

# New and emerging applications of speech synthesis

---

Junichi Yamagishi  
The Centre for Speech Technology Research  
University of Edinburgh

[www.cstr.ed.ac.uk](http://www.cstr.ed.ac.uk)

Speech synthesis seminar series 9th February 2011

# The approach taken in this talk

---

- Instead of
  - focussing on techniques which improve the quality of synthetic speech
- we will structure the tutorial around
  - scientifically and commercially novel and interesting technologies and applications
- and in doing this, we will
  - describe the underlying techniques
  - try to place speech synthesis on a more scientific basis
- *Of course, quality is not a solved problem - it's just not the focus of this tutorial*

# Current, new and emerging applications

---

- Currently, synthesis is an “optional extra” allowing **text** to be read out loud
  - Accessibility, screenreader
  - Telephone services, dialogue systems
  - Basic (and rather boring) e-Book reader
  - In-car navigation
  - Basic voice communication aids for people with disabilities
- But none of these seem to be ‘*killer applications*’
- New and emerging applications: focus on the **voice** not the text
  - Voice cloning
  - Voice reconstruction for people with disordered speech
  - Personalised speech-to-speech translation
  - Noise-adaptive speech synthesis

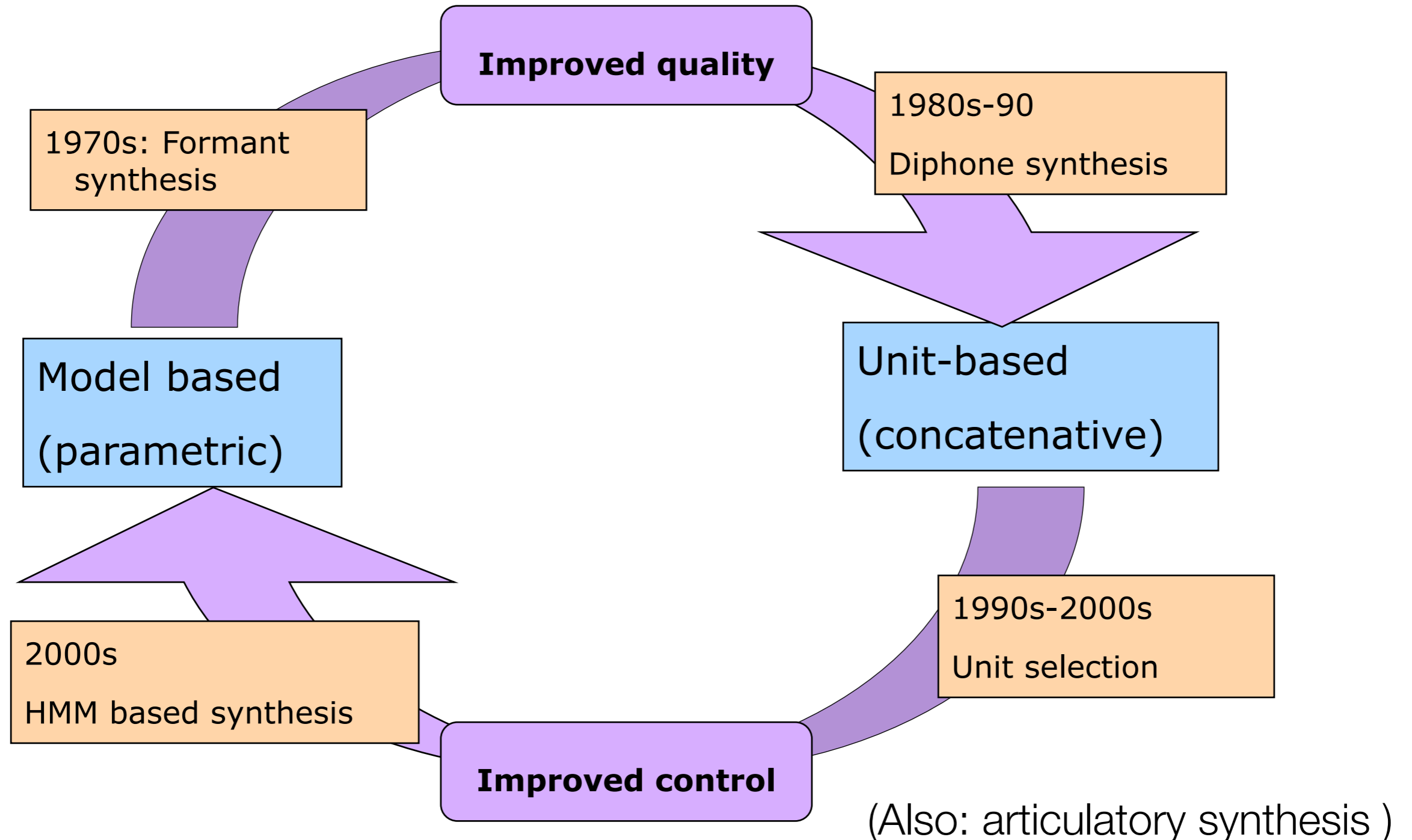
# Text-to-Speech (TTS)

---

- Definition: a text-to-speech system must be
  - able to read any text
  - intelligible
  - natural sounding
- Methods available
  - **Model-based**
    - simplified model of speech production (e.g., Klatt vocal tract model)
    - typically driven by rules
  - **Concatenative**
    - cut & paste recorded examples of real speech - unit selection
  - **Vocoder + statistical parametric model**
    - functional model of speech signal
    - driven by statistical parametric model (e.g., HMM)

# History of TTS

---



# Scientific basis

---

- Current TTS uses very limited knowledge of human
  - **speech production**
    - explicit (physical) models of production still inadequate
    - but there are other ways to make use of production knowledge
    - deeper models could enable better generalisation
  - **speech perception**
    - essential, but almost entirely absent from current systems
    - after all, the synthetic speech is intended to be **heard**
- Interplay with related fields: speech perception & production, animation, ...
  - **articulatory-controllable** statistical speech synthesis
  - TTS adaptive to listener & environment **without requiring new data**

# Contents

---

- Background
  - a brief introduction to HMM-based speech synthesis
  - basic principles of adaptation
- Applications
  1. Voice cloning
  2. Voice reconstruction
  3. Personalised speech-to-speech translation
  4. Articulatory-controllable speech synthesis

# Contents

---

- Background
  - **a brief introduction to HMM-based speech synthesis**
  - basic principles of adaptation
- Applications
  1. Voice cloning
  2. Voice reconstruction
  3. Personalised speech-to-speech translation
  4. Articulatory-controllable speech synthesis





# A brief introduction to HMM-based speech synthesis

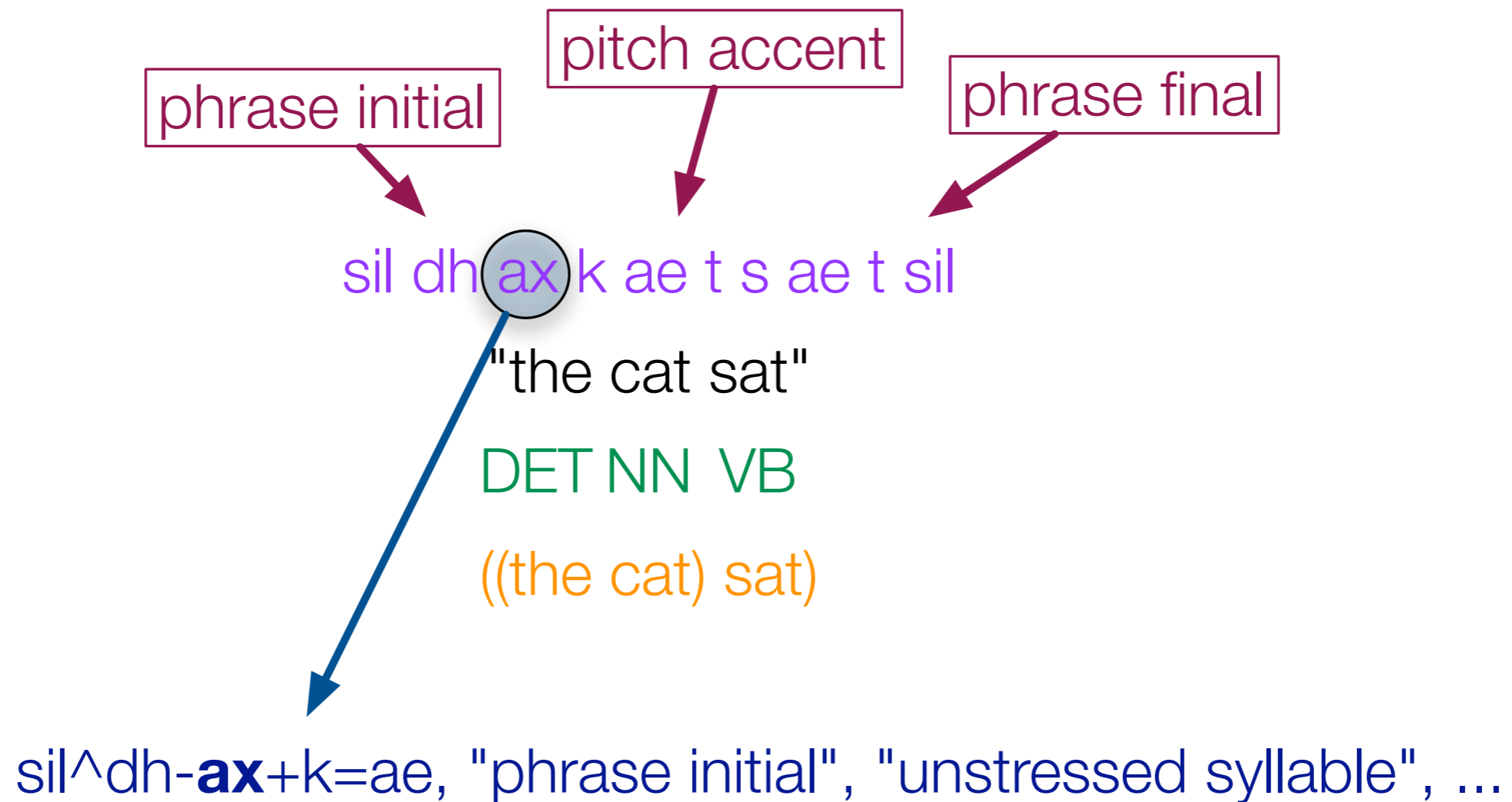
---

# Text-to-speech synthesis

---

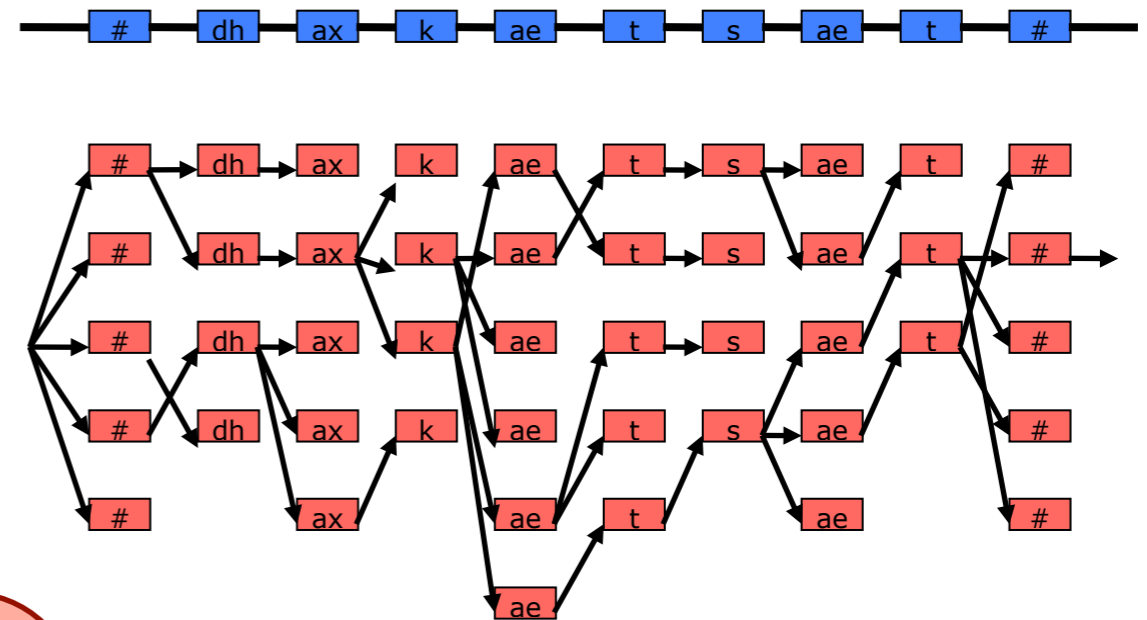
- Text to speech
  - *input:* text
  - *output:* a waveform that can be listened to
- Two main components
  - *front end:* analyses text and converts to linguistic specification
  - *waveform generation:* converts linguistic specification to speech

# From words to linguistic specification



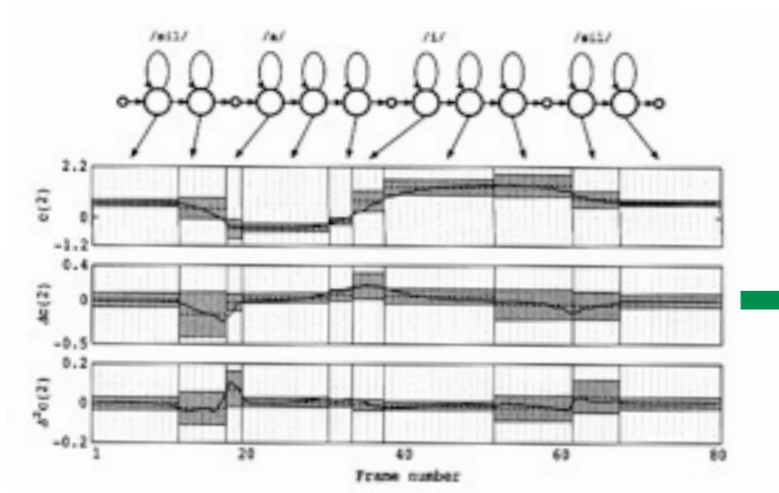
# From linguistic specification to speech

Concatenate  
pre-recorded speech

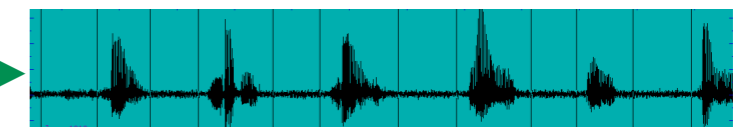


or

Drive a vocoder with a  
statistical model



vocoder



# Flattening the linguistic specification: attaching all features to the segment level

---



Phrase

Phrase

Word

Word

Word

Pitch accent

Boundary tone

Syllable

Syllable

Syllable

Syllable

P P P P P P P P P P P P P P P P

# Flattened specification: context-dependent phones



pau^pau-pau+ao=th@x\_x/A:0\_0\_0/B:x-x-x@x-x&x-x#x-x\$. . . . .  
pau^pau-ao+th=er@1\_2/A:0\_0\_0/B:1-1-2@1-2&1-7#1-4\$. . . . .  
pau^ao-th+er=ah@2\_1/A:0\_0\_0/B:1-1-2@1-2&1-7#1-4\$. . . . .  
ao^th-er+ah=v@1\_1/A:1\_1\_2/B:0-0-1@2-1&2-6#1-4\$. . . . .  
th^er-ah+v=dh@1\_2/A:0\_0\_1/B:1-0-2@1-1&3-5#1-3\$. . . . .  
er^ah-v+dh=ax@2\_1/A:0\_0\_1/B:1-0-2@1-1&3-5#1-3\$. . . . .  
ah^v-dh+ax=d@1\_2/A:1\_0\_2/B:0-0-2@1-1&4-4#2-3\$. . . . .  
v^dh-ax+d=ey@2\_1/A:1\_0\_2/B:0-0-2@1-1&4-4#2-3\$. . . . .

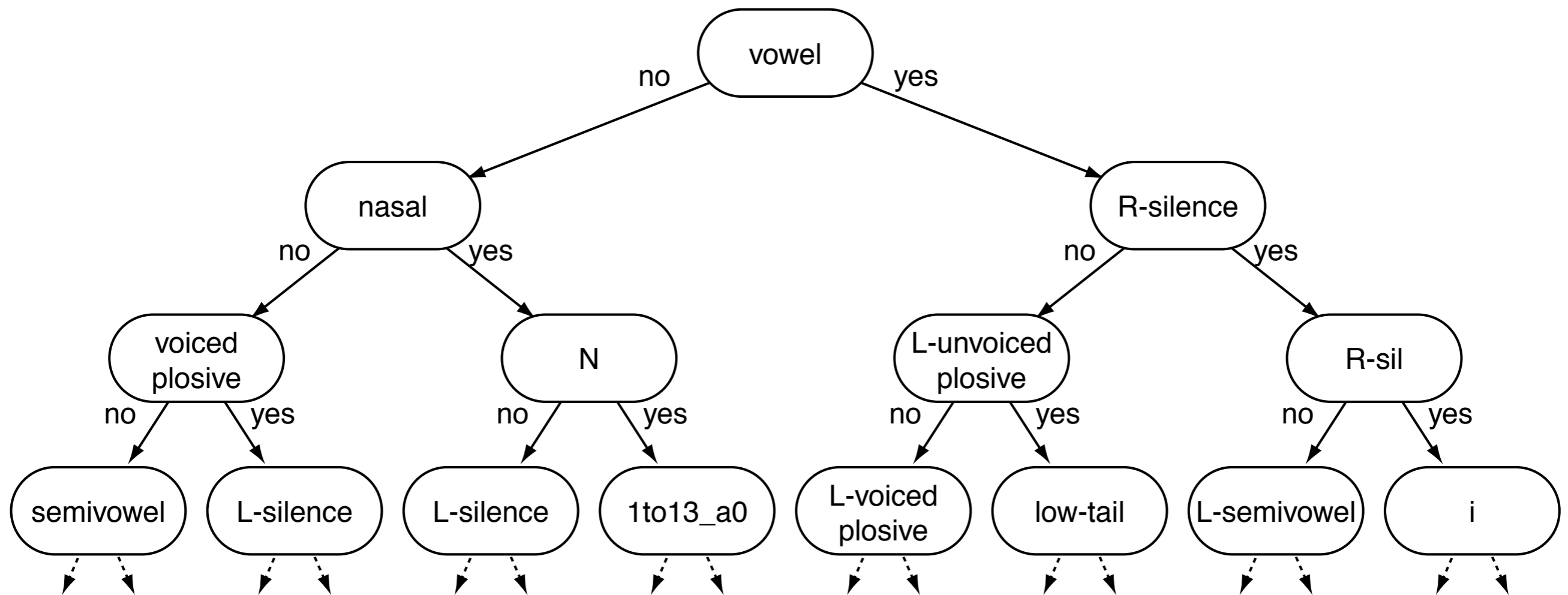
“Author of the . . . .”

# Too many model parameters

---

- The model is highly context dependent
  - as a consequence, many (in fact, most) model parameters will not have any corresponding training data
- The standard solution to this comes from ASR
  - share parameters across similar contexts
  - but how to determine which contexts are similar?
- Decision-tree based **context clustering**
  - Cluster together groups of parameters which have the same values for some subset of the linguistic specification (i.e., their contexts are similar)
  - Allows us to create parameters for contexts never seen in the training data
  - Automatically scales the model complexity (number of free parameters) to suit the available data

# Decision tree context clustering



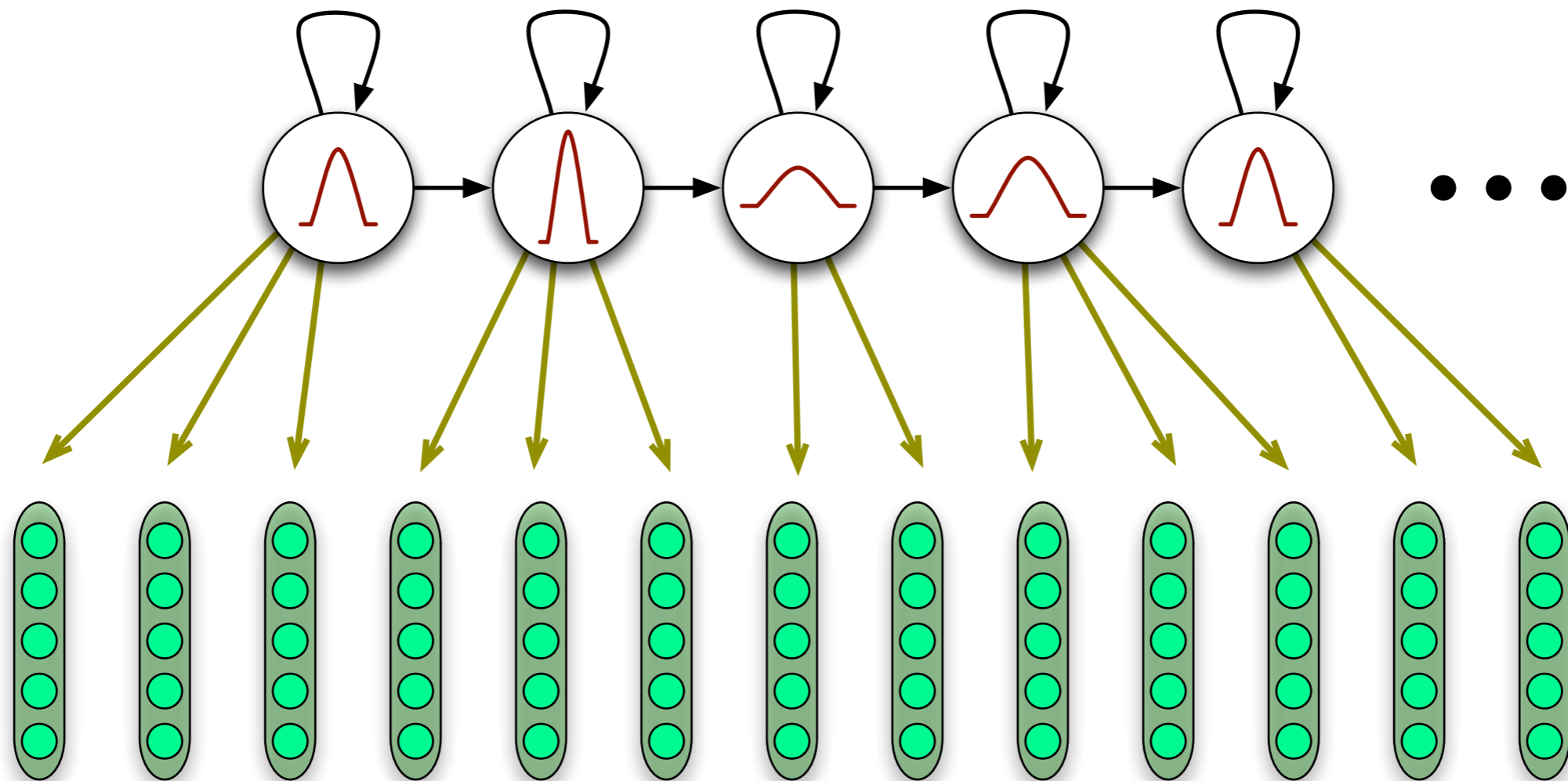


# HMM-based speech synthesis

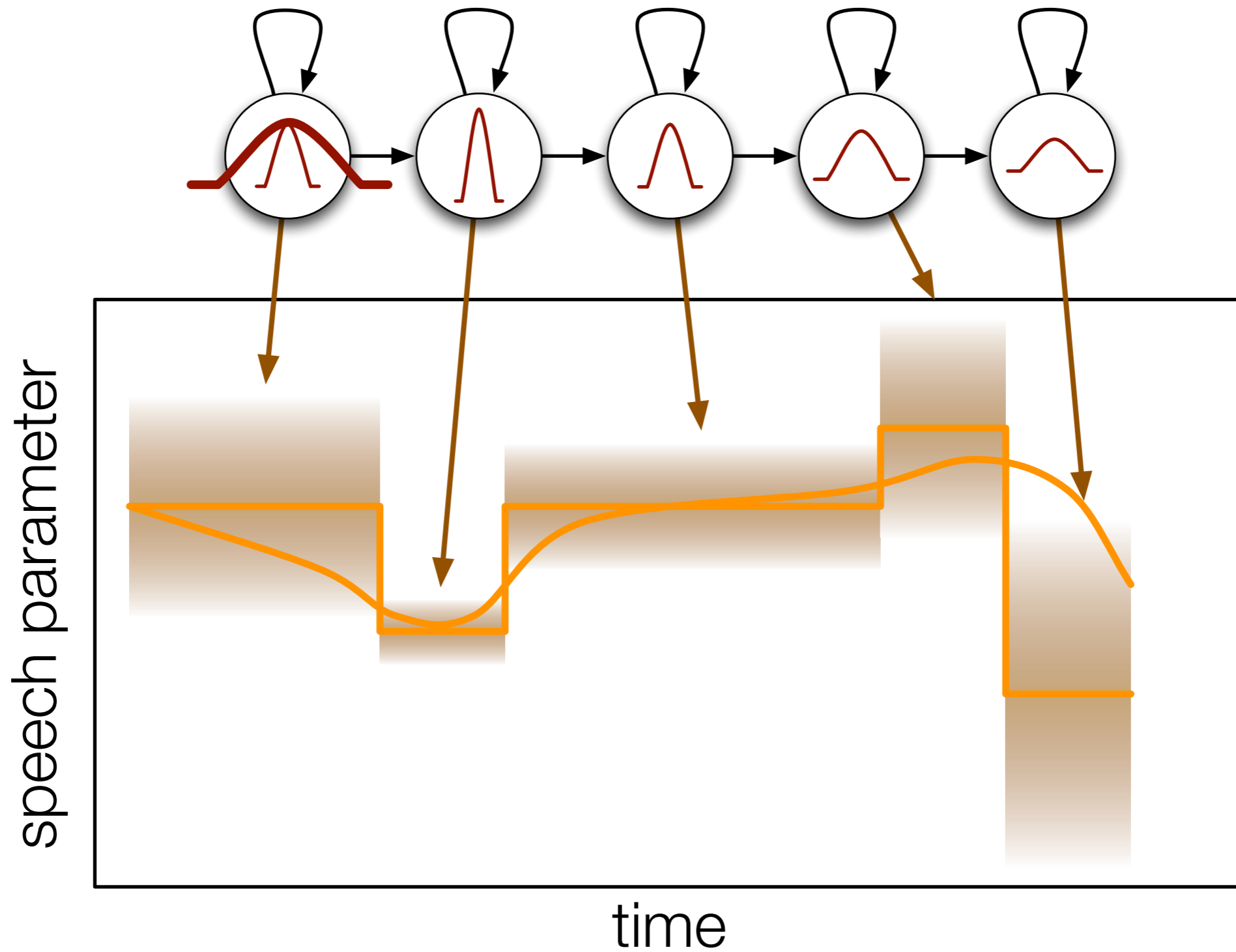
---

- Front end processes text and produces a linguistic specification
- Linguistic specification is flattened: a sequence of context-dependent phonemes
- HMMs are used to generate sequences of speech (in a parameterised form which we call ‘speech features’)
- From the parameterised form, we can generate a waveform using a vocoder
- The parameterised form contains sufficient information to generate speech:
  - spectral envelope
  - fundamental frequency (F0) - sometimes called ‘pitch’
  - aperiodic (noise-like) components (e.g. for sounds like ‘sh’ and ‘f’)

# HMMs are generative models



# Trajectory generation



# Comparison with ASR

---

- Differences from automatic speech recognition include
  - Synthesis uses a much richer model set, with a lot more context
    - For speech recognition: triphone models
    - For speech synthesis: “full context” models
      - “Full context” = both phonetic and prosodic factors
  - Observation vector for HMMs contains the necessary parameters to generate speech using a vocoder, such as spectral envelope, F0 & multi-band aperiodic energy amplitudes

# Comparison with unit-selection synthesis

---

- System construction
  - Training and optimising an HMM-based speech synthesiser is
    - almost entirely **automatic**
    - based on **objective** measures (maximum likelihood criterion, minimum cepstral distortion, etc)
  - Optimising a unit selection system (e.g., choosing target cost weights) is usually
    - done by **trial and error**
    - based on **subjective** measures (listening).
  
