

# The State of the Art in ASR (and Beyond?)

Steve Young

Microsoft Cambridge

and

Cambridge University Engineering Department

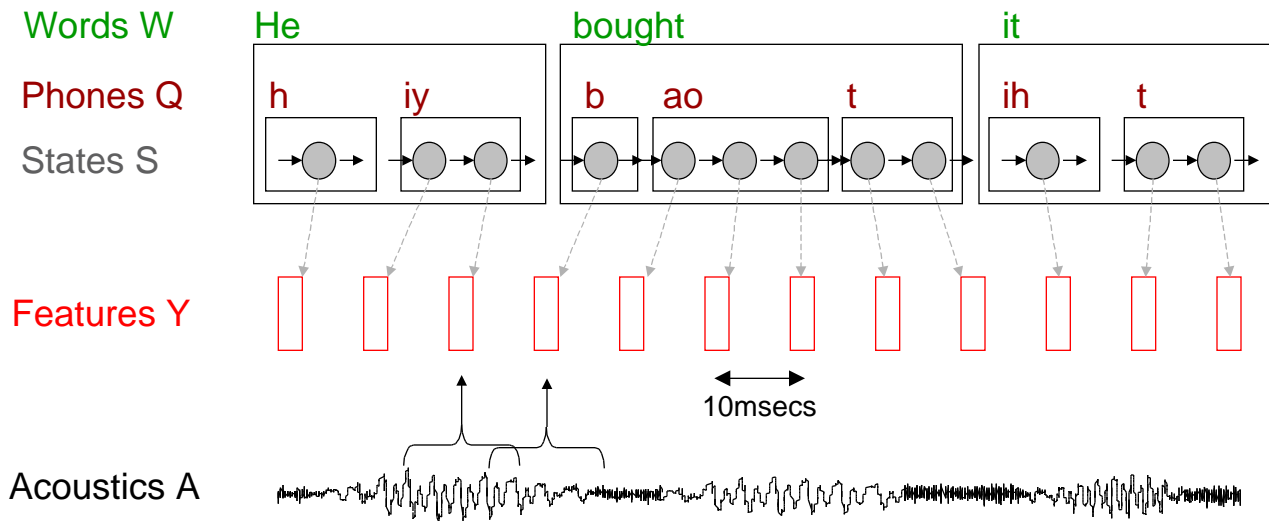
Speech Vision and Robotics Group



Microsoft

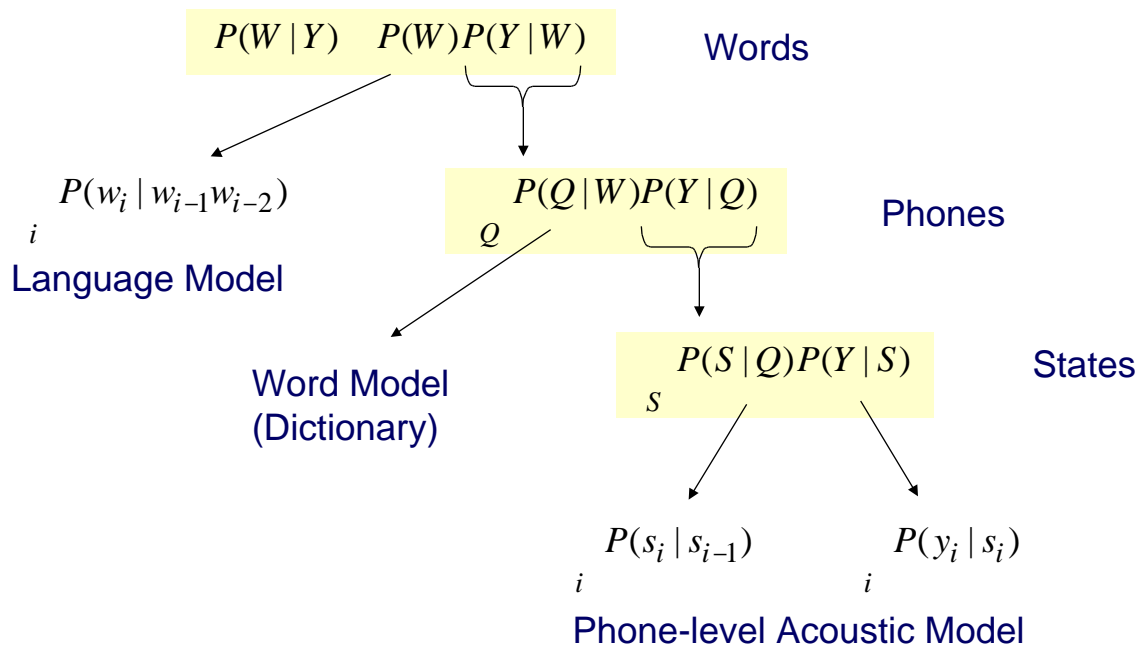


# Standard Model for Continuous Speech



Microsoft

Or if you prefer .....



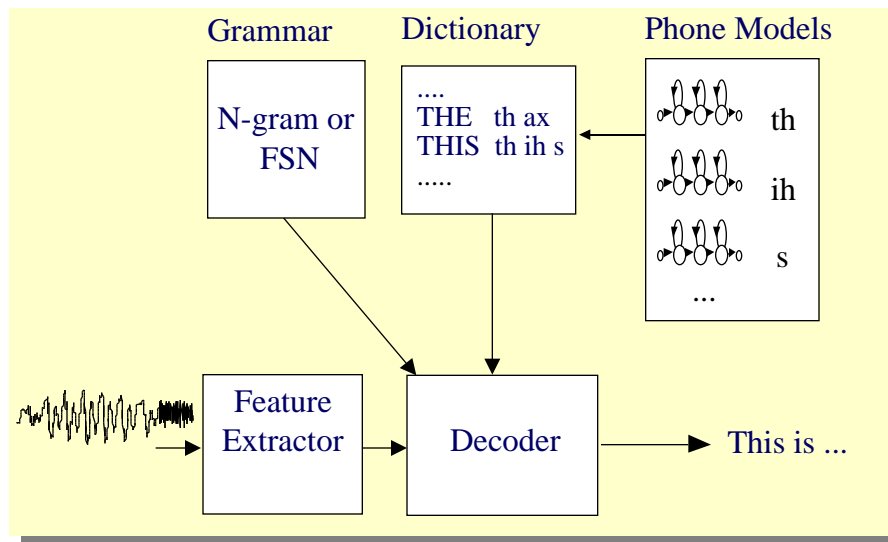
“Beads on a string model”



Microsoft



# Structure of a Simple Decoder



Microsoft



## A few problems ....

- Speech is not stationary over 10 msec intervals
- Every speaker is different
- Microphone, channel and background noise vary
- Segments are highly context dependent
- Pronunciations vary with speaker, style, context, etc
- Speech is not a sequence of segments, articulators move asynchronously
- Prosody makes a difference
- English language is not Markov

So considerable ingenuity is needed to make the “beads on a string” model work

Microsoft



# Building a State of the Art System (1)

## (a) Front-End Processing

- Mel-spaced filter bank + cepstral transform (MFCC or PLP)
- If telephone, restrict bandwidth to 100-3800 Hz
- Cepstral mean and variance normalisation
- First and second differences appended to static coefficients
- Vocal tract length normalisation (ML warping of frequency axis)

Result is 39-D feature vector every 10msecs. Zero mean and unit variance, with frequency axis warped to match speaker.

