UNIVERSITY OF SURREY
Centre for Vision Speech & Signal Processing
University of Surrey, Guildford GU2 7XH.

Speech Recognition Part 3

Dr Philip Jackson

G. Grammars for speech recognition
H. HMM likelihoods: tasks 1 & 2
I. Vocal-tract acoustics

Grammar outline

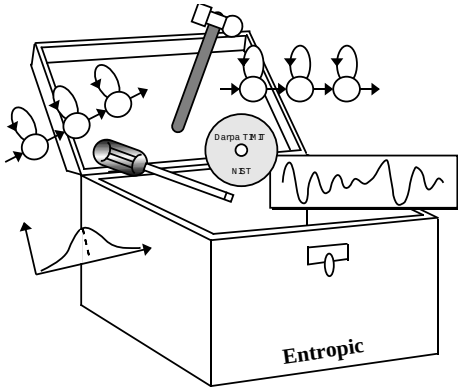
- **Isolated word recognizer**
  - Isolated digit recognition
  - Building the grammar
  - Training & testing
- **Task grammars**
  - Connected word recognition
  - Null states and word networks
  - Continuous speech recognition

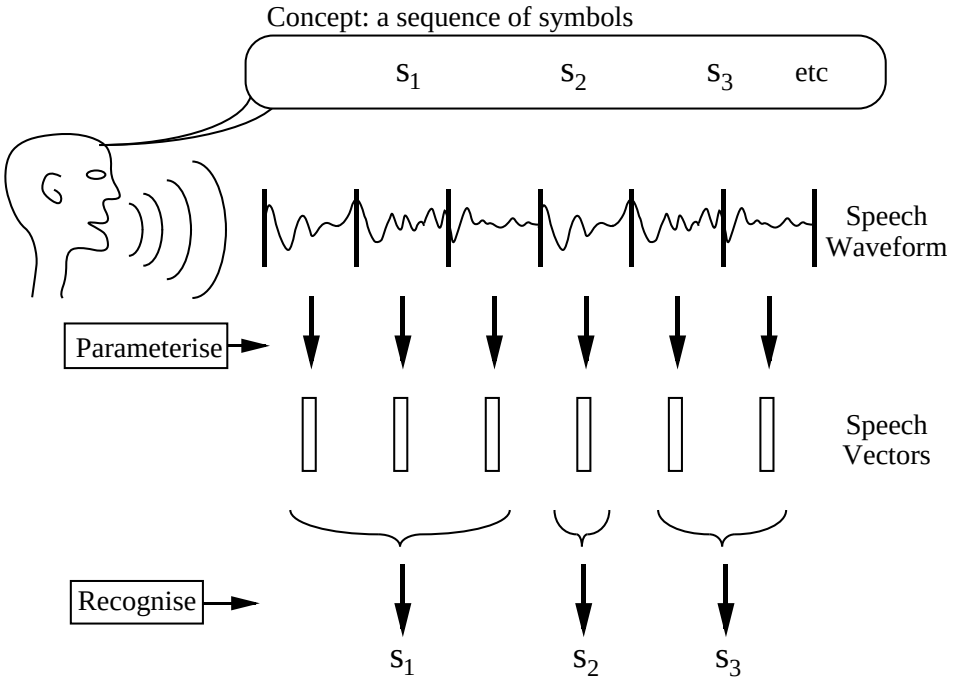
Key concept:

