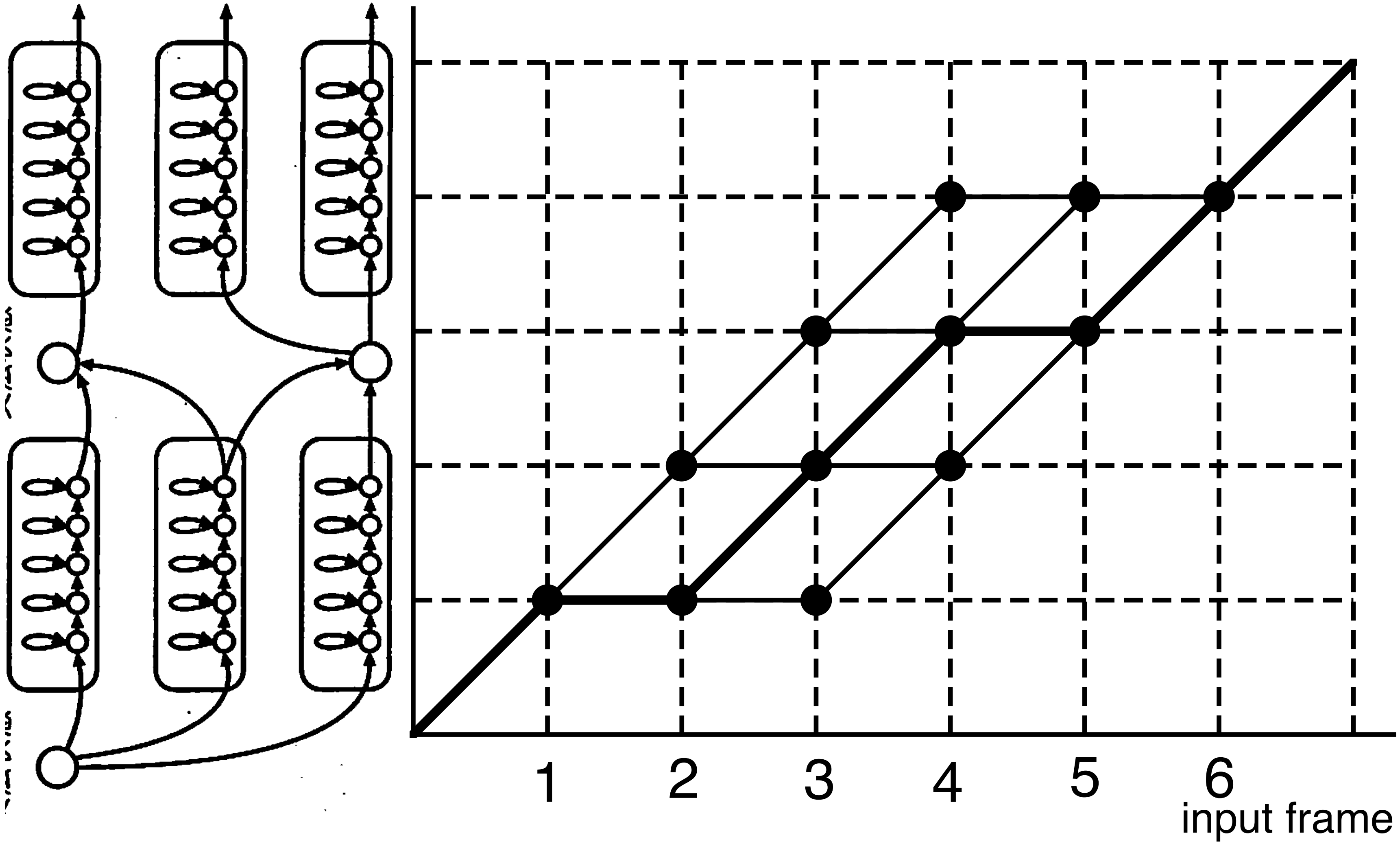
2-gram as network grammar

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





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 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)

N-gram language model

$P(s_1|o)$
$P(s_2|o)$
$P(s_3|o)$

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

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

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

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

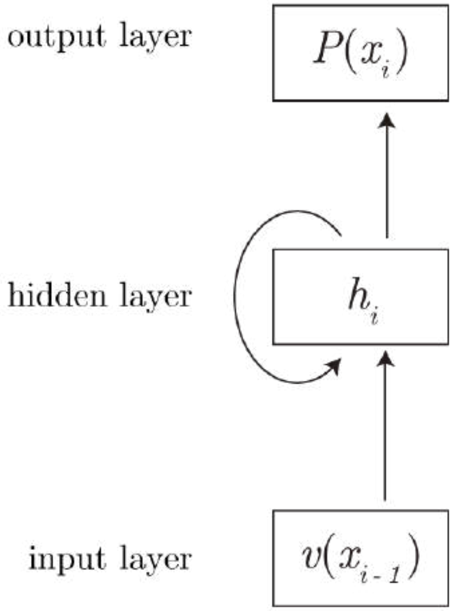
N-1 = 1 --> bi-gram
N-1 = 2 --> tri-gram

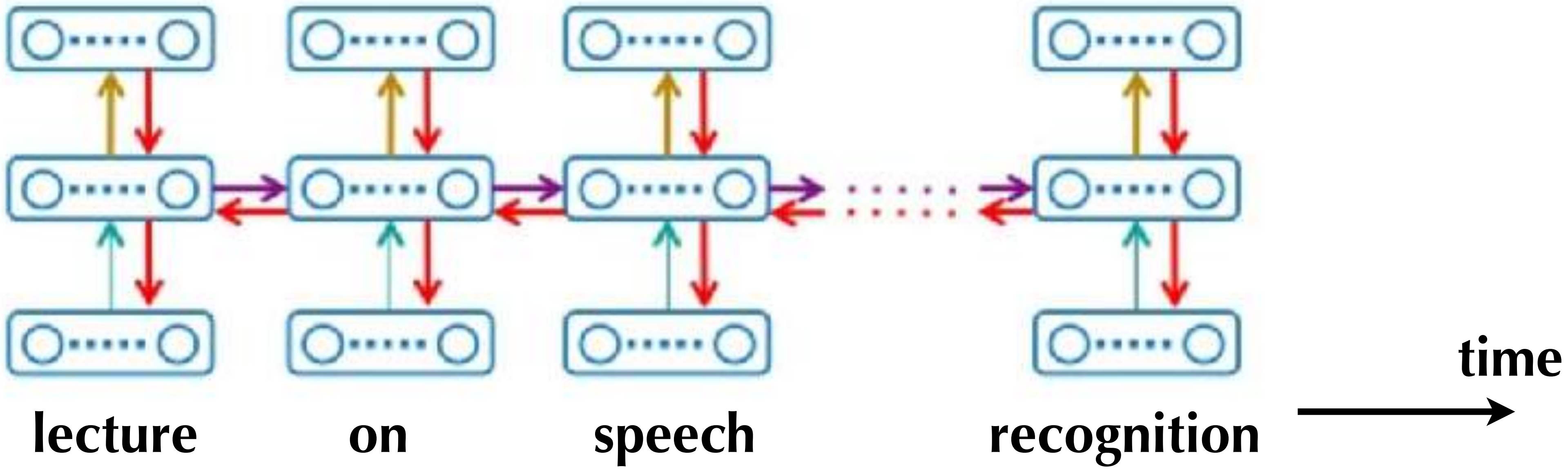
I’m giving a lecture on speech recognition technology to university students.

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

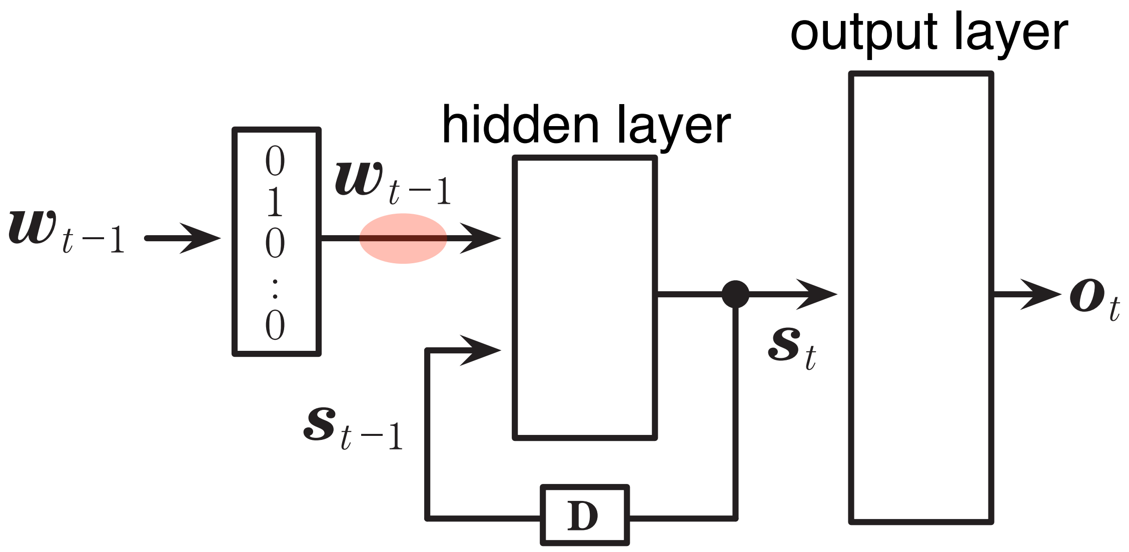
Recurrent NN-based LM

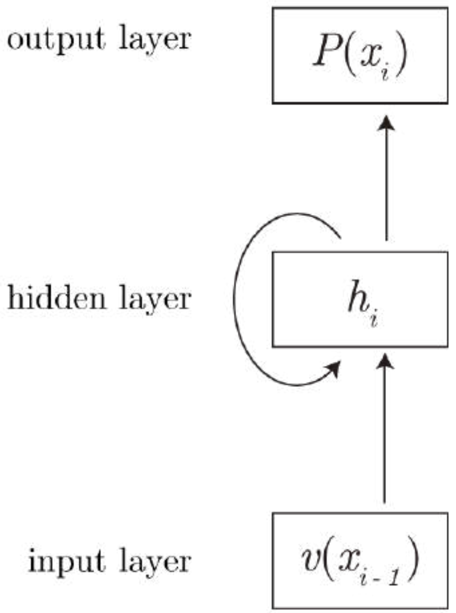
- v(x_t) = word features related to word at t
  P(x_t+1) = probability of word at t+1
  h = hidden layer

