Speech Recognition (Robustness)

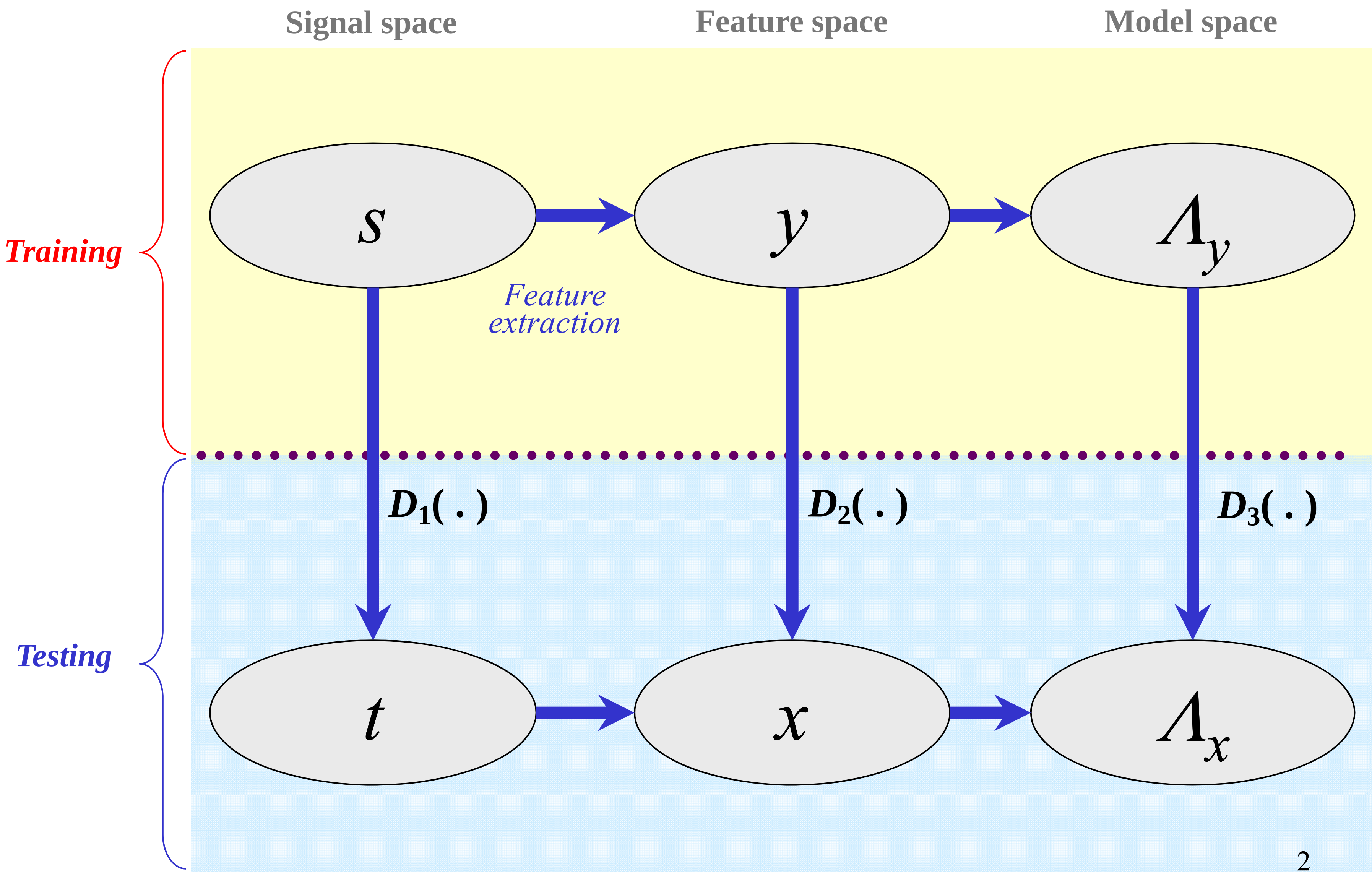
Sadaoki Furui
Tokyo Institute of Technology
Department of Computer Science
furui@cs.titech.ac.jp

Robust speech recognition

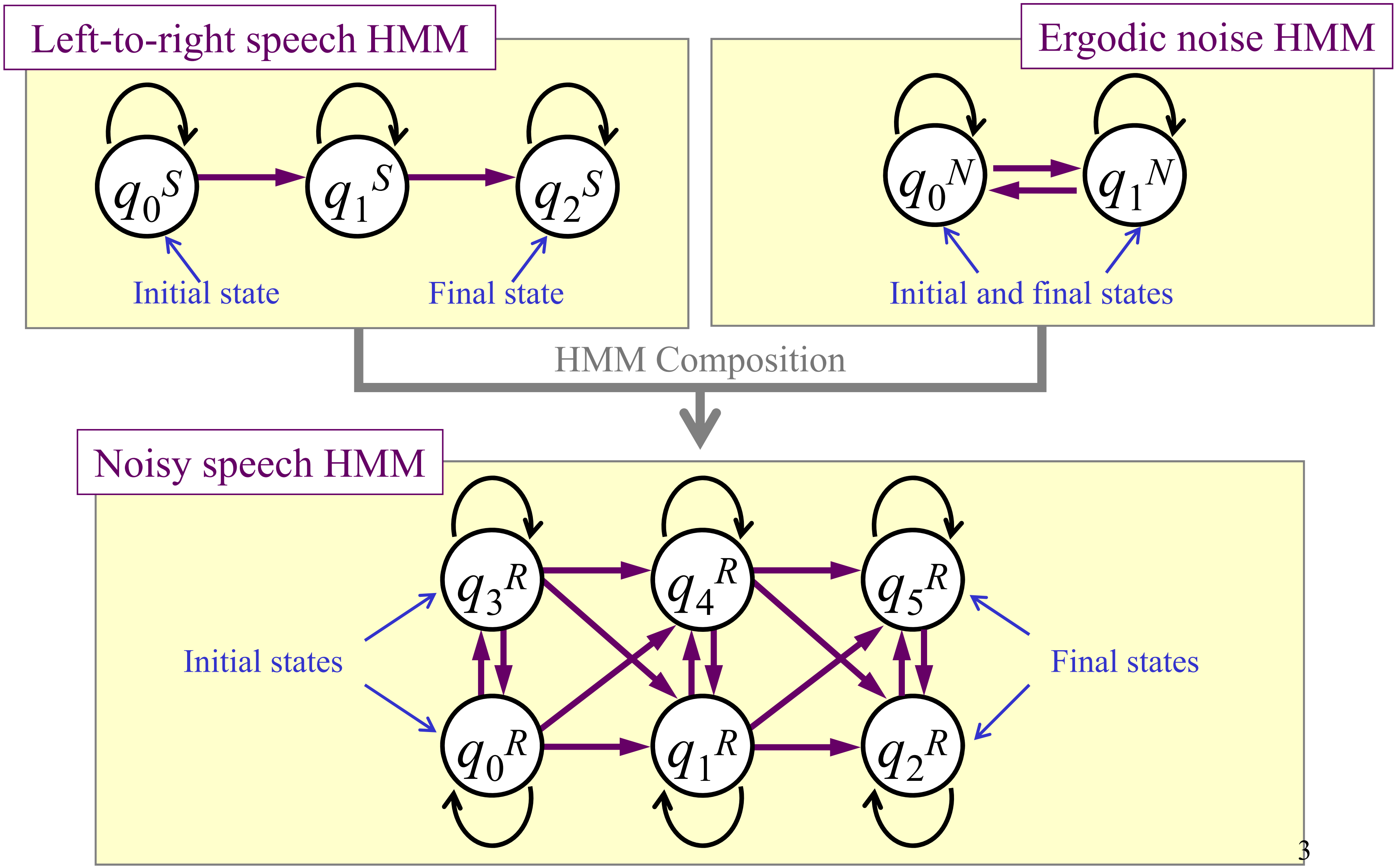
- Robust against **voice variation** due to individuality, the physical and psychological condition of the speaker, telephone sets, microphones, network characteristics, additive background noise, speaking styles, etc.
- Few restrictions on **tasks and vocabulary**
- Essential to develop automatic **adaptation** techniques
- **Unsupervised, on-line, incremental adaptation** is ideal

⇒ the system works as if it were a speaker/task-independent system, and it performs increasingly better as it is used

Mismatch between training and testing



HMM composition (PMC) process for creating a noisy speech HMM as a product of two source HMMs



Adaptation issues

- What's given – HMM, codebook, adaptation data, etc.
- Assumptions – Bayesian, transformation, combined
- Correlation – Unit and model parameter dependency
- Supervision – Supervised vs. unsupervised,
  text dependency
- Strategy – Batch, incremental, instantaneous (self
  adaptation), on-line vs. off-line
- Efficiency – Rate of adaptation, performance
- Combined equalization, normalization, and adaptation
- Speaker, environment, channel, transducer, task adaptation

Use of constraints in adaptation

- Exploiting correlation structure between parameters:
  - (Hierarchical) spectral clustering and smoothing
  - Mixture tying
  - Codebook mapping
  - Probabilistic spectral mapping
  - Acoustic bias normalization and context bias modulation
  - Stochastic matching

- Set of constraints on model parameters:
  - Multiple-regression-based prediction
  - Linear transformation between reference and adaptive vectors (translated into a bias vector and a scaling matrix, which can be estimated with an EM algorithm)

MLLR (maximum likelihood linear regression) for speaker adaptation of continuous density HMMs

$\hat{\mu} = \Gamma\zeta$

$\zeta = [\omega, \mu_1, \ldots \mu_n]'$: $(n+1)$-dimensional extended mean vector

μ : n-dimensional mean vector

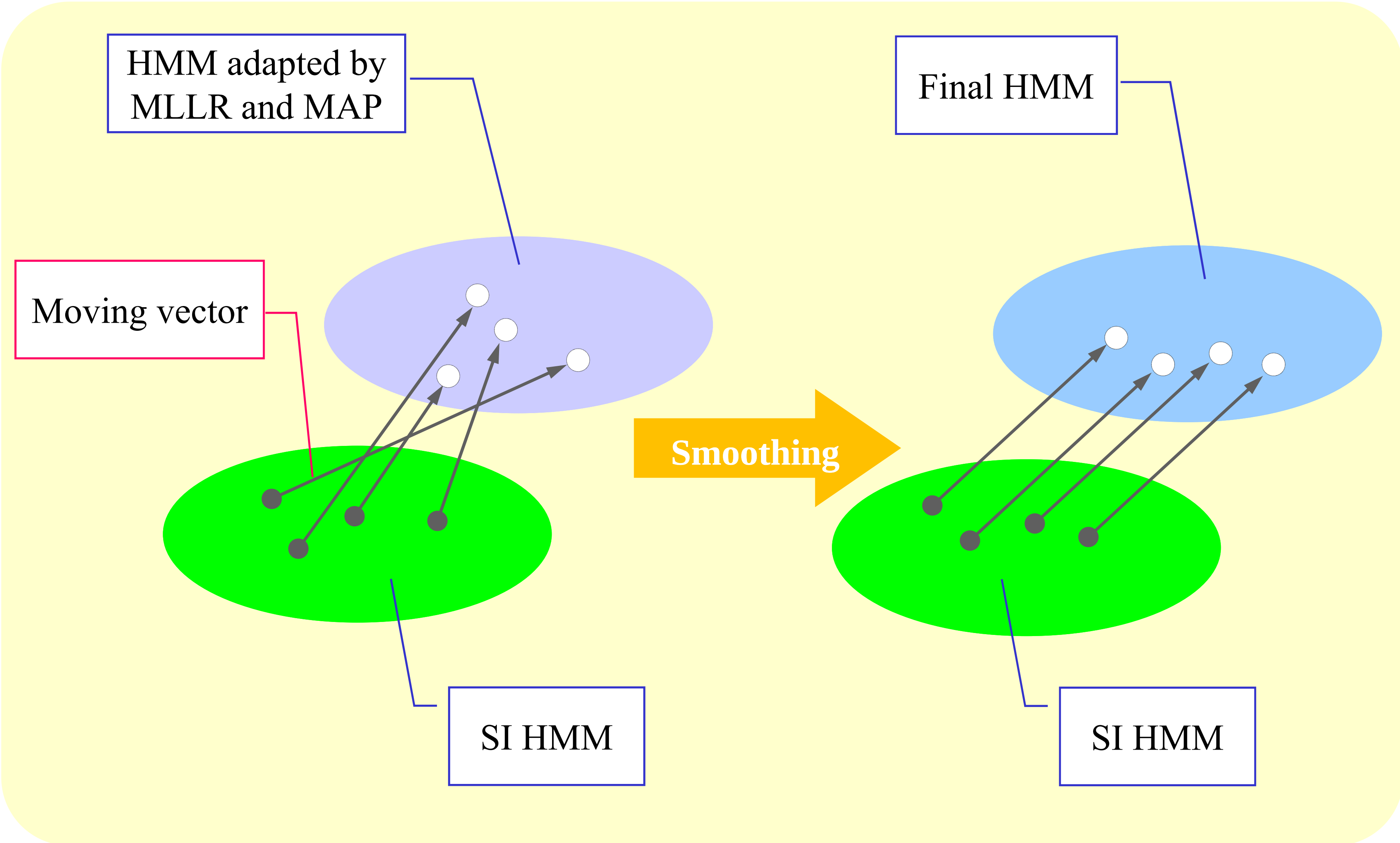
ω : offset term

$$\begin{cases} \omega = 1 : \text{include an offset in the regression} \\ \omega = 0 : \text{ignore offsets} \end{cases}$$

$\hat{\mu}$: adapted mean vector

Γ : $n \times (n+1)$ transformation matrix maximizing the likelihood of the adaptation data

Vector Field Smoothing (VFS)



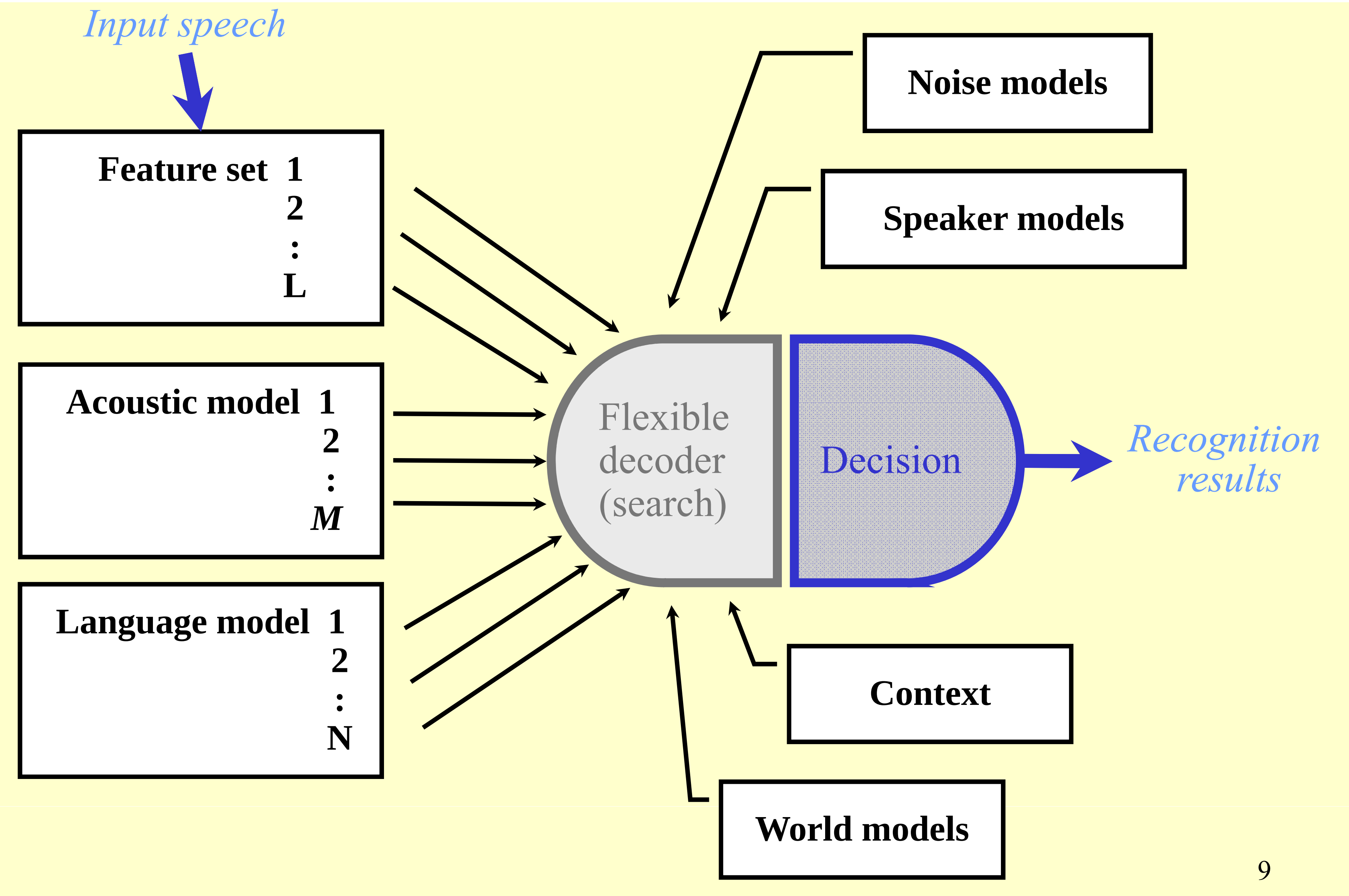
Noise/speaker/task adaptation in the search framework