- Data
  - Unit selection needs 5-10,000+ sentences of data from one speaker
  - Training a speaker-dependent HMM needs 1000+ sentences
  - Adapting a speaker-independent HMM needs 1-100 sentences

# Latest Samples

---



**English Sample**



**Spanish Sample**

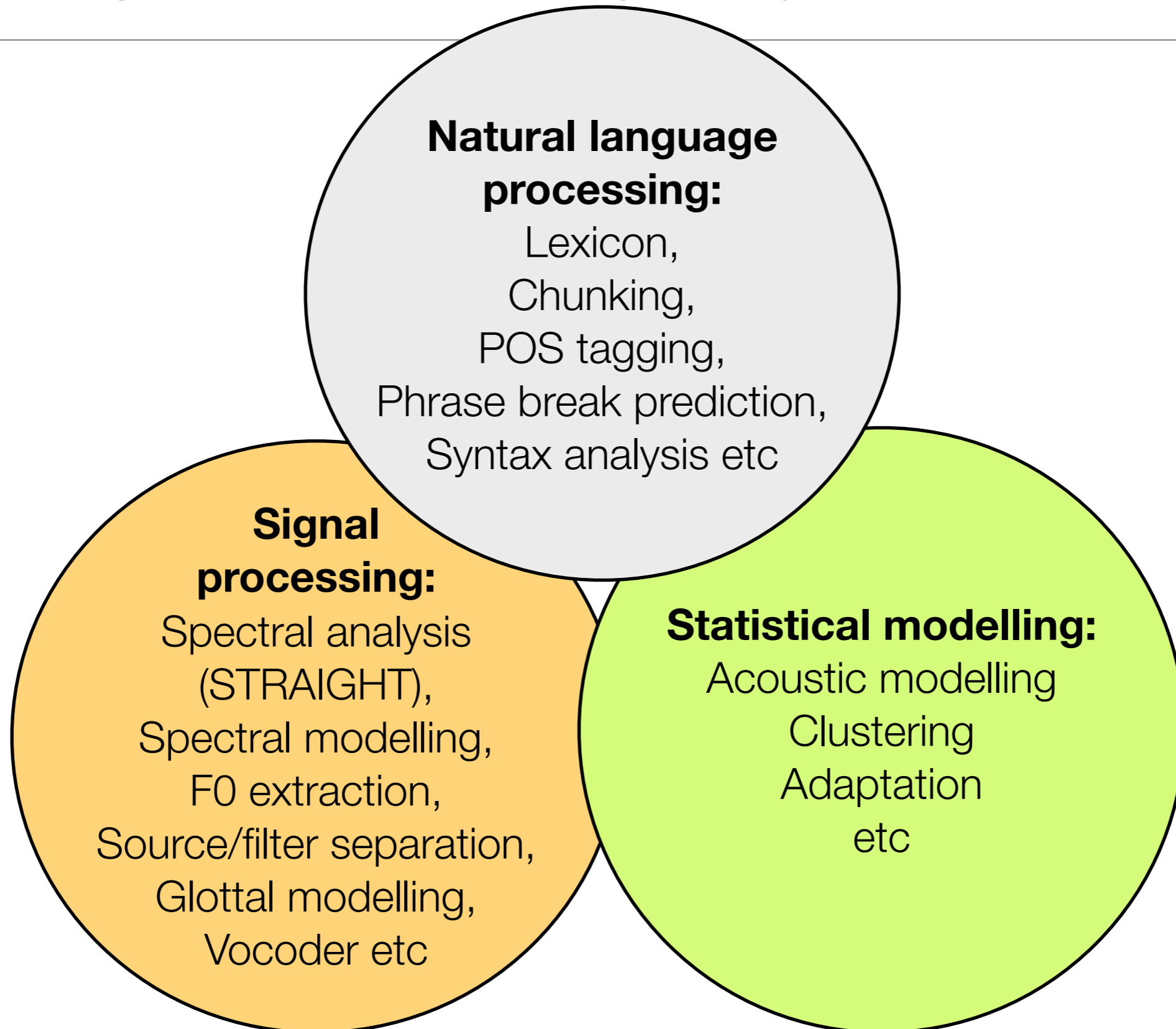


**Romanian Sample**

- 48kHz sampling frequency,
- STRAIGHT bark cepstrum, pitch in mel, Bark-critical-band limited aperiodicity measures, Mixed excitation plus group delay phase manipulation,
- Context-dependent state-tied MSD-HSMM with semi-tied covariance

# Text-to-speech: multidisciplinary research

---



# Contents

---

- Background
  - a brief introduction to HMM-based speech synthesis
  - **basic principles of adaptation**
- Applications
  1. Voice cloning
  2. Voice reconstruction
  3. Personalised speech-to-speech translation
  4. Articulatory-controllable speech synthesis



# The basic principles of adaptation

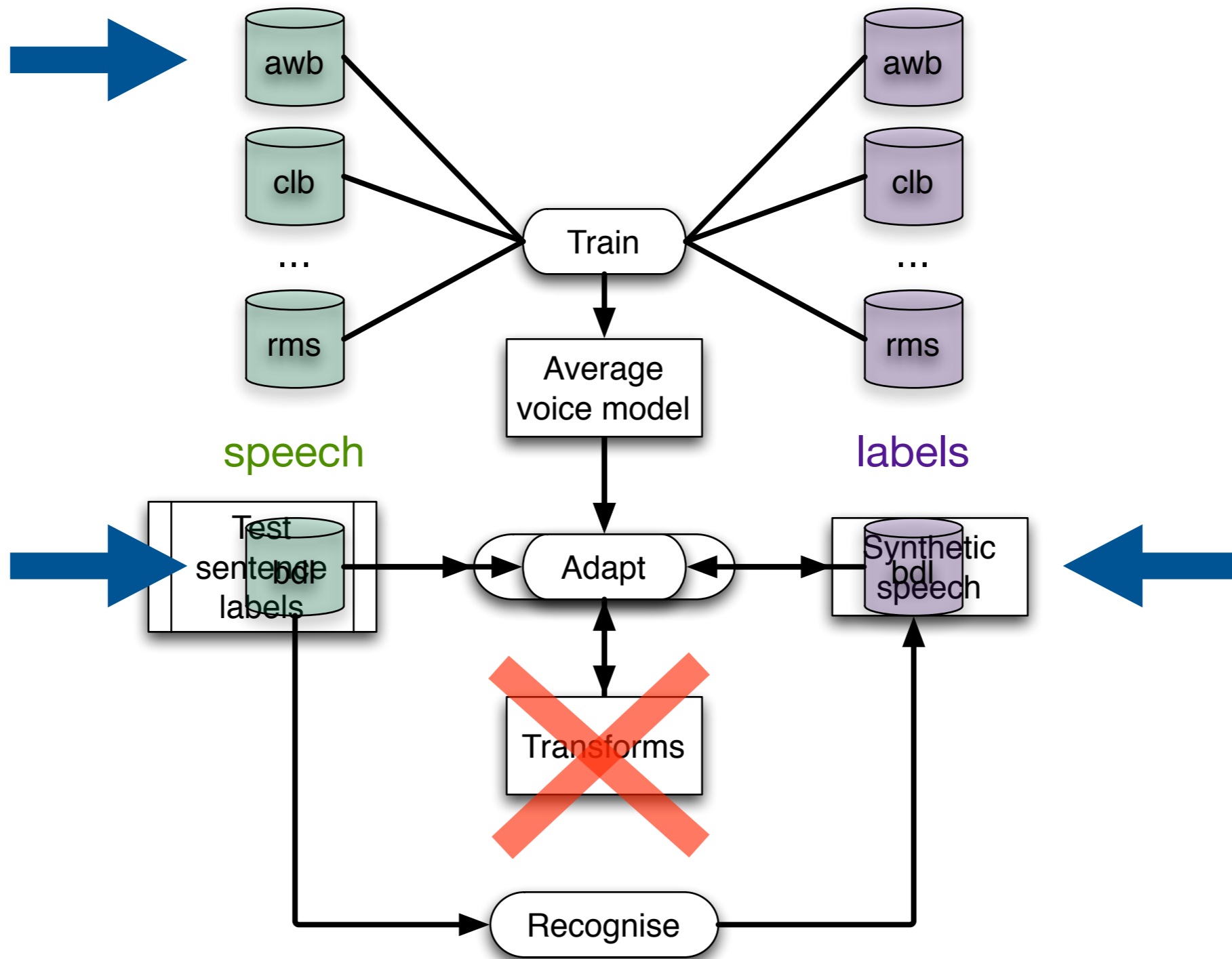
---

Most common use: speaker adaptation

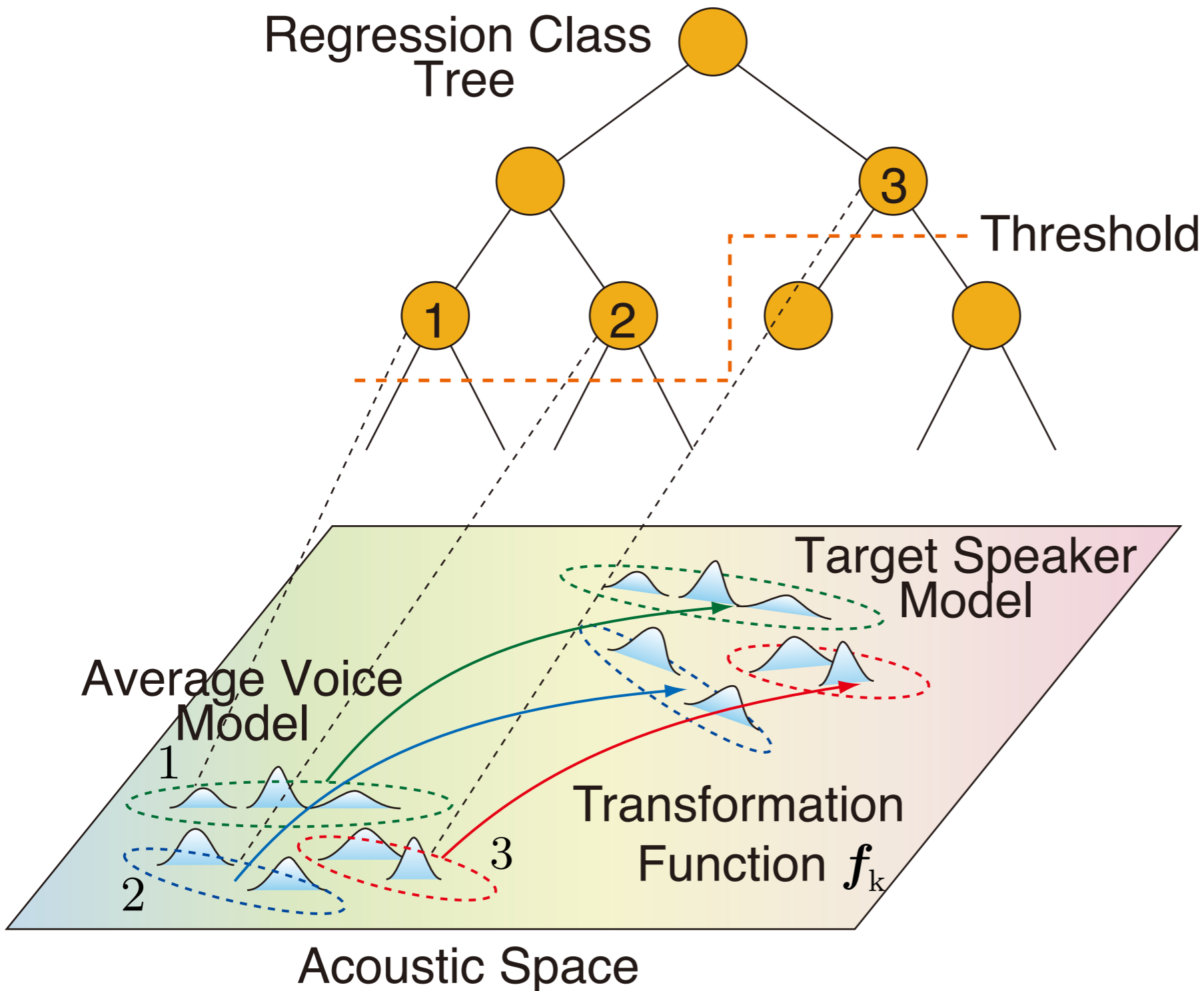
# Speaker adaptation

---

- One of the most important recent developments in speech recognition
- A **linear transform** is applied to every HMM parameter (Gaussian mean and variance) in order to adapt the model to new data
- Can be used to create new voices for speech synthesis:
  - Train HMM on lots of data from multiple speakers
  - Adapt this HMM using a small amount of data from the target speaker
    - 100 sentences is usually enough
- This is a very exciting development in speech synthesis
- Provided data are available, any other acoustic difference can be adapted
  - speaking style/emotion
  - dialect/accents
  - speaker identity across languages
  - and so on...



# ML-based linear regression



Transformation Function  $f_k$

AMCC

$$\bar{\mu}_i = \mu_i + \epsilon_k$$

MLLR

$$\bar{\mu}_i = \zeta_k \mu_i + \epsilon_k$$

CMLLR

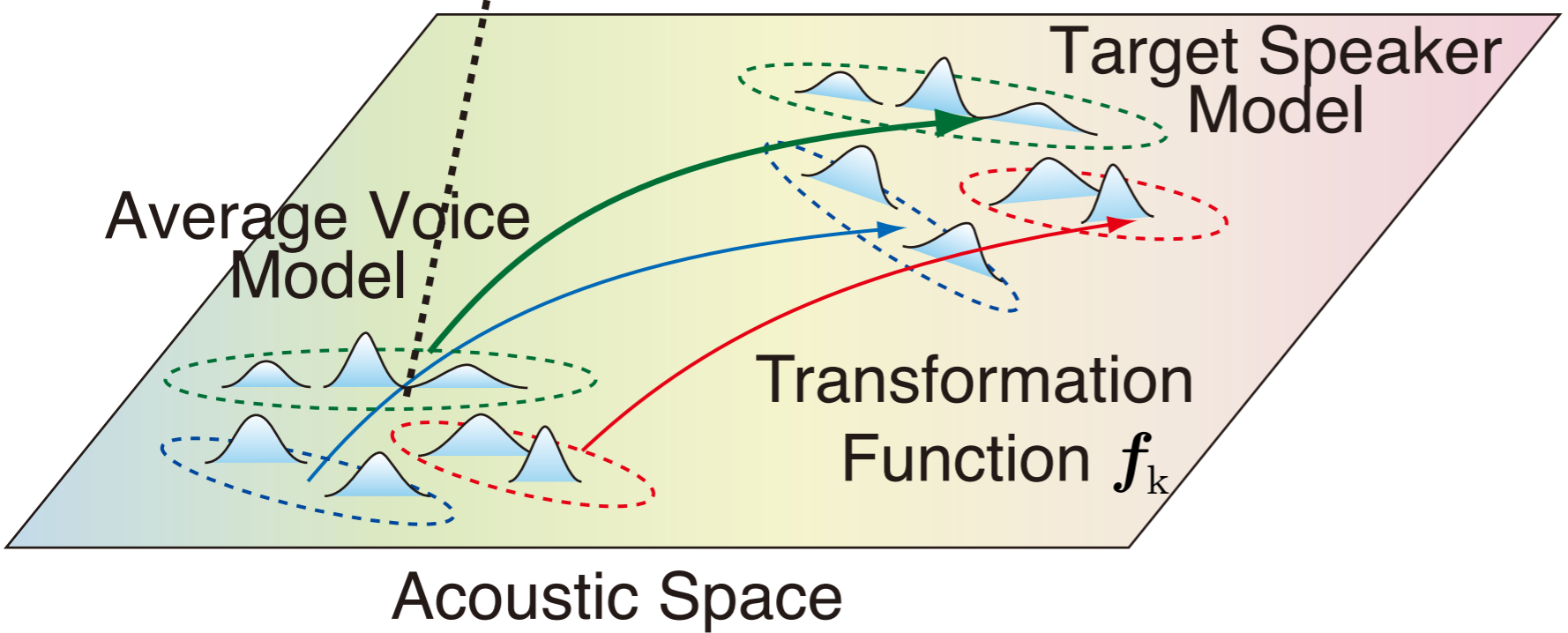
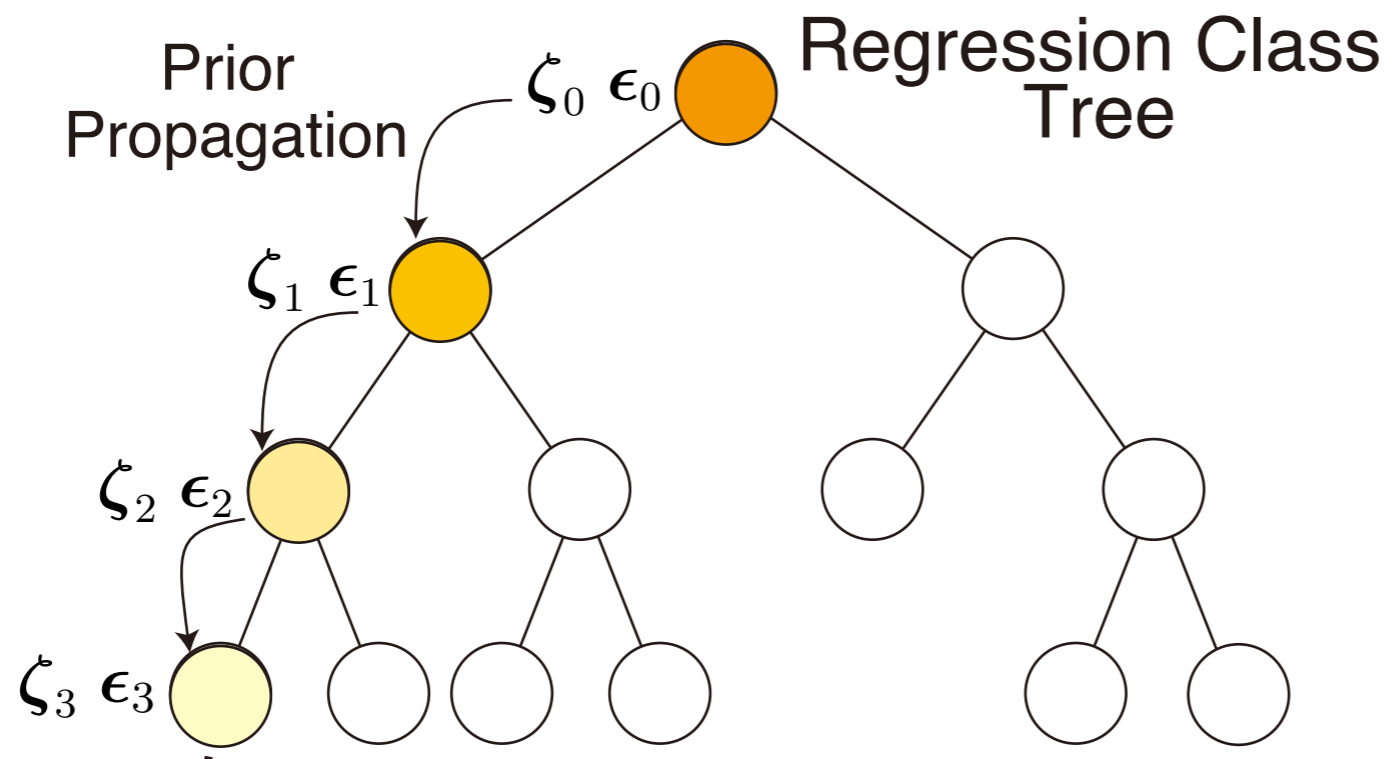
$$\bar{\mu}_i = \zeta_k \mu_i + \epsilon_k$$

$$\bar{\Sigma}_i = \zeta_k \Sigma_i \zeta_k^T$$

$\mu_i$  Mean Vector of Gaussian pdf  $i$

$\Sigma_i$  Covariance Matrix of Gaussian pdf  $i$  28

# MAP-based linear regression



Transformation Function  $f_k$

**SMAP**

$$\bar{\mu}_i = \mu_i + \epsilon_k$$

**SMAPLR**

$$\bar{\mu}_i = \zeta_k \mu_i + \epsilon_k$$

**CSMAPLR**

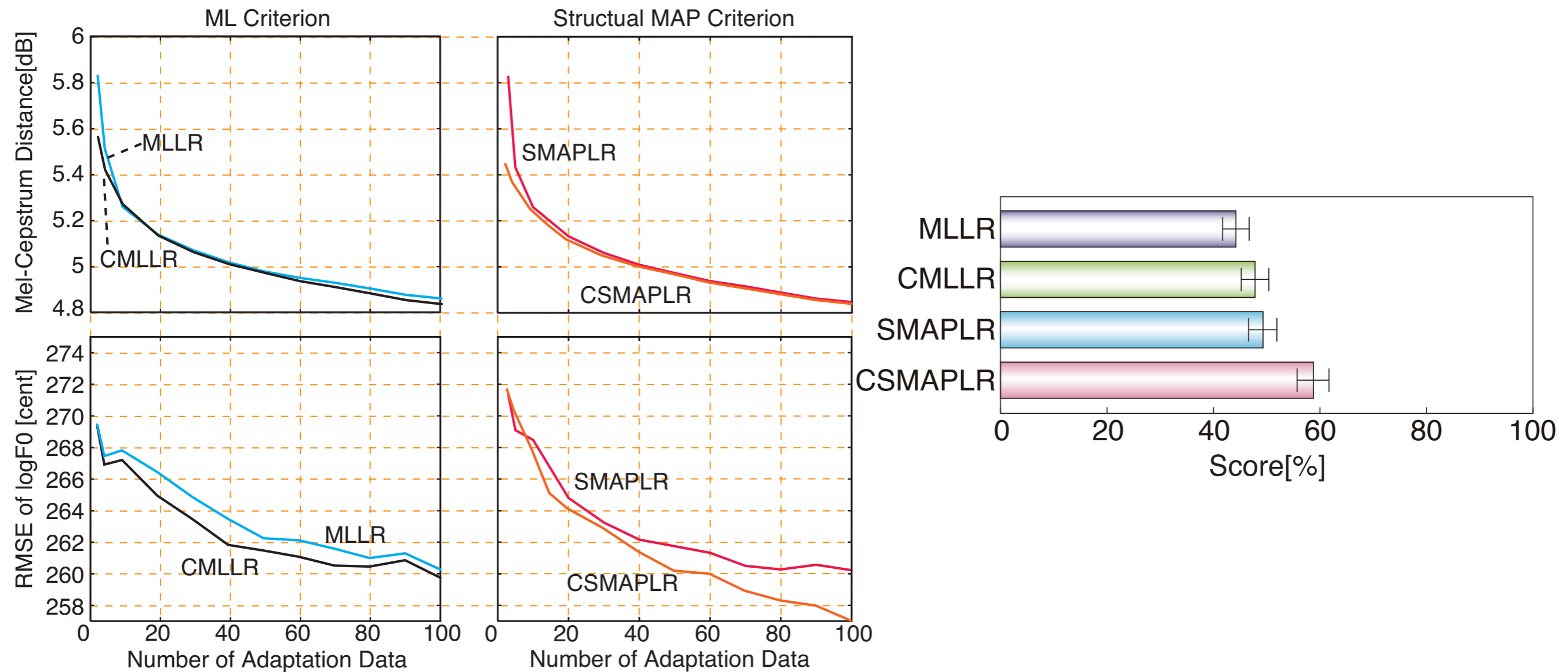
$$\bar{\mu}_i = \zeta_k \mu_i + \epsilon_k$$