Microsoft



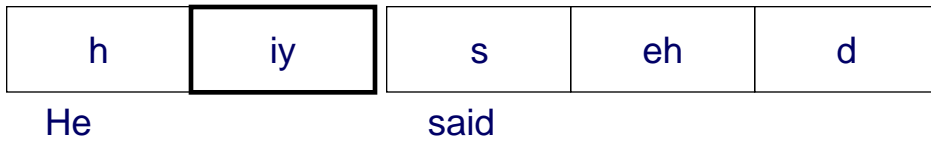
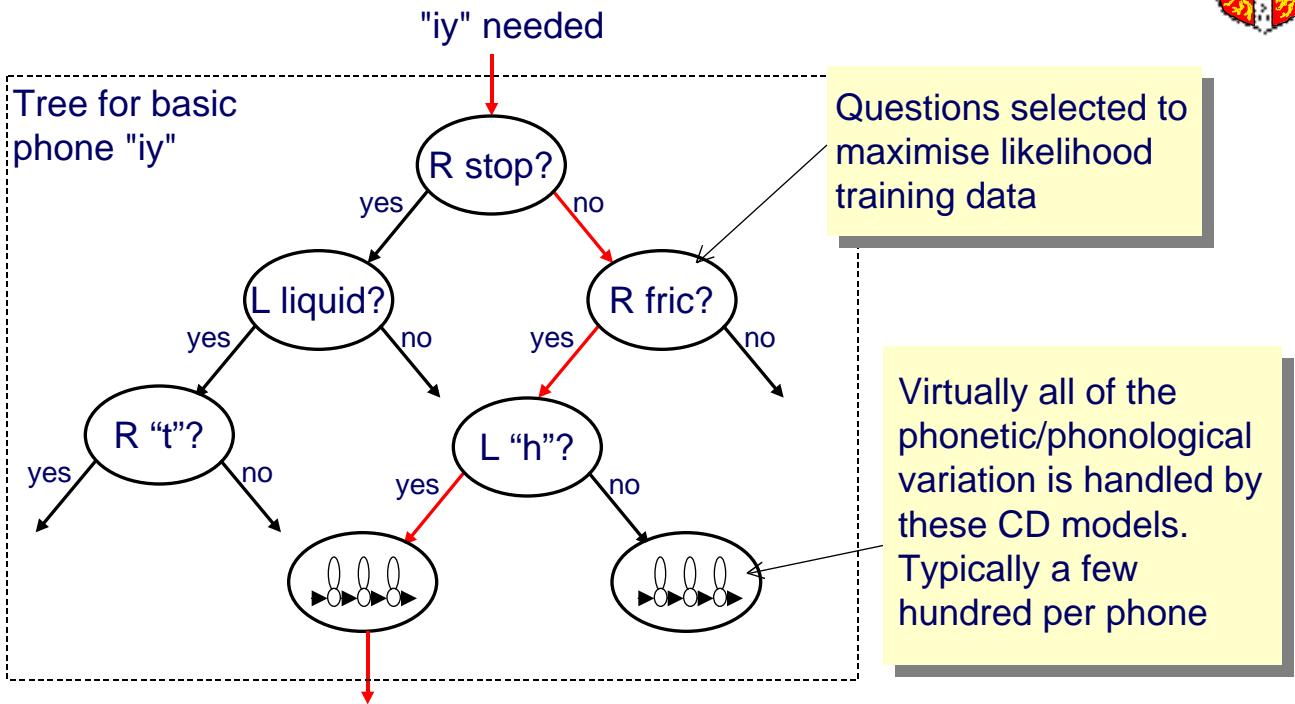
## Building a State of the Art System (2)

### (b) Acoustic Models

- 50 to 250 hours training data
- Observation probabilities are Gaussian mixture components
- Phone models context dependent (CD) - either triphone or quinphone
- CD models clustered according to phonetic context using decision trees
- State-based clustering with soft-tying between nearby states
- Gender dependent models
- Parameter estimation using MLE and MMIE

Result is a set of decision trees and one or more sets of context-dependent HMMs. Typically 10000 distinct states, 16-32 mixture components per state ie ~20,000,000 parameters per set.

Microsoft



Microsoft



## Building a State of the Art System (3)

### (c) Word Models

- Typically one or two pronunciations per word
- Pronunciation dictionaries have hand-crafted core
- Bulk of pronunciations generated by rule
- Probabilities estimated from phone-alignment of acoustic training set
- Type of word boundary juncture is significant

Result is a dictionary with multiple pronunciations per word, each pronunciation has a "probability".