- **Dialogue state/task/speaker-dependent LMs**
  (e.g. for mixed-initiative dialogue)
- **Speaker-class-dependent AMs**
- **Noise-class-dependent AMs**

- **Maximum likelihood model selection**

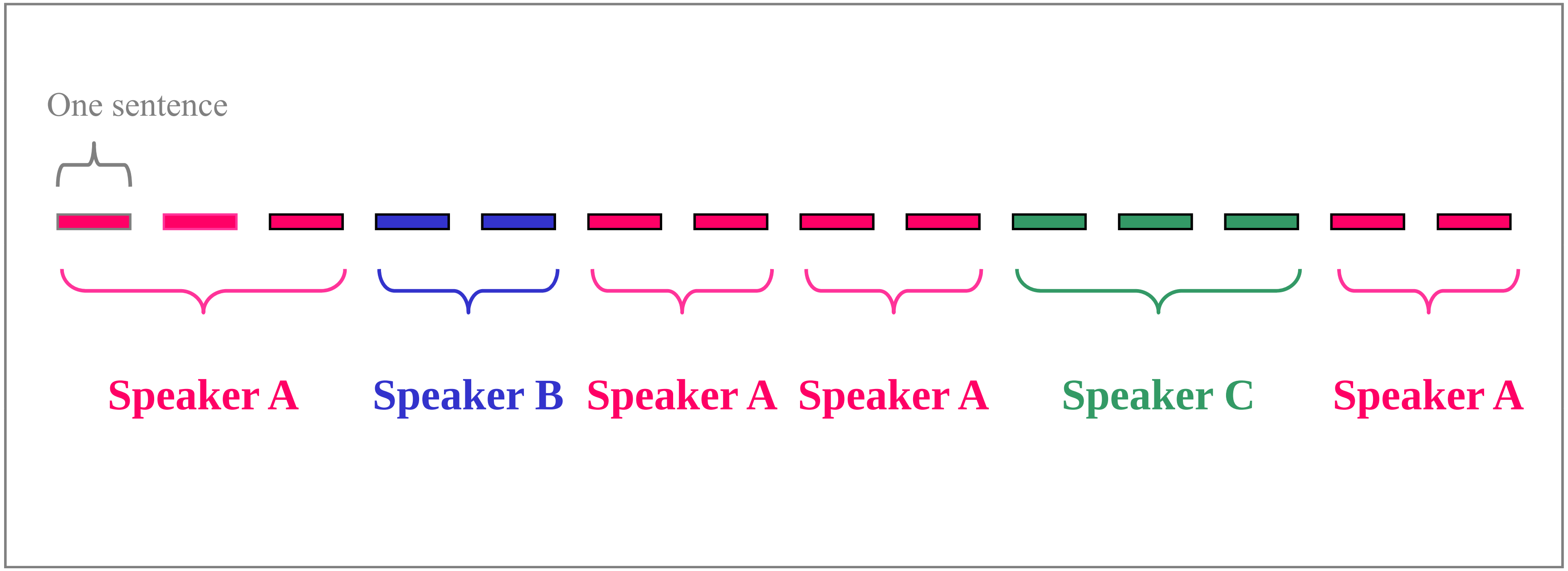
- **Maximum likelihood model adaptation**

Flexible speech recognition (adaptive search)



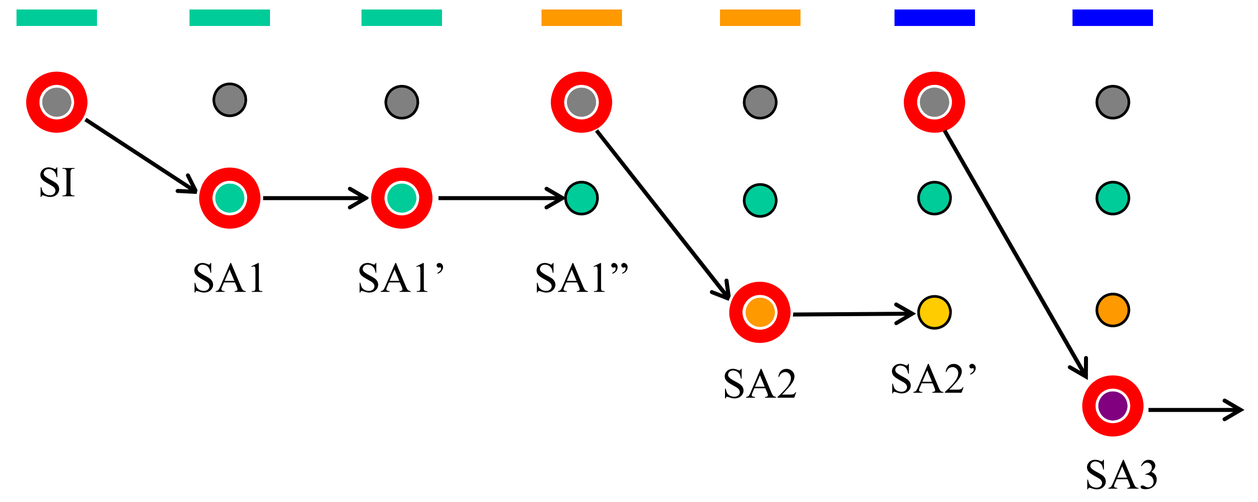
Features of broadcast news and meeting speech

- Frequent speaker changes
- Each speaker continuously utters several sentences



- Online, incremental adaptation within a segment in which one speaker utters continuously is ideal

Speaker adaptation process



Sentence utterance
SI : Speaker Independent model
SA : Speaker-Adapted model
Largest likelihood model

Adaptation algorithm

- ***MLLR-MAP-VFS* algorithm**
  - ***MLLR* : Maximum Likelihood Linear Regression**
  - ***MAP* : Maximum A Posteriori estimation**
  - ***VFS* : Vector Field Smoothing**

- **Phone clustering for MLLR**
  - **7 clusters (silence, consonants, and five Japanese vowels)**

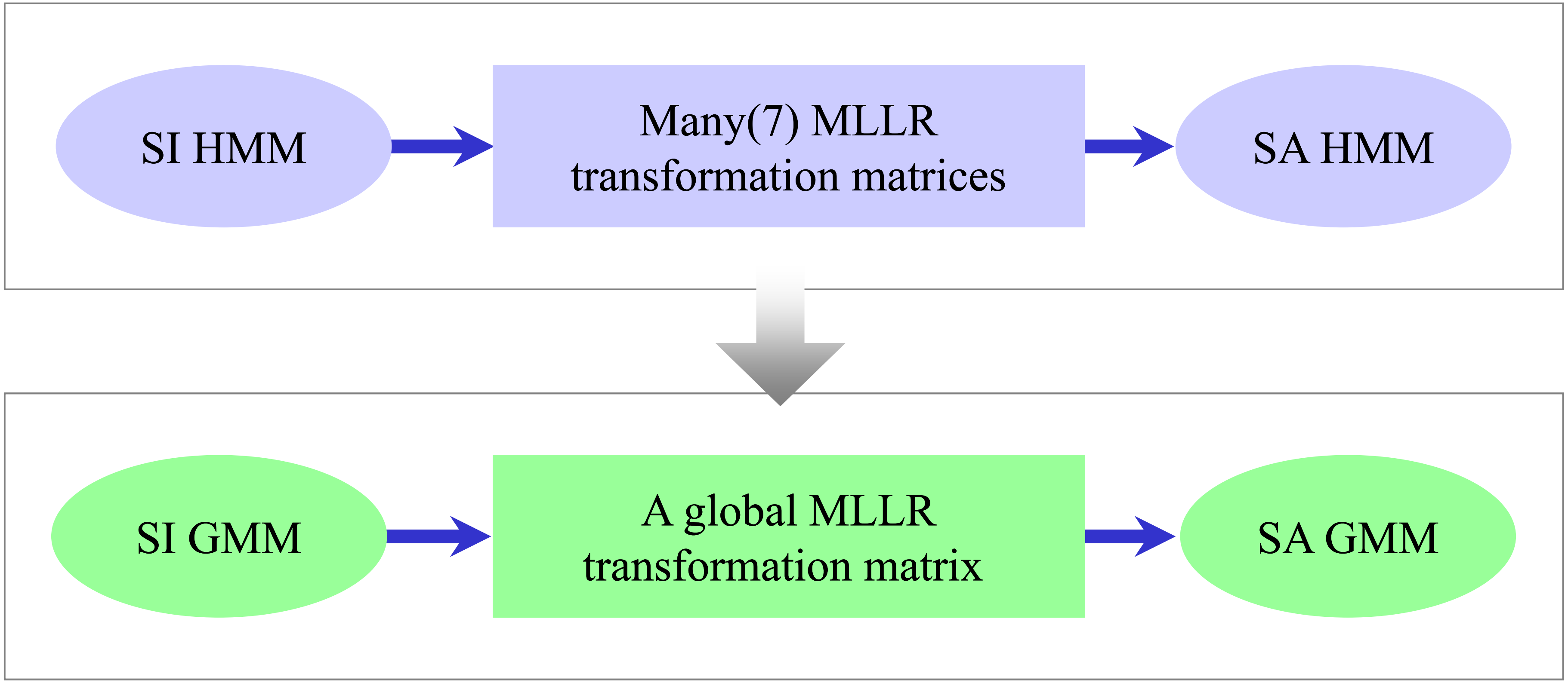
Speaker adaptation with GMM-based speaker change detection

- Reduction of the amount of computation :
  - Using a single state GMM instead of HMM
  - Advantage of GMM : simple structure

- GMMs for speaker change detection
  - Speaker adapted GMMs (SA GMMs)

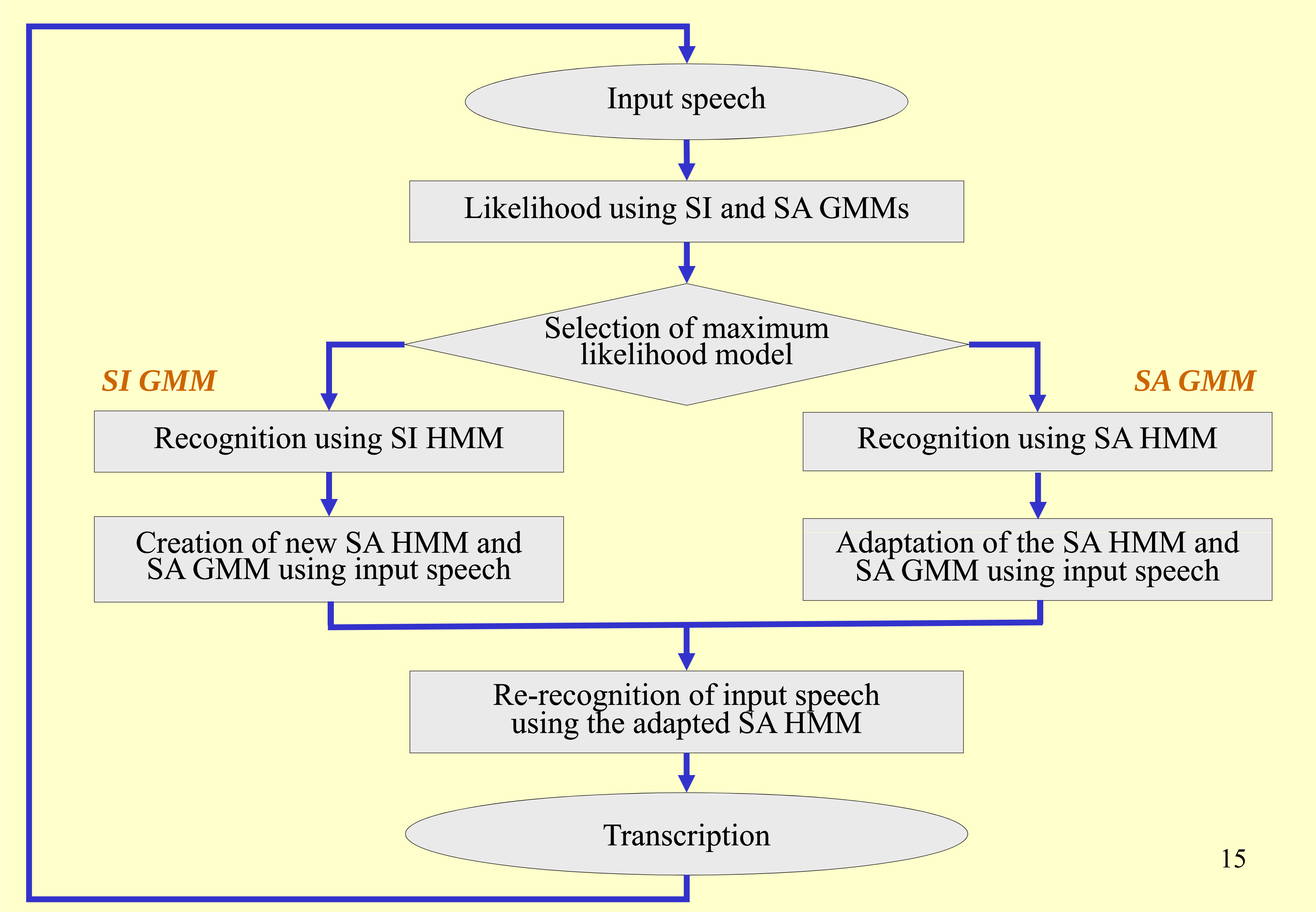
Problem: how to make the SA GMMs

Construction of SA GMM



- HMM : phone-class clustering
- GMM : transforming SI GMM into SA GMM using a single global HMM transformation matrix

On-line incremental speaker adaptation including speaker change detection (SI: speaker independent, SA: speaker adapted)