$$\bar{\Sigma}_i = \zeta_k \Sigma_i \zeta_k^T$$

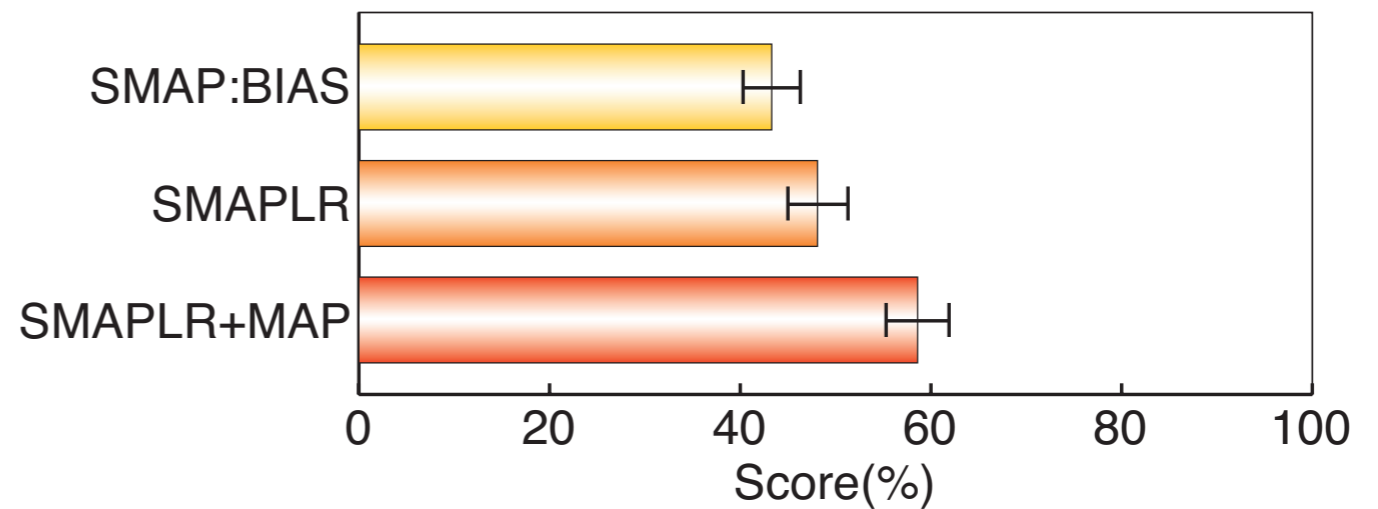
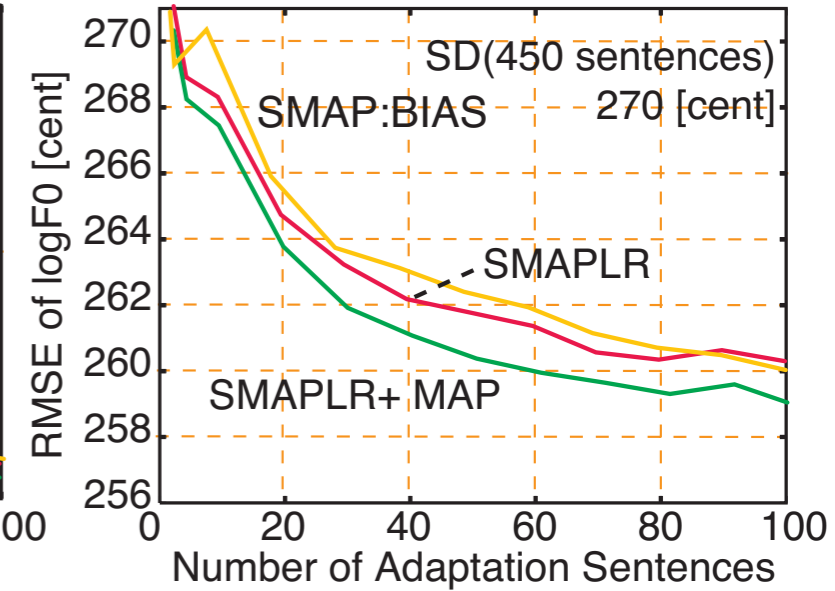
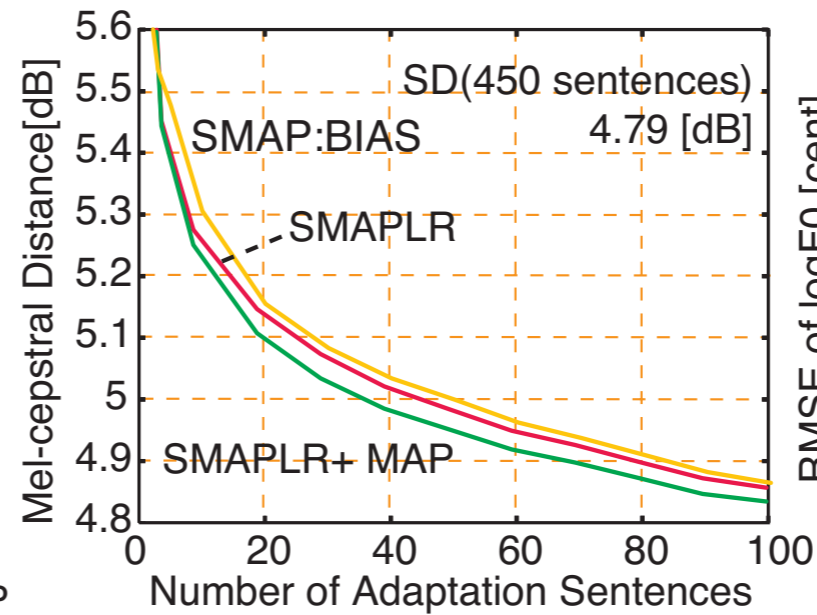
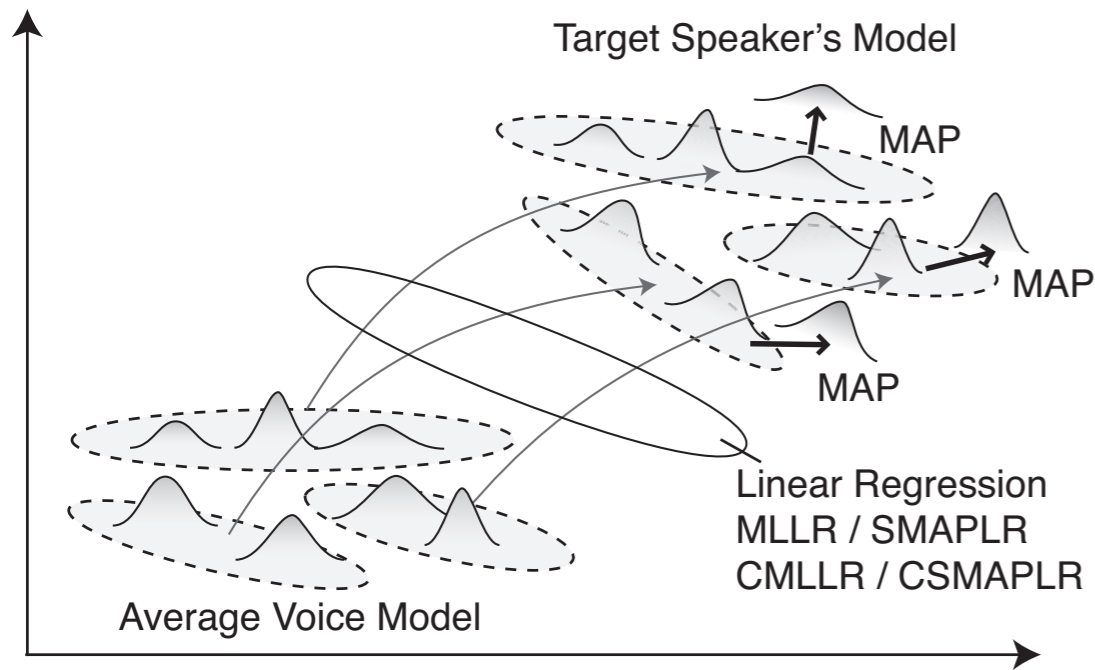
$\mu_i$  Mean Vector of Gaussian pdf  $i$

$\Sigma_i$  Covariance Matrix of Gaussian pdf  $i$

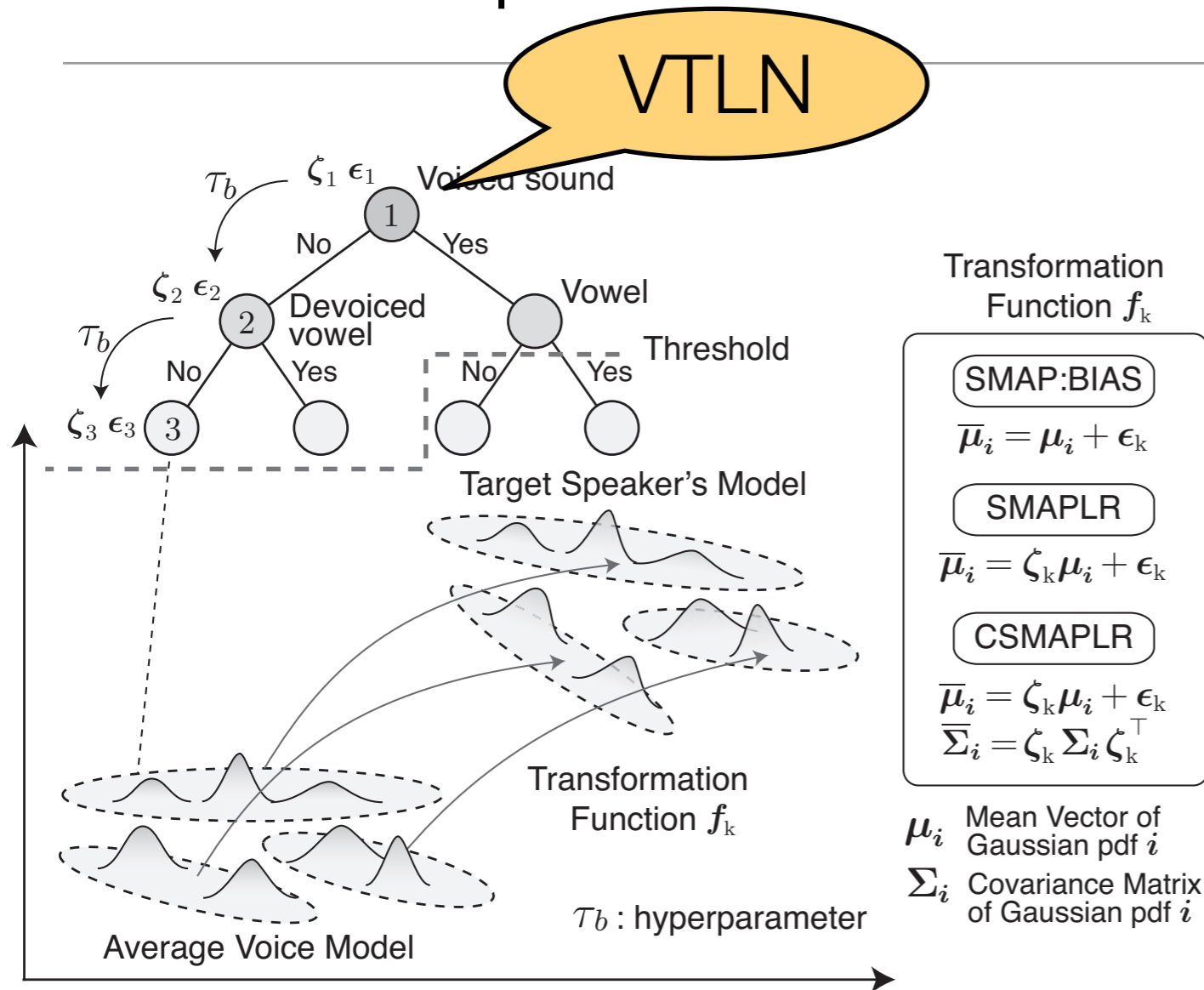
# Comparative study of linear transforms



# MLLR followed by MAP



# The VTLN prior for CSMAPLR



- EM estimation of an optimal warping parameter for VTLN (all-pass filter)
- Represent the estimated VTLN function as a CMLLR feature space matrix
- Use the VTLN-CMLLR matrix as a prior for the MAP estimation of CMLLR matrices at the root nodes
- Structural recursive MAP estimation of CMLLR matrices at lower child nodes

Prior at the root node	5 sentences	2 sentences	1 sentence
Identity	●	●	●
VTLN			●



# Speaker adaptation techniques for TTS

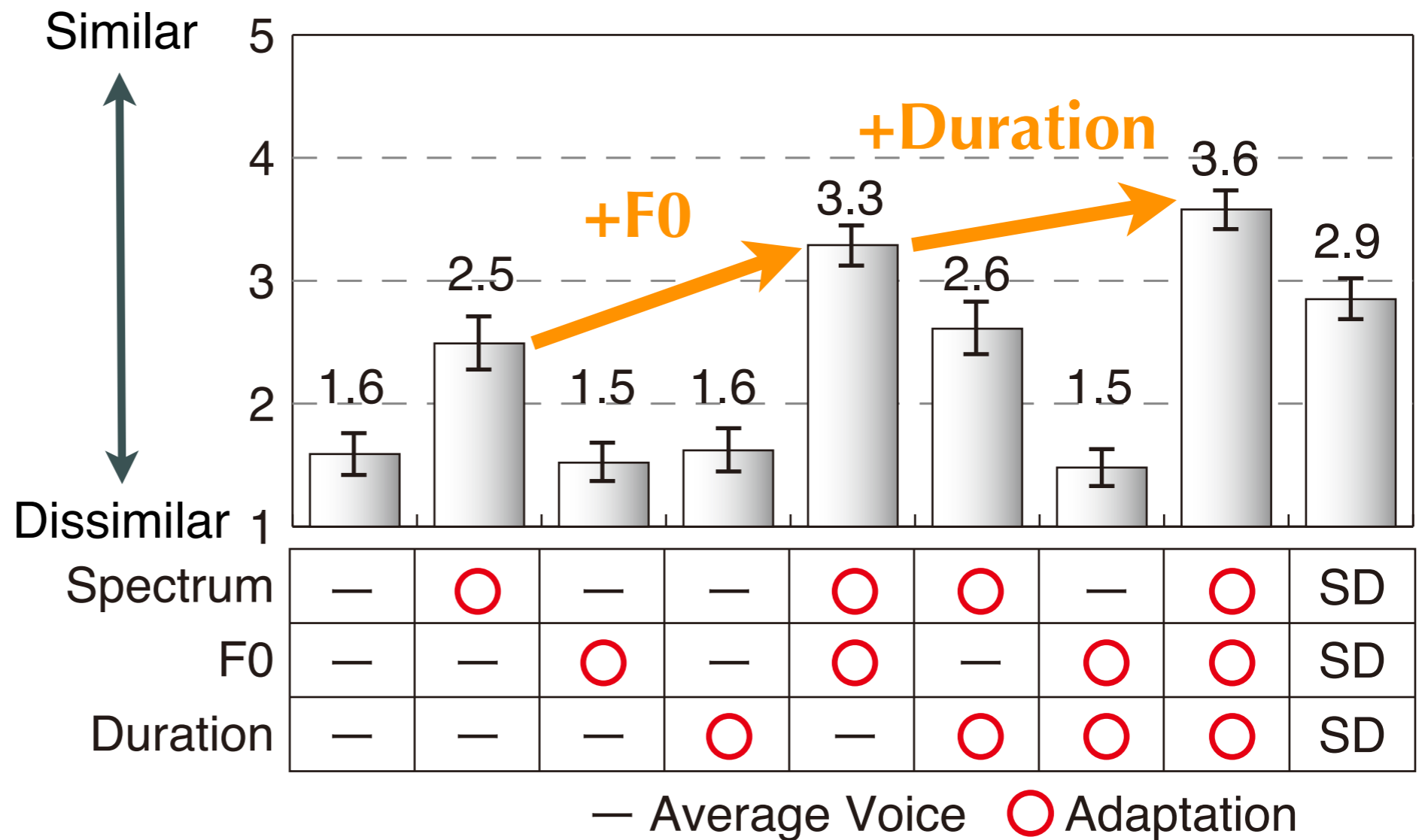
---

Differences between ASR adaptation and TTS adaptation

# Simultaneous adaptation of spectrum, fundamental frequency and duration



Method	CCR Test
# of Subjects	8 persons
# of Test Sentences	5 sentences
# of Adaptation Data	100 sentences

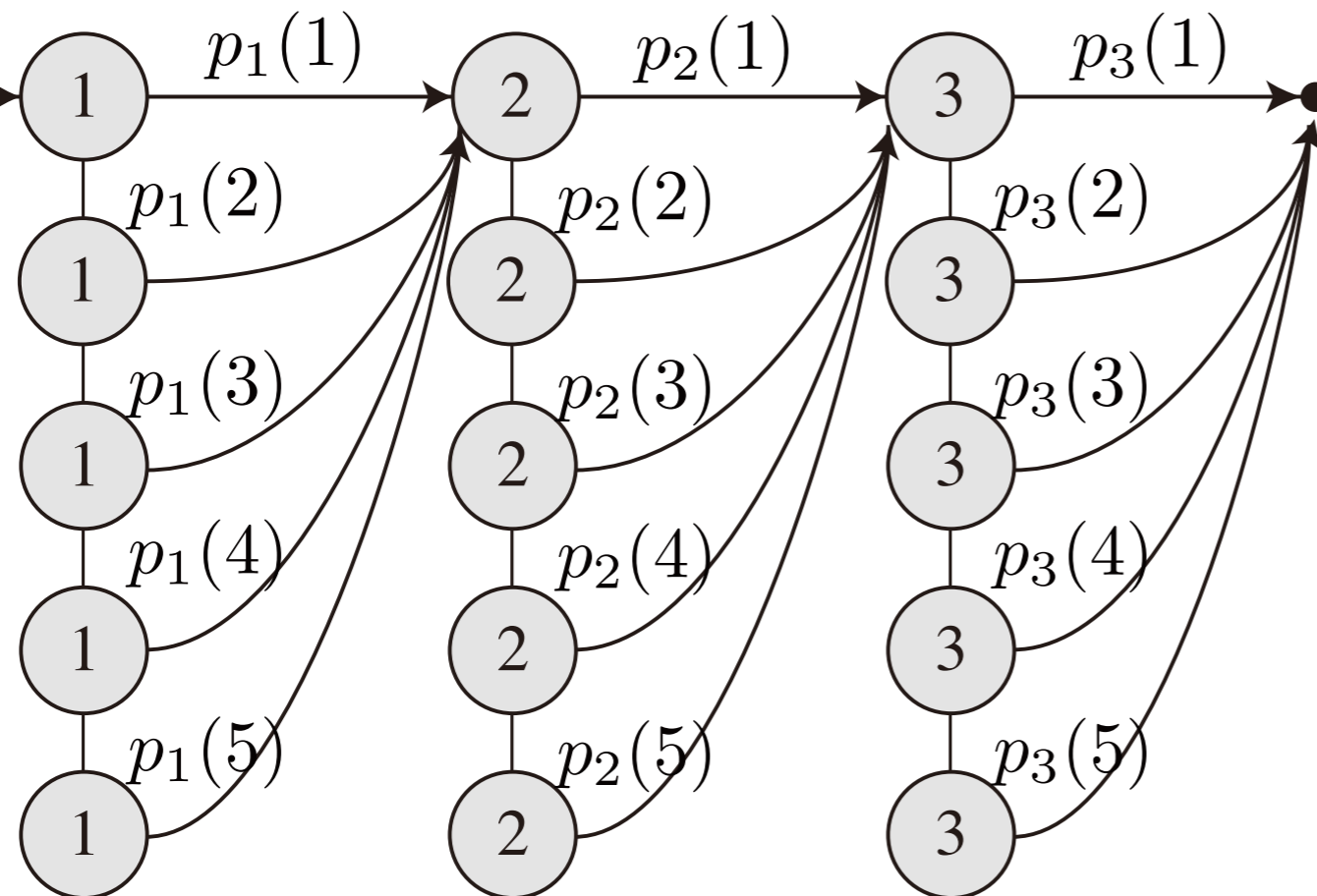


# Duration adaptation



## Application to hidden-semi Markov models (HSMM)

[S.E. Levinson '86]



HSMM: HMM with explicit duration pdfs

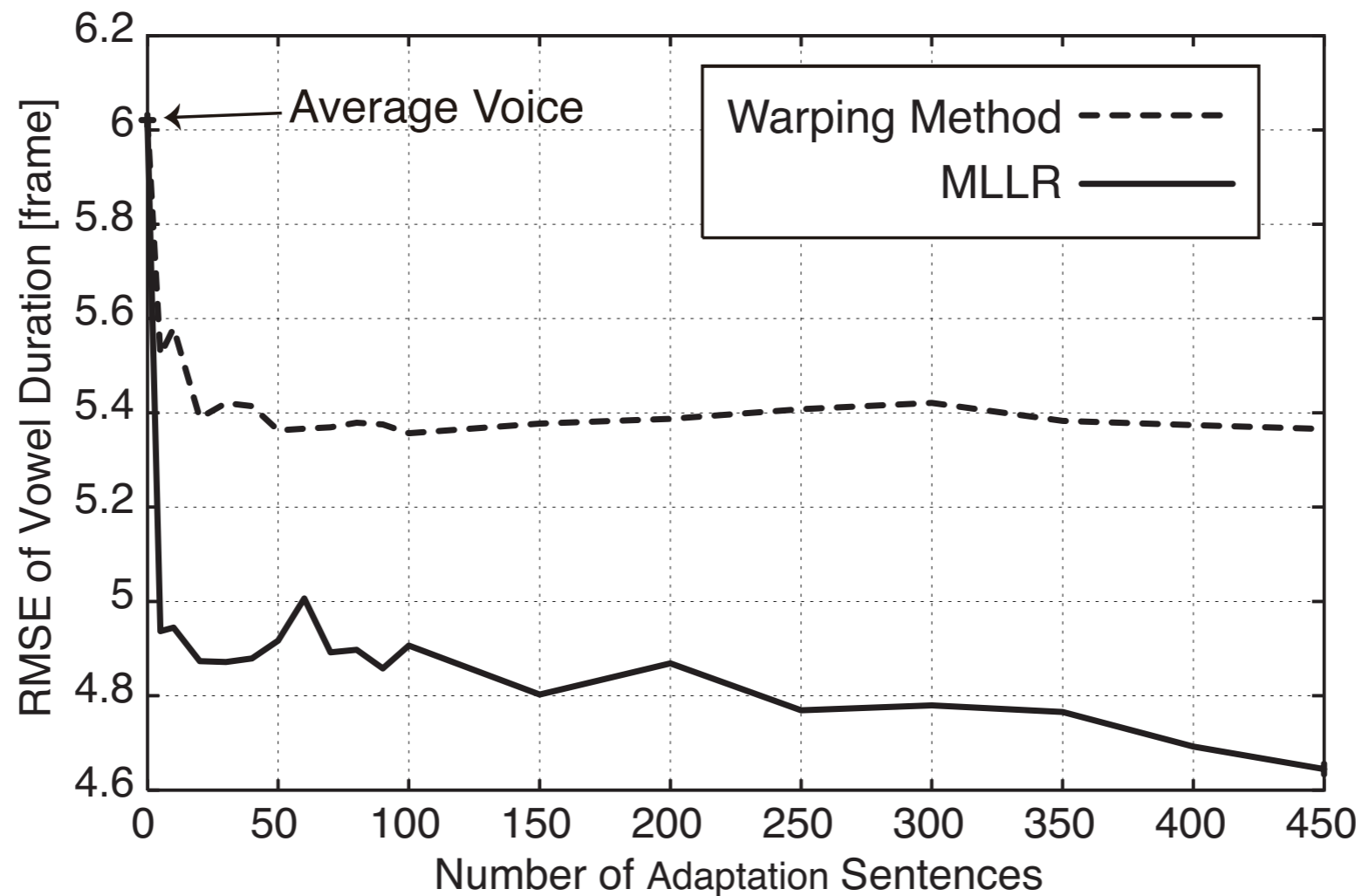
SAT for duration pdfs:

- (1) Linear transform of duration
- (2) Inverse estimation of parameters for duration pdfs

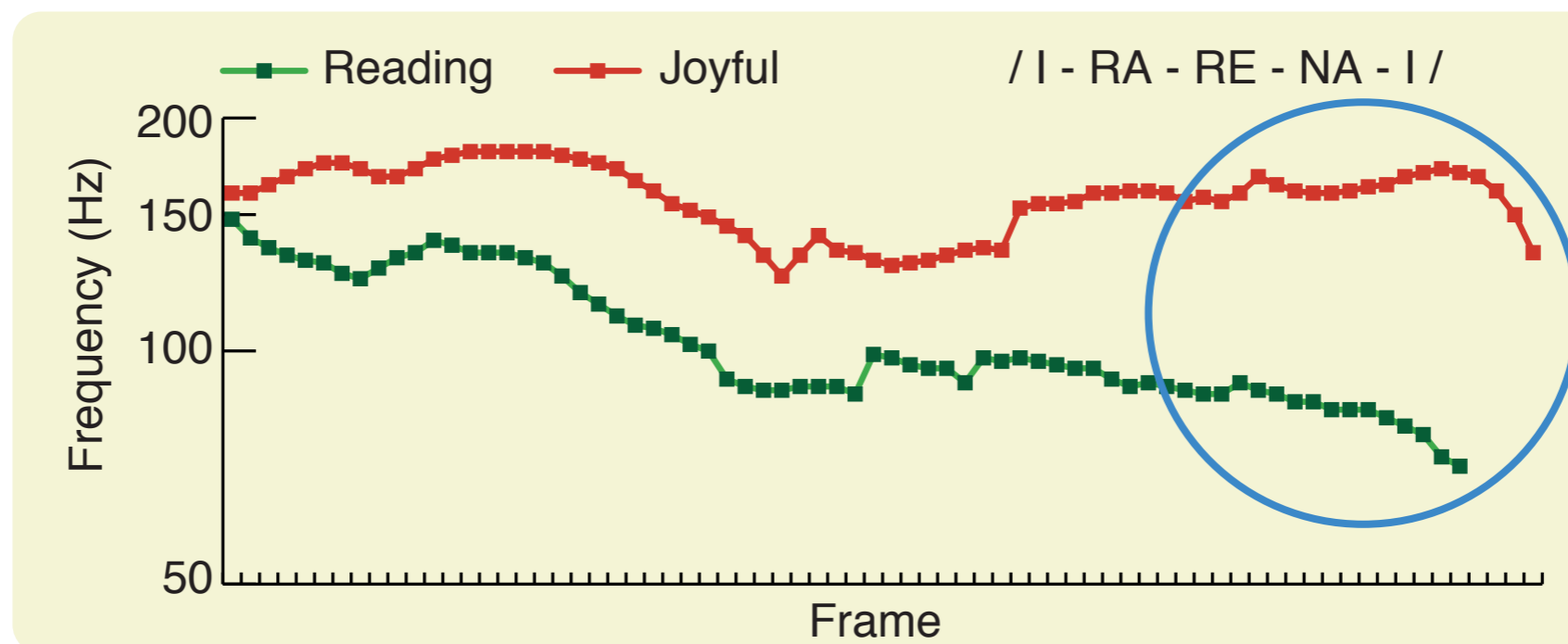
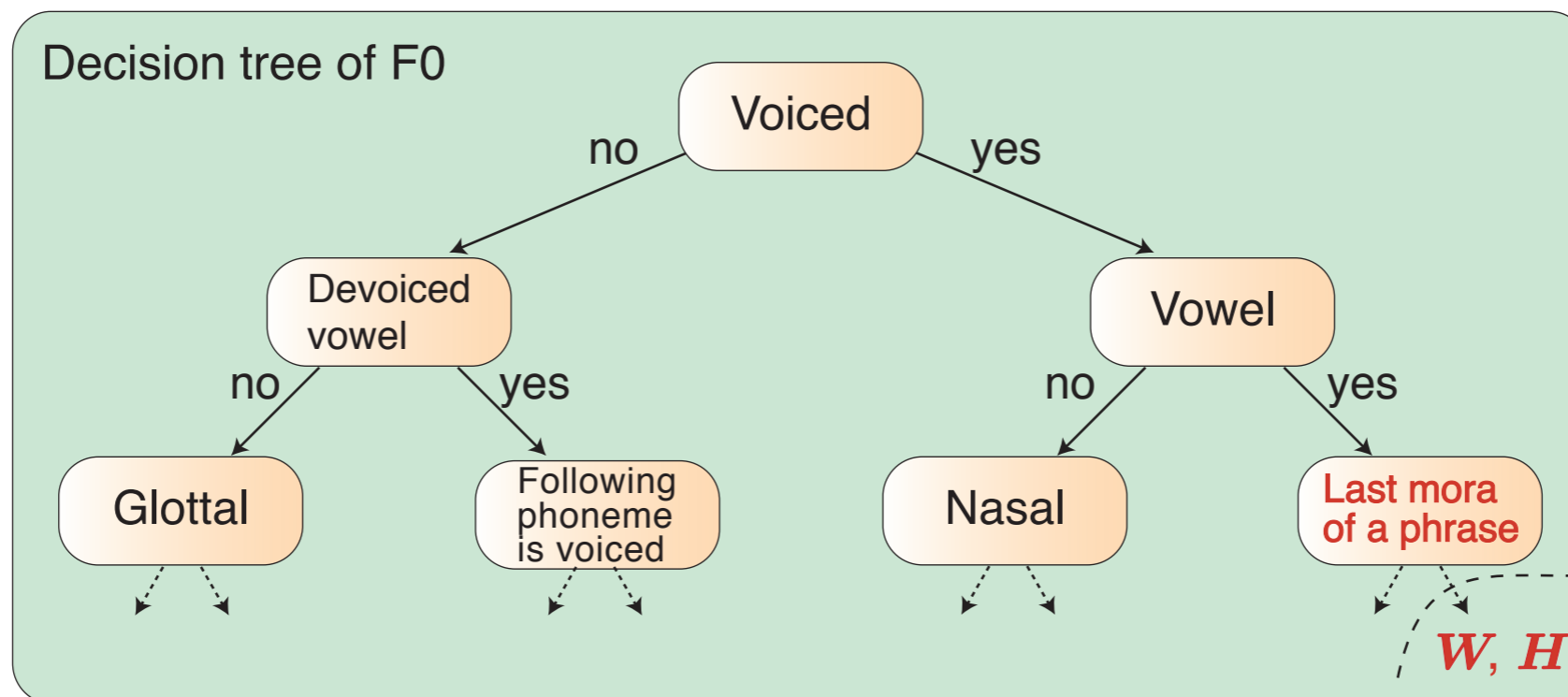
$p_i(d)$  : Duration Probability

$$p_i(d) = |a| \mathcal{N}(ad + b; \mu_i, \sigma_i^2)$$

# Comparison with a simple duration warping method



# The use of context decision trees as regression class trees





# A few examples of adaptation

---

- An American English female average voice to a 7-year-old girl
  - Average voice
  - Synthetic speech generated from adapted models
- An American English male speakers' average voice to an Indian English speaker
  - Average voice
  - Synthetic speech generated from adapted models
- Adaptation from neutral style to anger with gradual linear interpolation

# Contents

---

- Background
  - a brief introduction to HMM-based speech synthesis
  - basic principles of adaptation
- Applications
  - 1. Voice cloning**
  2. Voice reconstruction
  3. Personalised speech-to-speech translation
  4. Articulatory-controllable speech synthesis

# Voice cloning

---

Celebrity voices

Virtually unlimited number of voices

Voice banking



# Voice cloning

---

- What do we mean by ‘voice cloning’ ?
  - automatically creating synthetic voices from relatively small recordings
- How is it different from conventional voice building?
  - less data
  - lower quality recordings
  - non-professional speakers
  - less (or no) manual intervention
- What’s the point?
  - fully automatic voice creation
  - cheap mass-production of voices
  - huge variety of voices possible using existing data

# Application 1 – Celebrities voices

---

Speech data can be acquired from broadcast, podcasts, lectures, telephone

Synthetic speech samples created in this scenario

George W Bush podcast:

Synthetic speech samples generated from HMMs adapted using speech data found on his podcasts

Sample 

[Real-time demo \[web\]](#)

Queen Elizabeth-II's podcast

Synthetic speech samples generated from HMMs adapted using speech data found on her podcasts

Sample 

# Technical problems behind this app

---

- Technical problems of speech synthesis using found audio (e.g. for celebrity voice ) are conceptually similar to those of LVCSR
  - VAD
  - Noise reduction (spectral subtraction, noise gate etc)
  - **Unsupervised speaker adaptation**
  - etc

# Unsupervised adaptation systems [for BL 2009]

---



- Multi-pass architecture for unsupervised adaptation using AMIRT06 LVCSR
  - LVCSR
    - P1: VAD, speaker diarization followed by initial decoding using WFST decoder
    - P2: VTLN and calculation of posteriori/tandem features
    - P3: CMLLR estimation, followed by using MPE/VTLN/SAT model
    - P4: Lattice expansion (2gram to 4gram)
    - P5: CMLLR estimation using rescored 1-best hypothesis
    - P6: Confusion network using word posteriors
  - TTS
    - P7: Pruning based on confidence scores calculated from CN
    - P8: CMLLR/CSMAPLR estimation using TTS ML/SAT model

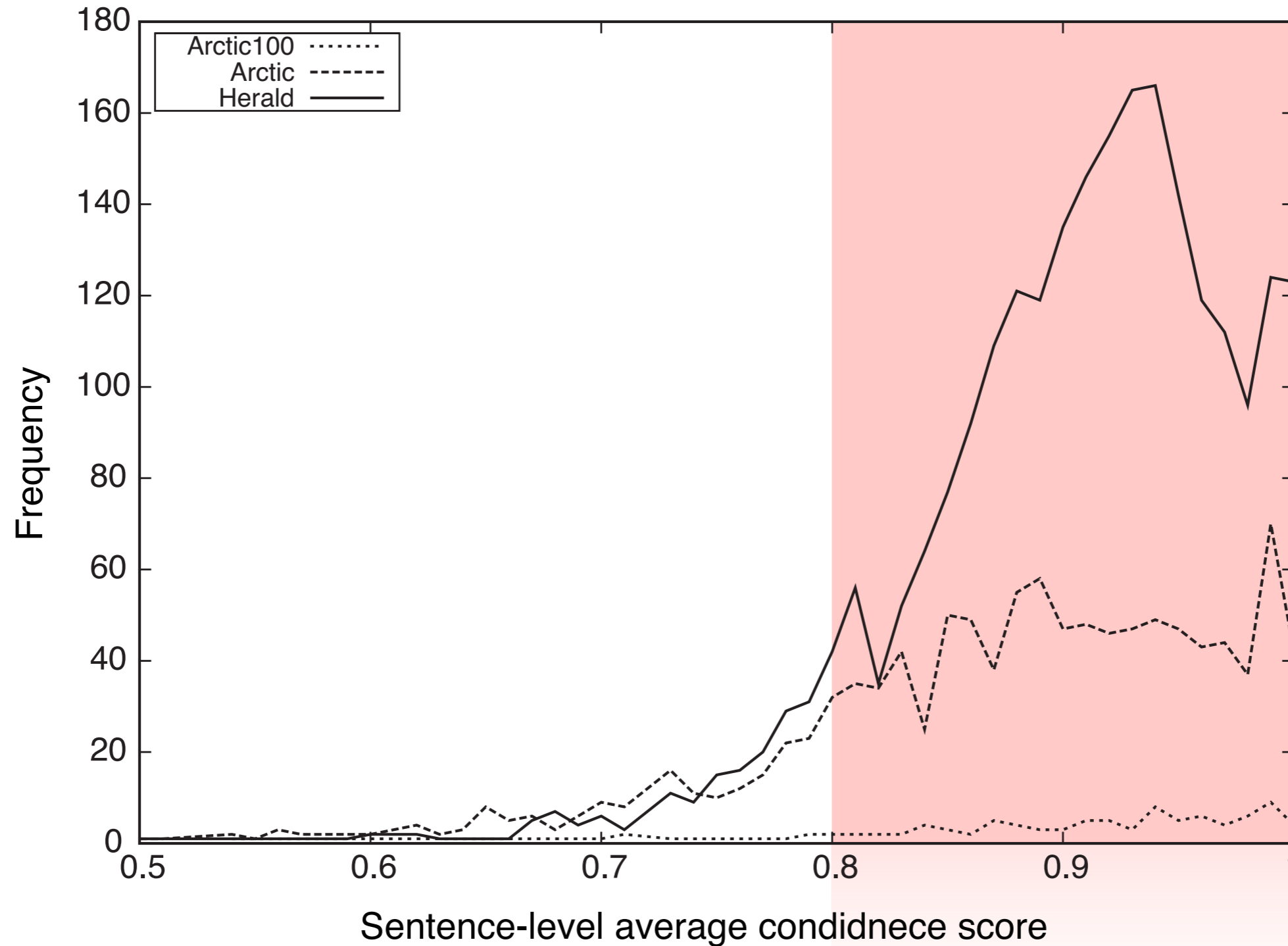
# ASR WERs

---

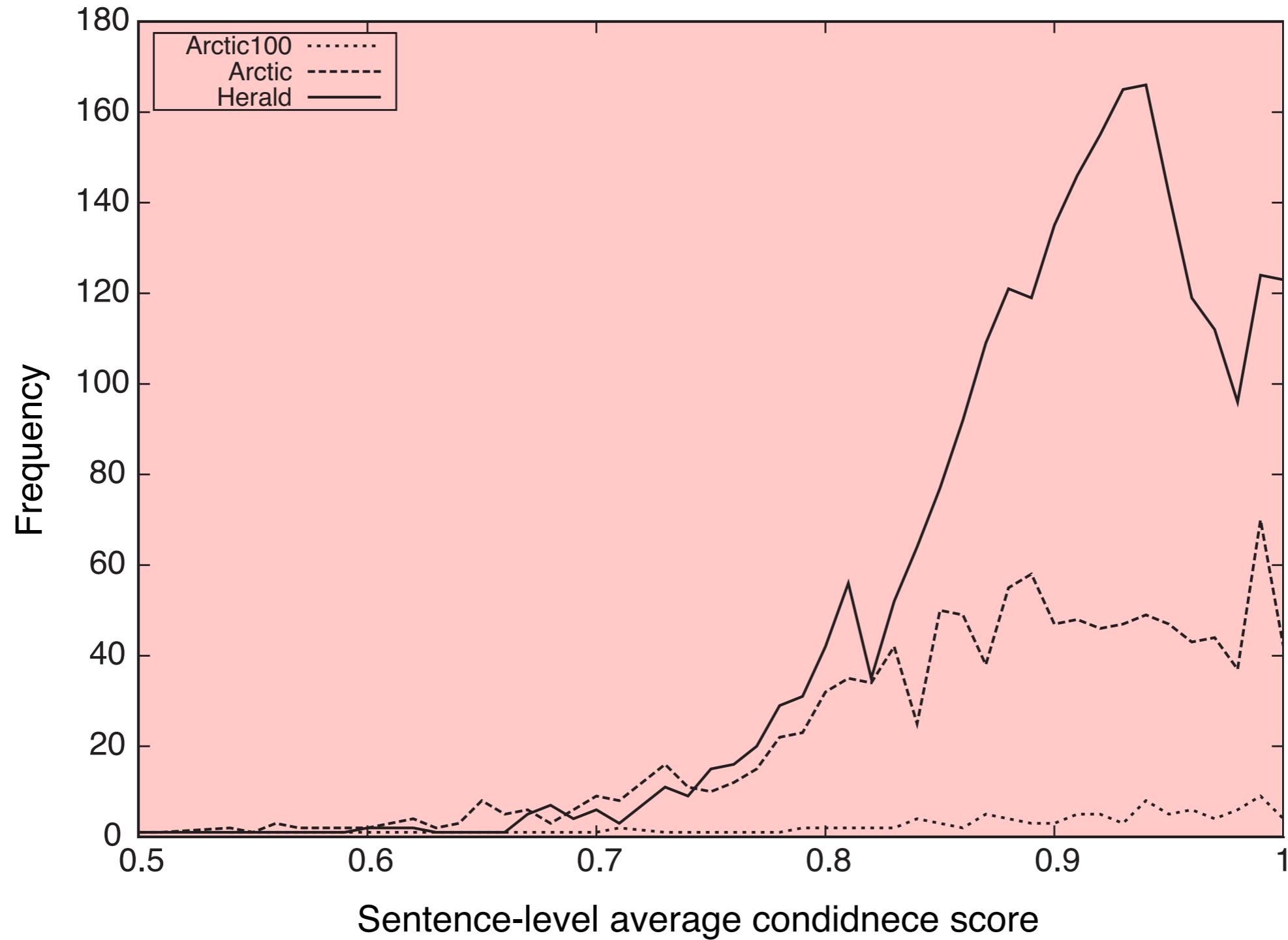
Table 7: WERs [%] obtained from the AMI 2006 RT system for each genre and pass.

Pass	Genre					
	Address	Arctic	Carroll	Herald 1	Herald 2	Herald 3
P1	47.3	41.7	58.4	40.2	36.6	34.3
P3	41.0	25.5	47.9	26.8	23.5	23.3
P6	40.8	27.5	47.8	28.1	24.8	24.2

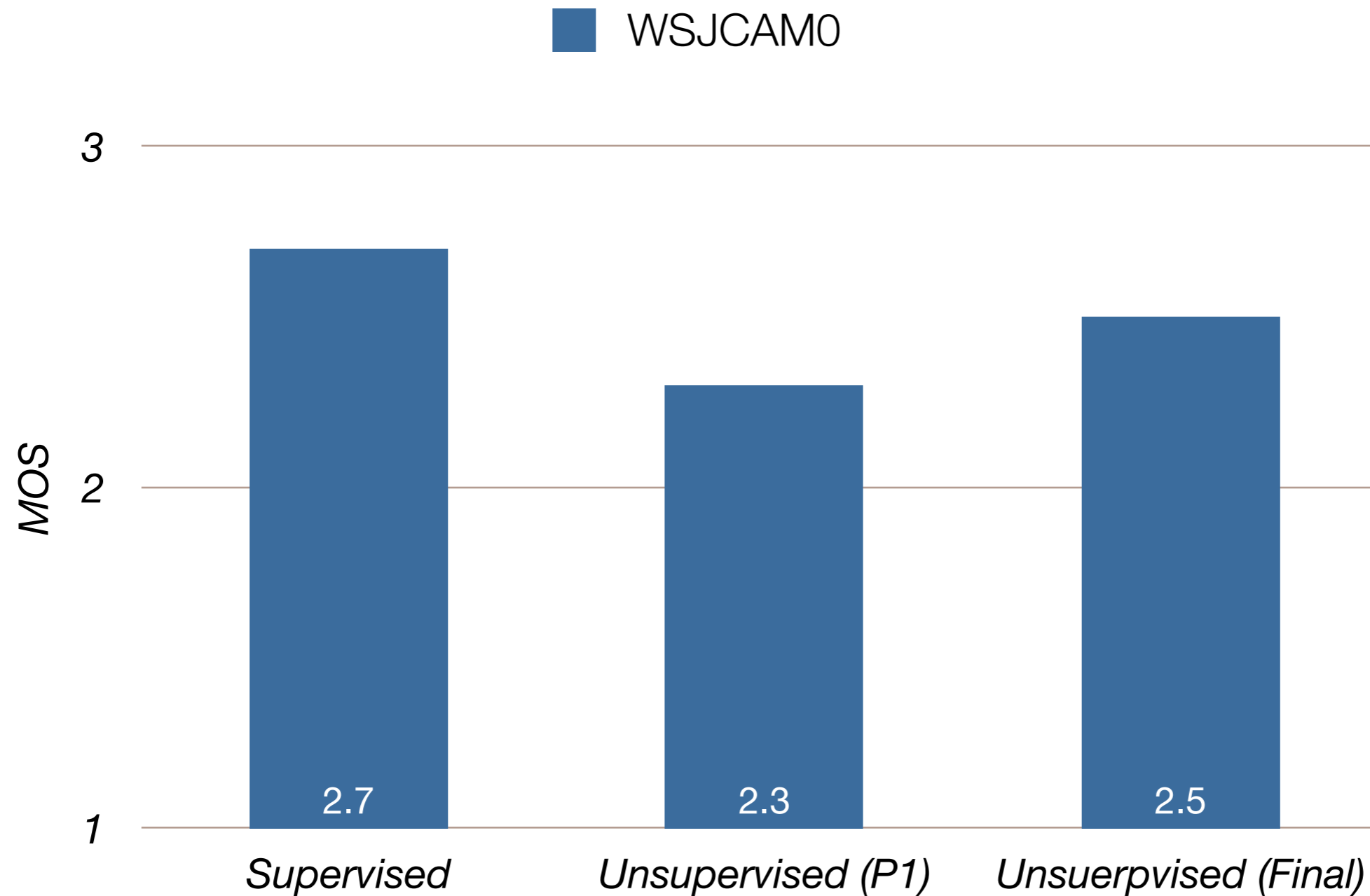
# Pruning based on confidence scores calculated from confusion network



# Audio samples

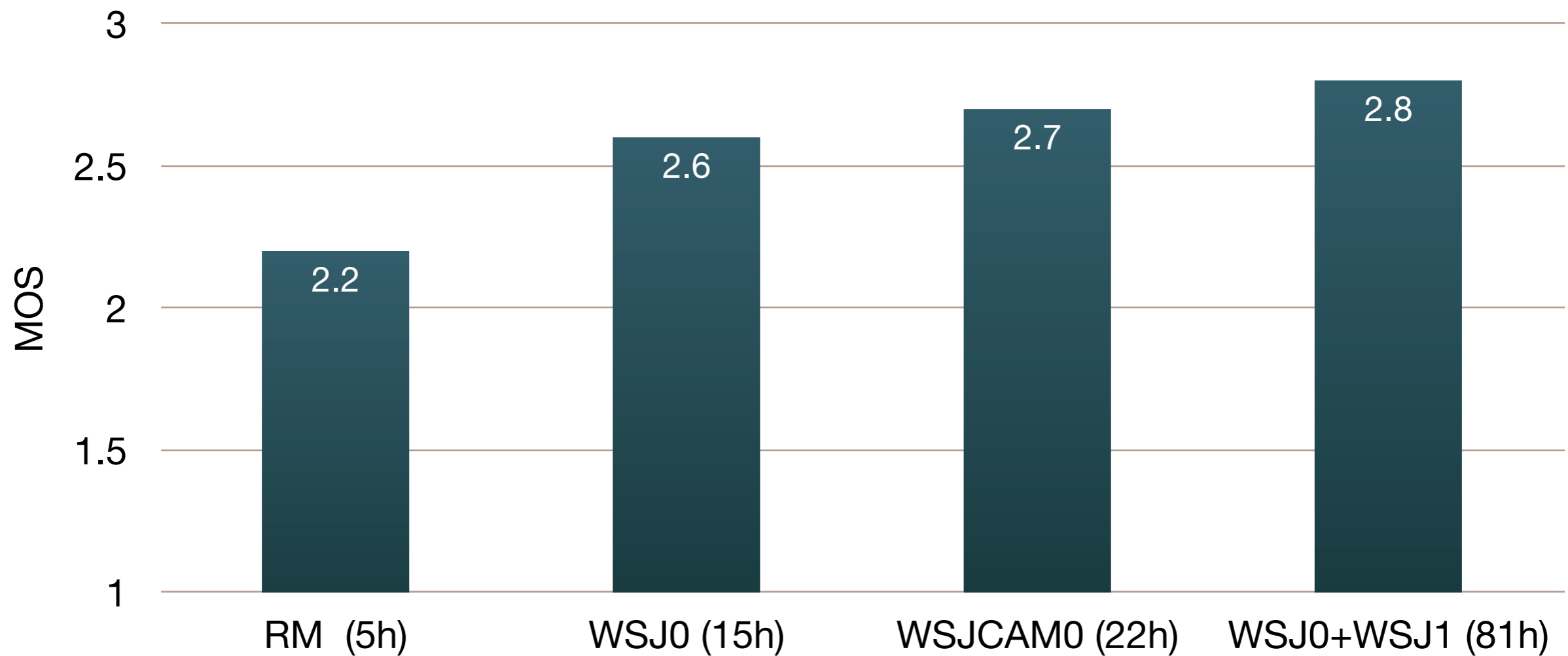


# Subjective scores of synthetic speech





Another solution to improve the quality:  
the use of larger average voice models