use a task's syntax to connect models that comprise a complete utterance

Isolated word recognizer development

- Data preparation
- Training
- Testing
- Analysis





Message encoding and decoding.*

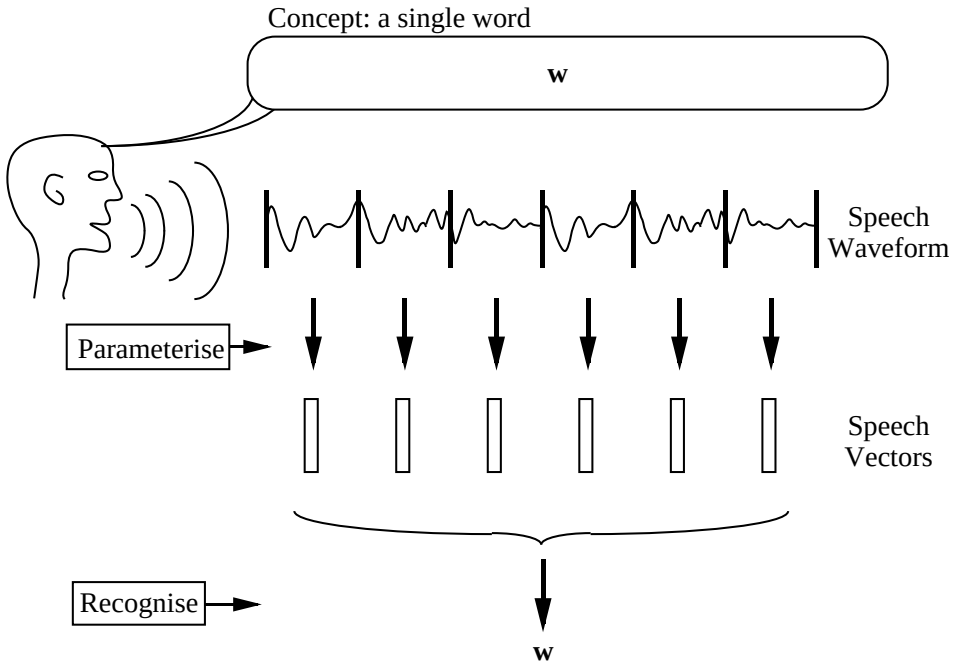
Isolated Word Recognition (IWR) task

The problem is to find

$$\hat{w} = \arg\max_i \{P(w_i|\mathcal{O})\} \tag{1}$$

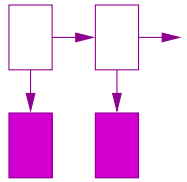
where according to Bayes

$$P(w_i|\mathcal{O}) = \frac{P(\mathcal{O}|w_i)P(w_i)}{P(\mathcal{O})} \tag{2}$$



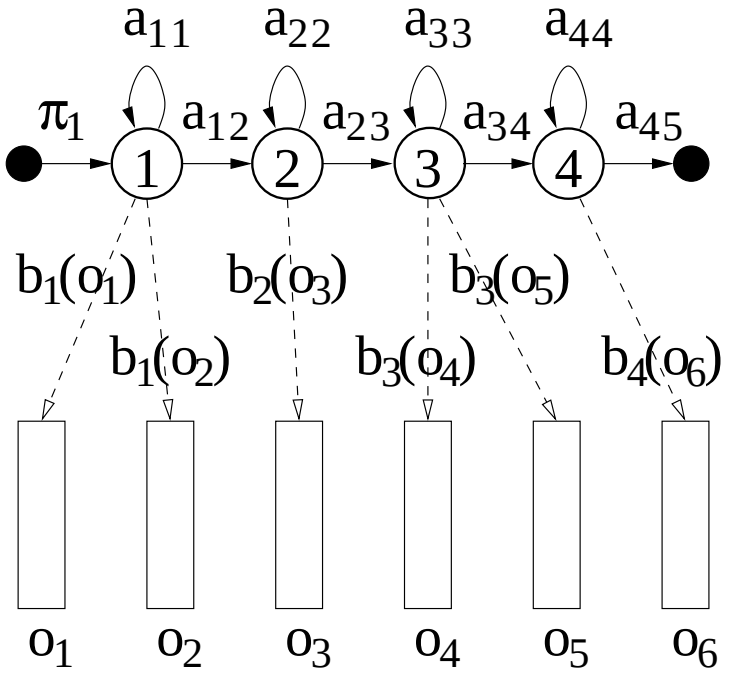
Gsolated word problem.*

The hidden Markov model



In IWR, we treat each HMM as a word template

$$
\begin{aligned}
P(\mathcal{O}|w_i) &\approx P(\mathcal{O}|\lambda_i) \\
&\approx \max_X \left[\left(\prod_{t=1}^{T} a_{x_{t-1}x_t} \, b_{x_t}(o_t) \right) \eta_{x_T} \right]
\end{aligned}
\tag{3}
$$

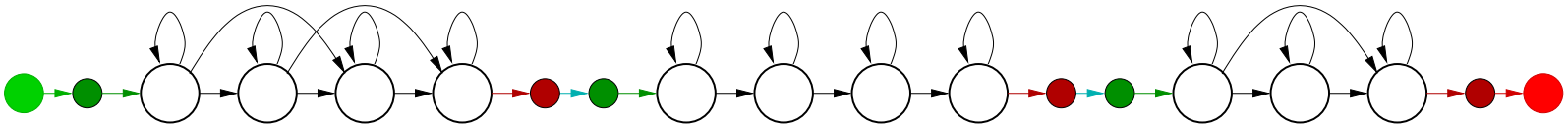


The HMM generation model.

Null states in an HMM

Properties of non-emitting null states:

- determine beginning and end of a model
- do not generate any observations
- can align a given model with an isolated test utterance
- useful for joining models together in a grammar
- transitions associated with null states can be modified to incorporate language model



IWR: Building the grammar

Example utterances:

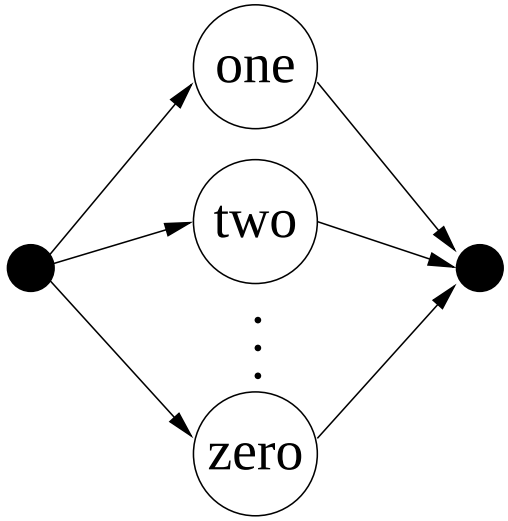
- eight
- oh
- six

Task grammar:

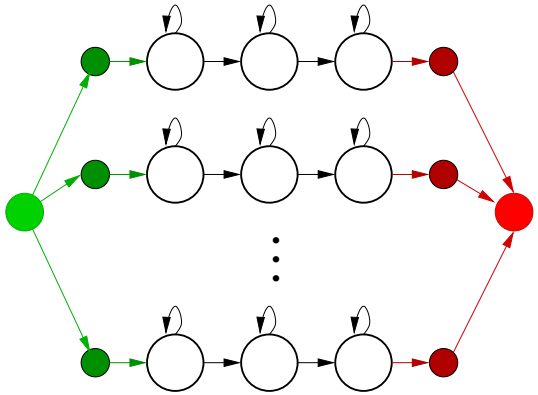
```
$digit = ONE | TWO | THREE |
         FOUR | FIVE | SIX |
         SEVEN | EIGHT | NINE |
         OH | ZERO;
( SENT-START $digit SENT-END )
```

Key:

| alternatives

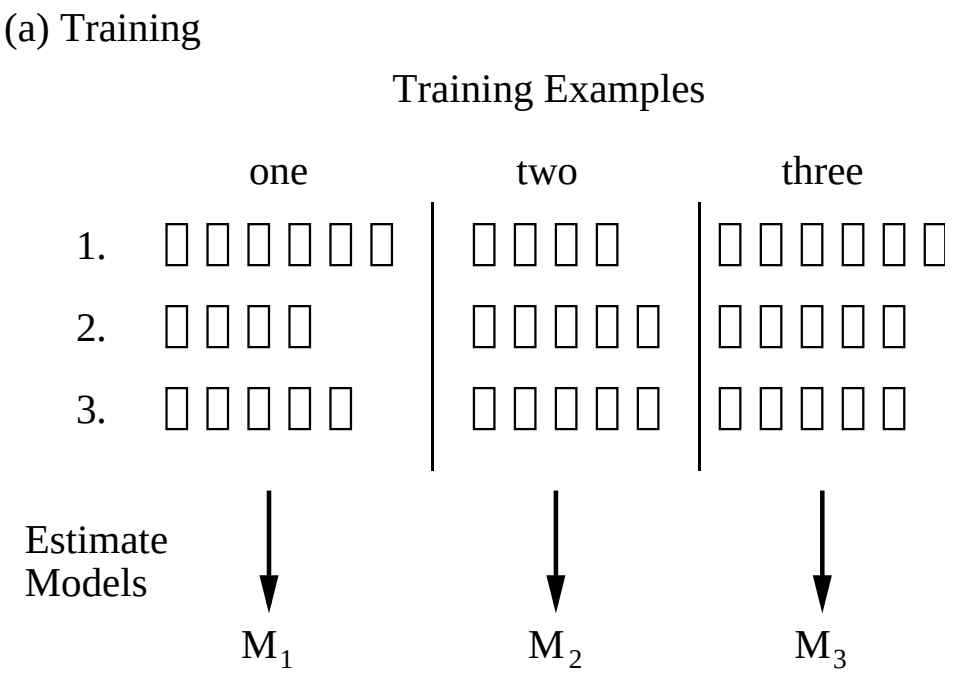


Grammar for IWR



Connected HMMs.

