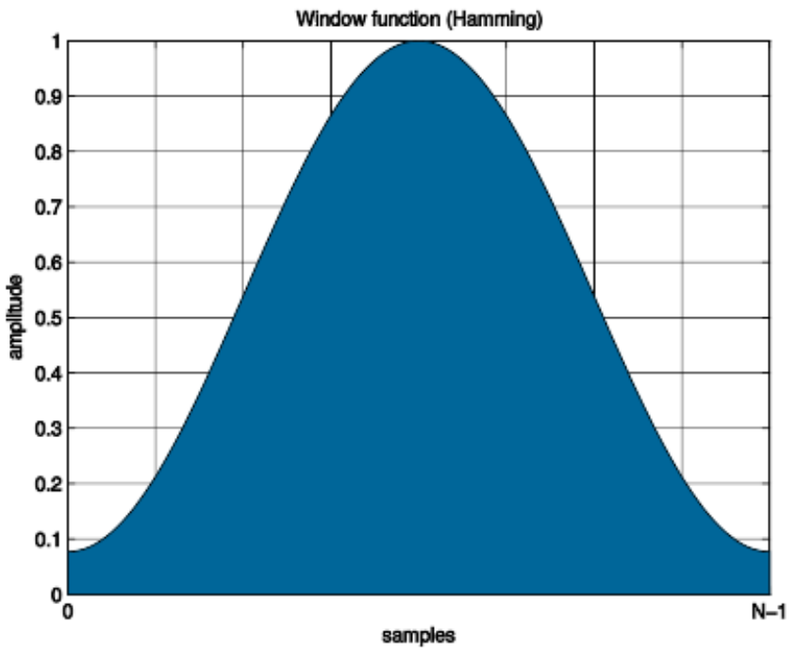
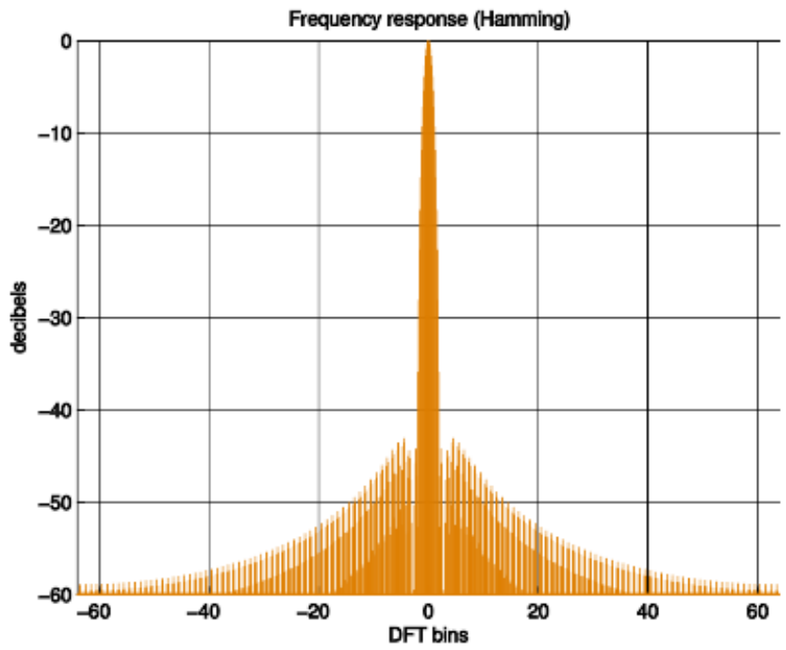
Natural Speech Technology

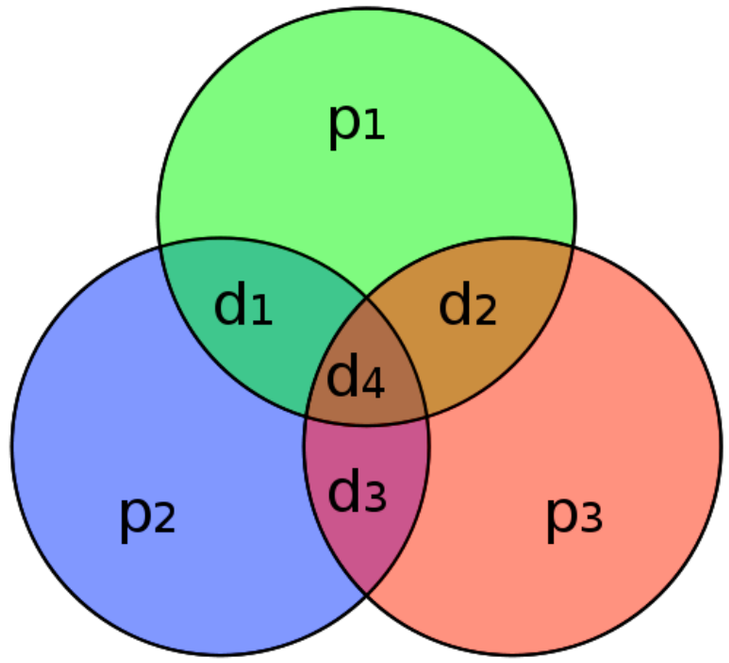
Steve Renals

Hamming Seminar

23 February 2011







$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right)$$

1. Drop modesty
2. Prepare your mind
3. Brains and courage
4. Age is important
5. Make the best of your working conditions
6. Work hard & effectively
7. Believe and doubt your hypotheses
8. Work on the important problems
9. Be committed
10. Leave your door open

UPDATED MAP OF ONLINE COMMUNITIES

SIZE ON MAP REPRESENTS VOLUME OF DAILY SOCIAL ACTIVITY (POSTS, CHAT, ETC). BASED ON DATA GATHERED OVER THE SPRING AND SUMMER OF 2010

EMAIL

SMS

SPOKEN LANGUAGE

FACEBOOK

NORTHERN WASTELAND OF UNREAD UPDATES

FARMVILLE

HAPPY FARM

DOPAMINE SEA

MMO ISLE

WoW

GULF OF LAG

CDC GAMES

PLAINS OF AWKWARDLY PUBLIC FAMILY INTERACTIONS

LAMEBOOK BAY

BLATHERSKITE RIVER

JUNGLE BAY MOUNTAINS OF "IT'S COMPLICATED"