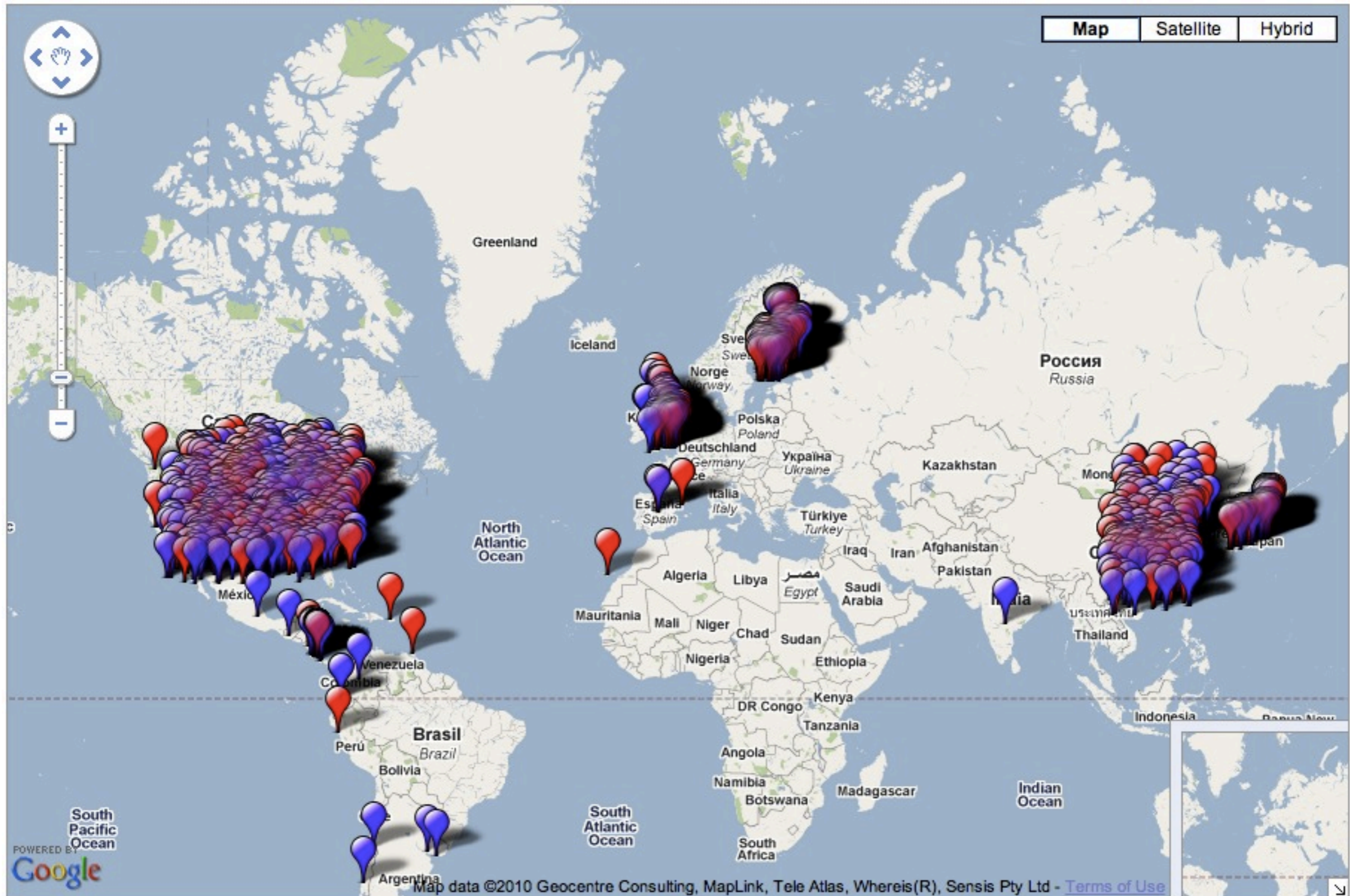
# Application 2 – Virtually unlimited number of speakers

---

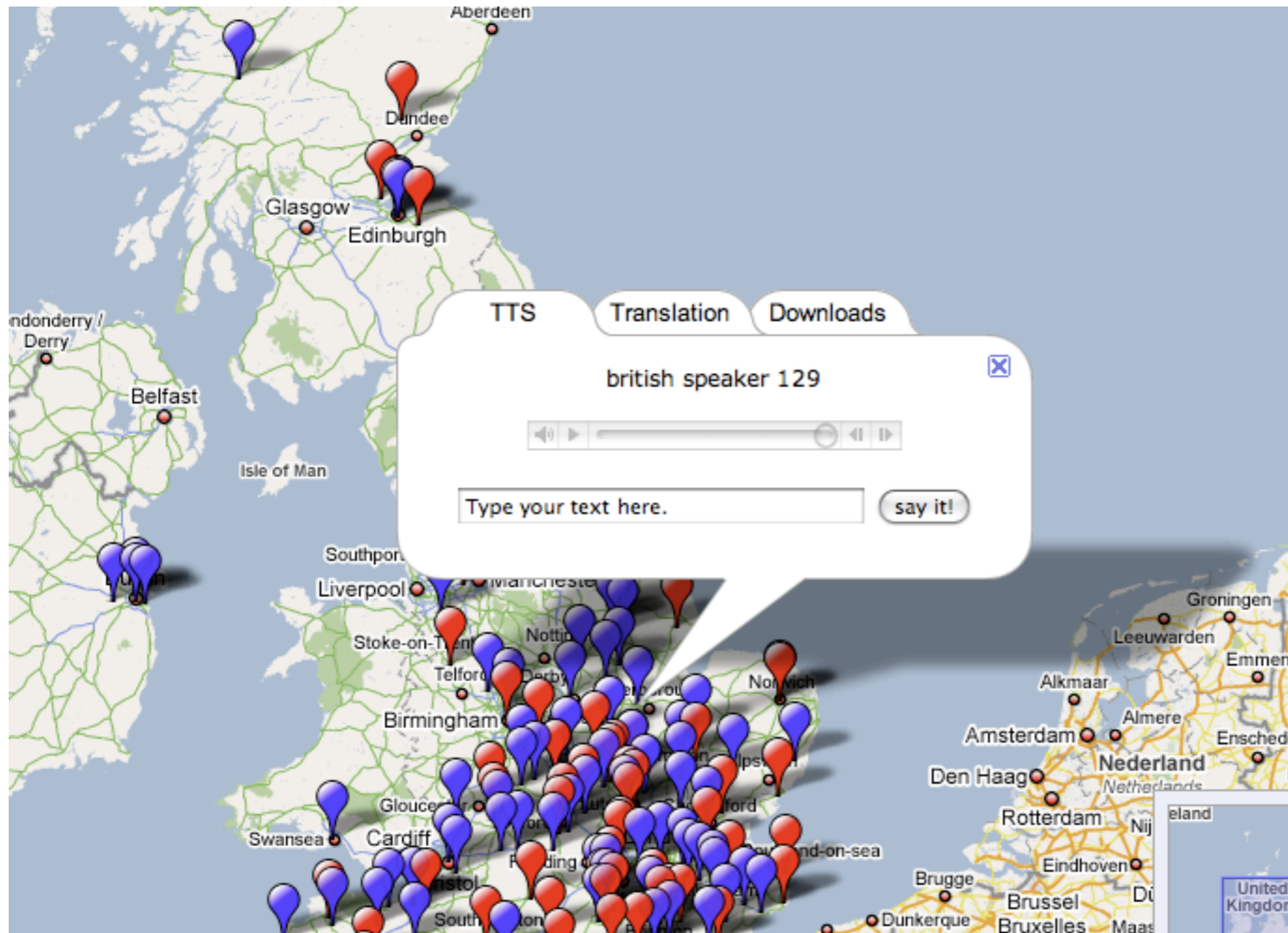


- TTS systems having unlimited number of speakers
  - eBook reader: select voices depending on the character, subject, style, ..
- ASR corpora typically contain a large number of speakers
- Good domain in which to demonstrate the impact of this scenario
- ASR corpora we have used so far
  - English: Resource management (RM), Wall-street journal WSJ0, WSJ1, & Cambridge version of WSJ0 (WSJCAM0)
  - Spanish: Globalphone
  - Mandarin: Speecon (Speech databases for consumer devices)
  - Finnish: Speecon
  - Japanese: JNAS (Japanese newspaper article sentences)

# Geographical view of many voices



# Online TTS with the largest number of voices in the world (we think...)



Total number of voices

- English: 675
- Spanish: 106
- Mandarin: 500
- Finnish: 203
- Japanese: 106

**- Total 1590**

# Technical problems behind these apps

---

- Speaker adaptive training
- Selection of average voice models
- How to select good data (low quality)
- **Vulnerability of speaker verification systems to voice clone**

# “Vocal Terror” [BBC 2007]

---

**BBC** NEWS

---

## Scientists warn of 'vocal terror'

By Liz Seward  
Science reporter, York

**Computers could mimic human speech so perfectly that vocal terrorism could be a new threat in 10-15 years' time, scientists suggest.**

In the future, it may be possible to mimic someone's voice exactly after recording just one sentence.

Such technologies would pose a danger if it were not possible to verify who was speaking, researchers believe.

Scientists were predicting the future at the British Association (BA) Festival of Science in York.

Dr David Howard from the University of York said: "The reason things are changing is because no longer are we using an acoustic model proposed in the 1950s."

**" It's not scaremongering; it's trying to say to people, 'we have to think about these things' "**

David Howard

New methods of creating computerised speech use models of a vocal tract to create a realistic sound, replacing the existing technique of copying sounds.

# Speaker verification vs Text-to-speech

---



- Voice cloning could be used for “breaking in” text-prompted speaker verification systems
- Attacking Scenarios to be assumed
  - Speech data is acquired from broadcast, podcasts, lectures, telephone
  - Using the acquired speech data, adapt HMM-based TTS systems in advance
  - Using the adapted models, synthesise speech for verifications
- How accurately can the HTS voices break in speaker verification system?

# GMM-UBM speaker verification system

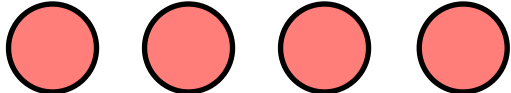
---

- GMM-UBM
  - 1024 components
- Features
  - 15 MFCC, 15  $\Delta$ -MFCC, log-energy,  $\Delta$  log-energy
  - Feature warping to improve robustness [J. Pelecanos and Sridharan]
- Adaptation
  - MAP adaptation (mean vector only)
- Performance on the NIST 2002 corpus
  - 330 speakers
  - 12.10% EER
  - Comparable performance with [C. Longworth and M. Gales 2009]



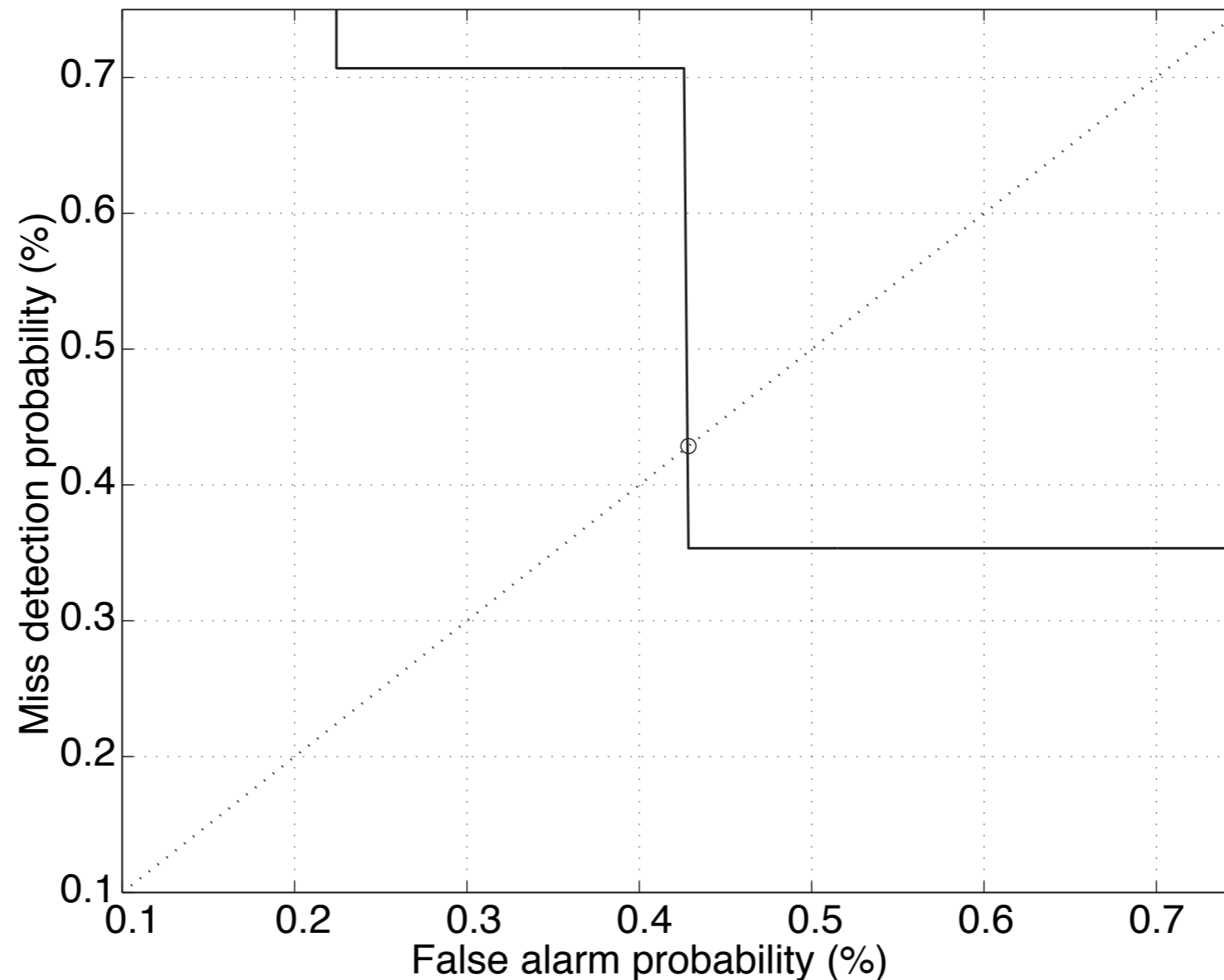
# Experiments – Data

---

- Our scenario is not building TTS systems on speaker verification databases, which are normally narrow band with noises
- Wall street journal corpora (WSJ0 and WSJ1)
  - 283 speakers (included in SI-284 set)
  - Divide the SI-284 speaker material into 3 sets, A, B, and C
  - Set A: TTS training data
    - Training of average voice models
    - Speaker adaptation (CMLLR) to individual speakers
  - Set B: SV training data
    - Training of the universal background model
    - Speaker adaptation (MAP) to individual speakers
  - Set C: Test data (30 sec/speaker)
    - Assumed to be speech reading text-prompts used for verifications
- Samples of synthetic speech created 

# Experiments – Performance of SV systems

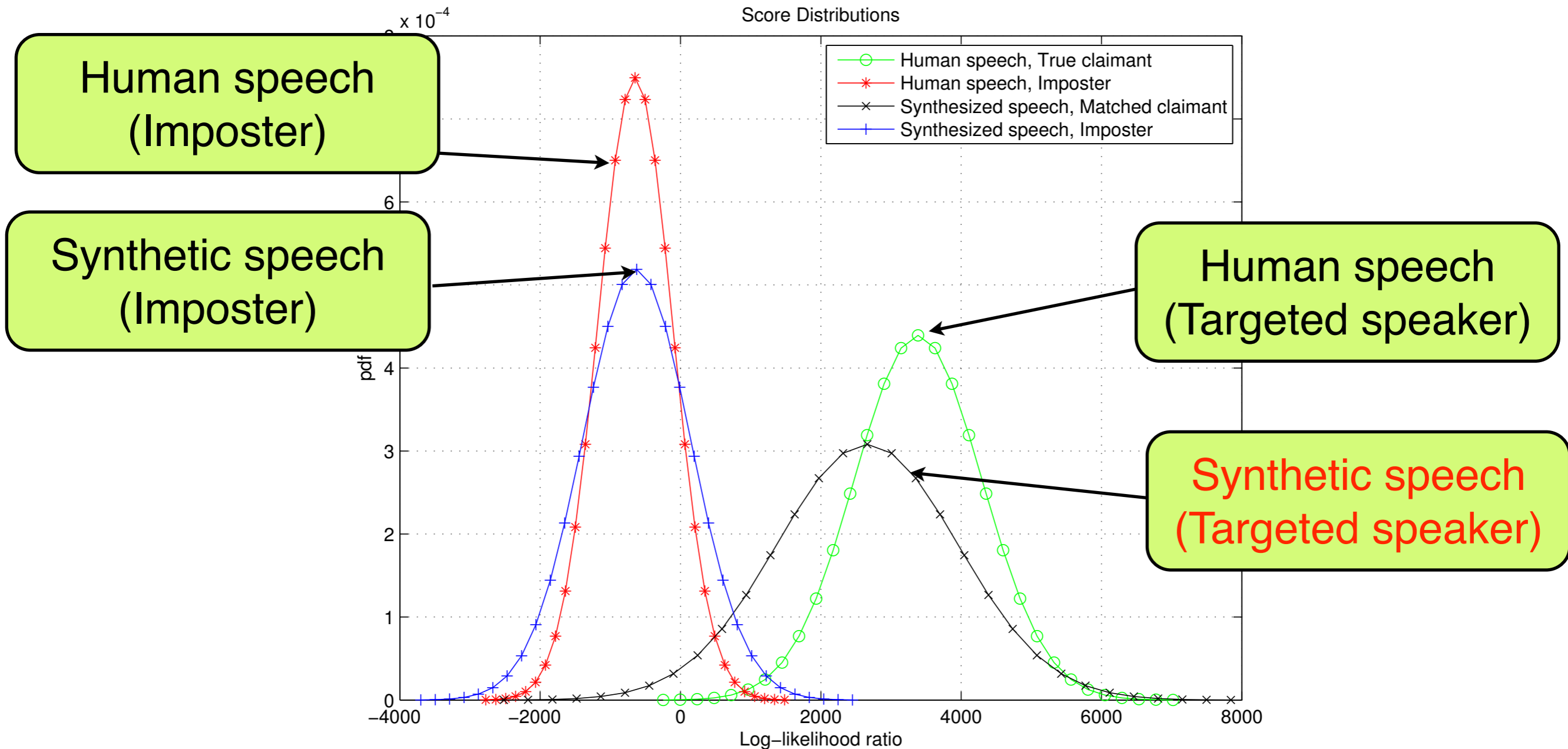
- Decision-error-tradeoff (DET) curve for human speech



- Equal-error-rate is 0.4% (speaker verification of human speech on WSJ corpus is relatively easy)

# Experiments – Human speech vs. Synthetic speech

- Score distributions of human speech and synthetic speech



- In matched claimant tests (synthesized voices claim to be their human counterparts), about **90%** of synthetic speech claims was accepted!

# How about SVM?

---

- Worse!! It accepts claims from synthetic speech more.

	GMM-UBM	SVM using GMM super-vectors
EER (human speech)	0.35%	0.35%
min DCF (human speech)	4.00E-03	2.40E-03
accepted claims from synthetic speech	<b>91.5%</b>	<b>95.8%</b>

# Can we block the attack?

---

- YES!
- The current major vocoders: minimum phase vocoders
  - Human perception is less sensitive to phase differences
  - Assumed as not worth modelling
- Hint: Phase, Phase spectrum etc
- Human is less sensitive perceptually, but differences between phase information of real and current synthetic speech are “visible”
- GMMs trained on an analysis method called “relative phase shift (RPS)”
- Results
  - Correctly classify human speech in 95%
  - Correctly classify synthetic speech in 88%

# Application 3 – Voice banking

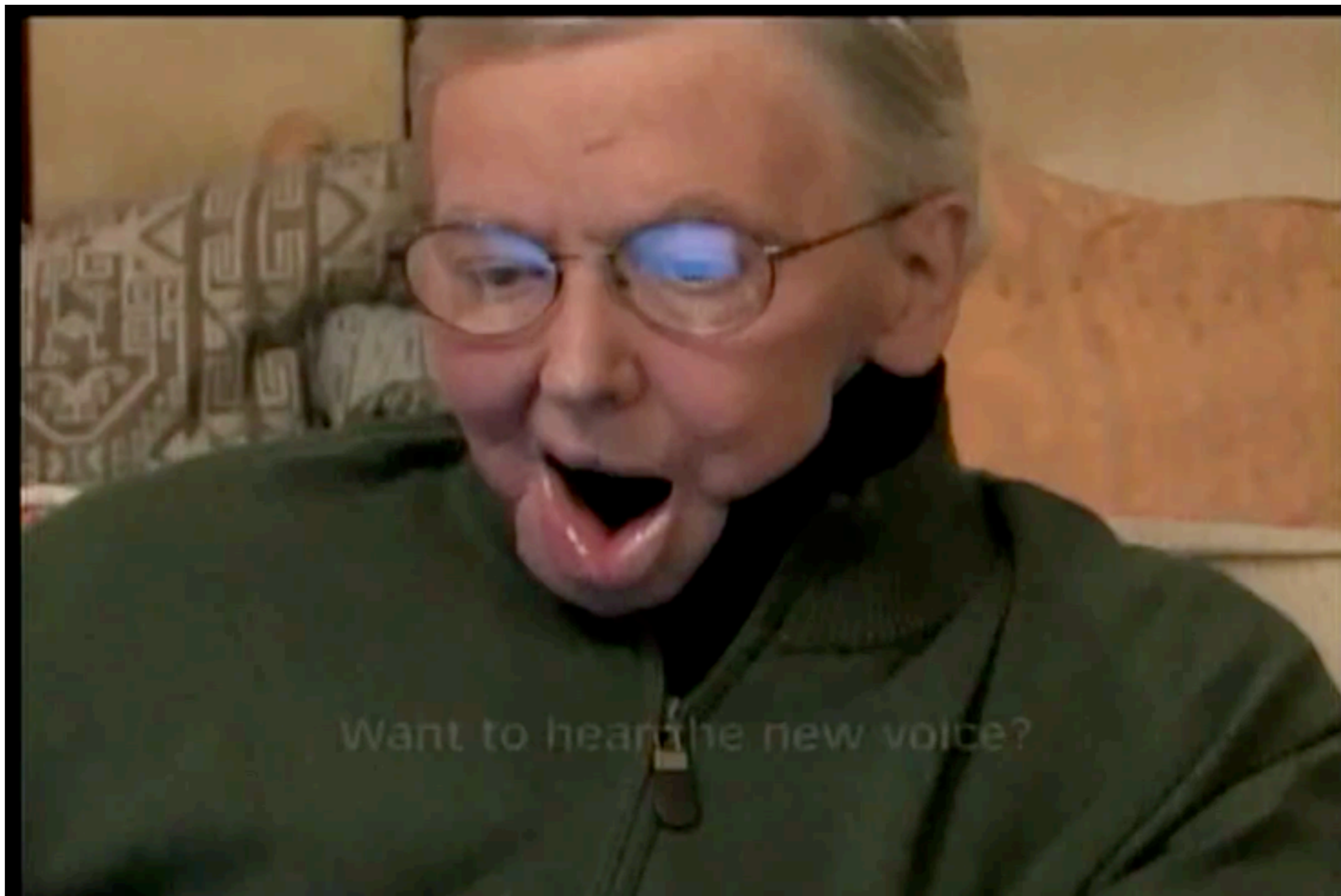
---

- If people record a small amount of their speech data, they could have personalised voice communication aids in case of any vocal problems such as speech disorder in the future
- CSTR members who “voice banked”



# Roger Elvert voice by Cereproc Ltd. [March 2010]

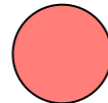
- Unit selection system using audio (commentary) included in many films



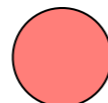
# Clinical trial by the Sheffield group

HTS voice can be adapted only from 100 sentences

Chapman's real speech



Chapman's synthetic speech



Times 2009

From The Times

September 5, 2009

## Patients without vocal chords could have voice restored

David Rose, Health Correspondent

RECOMMEND?

Patients who have had vocal cords removed could soon get their own voice restored with a synthesised version.

Students and academics at the University of Sheffield have used recordings and sampling technology to reconstruct the voice of Bernadette Chapman, who had a laryngectomy operation to remove her vocal cords after developing cancer.

Voice synthesising technology is already used by some patients who have lost the ability to speak, such as Sir Stephen Hawking, the eminent physicist.

But users have complained that the results sound like a "dalek" and the latest research attempted to recreate speech patterns that come close to sounding like a patient's own voice.

### RELATED LINKS

Trainee doctors lack educational foundations  
NHS expects 40,000 fewer swine-flu deaths

Researchers took recordings of Mrs Chapman's voice prior to the operation and, in collaboration with Edinburgh University's Centre for Speech Technology Research, used a speech synthesis technique to adapt an "average voice model" to sound like the person concerned.

