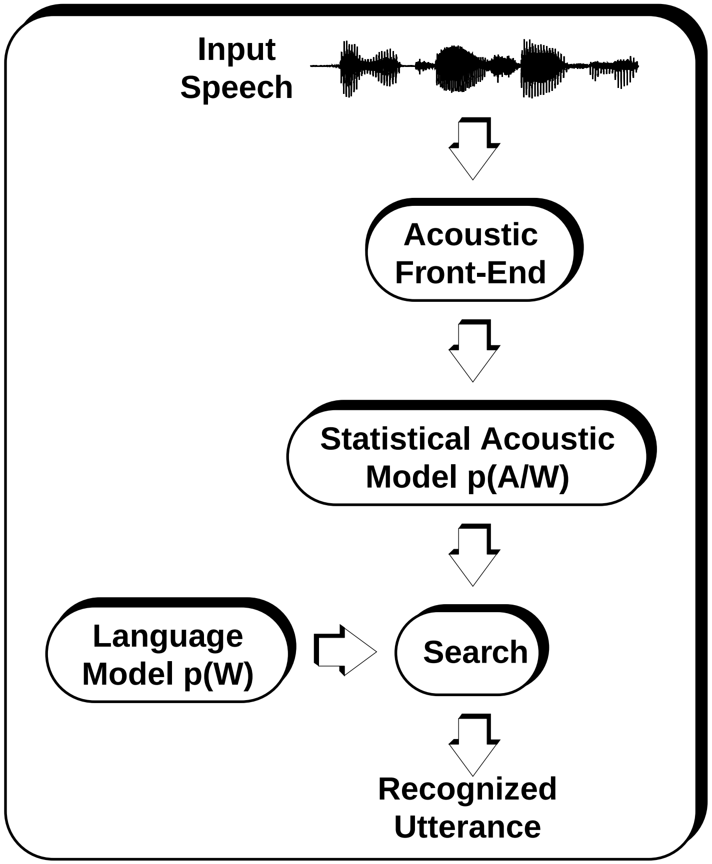
Generalized Hierarchical Search in the ISIP ASR System

Bohumir Jelinek, Feng Zheng, Naveen Parihar,
Jonathan Hamaker, Joseph Picone

11/07/2001
http://www.isip.msstate.edu/publications/conferences/asilomar/2001/presentation.pdf

Speech Recognition Problem Formulation



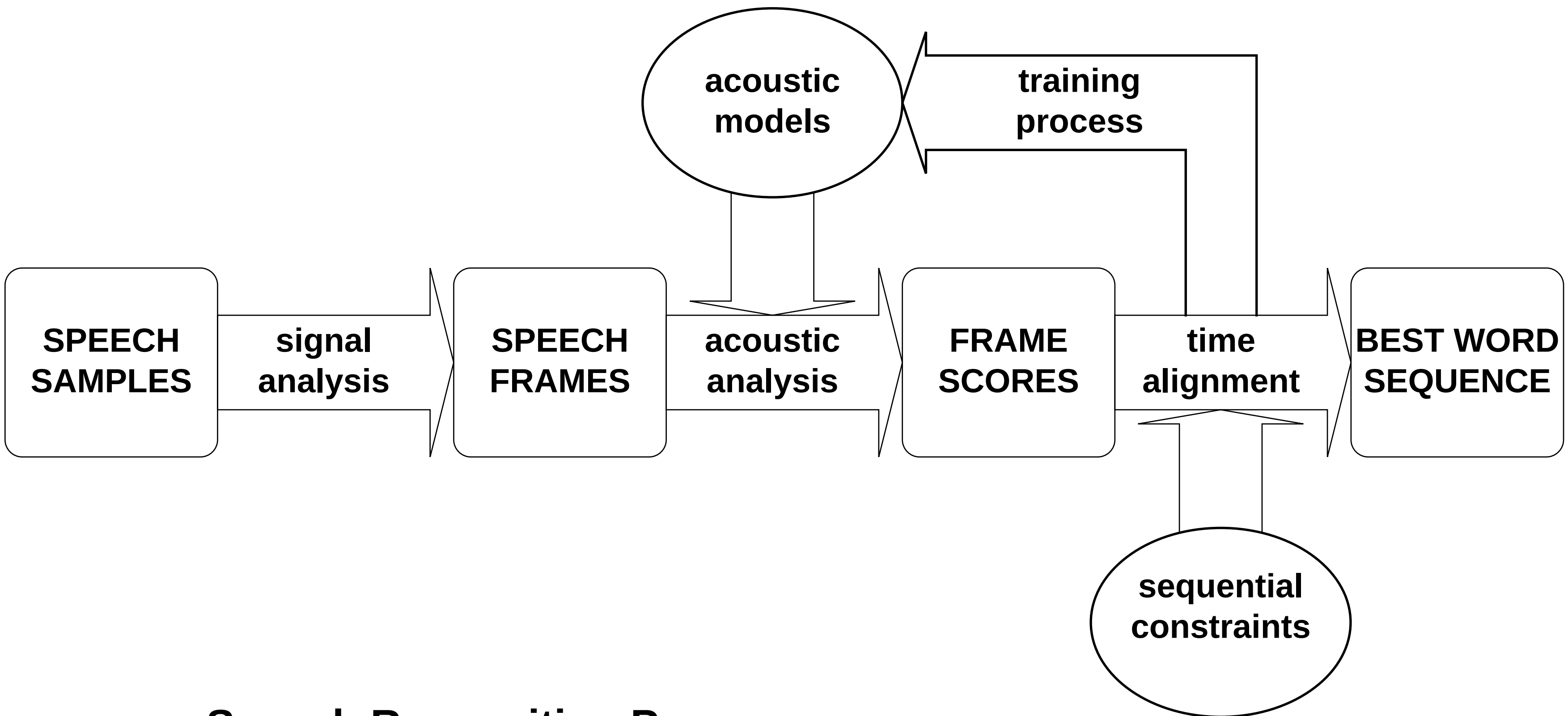
Bayesian framework:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\ p(W/A)$$

