Multi-Modal User Interaction
Fall 2008

Lecture 3: Speech Recognition II

Zheng-Hua Tan

Department of Electronic Systems
Aalborg University, Denmark
zt@es.aau.dk

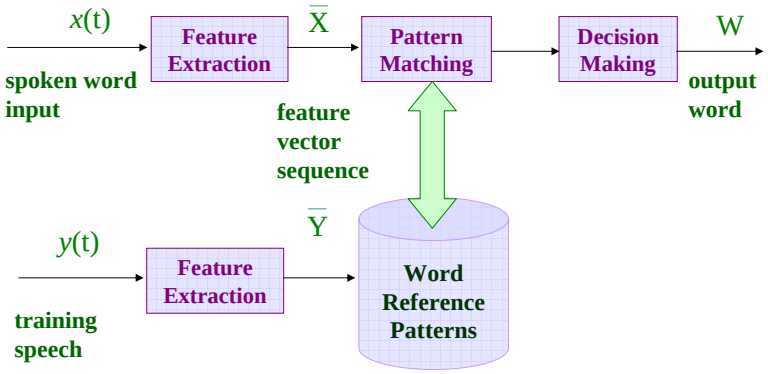
Part I: Types of speech recognizers

- Types of speech recognizers
  - IWR
  - Rule grammar
  - N-gram
- Acoustic modelling
- Sphinx-4

Types of speech recognisers

- Isolated word recognition
- Grammar based recognition
- Large vocabulary continuous speech recognition (N-gram)

Template based method for IWR

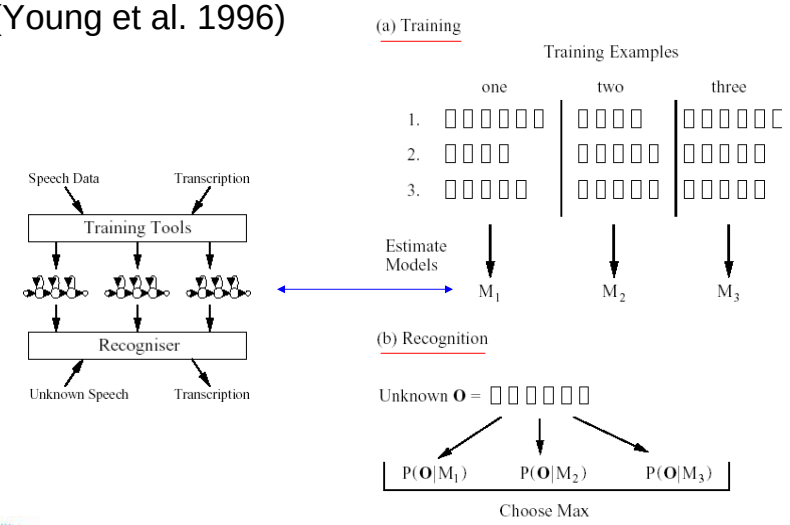


Template matching mechanism

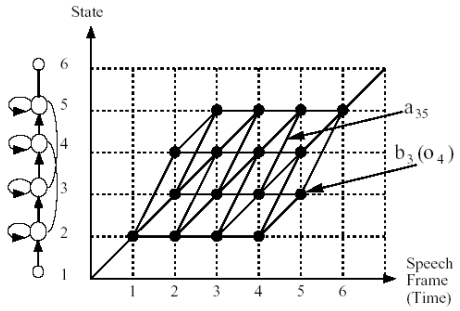
- Calculate the distance between two patterns
- Dynamic time warping (DTW)

HMM for IWR

(Young et al. 1996)



The Viterbi algorithm for IWR



Language modelling – word looping?