Training and test data



Using HMMs for isolated word recognition.*

Calculating recognition performance

Types of recognition error (e.g., for ground truth "A-B-C"):

- Substitution, S (e.g., "A-D-C")
- Deletion, D (e.g., "A-C")
- Insertion, I (e.g., "A-B-E-C")

$$\% \text{ Correct } = 100 \times \frac{N - S - D}{N} \tag{4}$$

$$\% \text{ Accuracy } = 100 \times \frac{N - S - D - I}{N} \tag{5}$$

$$\text{Error rate } \% = 100 \times \frac{S + D + I}{N} \tag{6}$$

Task grammars

IWR: Binary word grammar

Example utterances:

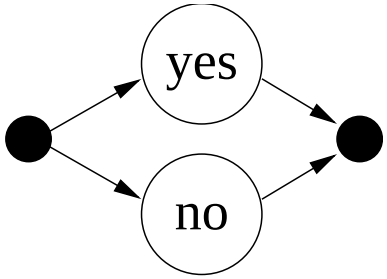
- yes
- no

Task grammar:

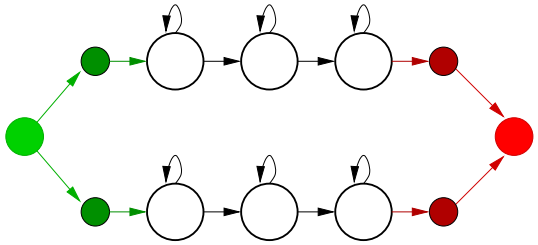
```
$answer = YES | NO;
( SENT-START $answer SENT-END )
```

Key:

| alternatives

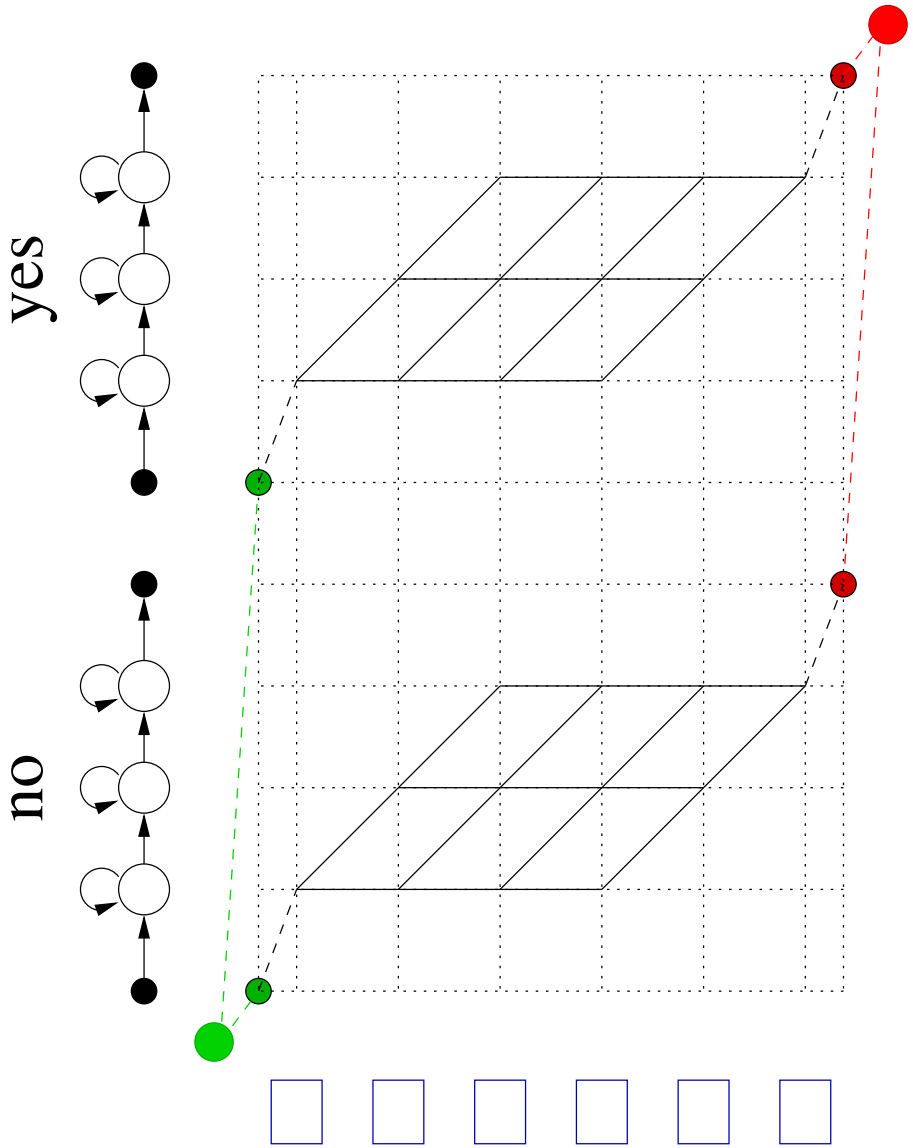


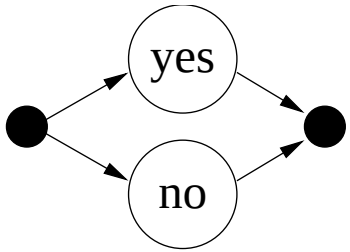
Binary IWR grammar.



Connected HMMs.

IWR: binary word trellis





Trellis (left) and grammar (right) for a two-word IWR network.

IWR: Isolated digit grammar

Example utterances:

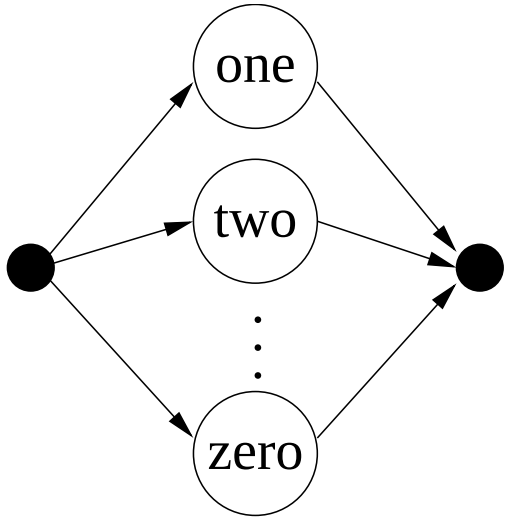
- eight
- oh
- six

Task grammar:

```
$digit = ONE | TWO | THREE |
         FOUR | FIVE | SIX |
         SEVEN | EIGHT | NINE |
         OH | ZERO;
( SENT-START $digit SENT-END )
```

Key:

| alternatives



Grammar for isolated digit recognition.

