



Speech recognition and understanding in realistic environments

Steve Renals

<http://homepages.inf.ed.ac.uk/srenals>

A Big Question

- How can machines make sense of human communication?
- This is a major scientific challenge. Solutions to the problem will lead to advances such as:
 - Richer, more humane interfaces to computers
 - Perceptual computers that can interpret their environment
 - Technological enhancements to human-human interactions (eg effective remote meetings)

A signal-based approach

- Coping with the richness of human communication means coping with signals spread across multiple modalities
- Statistical models of signals learned from streams of multimodal data
- The models and associated algorithms should
 - scale to huge amounts of data
 - be capable of adapting to data that has not been annotated or labelled by humans

How close are we?

- Speech recognition works with known speakers, benign environments or limited domains:
 - Commercial systems for dictation in a quiet environment (adapted to a particular speaker) - low error rates (for many users)
 - The best research systems for conversational speech recognition over the phone 15-30% word error rate; for broadcast news 10-20% word error rate

How far are we?

- Recognizing speech in a realistic environment:
 - No microphones attached to talkers
 - Multiple acoustic sources (eg overlapping talkers)
- Extending the problem:
 - Not just a single channel of audio: multiple microphones, also video information
 - Not just speech transcription, but interpretation, understanding or information access

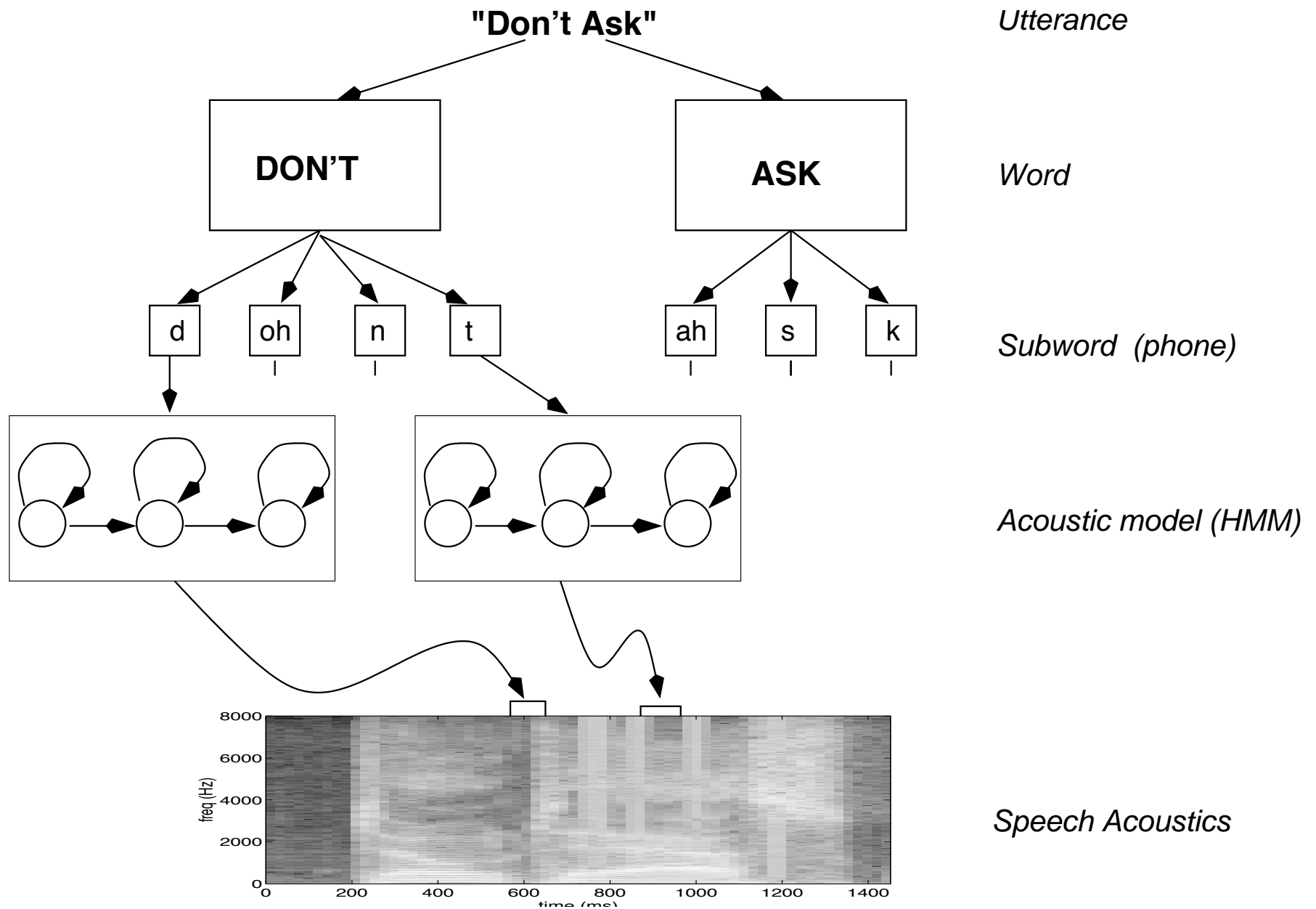
Overview

- Statistical modelling for speech recognition - hidden Markov models (HMMs)
- Applying HMM-based approaches to information extraction from and summarization of speech
- Multistream models for speech and multimodal data - processing multiparty meetings