- The allowed sequence of phoneme-based HMMs is defined by a finite state network and all of the words are placed in a loop

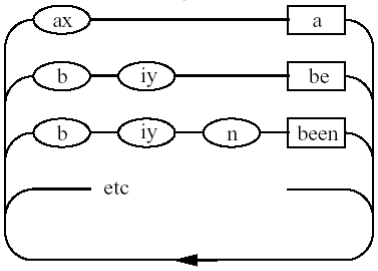


Fig. 1.7 Recognition Network for Continuously Spoken Word Recognition

Grammar – constraining search space

IWR, grammar-based ASR, N-grams

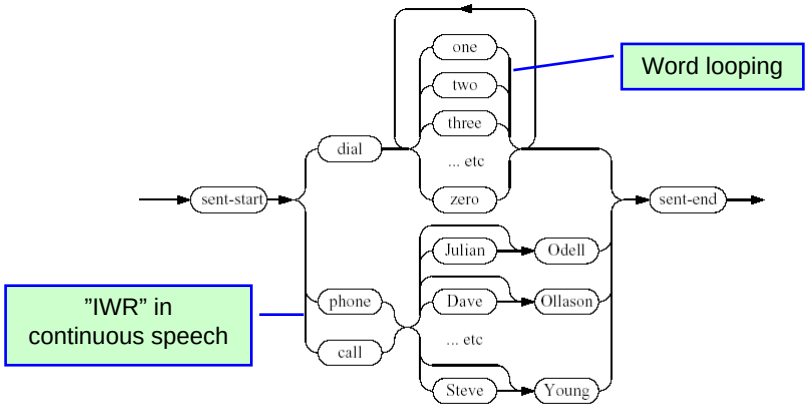


Fig. 3.1 Grammar for Voice Dialling

N-grams

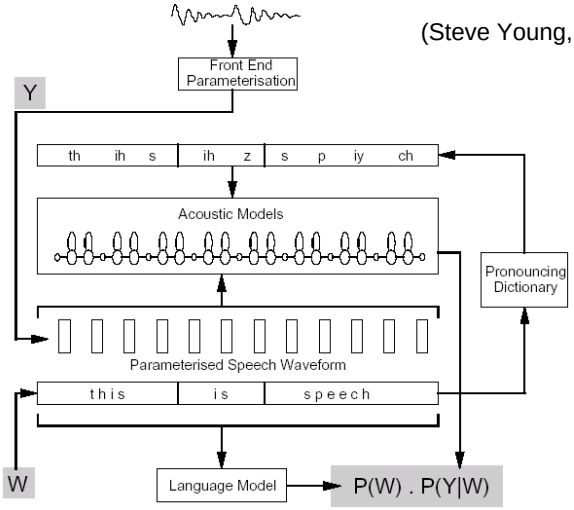
- LM is estimating the probability of word in an utterance given the preceding words.
- N-grams (bigrams, trigrams, etc.)

$$P(w_k \mid w_1 \ldots w_{k-1}) = P(w_k \mid w_{k-n+1} \ldots w_{k-1})$$

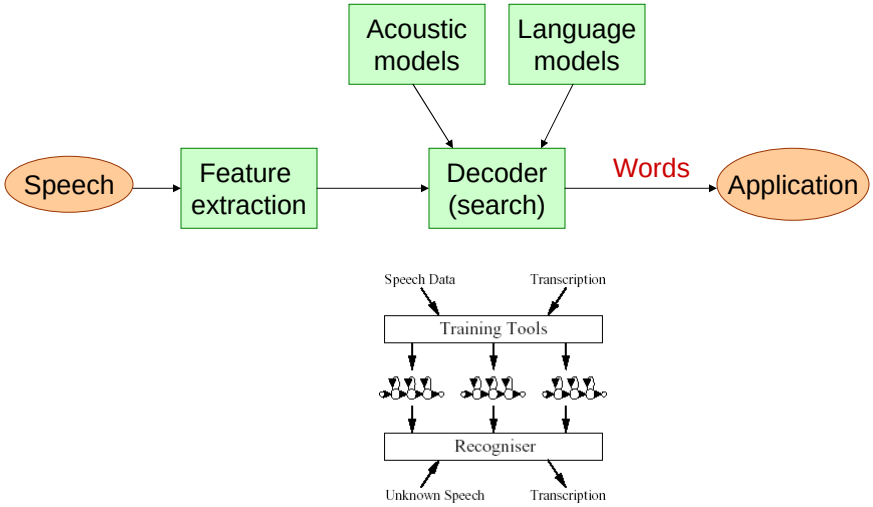
- Discounting and backing-off

LVCSR system overview

(Steve Young, 1996)