CHARRED WASTELAND OF ABANDONED SOCIAL NETWORKS

MYSPACE

ORKUT

LINKEDIN

HI5

OZONE

BUZZWORD BAY

REGIONAL BAY

FLICKR

YOUTUBE

VIRAL SHORES

MUSIC VIDEO BAY

SNOB SOUND

GOOGLE BUZZ

BIEBER BAY

TWITTER

SEA OF OPINIONS

BAY OF DRAMA

PHOTO BLOGS

BLOGOSPHERE

DIARY BLOGS

GOSSIP BLOGS

POLITICAL BLOGS

MUSIC BLOGS

TECH BLOGS

SEA OF PROTOCOL CONFUSION

WHALER SOUND

TROLL BAY

SEA OF MEMES

FORUMS

4CHAN

AIM

WINDOWS LIVE MESSENGER

CRAIGSLIST

2CHANNEL

GREAT FIREWALL

GULF OF CHINA

GRASS MUD HORSE BAY

http://xkcd.com/802/

http://xkcd.com/802/

SPOKEN LANGUAGE

INTERNET

EMAIL

SMS

CELL PHONES

PLAINS OF AWKWARDLY PUBLIC FAMILY INTERACTIONS

GONDOLIN

NING

NASA-KLASZA

WER-KENNT-WEN

CLASSMATES.COM

EURO

SKYROC

http://xkcd.com/802/

Speech technology seems to evoke two types of response

1. *It's a solved problem*
2. *It's hopeless*

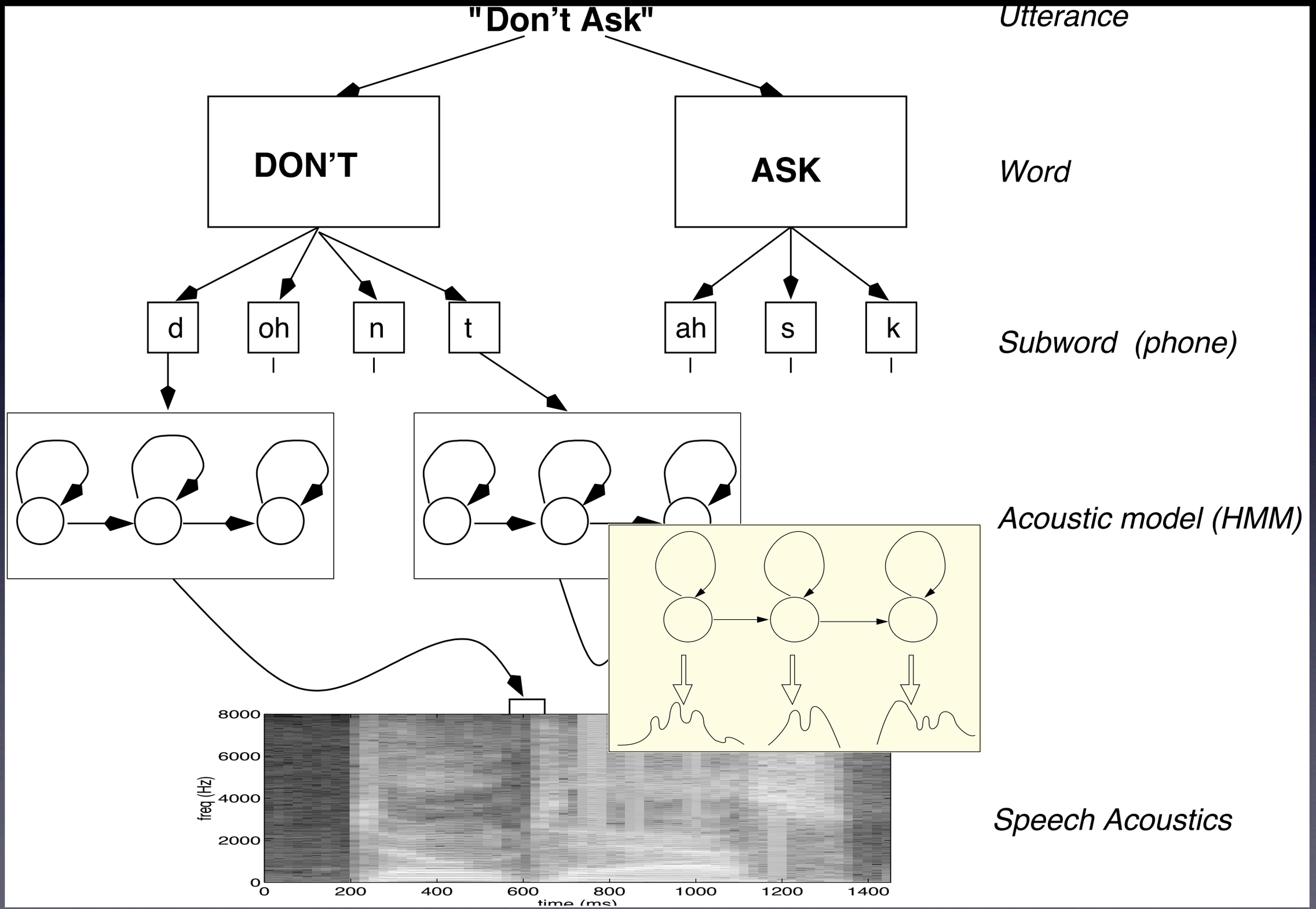
A natural speech technology

- Speech recognition
  - systems that can detect “who spoke what, when and how” for any acoustic environment and task domain
- Speech synthesis
  - controllable systems capable of generating natural and expressive speech in a given voice
- Adaptation, Personalisation, Expression

Speech recognition

- Dictated newspaper text (“Wall Street Journal”)
- Conversational telephone speech (“Switchboard”)
- Multiparty conversations (“AMI Meetings”)

HMM/GMM


"Don't Ask"
Utterance
DON'T
ASK
Word
d
oh
n
t
ah
s
k
Subword (phone)
Acoustic model (HMM)
8000
6000
4000
2000
0
freq (Hz)
0
200
400
600
800
1000
1200
1400
time (ms)
Speech Acoustics