$$= \underset{W}{\operatorname{argmax}}\ p(A/W)\,p(W)$$

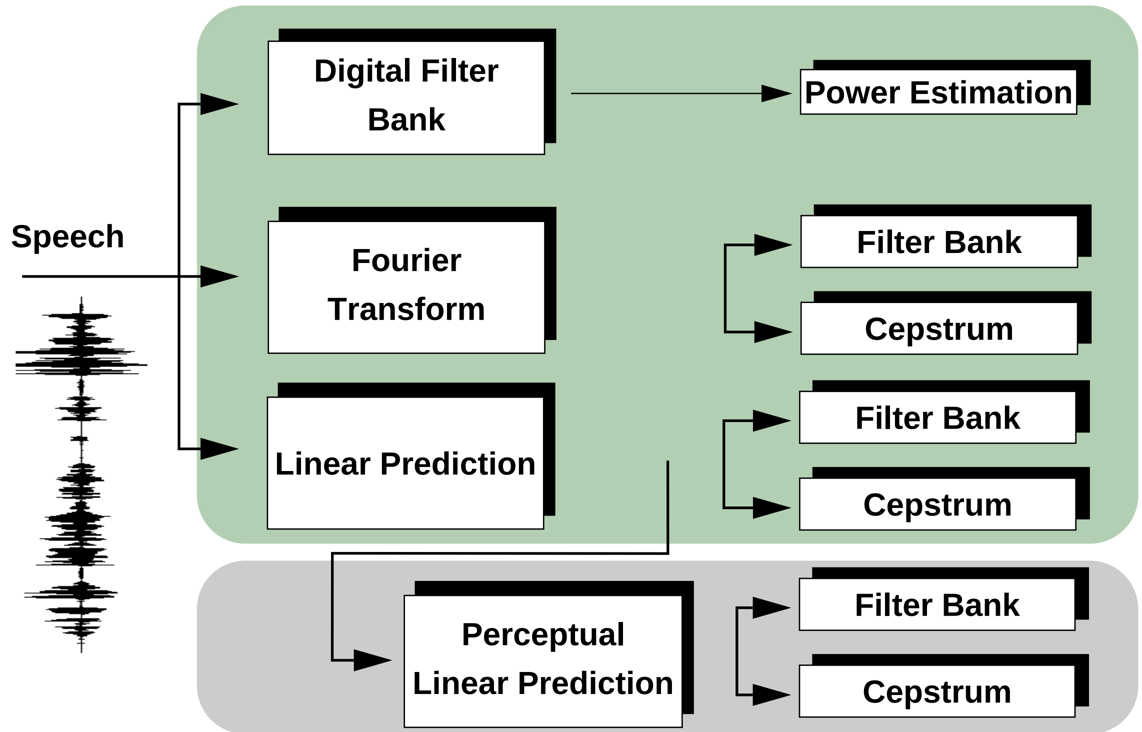
- **P(A|W)... acoustic component**
- **P(W)... language model component**

Block Diagram of Speech Recognizer



Speech Recognition Process

Front-End



- MEL-frequency cepstra coefficients + energy
- added delta and acceleration

Search

Find the most likely word sequence given acoustic and linguistic data.

