Selected topics from 40 years of research on speech and speaker recognition

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

Generations of ASR technology

1950 1960 1970 1980 1990 2000 2010

1952 **1G** 1968

Heuristic approaches
(analog filter bank + logic circuits)

1968 **2G** 1980

Pattern matching
(LPC, FFT, DTW)

1980 **3G** 1990

Statistical framework
(HMM, n-gram, neural net)

1990 **3.5G**

Discriminative approaches, robust training, normalization, adaptation, spontaneous speech, rich transcription

? **4G**

Extended knowledge processing

Prehistory

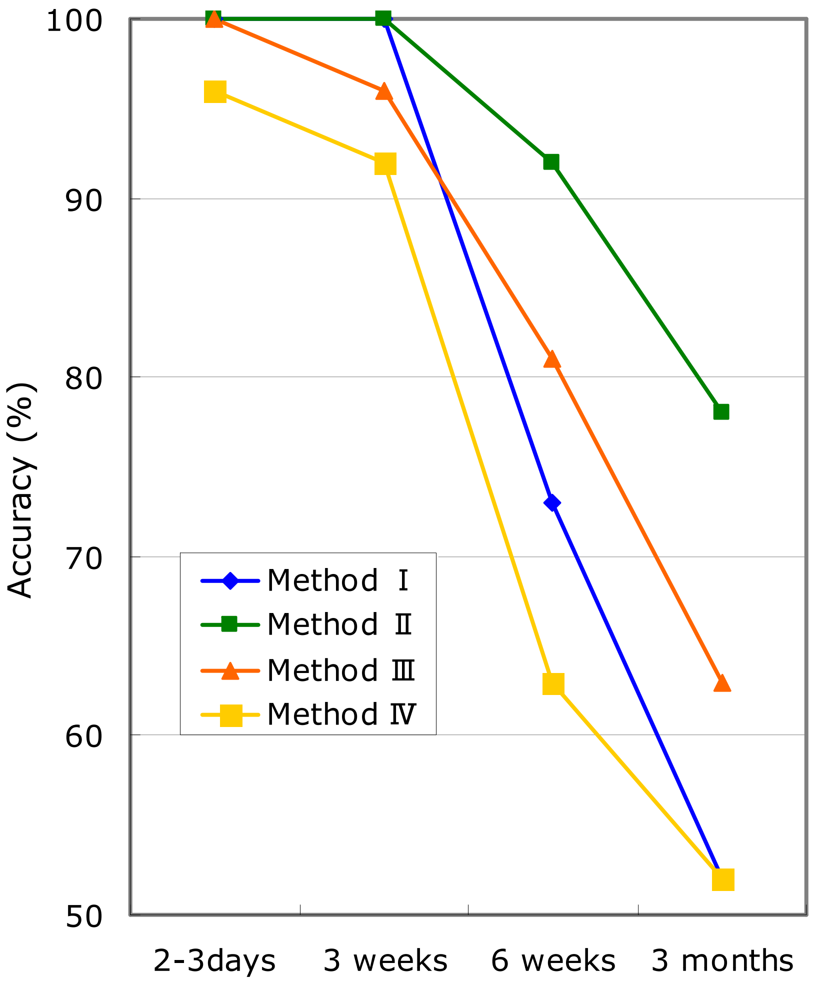
Our research
NTT Labs (+Bell Labs), Tokyo Tech
Collaboration with other labs

Japanese traditional cuisine “Kaiseki-ryori”

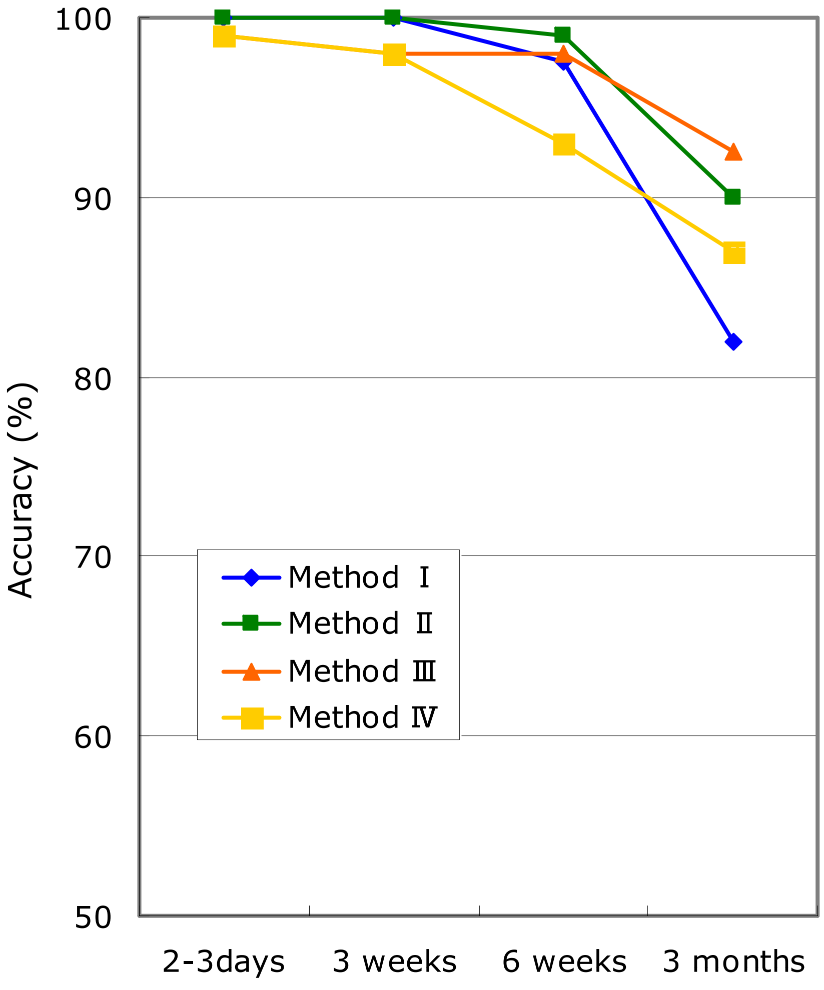
1970s

- Speaker recognition by statistical features
- Speaker recognition by cepstral features

Speaker recognition by long-term averaged spectrum

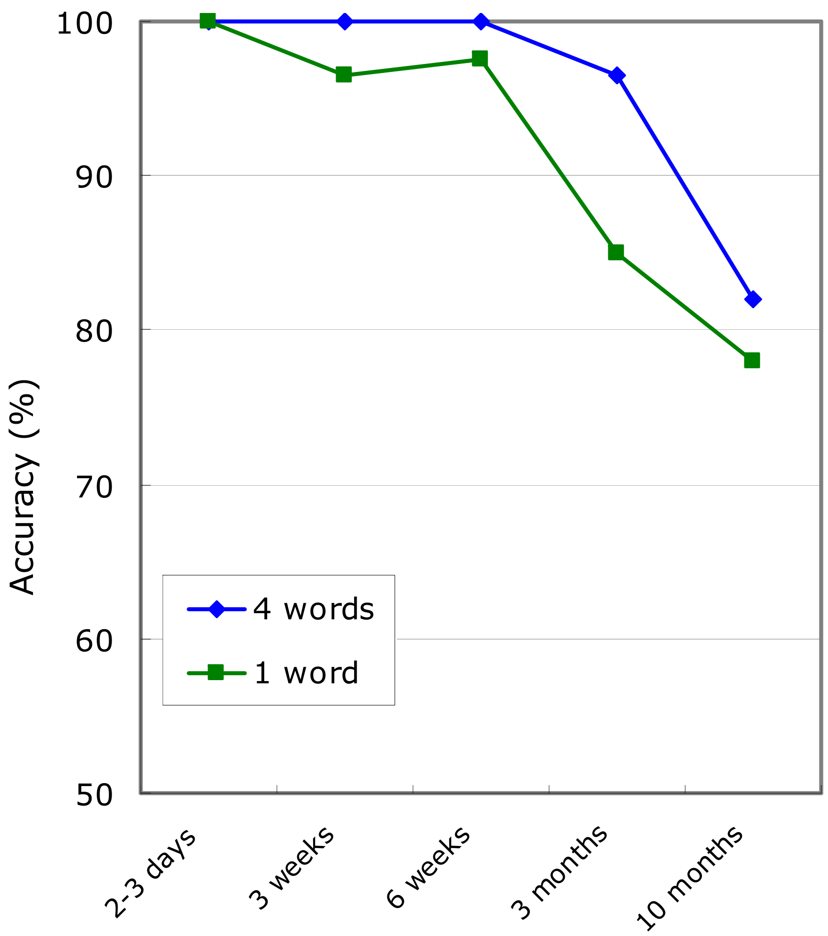


(a) Speaker identification

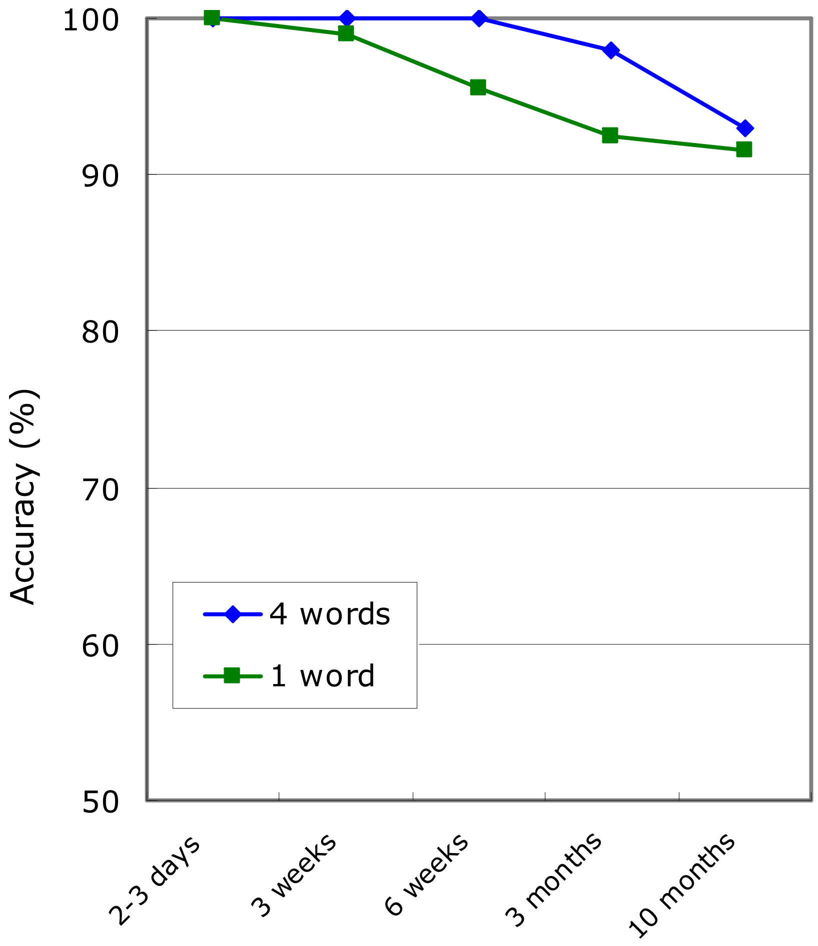


(b) Speaker verification

Speaker recognition by using LPC features



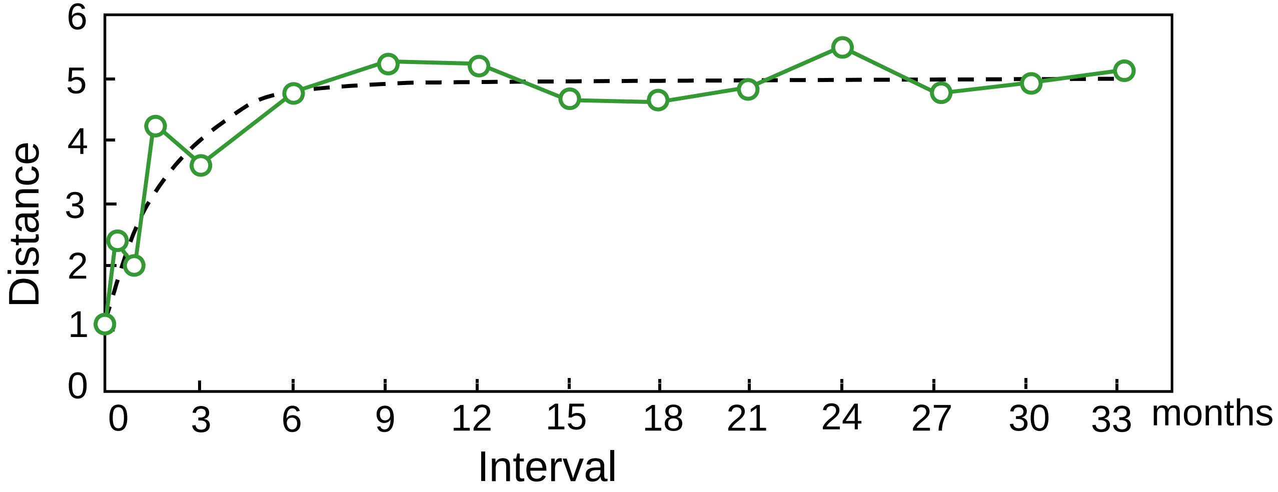
(a) Speaker identification



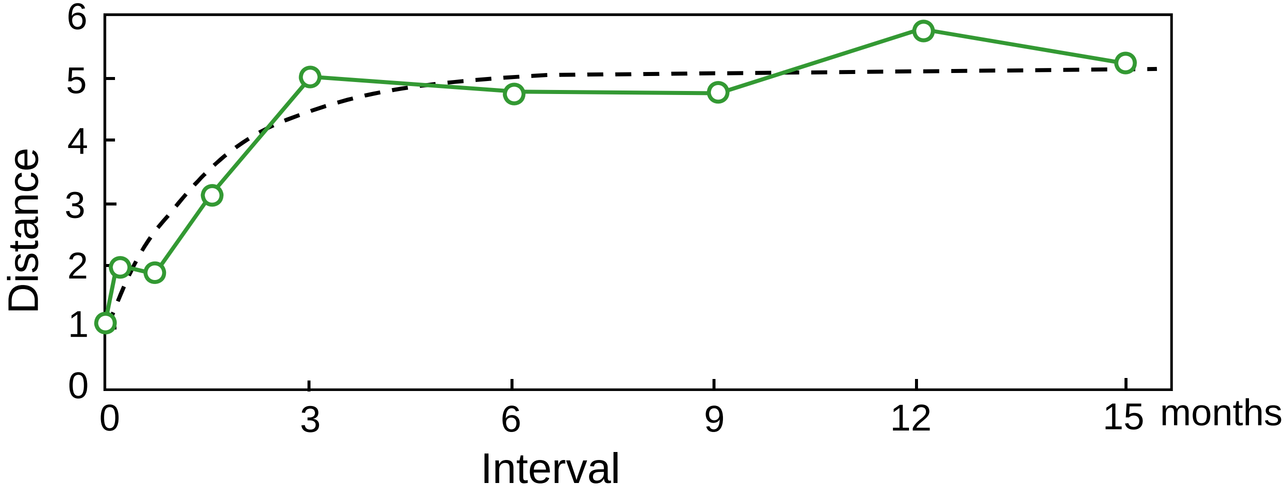
(b) Speaker verification

The amount of spectral variation as a function of time interval

Log-area-ratio parameters



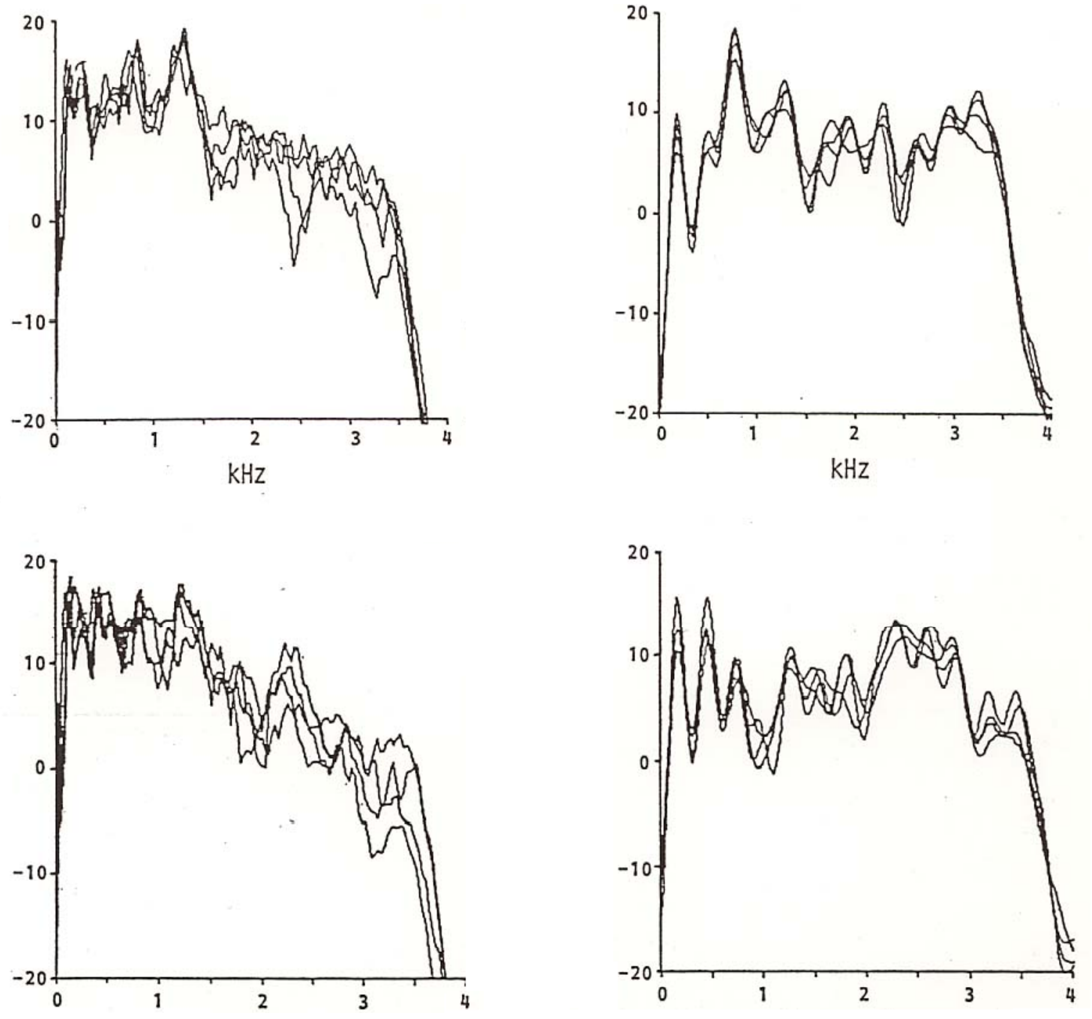
Long-term averaged spectrum



Variation of the long-time averaged spectrum from four sessions over eight months, and corresponding spectral envelopes derived from cepstrum coefficients weighted by the square root of inverse variances