Acoustic modelling

Basic Framework

HMM

Acoustic modelling

Acoustic Features

MFCC

HMM

PLP

Acoustic modelling

Objective function

MFCC

MLE

HMM

PLP

Acoustic modelling

Feature Normalisation

MFCC

CMN

VTLN

MLE

CVN

HMM

PLP

Acoustic modelling

Adaptation

MFCC

CMN

VTLN

MLE

CVN

HMM

MLLR

PLP

CMLLR

Acoustic modelling

Adaptive Training

CHAT

MFCC

CMN

VTLN

MLE

CVN

HMM

MLLR

PLP

SAT

CMLLR

Acoustic modelling

Feature Transformation

HLDA

CHAT

MFCC

CMN

VTLN

MLE

HMM

CVN

MLLR

PLP

SAT

CMLLR

Acoustic modelling

Discriminative Training

HLDA

CHAT

MFCC

MPE

CMN

VTLN

MLE

HMM

CVN

MLLR

fMPE

PLP

SAT

RDLT

CMLLR

Acoustic modelling

Task adaptation

MAP

HLDA

CHAT

MFCC

MPE

CMN

VTLN

MLE

HMM

MPE-MAP

CVN

MLLR

fMPE

PLP

SAT

RDLT

CMLLR

Acoustic modelling

Posterior features

MAP

HLDA

CHAT

MFCC

MPE

LCRC

CMN

VTLN

MLE

MPE-MAP

HMM

CVN

MLLR

fMPE

PLP

SBN

SAT

RDLT

CMLLR

Acoustic modelling

Model Combination

MAP

HLDA

CHAT

MFCC

MPE

LCRC

CMN

VTLN

MLE

HMM

MPE-MAP

CVN

ROVER

MLLR

fMPE

PLP

SBN

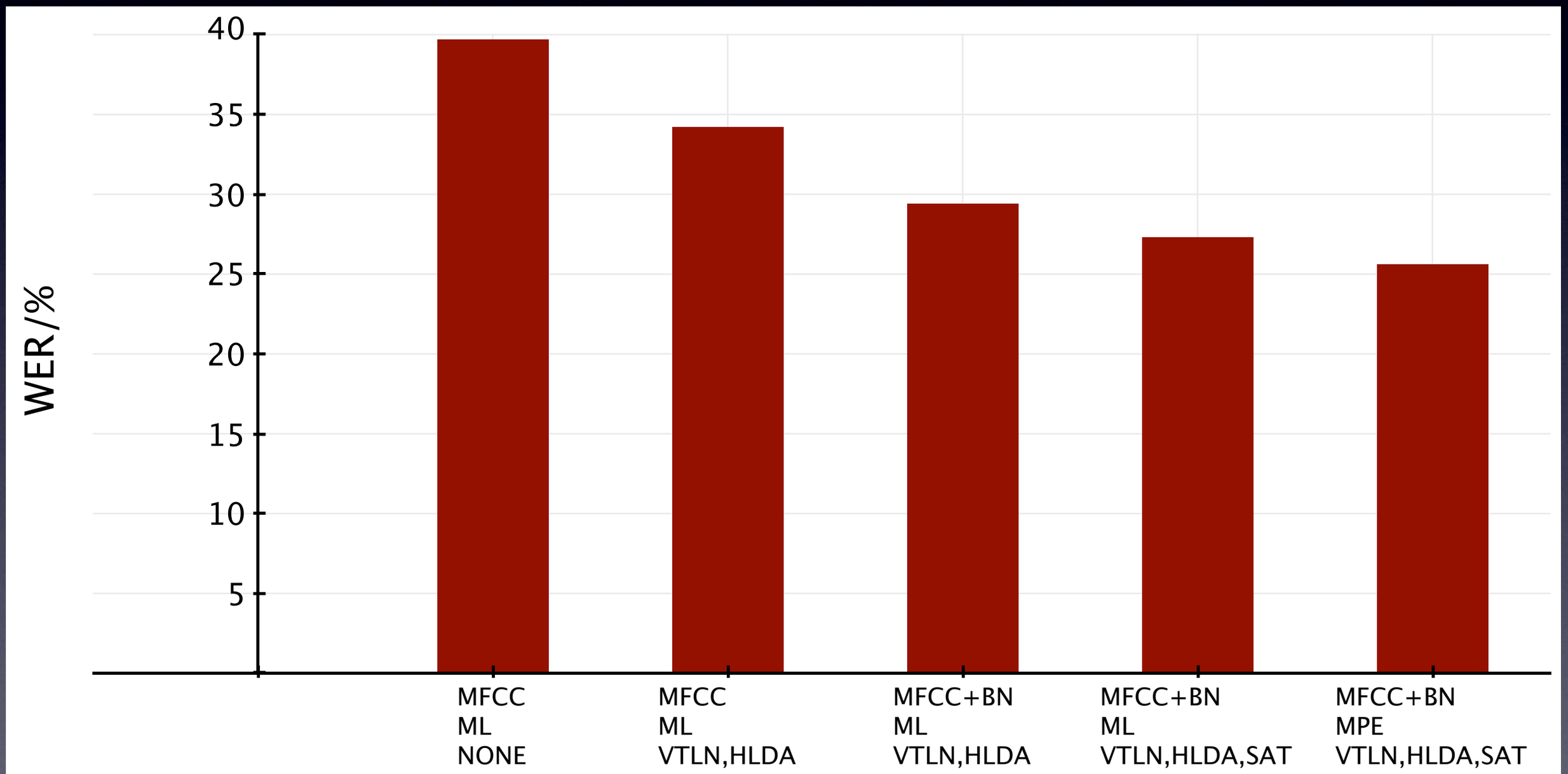
CN

SAT

RDLT

CMLLR

Additive gains on meeting recognition



Hain et al 2009

Features / Training / Adapt

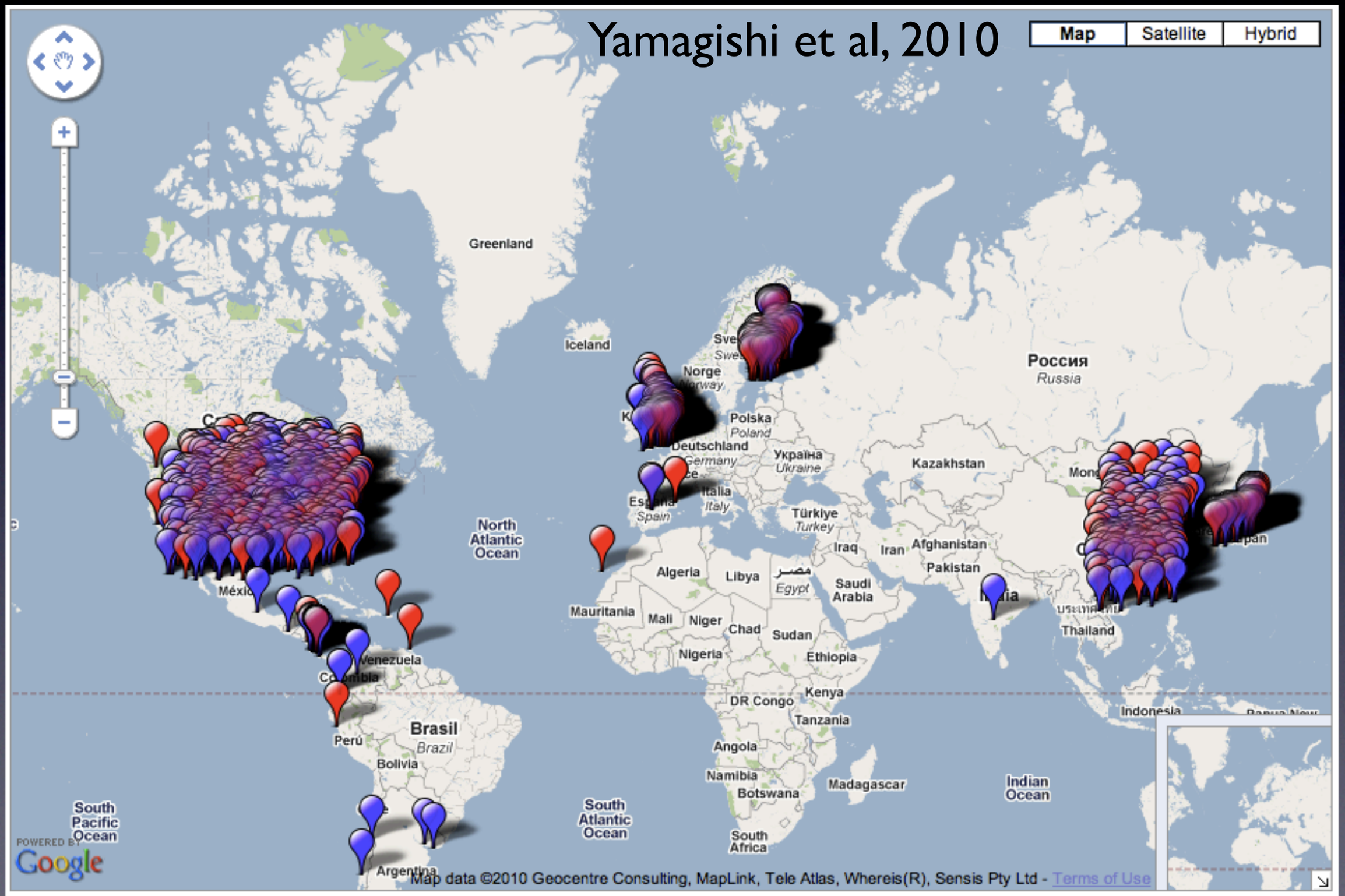
Speech synthesis

- 1970-80s: parametric, rule-based
- 1980-90s: data-driven, concatenate diphones
- 1990-2000s: data-driven, concatenative, unit selection
- 2000-2010s: statistical parametric (HMM) synthesis