Example: enumerative search for digit string recognition

- **10 word vocabulary, 6 digit strings**
- **10^6 possible paths**

Efficient decoder modifies search space:

- **reduction**
- **transformation**
- **suboptimal decisions**

Search Space Specification

N-gram language models:

$P(W_n) = P(W_n \mid W_{n-1}, W_{n-2} \ldots W_{n-N})$

- Unigrams
- Bigrams
- Trigrams

Method of generation:

- counting of word sequence occurrences in large text corpus

Network grammars:

Specify full search space by

- acceptable number sequences
- possible questions and answers
- correct sentence syntax

Method of generation:

- expert knowledge

Search Space Complexity

How to evaluate the difficulty of a recognition task?

Perplexity is entropy based measure of the task complexity:

$$PP = 2^{LP}$$

where

$$LP = \lim_{n = \infty} -\frac{1}{n} \sum_{i = 1}^{n} \log Q(w_i | w_1, \ldots, w_{i-1})$$

Perplexity represents average number of words that can follow any given word (branching factor)

Search Space Reduction

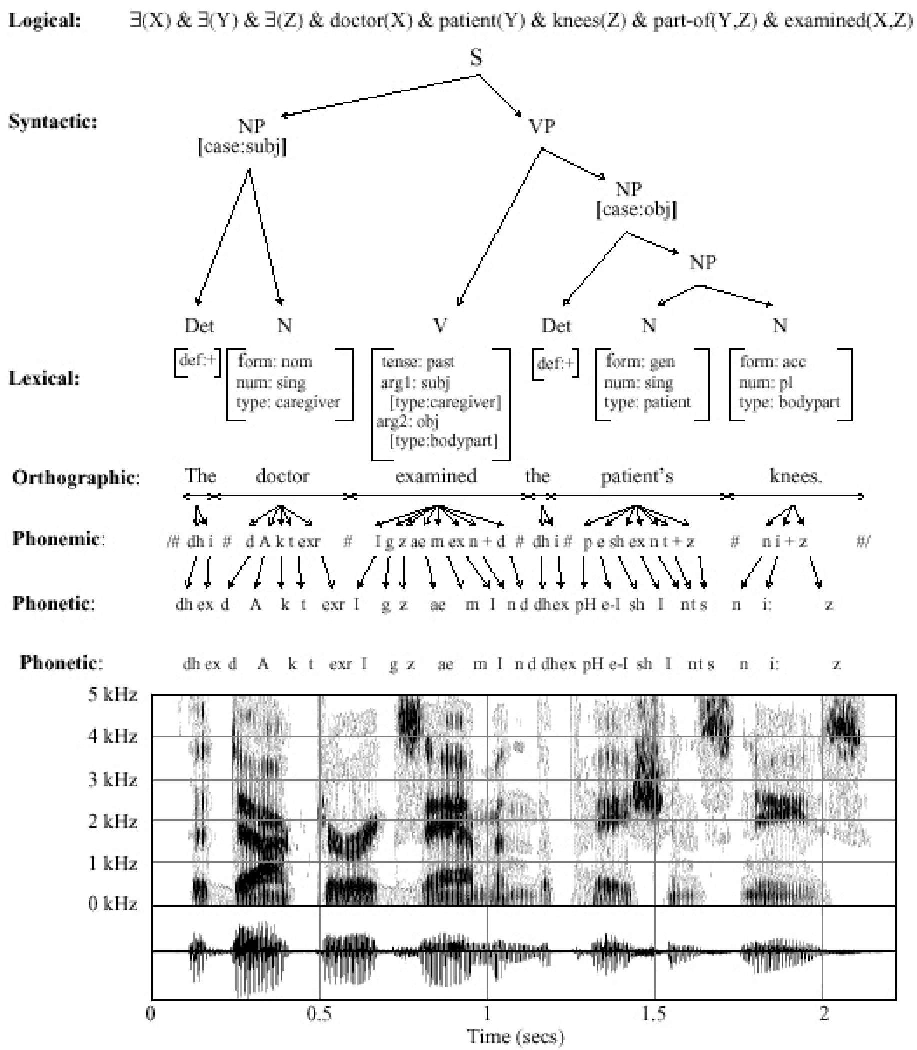
1.) Search space is constrained by the grammar