Sub. W

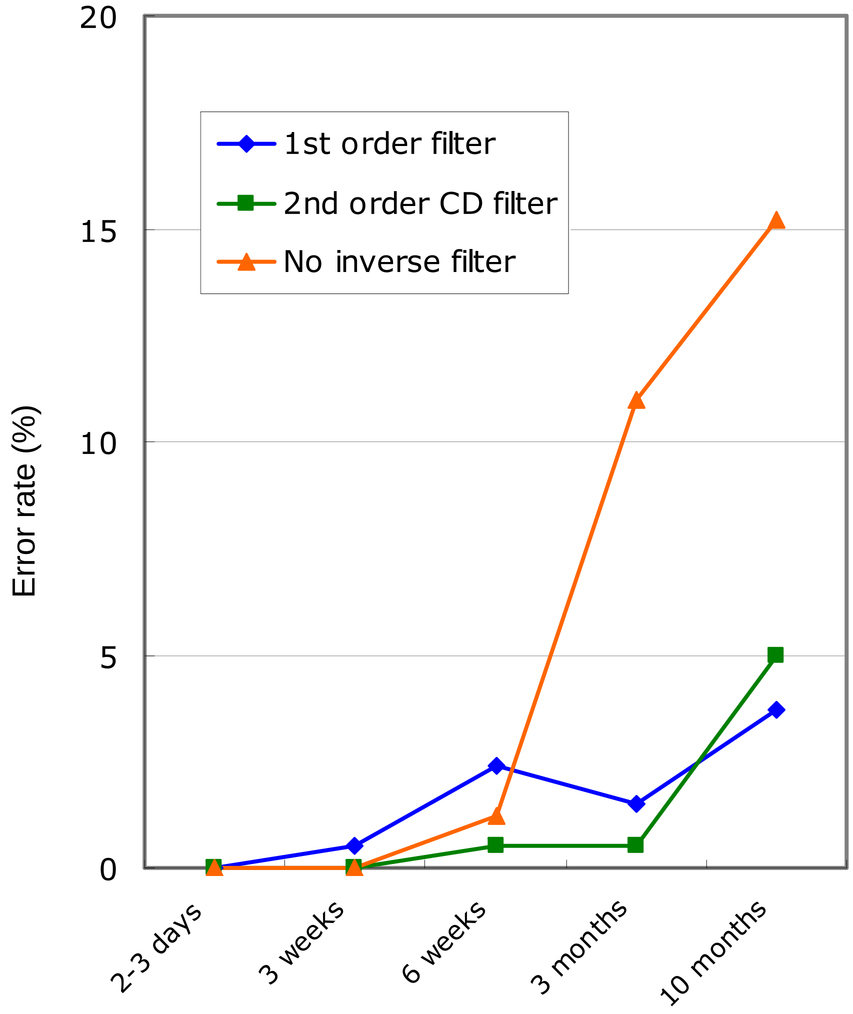
Sub. T



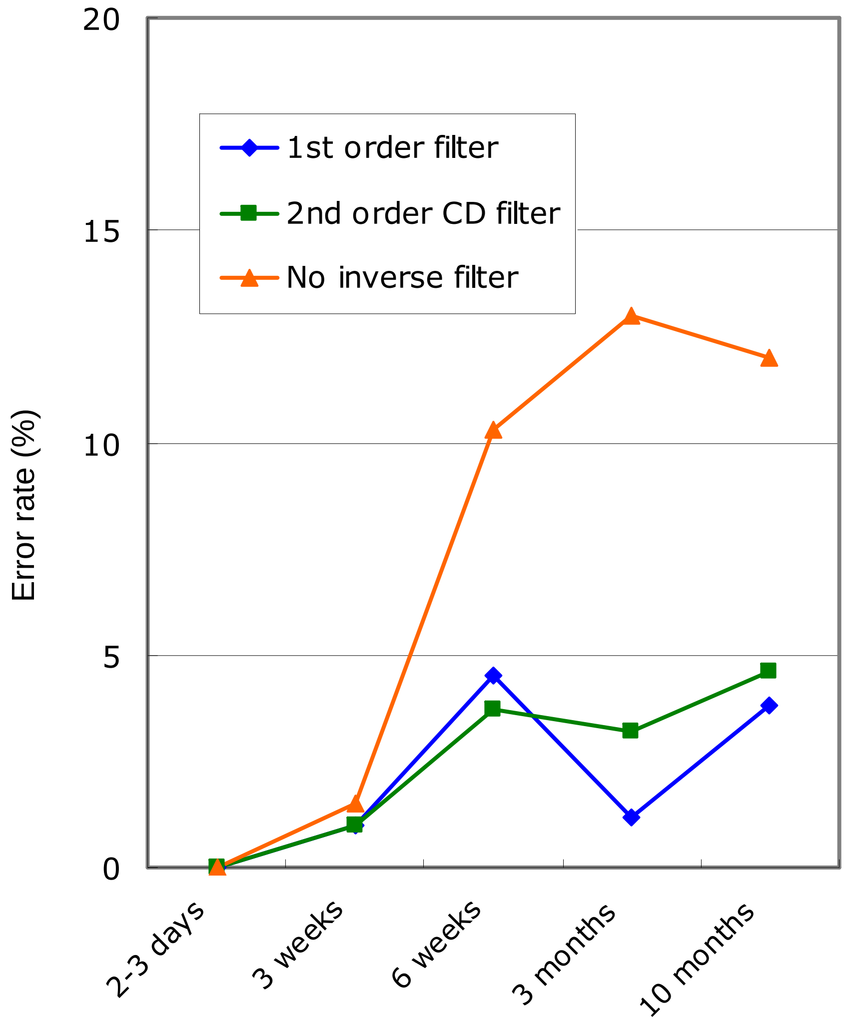
(a) Long-time averaged spectra

(b) Envelopes by weighted cepstrum

Speaker recognition by using LPC features (Effectiveness of inverse filtering)

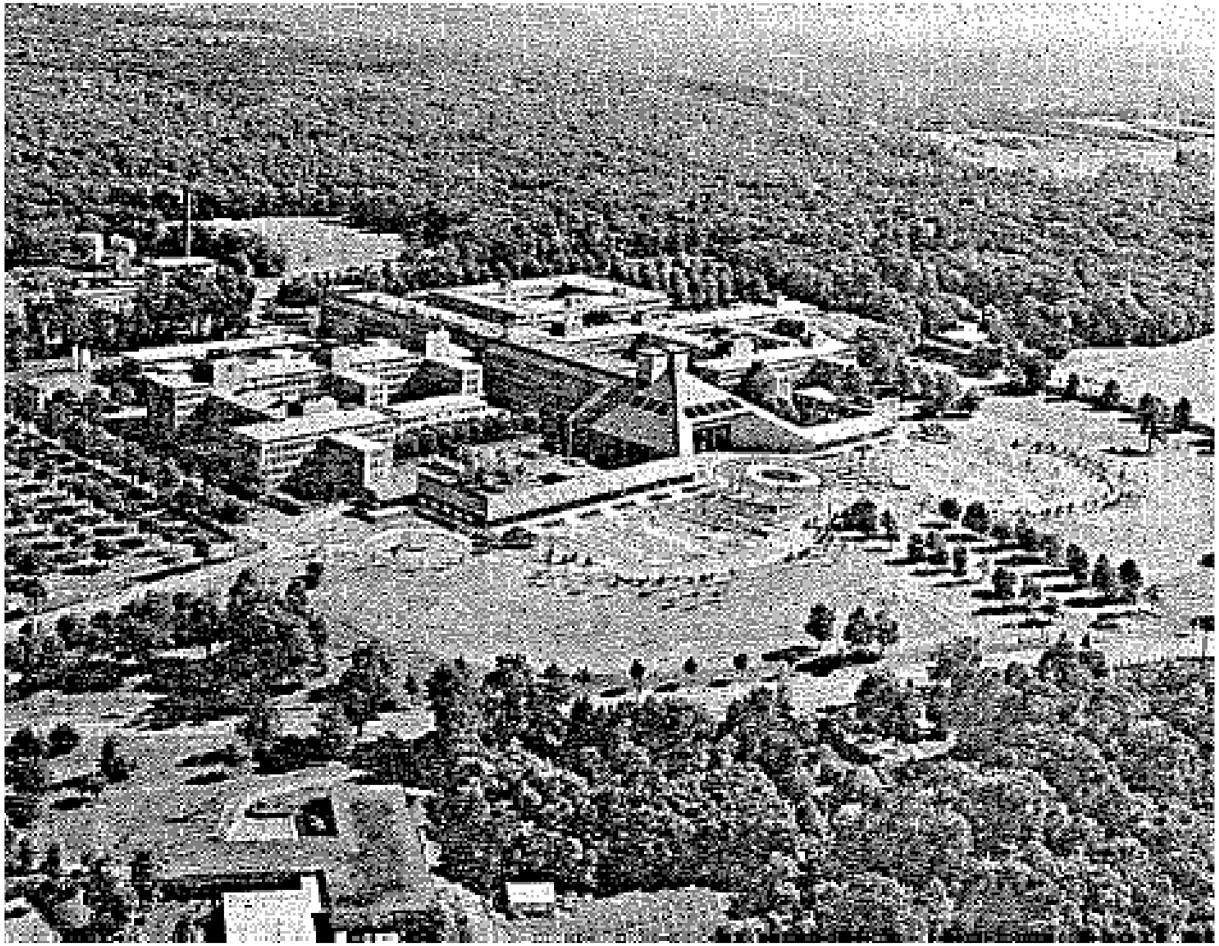


(a) Two words

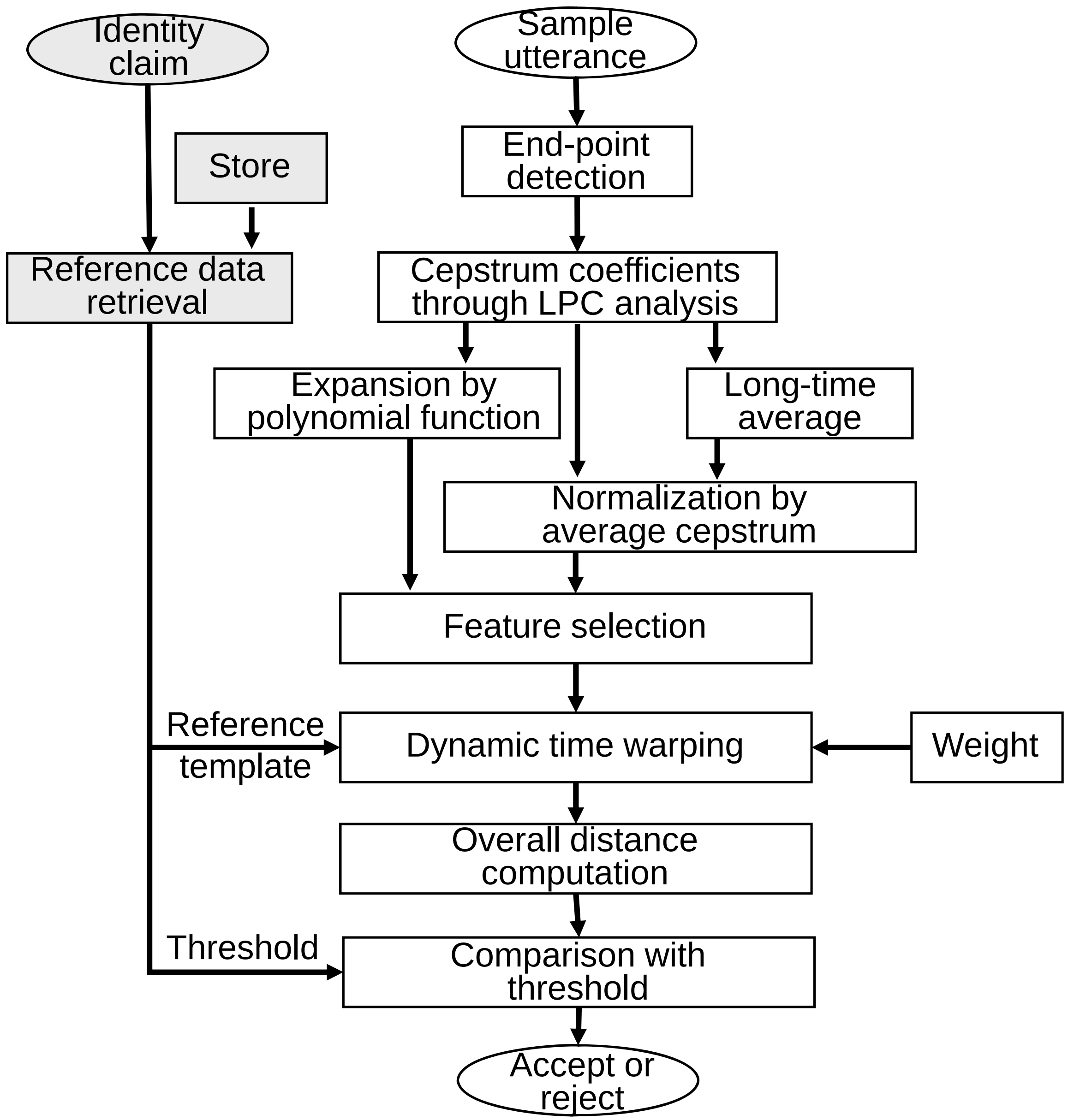


(b) One word

Research at Bell Laboratories, Murray Hill, from 1978 to 1979



Speaker verification using cepstrum features



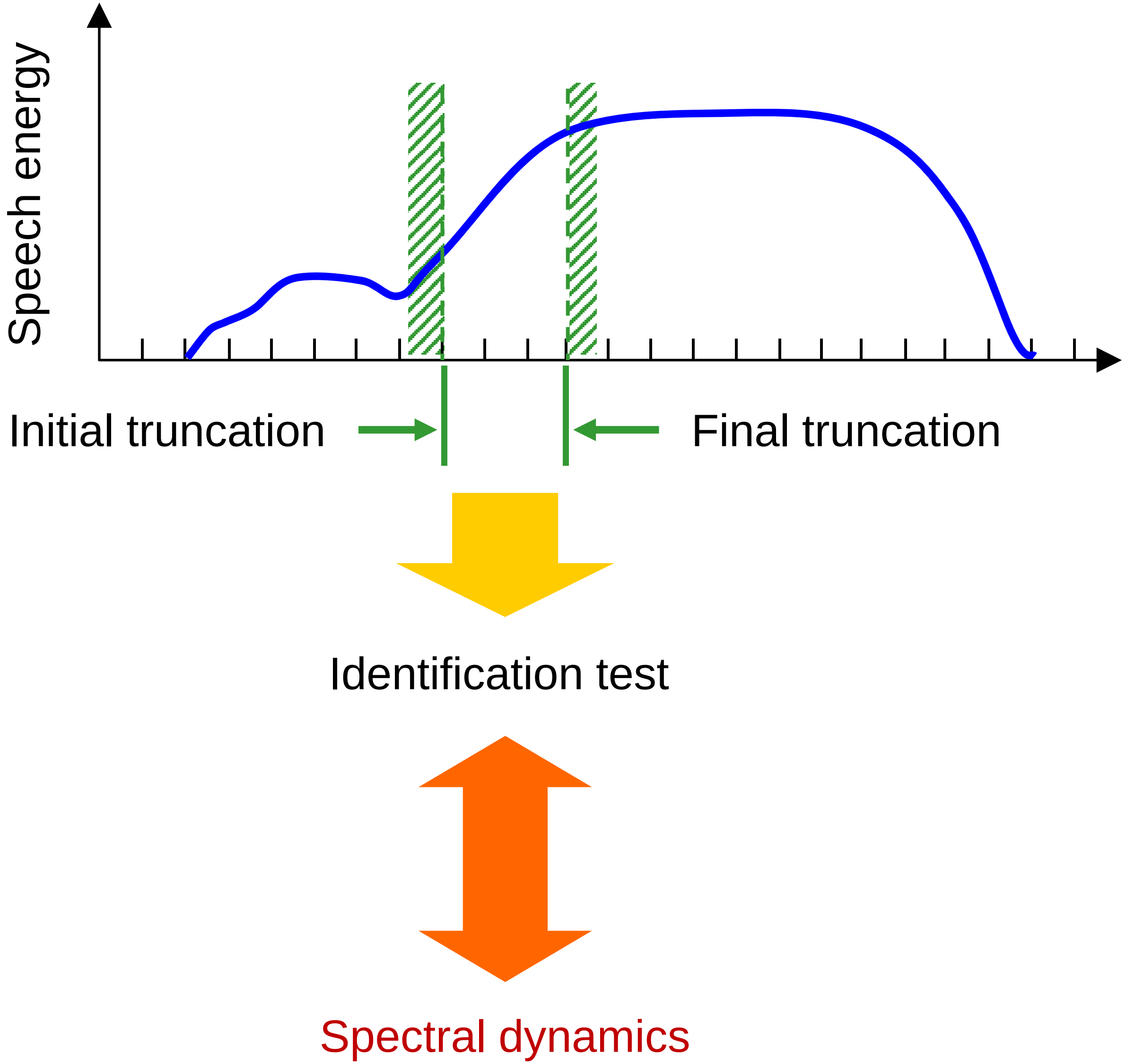
On-line speaker verification experiments using 120 Bell Labs employees

User: "We were away a year ago."
System: "Stand by for analysis."
System: "Your identity has been verified. Thank you."

