Return to Main

Objectives

Topolgy:
Transition Matrix
DTW Analogy
Alternatives

Duration Modeling:
State Duration Probabilities
Number of States

Training Schedules:
Simple Schedule
More Complex Schedules
Complex Models

On-Line Resources:
Training Workshop
Using HMMs
The HTK Book

LECTURE 26: PRACTICAL ISSUES

- Objectives:
  - Discuss common model topologies
  - Provide model design guidelines
  - Introduce typical training schedules

This lecture combines material from this paper:

J. Picone, "Continuous Speech Recognition Using Hidden Markov Models", *IEEE ASSP Magazine*, vol. 7, no. 3, pp. 26-41, July 1990.

and information found in this workshop:

> *Speech Recognition System Training Workshop*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, 39762, USA, January 2002.

ECE 8463: FUNDAMENTALS OF SPEECH RECOGNITION

Professor Joseph Picone
Department of Electrical and Computer Engineering
Mississippi State University

email: picone@isip.msstate.edu
phone/fax: 601-325-3149; office: 413 Simrall
URL: http://www.isip.msstate.edu/resources/courses/ece_8463

Modern speech understanding systems merge interdisciplinary technologies from Signal Processing, Pattern Recognition, Natural Language, and Linguistics into a unified statistical framework. These systems, which have applications in a wide range of signal processing problems, represent a revolution in Digital Signal Processing (DSP). Once a field dominated by vector-oriented processors and linear algebra-based mathematics, the current generation of DSP-based systems rely on sophisticated statistical models implemented using a complex software paradigm. Such systems are now capable of understanding continuous speech input for vocabularies of hundreds of thousands of words in operational environments.

In this course, we will explore the core components of modern statistically-based speech recognition systems. We will view speech recognition problem in terms of three tasks: signal modeling, network searching, and language understanding. We will conclude our discussion with an overview of state-of-the-art systems, and a review of available resources to support further research and technology development.

Tar files containing a compilation of all the notes are available. However, these files are large and will require a substantial amount of time to download. A tar file of the html version of the notes is available here. These were generated using wget:

> wget -np -k -m
> http://www.isip.msstate.edu/publications/courses/ece_8463/lectures/current

A pdf file containing the entire set of lecture notes is available here. These were generated using Adobe Acrobat.

Questions or comments about the material presented here can be directed to help@isip.msstate.edu.

LECTURE 26: PRACTICAL ISSUES

- Objectives:
  - Discuss common model topologies
  - Provide model design guidelines
  - Introduce typical training schedules

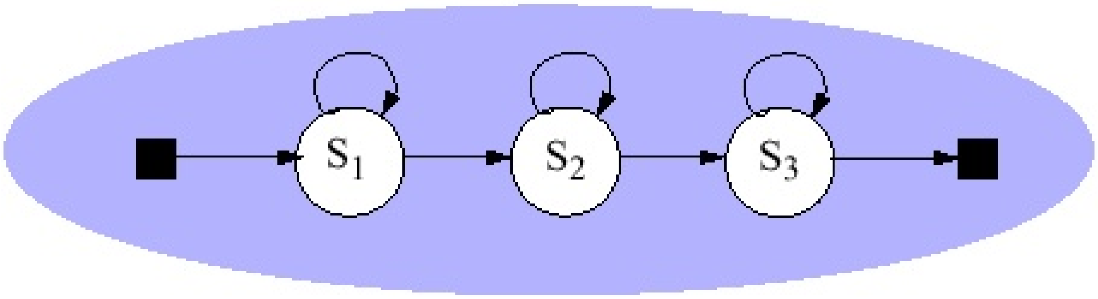
This lecture combines material from this paper:

> J. Picone, "Continuous Speech Recognition Using Hidden Markov Models", *IEEE ASSP Magazine*, vol. 7, no. 3, pp. 26-41, July 1990.

and information found in this workshop:

Speech Recognition System Training Workshop, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, 39762, USA, January 2002.

MODEL TOPOLOGY



Each phoneme has a transition and a state observation PDF

Transition Matrix:

$$\begin{array}{c} \text{From Entry State 0} \\ \text{From State 1} \\ \text{From State 2} \\ \text{From State 3} \\ \text{From Exit State 4} \end{array} \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & \alpha & 1-\alpha & 0 & 0 \\ 0 & 0 & \beta & 1-\beta & 0 \\ 0 & 0 & 0 & \delta & 1-\delta \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

State Observation PDF:

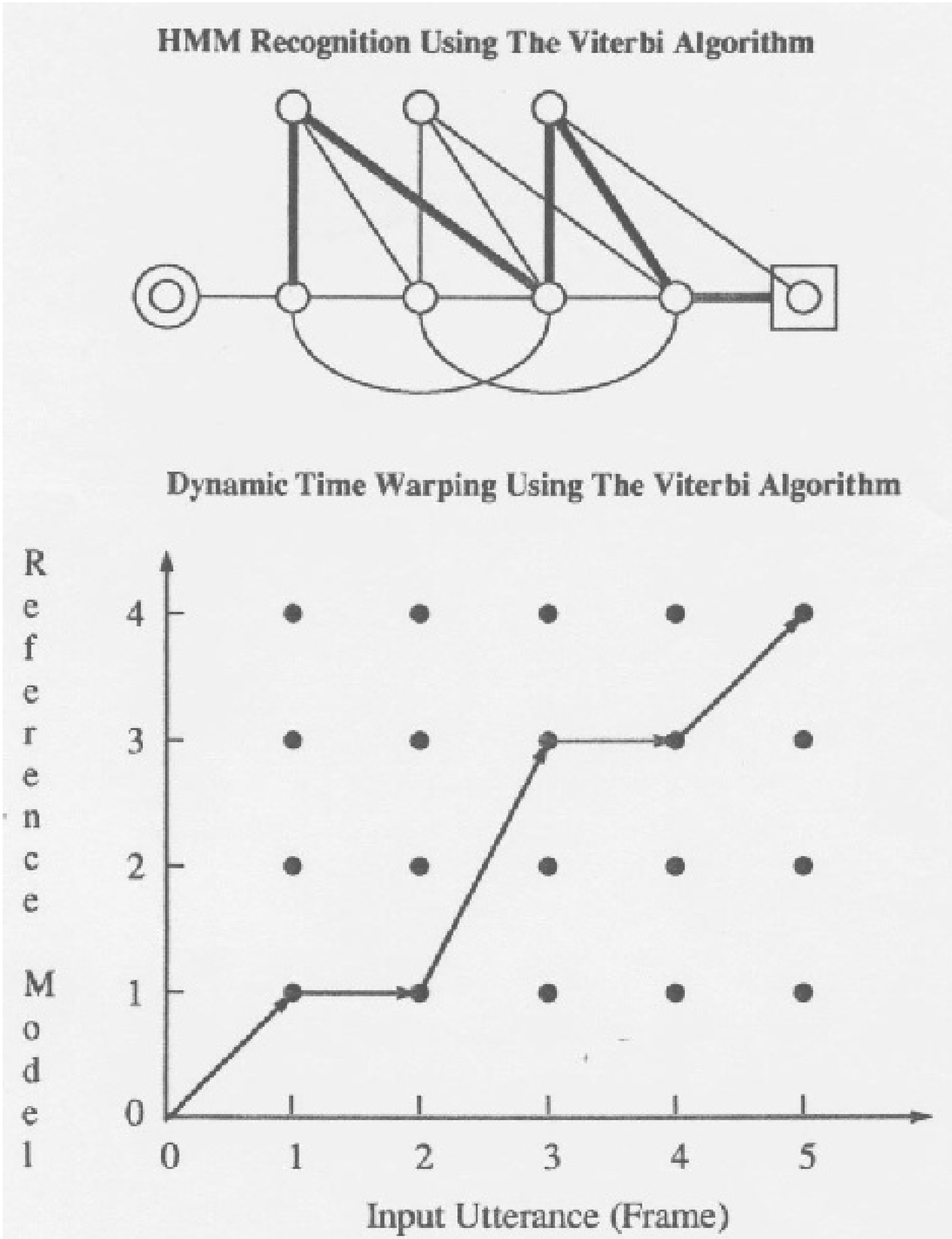
Multivariate Mixture Gaussians for states 1-3.

Parameters: mean vector, diagonal covariance matrix, mixture weights.

What is the total number of parameters to estimate per state?

10 mixtures (39 for mean vector + 39 variances) + 10 mixture weights = 790

THE DTW ANALOGY



- Note the similarity to DTW with slope constraints

ALTERNATIVE MODEL TOPOLOGIES

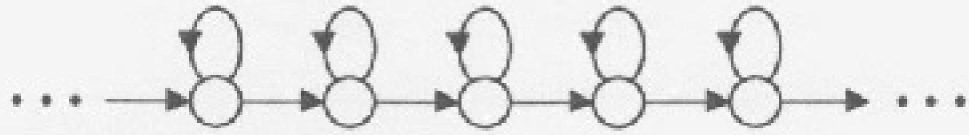


Figure 8(a). A simple progressive HMM topology. In general, the duration probability density function at a state has an exponential behavior.

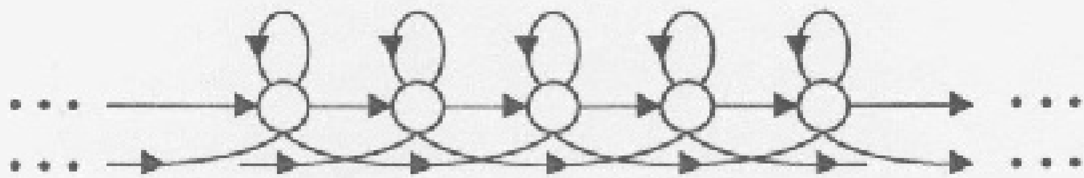


Figure 8(b). The Bakis topology (a progressive model with skip states).