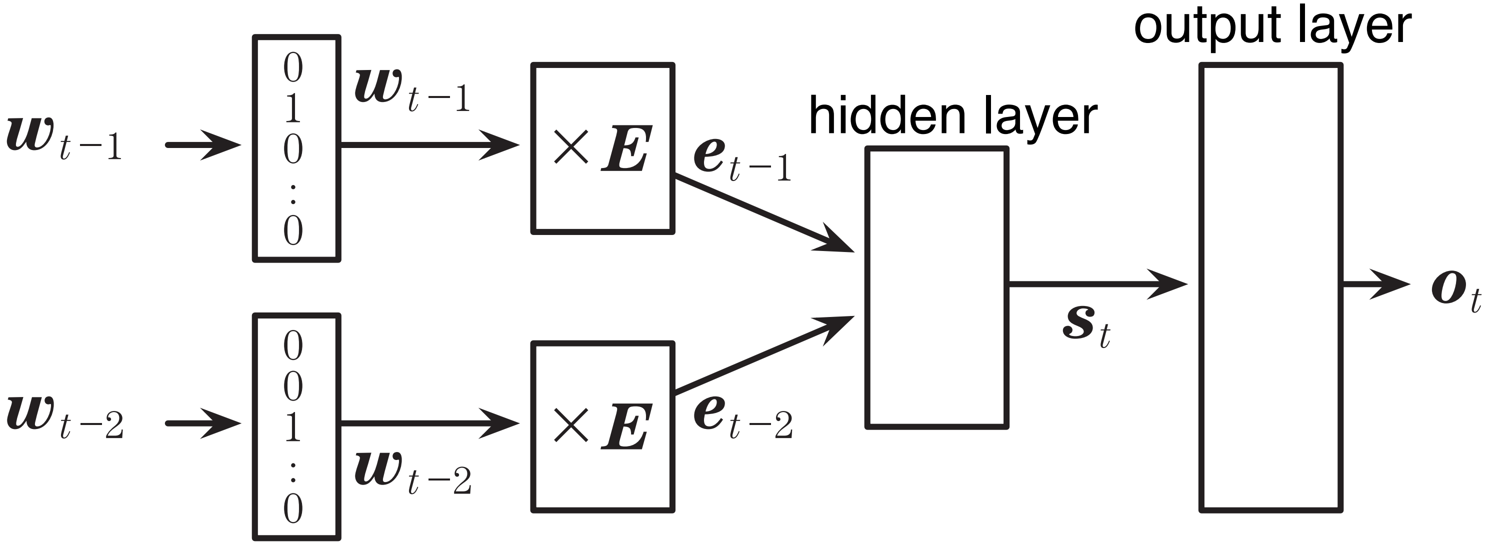




RNN-based LM and DNN-based LM

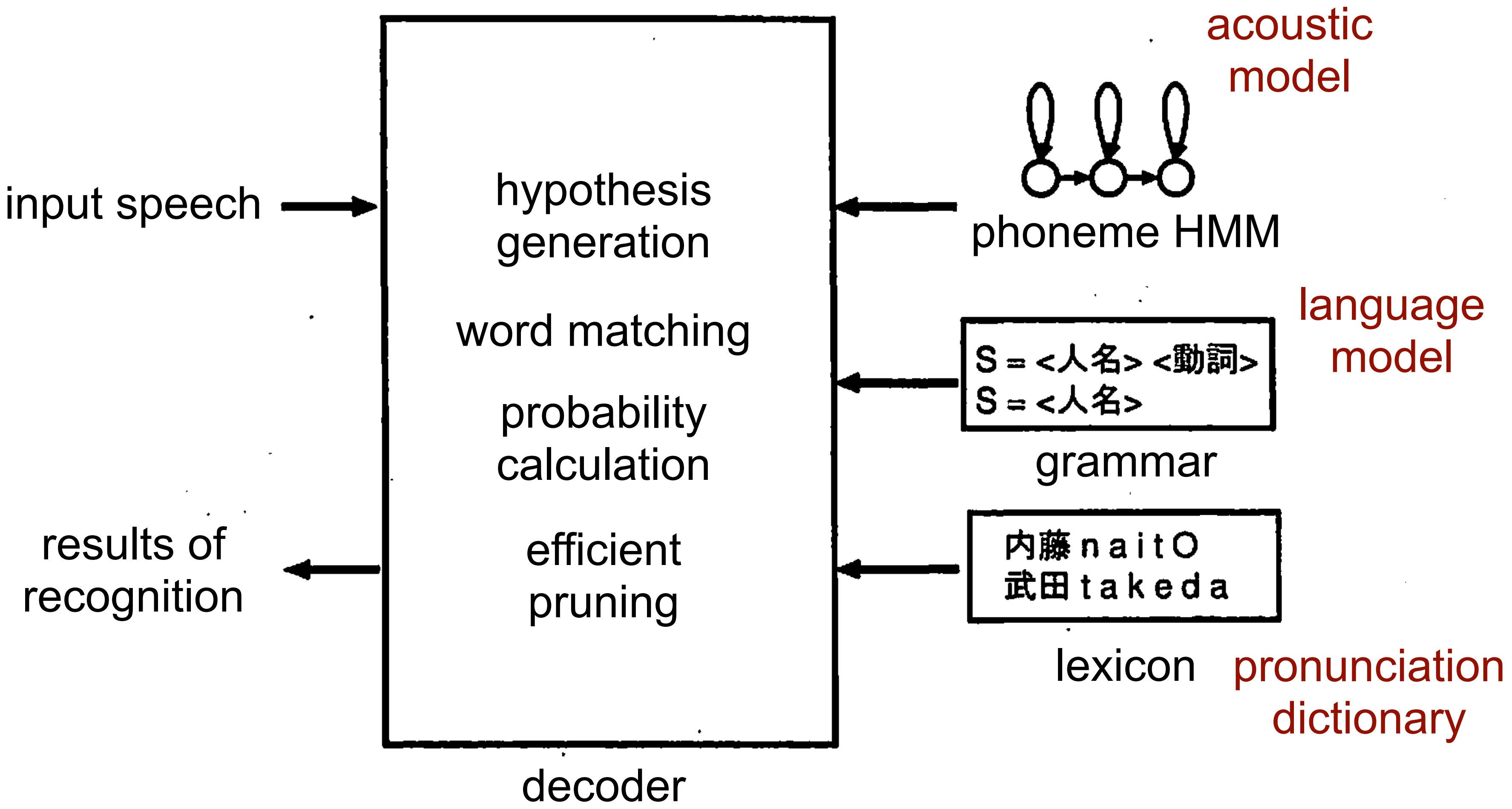






$$\underset{(\vec{e_1}, \vec{e_2}, \ldots, \vec{e_V})}{\boldsymbol{E}} \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \vec{e_2}$$

Development of a speech recognition system



input speech

hypothesis generation

word matching

probability calculation

efficient pruning

decoder

results of recognition

acoustic model

phoneme HMM

language model

S = <人名> <動詞>
S = <人名>

grammar

内藤 naito
武田 takeda

lexicon

pronunciation dictionary

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

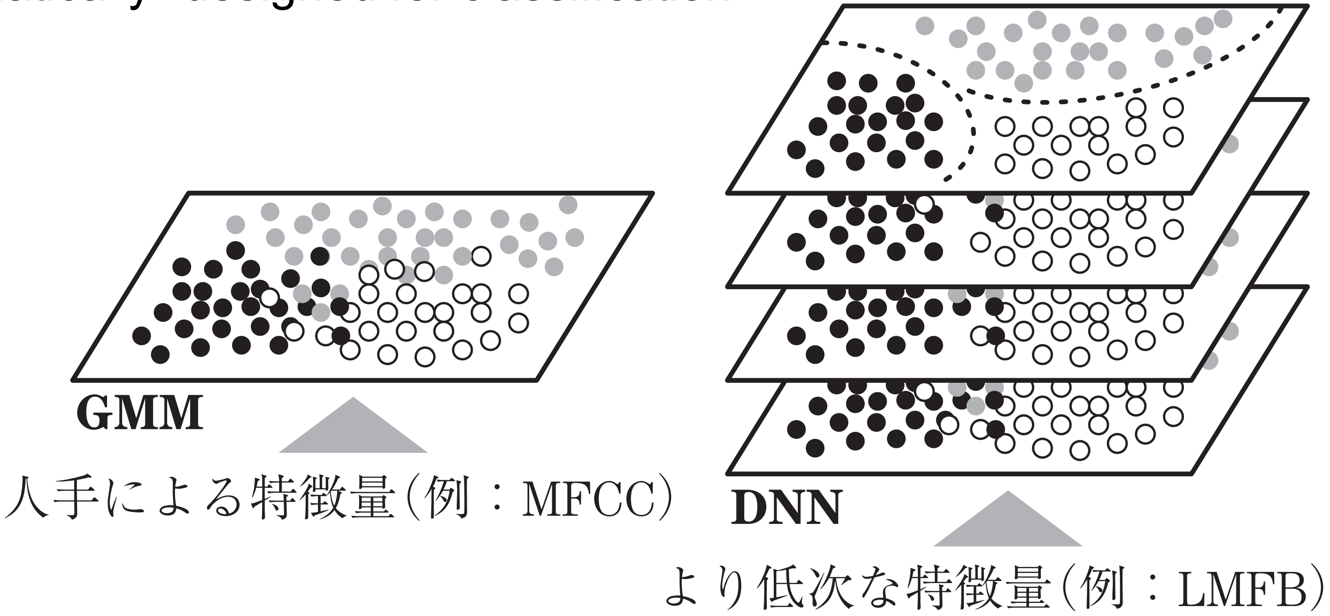
- Module-based ASR to one-package (E2E) ASR (next week)

One package (E2E) ASR

- Front-end of ASR (input)
  - A temporal sequence of samples (1-dimensional integer)
    - 16 bit integers with 16 / 44.1/ 48 [kHz] sampling
    - Temporal resolution is 1/16k, 1/44.1k, or 1/48k [sec]
  - A temporal sequence of feature vectors
    - Spectrum-based or cepstrum-based feature vectors
    - Temporal resolution is 10 [msec]
- Back-end of ASR (output)
  - A sequence of characters or phonemic symbols
    - "Cognitive Media Processing is given all in English."
    - k ɑ́ g n ə t ɪ v m íː d i ə p r ɑ́ s e s ɪ ŋ ɪ z g í v n ɔː l ɪ n í ŋ g l ɪ ʃ
    - Temporal resolution is ?? [msec] (1 syllable is about 150 [msec]).
- E2E ASR is a mapping problem between two sequences with different temporal resolution.
  - E2E ASR is trained only with a speech corpus, which has to be very large.
  - Generally speaking, however, |text corpus| >> |speech corpus|

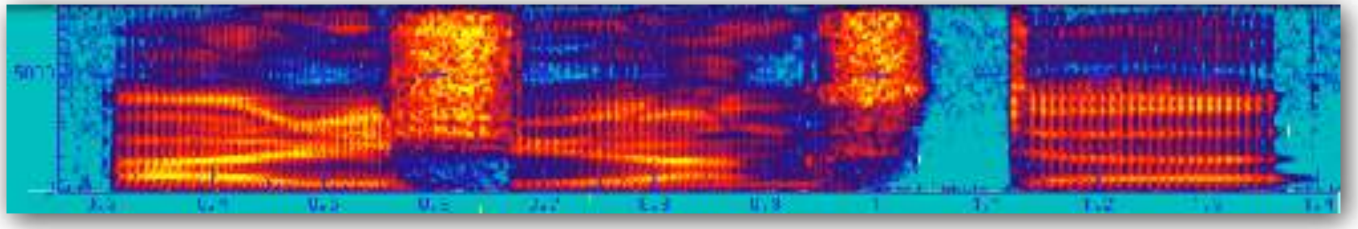
Why GMM-HMM < DNN-HMM?