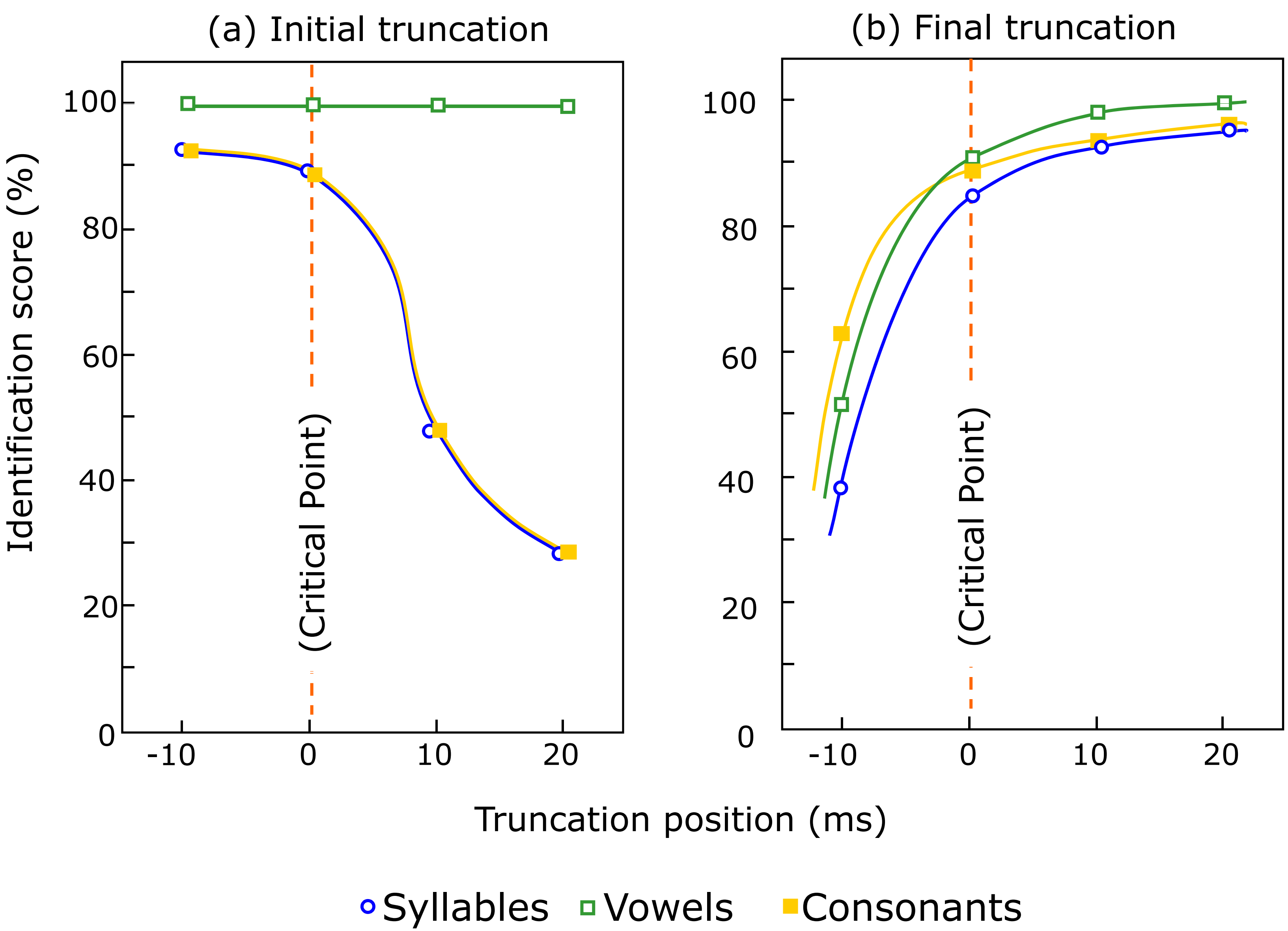
1980s

- Spectral dynamics in speech perception and recognition
- Speaker recognition by HMM/GMM

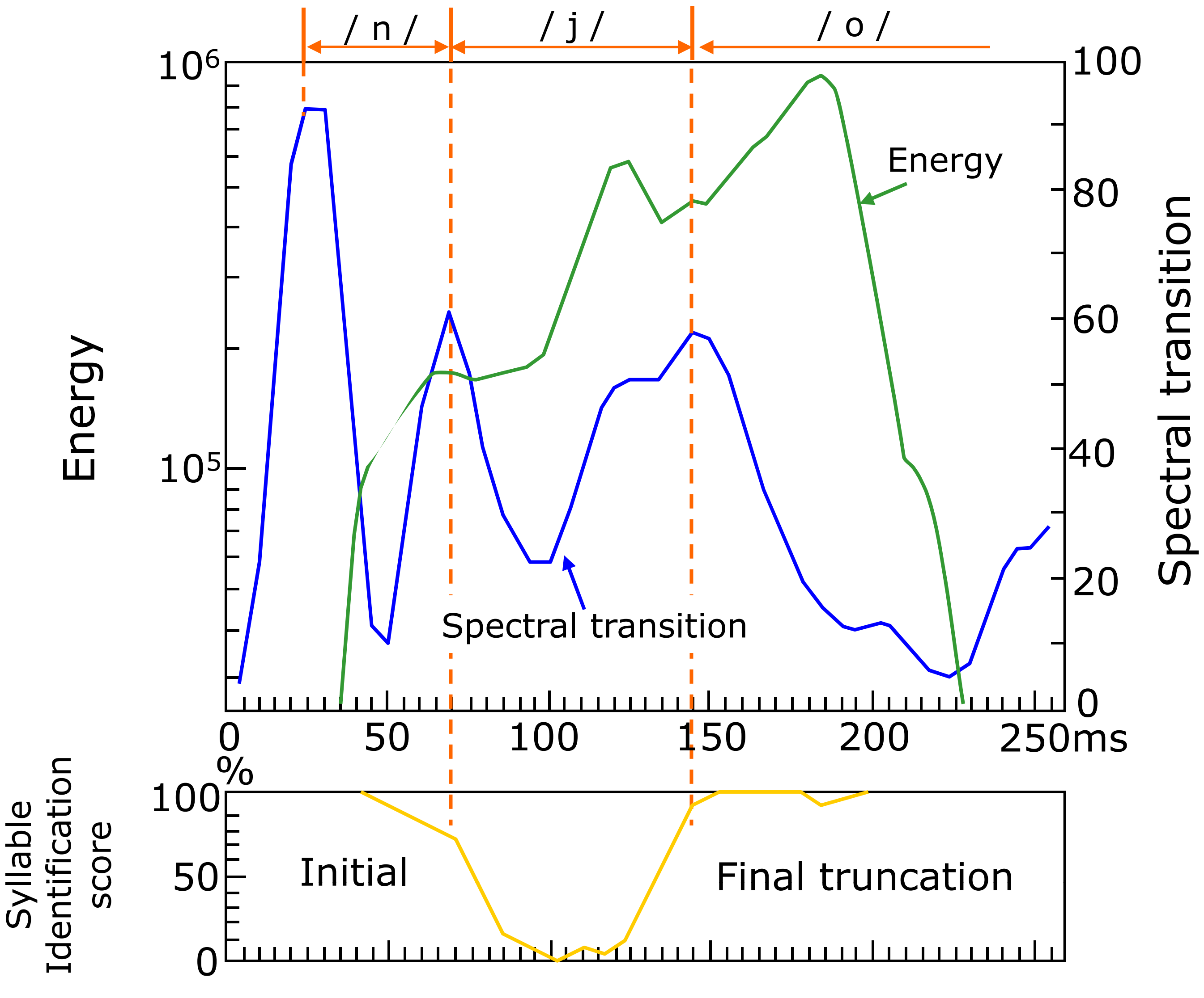
Analysis of relationships between spectral dynamics and syllable perception



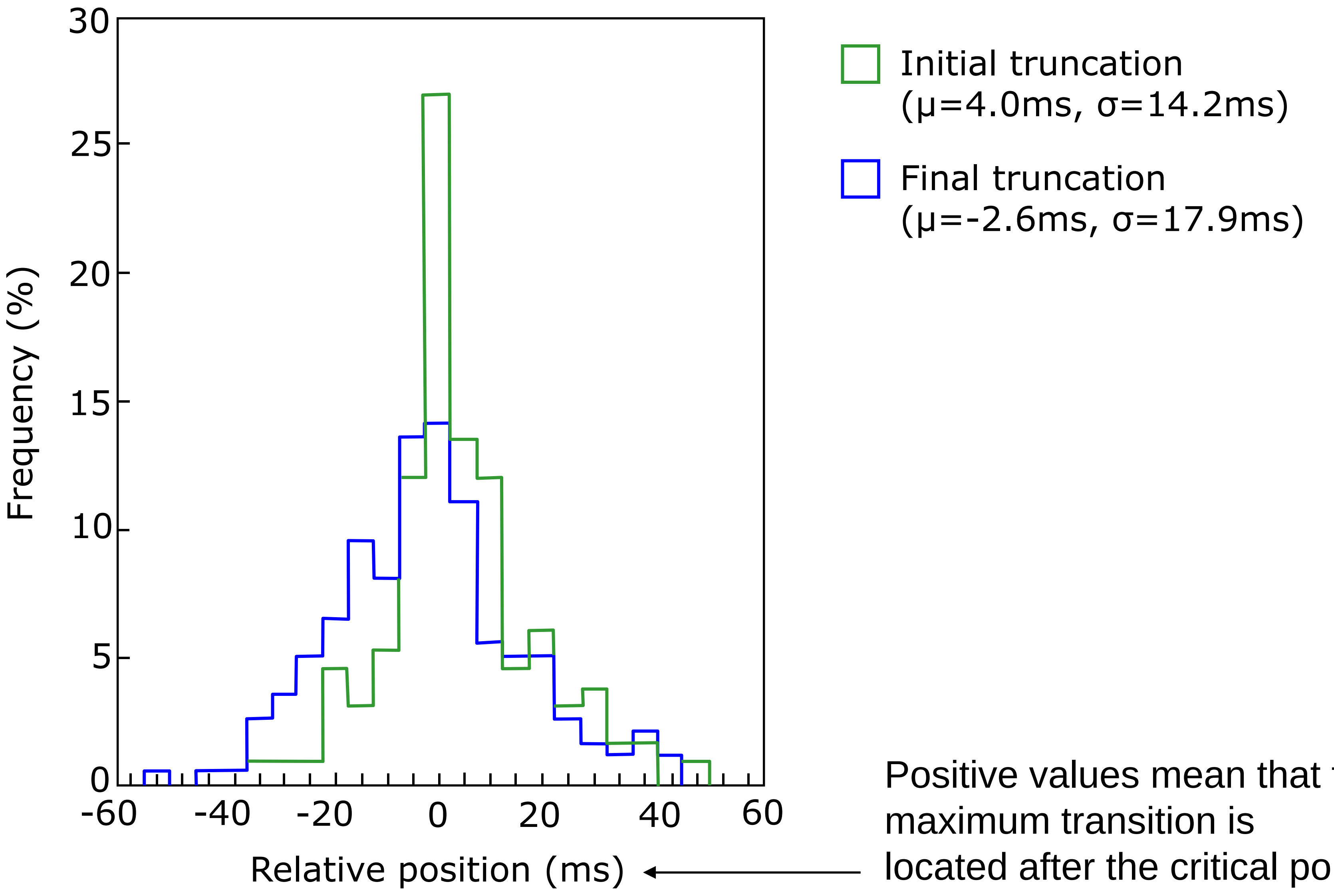
Relationship between truncated CV syllable identification scores and truncation position relative to the perceptual critical point



Relationship between spectral transition and syllable identification scores as a function of the truncation position for the syllable /njo/

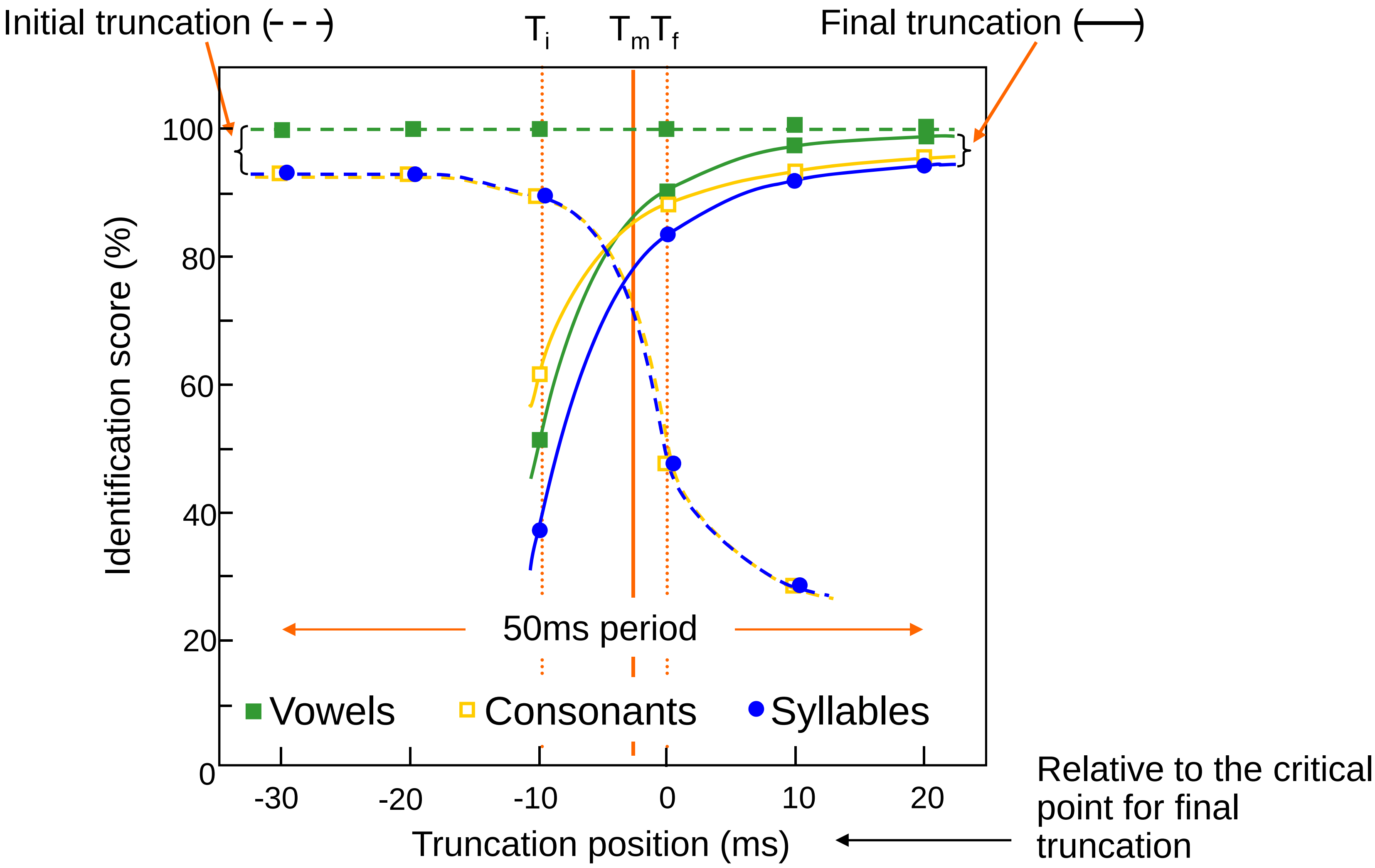


Distribution of the difference between the perceptual critical point and the maximum spectral transition position for all 100 syllables



Positive values mean that the maximum transition is located after the critical point

Relationship between truncation position and identification scores for the truncated CV syllables



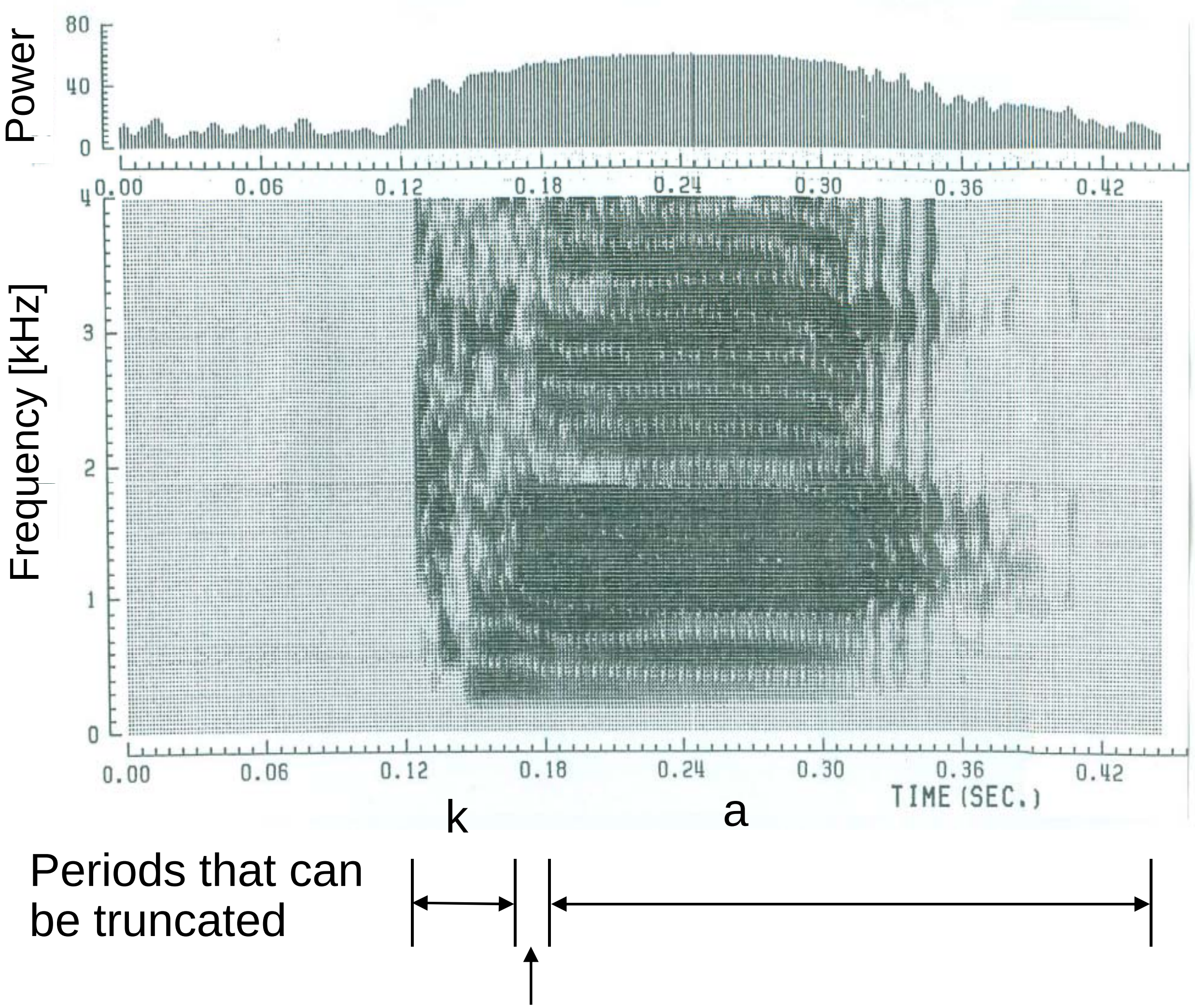
T_i, T_f : Perceptual critical point for initial & final truncation