HMM Speech Synthesis

- Use the HMM generative model to generate speech
  - automatic estimation of parameters
  - different objective functions possible
  - HMM/GMM adaptation algorithms – possible to develop new synthetic voices with a few mins of data
  - uses highly context dependent models
  - need to model duration, F0, multiband energy amplitude

A world of synthetic voices



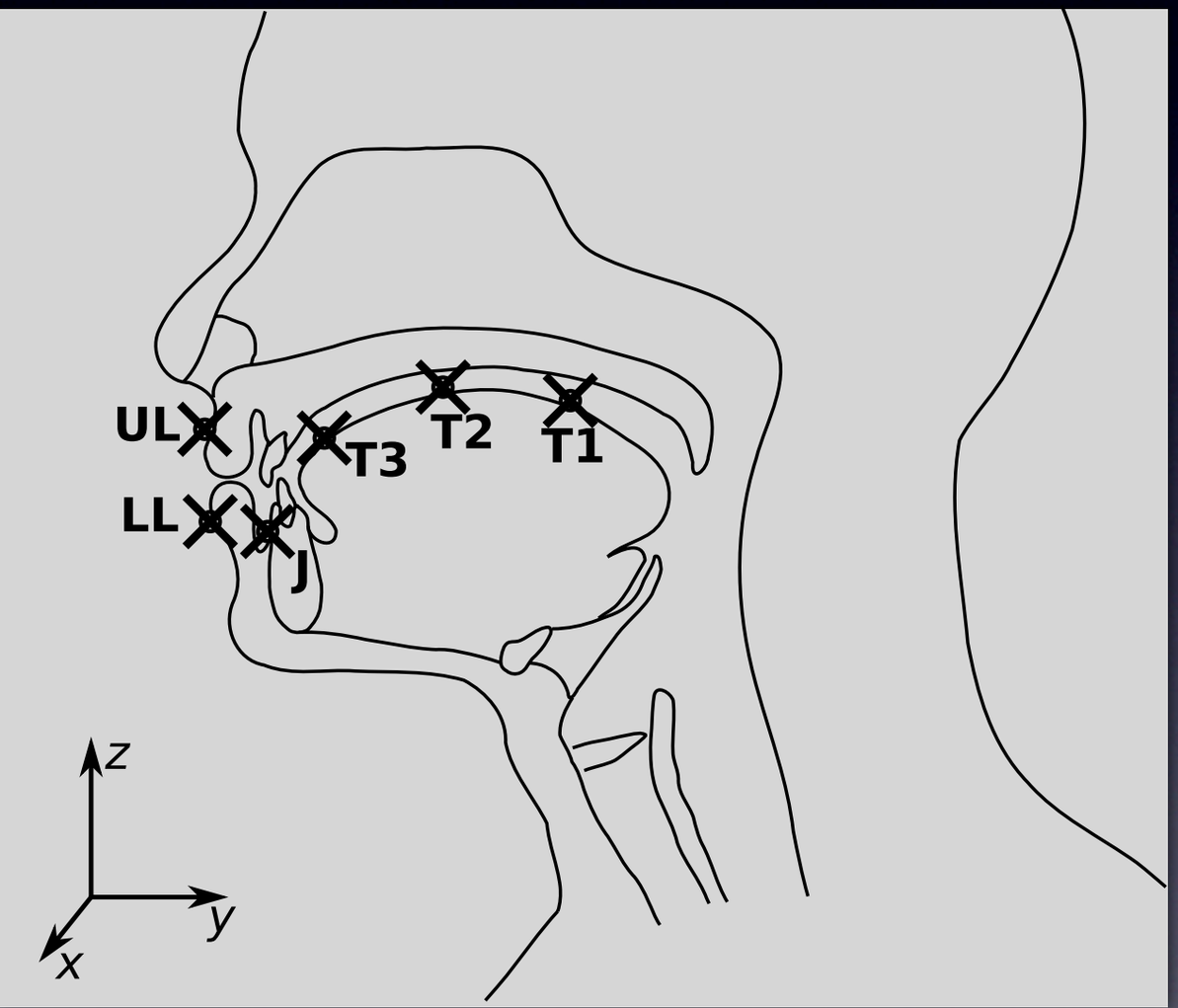
Key advances

- Speaker adaptation: MLLR and MAP families
- Context-dependent modelling: divide and conquer using phonetic decision trees
- Different training criteria: maximum likelihood, minimum phone error, minimum generation error
- Discriminative long-term features – “posteriograms”

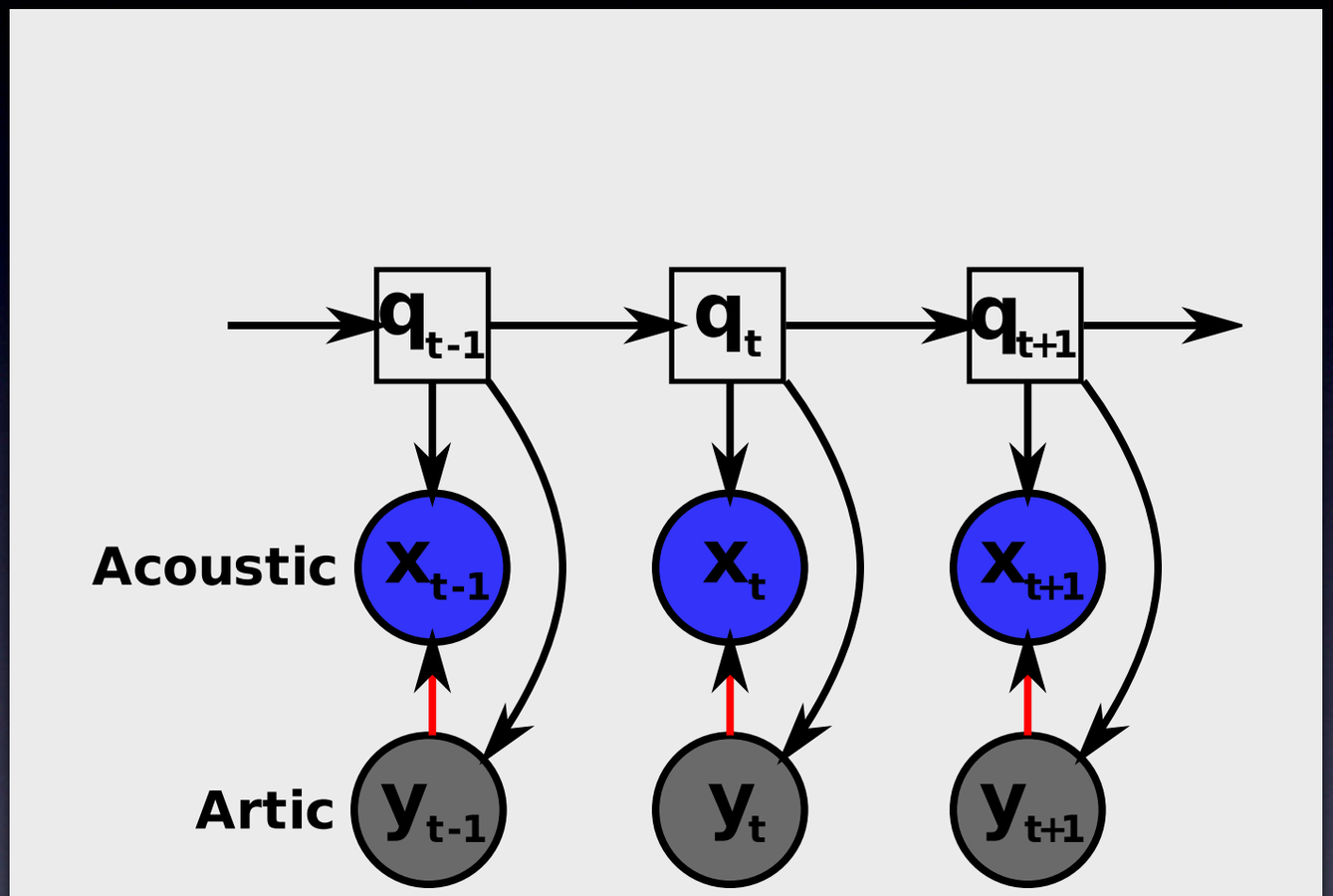
What's lacking?