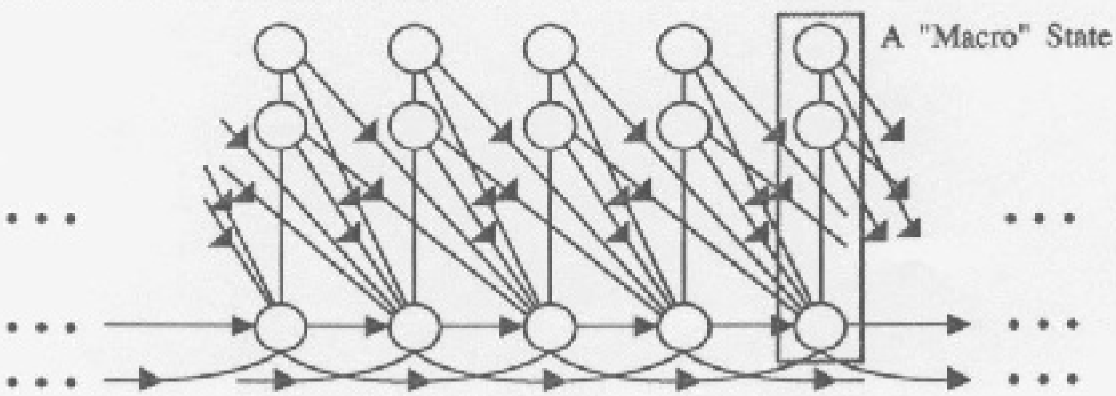


Figure 8(c). A finite duration topology. This topology is most analogous to DTW.

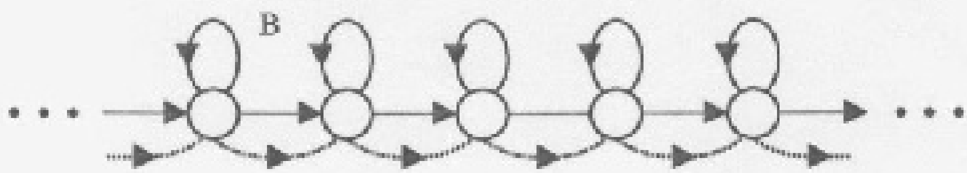


Figure 8(d). A fenonic baseform topology. The dashed line indicates a transition that produces no output.

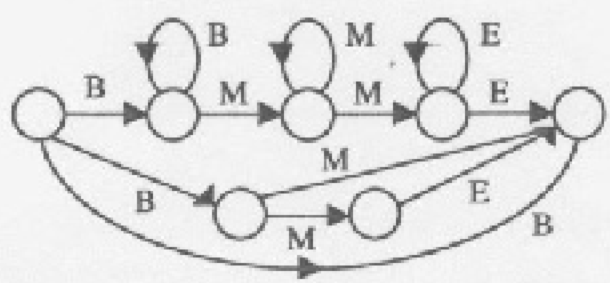


Figure 8(e). A modified fenonic baseform with tied transitions. Transitions in the same group share output probabilities.

- Note the similarity to DTW with slope constraints

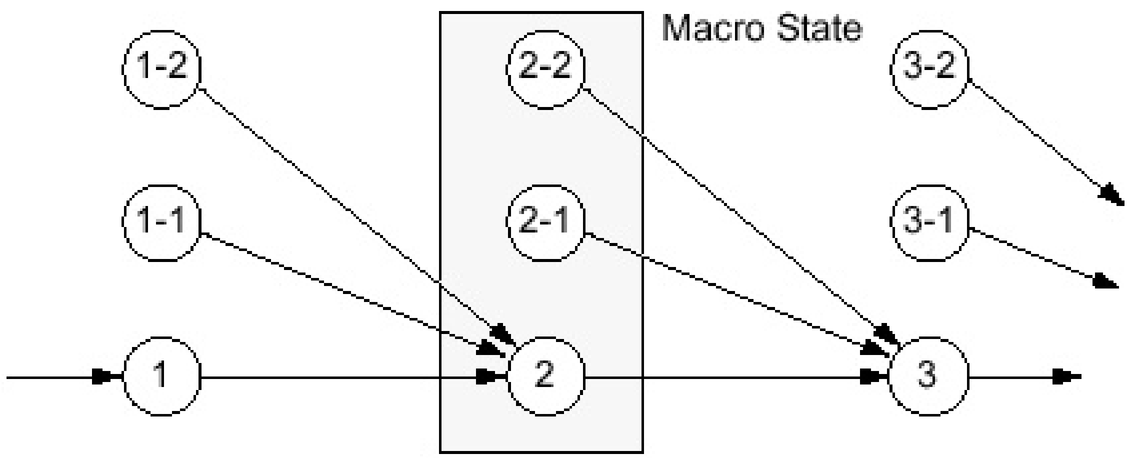
STATE DURATION PROBABILITIES

Recall that the probability of staying in a state was given by an exponentially-decaying distribution:

$$P(\bar{O}|Model, q_1 = i) = P(\bar{O}, q_1 = i|Model)/P(q_1 = i) = a_{ii}^{d-1}(1 - a_{ii})$$

This model is not necessarily appropriate for speech. There are three approaches in use today:

- Finite-State Models (encoded in acoustic model topology)



(Note that this model doesn't have skip states; with skip states, it becomes much more complex.)