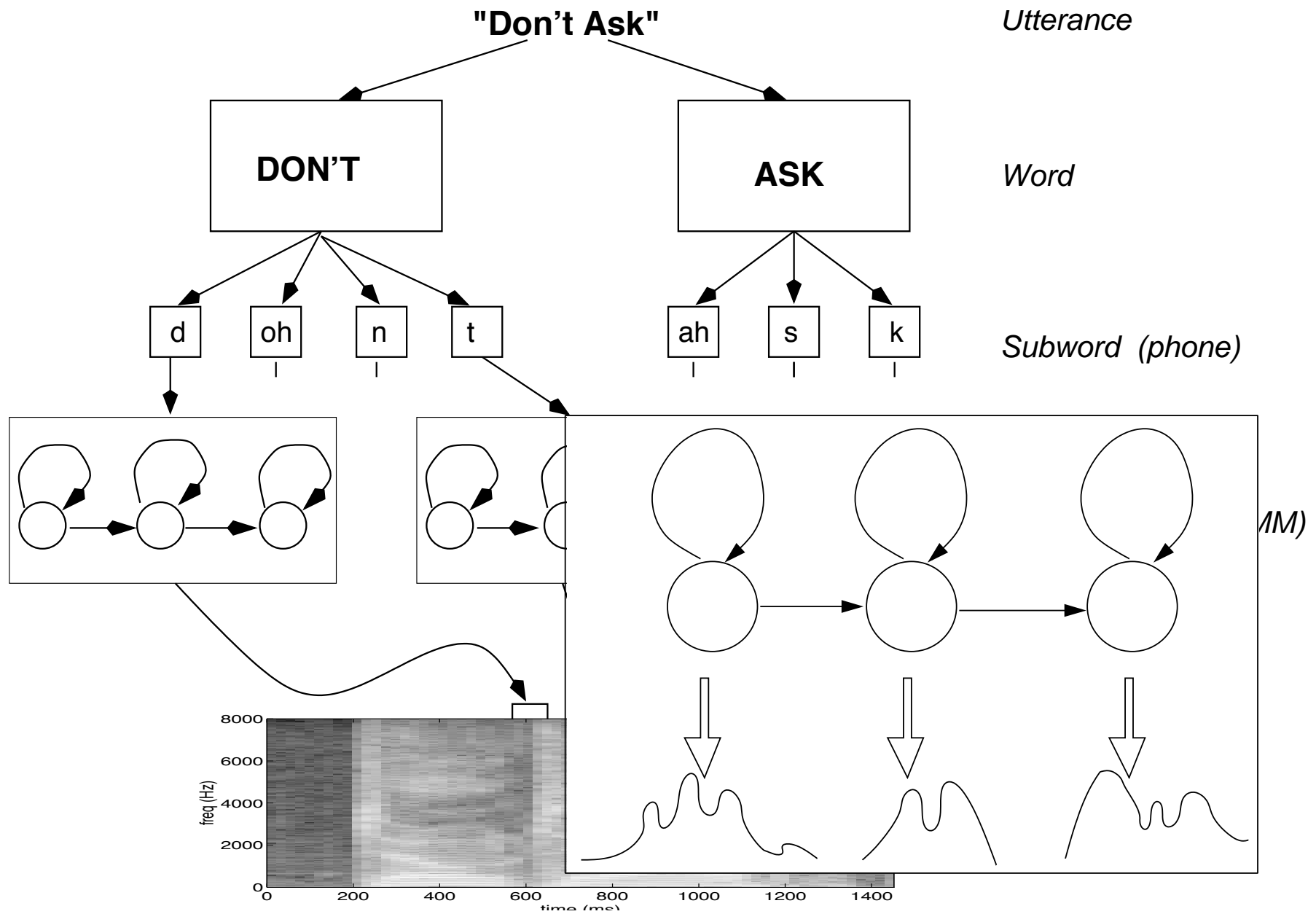
Hidden Markov Models

- Hidden Markov models (HMMs) form the foundations of all modern speech recognition systems (1970s: IDA - Baum/Ferguson/Poritz; IBM - Jelinek/Bahl/Mercer; CMU - Baker)
- Speech is produced by a hidden sequence of states which stochastically generate the observed acoustics (finite state generators)
- Recognition corresponds to finding the state sequence that generated the observed acoustics (and hence the phonemes and words)

HMM speech recognition



HMM speech recognition



HMM acoustic modelling: achievements

- Trainable from huge speech corpora
- Divide and conquer approach (context-dependent modelling)
- Automatic adaptation to new talkers
- Discriminative training criteria
- Confidence and rejection
- Years of research have resulted in very well optimised and engineered systems

HMM acoustic modelling: challenges

- HMMs are a bad model of speech production
- Speech is not a simple sequence of discrete units (“beads on a string”)
- The flat hidden structure has limited expressiveness
- Ongoing work at CSTR exploring streamed models, articulatory feature representations and richer hidden structures (Simon King, Mirjam Wester, Joe Frankel)

Information Access from Speech

- Speech-to-text is a very well defined problem... but it does not address many important issues in understanding spoken language
- We might want computers to