Mrs Chapman's new voice was "built" using about seven minutes of recorded speech, which amounted to 100 sentences. From those samples, the researchers say it is possible to synthesise any sentence by supplying the word sequence.

### EXPLORE HEALTH

EXPERT ADVICE  
HEALTH FEATURES  
MENTAL HEALTH  
ALTERNATIVE MEDICINE  
CHILD HEALTH  
HEALTH CLUB



### The NHS: enormous, expensive and still growing

Tim Glanfield looks behind the story of the NHS staff cuts



# Contents

---

- Background
  - a brief introduction to HMM-based speech synthesis
  - basic principles of adaptation
- Applications
  1. Voice cloning
  - 2. Voice reconstruction**
  3. Personalised speech-to-speech translation
  4. Articulatory-controllable speech synthesis



# Voice reconstruction – Personalised voice communication devices for people with vocal disabilities

---

*Credits: Sheffield University ; The Euan McDonald Centre for Motor Neurone Disease Research ; The Anne Rowling Regenerative Neurology Clinic*



# Neurological conditions or diseases which can result in a vocal pathology

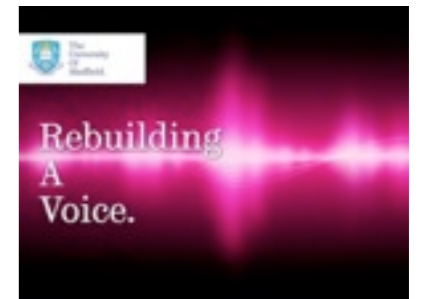
---

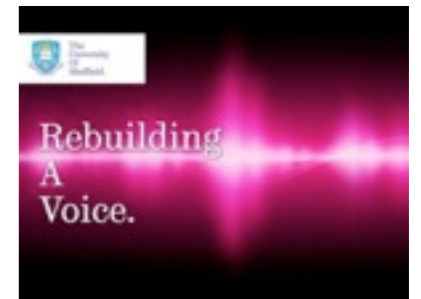


- Amyotrophic Lateral Sclerosis (ALS) / Motor Neurone Disease (MND)
  - Autism
  - Stroke Aphasia
  - Cerebral Palsy
  - Parkinson's Disease
  - Multiple Sclerosis
- 
- MND incidence (new cases per year) is about 2 per 100,000 people
  - Across all neuro-degenerative diseases: about 20 million new cases per year

# An interview with a patient with MND

---





# What do these people use?

---

- Augmentative and Alternative Communication (AAC)
- Voice Output Communication Aid (VOCA)
  
- Some diseases cause problems with the hands as well as the voice
  - so alternative forms of input are required
- Some diseases cause problems only with the voice
  - so 'type to talk' interfaces are possible
  
- The current VOCA on the market provide a small and inappropriate range of voices
  
- We want to provide personalised speech synthesis voices (which sound like the individual users)

# Can we use speaker adaptation?

## – Ideal and the real

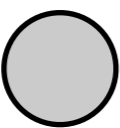
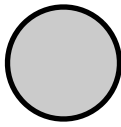


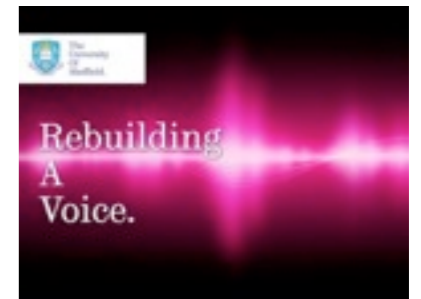
- For such people with vocal problems, speech synthesis is not just an optional extra to reading out text, but a critical function for
  - social communication, and
  - personal identity
- We can use speaker adaptation for creating the personalised voices only
  - If their speech was voice banked in prior,
  - if they are diagnosed at very very early stage, or
  - if they recorded their speech in acceptable clean conditions by themselves
- However, patients tend to arrive at the clinic **\*only\*** after vocal problems are moderate to severe

# What happens if you use speaker adaptation on disordered speech?

---



- Original data: 
  - 3 mins
  - previously recorded interview made in office environment
  - subject already had MND at that point - speech is already disordered
- Speaker adaptation
  - **voice clone of disordered speech** 



# Voice reconstruction

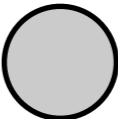
---

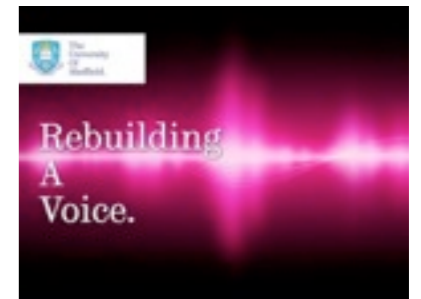
- We should extend our adaptation technique so that it can be applied even if speech has already be **disordered** at the time of recording
- Paradox
  - Recover speaker characteristics as much as possible
  - But we do not want to reproduce the symptoms due to vocal problems
- How ?
  - Separation speaker characteristics and vocal problems? (a hard problem)
  - Perform '**surgery**' on the adapted statistical models
    - fixing statistical models so that they can generate natural sounding speech while keeping speaker identity
- From voice cloning to voice reconstruction



# Voice reconstruction using average voice models [S. Creer et al. 2010]



- Some acoustic models are \*relatively\* less speaker-dependent
- Therefore, substitute
  - Adapted duration model
  - Adapted GV models (spectrum, logF0)
  - Adapted aperiodicity model
- into those of average voice models
- Example
  - Substitution using a WSJCAM0 average voice model 
- We can still hear problems of 1) the accent and 2) coarticulation
  - Reproduction of Edinburgh accent is not perfect
  - Bad coarticulation of diphthongs and long vowels



# Voice reconstruction using voice donor

---

- For coarticulation, we need to “fix” variances of static spectral features, delta features
- However, they are more speaker specific features
- Substitution by average voice model may result in lower similarity
  
- Substitution (excluding lower order part of static features) into models trained on speakers that match with a target patient in terms of
  - age, accent, social class, similar vocal tract shape etc
  - “Voice donor”
  
- Example of the voice donor
  - Close relative



# Advantages of voice donor

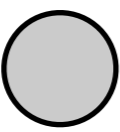
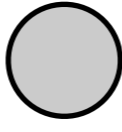
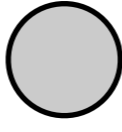
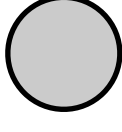
---

- Voice donor can help people with vocal problems simply by speaking
- The donor's speech is also automatically voice banked
  - Some diseases are genetic
  - Good and another motivation for the recording of close relative's speech
- They too could have a personalised voice communication aid in case of any vocal problems in the future
- Better awareness of the importance of "voice banking" amongst the public

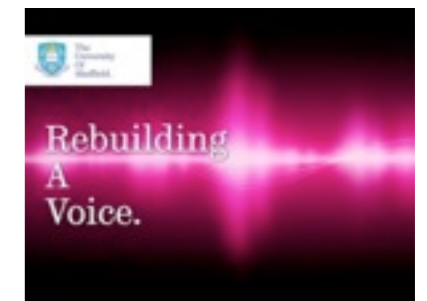


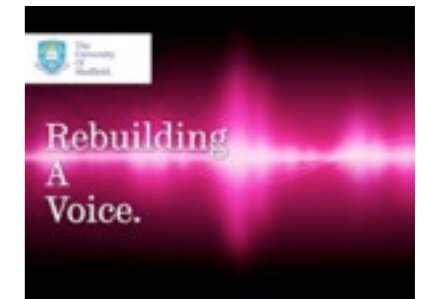
# Conceptual flow: Summary

---

- Original data: 
  - 3 mins
  - previously recorded interview made in office environment
  - subject already had MND at that point - speech is already disordered
- Speaker adaptation
  - voice clone of disordered speech 
- Voice reconstruction using voice donor
  - fixing statistical models so that they can generate natural sounding speech while keeping speaker identity 
- Voice banking of voice donor 

# A personalised voice output communication device with eye tracking





# Voice banking and voice reconstruction

---

- Voice banking and voice reconstruction will be carried out
  - Euan McDonald Centre for MND research
  - Anne Rowling Regenerative Neurology Clinic (under construction)
  - Barnsley Hospital (MyVOCA project of the Sheffield University)
- Voice reconstruction: 50 patients per year
- Voice banking and donor: over 50 people per year
- Future work
  - Speaker adaptation using hierarchical average voice models and/or automatic selection of donors
  - Automatic vocal function assessment, followed by automatic model substitution or interpolation

# Contents

---

- Background
  - a brief introduction to HMM-based speech synthesis
  - basic principles of adaptation
- Applications
  1. Voice cloning
  2. Voice reconstruction
  - 3. Personalised speech-to-speech translation**
  4. Articulatory-controllable speech synthesis

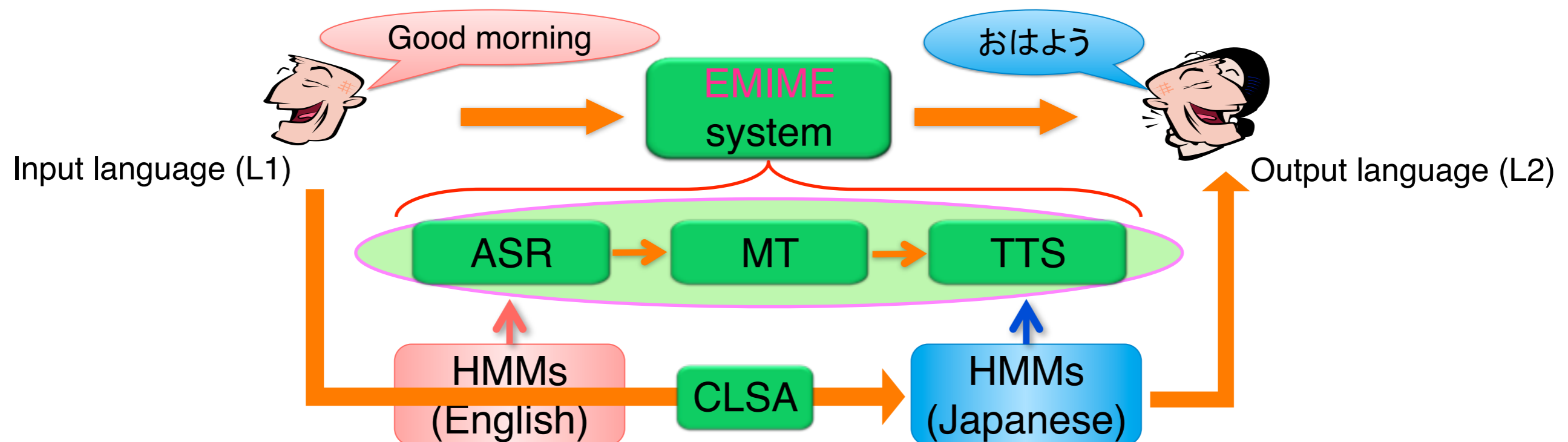
# Personalised speech-to-speech translation using cross-lingual speaker adaptation

---



# Speech-to-speech translation

## Speech-to-speech translation (S2ST): flow

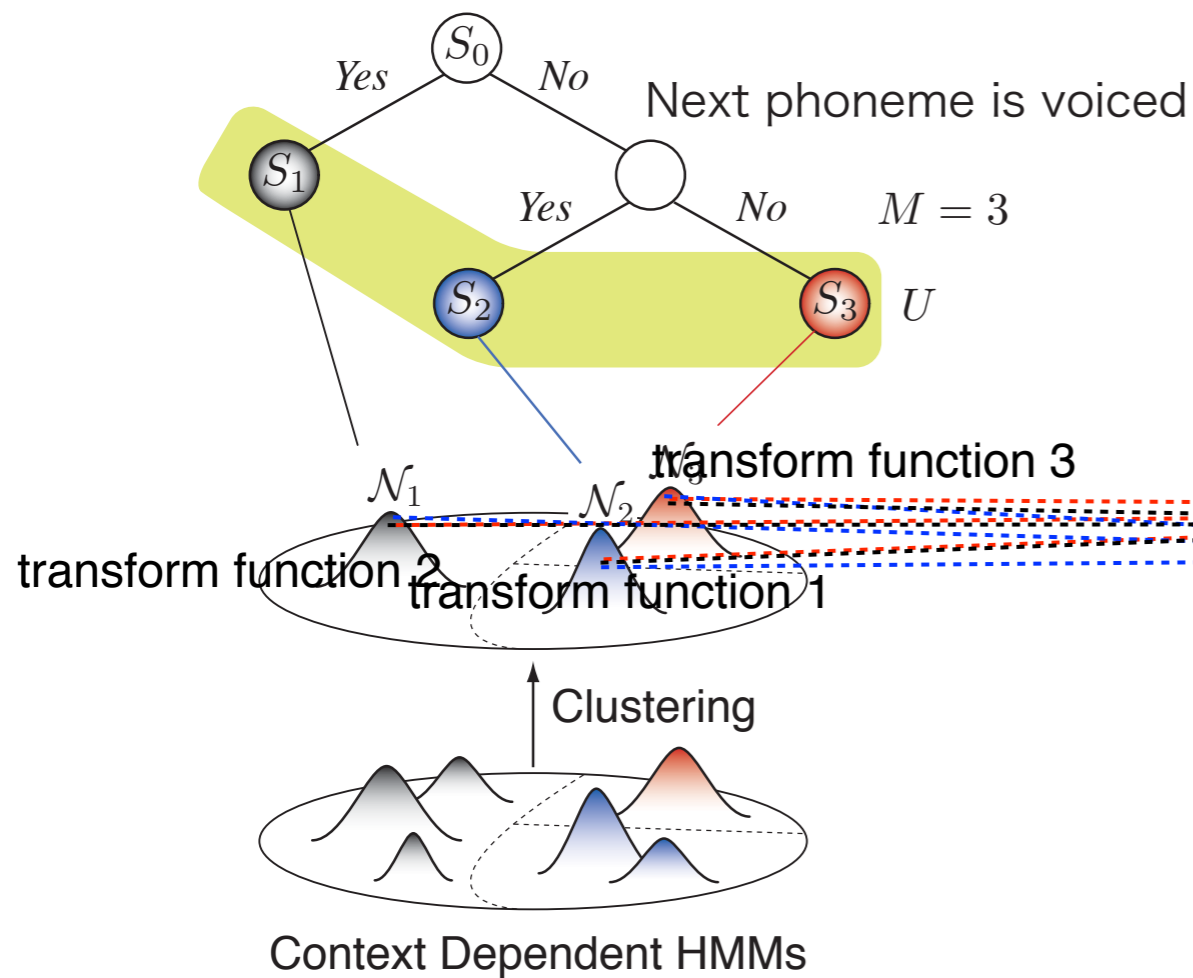


CLSA: cross-lingual speaker adaptation

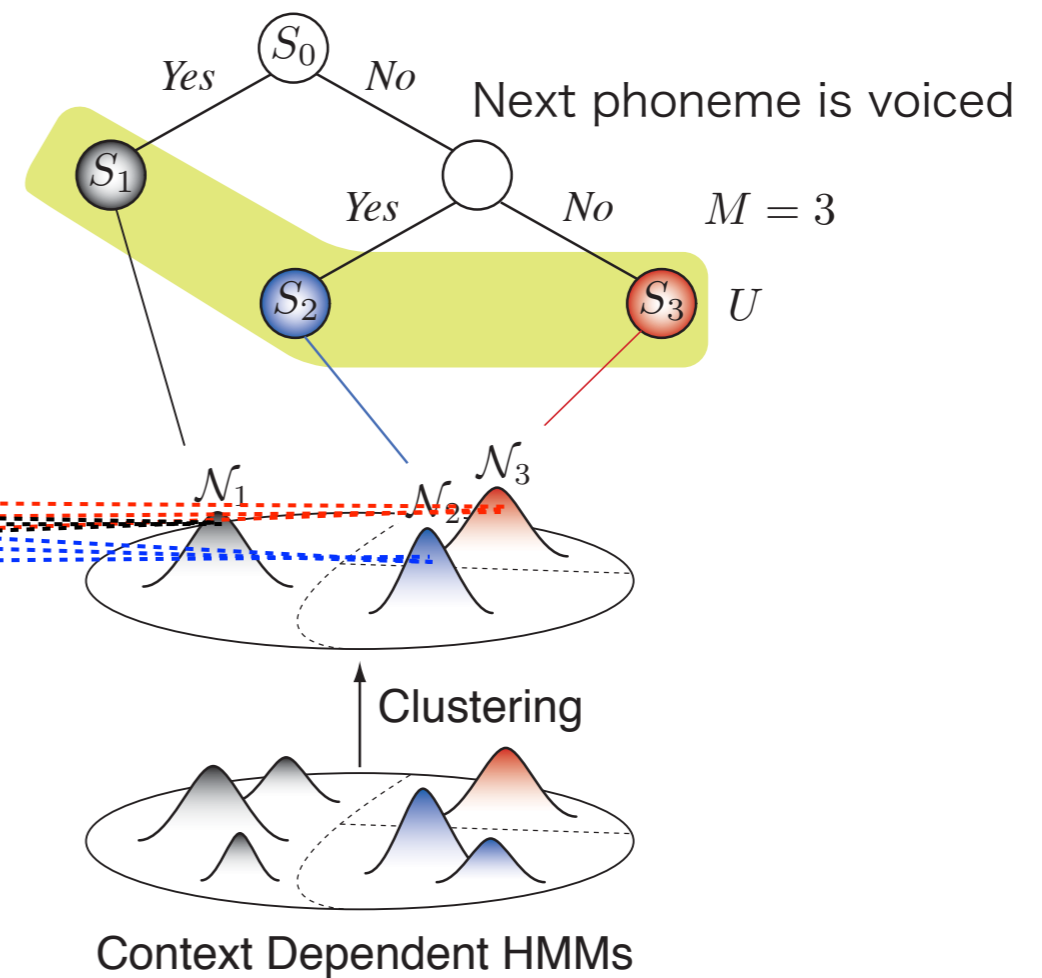
# Cross-lingual adaptation based on state-mapping



The current phoneme is voiced



The current phoneme is voiced



# Details of symmetrized KLD-based state-similarity measure



- Definition of the problem of KLD-based state mapping

$$f(i) = \arg \min_j D_{\text{KL}} \left( G_i^{(O)}, G_j^{(I)} \right)$$

$G_i^{(O)}$ : Average voice model for output language  
 $G_j^{(I)}$ : Average voice model for input language

- Definition of the symmetrized KLD between states i and j

$$D_{\text{KL}} \left( G_i^{(O)}, G_j^{(I)} \right) \approx D_{\text{KL}} \left( G_j^{(I)} \parallel G_i^{(O)} \right) + D_{\text{KL}} \left( G_i^{(O)} \parallel G_j^{(I)} \right)$$
$$D_{\text{KL}} \left( G_j^{(I)} \parallel G_i^{(O)} \right) = \frac{1}{2} \ln \left( \frac{|\Sigma_i^{(O)}|}{|\Sigma_j^{(I)}|} \right) - \frac{D}{2} + \frac{1}{2} \left( \Sigma_i^{(O)-1} \Sigma_j^{(I)} \right) + \frac{1}{2} \left( \mu_i^{(O)} - \mu_j^{(I)} \right)^\top \Sigma_i^{(O)-1} \left( \mu_i^{(O)} - \mu_j^{(I)} \right)$$

$\mu_j^{(I)}$   $\mu_i^{(O)}$ : Mean vectors of average voice models for input and output languages

$\Sigma_j^{(I)}$   $\Sigma_i^{(O)}$ : Covariance matrices of average voice models for input and output languages

# English-to-Japanese adaptation: Experimental conditions



## English ASR

Database	WSJ0 (15 hours, 84 speakers)	Sampling rate	16kHz
Analysis window	25ms Hamming window	Frame shift	10ms
Acoustic feature	13-th PLP + $\Delta$ + $\Delta\Delta$		

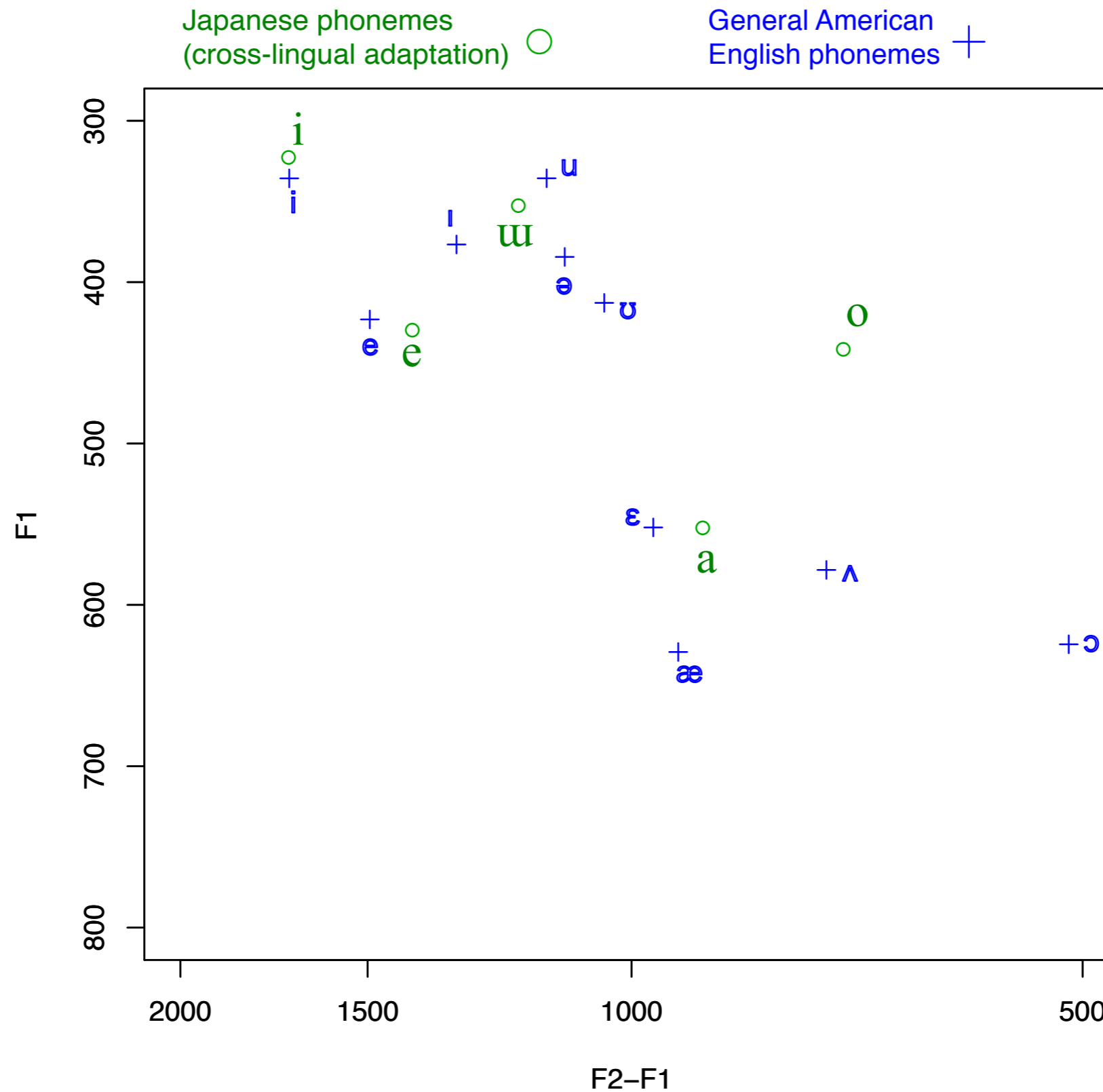
## English TTS

Database	WSJ0 (15 hours, 84 speakers)	Sampling rate	16kHz
Analysis window	25ms Hamming window	Frame shift	5ms
Acoustic feature	39-th STRAIGHT mel-Cepstrum + $\Delta$ + $\Delta\Delta$		
	Log F0 + $\Delta$ + $\Delta\Delta$		
	5 band-filtered aperiodicity measures + $\Delta$ + $\Delta\Delta$		