- 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.

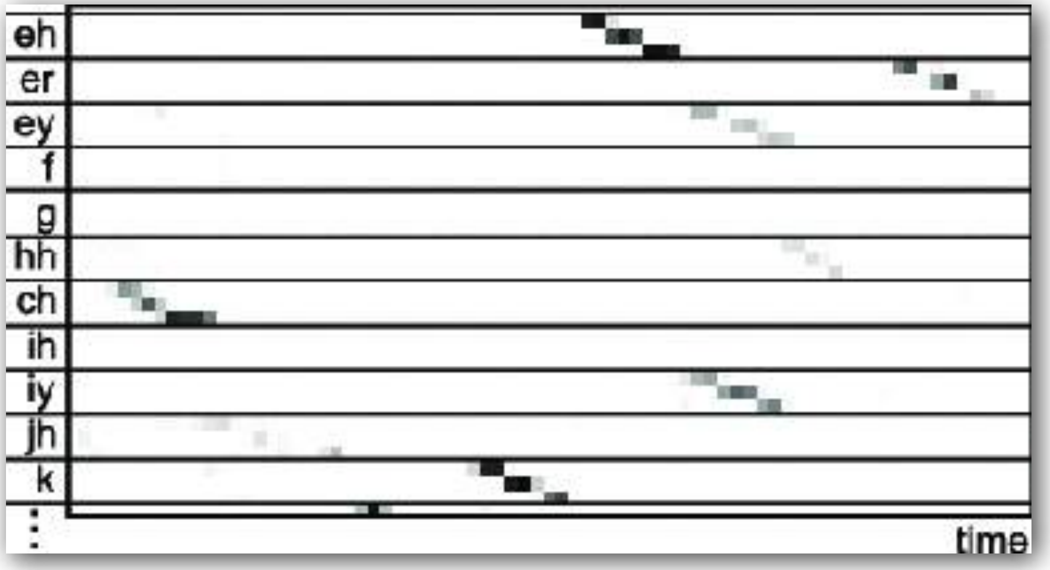


Direct use of DNN as phoneme classifier

- Phonemic (phonetic) posteriorgram (PPG)
  - Spectrogram is a temporal sequence of spectrum, which is an acoustic representation of speech at time t.



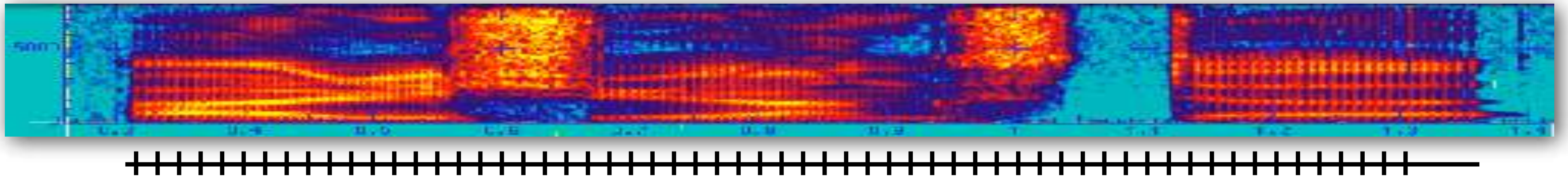
  - Posteriorgram is a temporal sequence of posterior probability distributions over the entire set of speech classes such as phonemes.



  - It is a linguistic or phonemic representation of speech.
  - Can be viewed as probabilistic (not deterministic) version of phonemic transcript
  - Difficult to realize alphabet-based posteriorgram

CTC (Connectionist Temporal Classifier)

- Adjustment of temporal resolution between source and target
  - Frame shift = 10 sec
  - Phoneme / alphabest = 30 - 150 msec
  - Introduction (insertion) of "blank" symbol (/) between consecutive symbols



aaaaaabbbbbbbccccaaaaa- - - -

aa/ /a/ / /bbb/ / / / /cc/ / /aa/ - - - -   --> abca

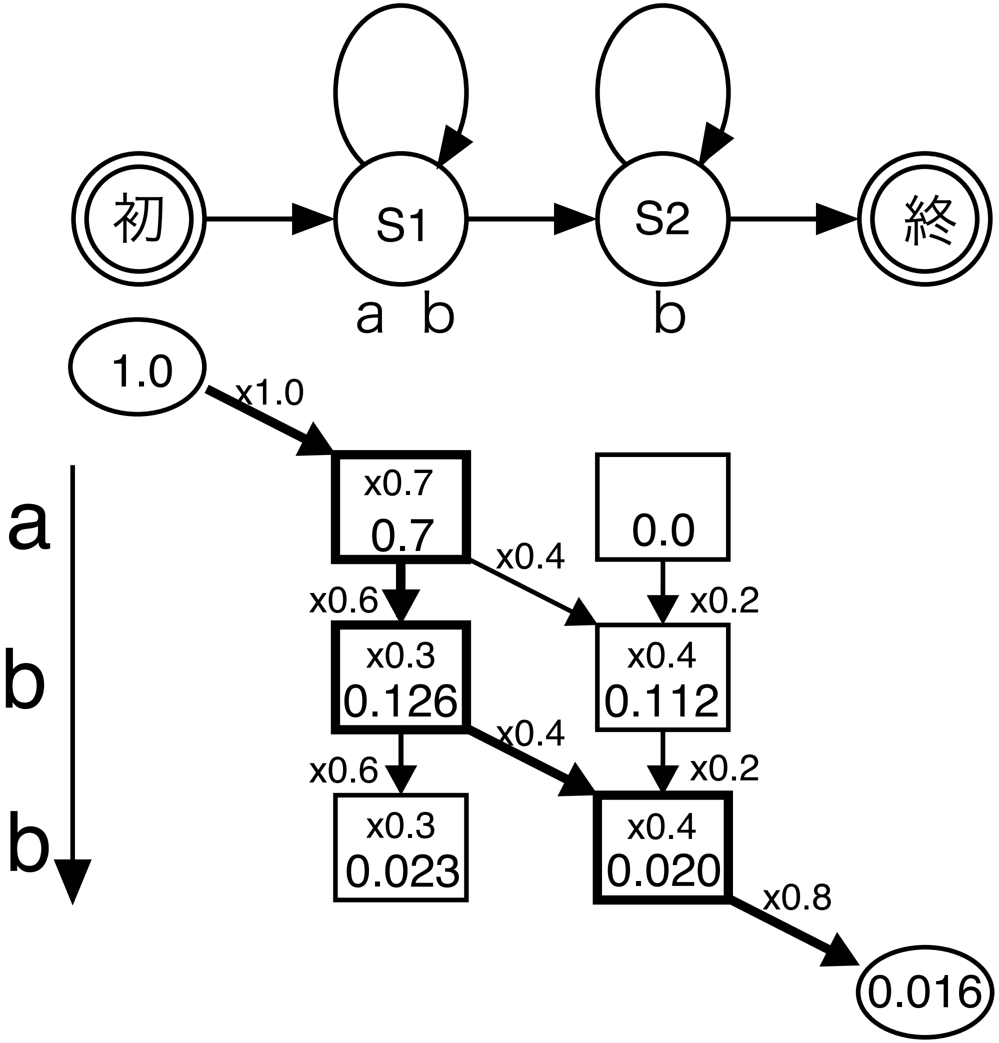
/a/ /aa/ /b/b/ /b/cc/ / /aa/ / - - - -   --> abca

aa/ /a/ / /bbb/ / / / /cc/ / /aa/ - - - -   --> abca

/ /a/ /a/b/ / /b/ /c/ / /c/ / /aa- - - -   --> abca

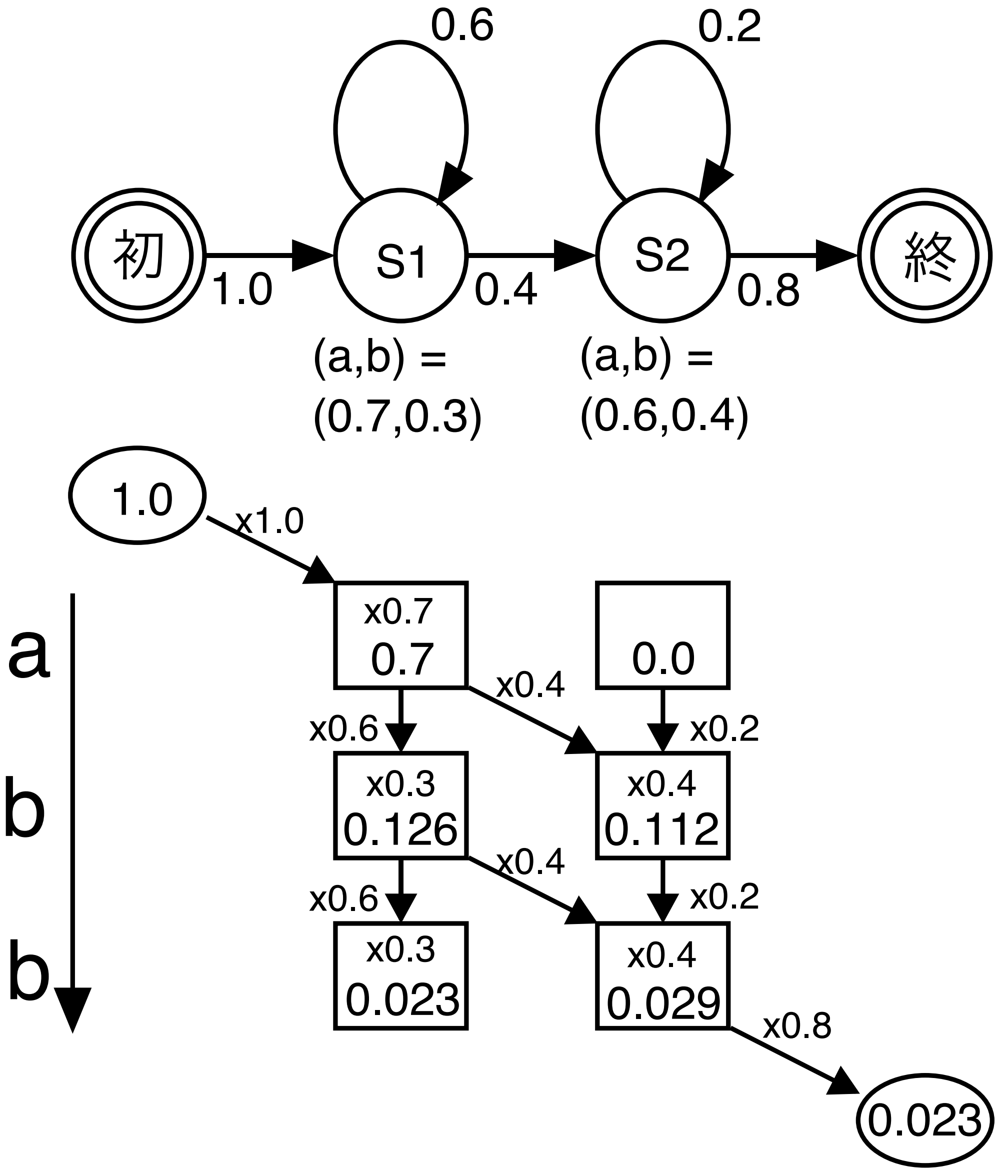
- Removal of "/" and merging symbol repetition make different symbol sequences correspond to a unique sequence of "abca".
- Each symbol except "/" may correspond to the core of the individual phonemes.

