From Isolated to Continuous Speech Recognition

陳嘉平

國立中山大學資工系

Statistical Automatic Speech Recognition

- 語音特性參數 (Speech Features)

  short-time signal processing on speech signals (*not* the focus today)

- 語音辨識模型 (Speech Recognition Models)
  - A recognizer does not know *what* is going to be said and *how* it is going to be said.
  - *What*: language models
  - *How*: acoustic models

System Training and Decoding

- 模型參數估算

$$\text{Maximum-likelihood: } \Theta^* = \arg\max_{\Theta} P(D|M,\Theta) \tag{1}$$
$$\Theta = \text{parameters},\ M = \text{model},\ D = \text{data}.$$

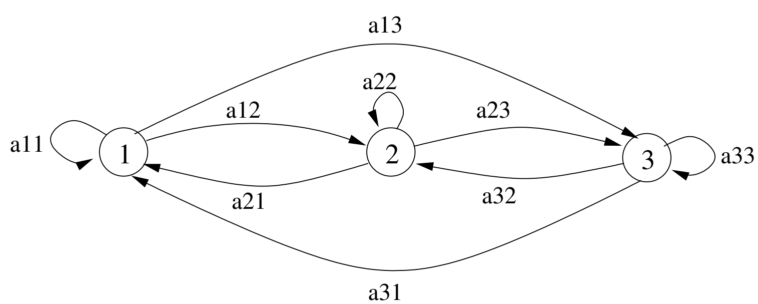
- 未知語音解碼

$$\begin{aligned}\text{Maximum A Posteriori: } S^* &= \arg\max_S P(S|X) \\ &= \arg\max_S \frac{P(S,X)}{P(X)} \\ &= \arg\max_S P(S,X) \\ &= \arg\max_S P(X|S)P(S)\end{aligned} \tag{2}$$

$P(X|S)$ = acoustic model score, $P(S)$ = language model.

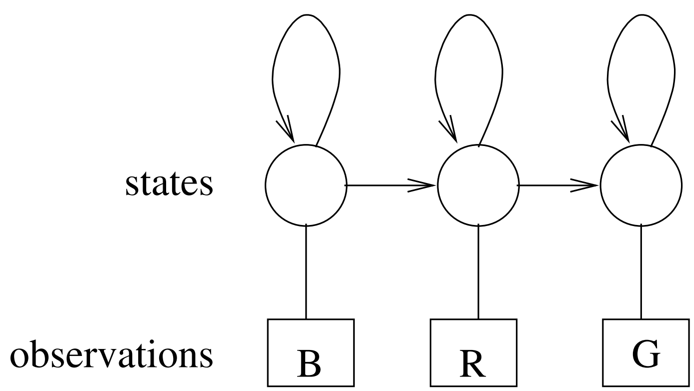
A Discrete Markov Model

- state space (狀態空間)
- initial probability (初始機率)
- transition probability matrix (轉移機率)
- joint probability (聯合機率)
- example: weather



Hidden Markov Model (HMM)

- observation (觀測值)
- emission probability (放射機率)
- joint probability
- example: urns (states) and balls (observations)



HMM Acoustic Models

- discrete-time = frame
- state = “(sub-)phone-like unit” and observation = features
- model parameters = initial probability + transition probability matrix + emission probability
- EM algorithm for parameter learning
- forward-backward algorithm for exact data likelihood computation
- Viterbi algorithm for optimal state sequence search

Isolated Speech Recognition

- Since each unknown utterance consists of one single word,

$$w^* = \arg\max_{w \in V} P(X|w)P(w), \tag{3}$$

  where V is the vocabulary.

- It is feasible to exhaustively compute the data-likelihood of each model if $|V|$ is small.

- The model data-likelihoods can be computed exactly via F/B algorithm or approximately via Viterbi algorithm.

Continuous Speech Recognition

- How many different sentences are there? Infinite!

  (give me a sentence and I can make a longer one.)

- What are the probabilities of these sentences?