 - make an intelligent response to a spoken query
 - search an archive of audio documents (eg TV and radio broadcasts)
 - extract relevant information from a message
 - summarize speech

Processing recognizer output

- Default approach is to perform speech recognition, then treat the recognizer output as text
- This works very well for some tasks
 - Broadcast news indexing and retrieval (finding the right clip has no degradation with up to 30% word error rate)
 - Automatic identification of names (accuracy scales linearly with word error rate)
- Less well for some others (eg summarizing meetings)

Broadcast news retrieval

THISL – BBC News Evaluation System

Enter your search string in the box below [Help on searching](#)

Stories on the European Commission

Start date: End date: [Help on dates](#)

Programmes available: [Help on restricting the search](#)

Midnight News (BBC1)
One O'Clock News (BBC1)
4pm News (BBC1)
5pm News (BBC1)

Top scoring stories


1 (16.3) 16-Mar-1999 BBC1 Nine O'Clock News ... Duration is 150 secs. Starts 9 minutes in.
2 (15.2) 20-Jul-1999 Radio 4 Midnight News ... Duration is 115 secs. Starts 9 minutes in.
3 (15.0) 15-Mar-1999 BBC1 Nine O'Clock News ... Duration is 66 secs. Starts 23 minutes in.
4 (14.8) 21-Jul-1999 Radio 4 Six O'Clock News ... Duration is 90 secs. Starts 20 minutes in.

[Help on viewing stories](#)

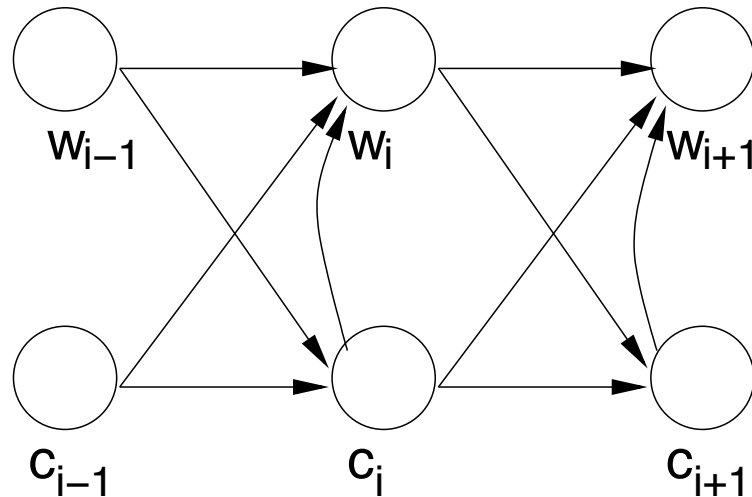
These results from: *commission european*