Output probability of observation sequence (Viterbi)



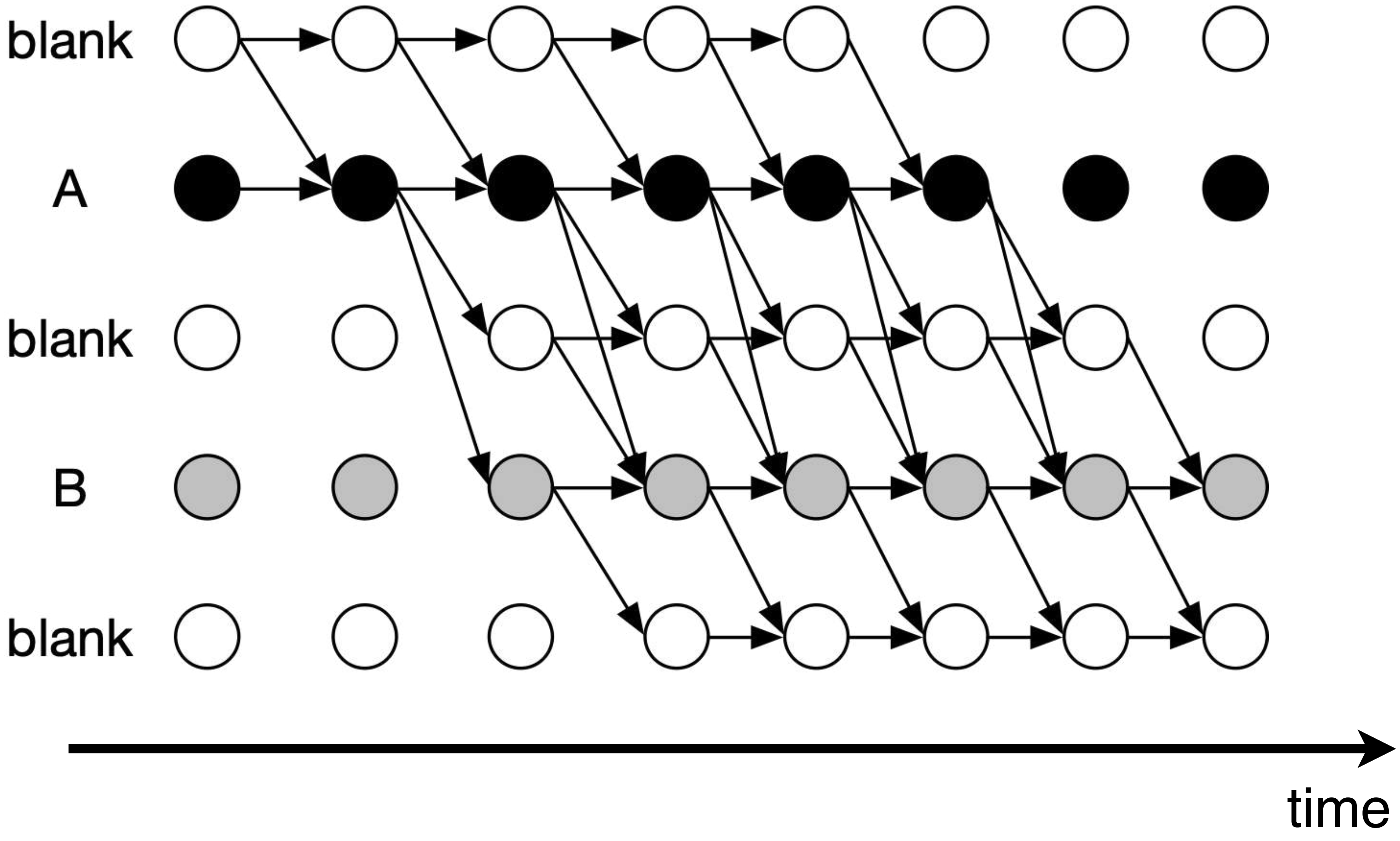
The maximum likelihood path is only adopted.

Output probability of observation sequence (Trellis)



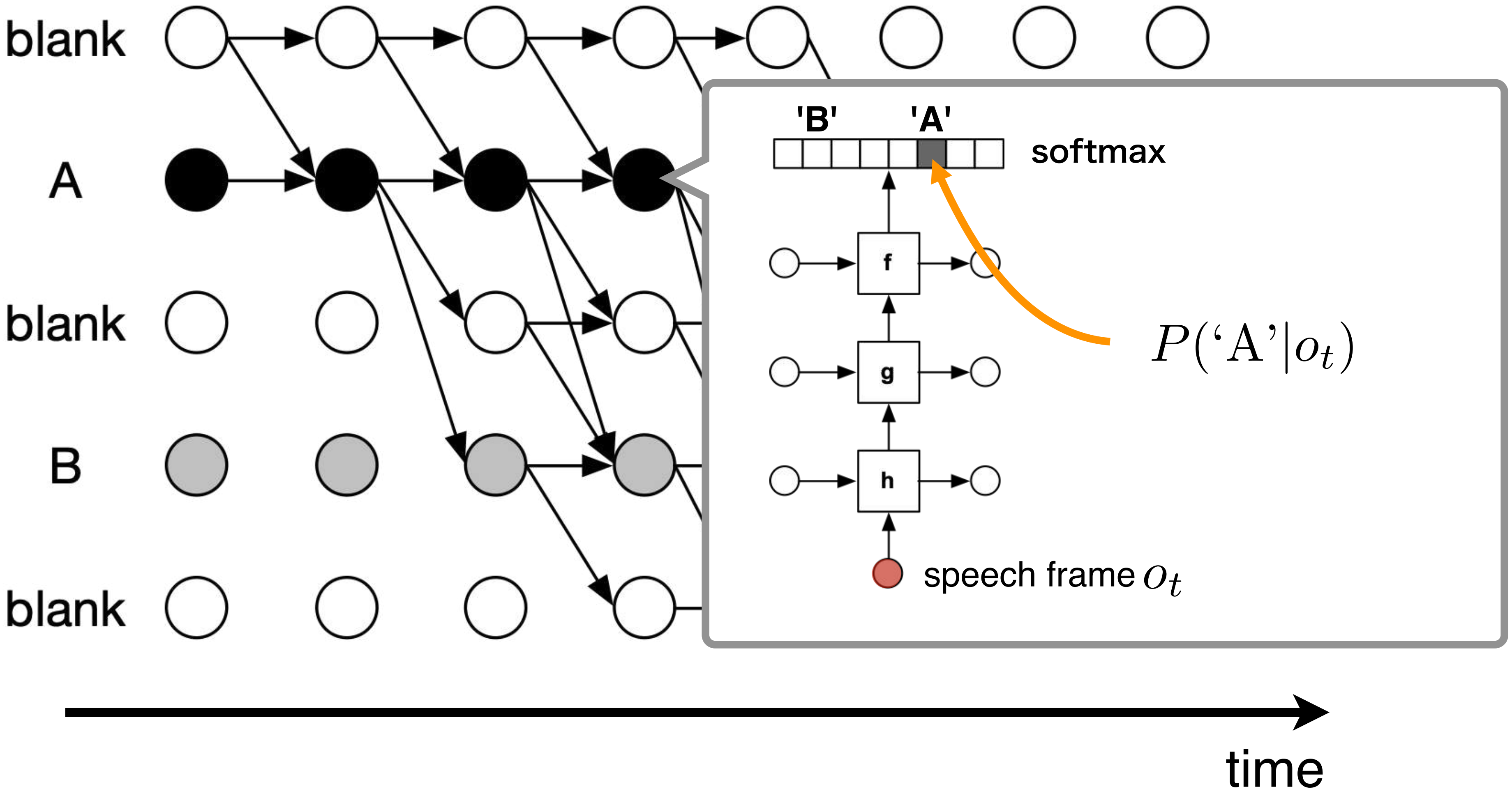
CTC (Connectionist Temporal Classifier)