  They must satisfy

$$\begin{cases} P(S) \geq 0, \\ \sum_S P(S) = 1 \end{cases} \tag{4}$$

  We use

$$P(S = w_1 \dots w_n) = P(w_1 w_2 \dots w_n) P(\text{eos} | w_1 w_2 \dots w_n). \tag{5}$$

Estimation of Sentence Probabilities

- Method 1 (brute-force)
  - Maximum-likelihood estimator for sentence s

$$P(S) = \frac{n(S)}{N}.$$

  - Problem: Some reasonable sentences have 0 probability. Need an extremely large *text corpus*.

- Method 2 (n-gram models)

$$P(S = w_1 \dots w_l) = \prod_{i=1}^{l} P(w_i | w_{i-n+1:i-1}) P(\text{eos} | w_{l-n+2:l}) \quad (6)$$

n-Gram Language Models

- word unigram
  REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE . . .

- word bigram
  THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO . . .

- The same idea can be applied to “letters”. In Chinese, it can be applied to “words”, “characters” or “syllables”.

- letter unigram: *ocro hli rgwr nmielwis eu ll nbnesebya th eei alhenhttpa oobttva nah brl . . .*

- letter bigram: *on ie antsoutinys are t inctore st be s deamy achin d ilonasive tucoowe at teasonare fuso tizin andy tobe seace ctisbe . . .*

- letter trigram: *in no ist lat whey cratict froure bers grocid pondenome of demonstures of the reptagin is regoactiona of cre . . .*

- letter four-gram: *the generated job providual better trand the displayed code . . .*

An n-Gram Example

- The number of probabilities in n-gram model grows exponentially with n. In practice, we start with bigram. Unigram is too rough.

- train set: {S1 = 我喜歡打羽毛球; S2 = 我甚麼球都可以打; S3 = 我甚至會空手道; S4 = 你喜歡打球嗎; S5 = 你至少會打桌球吧}

  test set: {T1 = 你會打羽毛球嗎; T2 = 你至會打}

$$P(\text{T1}) = P(\text{你}|\text{bos})P(\text{會}|\text{你})P(\text{打}|\text{會})P(\text{羽}|\text{打})P(\text{毛}|\text{羽})P(\text{球}|\text{毛})P(\text{嗎}|\text{球})P(\text{eos}|\text{嗎})$$
$$= \frac{2}{5} * 0 * \cdots * \frac{1}{5} = 0$$

$$P(\text{T2}) = P(\text{你}|\text{bos})P(\text{至}|\text{你})P(\text{會}|\text{至})P(\text{打}|\text{會})P(\text{eos}|\text{打}) = \frac{2}{5}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{5} = 0.01 > 0$$

Dealing with Data Sparsity

- Smoothing
  - additive smoothing
  - back-off smoothing
- class-based n-gram
- model interpolation

Applications of Language Models

- 自動語音辨識 (automatic speech recognition)
  $S^* = \arg\max_S p(X|S)p(S)$

- 中文輸入法 (Chinese input method)