CWR: Connected digit grammar

Example utterances:

- six eight six zero three one
- one oh one
- six oh four four

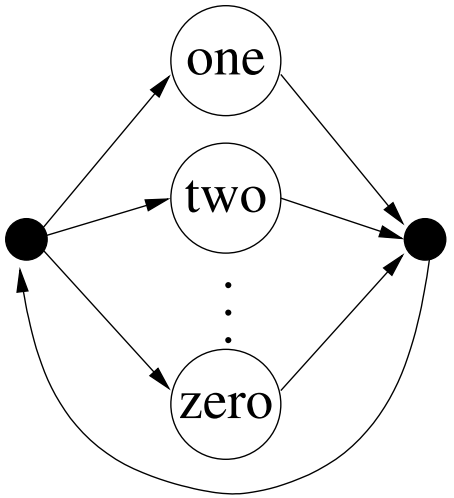
Task grammar:

```
$digit = ONE | TWO | THREE |
         FOUR | FIVE | SIX |
         SEVEN | EIGHT | NINE |
         OH | ZERO;
( SENT-START <$digit> SENT-END )
```

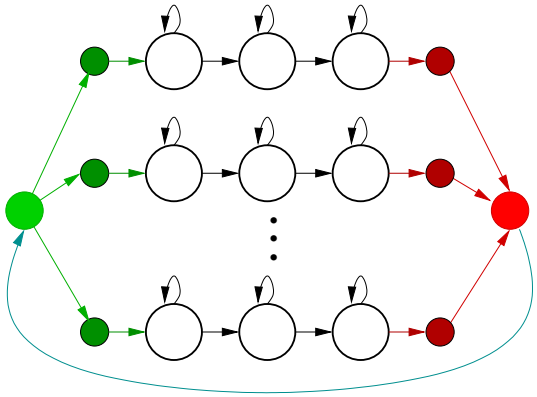
Key:

| alternatives

<·> one or more reps

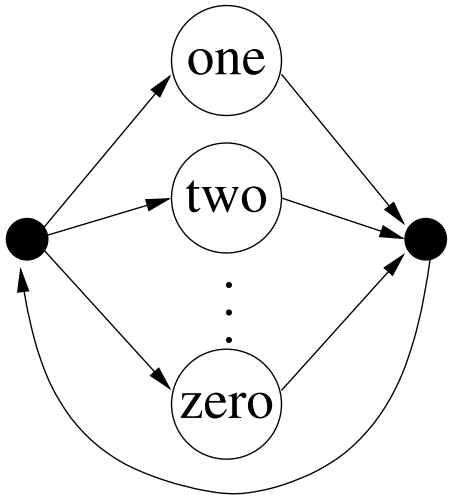


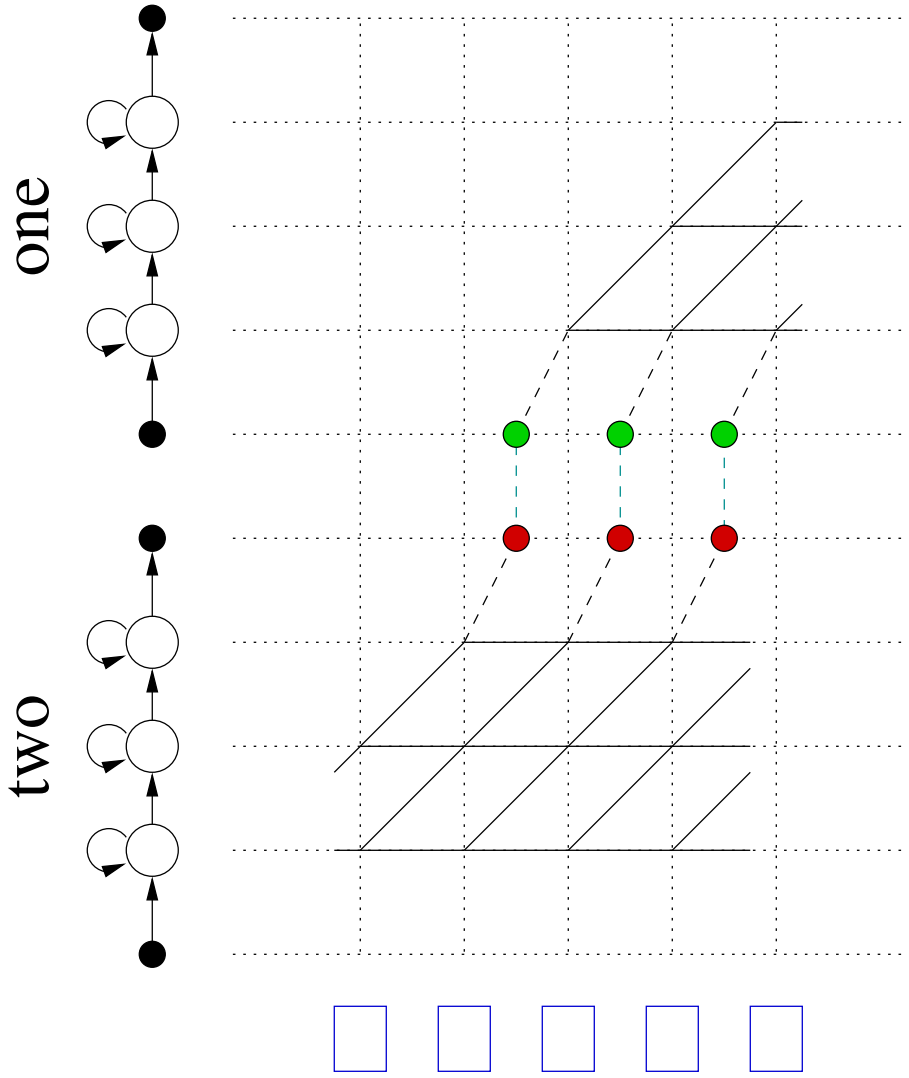
Grammar for connected digit recognition.



Connected HMMs.