MDL-based cluster number decision for speaker adaptation

- **MDL: minimum description length criterion**

$$l^{(i)} = -\log P_{\hat{\theta}(i)}(X^N) + \frac{\alpha_i}{2}\log N + \log I$$

$P_{\hat{\theta}(i)}(X^N)$: likelihood

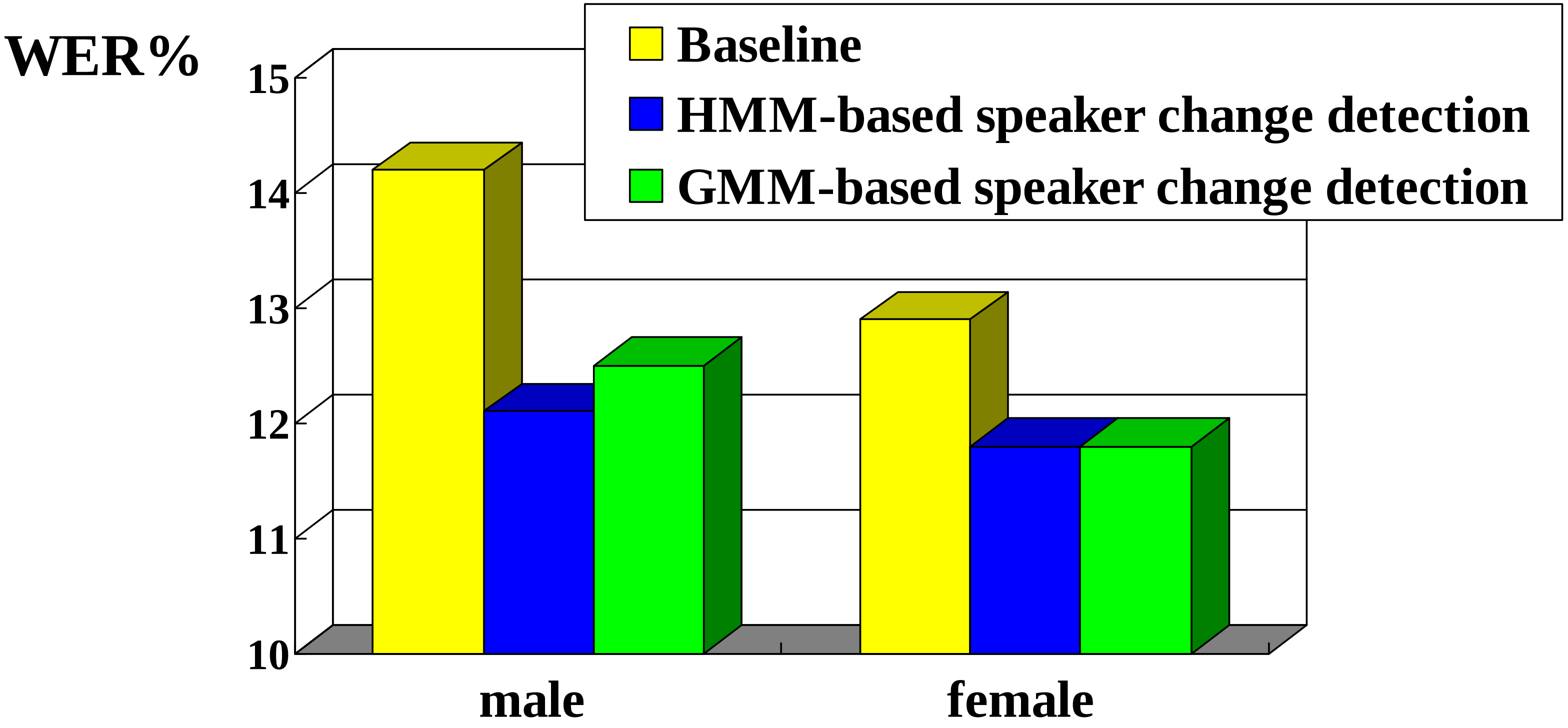
$\hat{\theta}_{(i)}$: maximum likelihood estimate for parameter θ of model i $(1 \le i \le I)$

$X^N = x_1, \ldots, x_N$: data

α_i : number of free parameters

- **MLLR-based speaker adaptation**
  - Number of phoneme clusters
  - Number of speaker clusters

Experimental results using GMM for speaker change detection

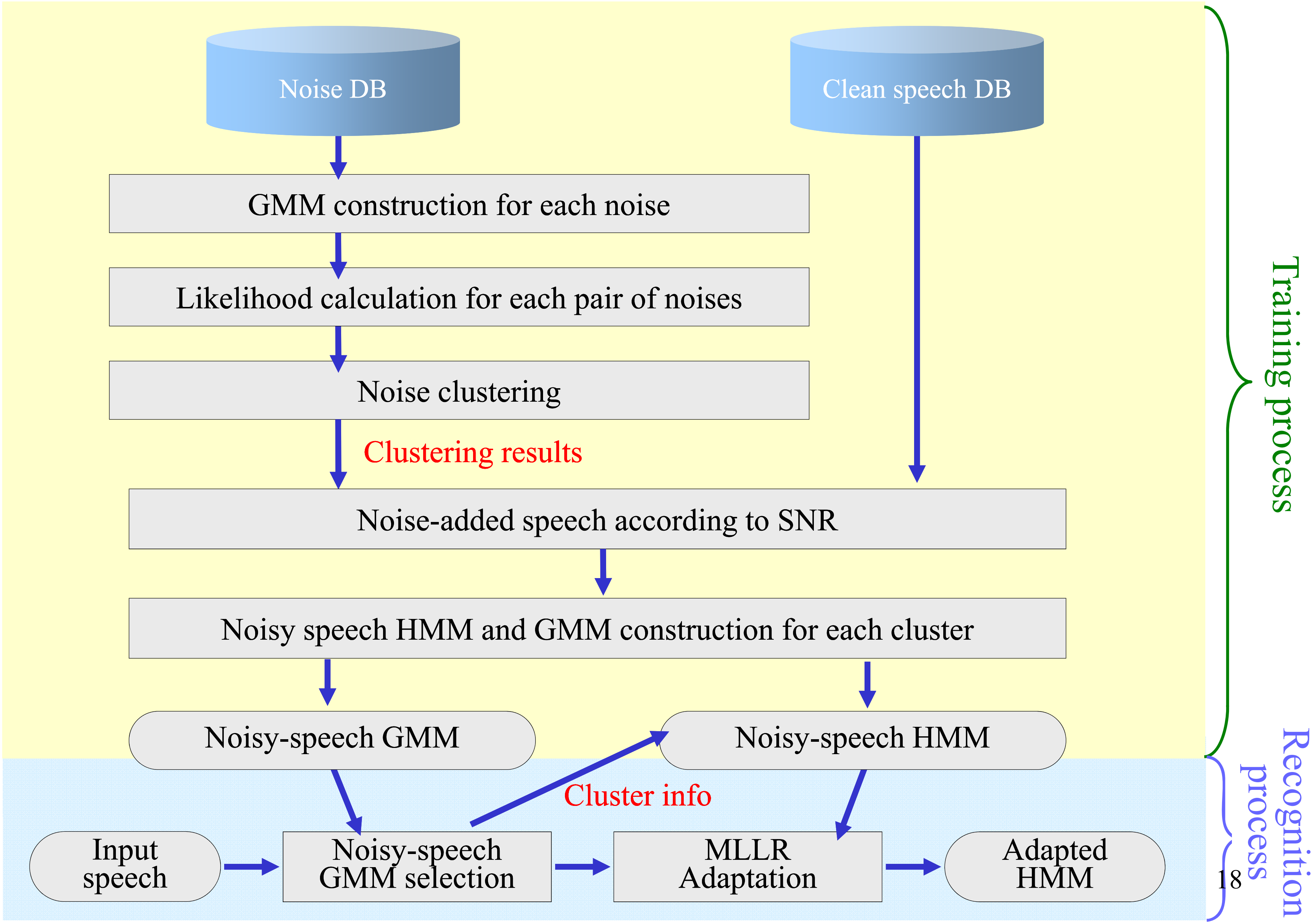


HMM-based method : 11.8% reduction of WER

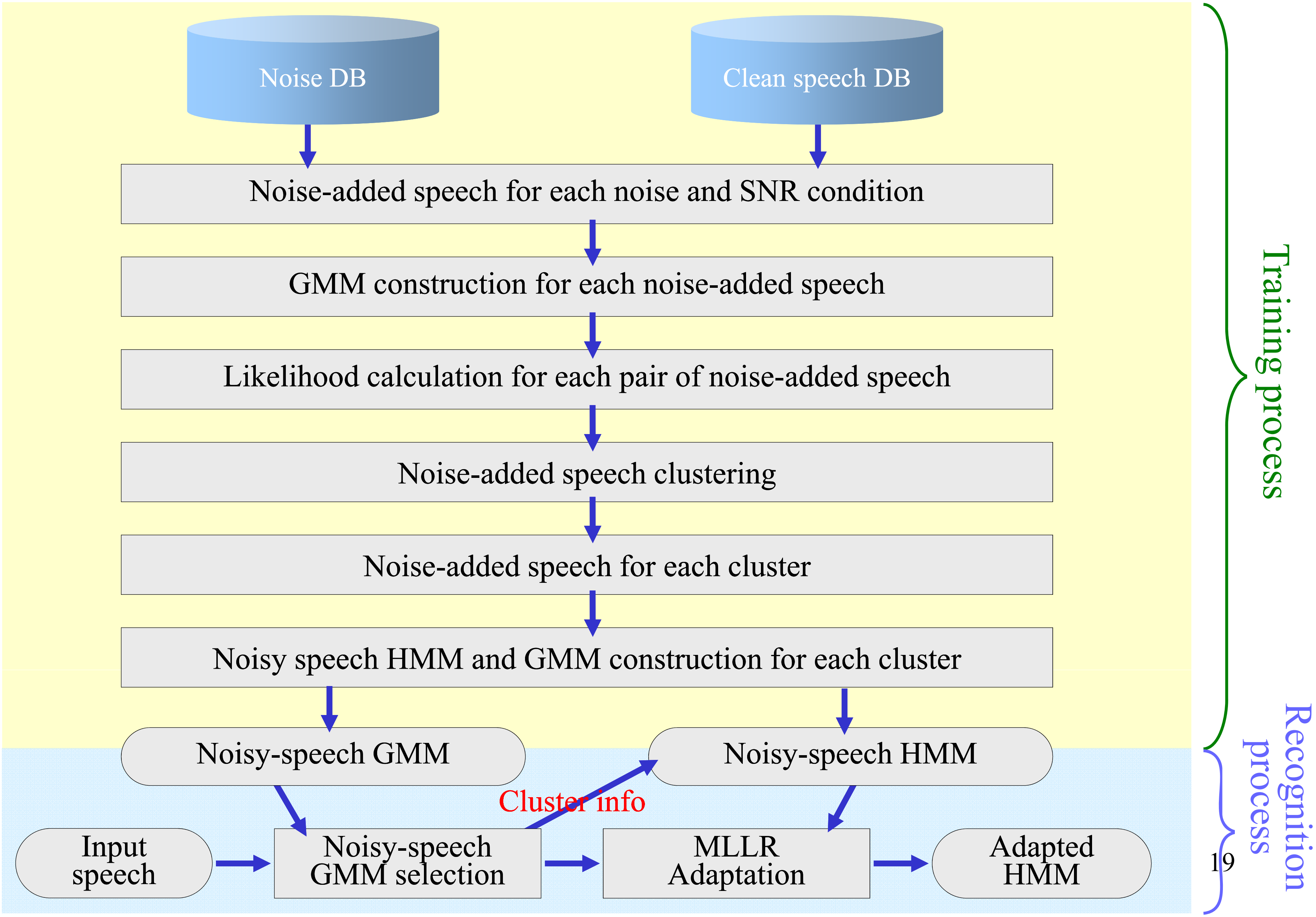
GMM-based method : 10.0% reduction of WER

Significant reduction in the amount of computation

Piecewise-linear transformation for HMM noise adaptation (Noise clustering)

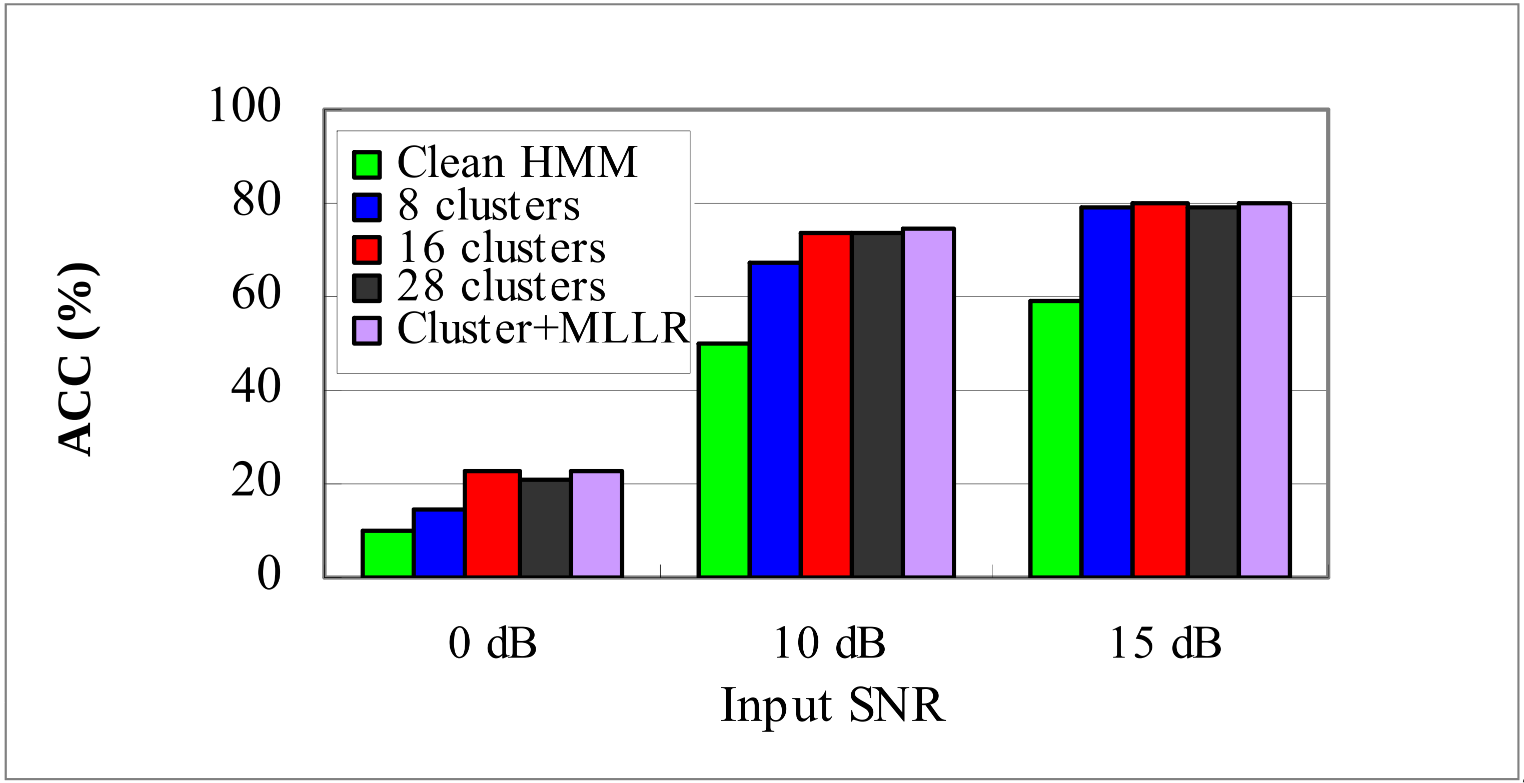


Piecewise-linear transformation for HMM noise adaptation (noisy-speech clustering)