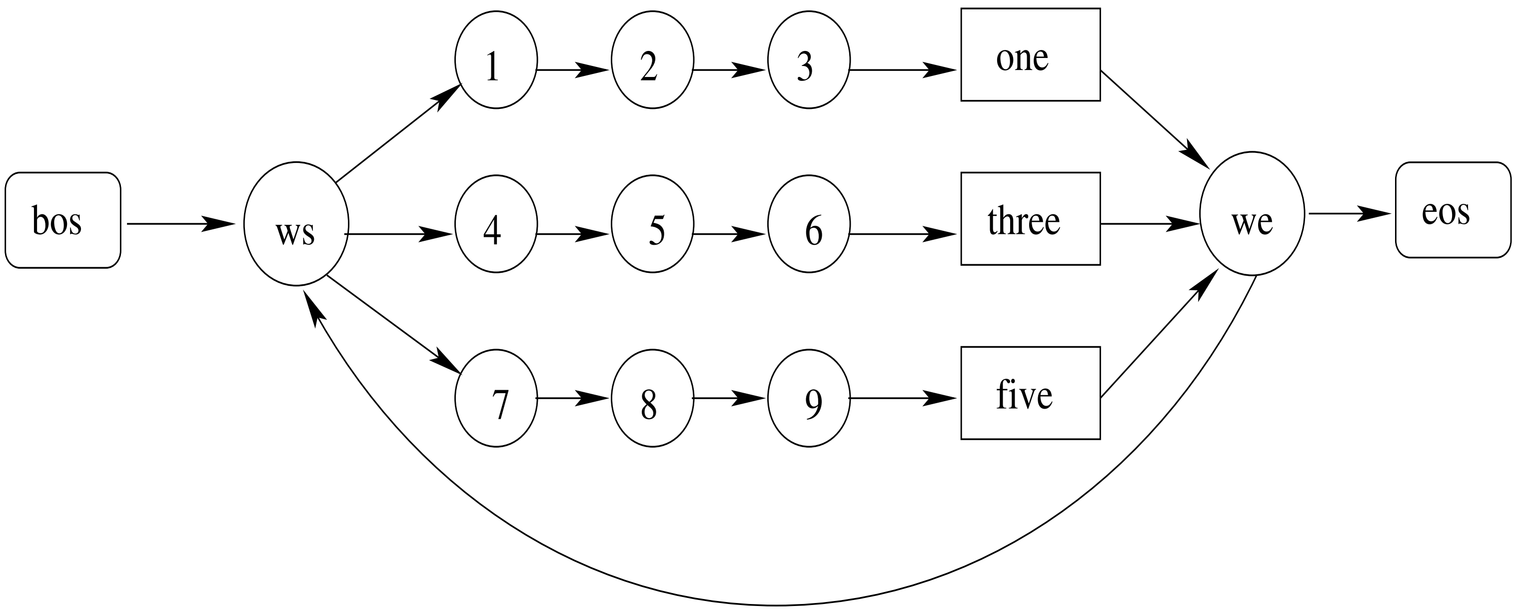
- 機器翻譯 (machine translation)
  $e^* = \arg\max_e p(f|e)p(e)$.
  http://www.systransoft.com/ http://google.com/language_tools

- 資訊檢索 (information retrieval)

Continuous Speech Decoder