T_m : Maximum spectral transition position

Experimental results

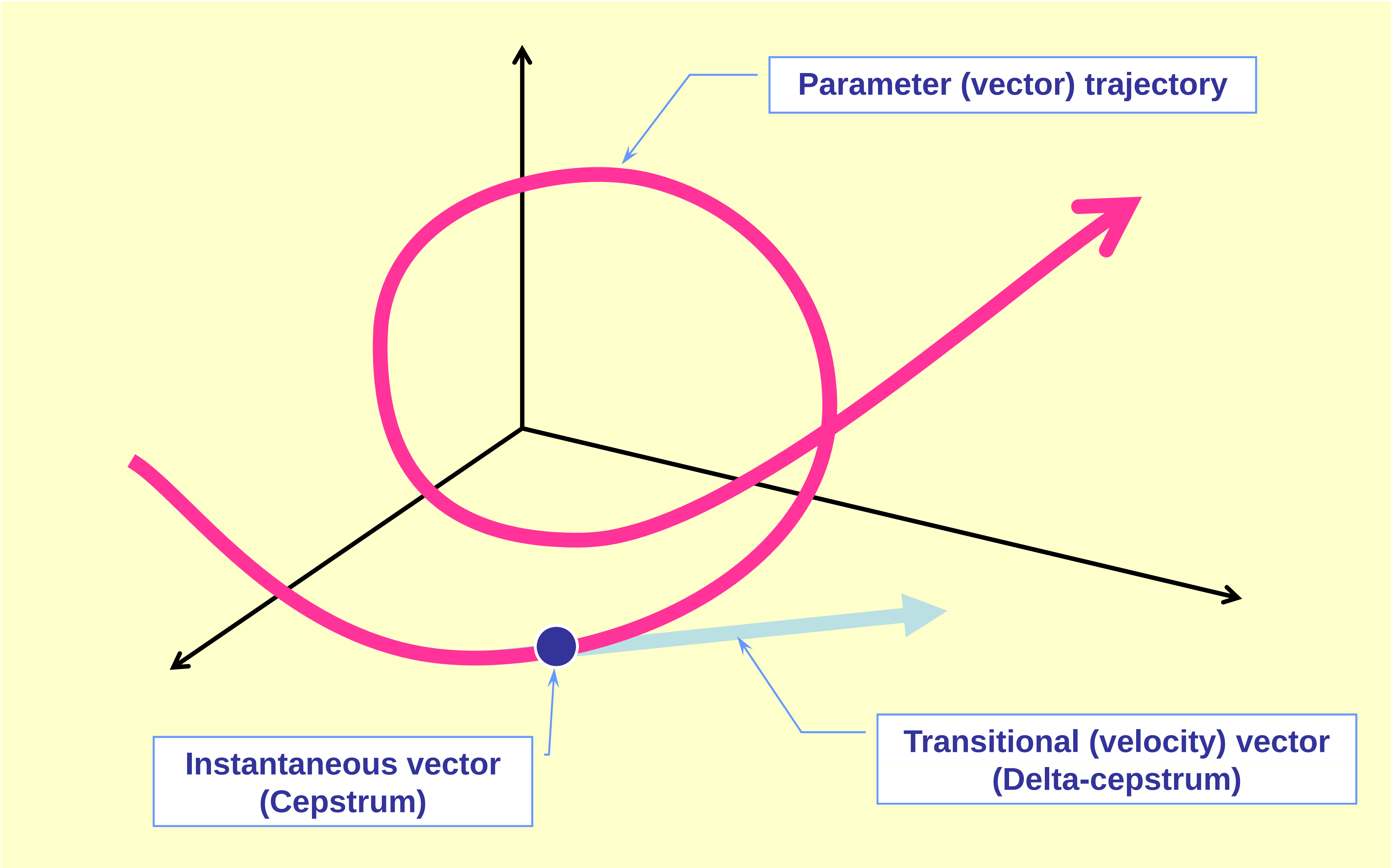
- "Perceptual critical points" (T_i, T_f) are related to maximum spectral transition positions (T_m).
- 10ms period including the T_m bears the most important information for consonant and syllable perception.
- Crucial information for both consonant and vowel identification is contained across the same transitional part of each syllable.
- The spectral transition is more crucial than unvoiced and buzz bar periods for consonant (syllable) perception.

Role of spectral transition for speech perception

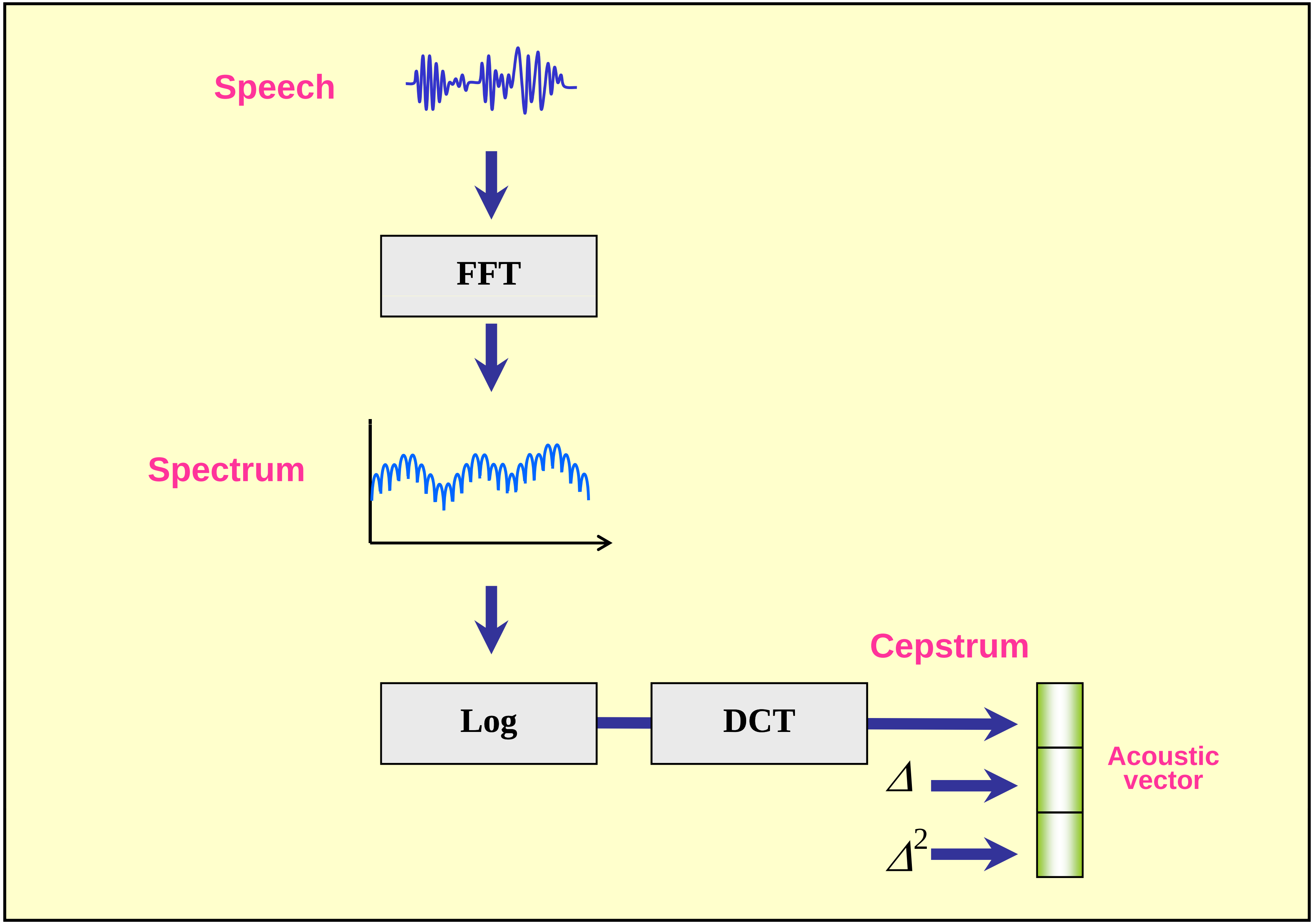


Maximum spectral change period: essential for syllable perception

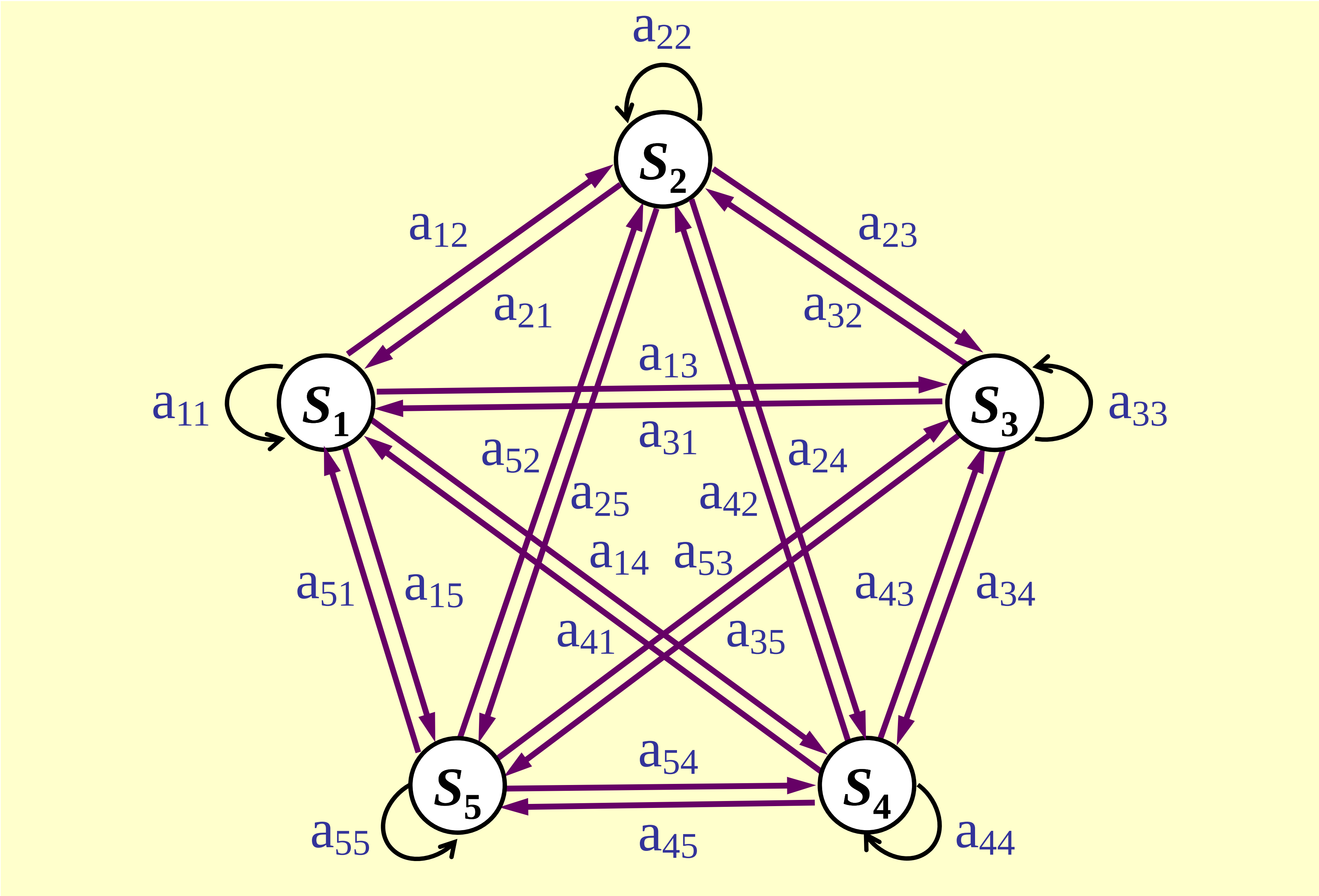
Cepstrum and delta-cepstrum coefficients



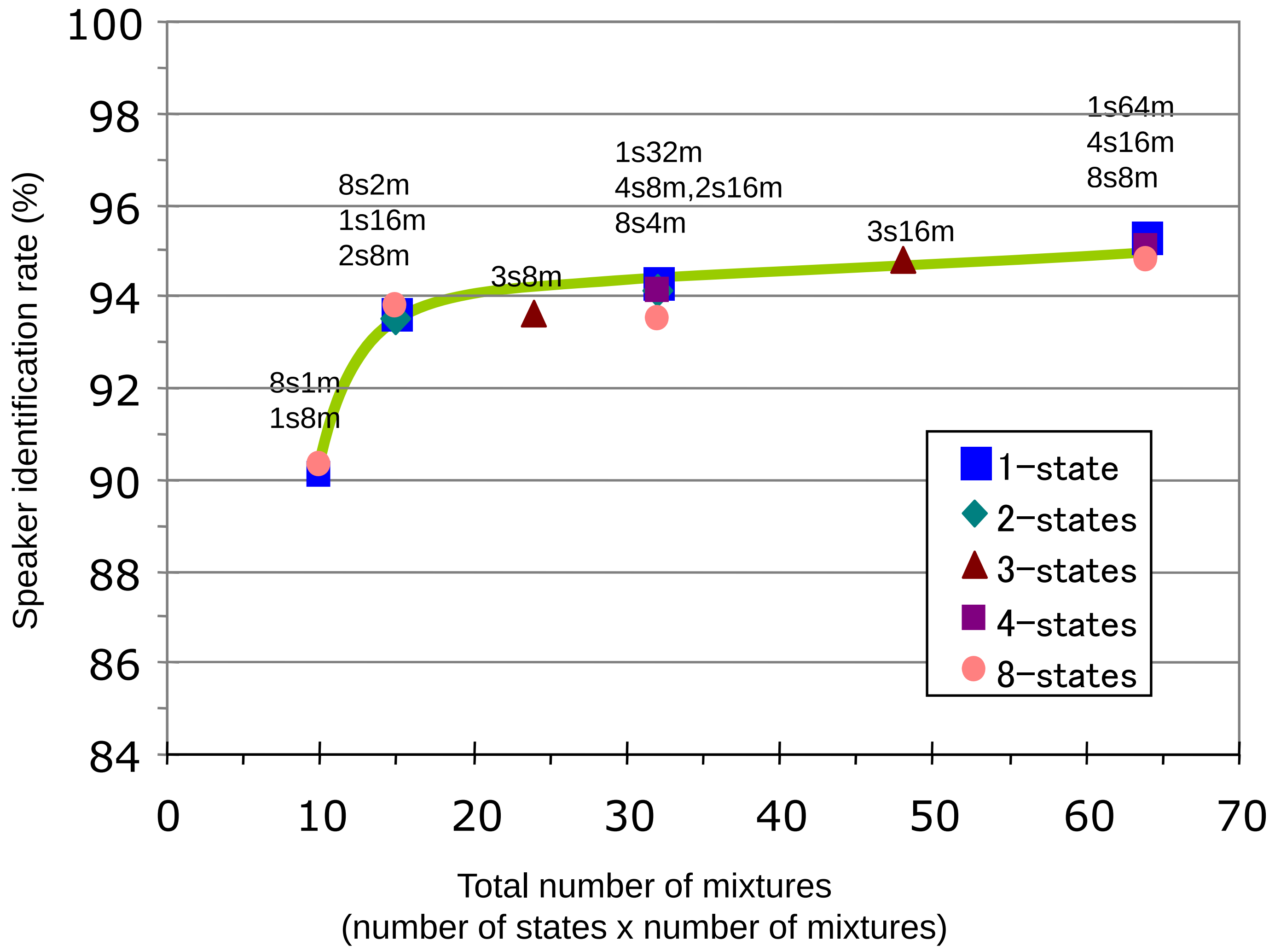
Instantaneous and dynamic cepstrum features



A five-state ergodic HMM
for text-independent speaker recognition



Speaker identification rates as a function of the number of states and mixtures in ergodic HMMs



1990s

- Japanese LVCSR using a newspaper corpus and broadcast news
- Robust ASR
- Text-prompted speaker recognition