- **N-gram language model**
- **network grammar**

2.) During the search for the best hypothesis allow only:

- **one best scored path coming to particular point in the search space (Viterbi pruning)**
- **reliably good hypothesis (beam pruning)**
- **certain maximum number of hypothesis (active instances pruning)**
- **hypothesize phonemes common for many words only once (lexical trees)**

Hierarchical Search



Goal:

- **incorporate as many knowledge sources as possible**
- **allow as much flexibility as possible**

Advantage:

- **ability to flexibly incorporate multiple knowledge sources improves performance**

Generalized Hierarchical Search

Comparison of generalized and standard approach:

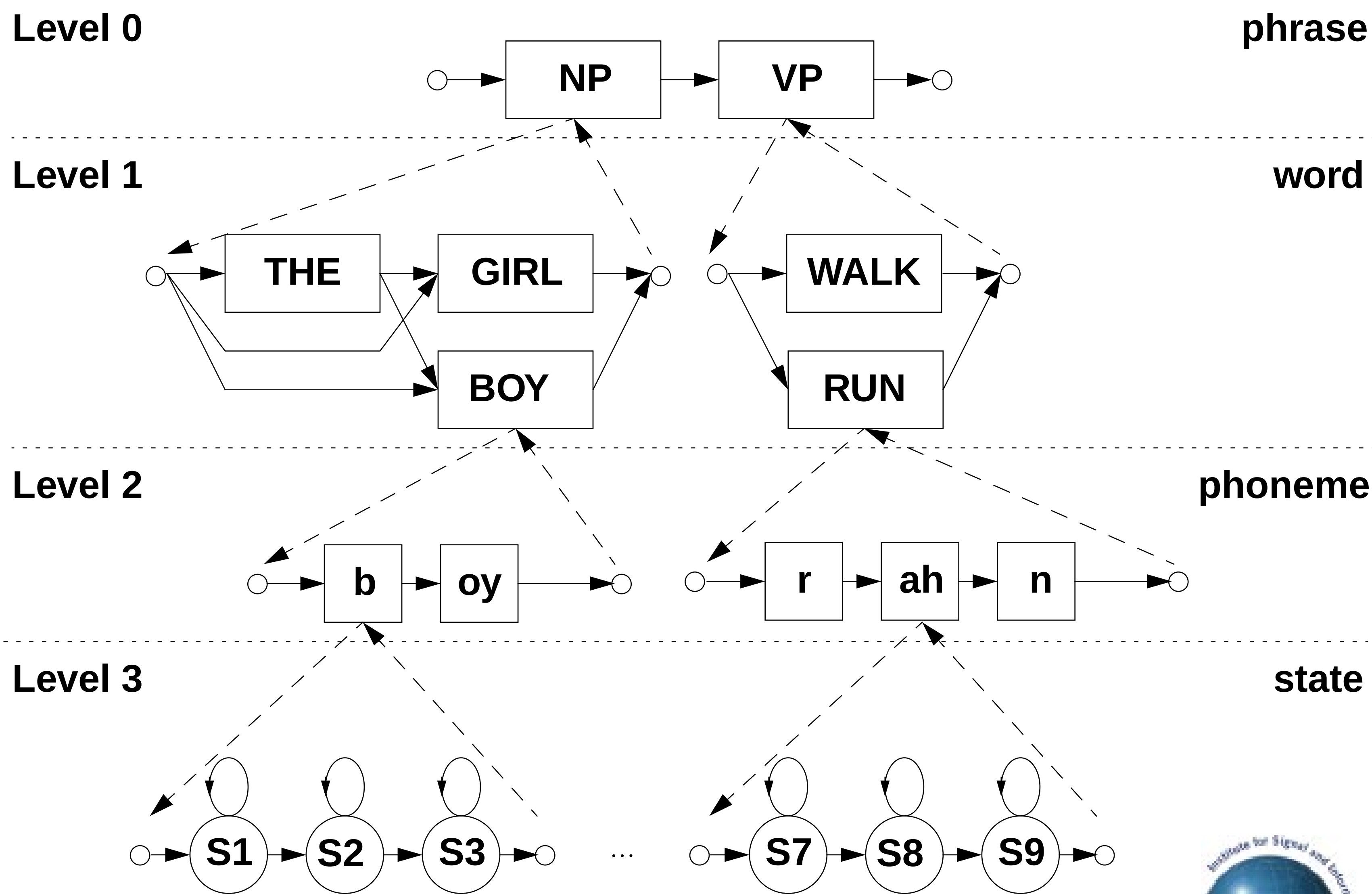
Generalized Hierarchical Search

- **general graphical specification**
- **allow any number of independent levels**
- **allow unlimited size context dependency at any level**
- **ability to change a search structure without changing the code**