Speech recognition system

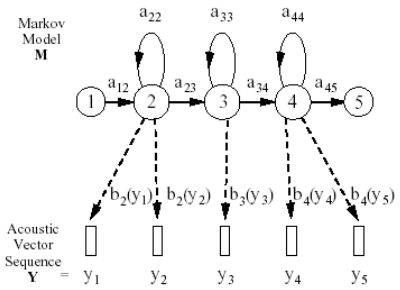


Part II: Acoustic modelling

- Types of speech recognizers
  - IWR
  - Rule grammar
  - N-gram
- Acoustic modelling
- Sphinx-4

Acoustic modelling

- Calculating $P(Y \mid W)$
- In LVCSR, word sequences are decomposed into phones
- Each phone is represented by an HMM with three emitting states and a simple left-to-right topology.



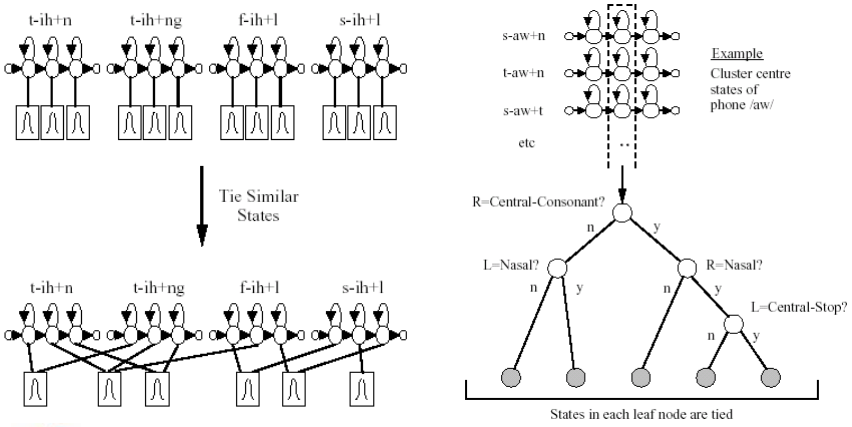
Context-dependent phone model

- Phrase "Beat it!"
  - Phone sequence: sil b iy t ih t sil
  - Triphone sequence:
    sil sil-b+iy b-iy+t iy-t+ih t-ih+t ih-t+sil sil
- Curse of dimension: Parameters of triphone modelling

  60 000 triphones (45 phones, 45^3= 91 125) * 3 states * (10 mixtures * 39 feature components * 2 Gaussian parameters + 10 mixture weights) = 142.2 million parameters!

State tying and decision tree clustering

- Too many parameters and too little training data
  → state tying + phonetic decision tree clustering



Part III: Sphinx-4

- Types of speech recongizers
  - IWR
  - Rule grammar
  - N-gram
- Acoustic modelling
- Sphinx-4

Sphinx

- Speech recognition software.
- Based on HMM (Hidden Markov Model)
- Four different versions since 1988.
  - Previous versions Sphinx-1, Sphinx-2, Sphinx-3 were programmed in C
  - Latest version (Sphinx-4) was programmed in JAVA

Sphinx-4

- A flexible open source framework for speech recognition
  - Object-oriented design making its integration with other modules simple
  - Support of a wide range of recognition tasks making it dynamically configurable
  - Support of live mode and batch mode recognition
  - Recognizing discrete and continuous speech
  - BSD-style license