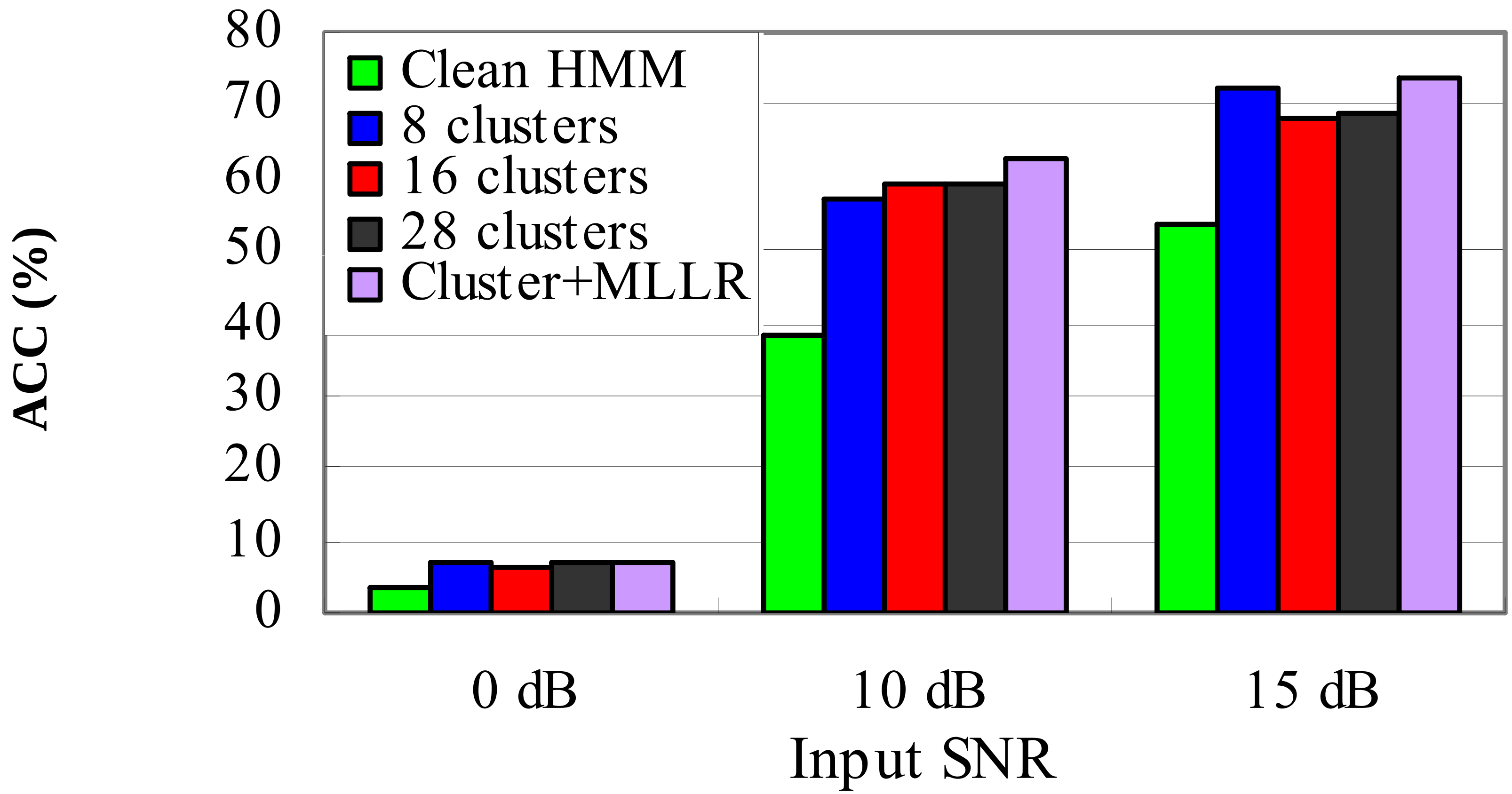


Effectiveness of the PLT method
(Artificially added crowd noise)

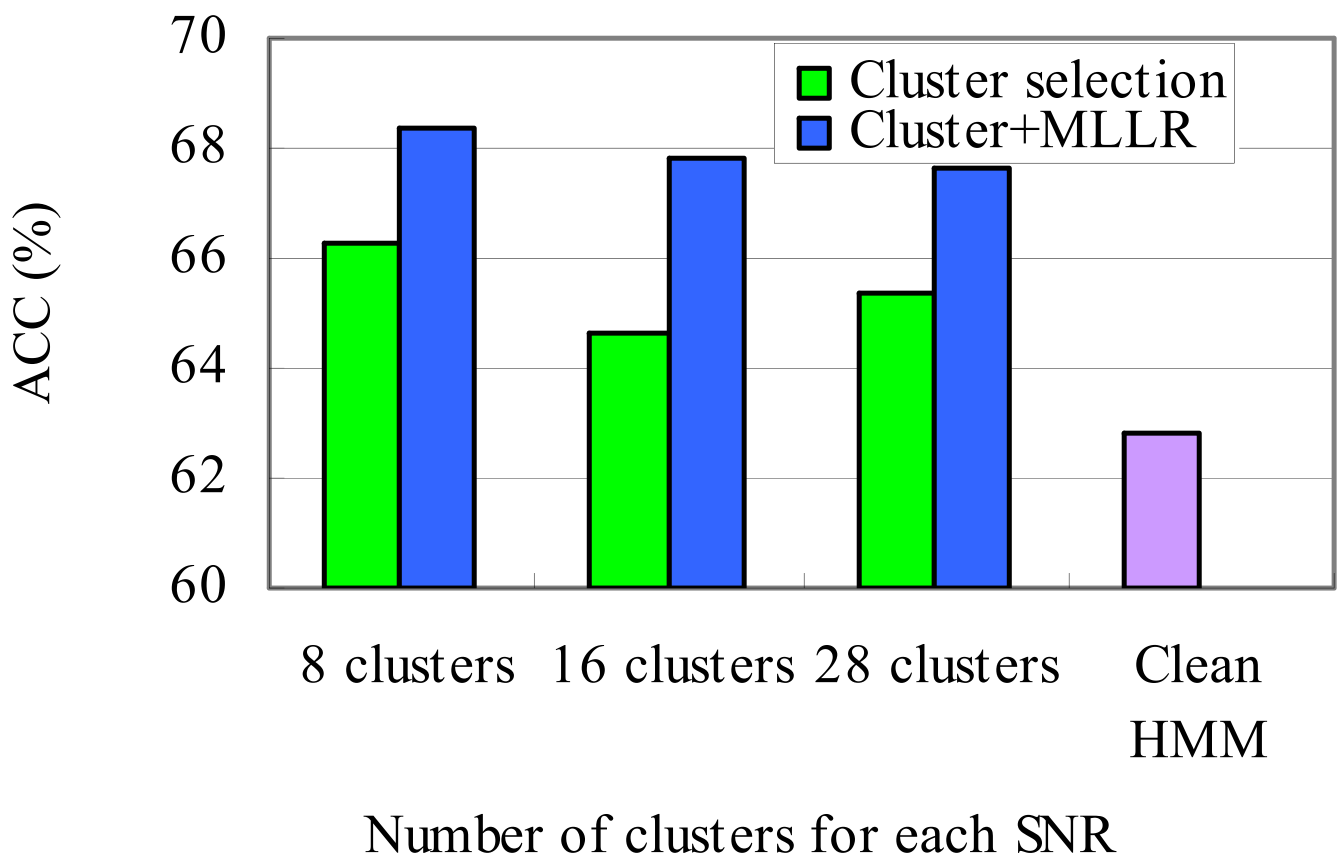
- Input utterance SNR is unknown: a noise-cluster HMM is selected from all noise-cluster HMMs with 0, 10, 15 or 20dB SNR
- Noise cluster GMM is used for cluster selection



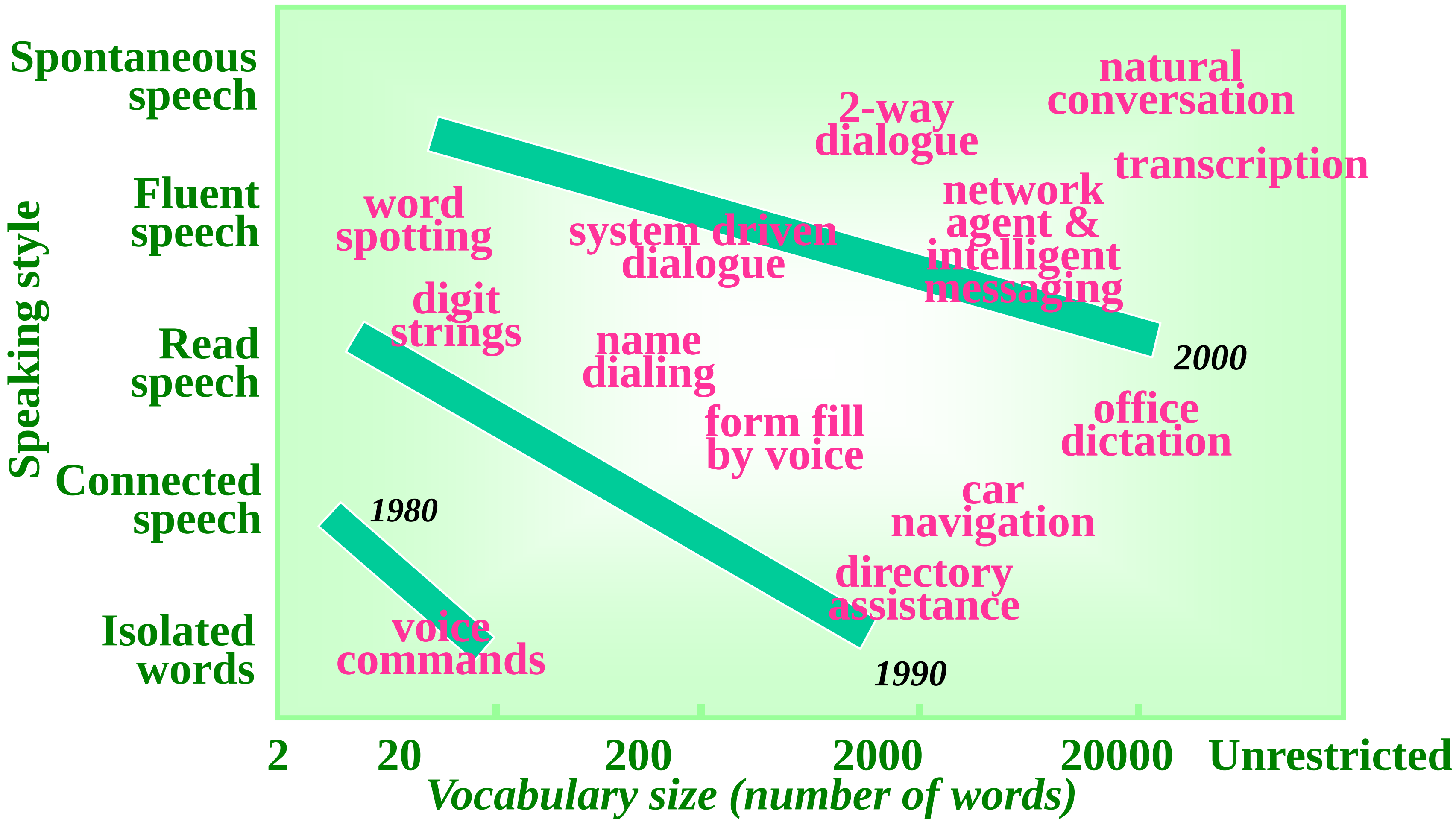
Effectiveness of the PLT method (Artificially added exhibition hall noise)



Effectiveness of the PLT method
(Real noisy speech from broadcast news)



Progress of speech recognition technology since 1980



Difficulties in (spontaneous) speech recognition

- Lack of systematic understanding in variability
  - Structural or functional variability
  - Parametric variability

- Lack of complete structural representations of (spontaneous) speech

- Lack of data for understanding non-structural variability

Spontaneous speech corpora

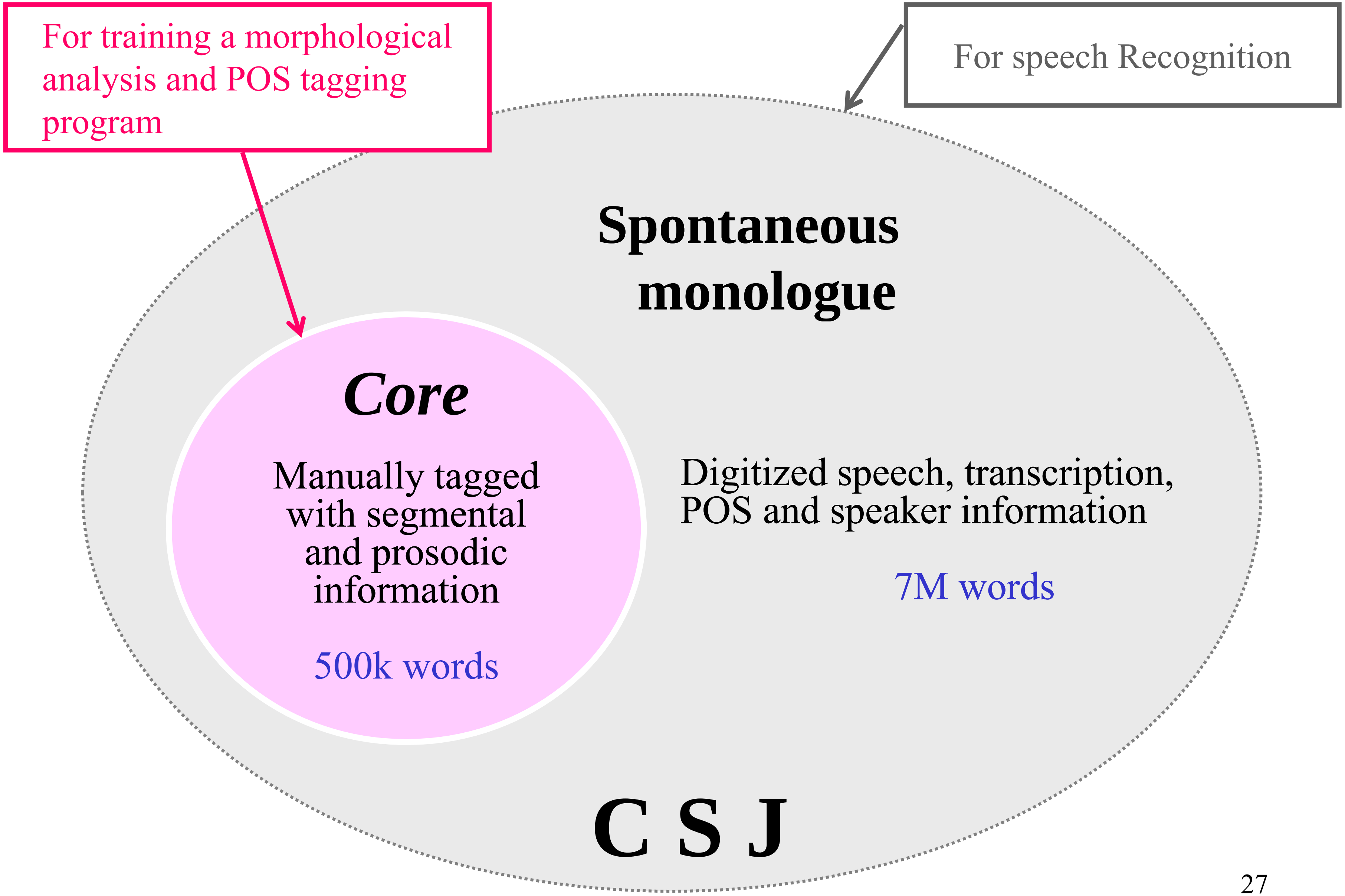
- **Spontaneous speech variations:** extraneous words, out-of-vocabulary words, ungrammatical sentences, disfluency, partial words, repairs, hesitations, repetitions, style shifting, ….
- **“There’s no data like more data”** – Large structured collection of speech is essential.
- **How to collect *natural* data?**
- **Labeling and annotation of spontaneous speech is difficult;** how do we annotate the variations, how do the phonetic transcribers reach a consensus when there is ambiguity, and how do we represent a semantic notion?

Spontaneous speech corpora (cont.)