- Discrete State Duration Models (D parameters per state)

$$P(\underline{d}_i = d) = \tau_d \qquad 1 \le d \le D$$

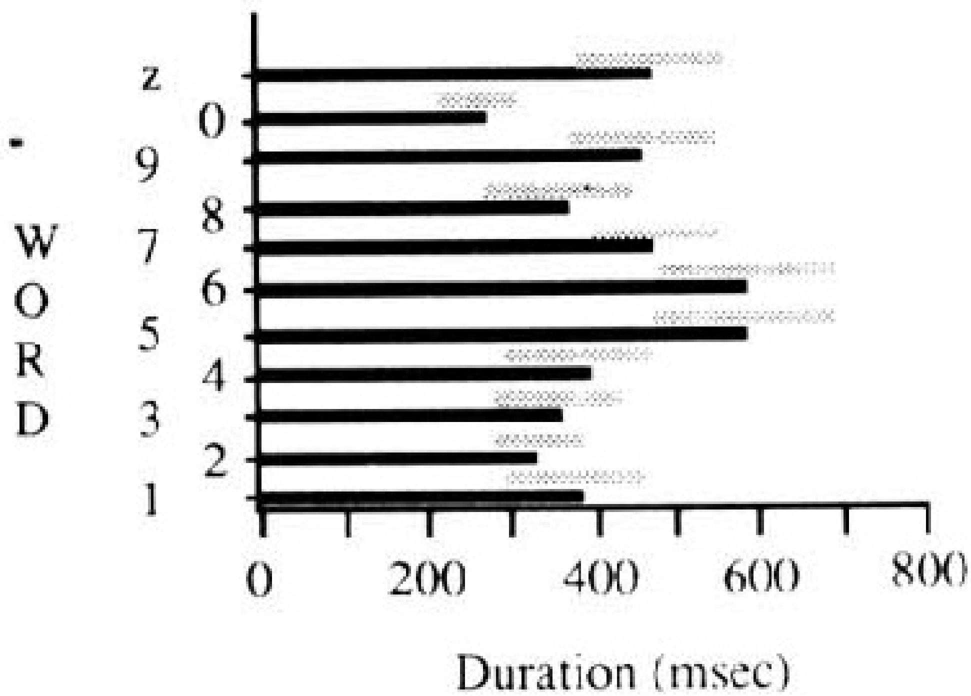
- Parametric State Duration Models (one to two parameters)

$$f(d_i) = \frac{1}{\sqrt{2\sigma_i^2}} \exp\left\{\frac{-\sqrt{2}|d|}{\sigma_i}\right\}$$

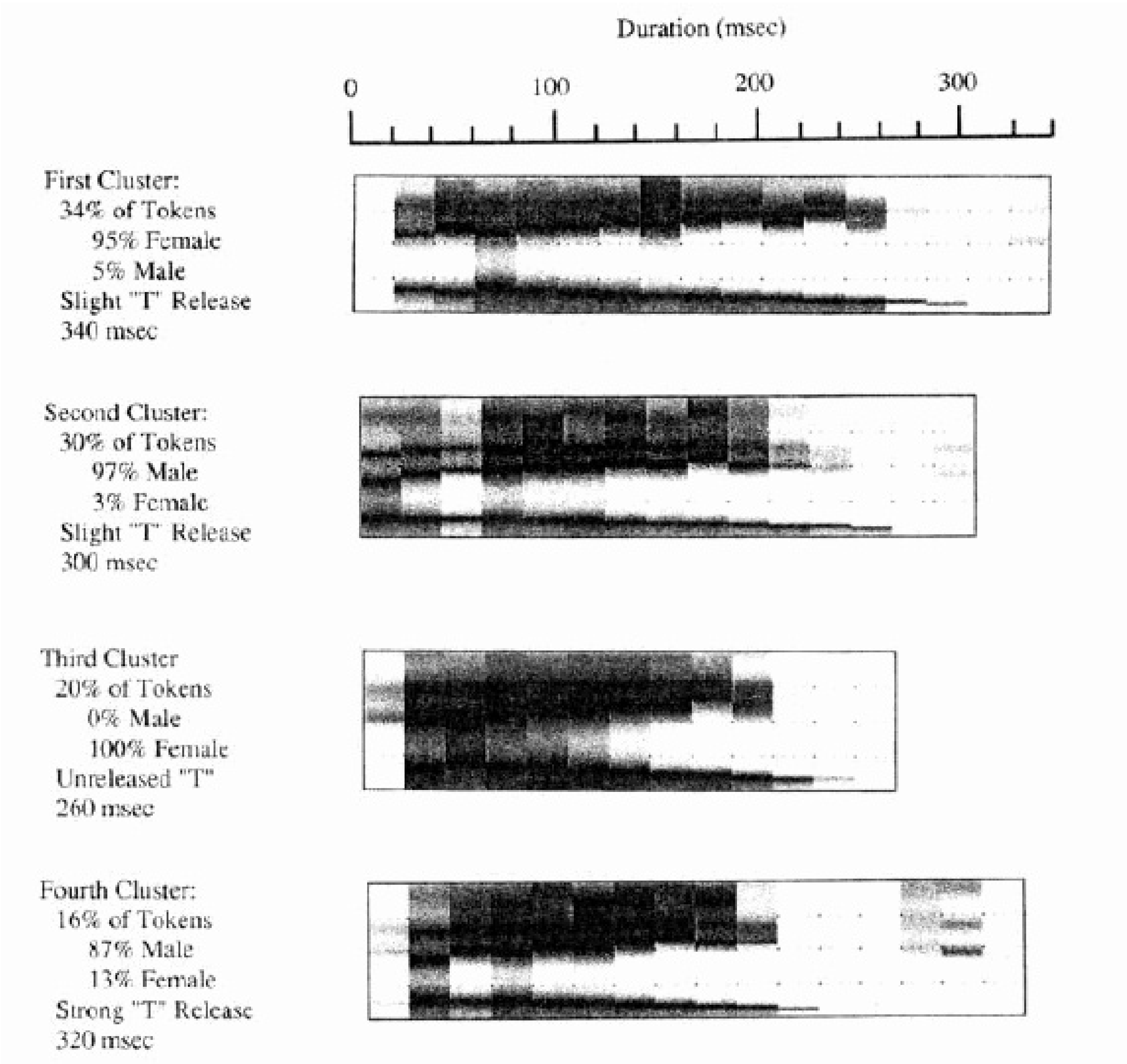
Reestimation equations exist for all three cases. Duration models are often important for larger models, such as words, where duration variations can be significant, but not as important for smaller units, such as context-dependent phones, where duration variations are much better understood and predicted.