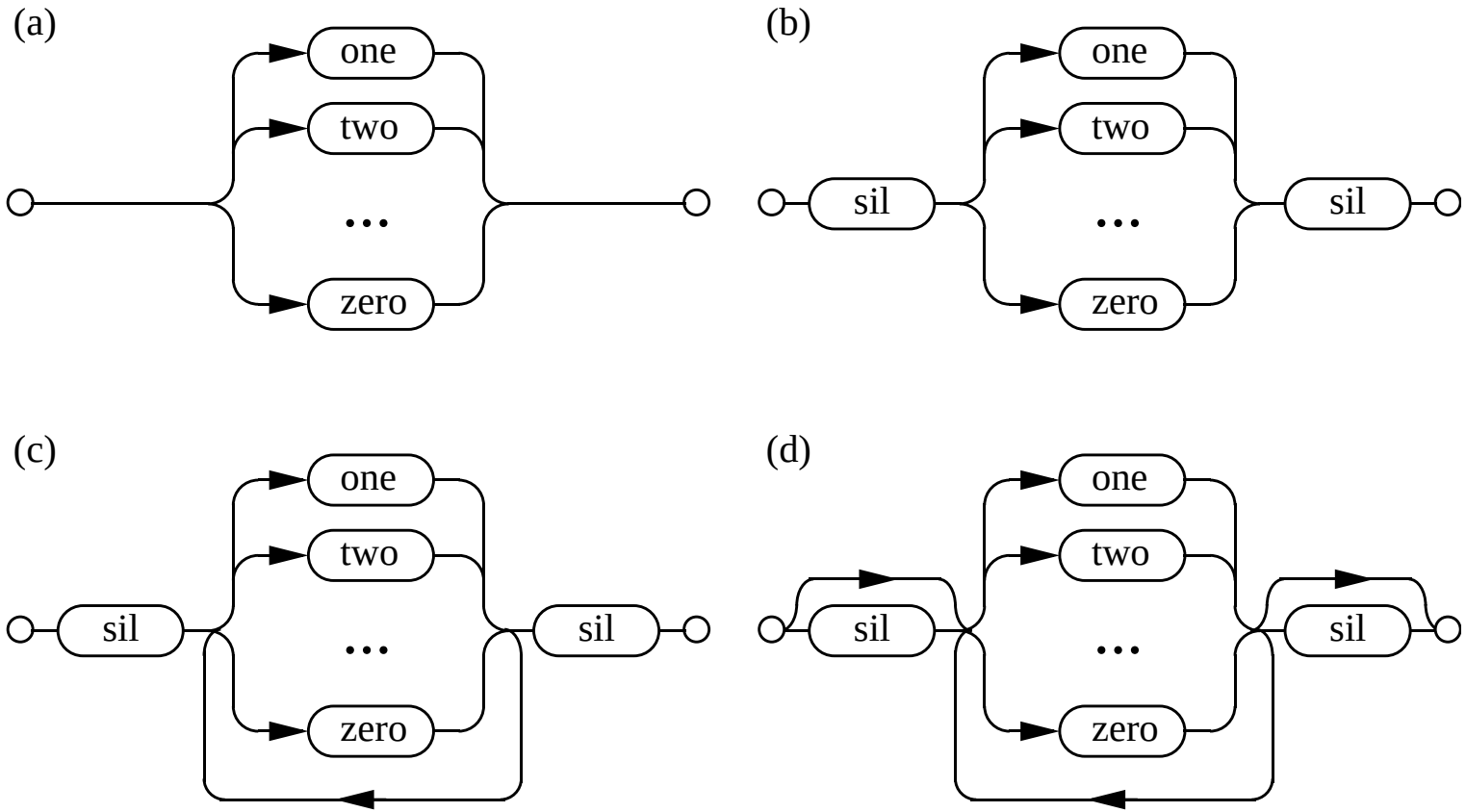
CWR: connected-digit trellis





Grammar (left) and trellis (right) for connected-digit recognition.

Isolated- & connected-digit grammars

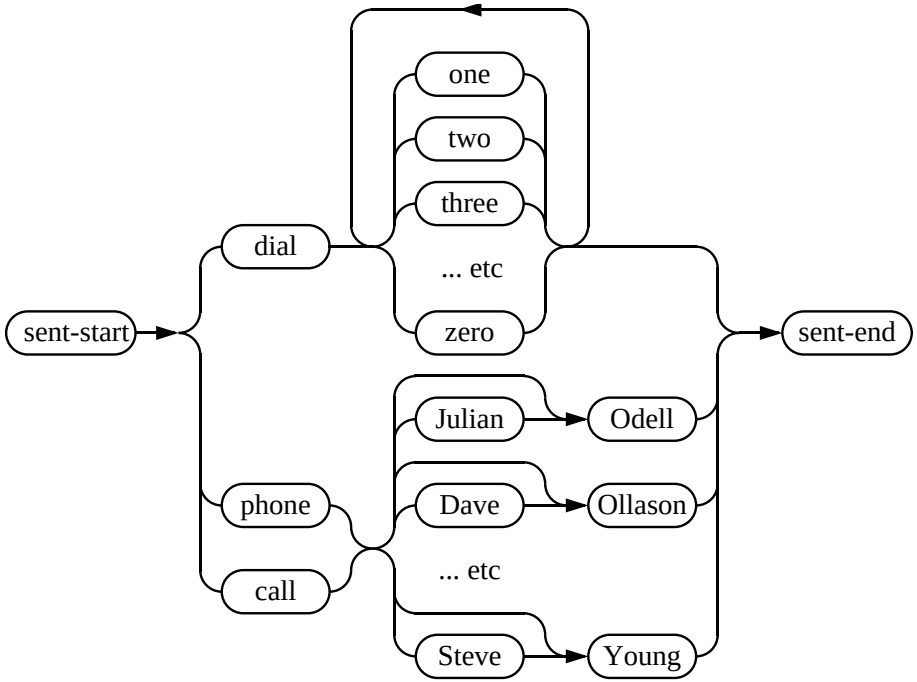


Example networks for digit recognition tasks:* (a) IWR, (b) IWR with end-point adjustment using silence model, (c) CWR with silence model, (d) with optional silence.

CWR: Connected word grammar

Example utterances:

- Dial three three two six five four
- Phone Woodland
- Call Steve Young



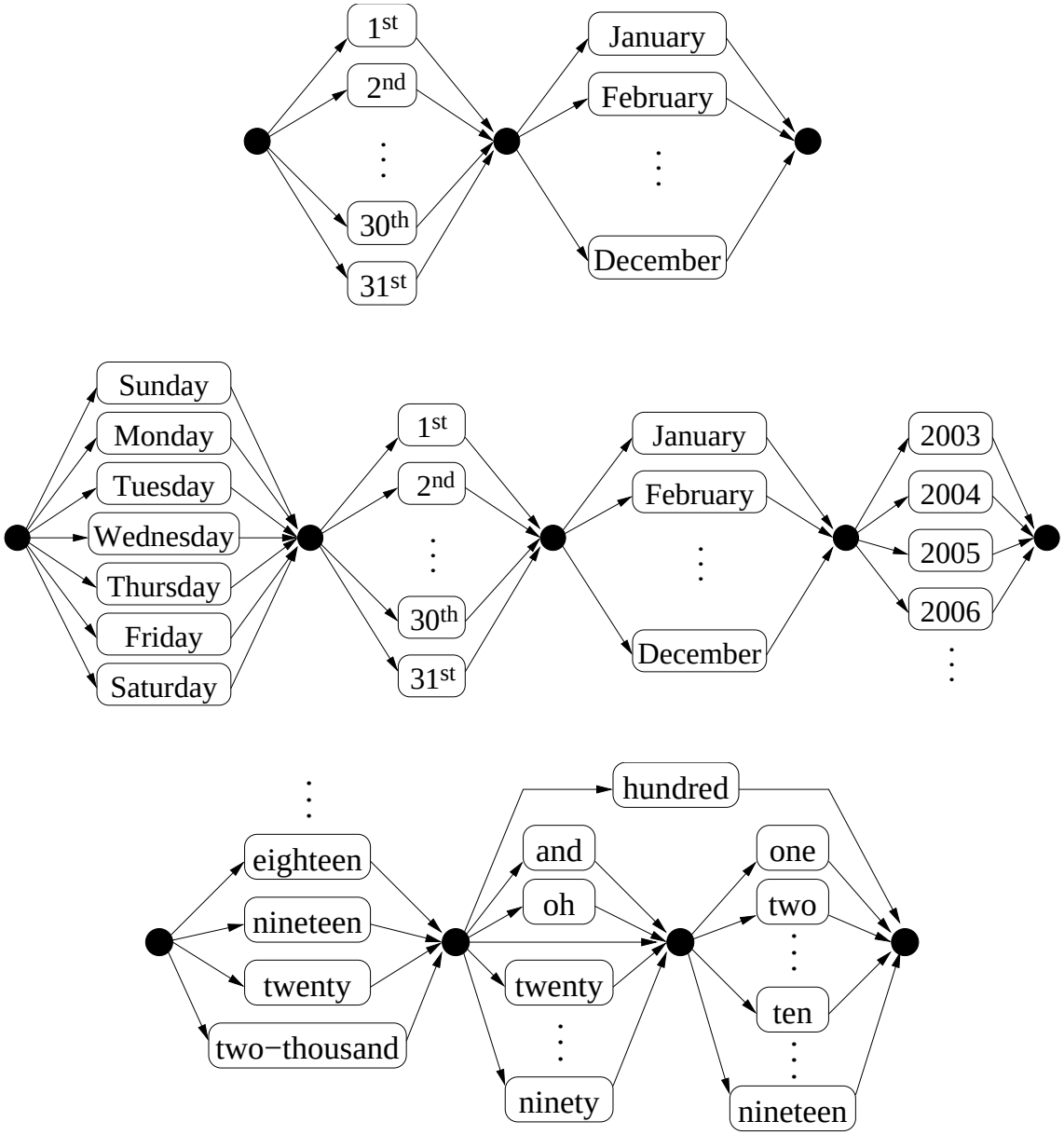
Grammar for voice dialling.*

Task grammar:

```
$digit = ONE | TWO | THREE | FOUR | FIVE |
         SIX | SEVEN | EIGHT | NINE | OH | ZERO;
$name  = [ JULIAN ] ODELL |
         [ DAVE ] OLLASON |
         [ PHIL ] WOODLAND |
         [ STEVE ] YOUNG;
( SENT-START ( DIAL <$digit> | (PHONE|CALL) $name) SENT-END )
```

Key: | alternatives, [·] optional, <·> one or more reps

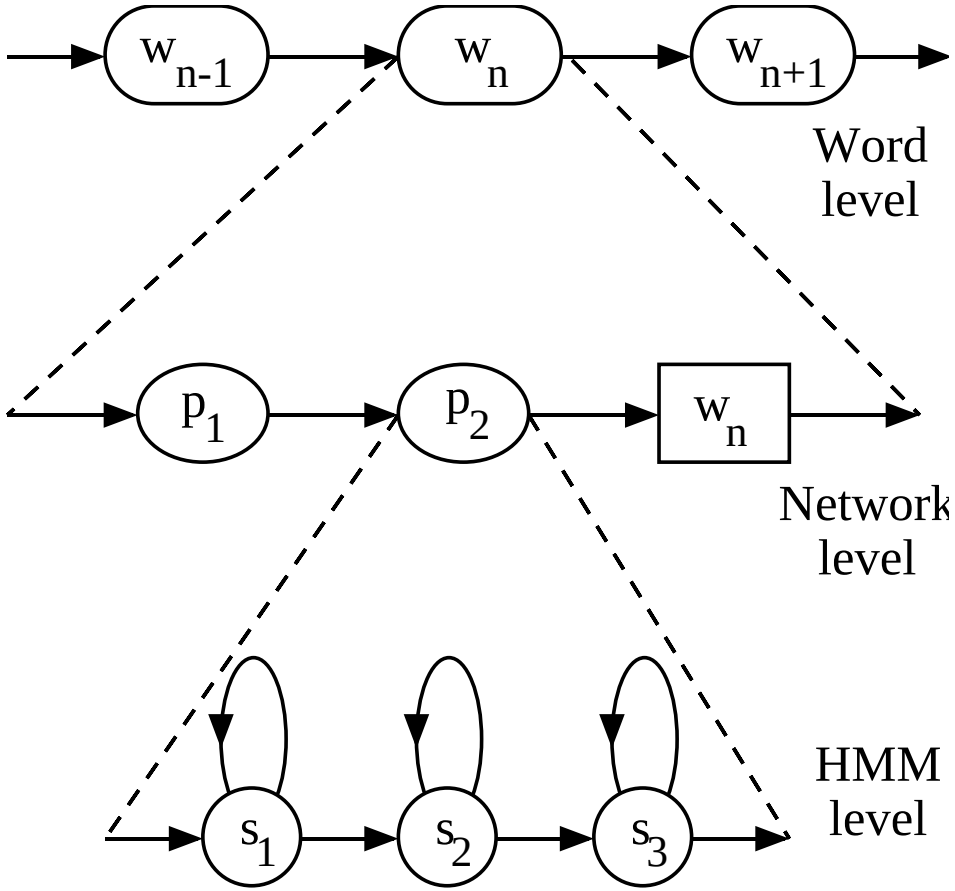
CWR: context-free grammars



Example grammars (from top): date, day and year.

Towards large-vocabulary speech recognition

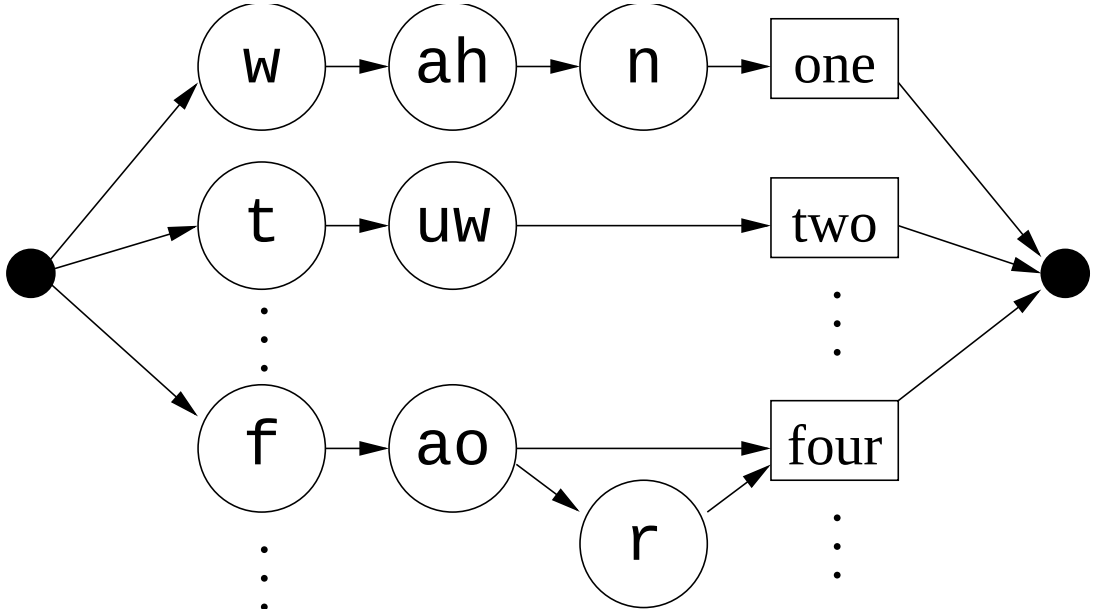
By nesting small models to make templates for larger units, we can exploit phonetics and word morphology