1. Speech knowledge
2. Factorisation in speech recognition, control in speech synthesis
3. Multilinguality
4. Rich transcription
5. Operating in complex acoustic environments
6. Unsupervised learning

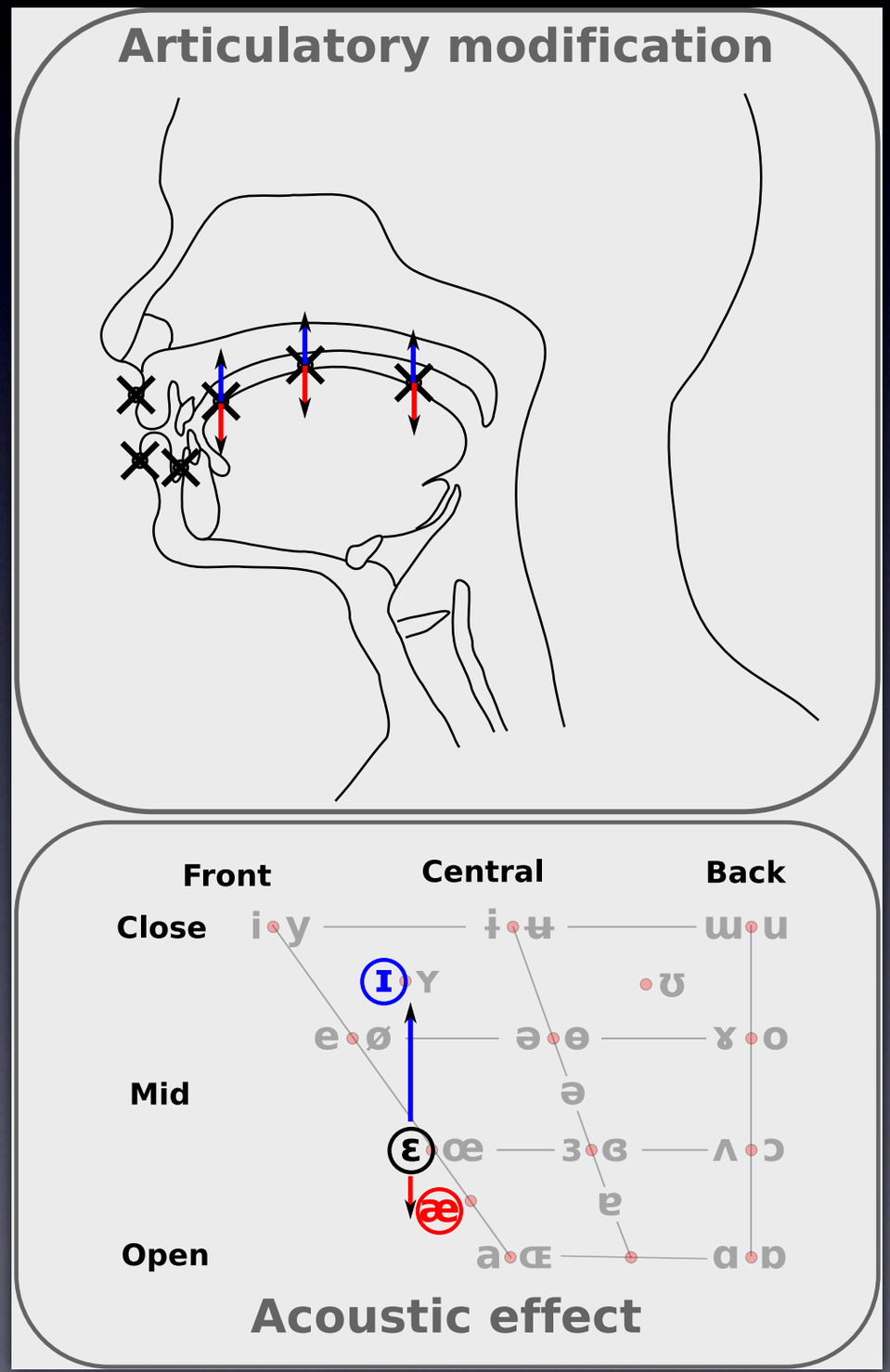
1. Speech knowledge



Acoustic-Articulatory HMM Synthesis



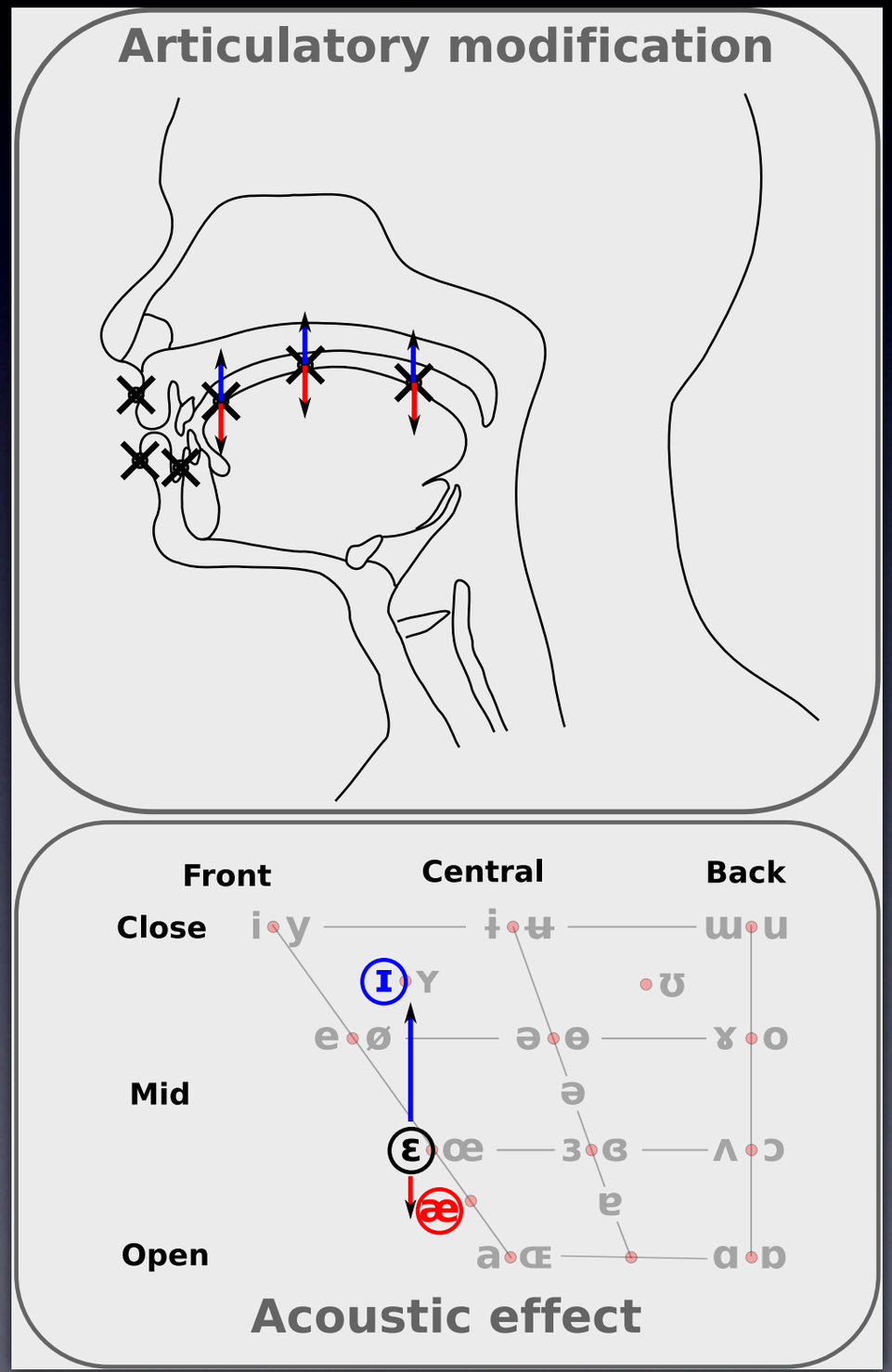
Ling, Richmond, Yamagishi & Wang, 2009



Acoustic-Articulatory HMM Synthesis

Tongue height (cm)		
	+1.5	○
	+1.0	○
	+0.5	○
	default	peck
	-0.5	○
	-1.0	○
	-1.5	○

Ling, Richmond, Yamagishi & Wang, 2009



2. Factorisation

- Adaptation algorithms successfully operate by transform model parameters (or features) based on small amount of data
- But they are a blunt instrument, adapting for whatever changes are in the data
  - channel
  - speaker
  - task
- Can we treat different factors separately?

JFA and Subspace Models

- Factorisation in speaker identification: verify the talker not the telephone channel!
  - Joint factor analysis – factor out the speaker and channel aspects of the model (Kenny et al, 2007)
- Factorisation in speech recognition
  - Subspace models – low dimensional global subspace, combined with state-specific parameters (Povey, Burget et al, 2010)

3. Multilinguality

- The power of the statistical framework:
  - we can use the same software to train systems in any language!
- But this assumes
  - transcribed acoustic training data
  - pronunciation dictionary
  - text data for language modeling
- Not all language are well resourced

Multilingual challenges

- Share common acoustic model information across languages
  - subspace models
- Cross-lingual adaptation (e.g. speak Japanese in your own voice)
  - EMIME
- Inference of pronunciations for new languages
- Automatic data collection