DURATION CONSIDERATIONS AND CLUSTERING

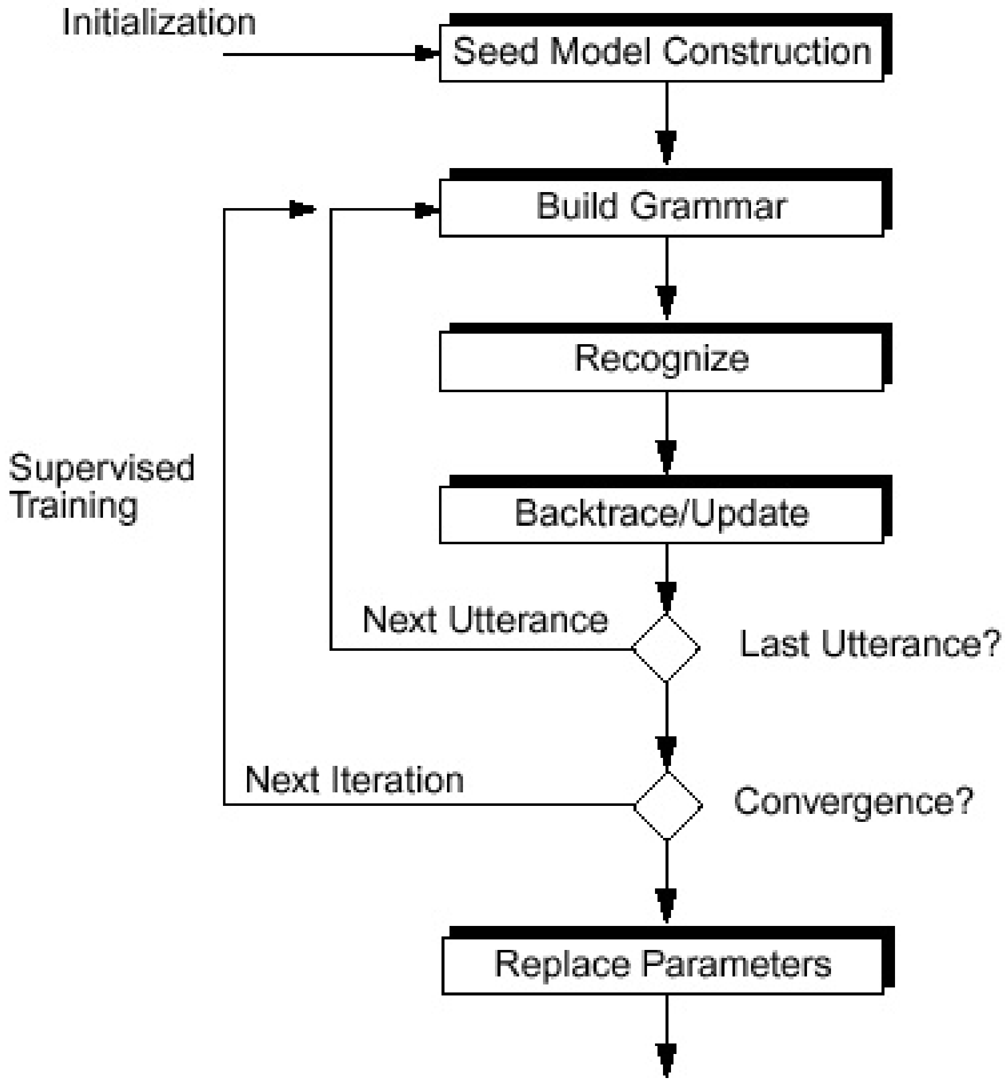
- For word model-based systems, we often will consider the duration of the word when assigning the initial number of states in the model:



- Clustering approaches can be used to learn pronunciation variants:



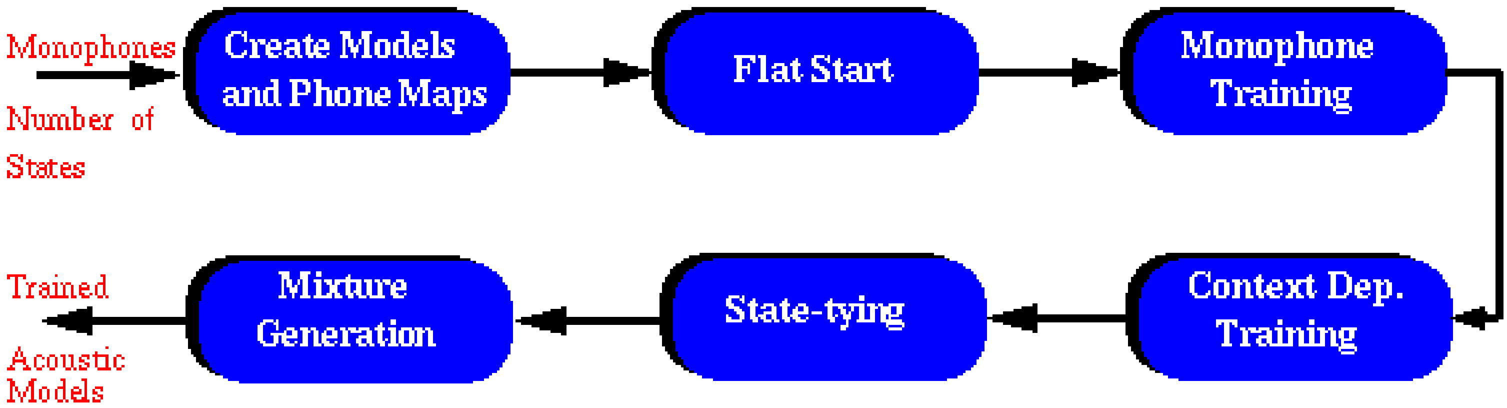
TYPICAL TRAINING SCHEDULES



Note that *a priori* segmentation of the utterance is not required, and that the recognizer is forced to recognize the utterance during training (via the build grammar operation). This forces the recognizer to learn contextual variations, provided the seed model construction is done "properly."

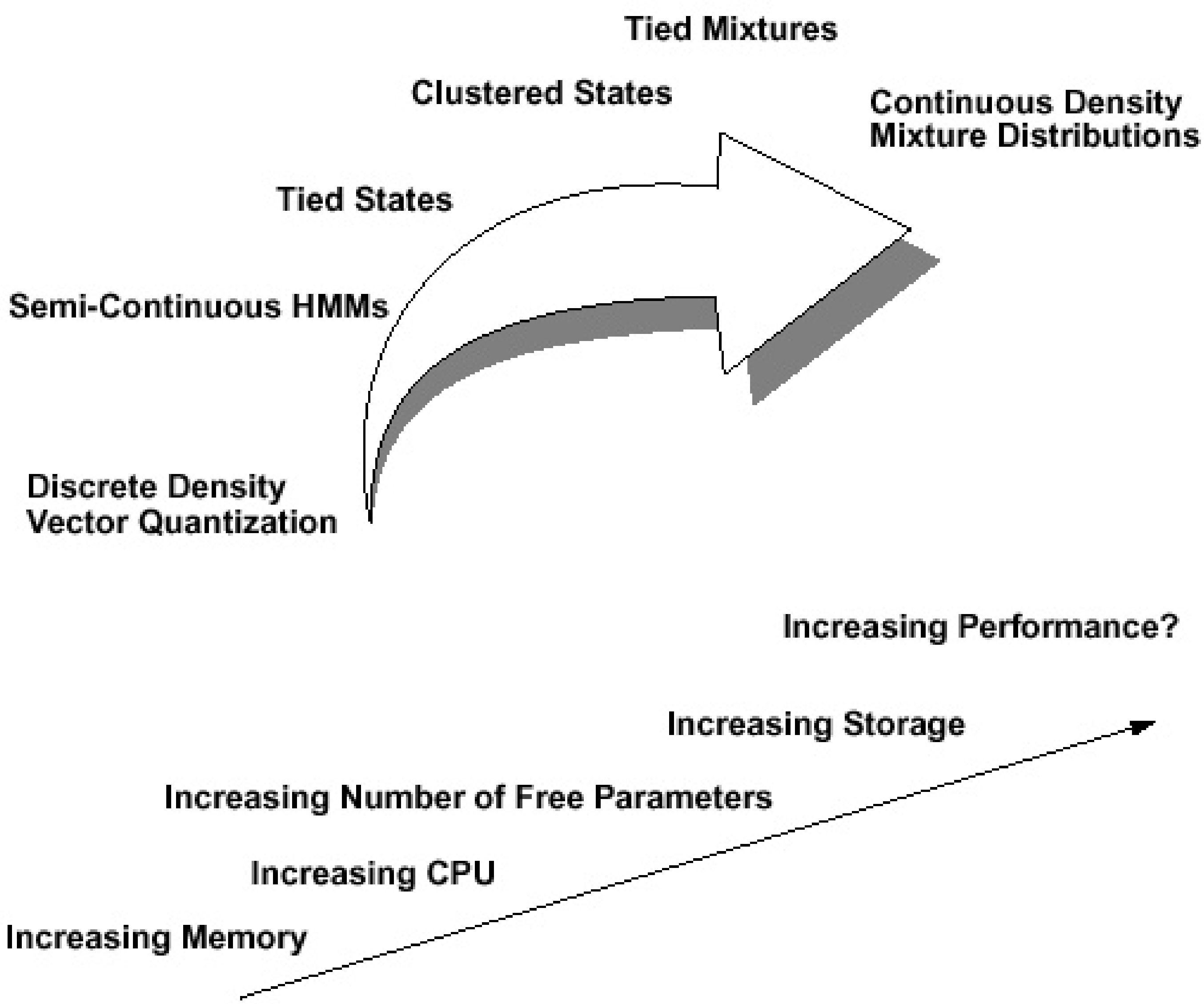
What about speaker independence?
Speaker dependence?
Speaker adaptation?
Channel adaptation?