Japanese LVCSR using a newspaper corpus and broadcast news

Comparison of lexica and LM training corpora for different languages

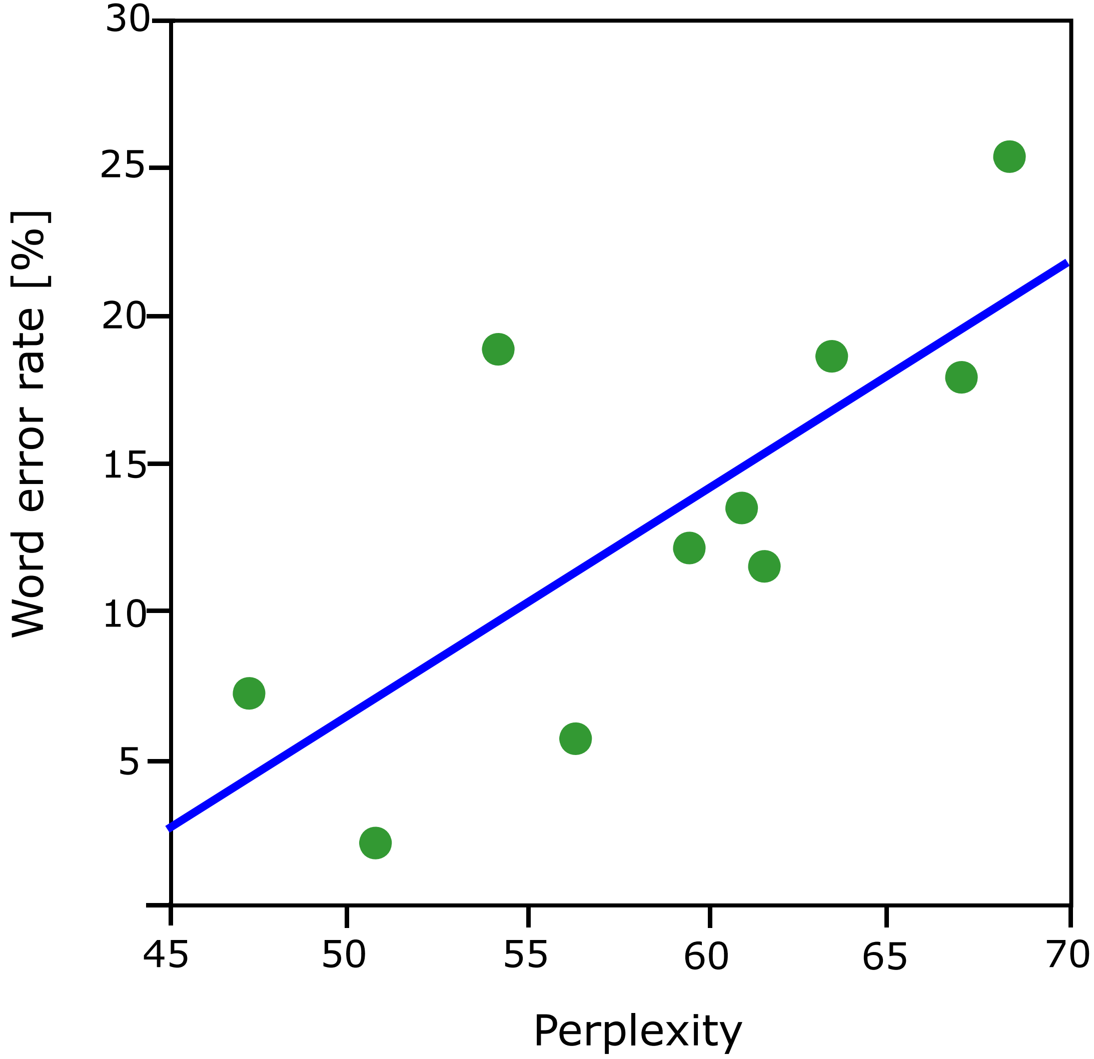
	Nikkei (Japanese)	WSJ (English)	Le Monde (French)	Frankfurter Rundschau (German)	Sole 24 (Italian)
Training test size [words]	180M	37.2M	37.7M	36M	25.7M
Distinct words	623k	165k	280k	650k	200k
5k coverage	88.0%	90.6%	85.2%	82.9%	88.3%
20k coverage	96.2%	97.5%	94.7%	90.0%	96.3%
40k coverage	98.2%	99.2%	97.6%	-	98.9%
65k coverage	99.0%	99.6%	98.3%	95.1%	99.0%
20k OOV rate	3.8%	2.5%	5.3%	10.0%	3.7%

LM units for Japanese: morphemes

Entry items corresponding to the number of homophone classes with *k* graphemic forms in the class

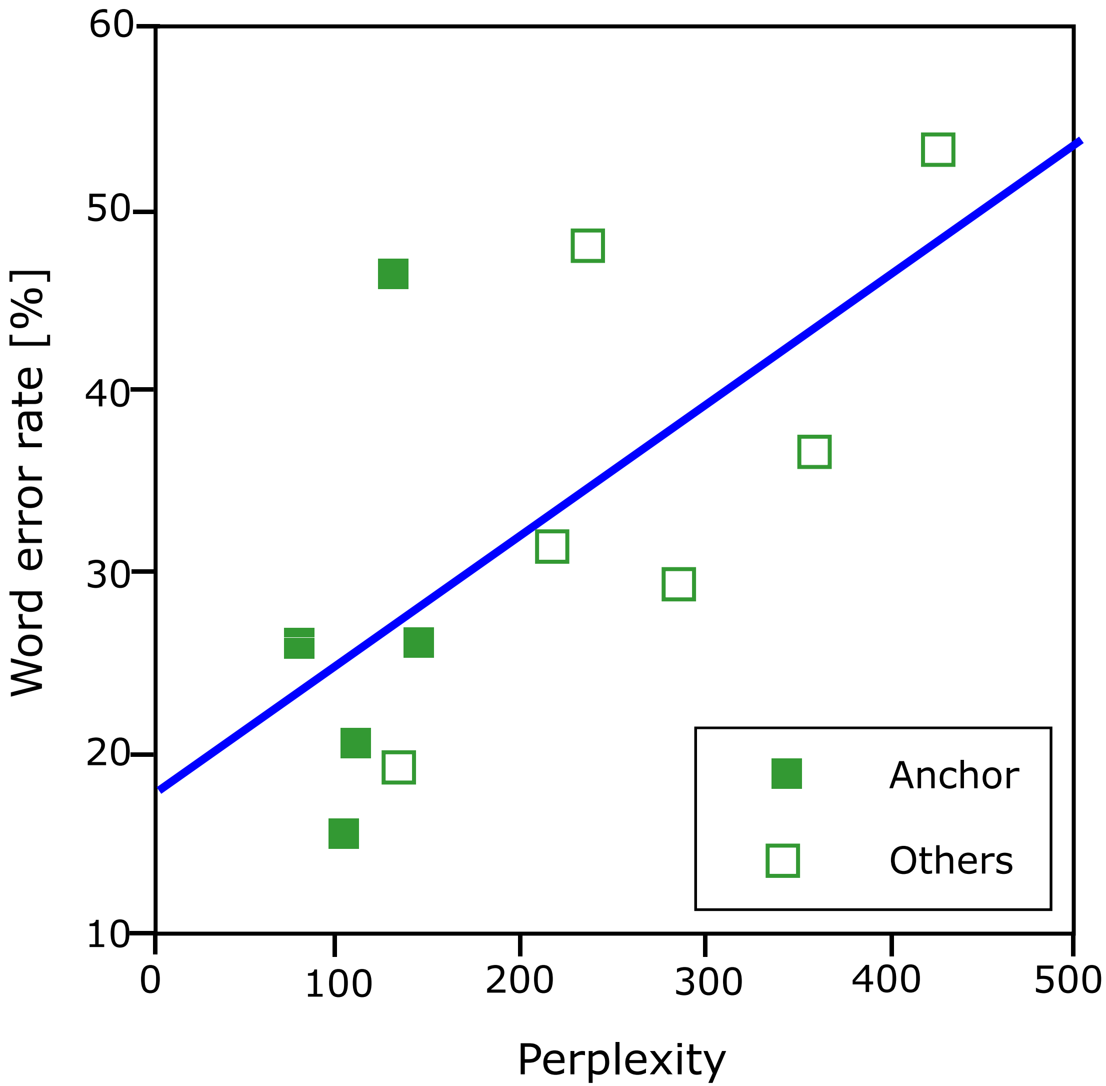
Corpus	Rate in Lexicon	Homophone class size (k)			
		1	2	3	>4
Nikkei (30k)	20%	24.1k	2438	706	565
BREF (10k)	35%	6686	1329	215	73
BREF (40k)	45%	23.7k	5361	1264	1039
WSJ (9k)	6%	8453	237	22	1
WSJ (65k)	15%	60.4k	3689	890	291
FR (64k)	10%	58.1k	2769	221	57
So24 (10k)	1.7%	9872	85	3	0

Relationship between perplexity (bigram) and word error rate for a read newspaper task



Mean word error rate by trigram: 9.5%

Relationship between perplexity (bigram) and word error rate for a broadcast-news task



Mean word error rate by trigram: 19.7% (Anchor)

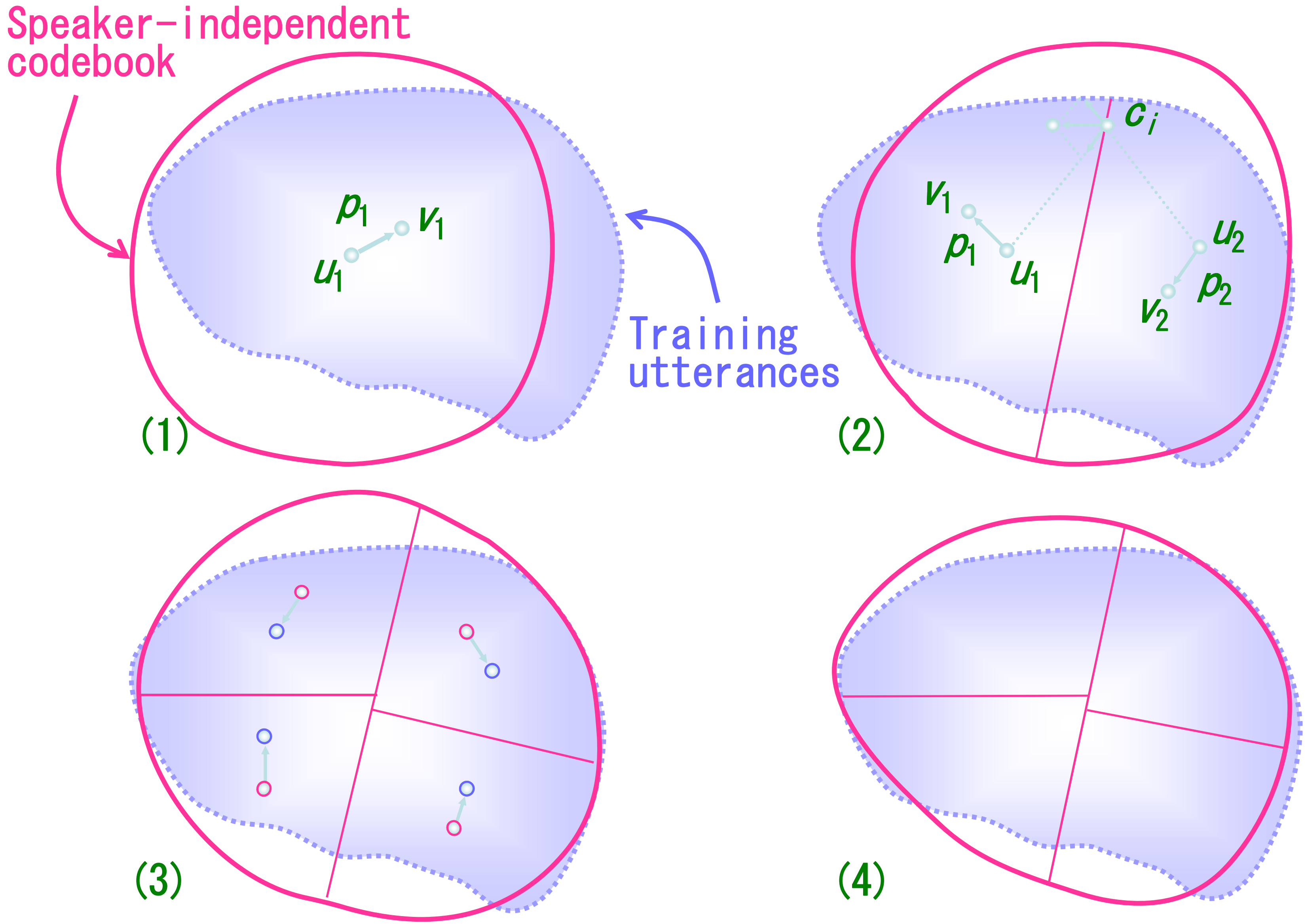
Robust ASR (Supervised/unsupervised acoustic model adaptation)

- Hierarchical spectral clustering-based unsupervised adaptation
- MAP+MCE (minimum classification error) training-based supervised adaptation
- N-best-based unsupervised adaptation

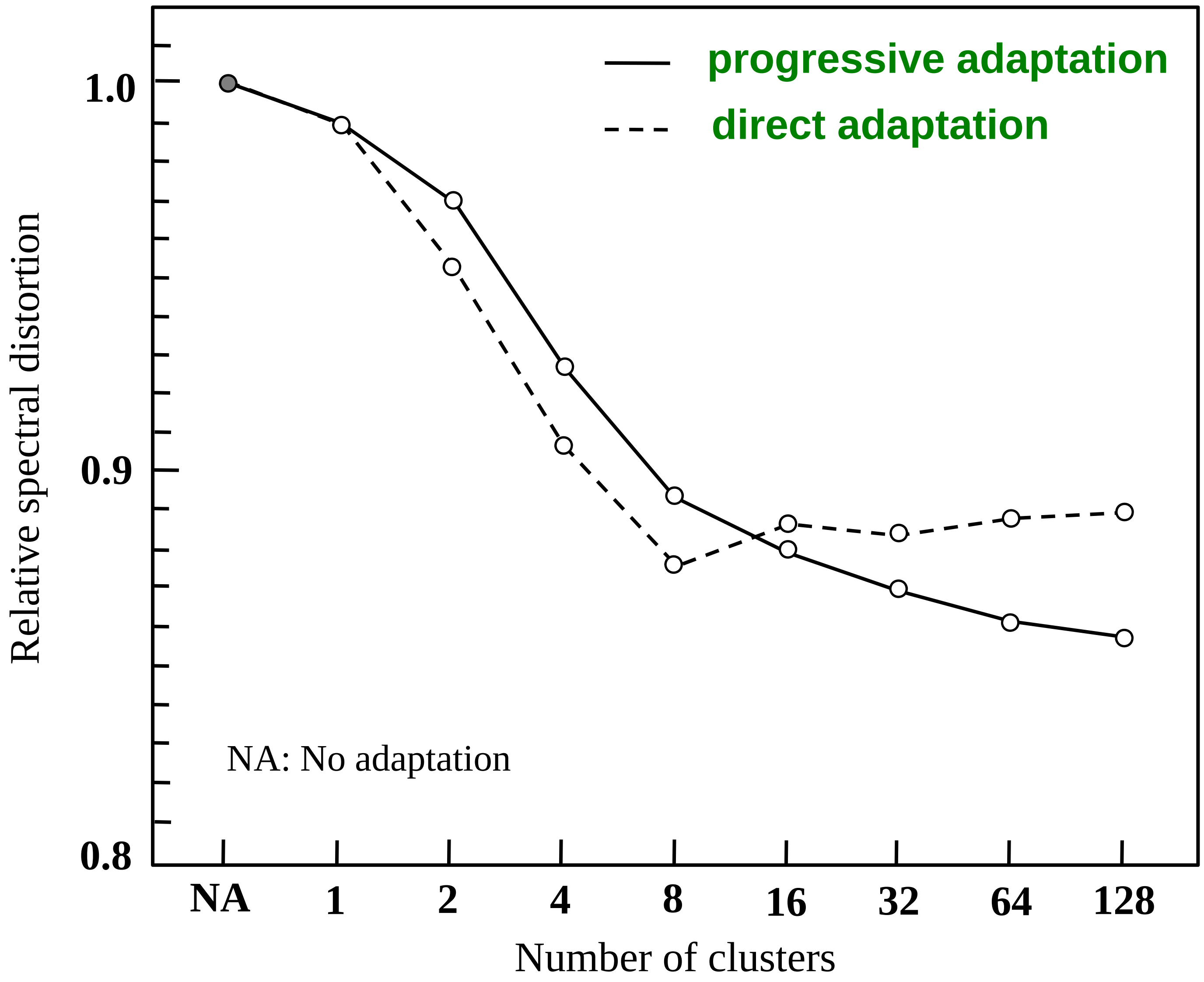
Robust ASR (Supervised/unsupervised acoustic model adaptation)

- Hierarchical spectral clustering-based unsupervised adaptation
- MAP+MCE (minimum classification error) training-based supervised adaptation
- N-best-based unsupervised adaptation

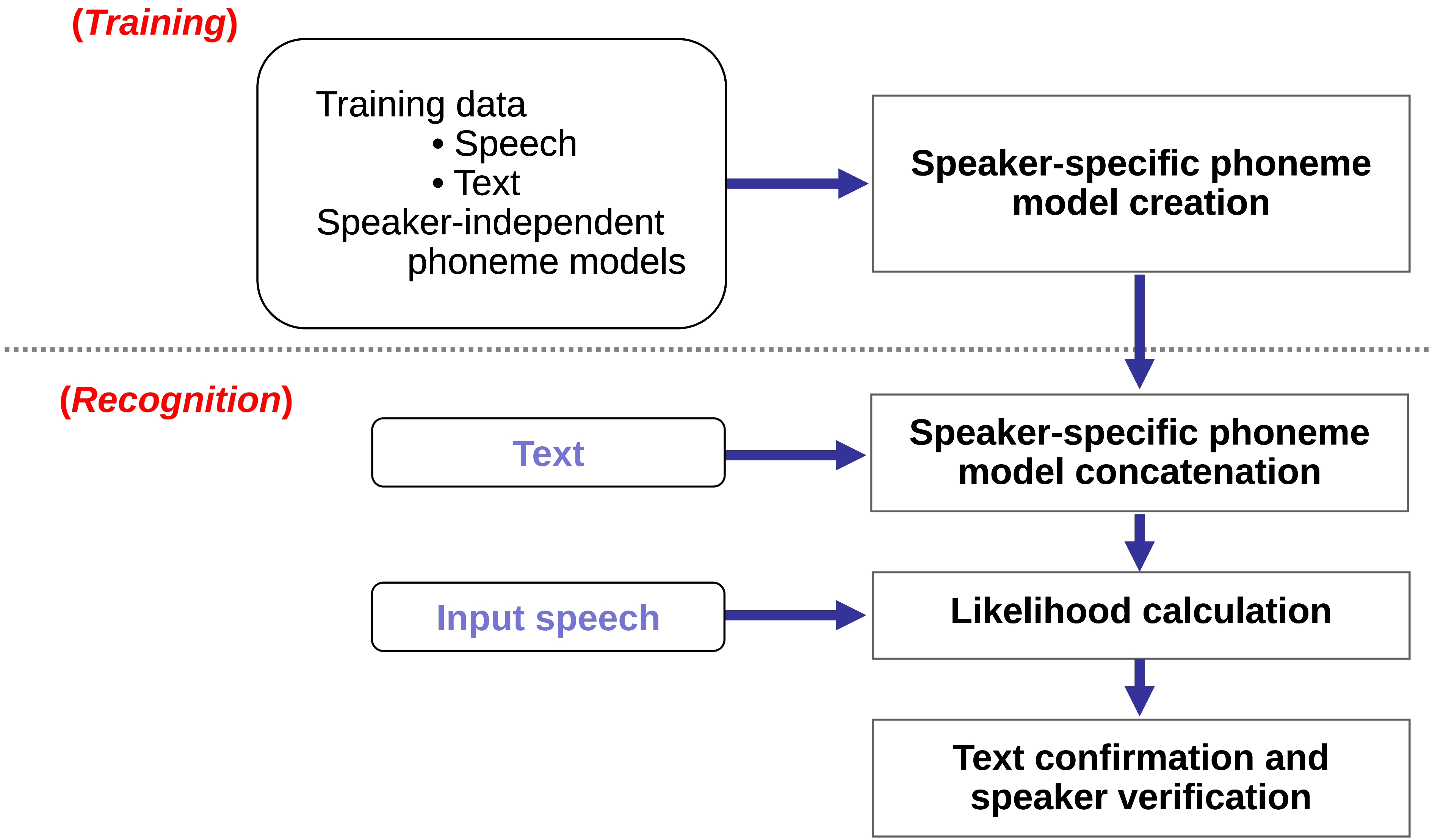
Hierarchical codebook adaptation algorithm maintaining continuity between adjacent clusters