Standard LVCSR Systems

- **N-gram based**
- **three levels: word, phone and state**
- **allow one left and one right context at phone level (triphone based)**
- **highly tuned to particular task, hard to modify structure**

Graph Representation



Context Dependency

Pronunciation of phoneme depends on the surrounding phonemes (allophone concept), pronunciation of word depends on the surrounding words (pronunciation modeling)

For general graphical representation it means:

- **each vertex of the search graph has several subgraphs specific for different contexts**
- **allow context dependency with unlimited depth at any level**

Advantages:

- **direct support of the pronunciation modeling and ability to implement N-gram language model using left context**

Parameter Sharing

Application of parameter sharing is necessary after we increase the number of system parameters (e.g. introduce full context dependent model set)

State tying:

- **if several states have similar parameters, we can force them to be identical**
- **if some state has not enough training data, we can force it to be identical with other state**

Technology for state tying:

- **phonetic decision trees (phonetic questions comes from expert knowledge source), state tying results in model clustering (several context dependent phones have the same physical model)**

Algorithmic Issues - Path Marker

Path Marker represents a dynamic component of the search algorithm.

Path Marker holds the following information:

- current location in the search space and graph vertices visited before
- backpointer - pointer to the previous trace
- path score
- frame when the trace was generated

Path Marker enables:

- search path generation and backtracking

Algorithmic Issues - History and Search Node

History (which is a component of trace) holds:

- **current location in the search space and graph vertices visited before at all higher levels**
- **for context dependent levels history do not stores also context vertices**
- **enables to propagate traces up and down through the levels of a graph structure**

Search Node (item in the vertex of the search graph) holds:

- **list of the paths that arrived in this vertex**
- **enables to do a Viterbi pruning: If more traces with the same history arrived in this vertex, keep only N (usually one) with the best score**

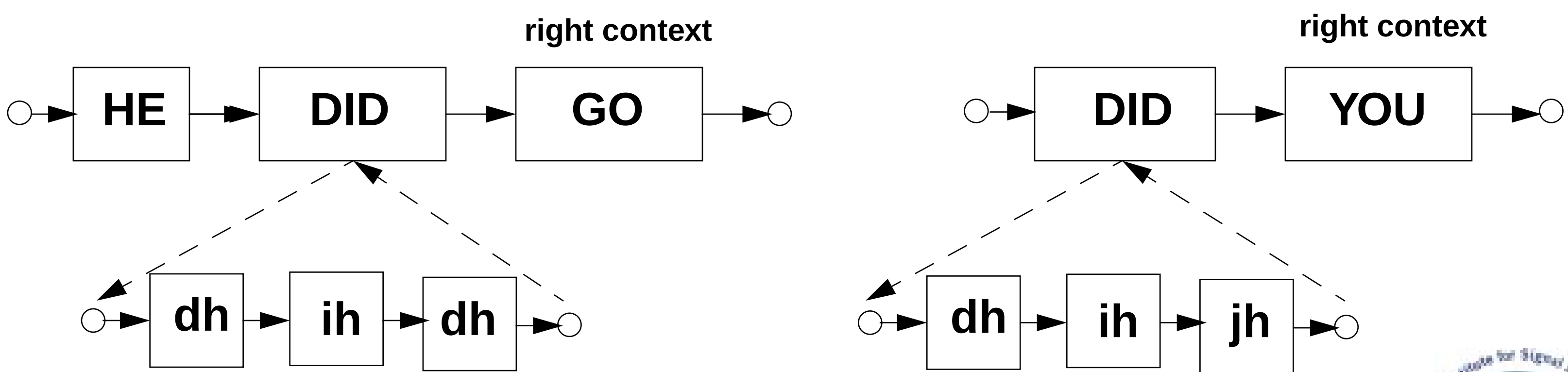
Algorithmic Issues - Context Dependency

Models of the symbols at any level can be either context-dependent or context-independent.