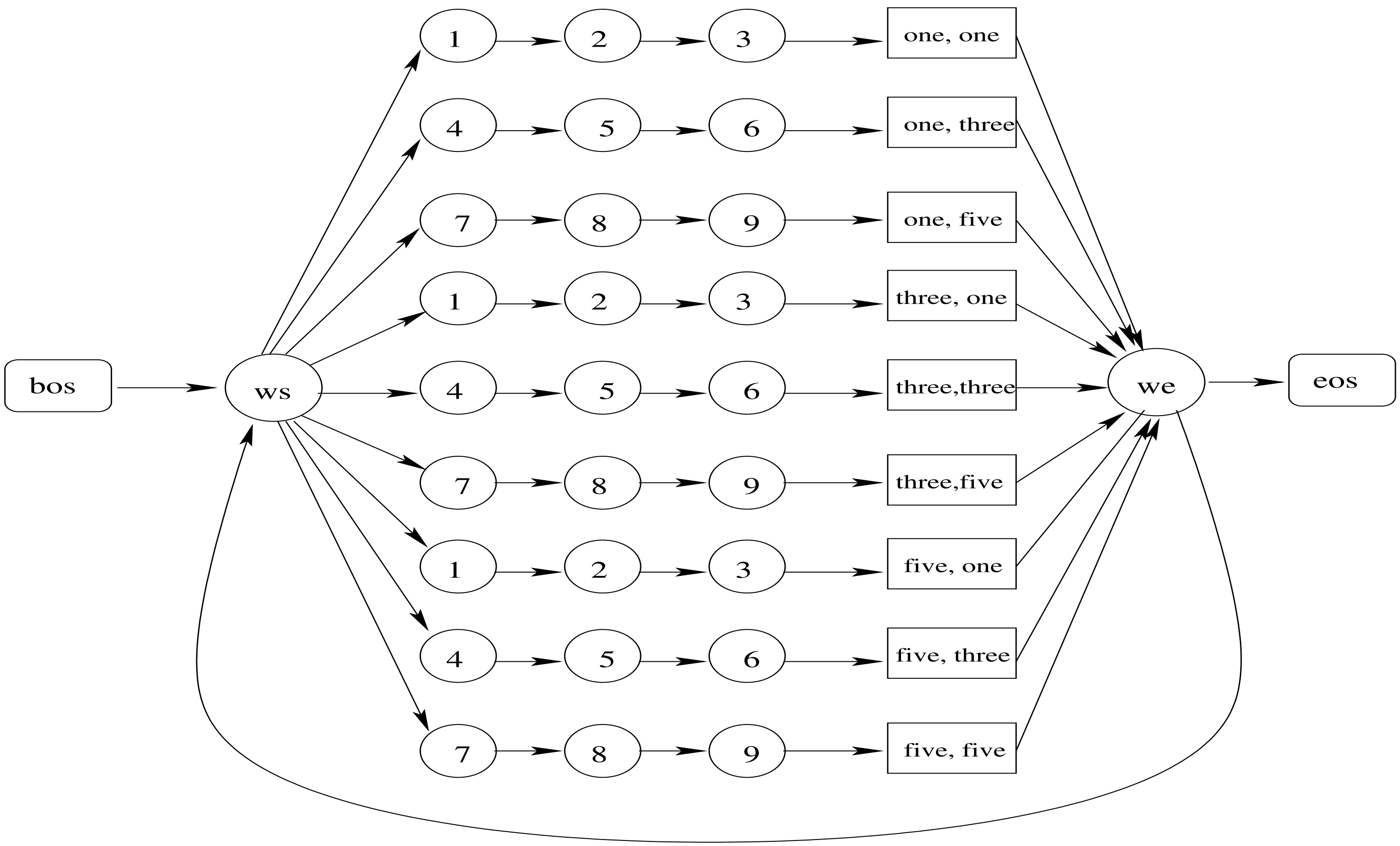
- A re-entrant network where the optimal path is searched for.
- Acoustic model scores are computed at the phone nodes.
- Language model scores are computed at the word nodes.