Recognition network levels.*

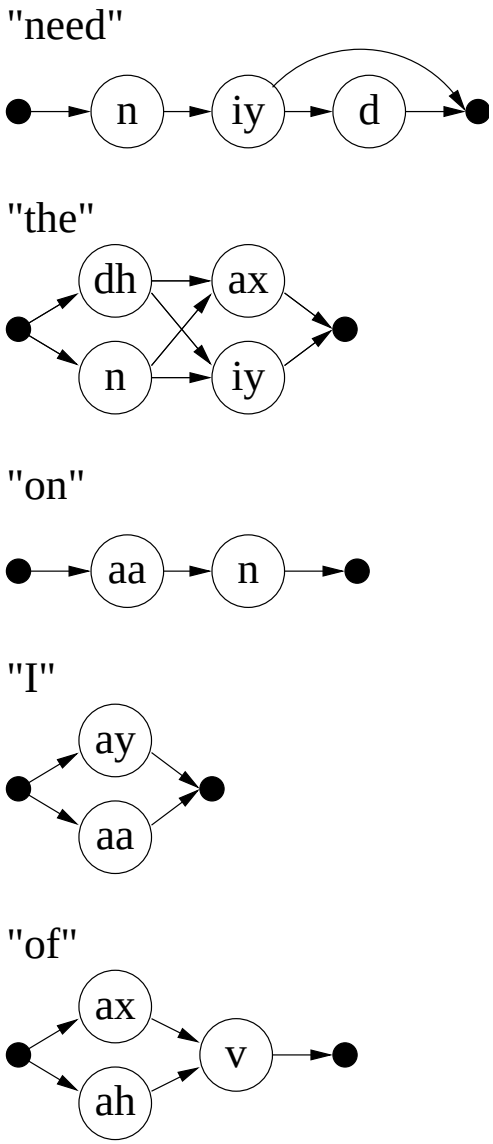
IWR: Phone-based digit dictionary

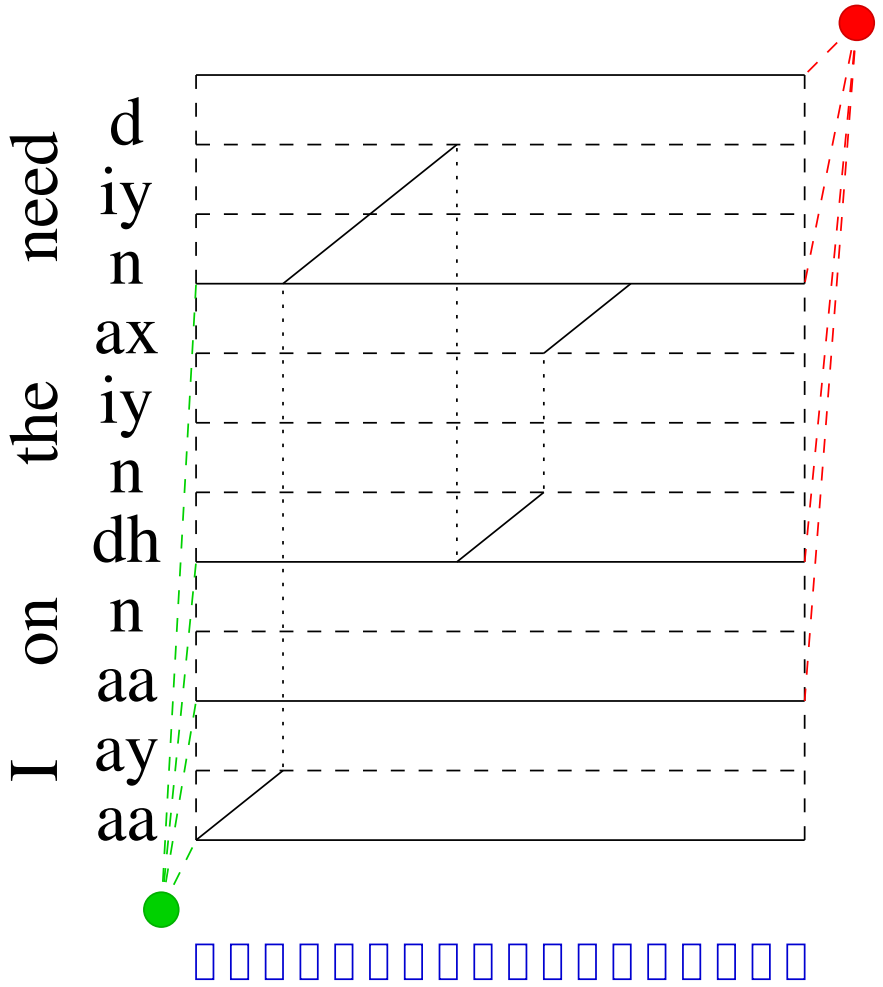
ONE	w ah n	SIX	s ih k s
TWO	t uw	SEVEN	s eh v n
THREE	th r iy	EIGHT	ey t
FOUR	f ao	NINE	n ay n
FOUR	f ao r	OH	ow
FIVE	f ay v	ZERO	z ia r ow



Grammar for phone-based isolated digit recognition.