Cepstral distortion between input speech and reference templates resulted from hierarchical codebook adaptation



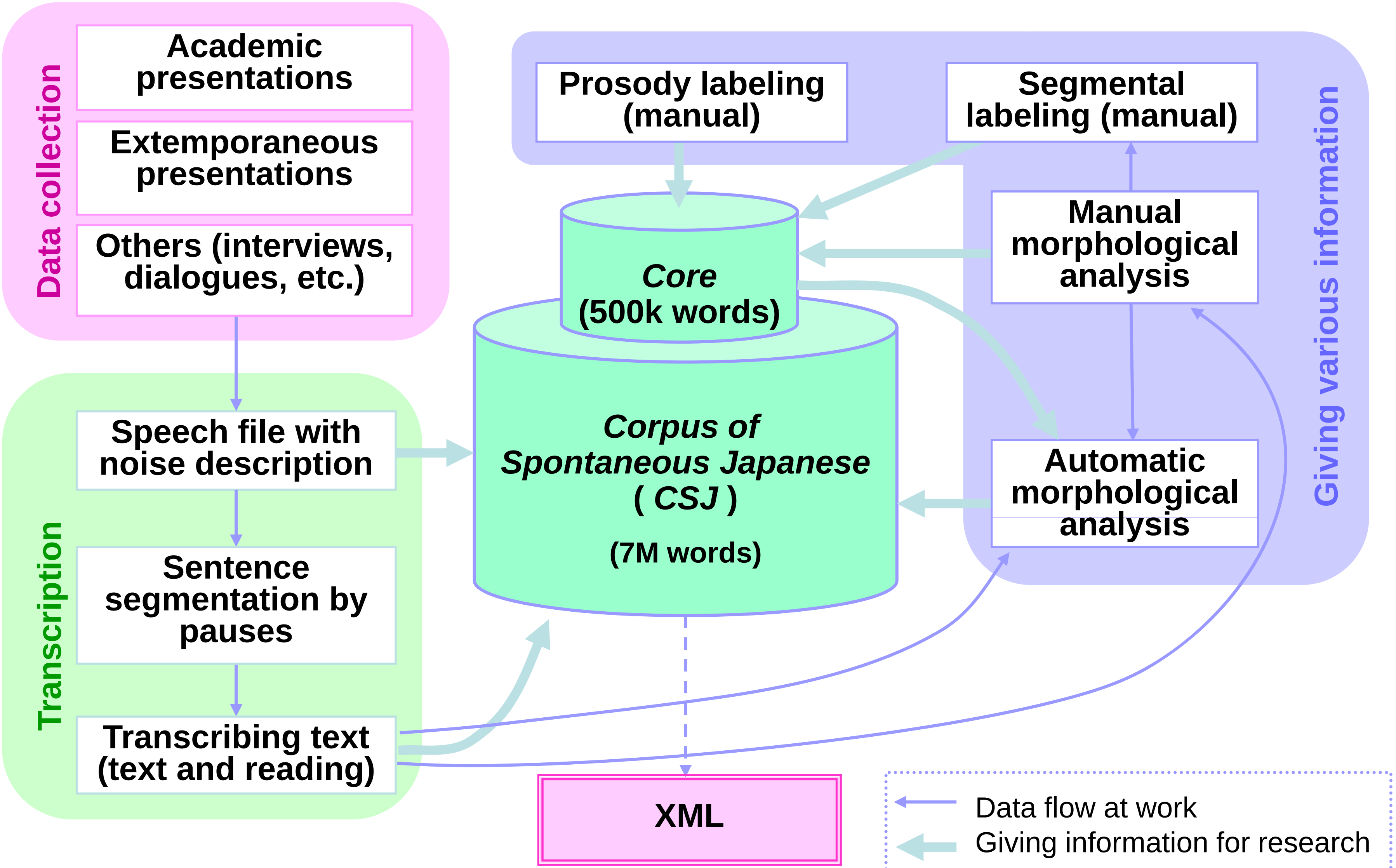
Text-prompted speaker recognition method



2000s (1)

- Spontaneous speech recognition project and CSJ corpus
- Spectral reduction in spontaneous speech
- Automatic speech summarization and evaluation

CSJ corpus construction

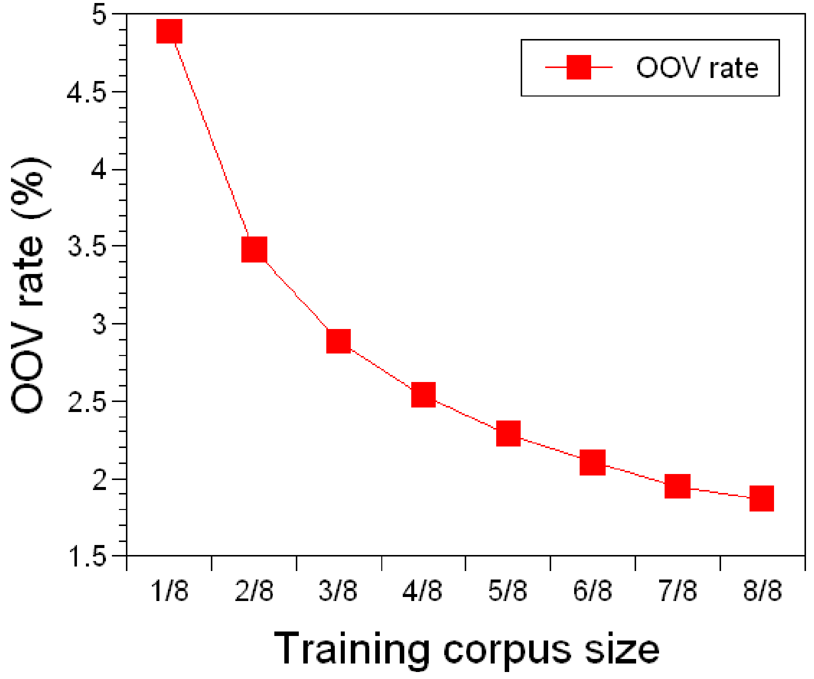


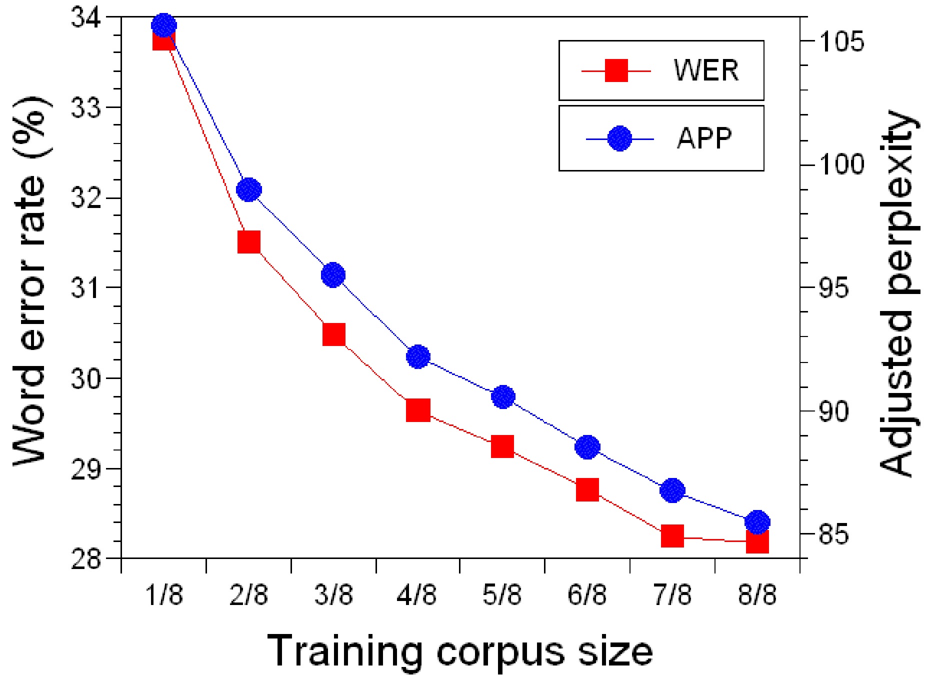
Contents of the CSJ

Type of Speech	# Speakers	# Files	Monologue/ Dialogue	Spontaneous/ Read	Hours
Academic presentations (AP)	838	1006	Monolog	Spont.	299.5
Extemporaneous presentations (EP)	580	1715	Monolog	Spont.	327.5
Interview on AP	* (10)	10	Dialog	Spont.	2.1
Interview on EP	* (16)	16	Dialog	Spont.	3.4
Task oriented dialogue	* (16)	16	Dialog	Spont.	3.1
Free dialogue	* (16)	16	Dialog	Spont.	3.6
Reading text	*(244)	491	Dialog	Read	14.1
Reading transcriptions	* (16)	16	Monolog	Read	5.5
				Total hours	658.8

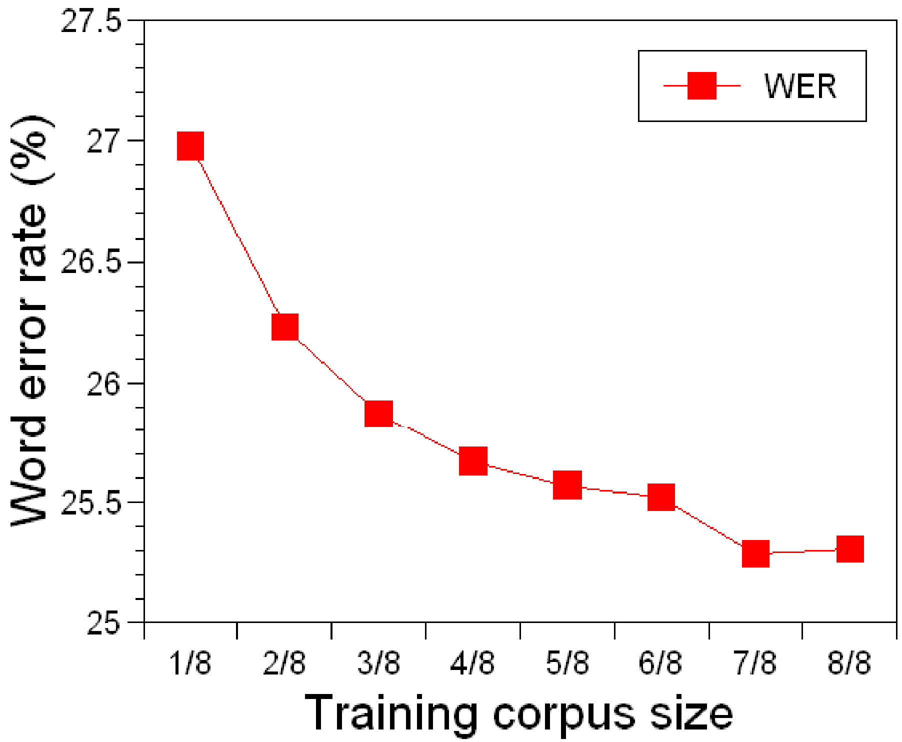
*Counted as the speakers of AP or EP

Out-of-vocabulary (OOV) rate, word error rate (WER) and adjusted test-set perplexity (APP) as a function of the size of language model training data (8/8 = 6.84M words)





Word error rate (WER) as a function of the size of acoustic model training data (8/8 = 510 hours)



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

Speaker-independent recognition

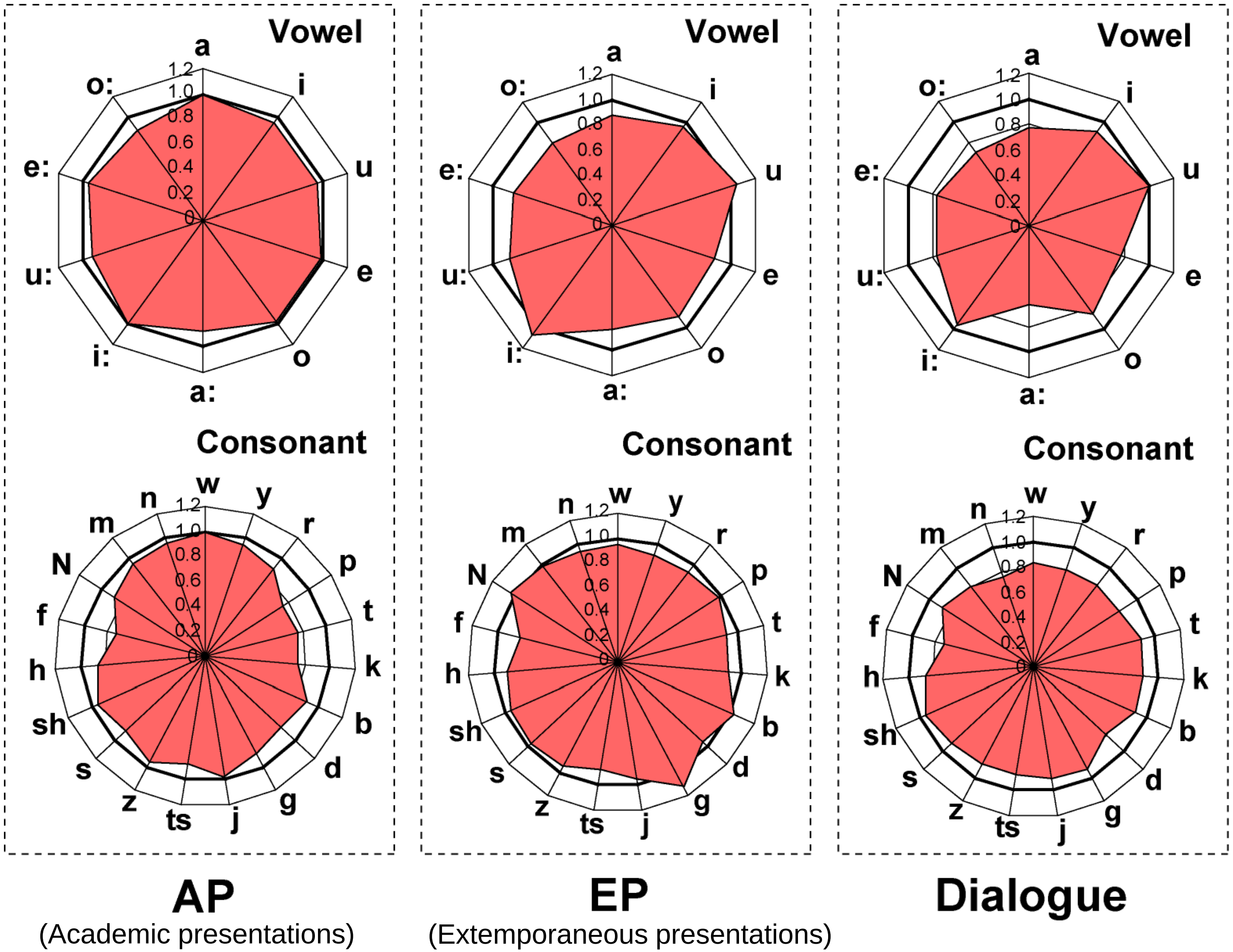
$$Acc = 0.12AL - 0.88SR - 0.020PP - 2.2OR + 0.32FR - 3.0RR + 95$$