lost credibility that's still going through the motions and say well until the political masters in the fifteen member states to agree what to do next so are these **commissions** and they get their jobs and how did they get into this mess our correspondent james robbins reports this evening here in the streets with **european** district with you in israel and many other fourteen thousand officials who were full **european commission** would be heading home trying to work out with this extraordinary crisis their masters these then the **commission** may serve so would this **european commission** in brussels actually do is widely misunderstood partly because the **commission** is neither parliament or purely a civil service other commissioners who resigned on their part for politicians who have many other trappings fire officers then they're not ministers either mr. carter today with a new clinic where britain's two commissioners could try to get on with business as usual bushey ordered proposals for a better chance for this special responsibility in the commissioner of the we just leaving o. p. c. r. would be them coming next optimistic in oakland a senior aide to **european** parliament if the **commission** witching ecas new **european** order but it's the parliament the national governments who may be or in europe these two **european** parliament a new clinic in **commissions** must answer for their work would include making sure transport operators stick to the reason for anyone who didn't think there's a lot of time were unlike any other item or worse yet half of those cheerful because members their politics their new clinic is not accused personally wrong doing we plan to transport committee his answer in detail policy questions you to meet again concede the **commission** has failed to find ways to stop some abusing their office the result of neglect earn up to deliberate militias neglect or dishonesty correct but simply that early buzz is a management other people's resources would never consider to be all that important it meant that more reform is being undertaken by this **commission** in the last four years them was undertaken in the previous thirty years so you give **european** commissioners their jobs in the first place their individual prime minister's maybe to some it's like the president to the **commission** it's a messy processed in nineteen ninety four john major famously so the first choice ([Play whole programme](#)).

BBC1
Nine O'Clock
News
16-Mar-1999



Named entity identification



- About 9% of broadcast news is names
- HMM/n-gram model of names and classes (person name, etc.)
- Some technicalities (smoothing, multi-word names)
- 89% precision and recall on hand transcription, 77% on recognizer output (21% WER)

Using prosodic information

- Prosodic structure is observed in the energy, intonation and timing of speech
- Information about emotion, syntax, turn-taking
- Features such as pitch contour and durational information can predict structural features such as boundaries and keywords
- Many candidate prosodic features - select automatically for a given task (eg by Parcel)
- Used successfully in sentence/topic segmentation, speech summarization

Speech summarization



- Keyword extraction from voicemail messages
- Automatic selection of lexical and prosodic features, pattern classification approach
- 10% improvement using prosodic features
- Broadcast news summarization
- How far do text-based extraction methods transfer to broadcast news?

Structural features

- Extractive summarization of text works best using both structural and content features: sentence position, sentence length, word distributions,...
- Structural features are easy to obtain in text (markup); they must be inferred in speech
- Sentence boundary identification - HMMs of lexical and prosodic features (esp. pause)
- Topic segmentation - HMM approaches; also maximum entropy modelling

Broadcast news summarization

- For text news, one or two structural features strongly predict which sentences to extract
- For broadcast news a larger set of content and structural features are required
- Dependence on style (eg read news vs spontaneous interviews)
- Relatively little degradation with word error rate
- Extraction is the easy part of summarization; generation of coherent summaries is harder....

Multiparty meetings

- Development of multimodal approaches to support human interaction in meetings (M4 and AMI projects)
- Instrumented meeting room. Capture multiparty meetings using multiple microphones, multiple video cameras, PC VGA capture, digital pens, e-beam whiteboard - all time-synchronized
- Technology targets: meeting browsers, remote meeting assistants



Right I didn't mean to imply that

Yeah

that we - that we shouldn't discuss this now, but I'm - I'm just saying that

Oh not right now, but I mean in the future. So at this meeting with Liz

Right

I - you know - I mean

Right

I - I do - I'd like to - I like that stuff

Sure sure

So when is she showing up?