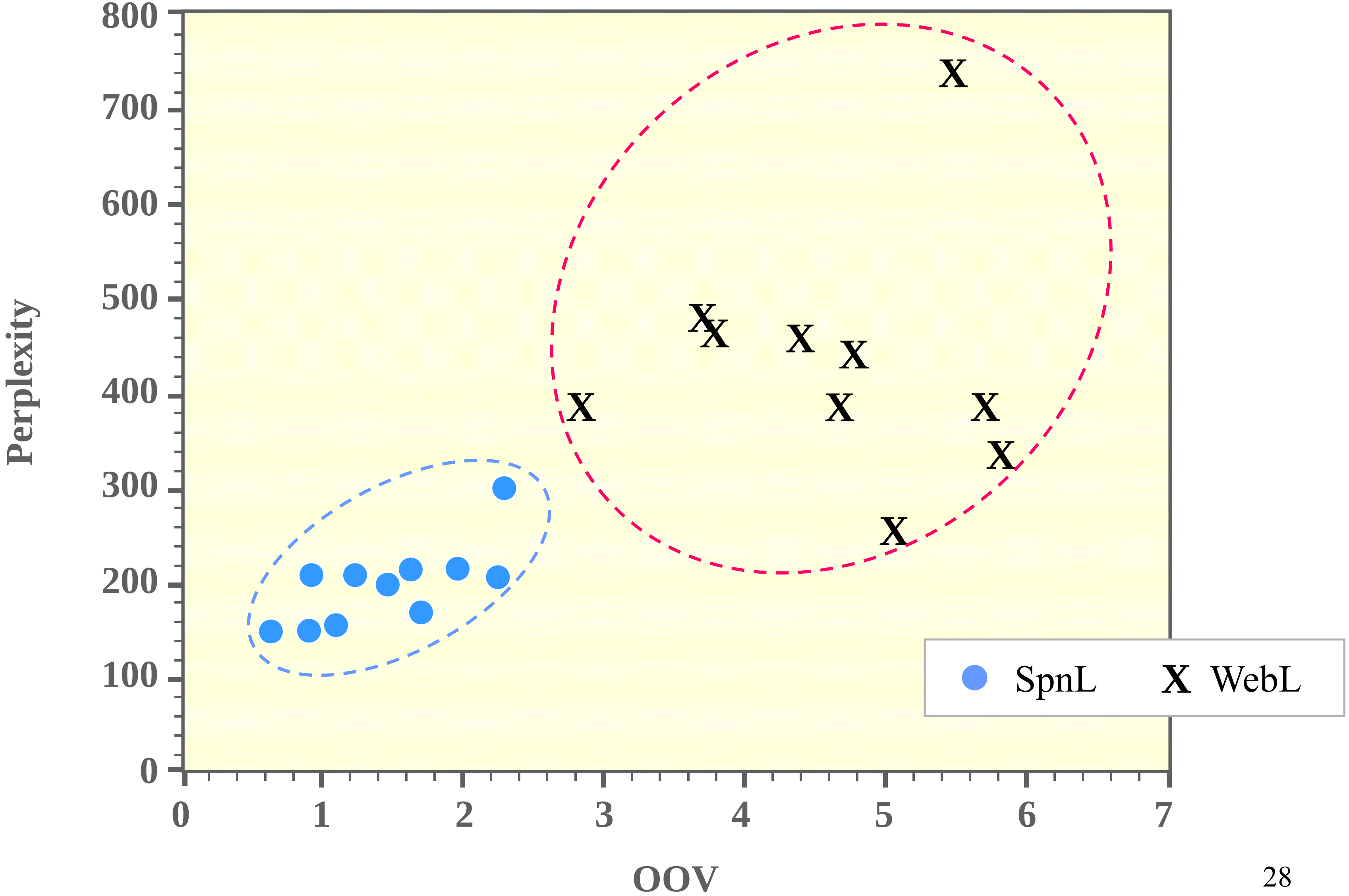
- **How to ensure the corpus quality?**
- Research in **automating or creating tools to assist the verification procedure** is by itself an interesting subject.
- **Task dependency:** It is desirable to design a task-independent data set and an adaptation method for new domains.

Benefit of a reduced application development cost.

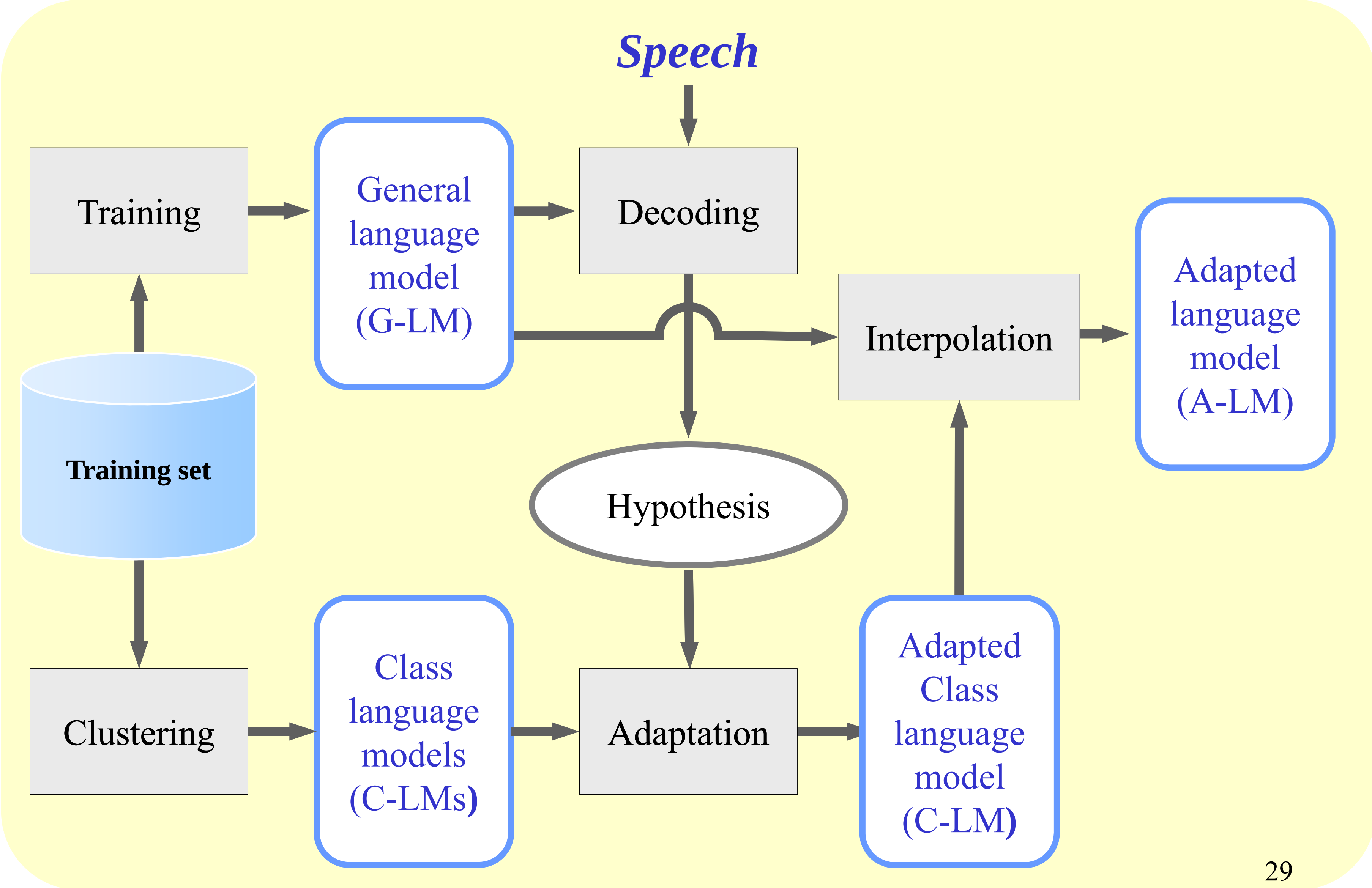
Overall design of the Corpus of Spontaneous Japanese (CSJ)



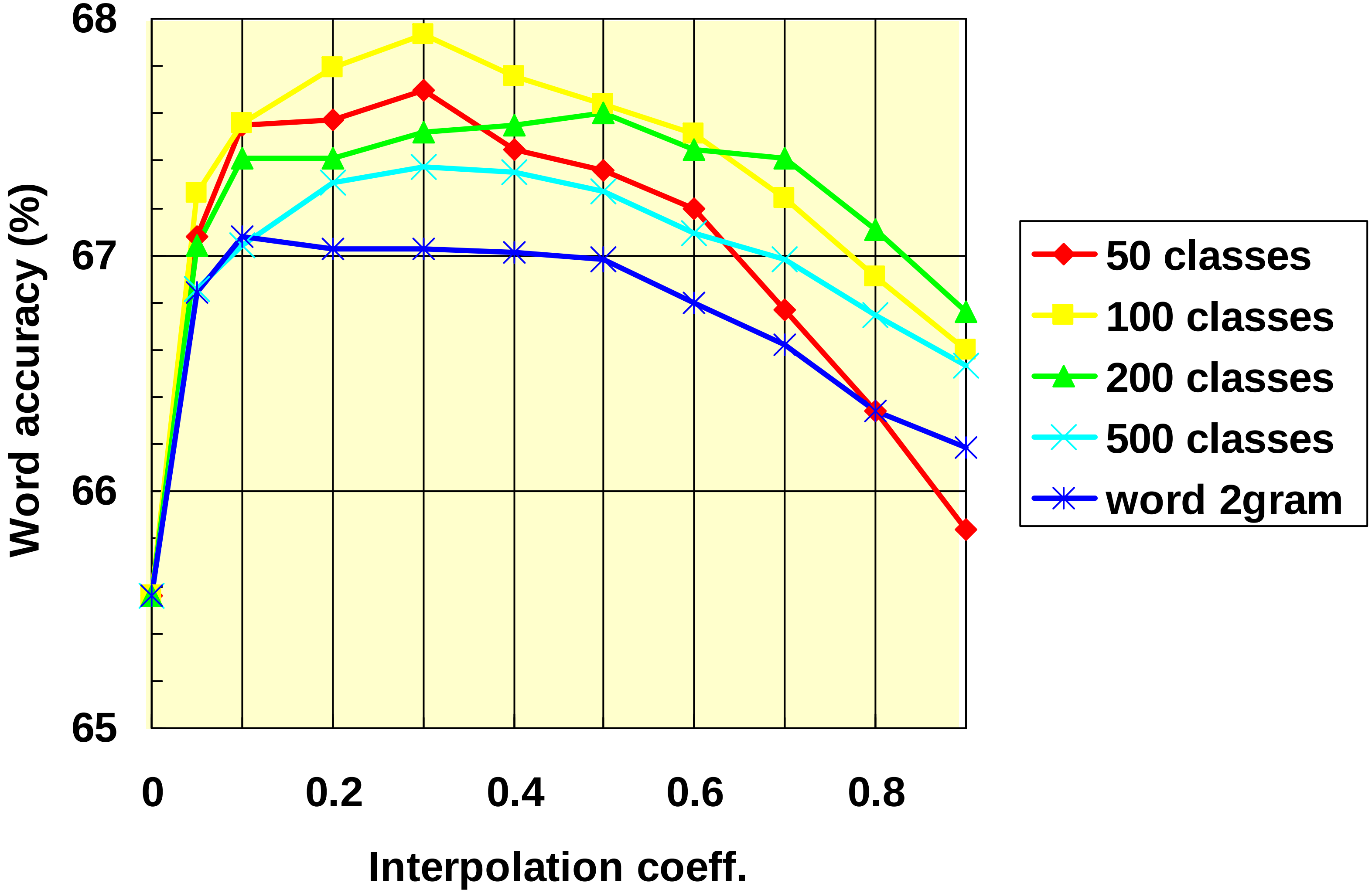
Test-set perplexity and OOV rate for the two language models



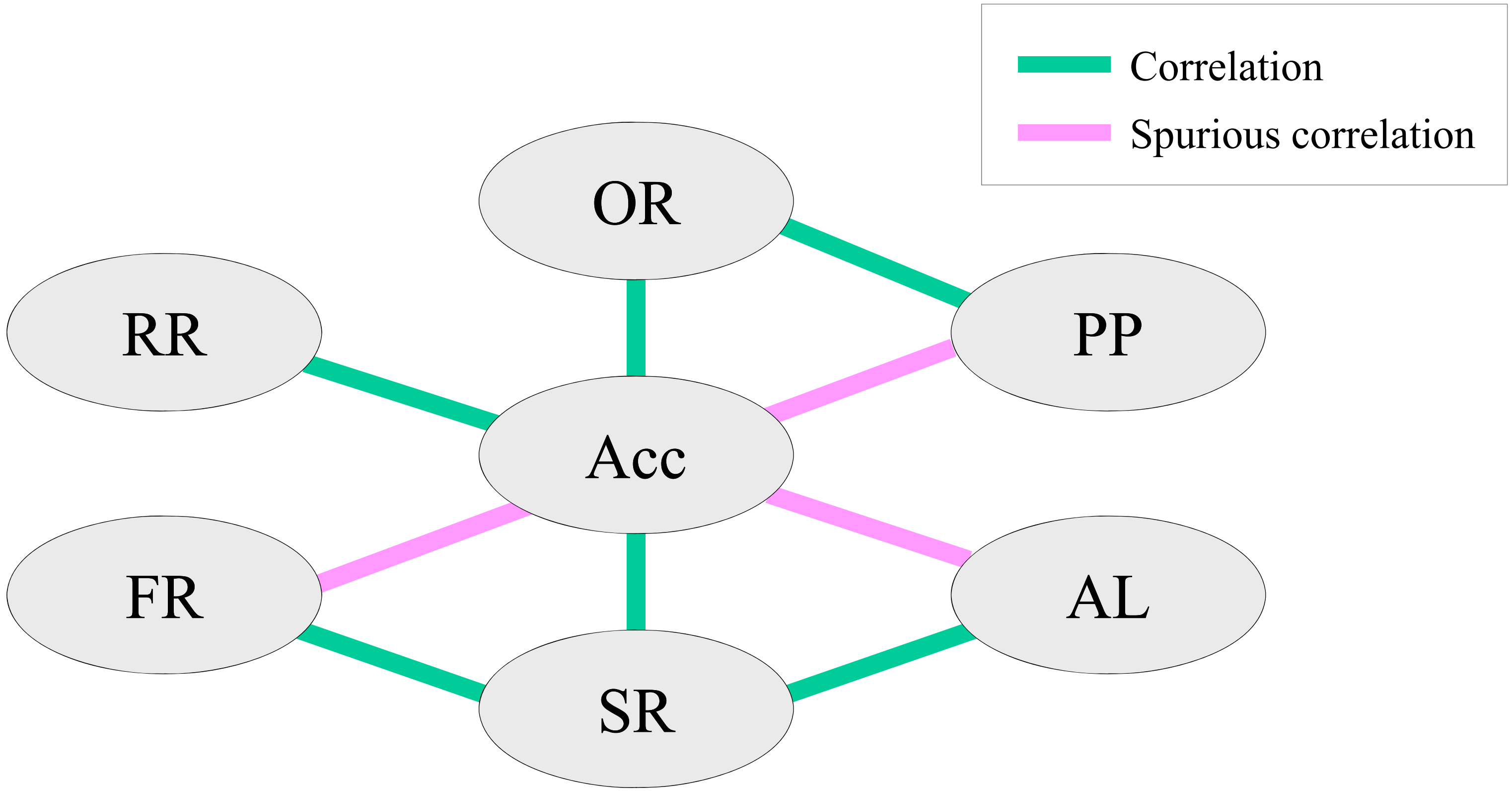
Unsupervised class-based language model adaptation



Word accuracy vs. interpolation coefficient



Summary of correlation between various attributes



Acc: word accuracy, OR: out of vocabulary rate,
RR: repair rate, FR: filled pause rate,
SR: speaking rate, AL: averaged acoustic frame likelihood,
PP: word perplexity

Linear regression models of the word accuracy (%) with the six presentation attributes