MORE EXTENSIVE TRAINING SCHEDULES



- Phone-based HMM systems require a more extensive training process.
- Mixture generation is usually done last, and is performed using a cluster-splitting approach.
- Retraining after mixture generation is important.

MODEL COMPLEXITY



- Numerous techniques to robustly estimate model parameters; among the most popular is deleted interpolation:

$$\boldsymbol{A} = \varepsilon_t \boldsymbol{A}_t + (1 - \varepsilon_t) \boldsymbol{A}_u$$

$$\boldsymbol{B} = \varepsilon_t \boldsymbol{B}_t + (1 - \varepsilon_t) \boldsymbol{B}_u$$

WORKSHOP PROGRAM

This workshop is the first in a series of training workshops intended to train entry-level researchers on the details of building speech recognition systems using the ISIP system. Because seating is limited, preference will be given to graduate students planning on using the ISIP system in their research. ISIP will subsidize the travel expenses for students planning to attend the workshop. Please see the registration page for more details.

We are now in the second year of our project to build a public domain system that is competitive with state of the art. An overview of our mission to develop Internet Accessible Speech Recognition Technology is available on the web. There is also a web site devoted to dissemination of information, and a mailing list used to promote discussions within our user community.

A preliminary agenda for the workshop is shown below. If you have comments or suggestions about the agenda, please feel free to contact us. We look forward to seeing you at SRSTW'00.

Location:

Morning Lectures:	**Simrall Auditorium, Simrall Enginering Building**
Afternoon Laboratories:	**ELI/ Giles Rooms, Mitchell Memorial Library**

Schedule:

DAY	TIME	SESSION	DESCRIPTION	PRESENTERS
SUNDAY MAY 12	15:00 - 17:00	1.1: Orientation	Registration, Computers	Staff
	17:00 - 19:00	1.2: Reception	BBQ, Frisbee, softball	Staff
MONDAY MAY 13	08:30 - 10:00	2.1: Introduction	Welcome Speech Recognition Foundation Classes	Joe Picone
	10:00 - 10:30	**BREAK**		
	10:30 - 12:00	3.1: Classes	DSP and Templates Data Structures Algorithms, I/O, and Audio	Jie Zhao
	12:00 - 13:30	**LUNCH (SELF-PAY)**		
	13:30 - 15:00	3.2: IFC Programming	Math Classes Templates Data Structures	Jie Zhao
	15:00 - 15:30	**BREAK**		
	15:30 - 17:00	3.3: Algorithms	Basic DSP Buffers and I/O Audio Front-End	Jie Zhao
TUESDAY MAY 14	08:30 - 10:00	4.1: Signal Processing	Measurements Statistical Modeling Typical Implementations	Joe Picone
	10:00 - 10:30	**BREAK**		
	10:30 - 12:00	5.1: Transformations	Algorithms and Recipes Front-End Overview A Graphical User Interface	Shivali Srivastava
	12:00 - 13:30	**LUNCH (SELF-PAY)**		
	13:30 - 15:00	5.2: Evaluations	Generating Features Running Evaluations Adding New Algorithms	Shivali Srivastava
	15:00 - 15:30	**BREAK**		