Well, I mean, they're coming in April

April. OK

Right. But, um

(Hand transcript)



right yeah race i didn't mean imply that that we'd did
that we should that that's just now but i'm i'm saying that
oh not right now i mean in the future
right
so at this meeting with with you know i mean
right
i i do i'd like to i'd like to stop
sure sure
when she showing
well i mean they're coming in april
april but in right
right but

(ASR Output)



Meeting Browser

Source localization

Meeting event recognition

Speech recognition



Topic segmentation

Object tracking

Meeting modelling

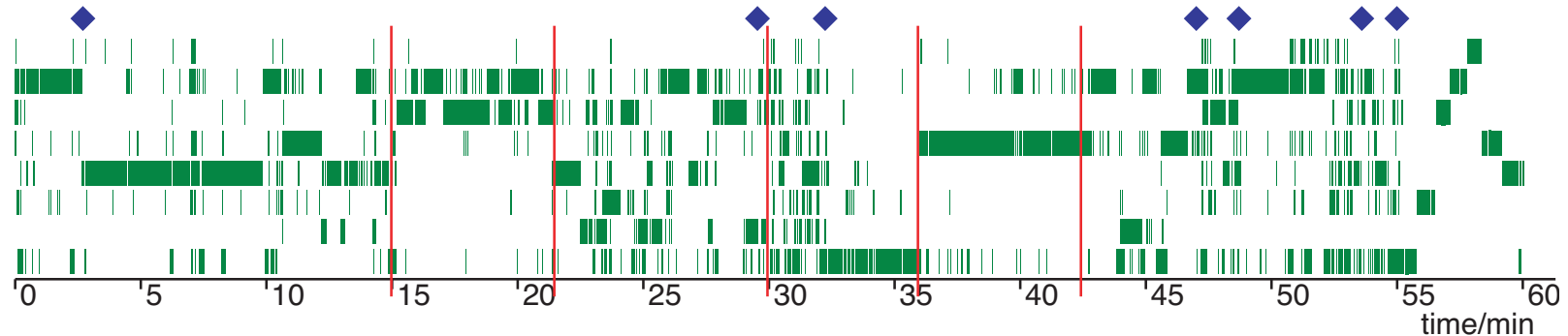
Summarization

Multimodal fusion

Discourse Analysis

Modelling speaker interactions

mr04: Hand-marked speaker turns vs. time + auto/manual boundaries



- Much of a meeting's content is contained in the interaction of participants, as well as in the words spoken
- Browse and segment meetings based on speaker interaction patterns
- Use finite state statistical models for such patterns