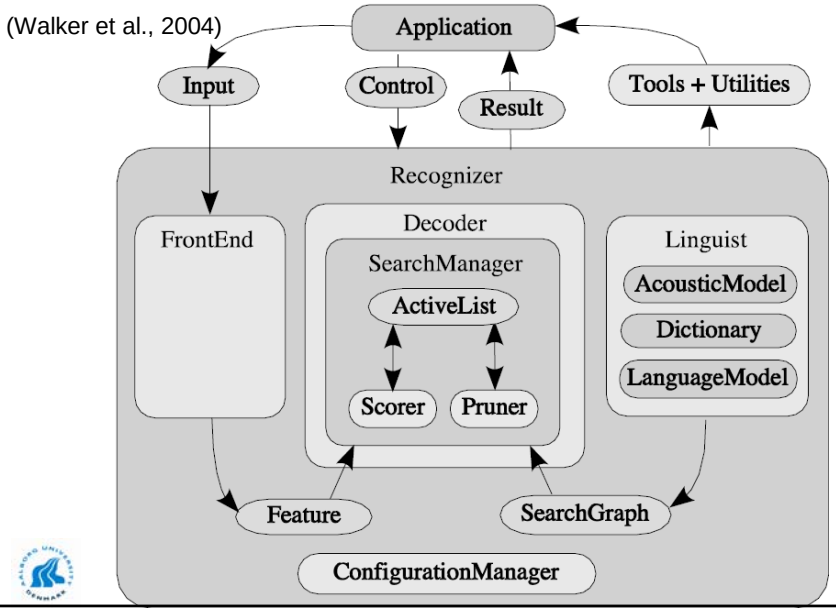
Sphinx-4 performance

(Walker et al., 2004)

Test	*WER*		*RT*		
	Sphinx-3.3	*Sphinx-4*	*Sphinx-3.3*	*Sphinx-4 (1 CPU)*	*Sphinx-4 (2 CPU)*
TI46 (11 words)	1.217	0.168	0.14	0.03	0.02
TIDIGITS (11 words)	0.661	0.549	0.16	0.07	0.05
AN4 (79 words)	1.300	1.192	0.38	0.25	0.20
RM1 (1000 words)	2.746	2.739	0.50	0.50	0.40
WSJ5K (5000 words)	7.323	7.174	1.36	1.22	0.96
HUB-4 (64000 words)	18.845	18.878	3.06	4.40	3.80

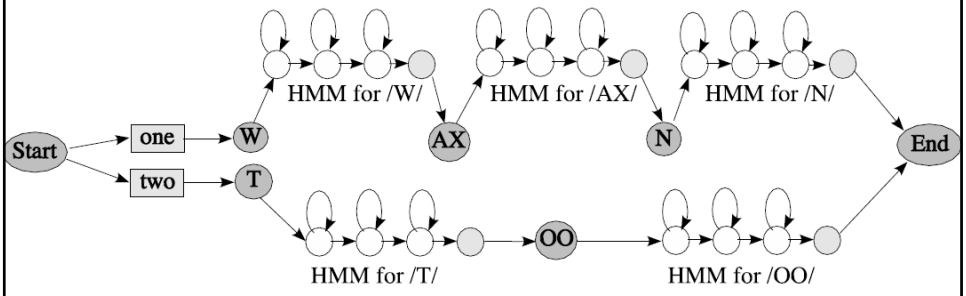
Sphinx-4 decoder framework

(Walker et al., 2004)



Example of search graph

(Walker et al., 2004)



Hello world application

- Robot control

```
#JSGF V1.0;
/**
 * JSGF Robot Grammar for Hello World example
 */
grammar robot;

public <move> =    (LIFT ARM | STEP FORWARD | SIT
   DOWN | ENTER STAIRS | FETCH THE CUP) * ;
```

Summary

- Types of speech recongizers
  - IWR
  - Rule grammar
  - N-gram
- Acoustic modelling
- Sphinx-4

- *Java Speech Grammar Format* (JSGF)
- http://java.sun.com/products/java-media/speech/forDevelopers/JSGF/index.html
- A rule expansion followed by the asterisk symbol indicates that the expansion may be spoken *zero or more times*. The asterisk symbol is known as the Kleene star (after Stephen Cole Kleene, who originated the use of the symbol). For example,
  - <command> = <polite>* don't crash;
- allows a user to say things like "please don't crash", "oh mighty computer please please don't crash", or to ignore politeness with "don't crash"