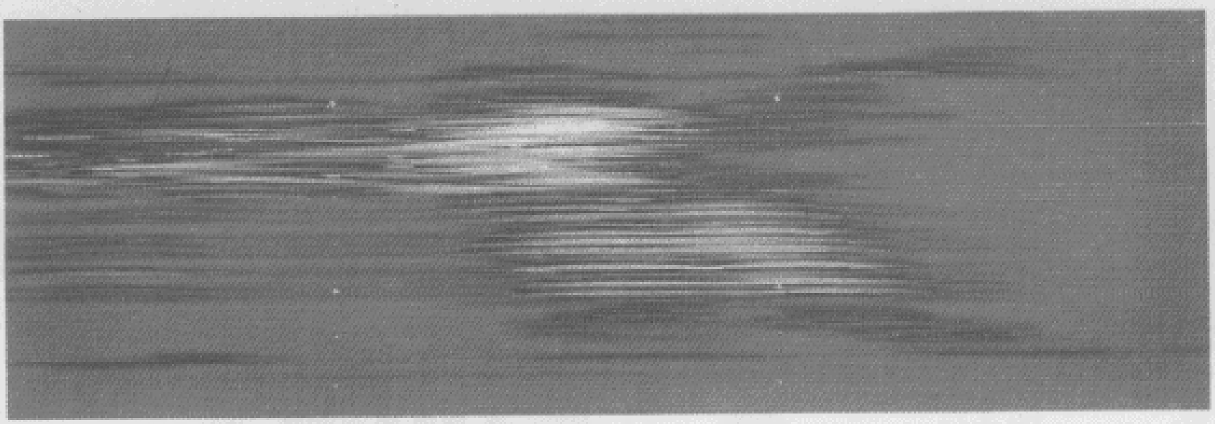
	15:30 - 17:00	5.3: TIDIGITS	Building Digit Recognizers Evaluating New Front-Ends Improving Performance	Shivali Srivastava
WEDNESDAY MAY 15	08:30 - 10:00	6.1: Acoustic Modeling	Hidden Markov Models Training Modeling Context	Mark Ordowski
	10:00 - 10:30	**BREAK**		
	10:30 - 12:00	7.1: Model Design	Model Initialization Context-Independent Context-Dependent	Ram Sundaram
	12:00 - 13:30	**LUNCH (SELF-PAY)**		
	13:30 - 15:00	7.2: Training	Baum-Welch Training State Tying Mixture Generation	Ram Sundaram
	15:00 - 15:30	**BREAK**		
	15:30 - 17:00	7.3: Alphadigits	Hierarchical Systems Lexicons and Pronunciations Adding Language Models	Ram Sundaram
	17:30 - 21:00	**ROAD TRIP (BRING YOUR CAMO!)**		
THURSDAY MAY 16	08:30 - 10:00	8.1: Language Models	Networks and N-grams Smoothing and Pruning Search Algorithms	Joe Picone
	10:00 - 10:30	**BREAK**		
	10:30 - 12:00	9.1: Search	Viterbi Beam Search N-gram Decoding Word Graphs and Rescoring	Aravind Ganapath.
	12:00 - 13:30	**LUNCH (SELF-PAY)**		
	13:30 - 15:00	9.2: N-gram Models	Generating N-gram Models Evaluating Complexity Pruning Language Models	Jie Zhao
	15:00 - 15:30	**BREAK**		

	15:30 - 17:00	9.3: Rescoring	Graph Generation and Compaction Acoustic Rescoring Language Model Rescoring	Jie Zhao
	19:00 - 21:00	**SRSTW'02 BASKETBALL TOURNAMENT**		
FRIDAY MAY 17	08:30 - 10:00	10.1: LVCSR Systems	Typical LVCSR Systems Multi-Pass Systems Open Discussion	Joe Picone
	10:00 - 10:30	**BREAK**		
	10:30 - 12:00	11.1: Real Speech	Switchboard Issues in Training Efficient Decoding	Aravind Ganapath.
	12:00 - 13:30	**LUNCH (SELF-PAY)**		
	13:30 - 15:00	11.2: Building Systems	Preparing Data Acoustic Training Word Graph Generation	Ram Sundaram
	15:00 - 15:30	**BREAK**		
	15:30 - 17:00	11.3: Switchboard	Decoding Scoring Optimizations	Ram Sundaram
	19:00 - 22:00	**BANQUET**		

1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - 10 - 11 - 12 - 13 - 14 - 15 - 16

Continuous Speech Recognition Using Hidden Markov Models

Joseph Picone



Stochastic signal processing techniques have profoundly changed our perspective on speech processing. We have witnessed a progression from heuristic algorithms to detailed statistical approaches based on iterative analysis techniques. Markov modeling provides a mathematically rigorous approach to developing robust statistical signal models. Since the introduction of Markov models to speech processing in the middle 1970s, continuous speech recognition technology has come of age. Dramatic advances have been made in characterizing the temporal and spectral evolution of the speech signal. At the same time, our appreciation of the need to explain complex acoustic manifestations by integration of application constraints into low level signal processing has grown. In this paper, we review the use of Markov models in continuous speech recognition. Markov models are presented as a generalization of its predecessor technology, Dynamic Programming. A

unified view is offered in which both linguistic decoding and acoustic matching are integrated into a single optimal network search framework.

Up | Home | Site Map | What's New | Projects | Publications
Speech | Administration | About Us | Search | Contact