Speaker-adaptive recognition

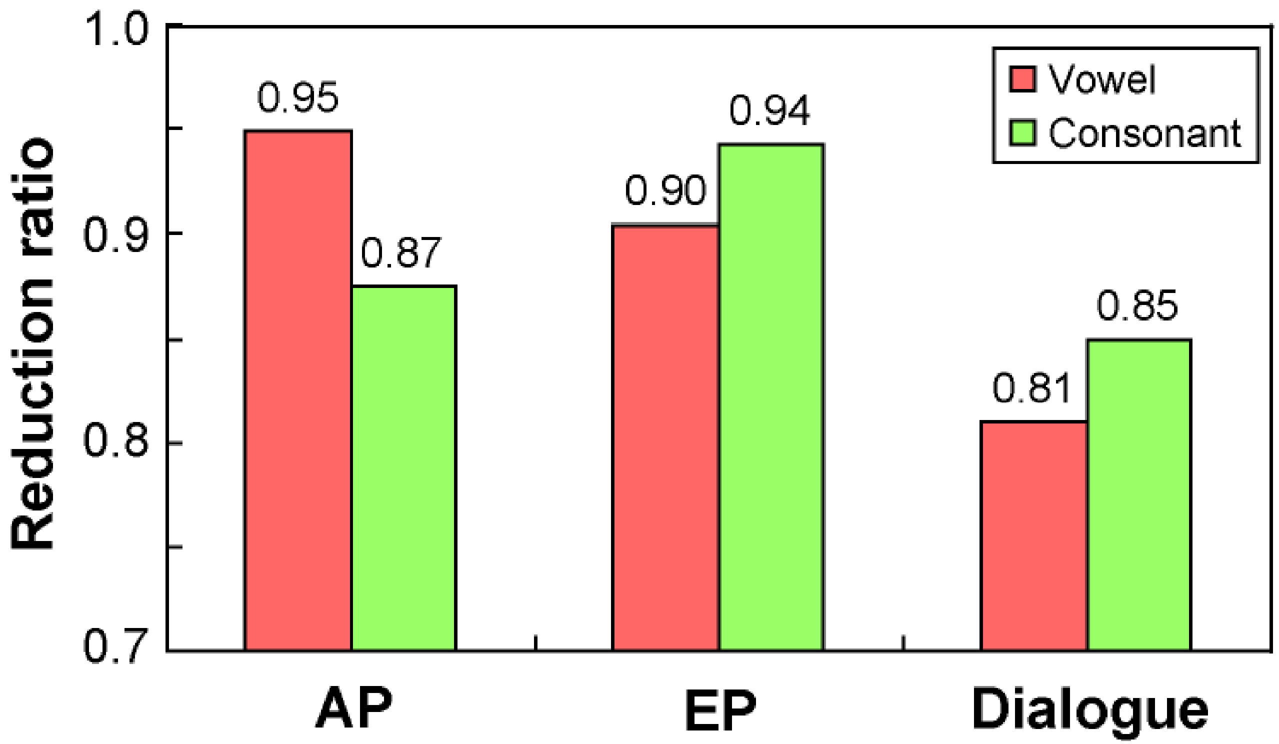
$$Acc = 0.024AL - 1.3SR - 0.014PP - 2.1OR + 0.32FR - 3.2RR + 99$$

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

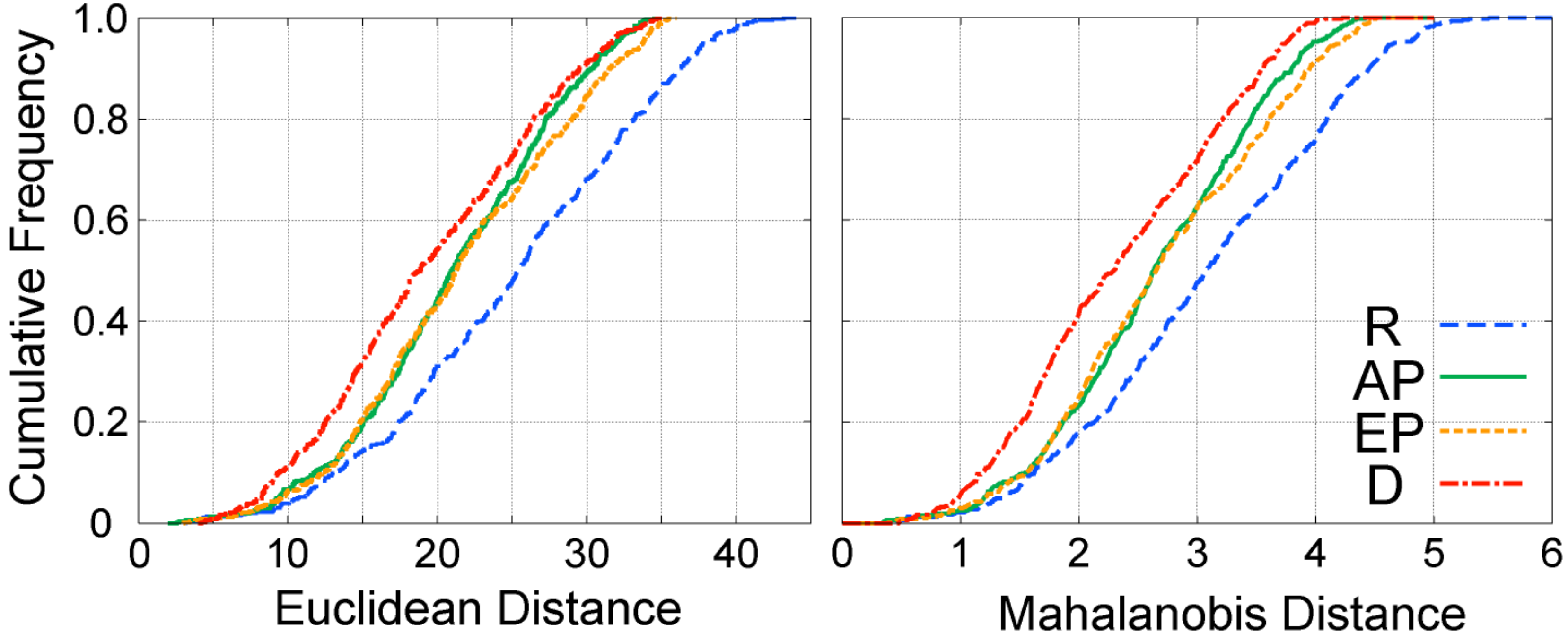
The reduction ratio of the vector norm between each phoneme and the phoneme center in the spontaneous speech to that in the read speech



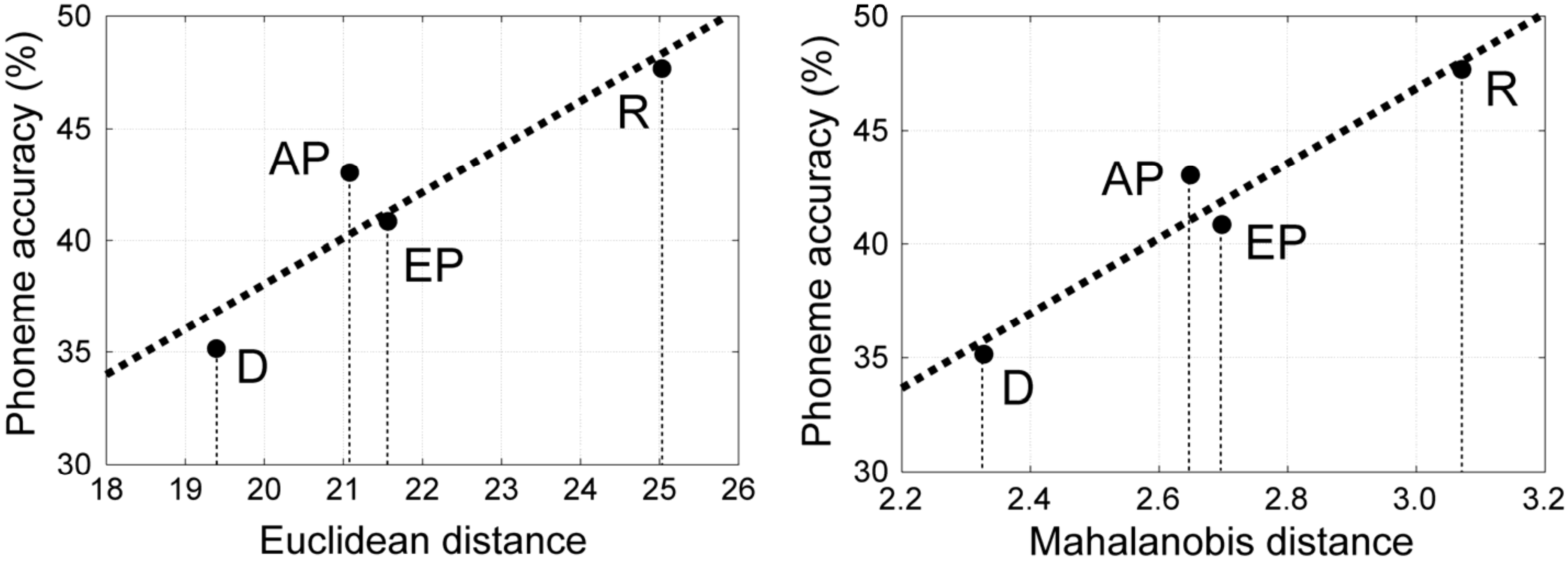
Mean reduction ratios of vowels and consonants for each speaking style



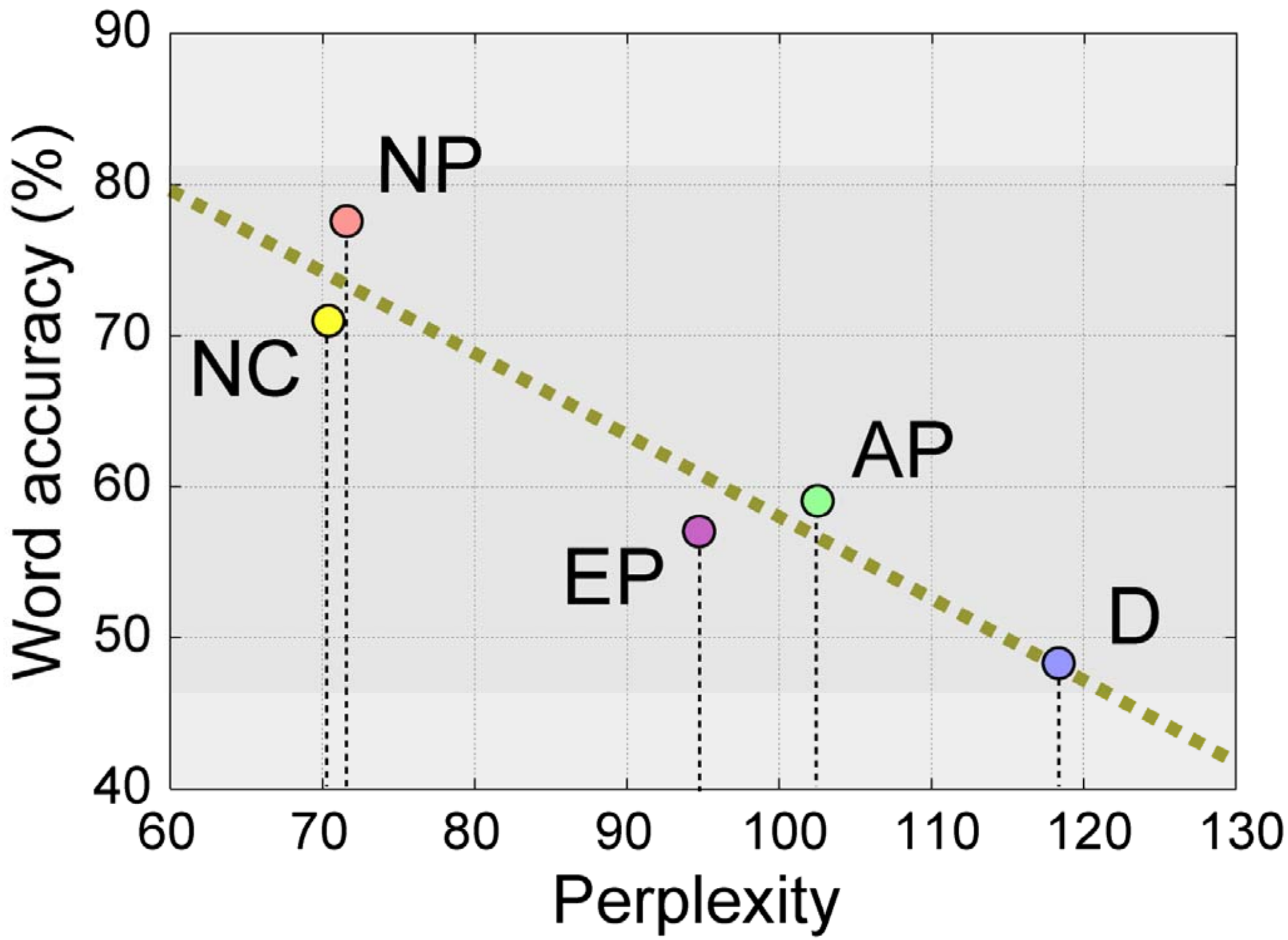
Distribution of distances between phonemes (R: read speech, D: dialogue)



Relationship between phoneme distances and phoneme recognition accuracy

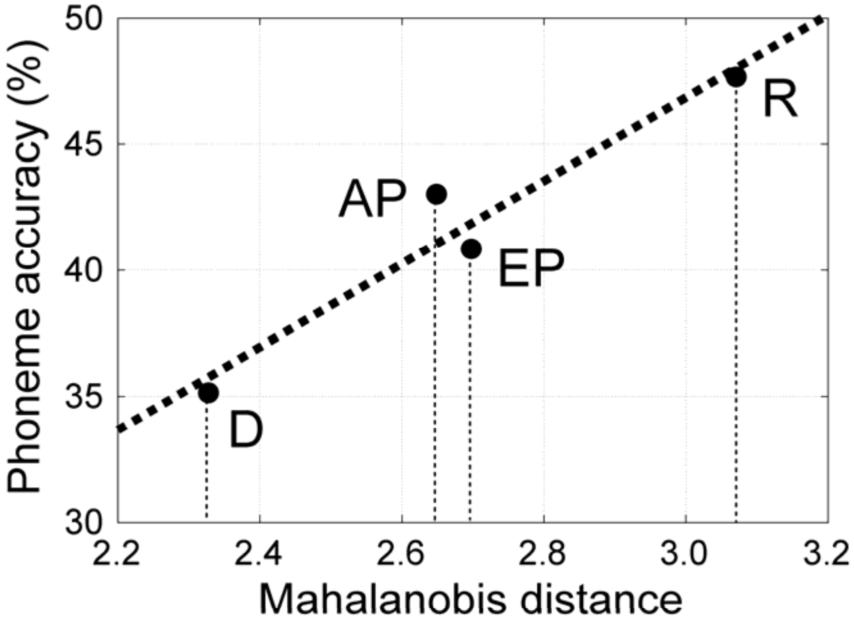


Relationship between test-set perplexity and word recognition accuracy (%) (NC: news commentary)

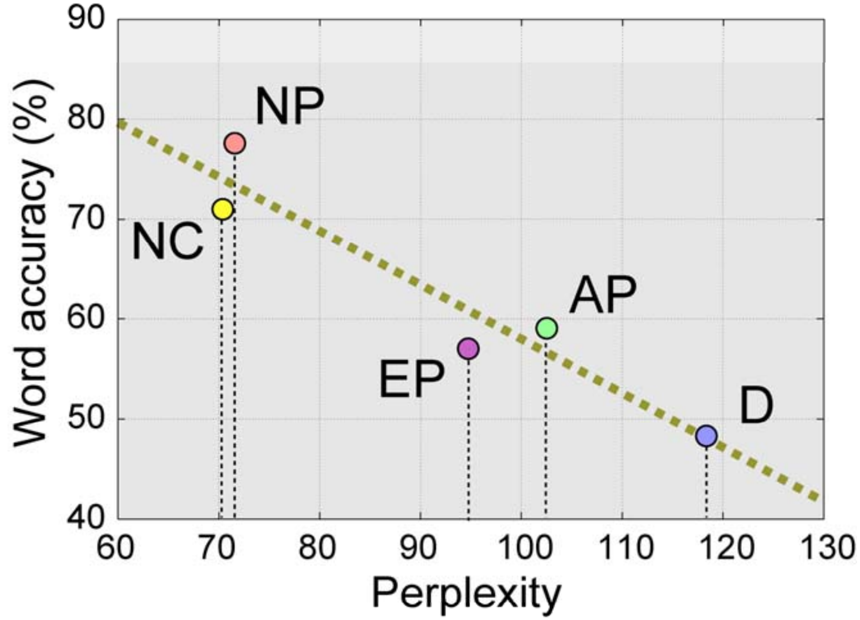


Equation for estimating word recognition accuracy

(Acoustic variation only)



(Acoustic and linguistic variations)