- Multiple paths (trellis) corresponding to a single string



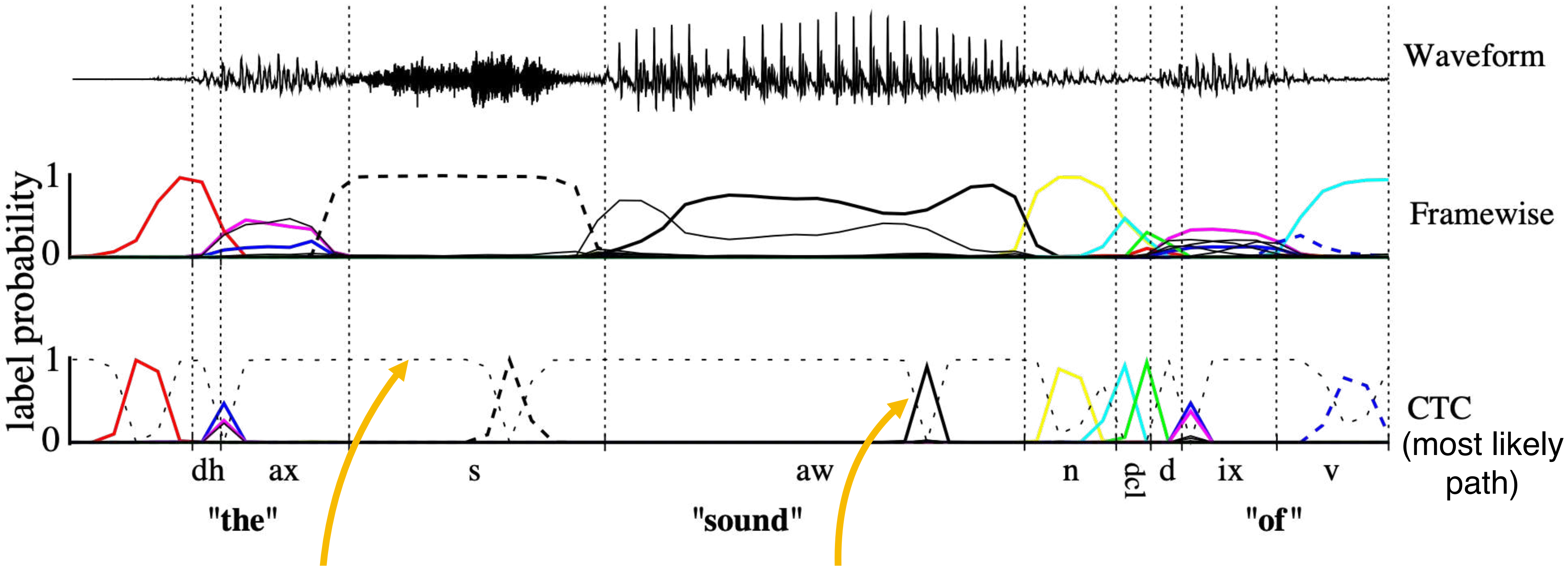
CTC (Connectionist Temporal Classifier)

- Multiple paths (trellis) corresponding to a single string



CTC (Connectionist Temporal Classifier)

- Frame-based phoneme classifier vs. phoneme-based CTC



The blank symbol is observed very frequently.

Non-blank symbols are often observed at once in a phoneme segment.

- CTC can be applied to alphabets, not phonemes.

Today’s menu

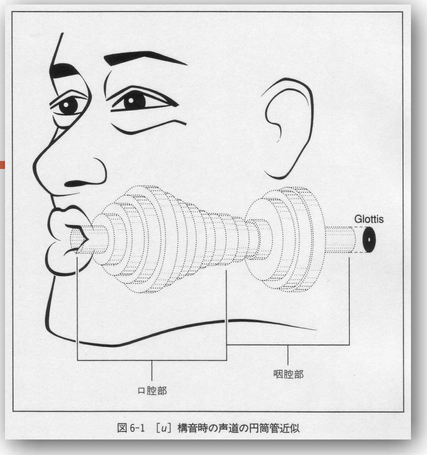
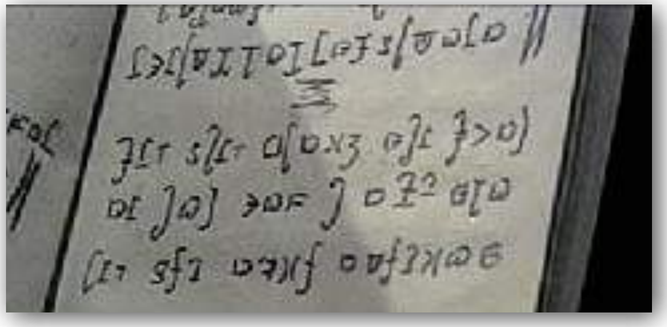
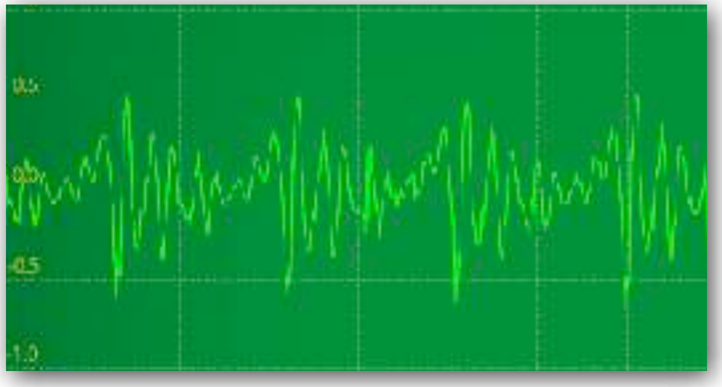


図 6-1 ［*u*］構音時の声道の円筒管近似

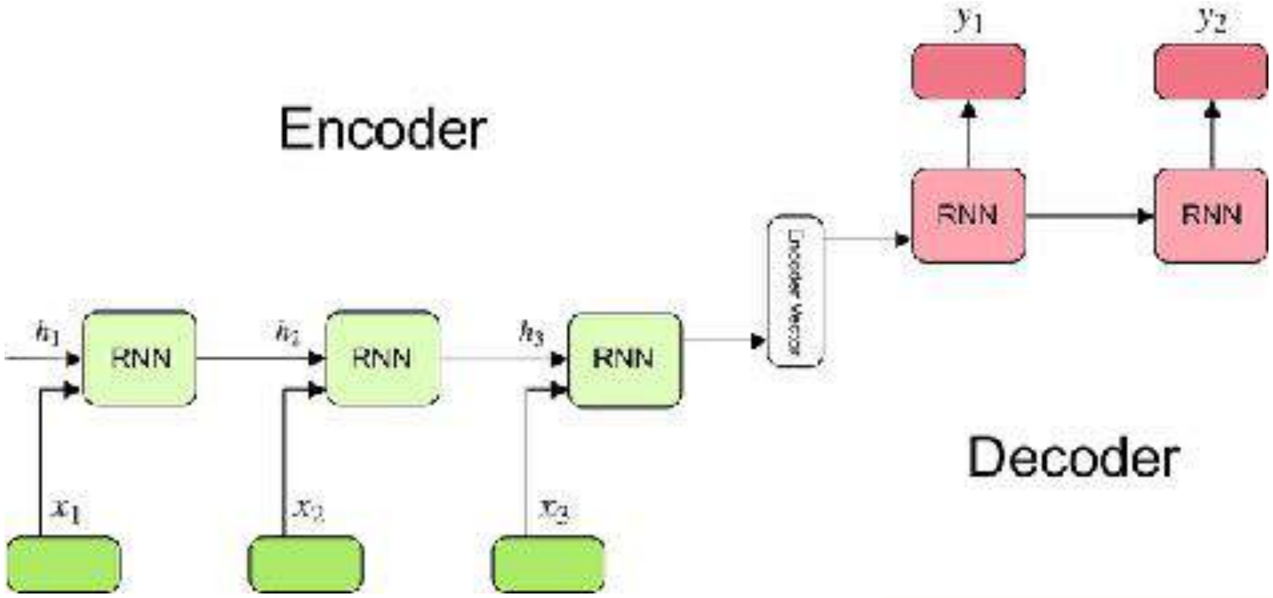
- Speech --> sounds --> vibrations (waves) of air particles
- Fundamentals of phonetics
  - How are vowel sounds produced?
  - Phonetics = articulatory phonetics + acoustic phon. + auditory phon.
- More on articulatory phonetics
  - Observation of speech organs
- More on general phonetics
  - General phonetics = language independent phonetics
  - How to symbolize language sounds found in any language?
- More on acoustic phonetics
  - Vowels as standing waves
    - Resonance frequency = formant frequency
  - Link between acoustic phon. and articulatory phon.
- Summary





Encoder-decoder model

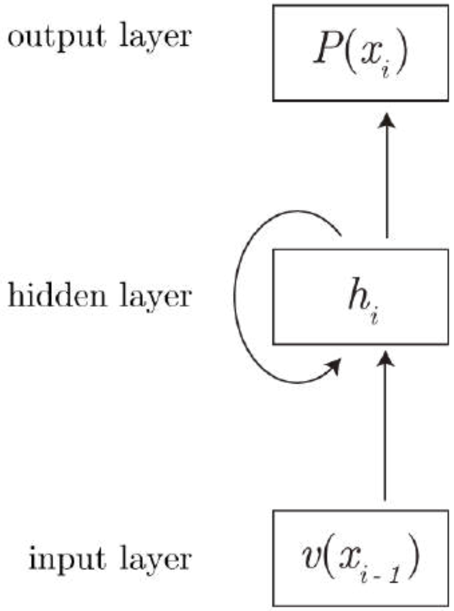
- Input sequence is converted to output sequence.
  - Encoder can be interpreted as extractor of abstract representation from input signals or tokens.
  - Decoder can be interpreted as embodiment of the abstract representation into actually observed output signals or tokens.
  - An example of the encoder-decoder model using Recurrent Neural Network (RNN).
    - Input = x1, x2, x3, ...   Output = y1, y2, y3, ...
    - Long-distance relations are *actually* difficult to be encoded or decoded.