(We would like to do more here, but increasing the number of pronunciations also increases the confusability of words and typically degrades performance)

Microsoft



## Building a State of the Art System (4)

### (d) Language Model

- 100M to 1G word training data
- Back-off word LM interpolated with class-based LM
- Typically ~500 classes derived from bigram statistics
- Vocabulary selected to minimise OOV rate
- Vocabulary size typically 64k

Result is word based LM with typically 20M grams  
interpolated with a class-based LM with typically 1M grams

Microsoft



## Building a State of the Art System (5)

### (e) Decoding

- Multipass sentence-level MAP decoding strategy using lattices
- Initial pass typically simple triphones + bigram LM
- Subsequent passes refine acoustic models (eg GD quinphones)
- Speaker adaptation of Gaussian means and variances using MLLR
- Global full variance transform using "semi-tied" covariances
- Word posteriors and confidence scores computed from lattice
- Final output determined using word-level MAP
- Final output selected from multiple model sets

Result is best hypothesis selected for minimum WER, with confidence scores

Microsoft

## Performance



Each refinement typically gives a few percent relative improvement but  
If done carefully, improvements are additive

Example - The CUED HTK 2000 Switchboard System  
(courtesy Hain, Woodland, Evermann, Povey)

Pass	Models	Ctxt	VTLN	MLLR	FV	MAP	SwbdII	CHE
1	GI-MLE	Tri				Sent	31.7	45.4
2	MMIE	Tri	Yes			Sent	25.5	38.1
3	MMIE	Tri	Yes	Yes		Sent	22.9	35.7
4a	MMIE	Tri	Yes	Yes	Yes	Word	20.9	33.5
4b	GD-MLE	Tri	Yes	Yes	Yes	Word	21.9	33.7
5a	MMIE	Quin	Yes	Yes	Yes	Word	20.3	32.6
5b	GD-MLE	Quin	Yes	Yes	Yes	Word	21.0	32.8
Final	Combine 4a+4b+5a+5b						19.3	31.4

Microsoft



## Some Limitations of Current Systems

- Spectral sampling with uniform time-frequency resolution loses precision
- Useful information is discarded in front-end eg pitch
- Minimal use of prosody and segment duration
- Single stream "string of beads" model deals inefficiently with many types of phonological effect

Moving to a parallel stream architecture provides more flexibility to investigate these issues

Multistream HMMs: Boulard, Dupont & Ris, 1996

Factorial HMMs: Ghahramani and Jordan, 1997

Buried HMMs: Bilmes, 1998

Mixed Memory HMMs: Saul & Jordan, 1999

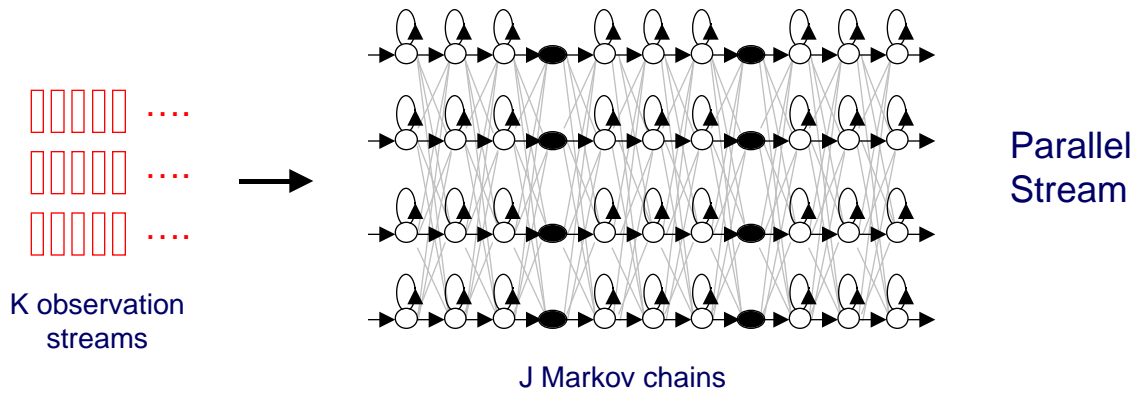
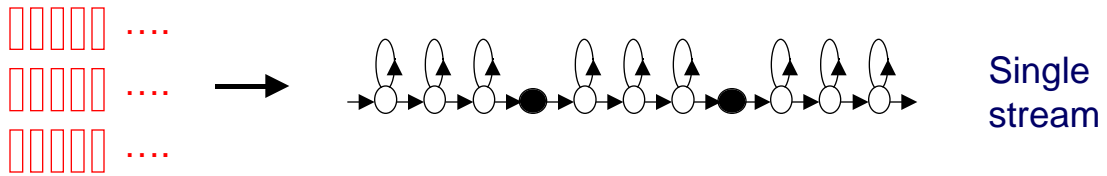