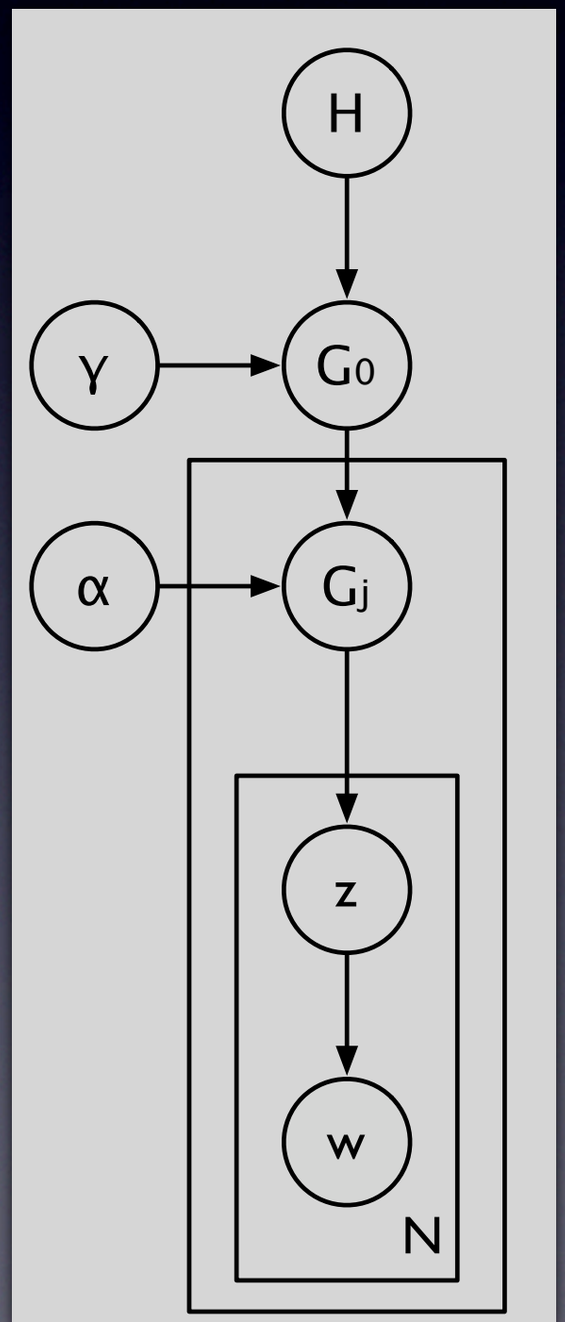
3 (a). Accents

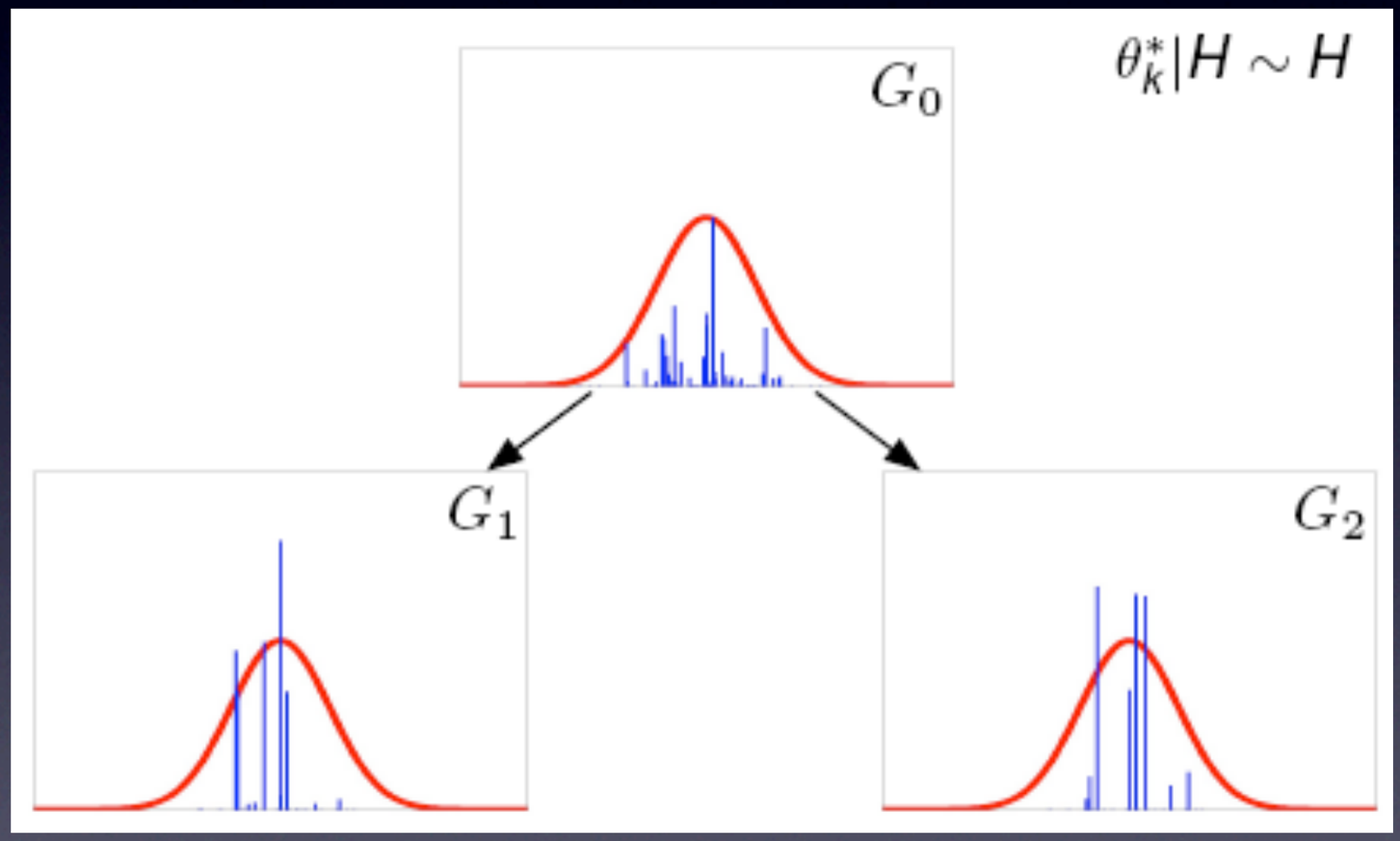
- Accents implicitly modelled by the acoustic model – treat accent as separately modelled factor?
- Structured accent models for synthesis and for recognition
- Can we make use of accent-specific pronunciations?

4. Rich transcription

- Speech contains more than just the words – recognise (and synthesise) metadata
- Analyse and synthesise social content
  - expression
  - subjectivity
  - social role
- Towards speech understanding
  - summarisation
  - topic extraction

Incorporate topics, social role, etc.