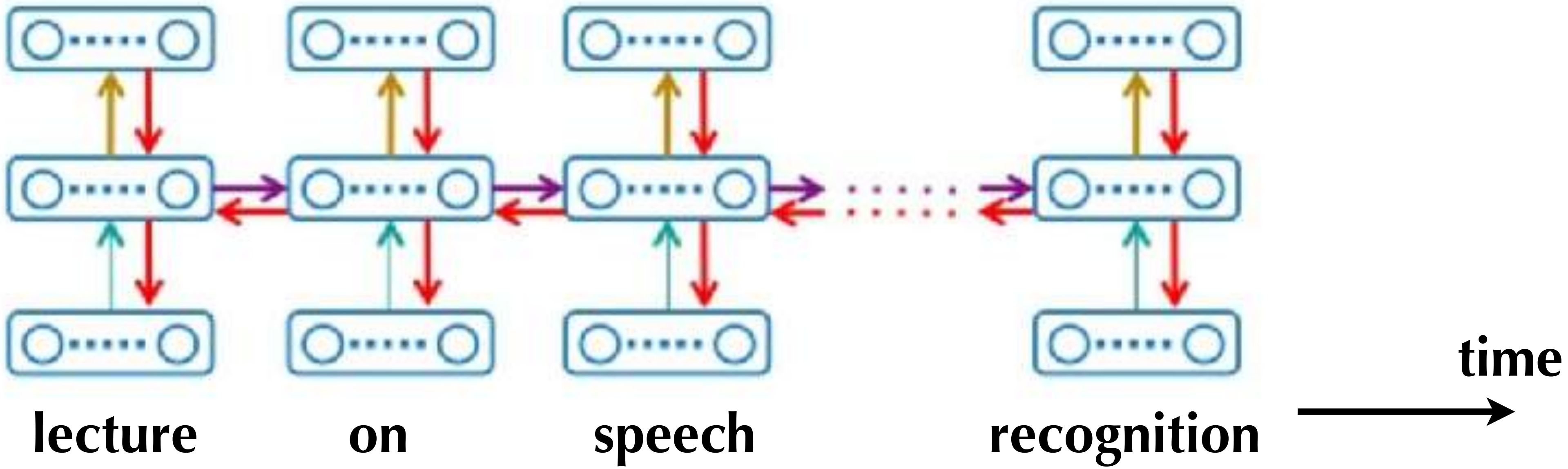


Source: https://arxiv.org/abs/2102.03218

Recurrent NN-based LM

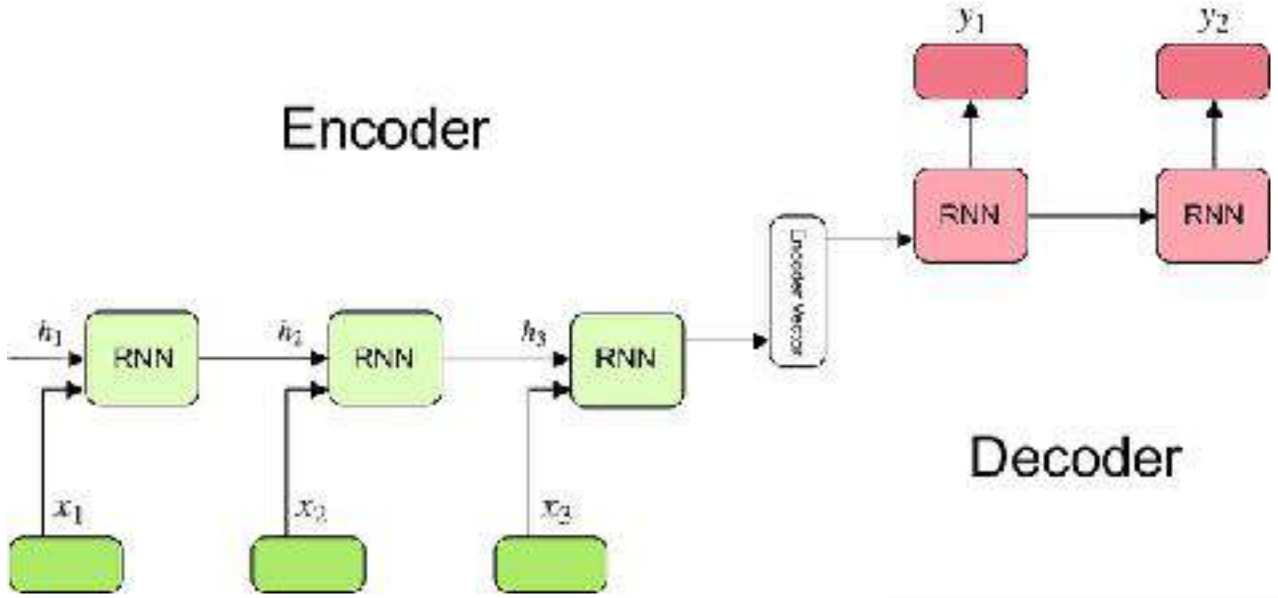
- v(x) = word features of related to word at t
  P(x) = probability of word at t+1
  h = hidden layer





Encoder-decoder model

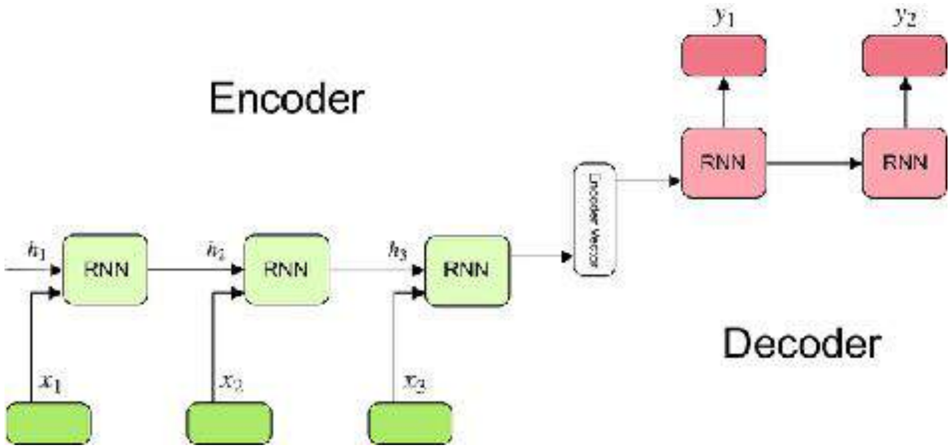
- Input sequence is converted to output sequence.
  - Encoder can be interpreted as extractor of abstract representation from input signals or tokens.
  - Decoder can be interpreted as embodiment of the abstract representation into actually observed output signals or tokens.
  - An example of the encoder-decoder model using Recurrent Neural Network (RNN).
    - Input = x1, x2, x3, ...   Output = y1, y2, y3, ...
    - Long-distance relations are *actually* difficult to be encoded or decoded.

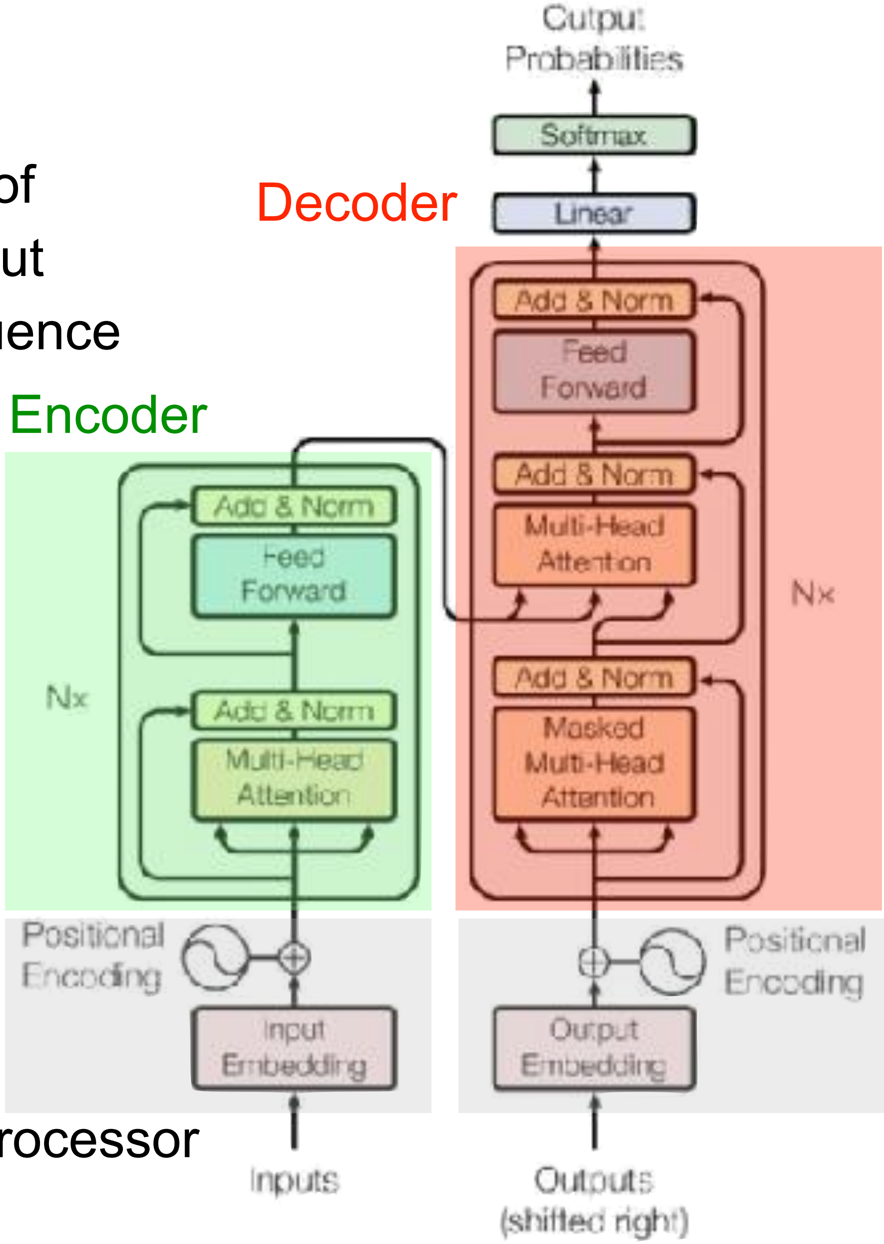


Source: https://arxiv.org/abs/2102.03218

Transformer model

- Attention is all you need !!
  - https://arxiv.org/abs/1706.03762
  - Explicit modeling of the relations (similarities) of the current input token to other ones in the input sequence and to the tokens in the output sequence generated so far.
    - Self-attention mechanism