## Japanese TTS

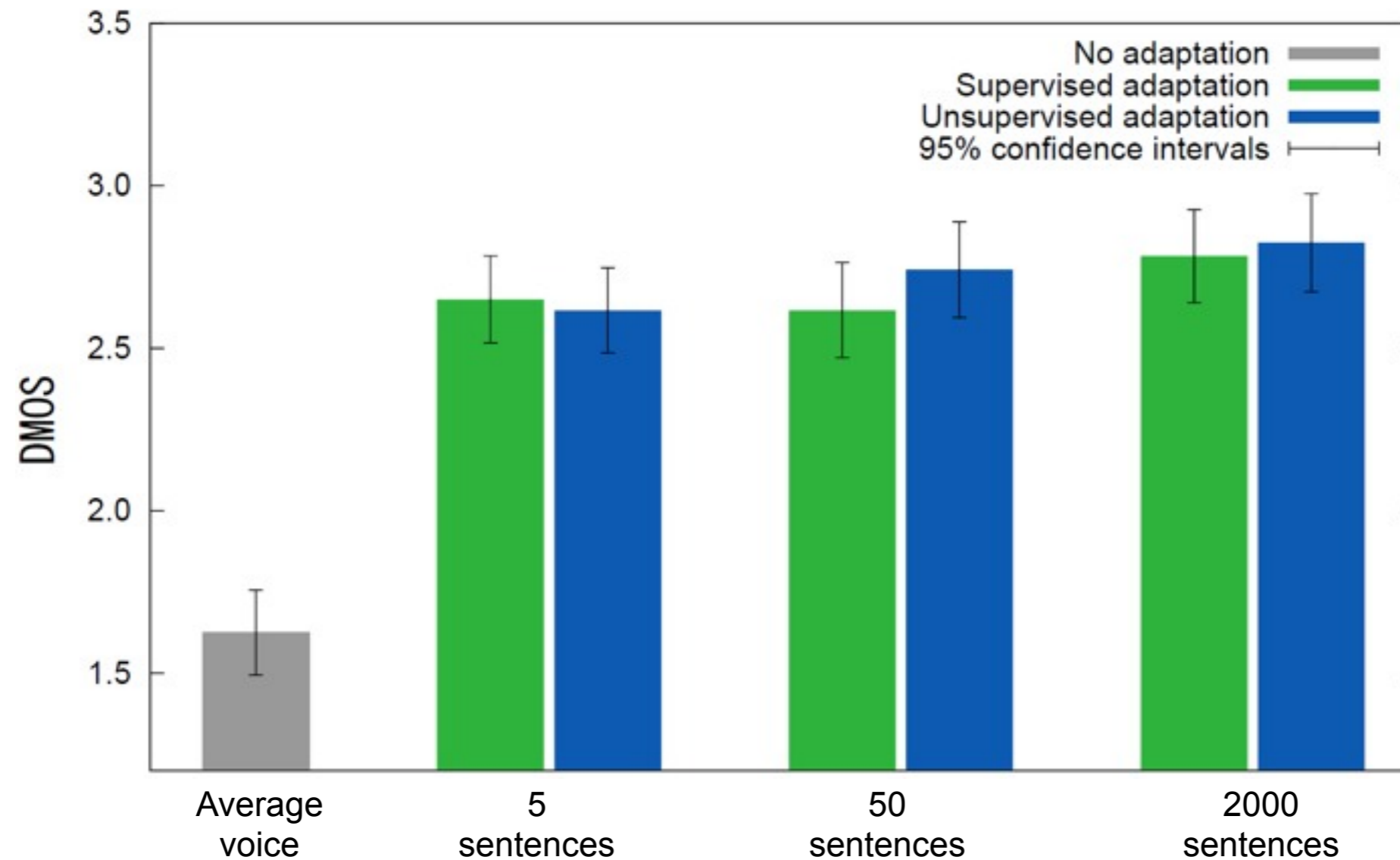
Database	JNAS (19 hours, 86 speakers)	Sampling rate	16kHz
Analysis window	25ms Hamming window	Frame shift	5ms
Acoustic feature	39-th STRAIGHT mel-Cepstrum + $\Delta$ + $\Delta\Delta$		
	Log F0 + $\Delta$ + $\Delta\Delta$		
	5 band-filtered aperiodicity measures + $\Delta$ + $\Delta\Delta$		

# Vowel comparisons



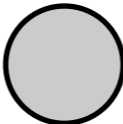
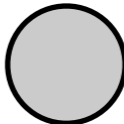
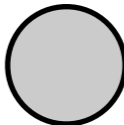
# Demo 1: WSJ0 speaker 001

Target speaker ←



# Demo 2: BBC presenter James Cook



	Sample 1	Sample 2
Original speech		
English TTS adapted to James		
Japanese TTS adapted to James		

- Ideal conditions
  - speech data recorded in controlled recording studio
  - oracle ASR and MT
- But quality of Japanese voice limited by lack of speech databases recorded at higher sampling rates

# Contents

---

- Background
  - a brief introduction to HMM-based speech synthesis
  - basic principles of adaptation
- Applications
  1. Voice cloning
  2. Voice reconstruction
  3. Personalised speech-to-speech translation
  - 4. Articulatory-controllable speech synthesis**



# Articulatory-controllable statistical speech synthesis

---

*Credits: Zhenhua Ling (University of Science and Technology of China) & Korin Richmond (CSTR) ; LISTA project*

**listeningTalker**

# Motivation

---

- Speaker adaptation requires “too much” adaptation data
  - because it’s a shallow method
  - adaptation takes place at a surface level
    - features or model parameters
    - no ‘deep model’ underlying the adaptation process
    - just a non-linear transform of the whole feature (or model) space
- How about speech modification **without requiring new speech data**?
  - perhaps based on other information, such as articulation, listener characteristics, environment, ...
  - our starting point: knowledge of speech production, in the form of articulatory measurement data

# Articulatory data used

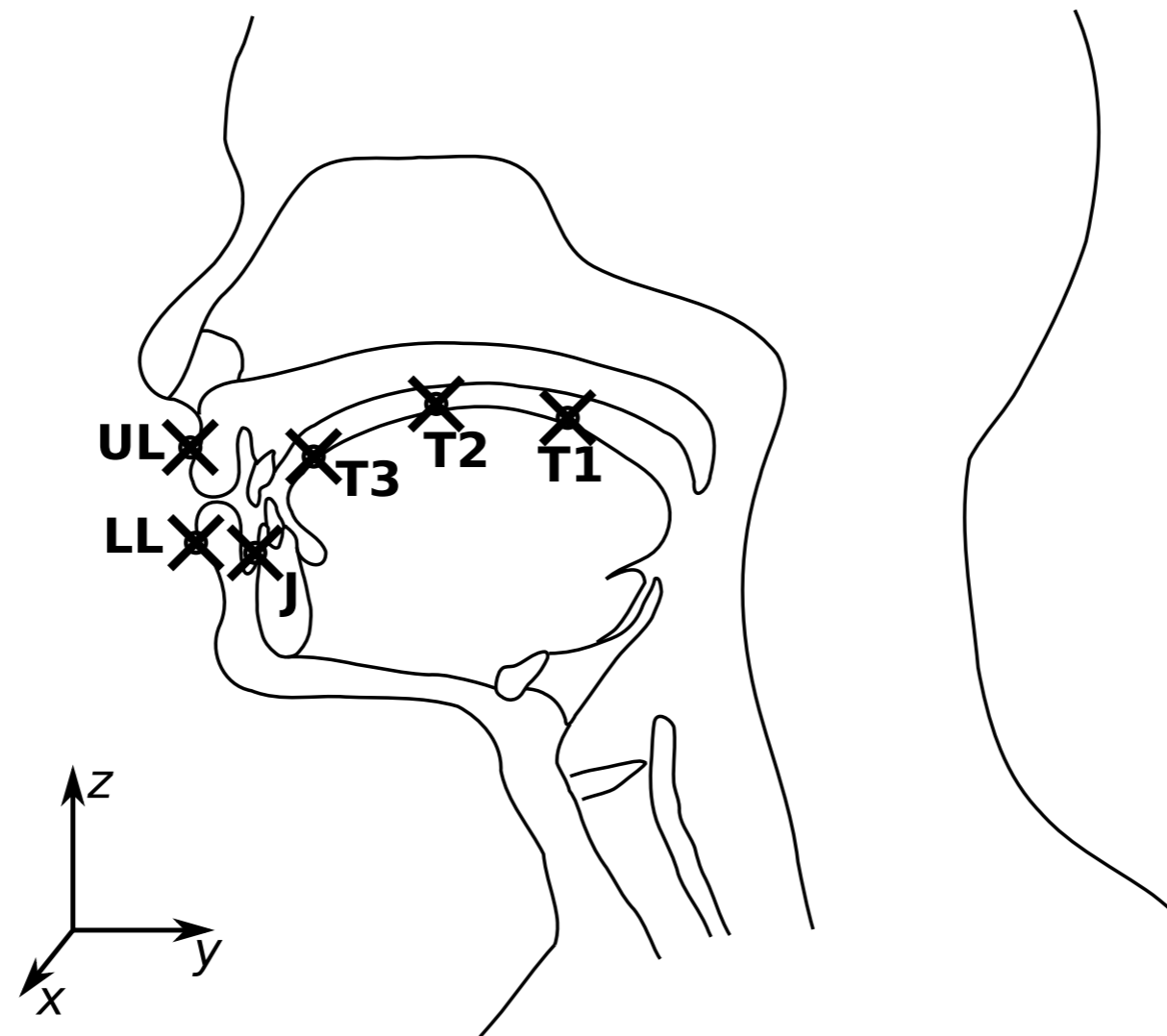
---

- Male native British English (RP accent) speaker
- 1,263 phonetically-balanced utterances
- 7 articulatory points: UL, LL, LI, TT, TB, TD
- Carstens 3D Electromagnetic Articulograph
- Audio suitable for speech synthesis



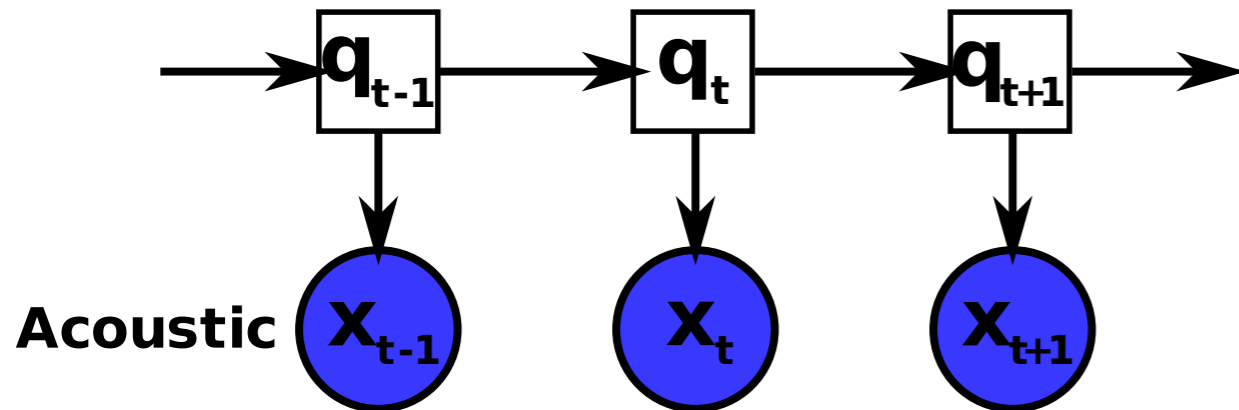
# Measurement points (coil locations)

---



# Introducing articulation into the HMM

Acoustic only

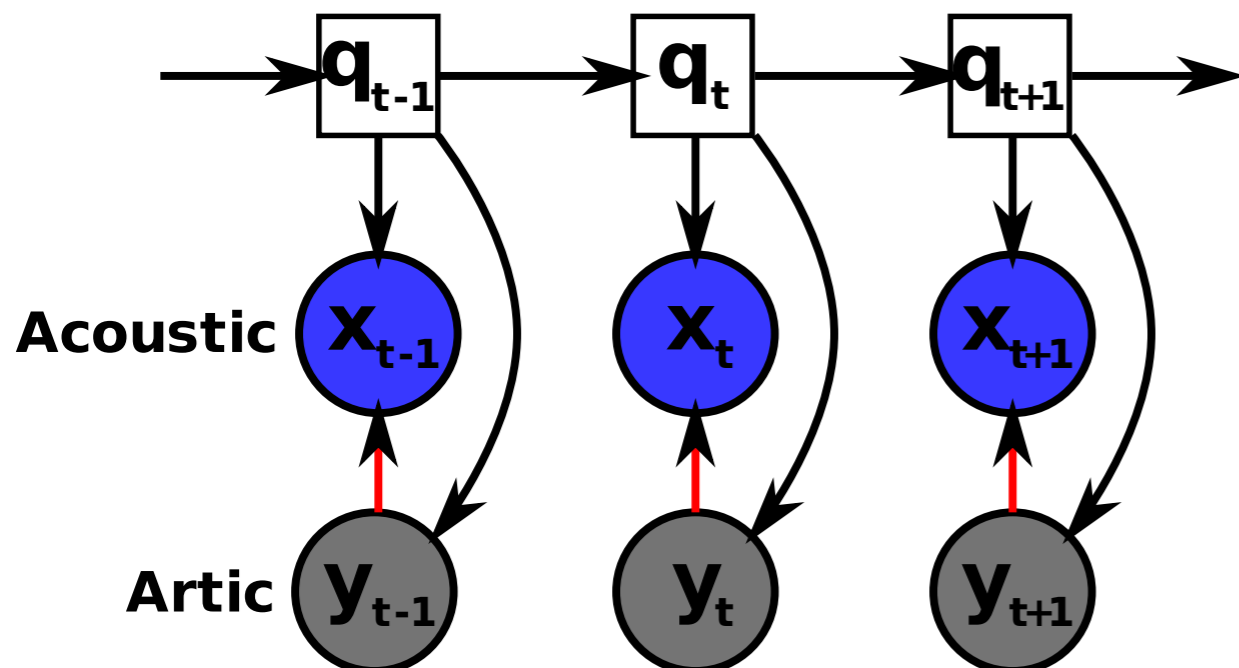


Model **joint distributions** of acoustic and articulatory parameters

Acoustic distributions (for spectral parameters) **dependent on articulation**

**Dependency = linear transform**

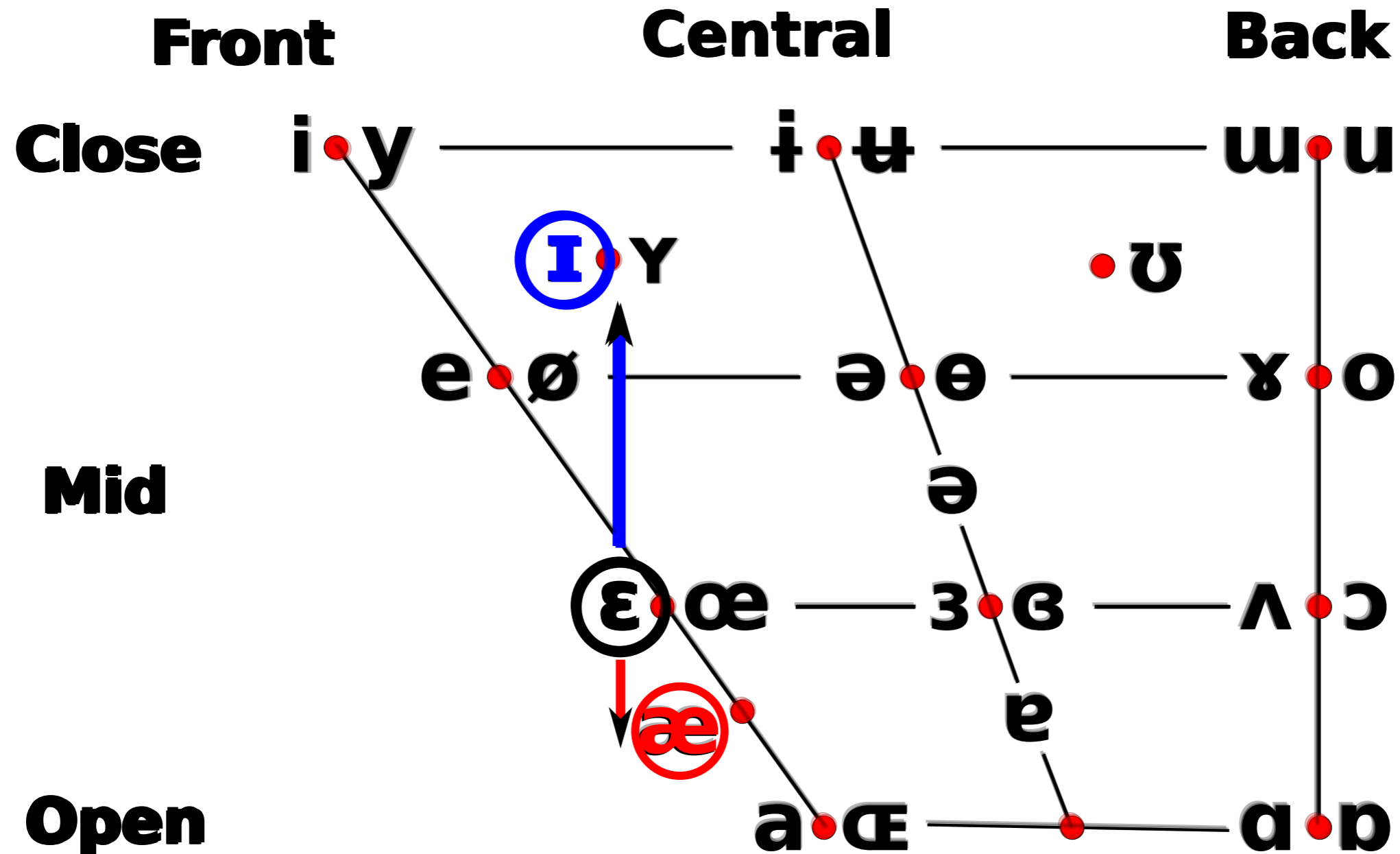
EMA + acoustic



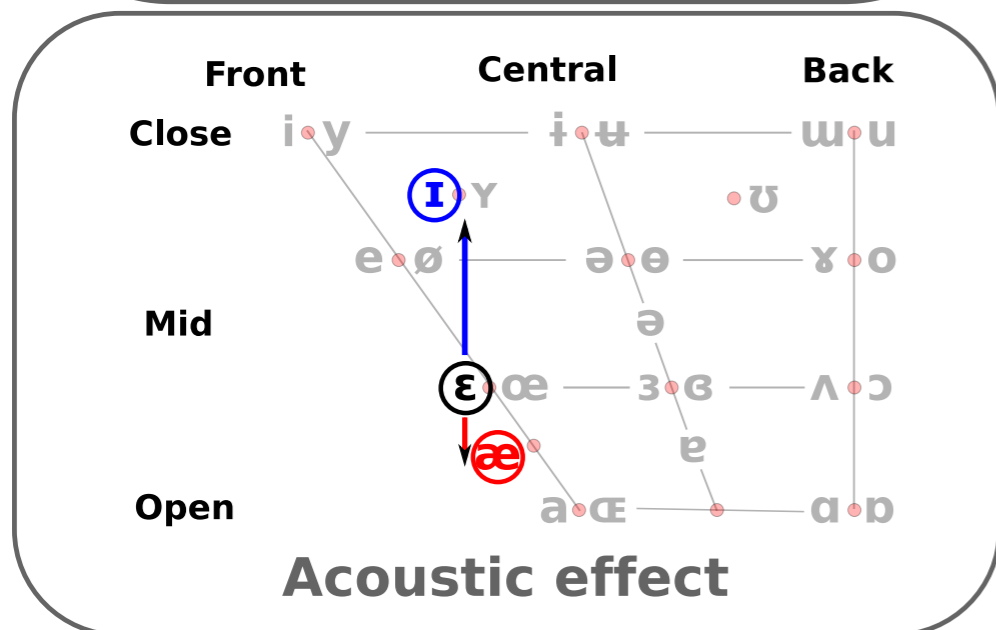
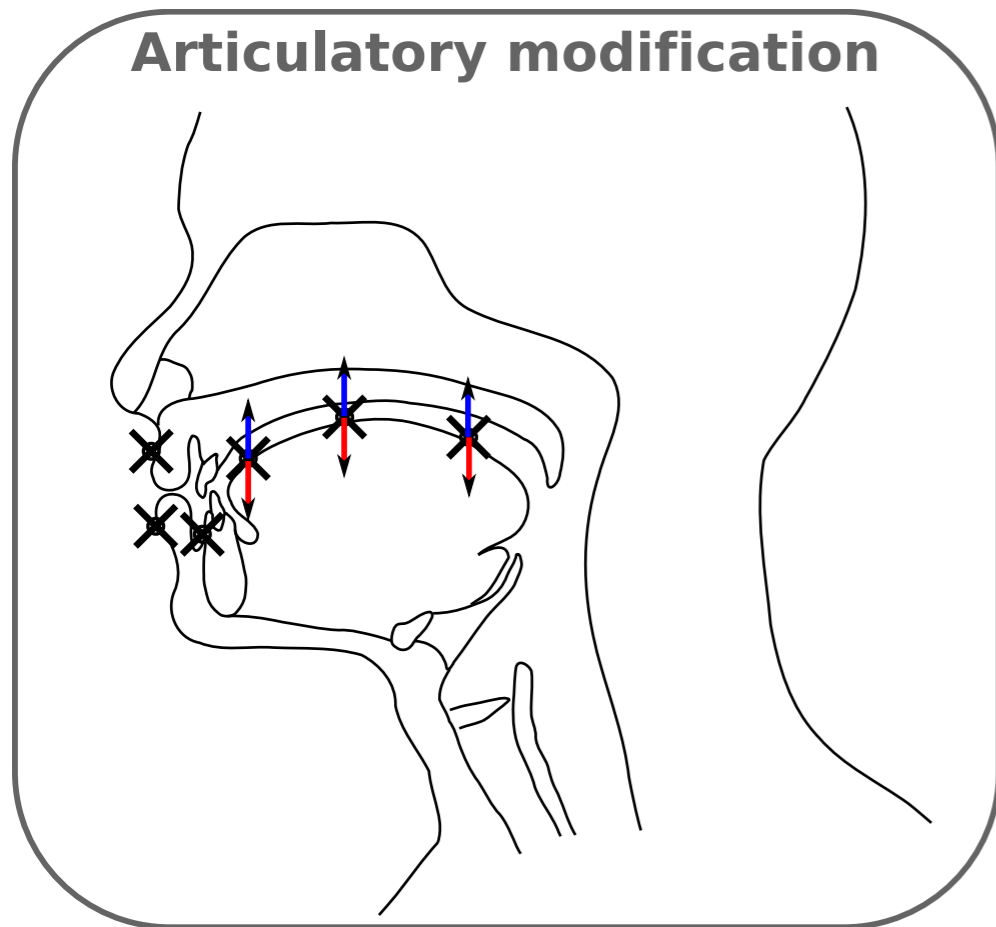
No loss of quality

*Note: can use articulatory function to modify  $y_t$*

# Articulatory modification of a vowel



# Change tongue height = change the vowel



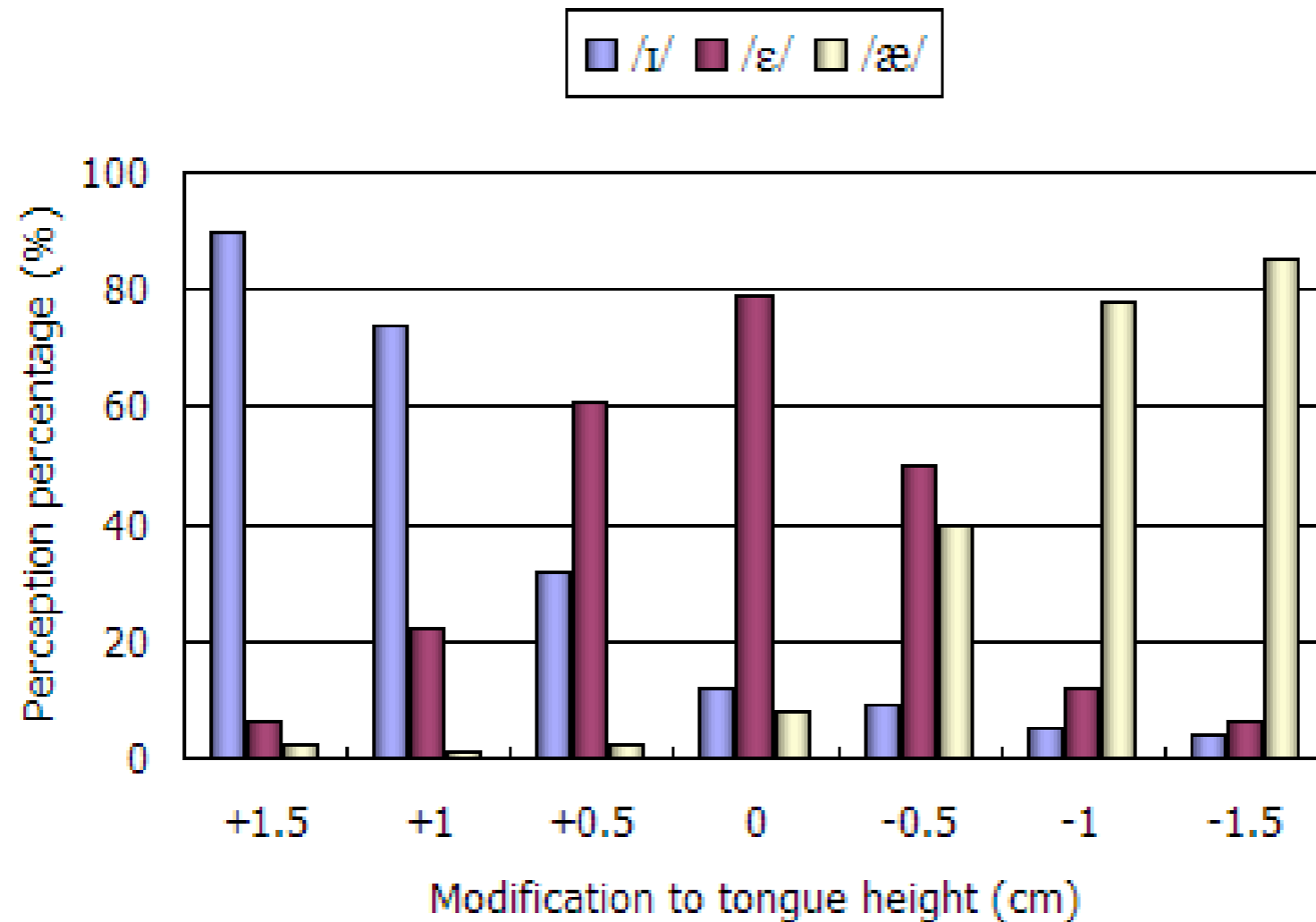
Tongue height (cm)