In case of context dependency, next lower level contains subgraphs that are specific for a particular contexts.

Mapping of the contexts to the index of the model at the next lower level is stored in context mapping hash table.

Word level with right context dependency of depth one:



Tidigits Database Results

TIDIGITS

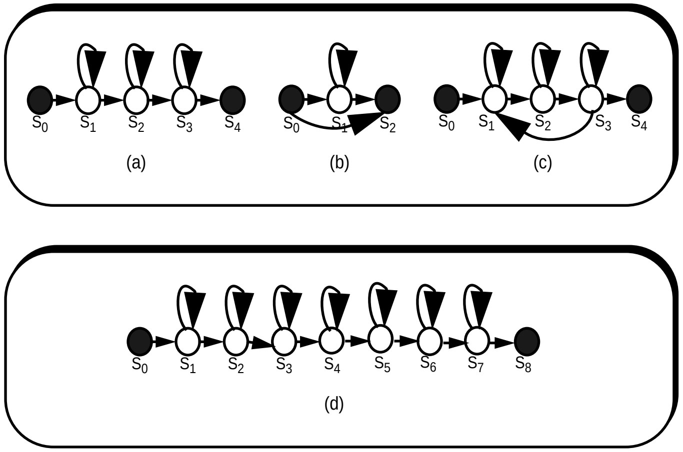
- 25 thousand digit sequences, studio quality data
- 8 kHz, 16bit/sample

Models

- 16 mixture Gaussian continuous density HMMs
- cross-word triphone models
- word models

Results

- WER = 0.6% for triphone models
- WER = 0.4% for word models



HMM model topologies (a) typical triphone model, (b) short pause, (c) silence, (d) typical word model

RM Database Results

Resource Management

- prompted queries into database
- 1000 word vocabulary
- perplexity 60
- very low background noise conditions
- 16kHz, 16bit/sample

Models

- 6 mixture Gaussian continuous density HMM
- cross-word triphone
- bigram language model

Results:

- 3.4% WER

WSJ Database Results

Wall Street Journal (WSJ0)

- read news database
- 5000 word vocabulary
- perplexity 147
- 16kHz, 16bit/sample
- very low background noise conditions

Models

- 16 mixture Gaussian continuous density HMM
- cross-word triphone
- bigram language model

Results:

- 8.3% WER

Tuning the System Parameters

Exp	Tied States	Word Error Rate				
		Feb89	Oct89	Feb91	Sep92	Average
13	1946	2.9%	4.4%	3.1%	5.5%	4.0%
14	3073	2.9%	3.9%	2.3%	5.2%	3.6%
15	3554	2.8%	3.5%	2.6%	5.2%	3.5%
16	4004	3.3%	4.4%	2.6%	5.9%	4.0%
17	4902	3.1%	3.7%	2.9%	6.3%	4.0%
18	8392	7.6%	9.5%	7.1%	12.3%	9.1%

Comparison of performance while tuning the number of tied states on RM (above) and WSJ0 database (below).

Number of Tied-States	State-Tying Thresholds			xRT	WER	Sub.	Del.	Ins.
	Split	Merge	Occup.					
1,882	650	650	1400	151	11.0%	8.0%	1.7%	1.2%
3,024	150	150	900	149	10.7%	8.0%	1.6%	1.1%
3,215	165	165	840	138	8.6%	6.8%	1.1%	0.7%
3,580	125	125	750	123	8.9%	6.7%	1.4%	0.8%
3,983	110	110	660	120	8.7%	6.6%	1.0%	1.1%
4,330	100	100	600	116	9.1%	6.5%	1.4%	1.2%

Conclusions

We have implemented Generalized Hierarchical Search algorithm in the ISIP ASR system. It employs a flexible and configurable multi-level search strategy capable of incorporating hierarchical knowledge sources with no changes to source code. It allows to incorporate higher-level knowledge sources such as discourse, part of speech, and understanding constraints to the speech recognition problem.

Future directions:

- **pronunciation modeling**
- **question answering**
- **discriminative acoustic classifiers - SVM**

References

1. N. Deshmukh, A. Ganapathiraju and J. Picone, *Hierarchical Search for Large Vocabulary Conversational Speech Recognition*, IEEE Signal Processing Magazine, vol. 16, no. 5, pp. 84-107, September 1999.

2. F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, Massachusetts, London, England, 1998.

3. J. Picone, “Tutorials”, *http://www.isip.msstate.edu/projects/speech/software/tutorials/*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, November 2001.