Decoder

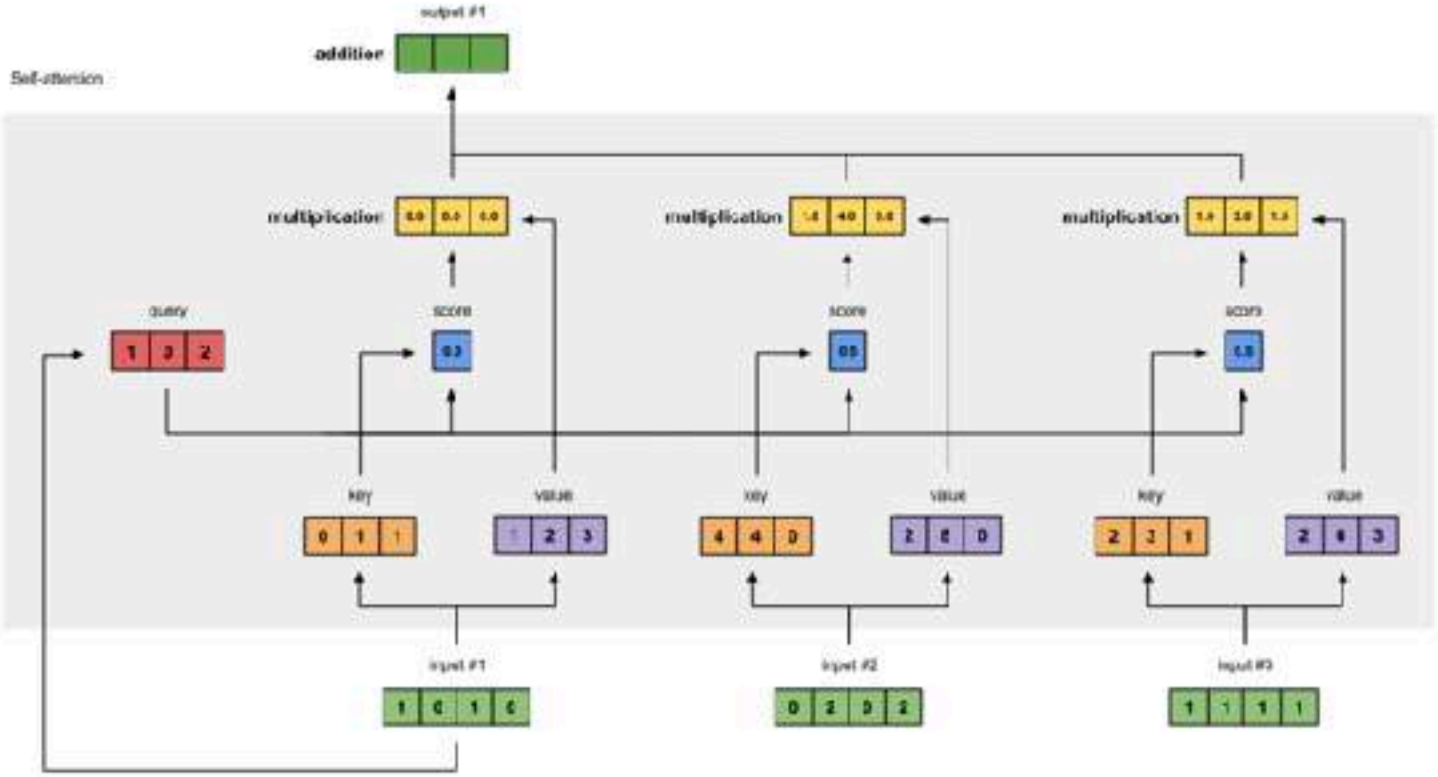
Encoder

Preprocessor

Self-attention mechanism

- Relatedness of the current input token
  - to the other tokens in the input sequence and
  - to the tokens in the output sequence generated so far.
- A token is converted to its three components.
  - Value vector, key vector, and query vector.

$$\underset{(\vec{e_1}, \vec{e_2}, \ldots, \vec{e_V})}{\boldsymbol{E}} \begin{pmatrix} 0 \\ 1 \\ 0 \\ : \\ 0 \end{pmatrix} = \vec{e_2}$$



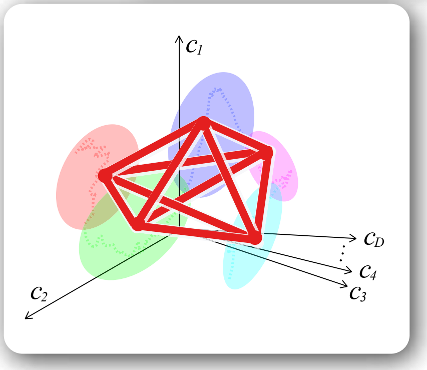
https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a

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 synthesis -

Nobuaki Minematsu

Today’s menu

- Overview of text-to-speech conversion
  - From speaking machine to reading machine
- Text analysis
  - Text processing using units of sentences, phrases, and words
- Reading analysis
  - Assignment of reading (phonetic symbol + prosody) to each phoneme
- Waveform generation
  - Conversion of phonetic symbols + prosody to acoustic waveforms
- Some demos
  - Unit selection synthesis + HMM-based synthesis

Today's menu

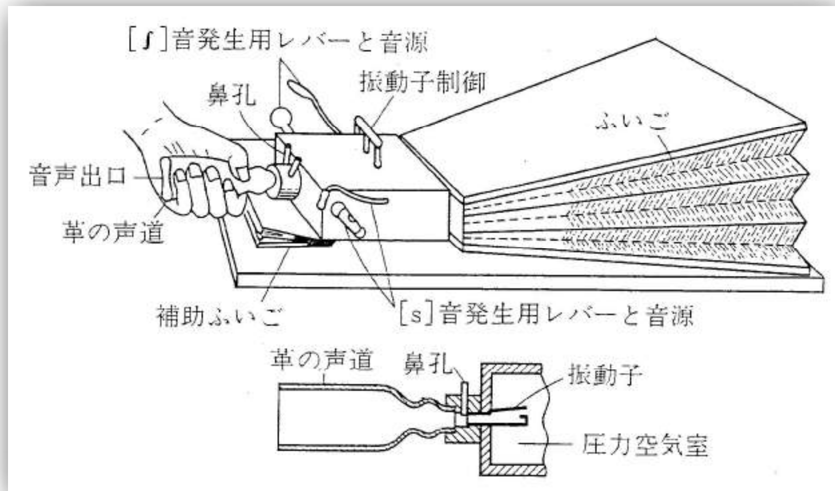
- Overview of text-to-speech conversion
  - From speaking machine to reading machine
- Text analysis
  - Text processing using units of sentences, phrases, and words
- Reading analysis
  - Assignment of reading (phonetic symbol + prosody) to each phoneme

- Application of text and reading analysis for TTS to Japanese education
  - How to read Japanese text with good and natural prosody?

- Waveform generation
  - Conversion of phonetic symbols + prosody to acoustic waveforms
- Some demos
  - Unit selection synthesis + HMM-based synthesis

The world oldest speech generator

- Speaking machine (Kempelen 1791)

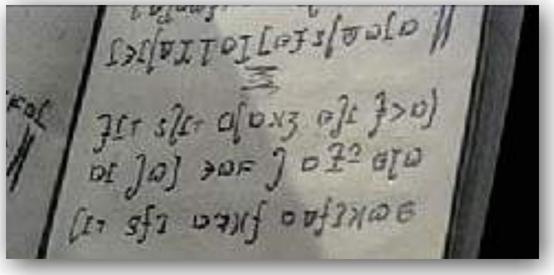


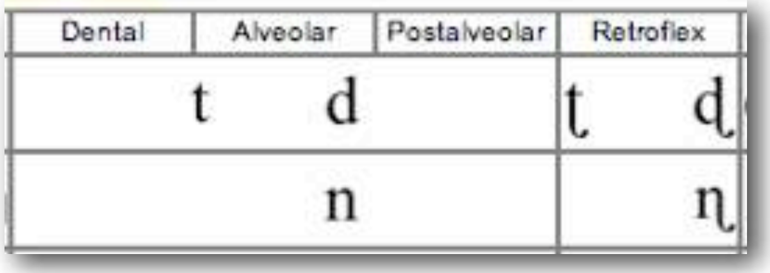
Difficulty of TTS

- Raw text alone is not sufficient “phonetically” for it to be read.
  - How to read Kanji? How to convert Kanji to Hiragana?
    - 今日の午後は，生物学の授業に出ました。
    - 今日＝きょう？こんにち？　　生物＝せいぶつ？なまもの？
  - Hiragana is similar to phonemic representation. It is enough for TTS?
    - とんぼ，とんねる，どんぐり
    - すいか，たべますか？
  - If text is represented by phonetic symbols, is it enough?
    - How about prosodic features which are needed for text-to-speech conversion?
    - Intonation, word accent, durational control (speaking rate), etc.
    - 赤（あか）＋えんぴつ　→　あかえんぴつ
  - Only “read”-style speech? Only native speech?
    - Expressive (emotional) speech
    - Non-native speech

Phones and phonemes

- Phones
  - A phone is the minimal unit of speech of any language.
  - Phonetic symbols are language-independent and used by phoneticians to transcribe speech of any language. Defined by by Int. Phonetic Association.
  - Should be used like [a b c d e f g].



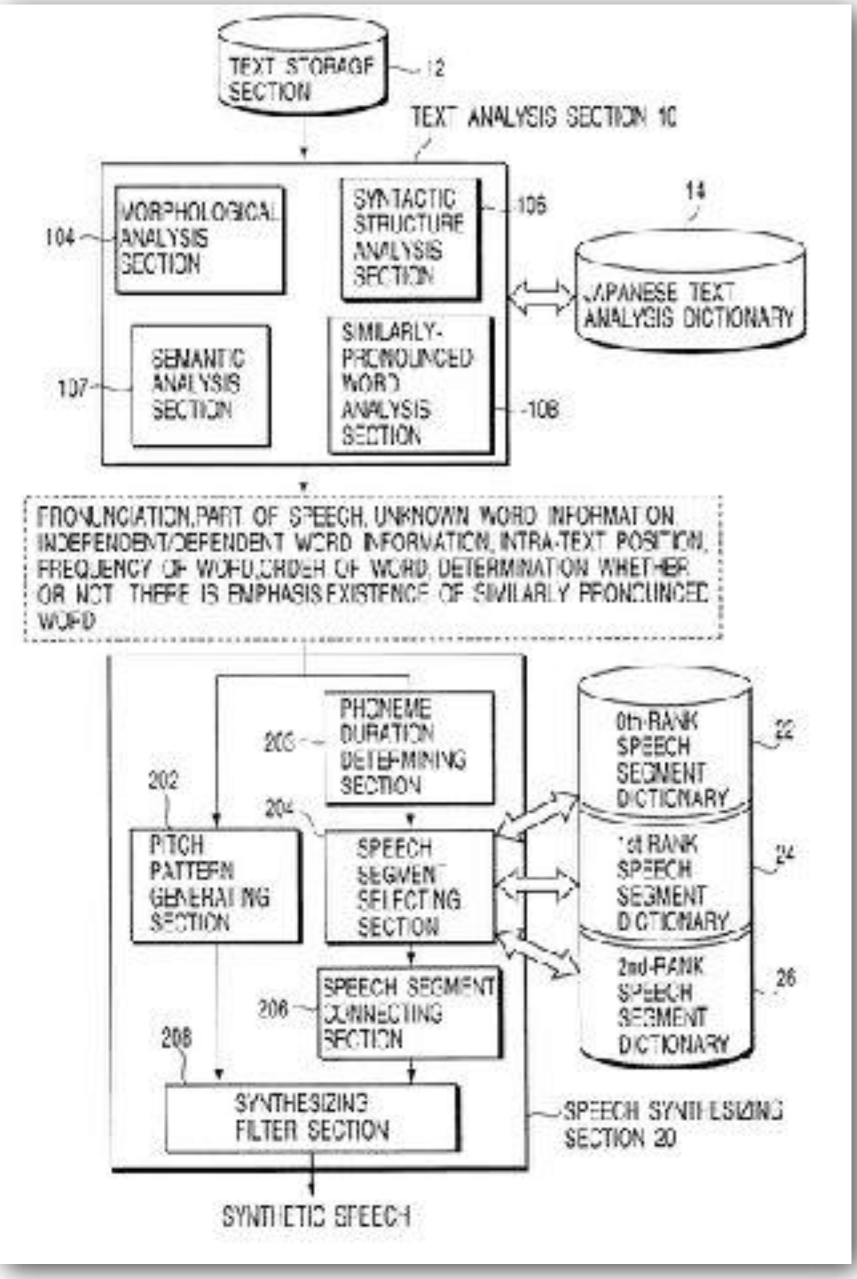


Dental	Alveolar	Postalveolar	Retroflex
t	d		ʈ ɖ
	n		ɳ

- Phonemes
  - A phoneme is the minimal unit of speech of a specific language, perceived by native speakers of that language.
  - Phonemic symbols are language-dependent and used by ordinary people to transcribe speech of that language. Can be defined by a user.
  - Should be used like / a b c d e f g /.