Speaker-independent recognition

$$Acc = 0.12AL - 0.88SR - 0.020PP - 2.2OR + 0.32FR - 3.0RR + 95$$

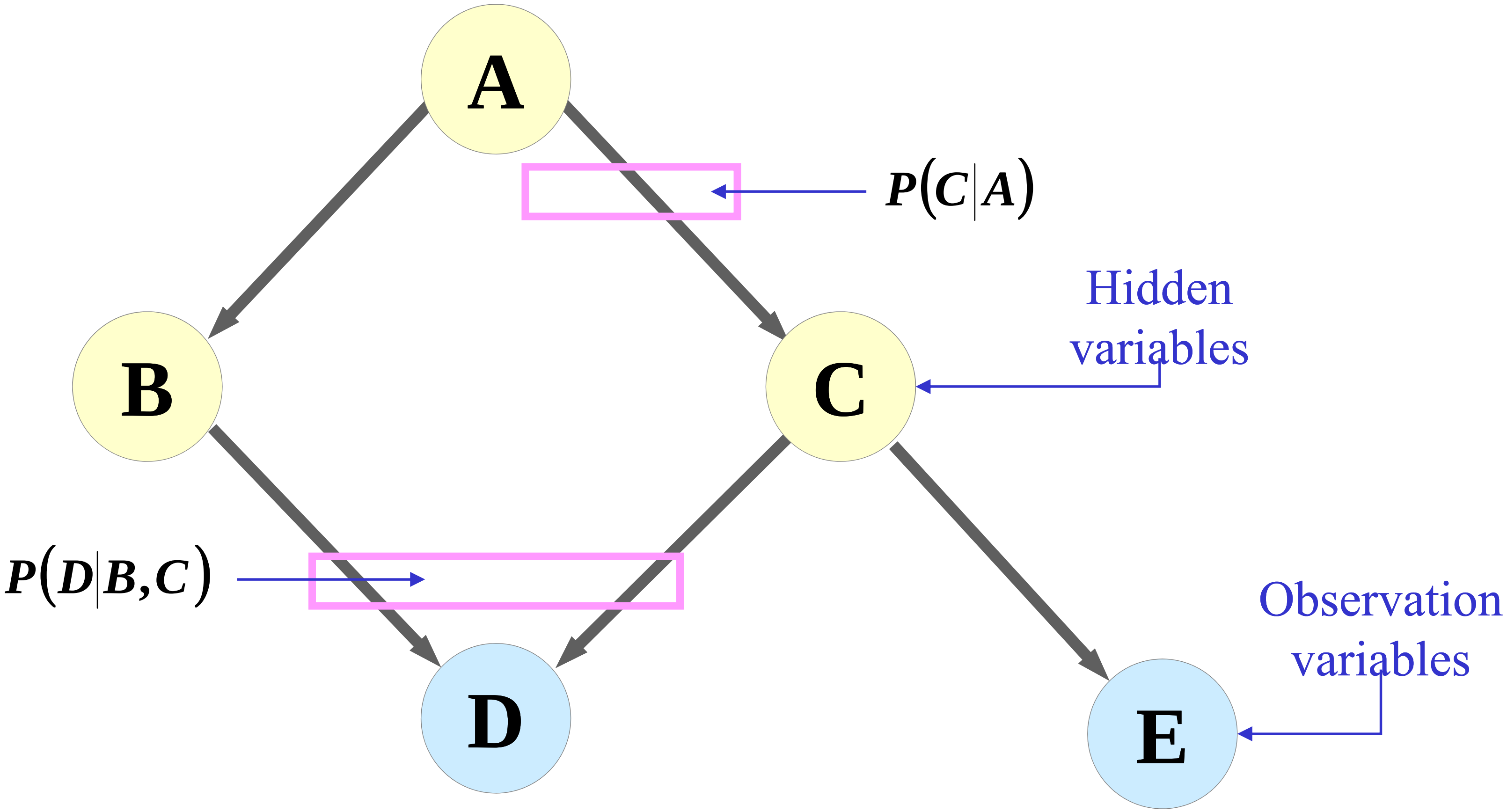
Speaker-adaptive recognition

$$Acc = 0.024AL - 1.3SR - 0.014PP - 2.1OR + 0.32FR - 3.2RR + 99$$

Acc: word accuracy, SR: speaking rate,
PP: word perplexity, OR: out of vocabulary rate,
FR: filled pause rate, RR: repair rate

A Bayesian network with five variables

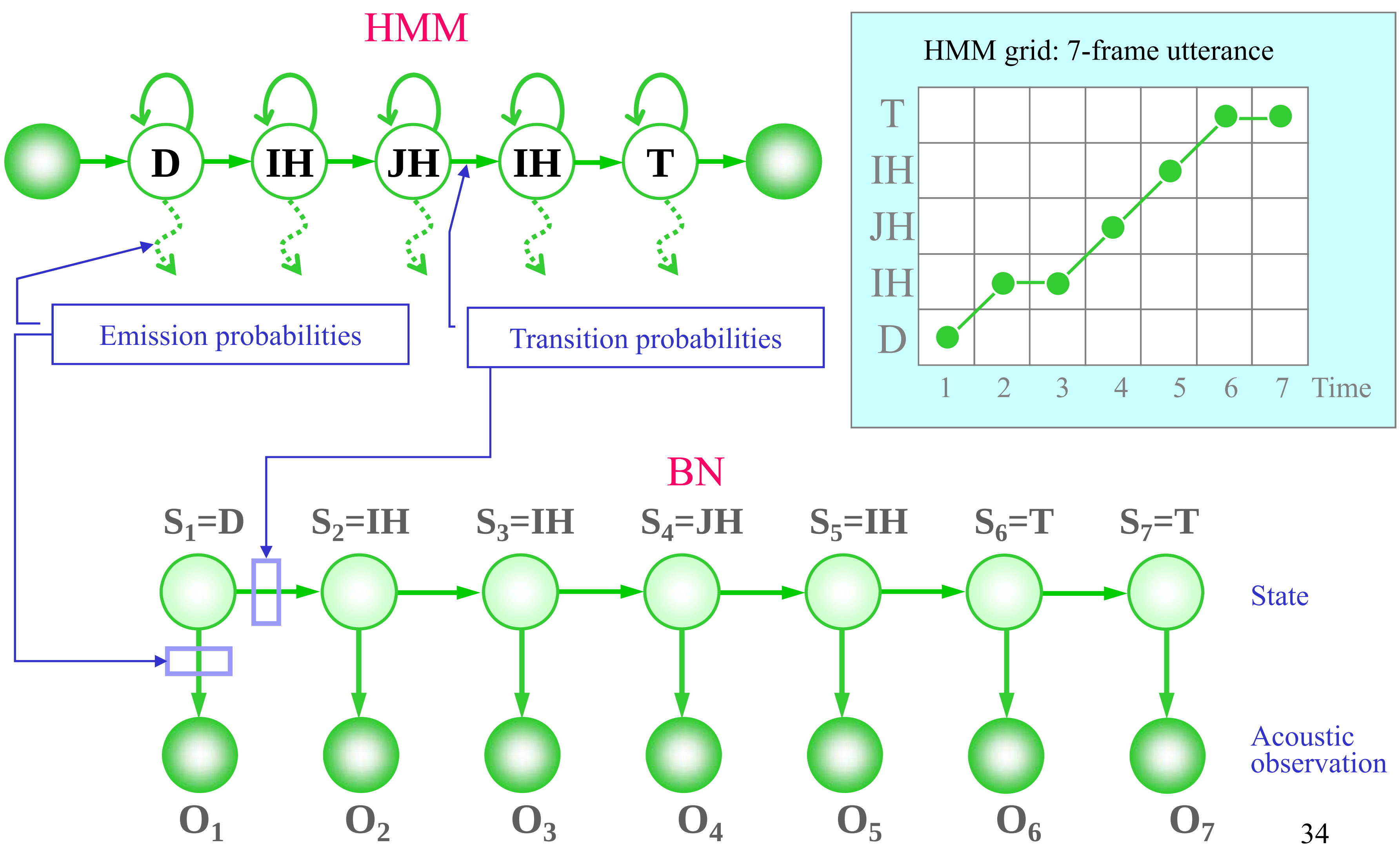
Variables with known values are shaded. Conditional probability functions (indicated by boxes) are associated with each variable and used to return numerical values for conditional probabilities.



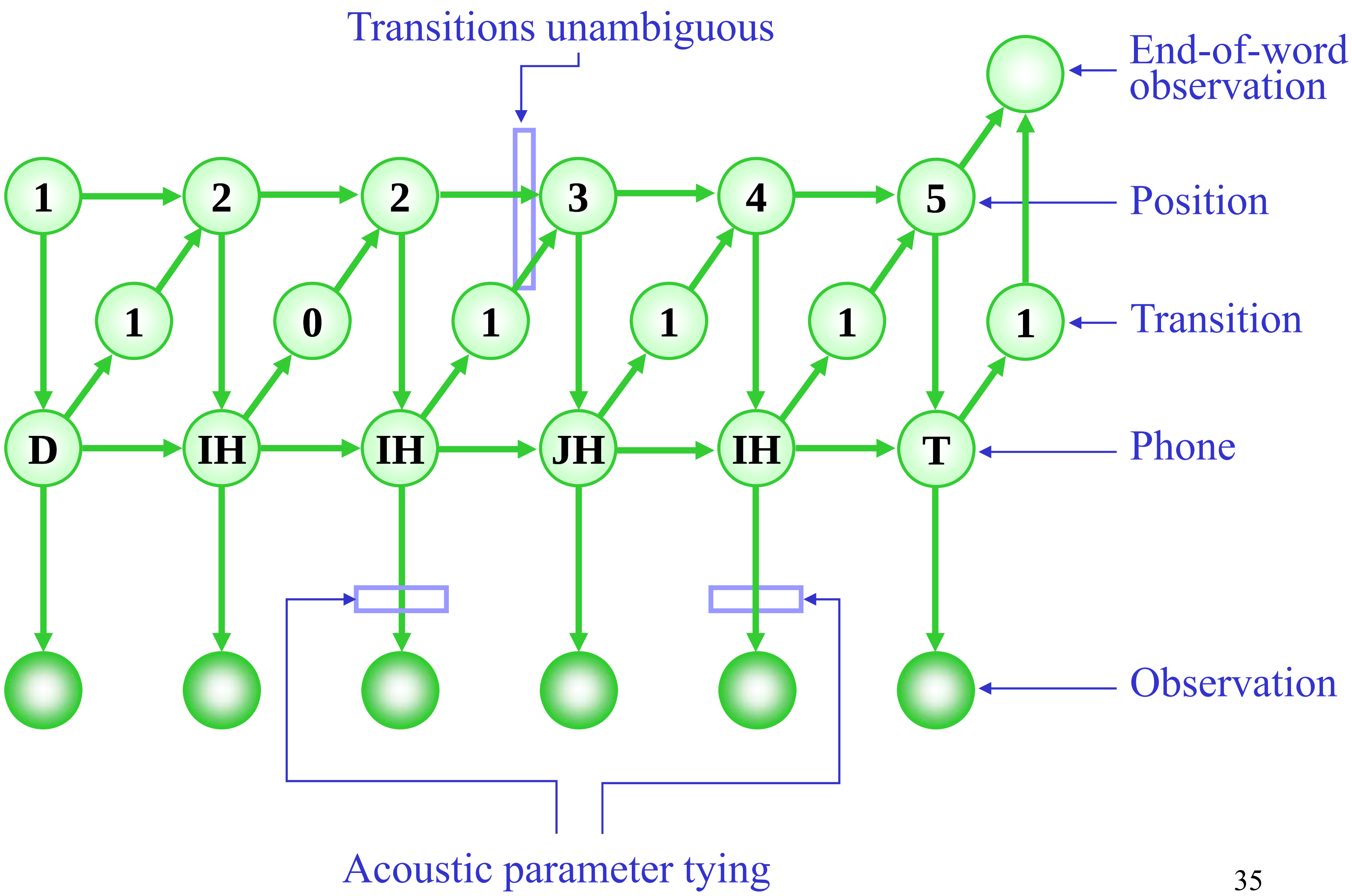
Joint distribution: $P(A,B,C,D,E) = P(A)P(B|A)P(C|A)P(D|B,C)P(E|C)$

Comparison of HMM and Bayesian network

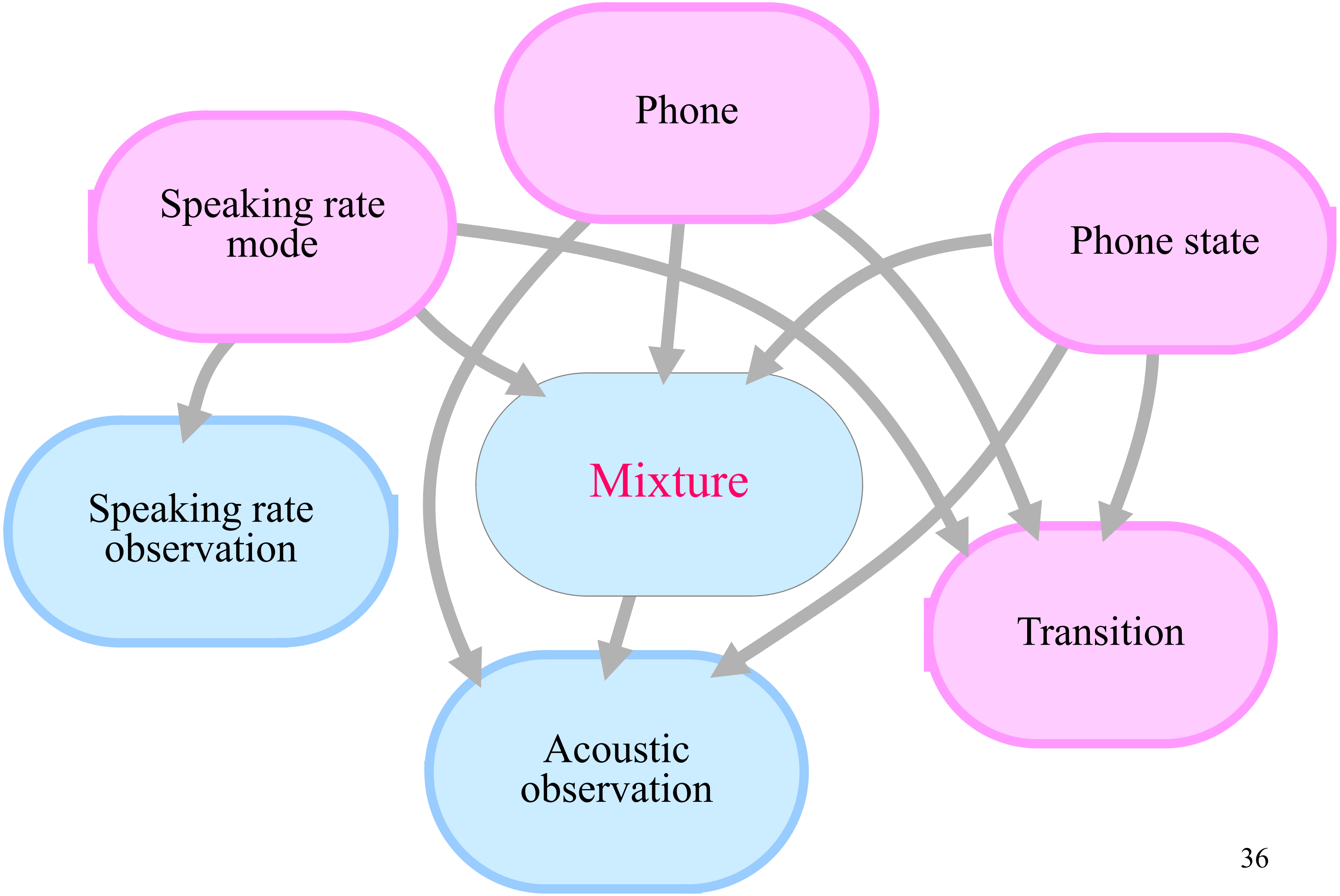
The dashed lines in the HMM represent the acoustic emissions that occur at each time frame