+1.5	○	○	○
+1.0	○	○	○
+0.5	○	○	○
default	set	peck	led
-0.5	○	○	○
-1.0	○	○	○
-1.5	○	○	○

# Perceptual test results

20 listeners, lab condition

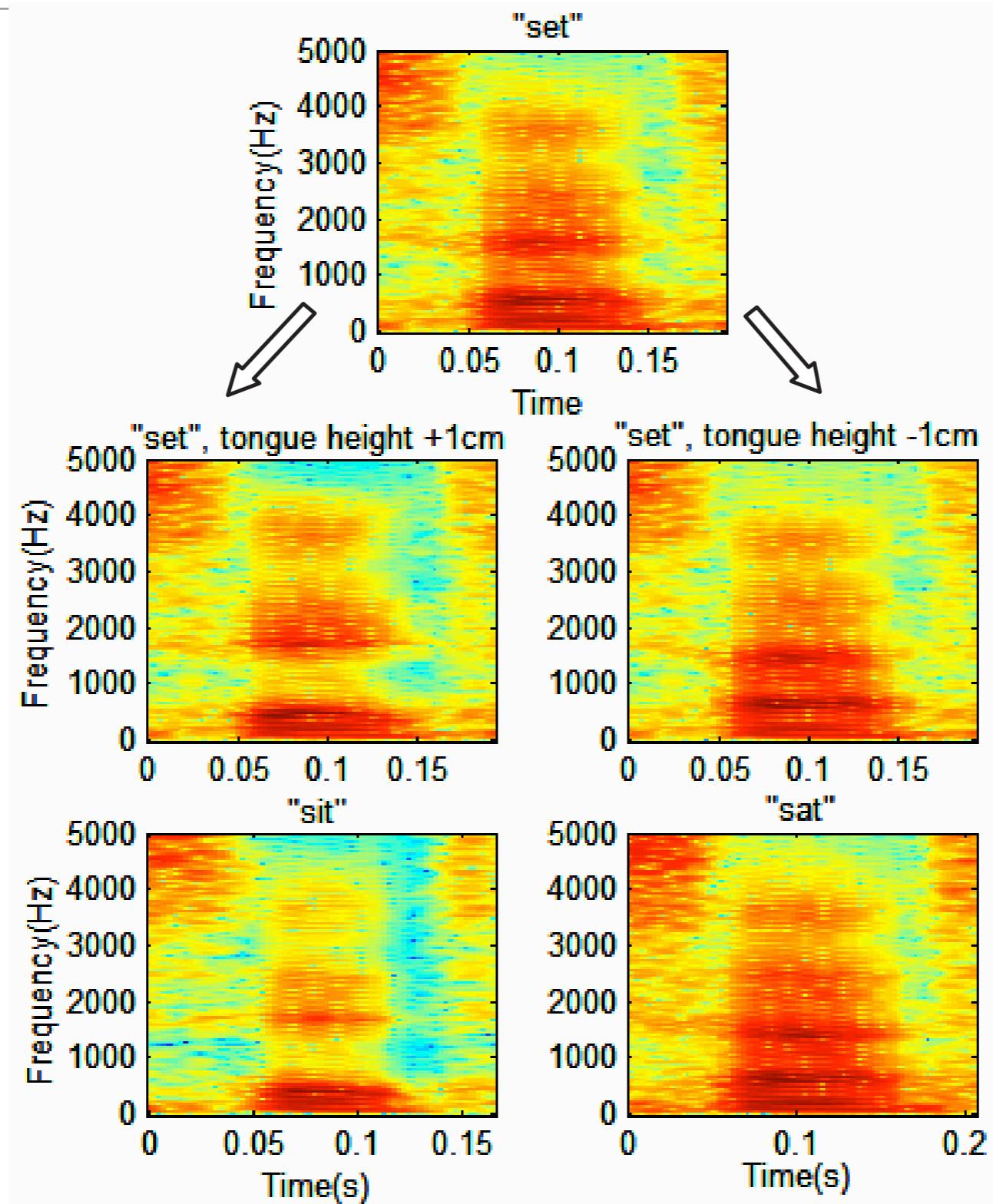
results pooled across speakers and words



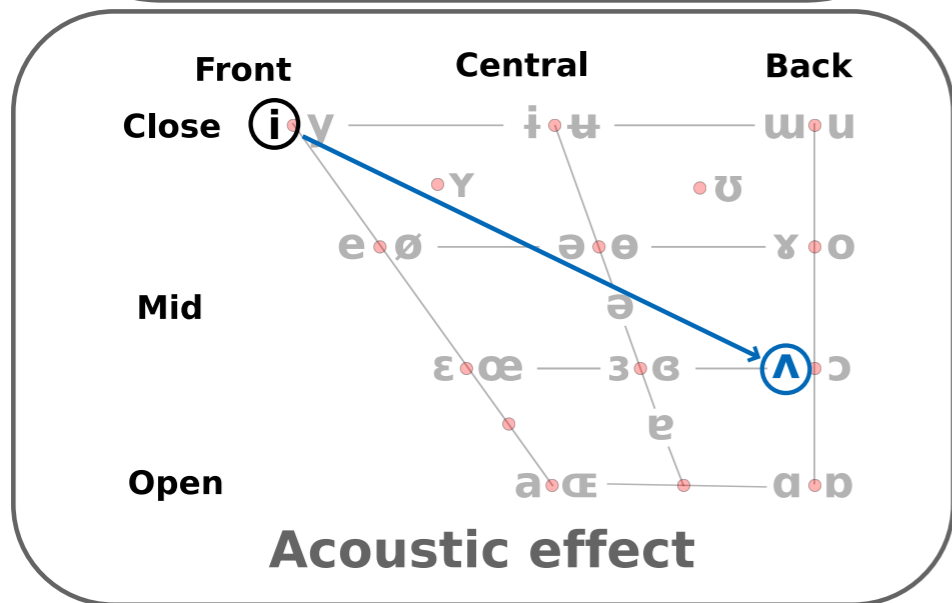
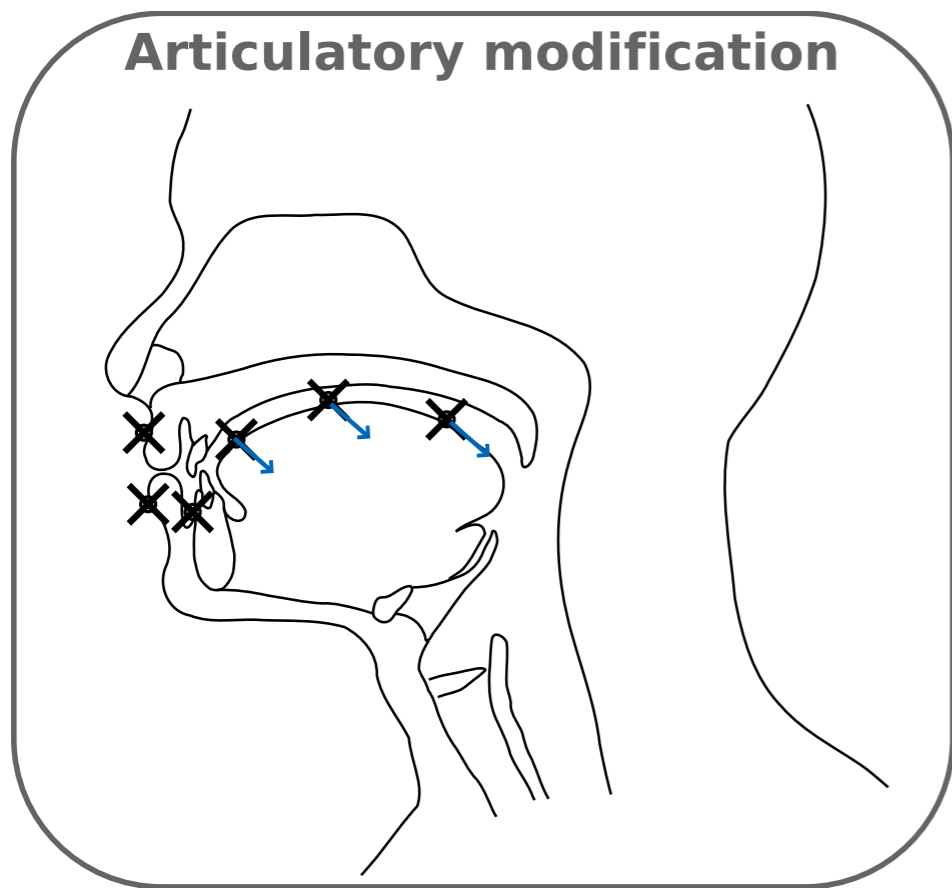
Articulatory modification changes vowel quality as we expected!



# Sample: spectrogram



# Change tongue position – Change vowel !! (Sample 2)



“him”	
default	Tongue height -1 cm Tongue position +2 cm
●	●

# Creating stimuli for speech perception experiments

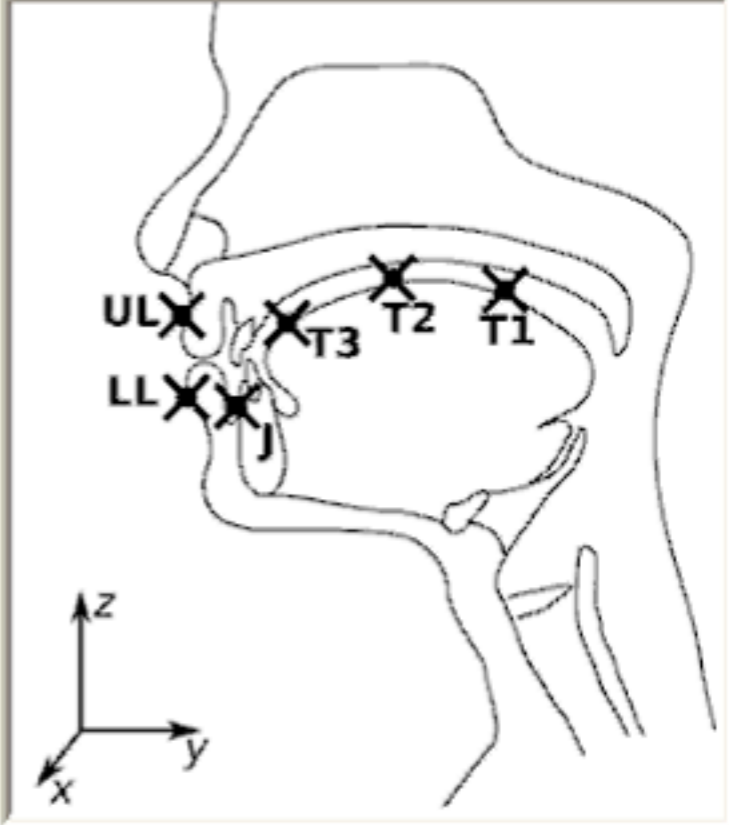
句子 音素 动画对比演示

Now we'll say  again.

动画演示

y轴统一调整: 舌根T1\_y: 舌面T2\_y: 舌尖T3\_y: 下颚J\_y: 上唇LL\_y: 下唇UL\_y:

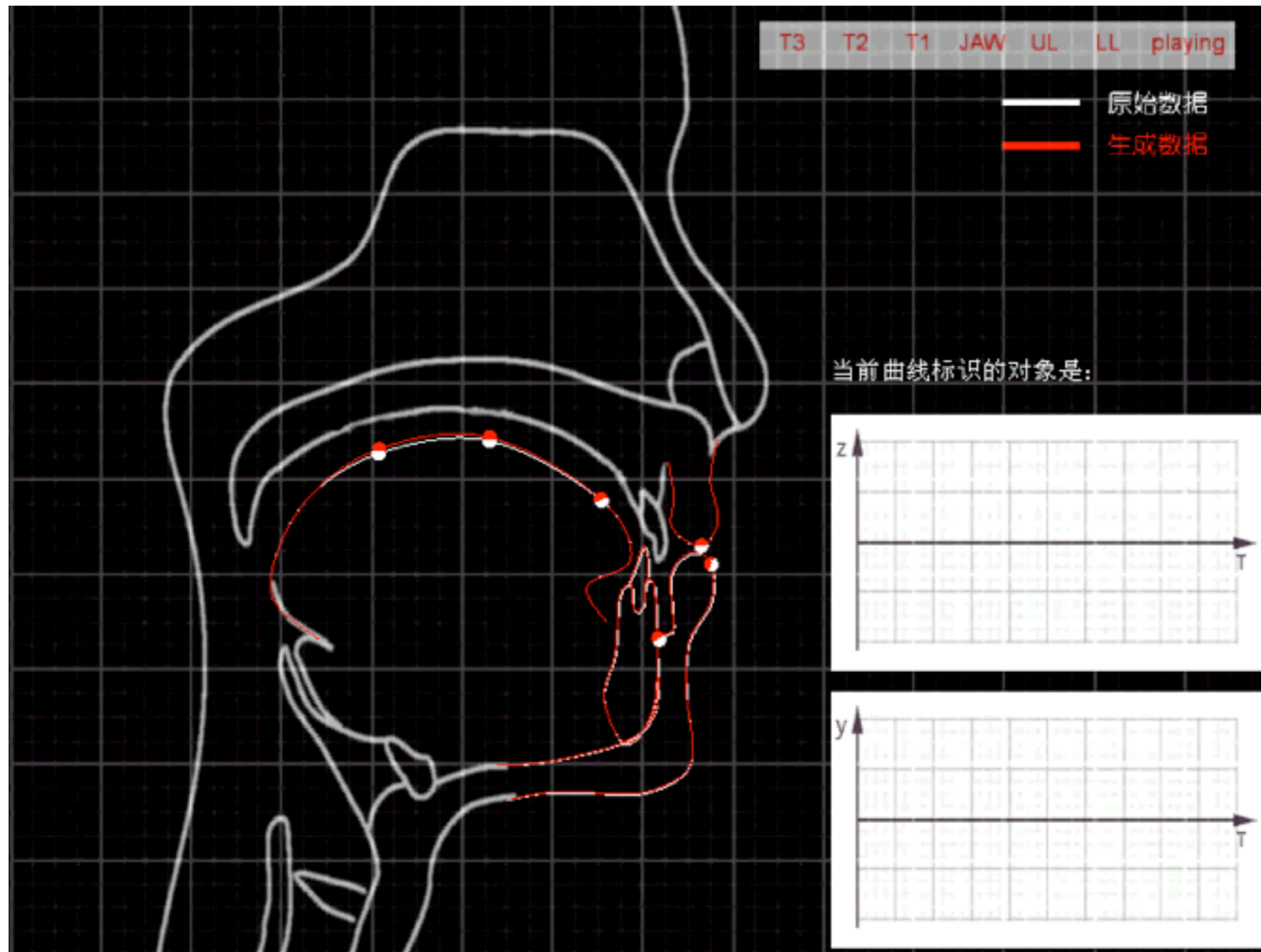
z轴统一调整: 舌根T1\_z: 舌面T2\_z: 舌尖T3\_z: 下颚J\_z: 上唇LL\_z: 下唇UL\_z:



默认参数 合成 播放原始语音 播放合成语音

Credit: Tian-Yi Zhao (USTC)

# Acoustic-to-articulatory inversion (application: computer-assisted pronunciation learning)



Credit: Tian-Yi Zhao (USTC)

# Change dynamics of tongue movement – Hypo- and hyper-articulation!

---

- Dynamic range scaling of the generated articulatory trajectory mimics hyper-articulation
- This can make the synthetic speech more intelligible, especially in noisy conditions

normal (1.0)

hypo-articulation (0.8)

hyper-articulation (1.2)

- Intelligibility tests of synthetic speech in noise
- babble noise recorded in a dining hall
- 5 dB speech-to-noise ratio
- 1.2 times scaling of z-axis led to a reduction in the WER of synthetic speech in noise from 63% to 48%

# Noise-adaptive speech synthesis

---

- Usefulness of hyper articulated speech
  - Car navigation systems
  - Dialog systems in noisy places
  - Any TTS devices in noise
- Examples
  - Car noises (Toshiba car noises)
    - Land Rover
    - Highway
    - Windy



**Normal speech**



**Hyper-articulated  
speech**



**Hyper-articulated +  
spectral-tilt-modified  
speech**

# Conclusions

---

# Conclusions

---

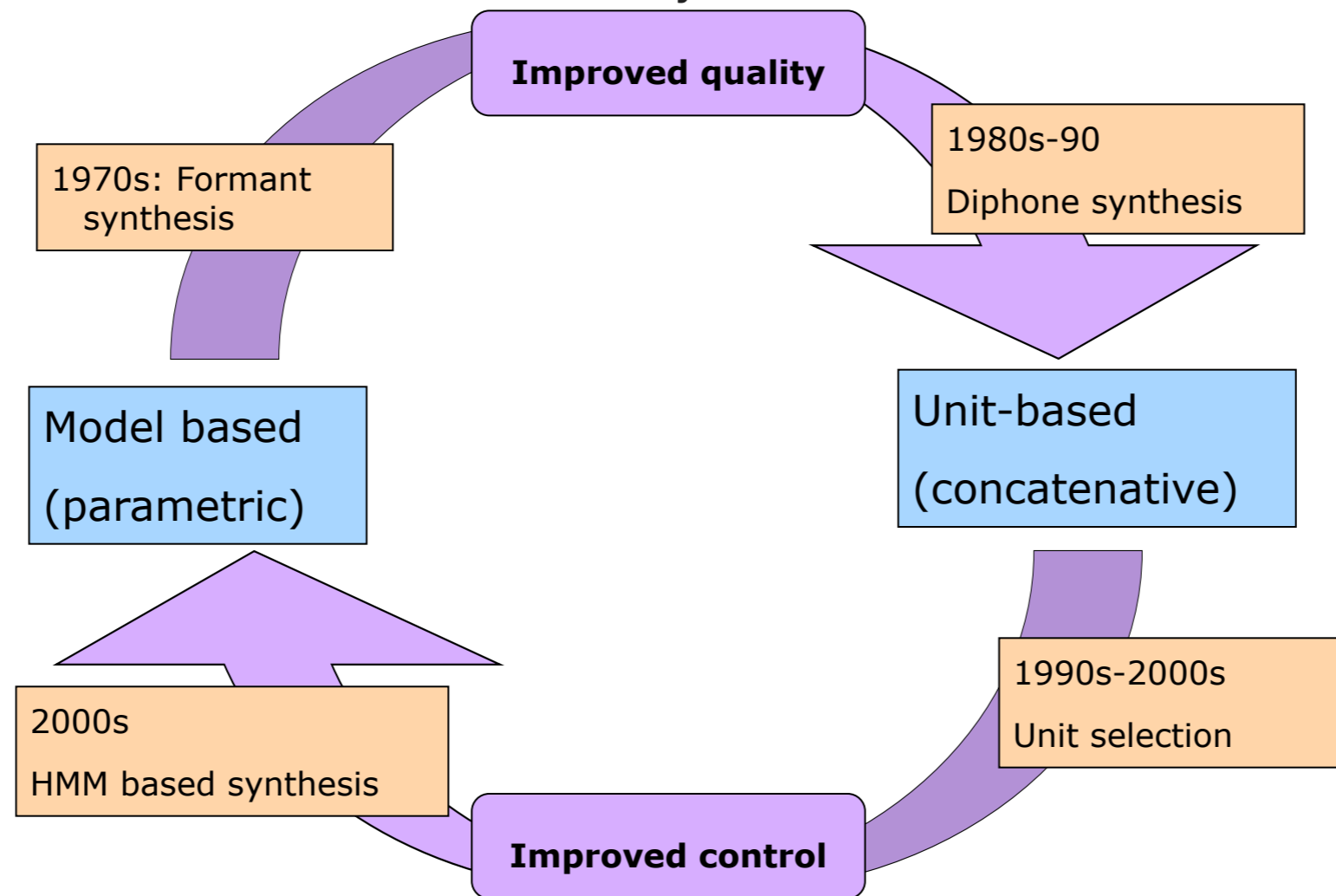
- Thanks to statistical and machine learning approaches, speech synthesis can do more than just read out text in a predefined voice.
- New research areas and interesting applications are emerging
- Now text-to-speech synthesis systems have
  - the **adaptability** and **flexibility** to
    - clone users' voices
    - reconstruct patients' voices from disordered speech
  - and the **controllability** to
    - manipulate speech via articulatory parameters
    - enhance intelligibility in noise



# A “take-home” message

---

- TTS research so far: the wheel of history



- TTS now has quality, flexibility *and* controllability
- Future TTS research should improve all these aspects simultaneously and should make the ‘next step’ rather just a ‘second turn’ of the wheel

# Credits and acknowledgements

---

- Core HMM synthesis techniques
  - K. Tokuda (NIT), H. Zen (Toshiba), T. Toda (NAIST), K. Oura (NIT), Y. Wu (Microsoft), M. Takashi (Toshiba), T. Kobayashi (TIT), T. Nose (TIT), M. Tachibana (Yamaha)
- Latest HTS samples in English, Spanish, and Romanian
  - O. Watts (CSTR), R. Barra Chicote (UPM), A. Stan (University of Cluj-Napoca)
- 1000s HTS voices
  - B. Usabaev, J. Tian, Y. Guan (Nokia), A. Suni, T. Raitio, M. Vainio, P. Alku, R. Karhila, M. Kurimo (TKK/AALTO), J. Dines (IDIAP)
- Speaker verification vs text-to-speech
  - P. De Leon (NMSU), I. Hernaez, I. Saratxaga (Univ. Basque Country), Michael Pucher (ftw)
- Celebrity voices
  - M. Aylett (Cereproc), Z. Ling (USTC)
- Child speech synthesis
  - O. Watts (CSTR), K. Berkling (Germany)
- Cross-lingual speaker adaptation
  - Y. Wu, K. Oura, K. Tokuda (NIT), J. Dines, H. Liang, L. Saheer (IDIAP), M. Gibson, W. Byrne (Cambridge), M. Wester (CSTR), and J. Cook (BBC)
- Voice banking and reconstruction
  - S. Creer, P. Green, S. Cunningham (University of Sheffield), E. MacDonald, S. Chandran (EMC)
- Articulatory modification of synthetic speech
  - Z. Ling (USTC), K. Richmond (CSTR), T. Zhao (USTC), C. Valentini (CSTR)

# Suggested follow-up reading

---

- 1000s HTS voices
  - J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, M. Kurimo “Thousands of Voices for HMM-based Speech Synthesis -- Analysis and Application of TTS Systems Built on Various ASR Corpora” IEEE Audio, Speech, & Language Processing, vol.18, issue.5, pp.984-1004, July 2010
- Speaker verification vs text-to-speech
  - P. De Leon, M. Pucher, J. Yamagishi “Evaluation of the vulnerability of speaker verification of synthetic speech”, Proc. Odyssey 2010 (The speaker and language recognition workshop), pp151-158, July 2010
- Robust speech synthesis and celebrity voices
  - J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, S. Renals, “A Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis,” IEEE Audio, Speech, & Language Processing, vol.17, no.6, pp.1208-1230, August 2009
- Child speech synthesis
  - O. Watts, J. Yamagishi, S. King, K. Berkling, “Synthesis of Child Speech with HMM Adaptation and Voice Conversion” IEEE Audio, Speech, & Language Processing, vol.18, issue.5, pp.1005-1016, July 2010
- Unsupervised Cross-lingual speaker adaptation
  - J. Dines, H. Liang, L. Saheer, M. Gibson, W. Byrne, K. Oura, K. Tokuda, J. Yamagishi, S. King, M. Wester, T. Hirsimaki, R. Karhila, M. Kurimo. “Personalising Speech-to-Speech Translation: Unsupervised Cross-lingual Speaker Adaptation for HMM-based Speech Synthesis,” under review
- Voice reconstruction
  - S. Creer, P. Green, S. Cunningham, and J. Yamagishi, “Building personalised synthesised voices for individuals with dysarthria using the HTS toolkit,” Computer Synthesized Speech Technologies: Tools for Aiding Impairment John W. Mullennix and Steven E. Stern (Eds), IGI Global press, Jan. 2010. ISBN: 978-1-61520-725-1
- Articulatory modification of synthetic speech
  - Z-H. Ling, K. Richmond, J. Yamagishi, R.-H. Wang “Integrating Articulatory Features into HMM-based Parametric Speech Synthesis, IEEE Audio, Speech, & Language Processing.vol.17 No.6 pp.1171-1185 August 2009 (**IEEE Signal Processing society Young Author Best Paper Award 2010**)