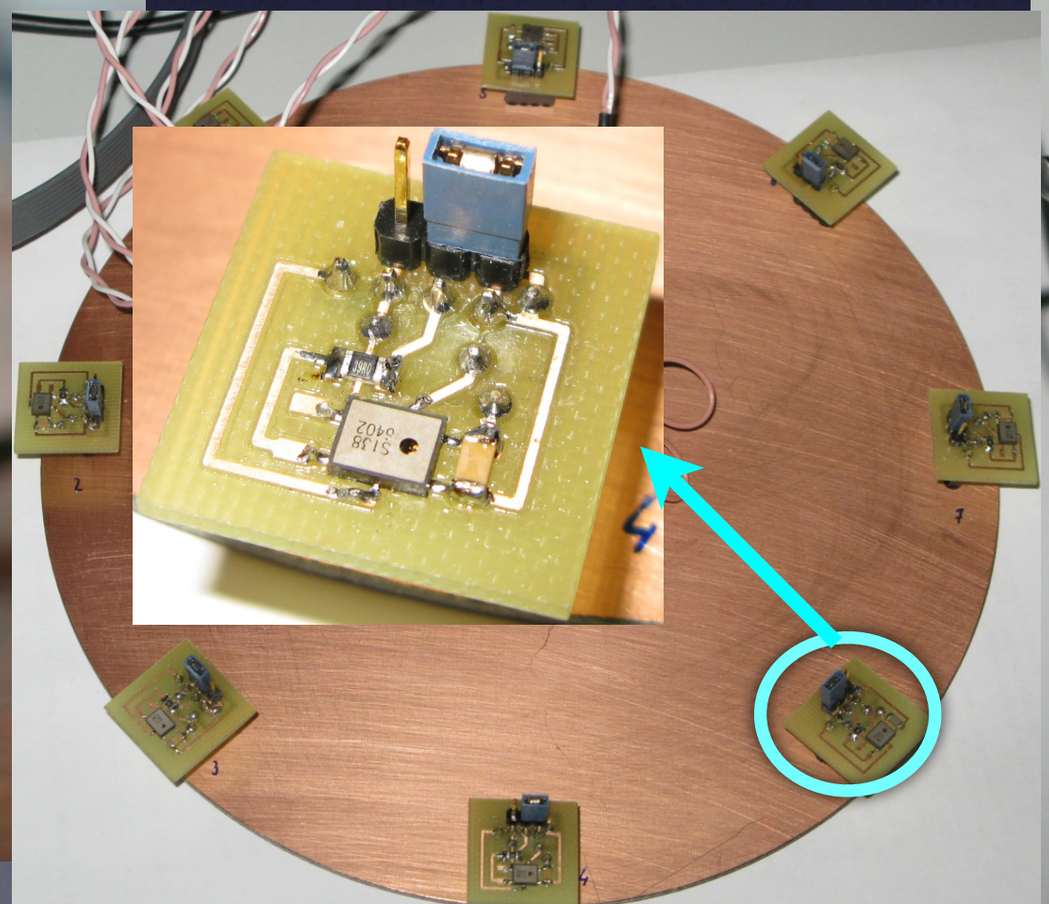


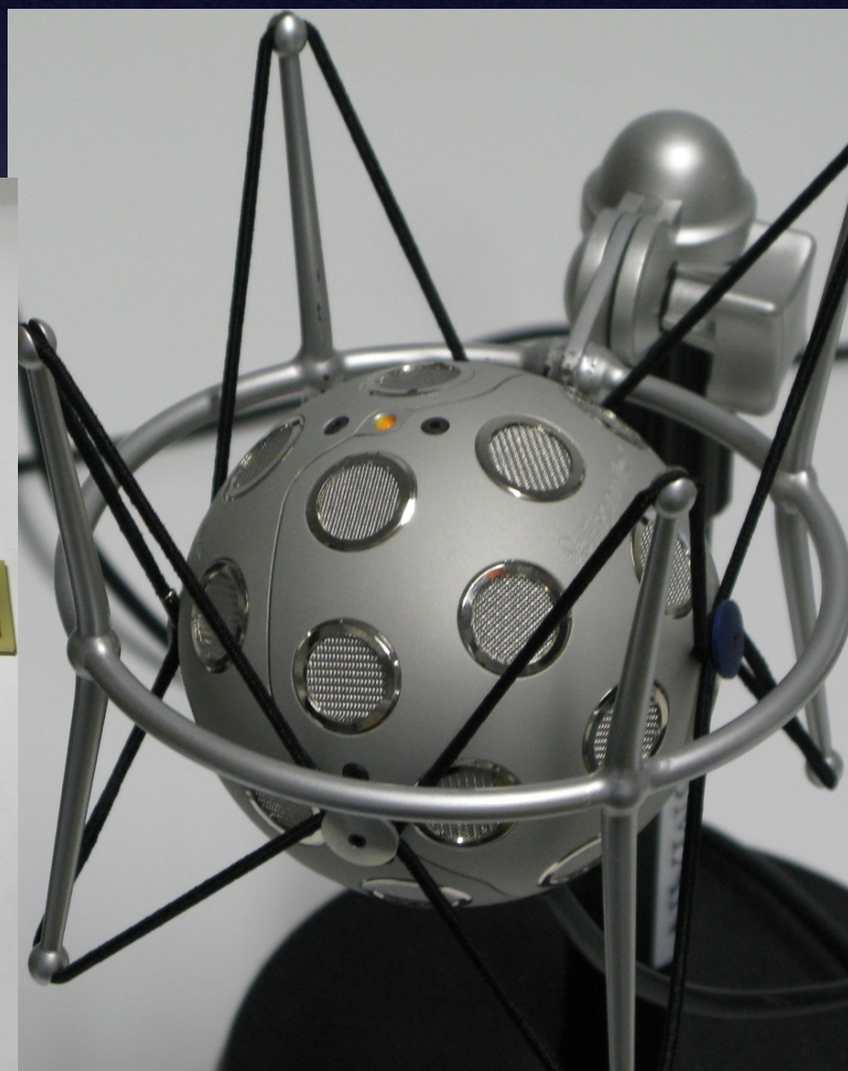


Huang, 2009

5. Complex acoustic environments

- Natural environments have many acoustic sources
- Capture and analyse the sounds in an environment

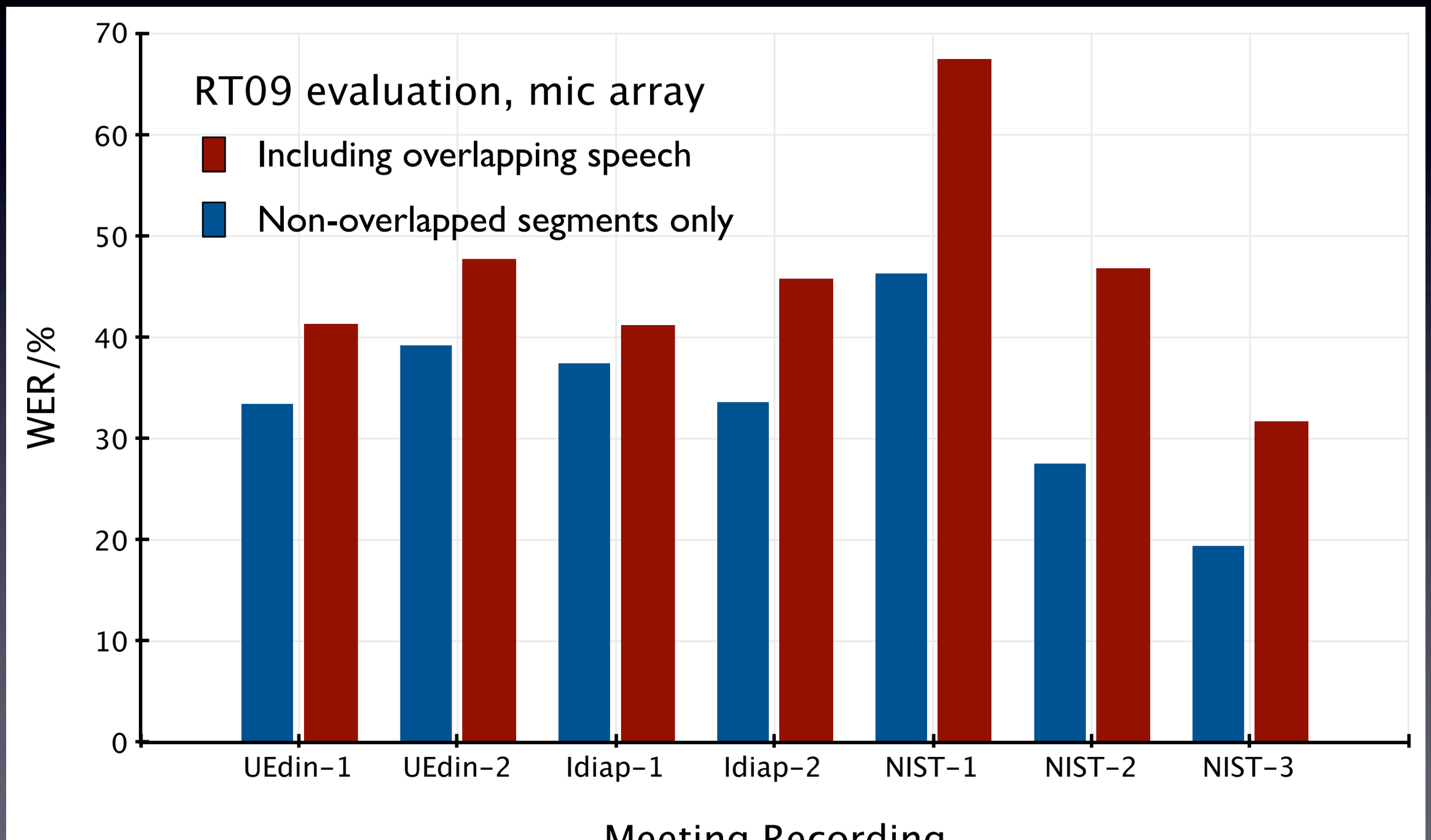




Distant speech recognition

- Using recognition models rather than enhancement models for (uncalibrated) microphone arrays
- Combining separation with recognition – overlapped speech
- Large arrays of inexpensive silicon microphones
- Uncalibrated arrays (no common clock, no position information)
- Analyse all the acoustic sources in a scene

Overlapped speech
NIST RT-2009 evaluation



6. Unsupervised learning

- It's not really economic to manually annotate the diversity of human speech
- Unsupervised / lightly supervised learning
  - web resources
  - combined generative/discriminative models
- One million hours of speech? - OK!
- Twenty-five million hour of annotation? - hmmm...
- Move from fixed corpora to never-ending streams?

Summary

- Adaptation, Personalisation, Expression
  - Incorporate speech knowledge
  - Factorisation and control
  - Multilinguality
  - Accents
  - Rich transcription
  - Complex acoustic environments
  - Unsupervised learning

Thanks.