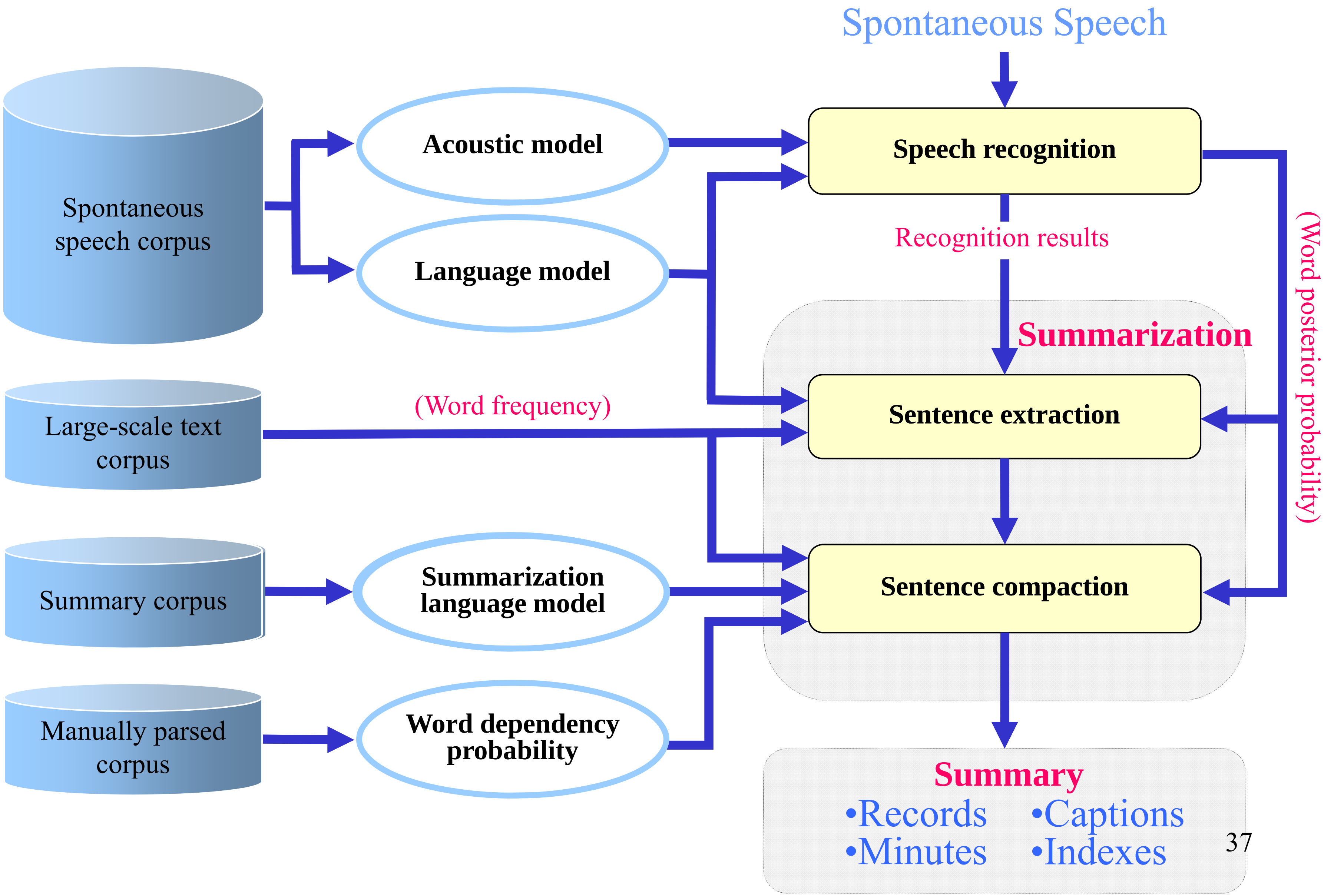
A Bayesian network representation of a typical speech recognition HMM



Bayesian network representation of HMM incorporating speaking rate variations



Automatic summarization based on the combination of important sentence extraction and sentence compaction



Modality-oriented classification of multimodal systems

Input modalities

- **Speech**
  - Continuous sp.
  - Discrete sp.
  - Isolated words
  - Spelled words
- **Gesture**
  - 2D
  - 3D
- **Handwriting**
  - Cursive script
  - Printed script
  - Isolated digits/ characters
- **Facial features**
  - Eye movement/ gaze
  - Lip movement
- **Conventional input**
  - Keyboard
  - Pointing device
  - Touch screen

Output modalities

- **Speech**
  - audio
  - visual
  - tactile
  - Talking face
- **Text**
  - Printed handwriting
- **Graphics**
  - Images video
- **Non-speech audio**
  - Sound clips
  - Music
  - Earcons
- **Haptic/ tactile**
  - Pneumatic
  - Vibrotactile
  - Electrotactile
  - Neuromuscular

Task-oriented taxonomy of multimodal applications

- **Interactive**
  - **Mediation of human-human interaction**
    - Collaborative work Translation Teleconferencing
  - **Human-computer interaction**
    - **Entertainment**
      - Games Animations Interactive TV
    - **Command & control**
      - Robot control Voice operation Security-access
    - **Translations & queries**
      - Call centers Electronic commerce
    - **Data entry & manipulation**
      - **Simple data entry**
        - Form filling PDAs
      - **Text & multimedia**
        - Dictation Programming Web authoring UI design tool User-based indexing
    - **Other**
      - Tutoring/education Smart rooms Virtual reality Wearable computing
- **Non-interactive**
  - Transcription
  - Indexing of multimedia data

Taxonomy of system-level evaluation techniques

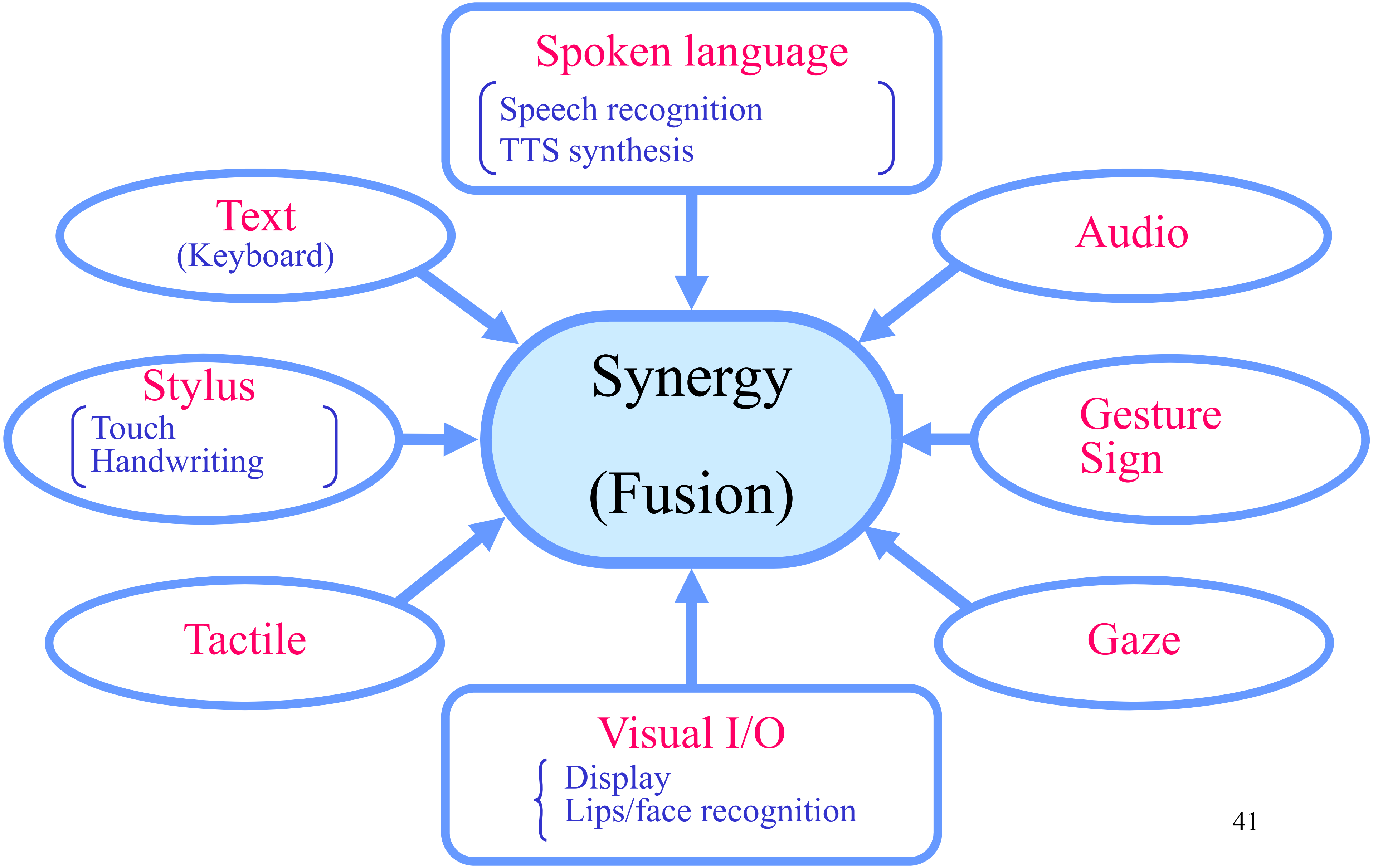
Evaluation techniques

- **Predictive models**
- **Experimental techniques**
- **Expert evaluations**

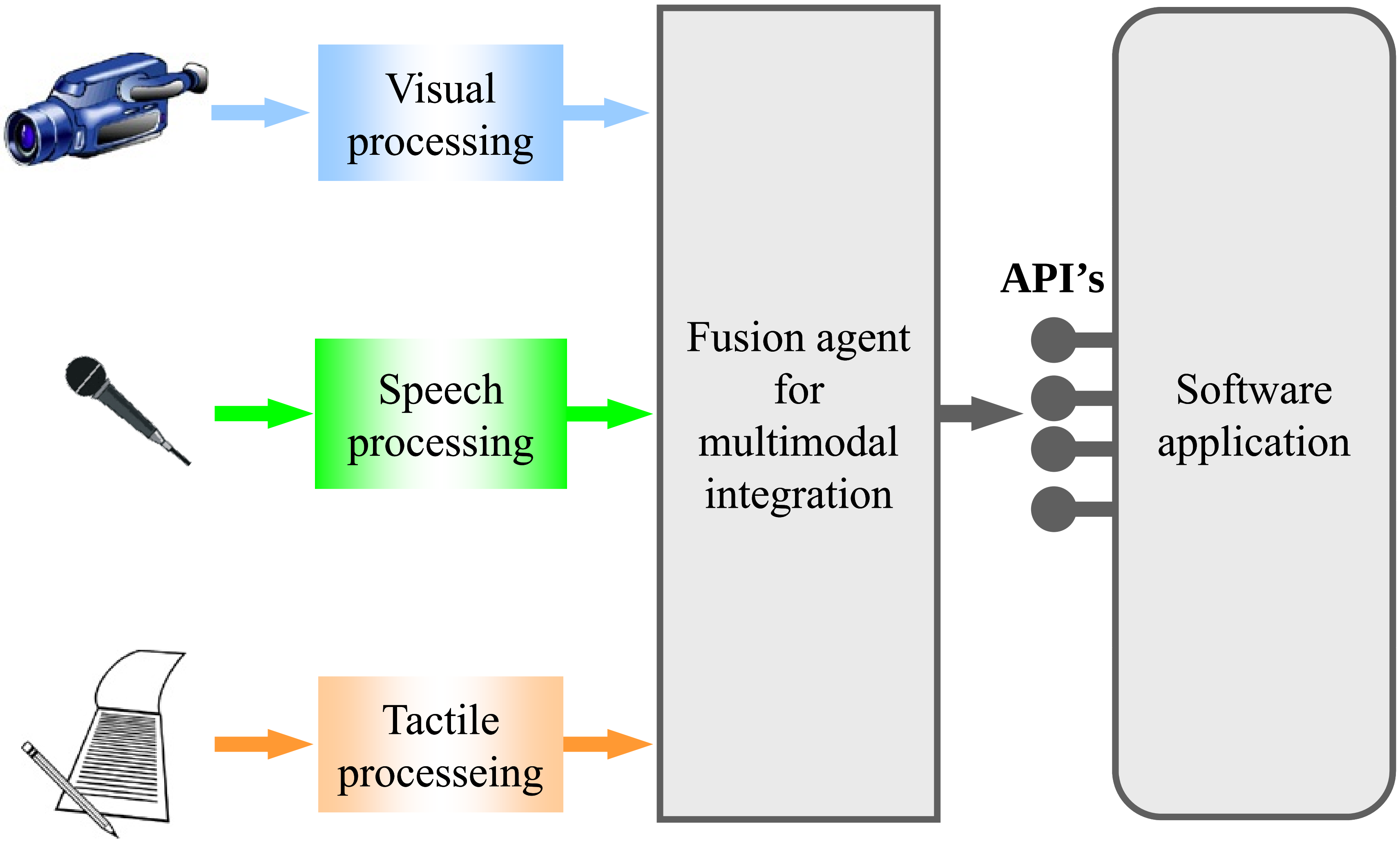
Empirical model
Theory-based model

Benchmark evaluation User study (Prototype)
Simulation (WoZ)
Interactive design

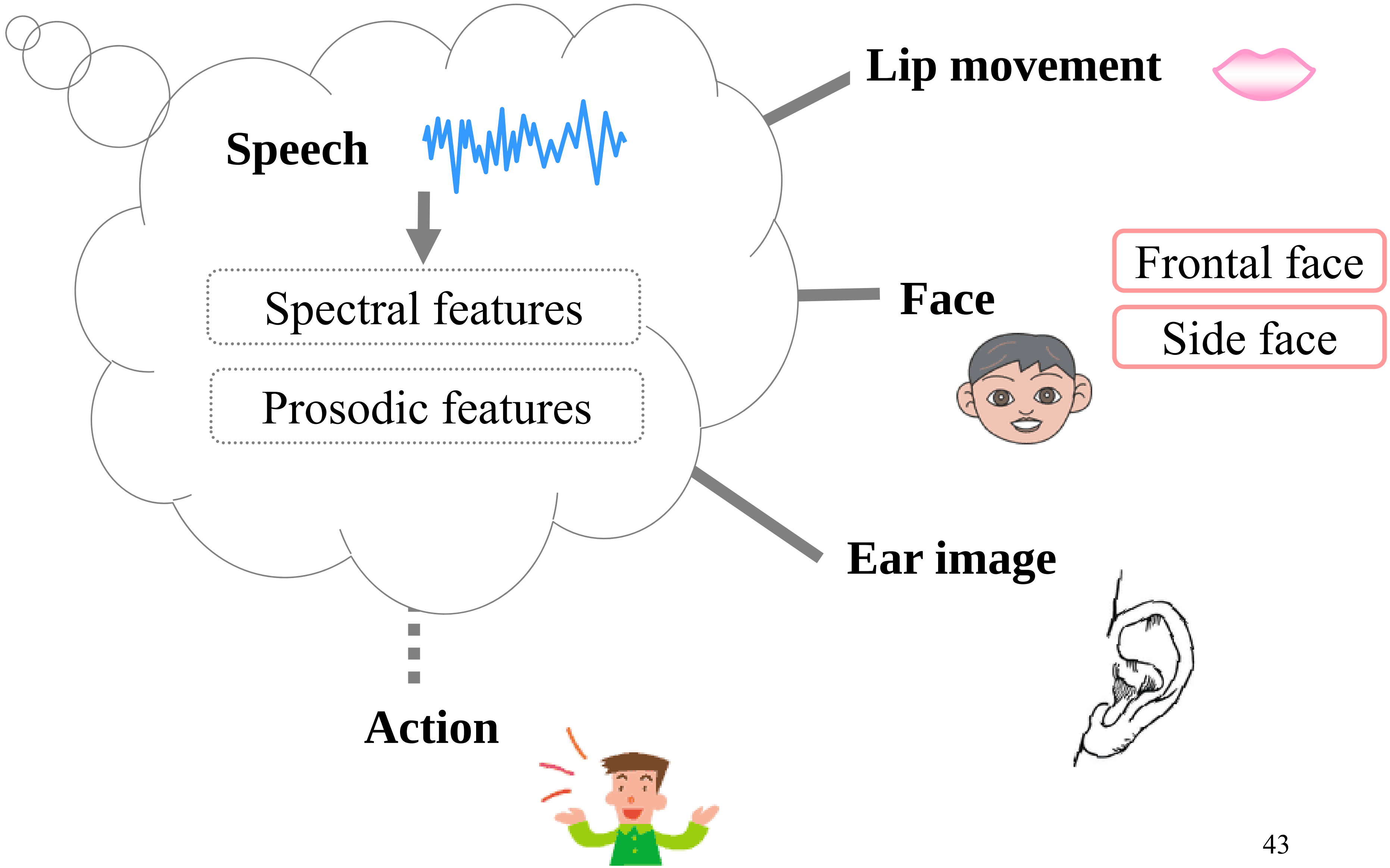
Multimodal human-machine communication (HMC)