/arajurugeNzituo/ → [ɐɾɐj̞ɨɾɨ̠ge̞ẽ̞ɲ̟d̠͡ʑit͡sɨo̜ʔ]

Overview of text-to-speech conversion


TEXT STORAGE SECTION
12
TEXT ANALYSIS SECTION 10
104
MORPHOLOGICAL ANALYSIS SECTION
SYNTACTIC STRUCTURE ANALYSIS SECTION
106
14
JAPANESE TEXT ANALYSIS DICTIONARY
107
SEMANTIC ANALYSIS SECTION
SIMILARLY-PRONOUNCED WORD ANALYSIS SECTION
108
PRONUNCIATION, PART OF SPEECH, UNKNOWN WORD INFORMATION, INDEPENDENT/DEPENDENT WORD INFORMATION, INTRA-TEXT POSITION, FREQUENCY OF WORD, ORDER OF WORD, DETERMINATION WHETHER OR NOT THERE IS EMPHASIS, EXISTENCE OF SIMILARLY PRONOUNCED WORD
205
PHONEME DURATION DETERMINING SECTION
202
204
PITCH PATTERN GENERATING SECTION
SPEECH SEGMENT SELECTING SECTION
0th-RANK SPEECH SEGMENT DICTIONARY
22
1st-RANK SPEECH SEGMENT DICTIONARY
24
2nd-RANK SPEECH SEGMENT DICTIONARY
26
206
SPEECH SEGMENT CONNECTING SECTION
208
SYNTHESIZING FILTER SECTION
SPEECH SYNTHESIZING SECTION 20
SYNTHETIC SPEECH

Conversion of any text input to its reading

今日の午後は，生物学の授業に出ました。

Conversion of reading to waveforms

今日＝キョー？コンニチ？

生物＝セイブツ？ナマモノ？

キョーノゴゴワ／セイブツガクノジュギョーニ／デマシタ

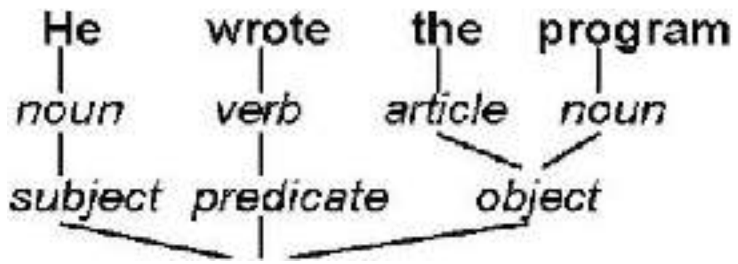
は＝ハ？ワ？

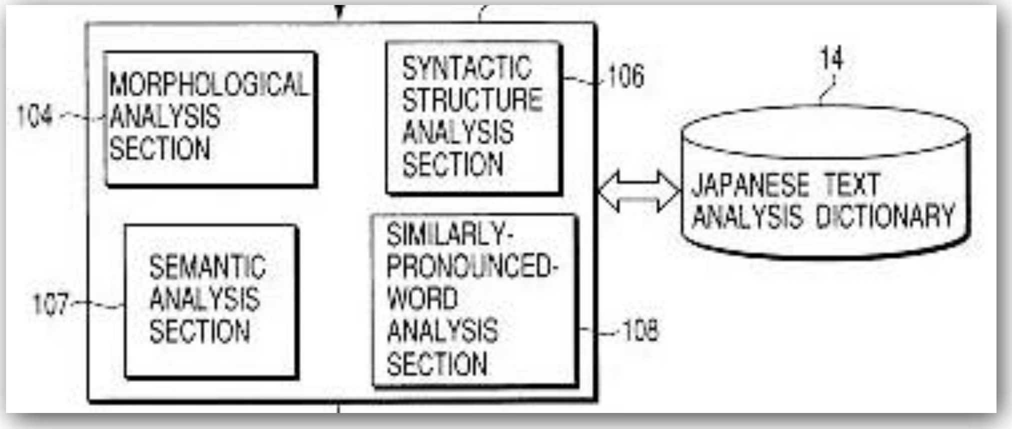
セイブツ＋ガク　→　セイブツガク

Vowel in シ of デマシタ＝voiced or unvoiced？

Text-to-speech synthesis

- Conversion from any input text to its sound sequence
  - In Japanese, Kanji have multiple ways of reading.
  - Kanji, Hiragana, Katakana, and Romaji
- Text analysis
  - Morphological analysis
    - An input sentence is divided into words (morphemes).
    - Part of speech (品詞) is assigned to each word.
  - Syntactic analysis
    - Syntactic structure in a sentence is extracted.
  - Semantic analysis
    - In many cases, it refers to correlation and co-occurrence among words.
- Reading analysis
  - Phonemic aspect
  - Prosodic aspect
- Waveform generation




MORPHOLOGICAL ANALYSIS SECTION
104
SYNTACTIC STRUCTURE ANALYSIS SECTION
106
14
JAPANESE TEXT ANALYSIS DICTIONARY
SEMANTIC ANALYSIS SECTION
107
SIMILARLY-PRONOUNCED-WORD ANALYSIS SECTION
108

Reading analysis

- Phonemic aspect (segmental aspect of speech)
  - Text to phonemic (Hiragana) representation
    - 今日の午後は → ky o: n o g o g o w a （きょうのごごは）
  - Phonemic representation to phonetic representation
    - ん(N) → n, m, ng ?
    - Some vowels have to be unvoiced.
    - :
- Prosodic aspect (supra-segmental aspect of speech)
  - Word-level processing
    - Word accent, accent sandhi in compound words
  - Phrase-level processing
    - Accent sandhi in connected words of a phrase
  - Sentence-level processing
    - Emphasis, phrasing

Raw text to Hiragana (phonemes)

- Two different reading styles of Japanese
  - On(音/オン)-reading and Kun(訓/くん)-reading
    - 生きる（い），生える（は），生物（セイ，なま），
- は and へ
  - は in 私は：“w a”, not “h a”
  - へ in 大学へ：“e” not “h e”
- Lengthened vowels
  - 消耗（しょうもう）：sh o: m o:
  - 映画（えいが）：e: g a
  - 大阪（おおさか）：o: sa ka
- Geminate consonant sounds (especially in numerical expressions)
  - 一巻（いっかん，ikkan）k becomes long (long consonant).
- Euphonic change of an unvoiced consonant to its voiced version (連濁)
  - 江戸川（えど＋かわ→えどがわ）

Numerical expressions

- Two different ways of reading numerical expressions
  - 03-5841-6662 : digit by digit
  - 123,456 yen : use of places, e.g. 12万3千4百5十6円
- Sound changes that are unique to numerical expressions
  - 523 is not ご　に　さん but ごおにいさん
- Geminate consonant sounds
  - 一本：いち＋ほん→いっぽん
  - １cm：いち＋せんち→いっせんち
- Euphonic change of unvoiced consonant to its voiced version (連濁)
  - 三本：さん＋ほん→さんぼん
  - 三階：さん＋かい→さんがい
  - 三回：さん＋かい→さんかい
- Exceptional cases
  - 一日：ついたち（いちにち is OK）
  - 二日：ふつか（ににち is OK）

Phoneme symbols to phonetic symbols

- Some vowels become unvoiced
  - If a vowel is surrounded by unvoiced consonants, it often become unvoiced.
    - アシカ（a sh i k a），エンピツ（e N p i ts u），スキヤキ（s u k i y a k i）etc
- Nasalized consonants
  - 株式会社（かぶしきがいしゃ，k a b u sh i k i g a i sh a)
- Syllabic nasal (撥音，ん)
  - 粘板岩（ねんばんがん）
    - [m] before p, b, and m
    - [ng] before k, g, and ng
    - [n] before t, d, n, and pause

Non-linguistic symbols and short forms

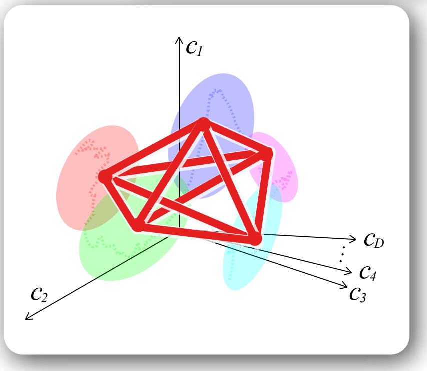
- Have to be converted into their phoneme sequences
  - %, kg, @
  - HMM, IT, IEEE
- Unknown words
  - Person’s names, city names (proper nouns)
  - Their reading have to be predicted using linguistic knowledge.
    - Grapheme-to-phoneme conversion

JEITA format

- Japan Electronics and Information Technology Industries Association
  - Text format designed specially for TTS input.
  - Proposed as standard and recommended by JEITA
  - Examples
    - 2006年の調査によると，日本全国で，約33%の家庭が，ペットを飼っているそうです。
    - ニセンロクネンノチョウサニヨルト，ニホンゼンコクデ，ヤクサンジュウサンパーセントノカテイガ，ペットヲカッテイルソウデス。
    - **ニセ'ンロクネンノ/チョ'ーサニヨルト_ニホ’ン_ゼ’ンコクデ_ヤ’ク/サ’ンジューサンパーセントノカテーガ_ペ’ットオ/カ’ッテイルソーデス%.**
    - ペットを家族のように考える人が増えたため，ペット関連の新しいビジネスも，生まれました。
    - ペットヲカゾクノヨウニカンガエルヒトガフエタタメ，ペットカンレンノアタラシイビジネスモ，ウマレマシタ。
    - **ペ'ットオ/カ'ゾクノヨーニカンガエルヒ%トガ_フ’エタタメ_ペットカ'ンレンノ_アタラシ'ー/ビ'ジネスモウマレマシ%タ.**

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



Assignment

- Assignment
  - Read a research paper related to the second four lectures of this class.
  - **Submit two PDF files: 1) the paper and 2) summarization of the paper and your comments on the paper**
  - All the materials used in the lectures are available at:
    - https://www.gavo.t.u-tokyo.ac.jp/~mine/japanese/CMP/class.html
- Length
  - Two or more pages of A4 size for 2)
- Submission
  - Your report should be submitted via. ITC-LMS.
  - **The filenames must be in the following format.**
    - **[student_id]_paper.pdf and [student_id]_[name].pdf**
    - **For example,**
      36-302439_paper.pdf (paper)
      36-302439_nobuaki-minematsu.pdf (summary and comments)
- Deadline = 23:59:59 on Dec 26.
  - You have two weeks to go.