Examples of Recognition Networks

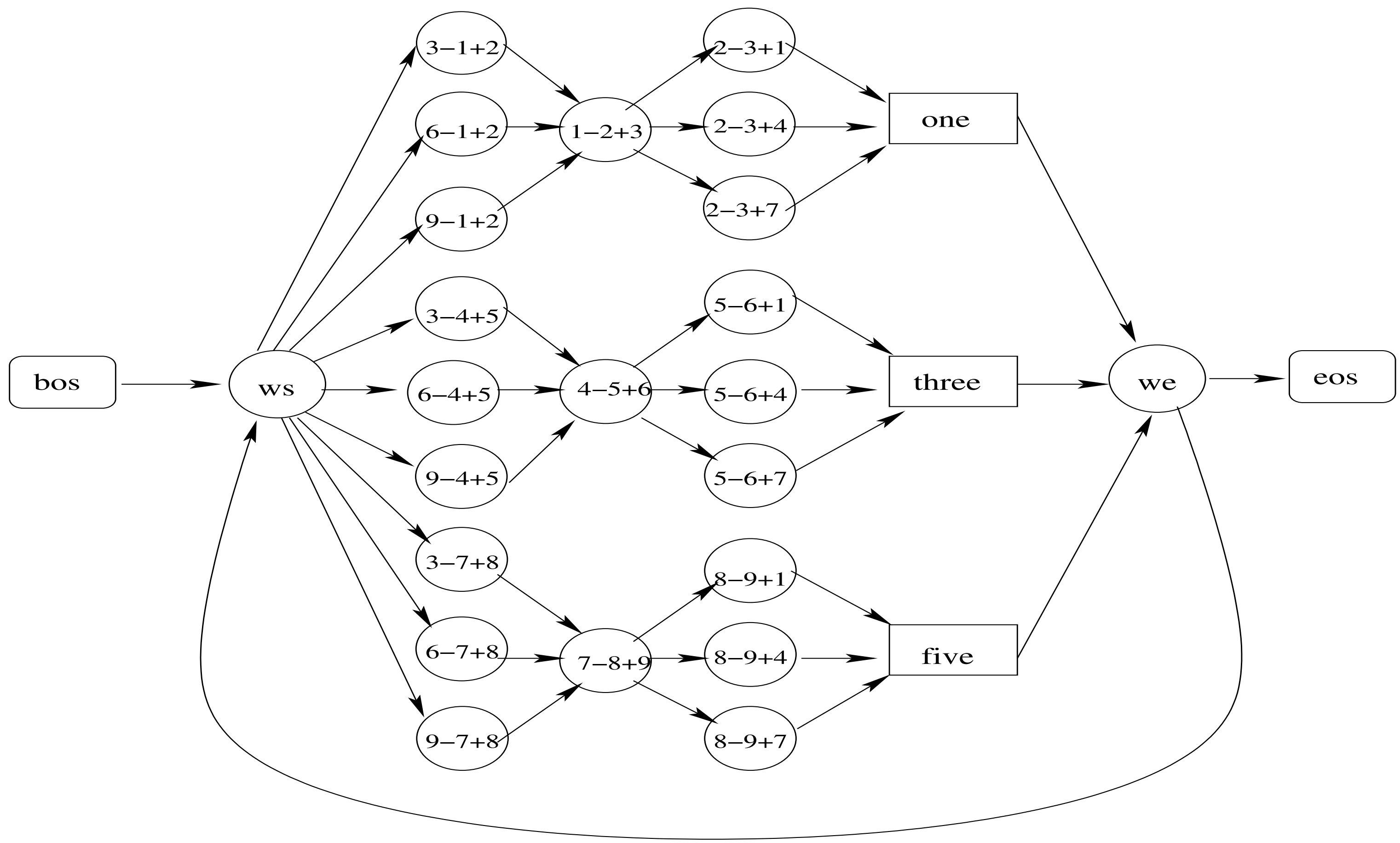


vocabulary set = {one, three, five}

Examples of Recognition Networks



Examples of Recognition Networks



Large-Vocabulary Continuous Speech Recognition (LVCSR)

- acoustic model refinement
  - context-dependent phone models
  - parameter typing
  - model adaptation
- long-range language models
- decoder design

A Decoder Design in LVCSR

- the search problem and maximum approximation

$$
\begin{aligned}
w^*_{1:N} &= \arg\max_{w_{1:N}} P(w_{1:N}, x_{1:T}) \\
&= \arg\max_{w_{1:N}} \sum_{s_{1:T}} P(w_{1:N}, s_{1:T}, x_{1:T}) \\
&\doteq \arg\max_{w_{1:N}} \left\{ P(w_{1:N}) \max_{s_{1:T}} Pr(x_{1:T}, s_{1:T} | w_{1:N}) \right\}
\end{aligned}
$$

- tree-structured lexicon

- time-synchronous word-conditioned Viterbi search
  - $Q_v(t, s)$: best partial match ending in s with predecessor word v
  - inter-tree recursion $Q_v(t, s) = \max_q \{ p(x_t, s|q) Q_v(t-1, q) \}$

- intra-tree recursion $Q_v(t, s = 0) = \max_u \{p(v|u) Q_u(t, S_u)\}$

- language model look-ahead (for bigram)

$$\pi_v(s) = \max_{w \in W(s)} p(w|v)$$

- beam pruning: Discard those hypotheses whose likelihood scores (AM and LM combined) too far behind the maximum

$$\tilde{Q}_v(t, s) < f_0 \; * \; \max_{v', s'} \tilde{Q}_{v'}(t, s'),$$

where $\tilde{Q}_v(t, s) = \pi_v(s) Q_v(t, s)$.

- both the maximum approximation and pruning can lead to sub-optimal hypothesis decoded. Yet they are much more efficient computationally.

Summary

- ASR = speech features + acoustic model + language model + decoder
- HMM acoustic models: state + emission + efficient algorithms
- n-gram language models: not satisfactory but acceptable
- an efficient decoder: tree-structured lexicon + language model look-ahead + pruning

研究概況

- 研究群
  - 台灣: 中研院, 台大, 清華, 交大, 成大, 師大, 長庚, 工研院, 中華電信, Acer, 台達電子, . . .
  - 世界: Cambridge, CMU, Berkeley (ICSI), MIT, Tokyo Institute of Technology, CUHK, University of Technology Aachen, IBM, . . .
- 研究領域
  語音辨識, 語者辨識, 噪音強健性, 聲控, 關鍵字搜尋, 資訊檢索, . . .