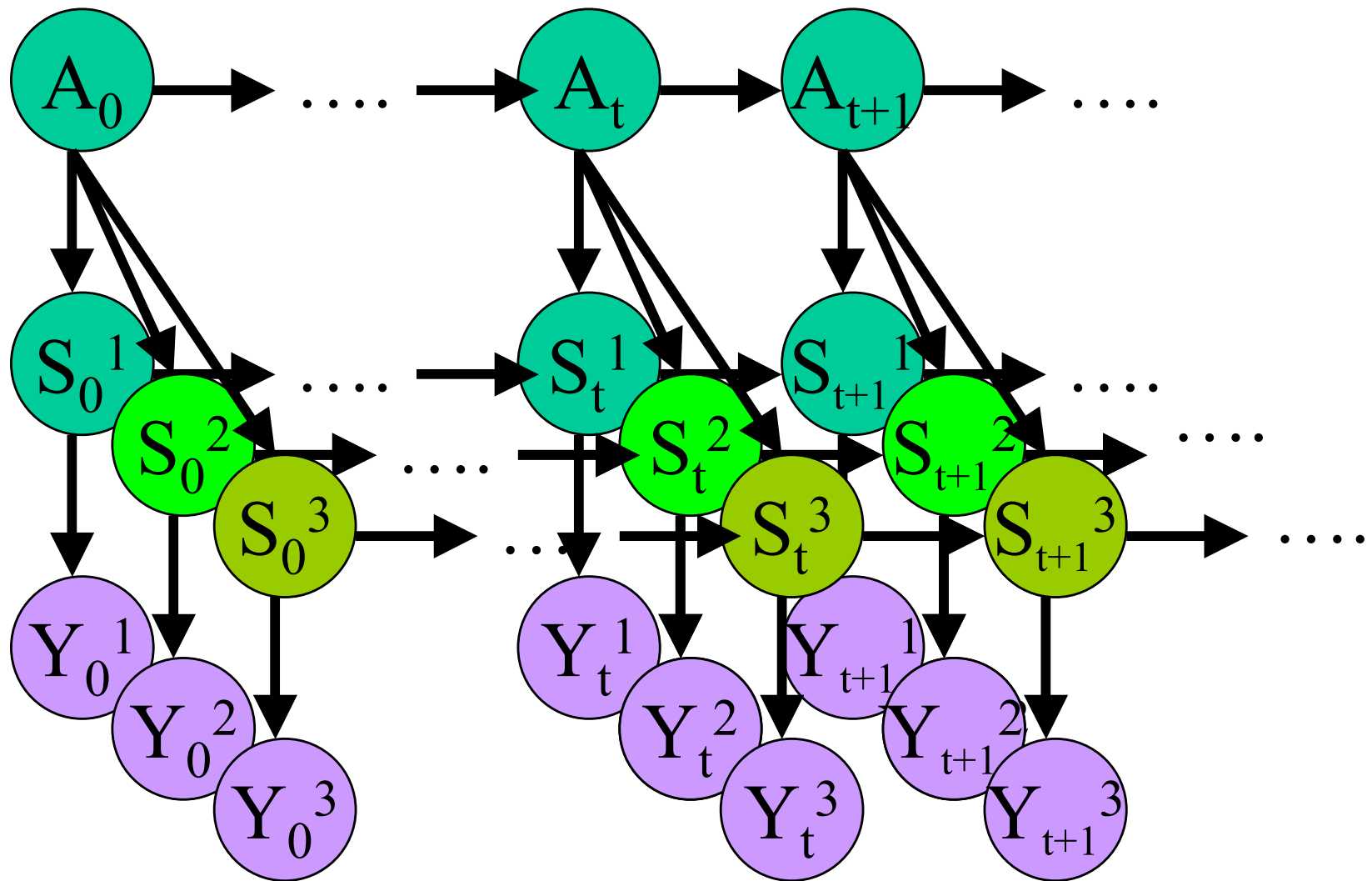
Meeting Event Detection

- Combine feature streams (speech, video, handwriting) to predict events in meetings
- Pilot study: detection of meeting actions (discussion, presentation, monologue,...) from a set of recorded meetings (M4 project)
- Features - speaker turn patterns, F0, rate, energy, lexical features
- HMM - treat the features as a single integrated feature vector - 44% action error rate

Multistream modelling

- Multistream dynamic Bayesian network (DBN) model (generalization of HMMs)
- Richer hidden structure, distributed state representation
- Feature streams processed independently and asynchronously
- Much greater degree of modelling flexibility
- Action error rate of around 9%

Multistream DBN



Multistream models

- Multistream models are well matched to multimodal data and audio-video speech recognition
- And they are well-matched to multiple channel recordings
- But they also offer a more sophisticated model of speech generation - no more “beads on a string” - a framework to develop models of speech better matched to what we know from experimental phonetics

Outlook

- **Similar models at different levels** - currently HMMs and other finite state models
- **Integration of multiple feature streams** - multimodality, use of prosody, multichannel recordings
- **Richer statistical models** - eg dynamic Bayesian networks (also latent variable models)
- **Data-driven feature extraction** - why should signal processing be separate from modelling?
- All this needs well-annotated data collections

Conclusion

- Signal-based approaches to human communication are powerful at several levels
- To understand human communication in real environments we need to make use of all observable aspects of communications - prosodics, interaction, other modalities
- Richer, more succinct models are required - large HMM systems simply “describe” the data
- An interdisciplinary problem - collaborations can be developed through common data sets

Credits

- Joint work with many people. In particular (and in rough chronological order):
Richard Rohwer, Nelson Morgan,
Hervé Boursard, Tony Robinson,
Mike Hochberg, Yoshi Gotoh,
Gethin Williams, Dave Abberley,
Miguel Carreira-Perpiñán, Costis Koumpis,
Vincent Wan, Dan Ellis, Heidi Christensen,
Yasser Abdel-Haleem, Alfred Dielmann,
and the CSTR and AMI teams