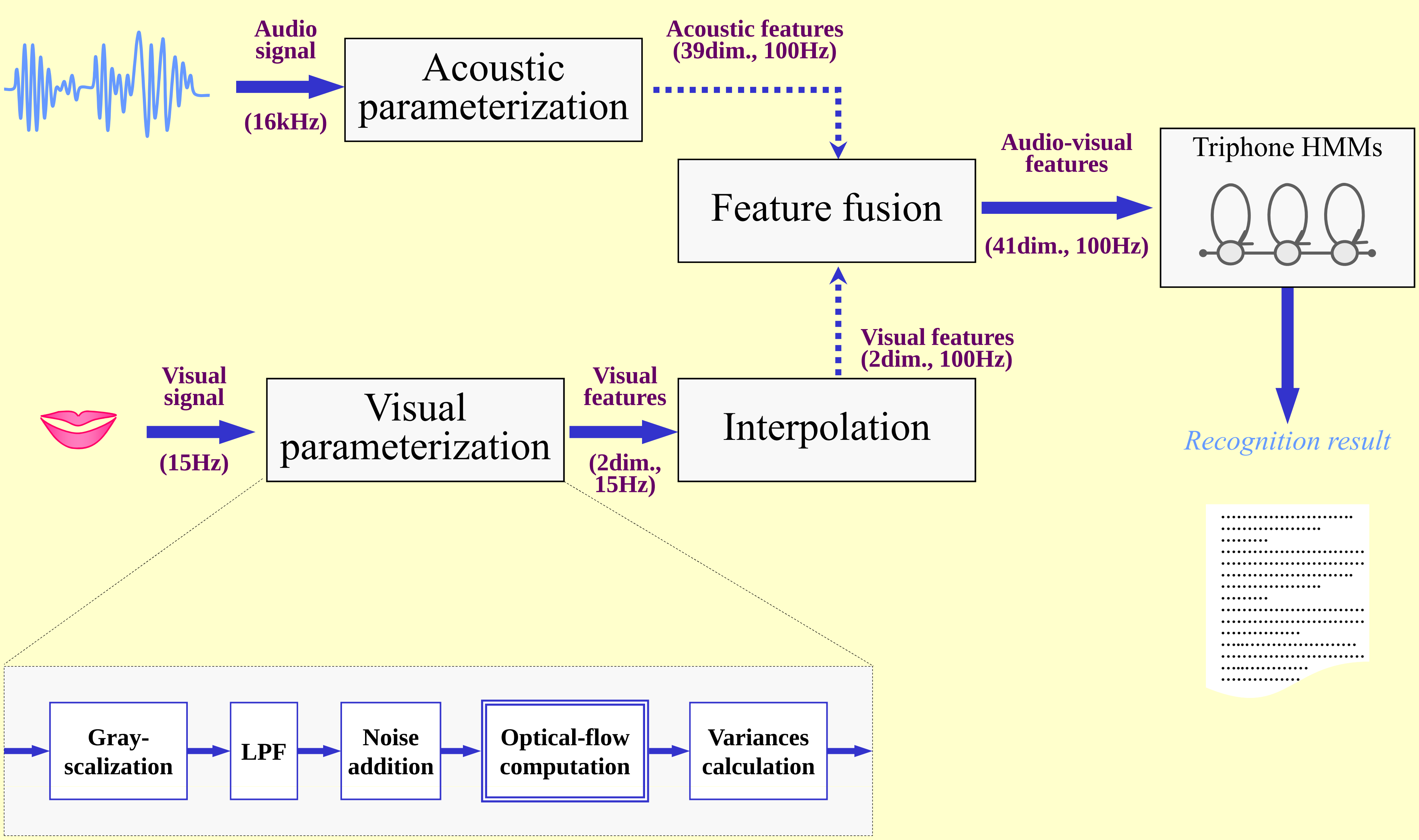
Architecture of multimodal human/computer interaction



Multimodal speech and speaker recognition



Audio-visual speech recognition system using optical-flow analysis



Audio-only and audio-visual connected digit recognition accuracy

SNR(dB)	Audio-only	Audio-visual (λ_a)
5	39.50%	44.95% (0.86)
10	58.55%	70.78% (0.90)
20	94.66%	94.51% (0.94)
∞	97.59%	97.96% (0.96)

(λ_a: optimum audio-weighting factor)

Recognition results using frontal-face images

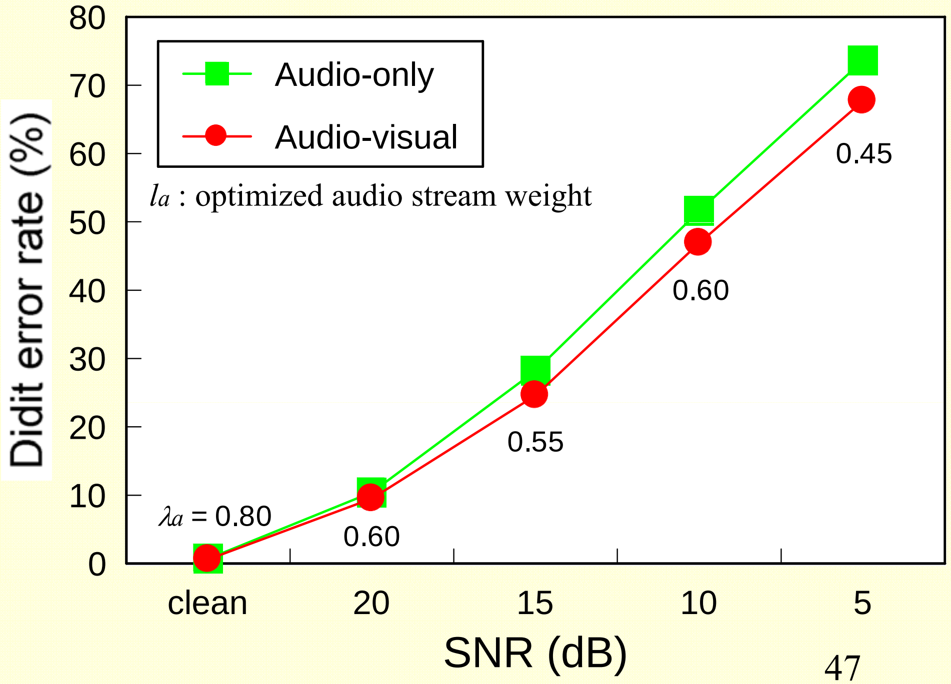
- **Training data**
  - 11 male speakers
  - Clean conditions
  - Strings of 2 to 6 connected digits in Japanese
    - e.g. "24", "899", "3723", …, "500241", …
  - 250 strings per speaker

- **Testing data**
  - 6 male speakers
  - Real car environment (10-15 dB SNR)
- **Visual parameters**
  - Maximum and minimum values of the integral of optical-flow vectors
- **Stream weight is optimized**
- **Digit error rate:**

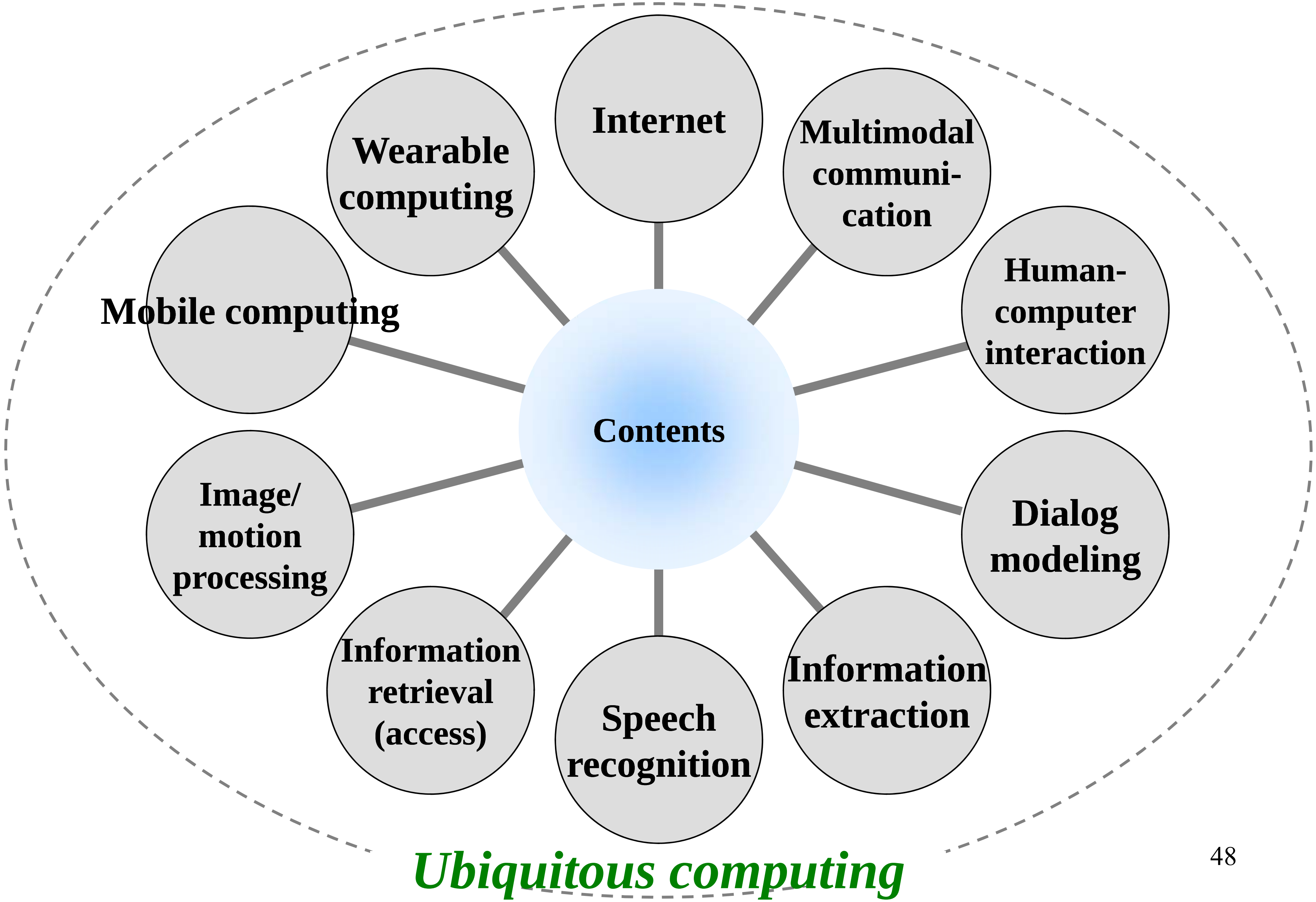
Audio-only	Audio-visual
41.5%	36.2%

Recognition results using side-face images

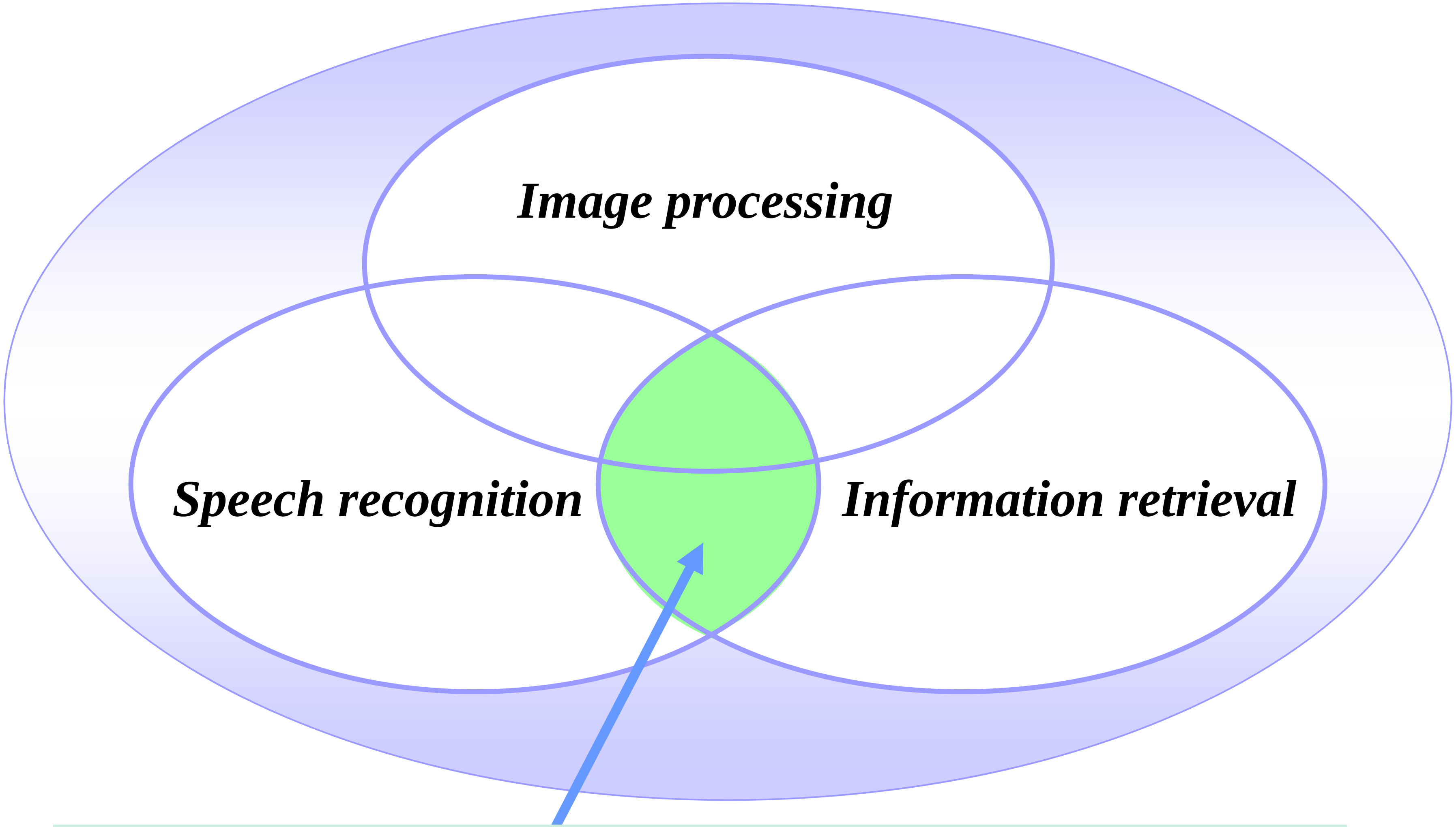
- **Database**
  - 38 male speakers
  - Clean conditions
  - Strings of 4 connected digits in Japanese
  - 50 strings per speaker
- **Testing data**
  - 19 male speakers
  - Contaminated with white noise
- **Visual parameters**
  - Horizontal and vertical variances of the flow vector components
- **Stream weight**
  - Optimized at each SNR condition



Multimedia contents technology



Multimedia information (Broadcast news, etc.)



Information extraction and retrieval of spoken language content
(Spoken document retrieval, information indexing, story segmentation, topic tracking, topic detection, etc.)

Summary

- **Two major speech recognition applications are *conversational systems for accessing information services* and *systems for transcribing, understanding and summarizing ubiquitous speech documents*.**
- **How to cope with additive *noise* and intra- and inter-*speaker variability* and how to model and recognize *spontaneous speech* are the most important issues.**
- **Speech recognition is a *search process* in a *super-high-dimensional non-linear space*.**
- **Construction of a large-scale *spontaneous speech corpus* is crucial.**
- ***Multimodal human-computer communication* and *information extraction* has a bright future.**