## Examples of pronunciation variability

- Feature spreading in coalescence:  
eg c æ n t -> c æ̃ t where æ is nasalised
- Assimilation causing changes in place of articulation:  
eg n -> m before labial stop as in input, can be, grampa
- Asynchronous articulation errors causing stop insertions:  
eg warm[p]th, ten[t]th, on[t]ce, leng[k]th
- r-insertion in vowel-vowel transitions:  
eg stir [r]up, director [r]of
- context dependent deletion:  
eg nex[t] week



# Parallel Stream Processing Models



Microsoft



## Architectural Issues

- Number and nature of input data streams
- Single sample-rate or multi-rate
- Number of distributed states
- Synchronisation points (phone, syllable, word, utterance?)
- State-transition coupling
- Observation coupling
- Stream weighting





Microsoft

## General “Single Rate” Model

States:

$$S = S_1, S_2, \dots, S_T \quad \text{where} \quad S_t = (s_t^1, s_t^2, \dots, s_t^J)$$

Observations:

$$Y = Y_1, Y_2, \dots, Y_T \quad \text{where} \quad Y_t = (y_t^1, y_t^2, \dots, y_t^K)$$

Assume conditional independence of observation and state components:

$$P(S_t | S_{t-1}) = \prod_{j=1}^J P(s_t^j | S_{t-1})$$
$$P(Y_t | S_t) = \prod_{k=1}^K P(y_t^k | S_t)$$

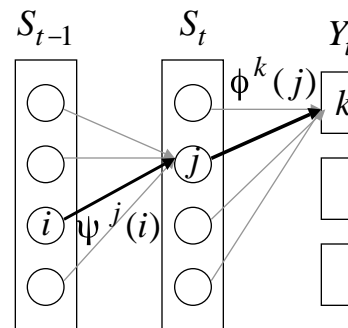


# Example Architectures

## Mixed Memory Approximation

$$P(s_t^j | S_{t-1}) = \prod_{i=1}^J \psi^{j(i)} a^{ji}(s_t^j | s_{t-1}^i)$$

$$P(y_t^k | S_t) = \prod_{j=1}^J \phi^k(j) b^{kj}(y_t^k | s_t^j)$$



Standard HMMs

$$J = K = 1 \quad \psi = \phi = [1]$$

Independent Streams

$$J = 1 \quad K > 1 \quad \psi = [1] \quad \phi = [1, 1, 1, \dots]$$

Multiband HMMs

$$J = K > 1 \quad \psi = \phi = I$$

Factorial HMMs

$$J > 1 \quad K = 1 \quad \psi = I$$

Microsoft



## Source Feature Extraction

- Single-stream (as in standard systems)
- Multiple Subbands (typically 2 to 4 streams)
- Discriminative Features (typically 10 to 15 streams)
- Hybrid (eg Mel-filter, auditory-filter, prosody, ..)
- Multi-resolution (eg wavelets)



Microsoft



## Training/Decoding Issues

Meta state space is huge, hence exact computation of the posterior probabilities is intractable.

Some options:

- Gibbs sampling
- Variational methods
- Chain Viterbi



Microsoft

## Some Simple Experiments



Isolet “alphabet” database

Exact EM training

(Courtesy of Harriet Nock)

Topology	2 SubBands		3 SubBands	
	3 – States	6 - States	3 - States	6 - States
Indep Streams	94.6	96.2	94.4	96.8
Obs Coupled	94.9	95.7	95.2	96.7
Fully Coupled	94.7	95.8	96.0	96.2
Multiband	94.0	95.3	95.2	95.8

Similar results obtained with Chain Viterbi training/decoding



Microsoft



## Conclusions

- State of the Art “Conventional HMM” systems continue to improve
- Many stages of complex processing, each giving small improvement
- Pronunciation modeling in conventional systems is unsatisfactory
- And still no effective use of prosody
- Parallel stream processing models warrant continued study