CSR: continuous-speech grammar and trellis

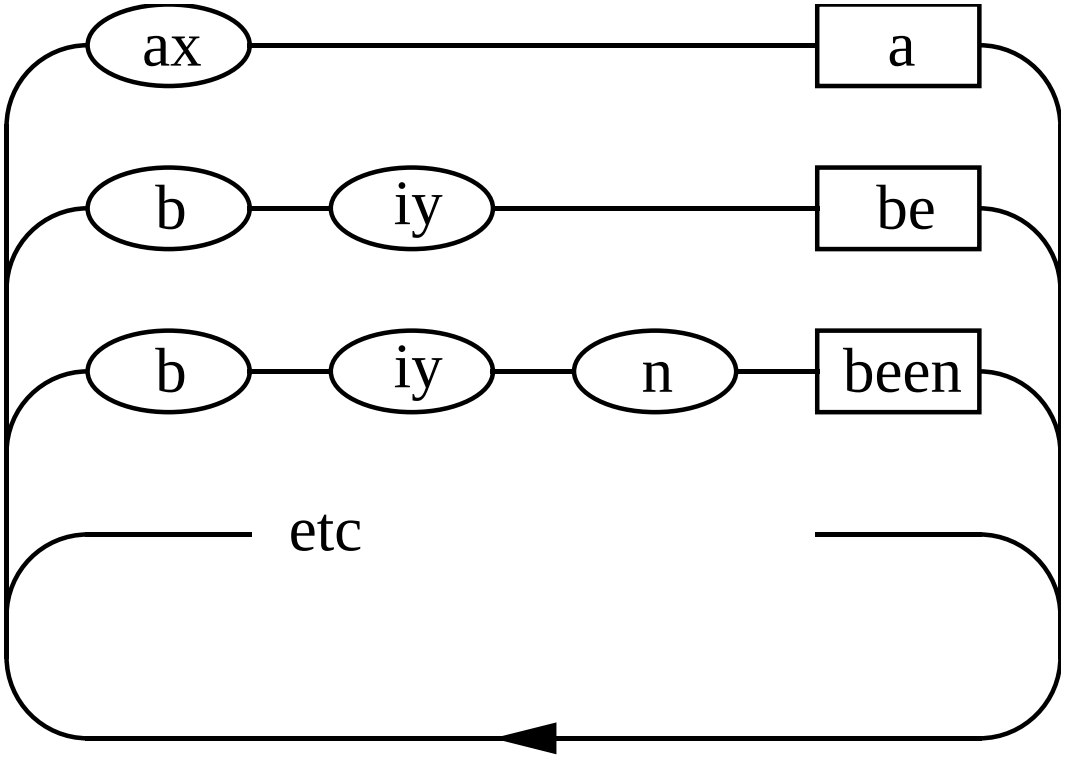




Continuous speech recognition networks (left) & trellis (right). G.20

Large-vocabulary continuous speech recognition (LVCSR)

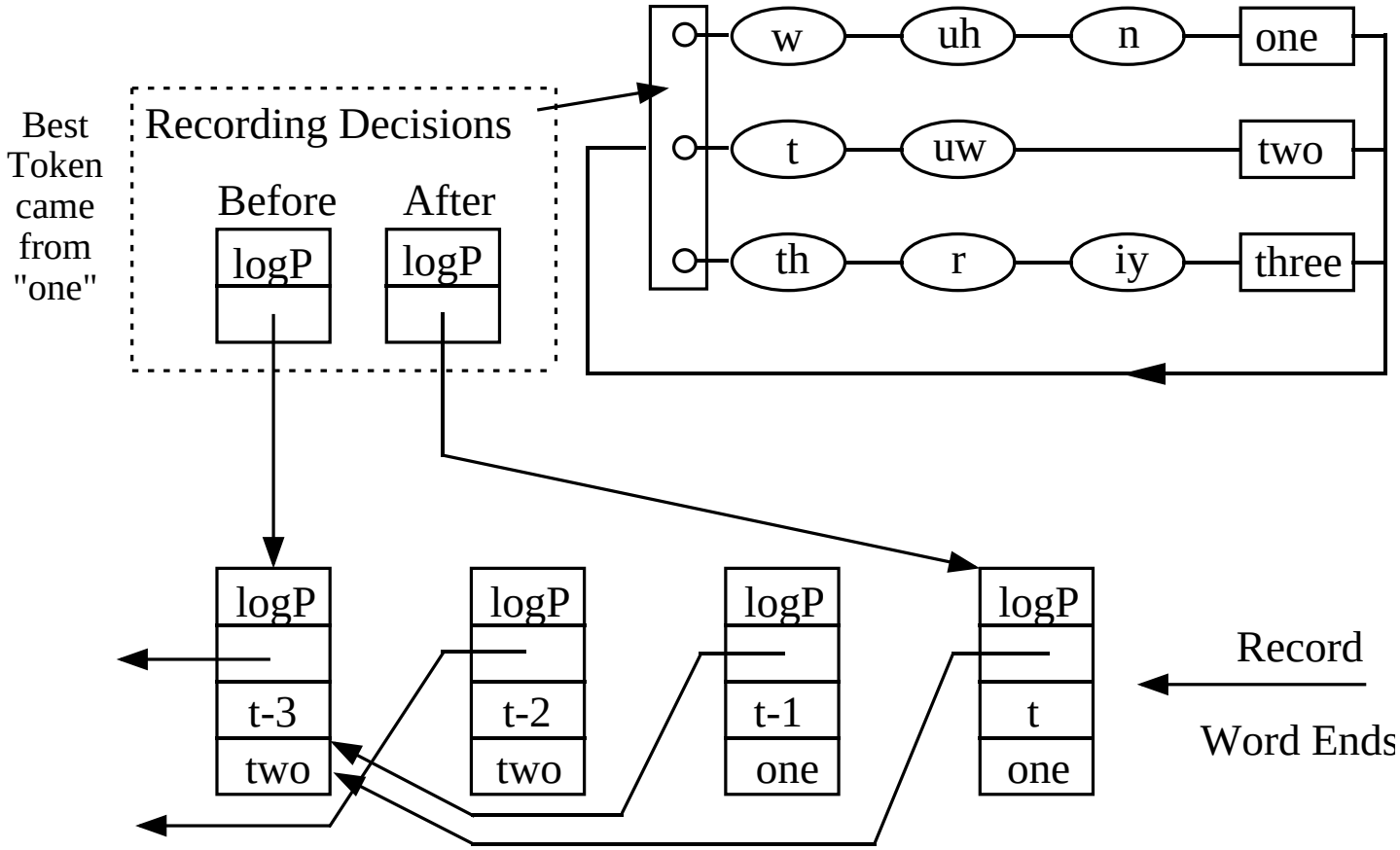
A pronunciation dictionary of phonemes and labels enables the recognition grammar to incorporate a long list of words



Word network for continuous speech recognition.*

CSR: recognition and traceback

During decoding, the recognizer records the scores as log-likelihoods and outputs labels at the word boundaries



Recording word-boundary decisions during continuous recognition.*

Grammar summary

- Isolated digit recognition
  - Null states connect grammars
  - Trellis diagrams
  - Correctness, accuracy & error rate
- Task grammars
  - Isolated word recognition (IWR)
  - Connected word recognition (CWR)
  - Context-free grammar
  - Enlarging vocabulary exploiting word structure
  - Building templates from phone models
  - Continuous speech recognition (CSR)