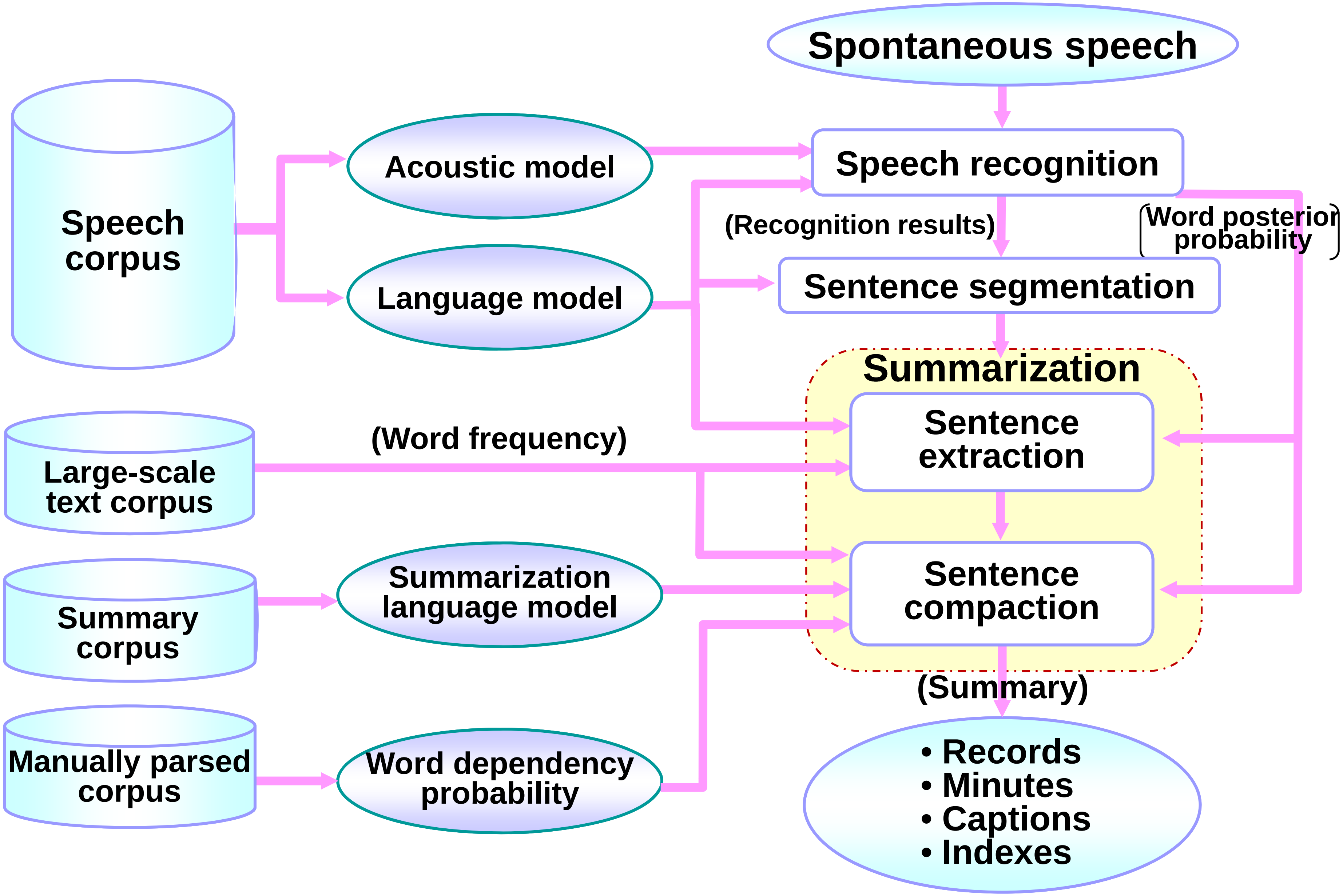
$$Accuracy \simeq ax + by + c$$

x : Mean Mahalanobis distance between phonemes

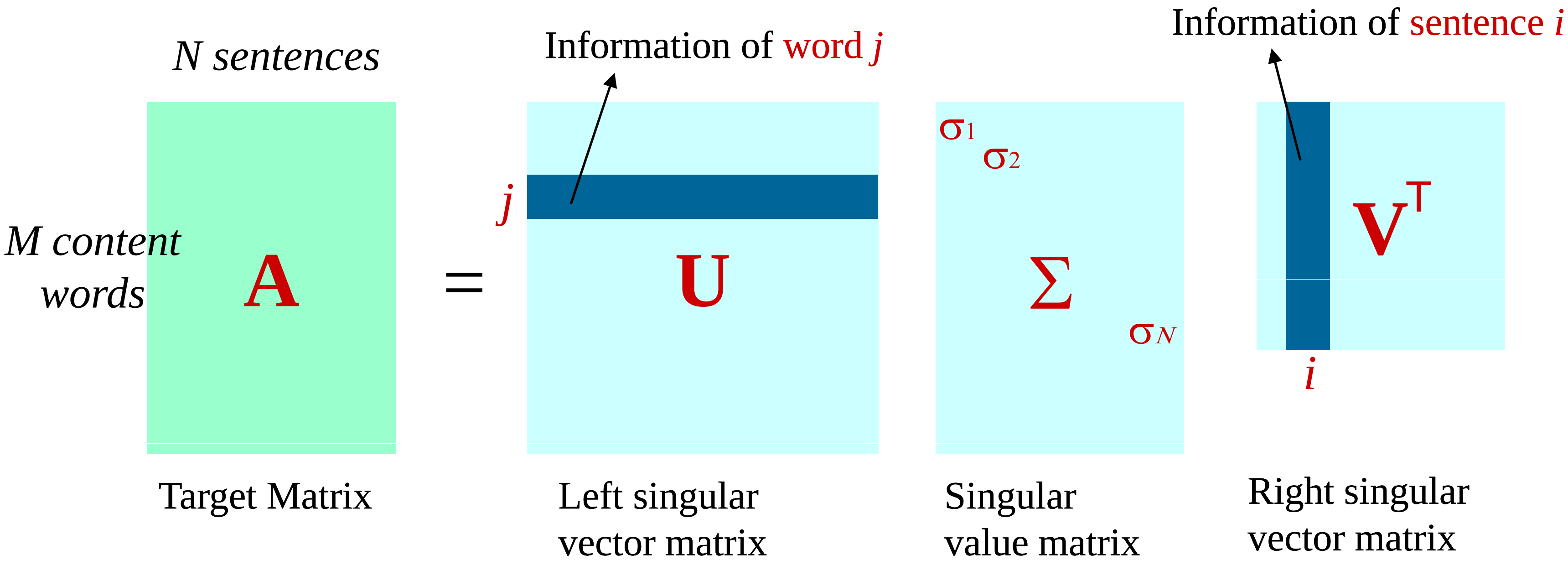
y : Test – set perplexity

a, b, c : Constant

Speech summarization by sentence extraction and compaction



Sentence clustering using SVD



- SVD semantically clusters content words and sentences
  - Deriving a latent semantic structure from a presentation speech represented by the matrix **A**
- Element a_{mn} of the matrix **A**

$$a_{mn} = f_{mn} \cdot \log (F_A / F_m)$$

f_{mn} : Number of occurrences of a content word (m) in the sentence (n)

F_m : Number of occurrences of a content word (m) in a large corpus

LSA-based sentence extraction

- Dimension reduction by SVD

$$A_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ a_{3i} \\ \vdots \\ a_{Mi} \end{bmatrix} \xrightarrow{\text{SVD}} \hat{A}_i = \begin{bmatrix} \sigma_1 v_{i1} \\ \sigma_2 v_{i2} \\ \vdots \\ \sigma_N v_{iN} \end{bmatrix} \xrightarrow{\text{Dimension reduction}} \psi_i = \begin{bmatrix} \sigma_1 v_{i1} \\ \vdots \\ \sigma_K v_{iK} \end{bmatrix}$$

- Each sentence is represented by a weighted singular-value vector
- In order to evaluate each sentence, the score of each sentence is calculated by the norm in the K dimensional space

$$\|\psi_i\| = \sqrt{\sum_{k=1}^{K} (\sigma_k v_{ik})^2} \Longrightarrow \text{Score for sentence extraction}$$

A fixed number of sentences having relatively large sentence scores in the reduced dimensional space are selected.

Word extraction score

Summarized sentence with M words $V = v_1, v_2, \ldots, v_M$

Score

$$S(V^M) = \sum_{m=1}^{M} \left\{ L(v_m \mid \cdots v_{m-1}) + \lambda_I I(v_m) + \lambda_C C(v_m) + \lambda_T T_r(v_m) \right\}$$

Linguistic score
Linguistic correctness
(Bigram/Trigram)

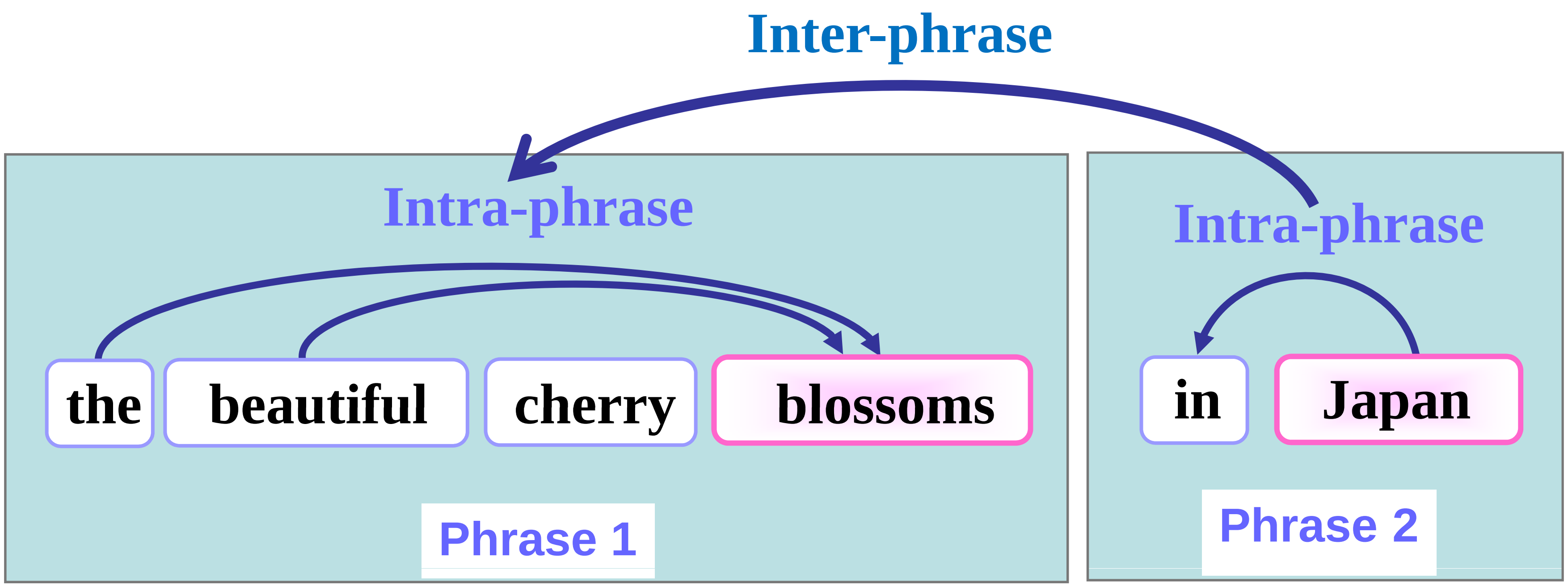
Significance (topic) score
Important information extraction
(Amount of information)

Confidence score
Recognition error exclusion
(Acoustic & linguistic reliability)

Word concatenation score
Semantic correctness
(Word dependency probability)

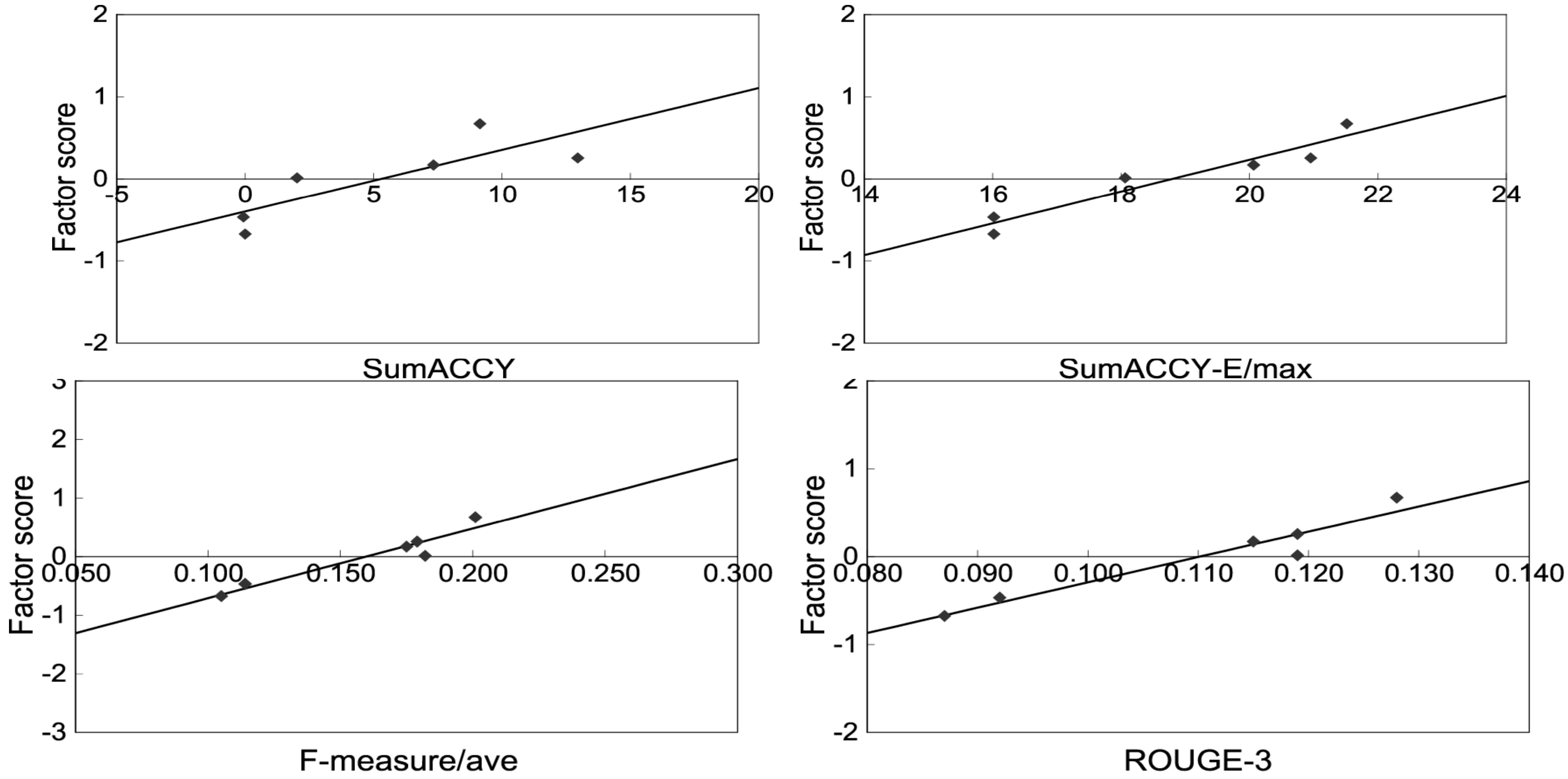
Word concatenation score

A penalty for word concatenation with no dependency in the original sentence



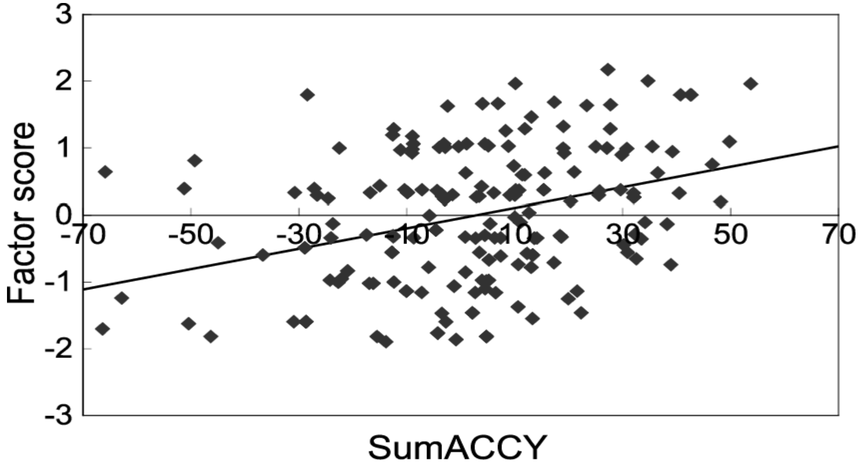
“the beautiful Japan” → Grammatically correct but incorrect as a summary

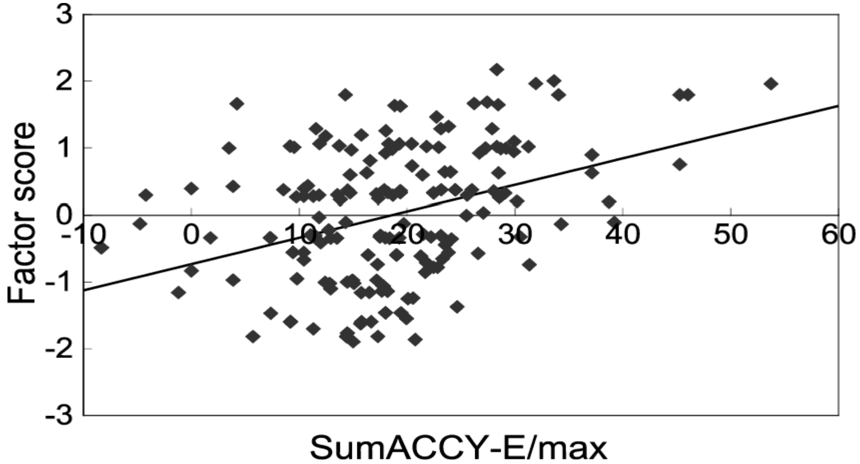
Correlation between subjective and objective evaluation scores (averaged over presentations)

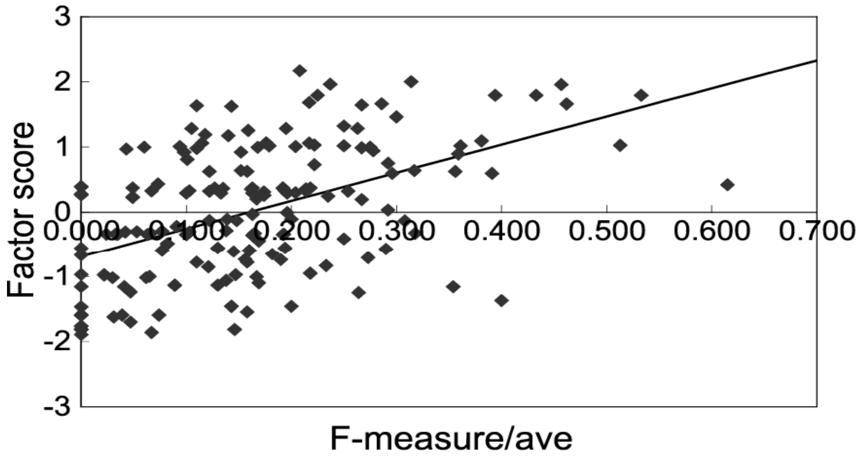


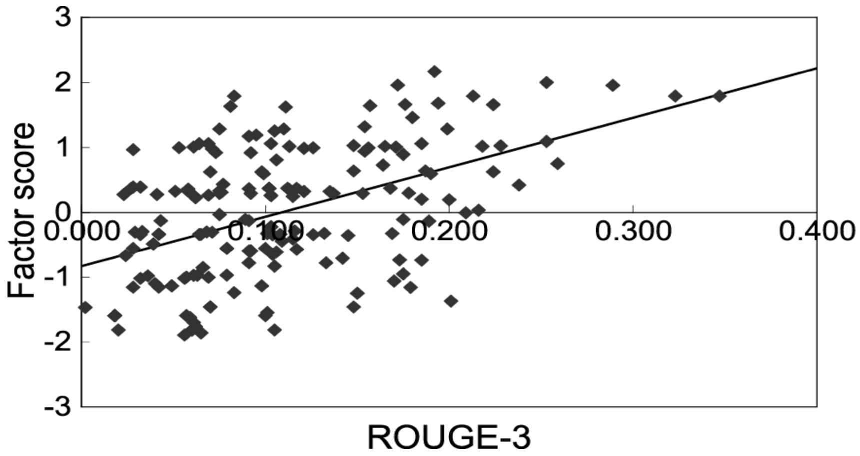
In the subjective evaluation, the summaries were evaluated in terms of ease of understanding and appropriateness as summaries on five levels.

Correlation between subjective and objective evaluation scores (each presentation)





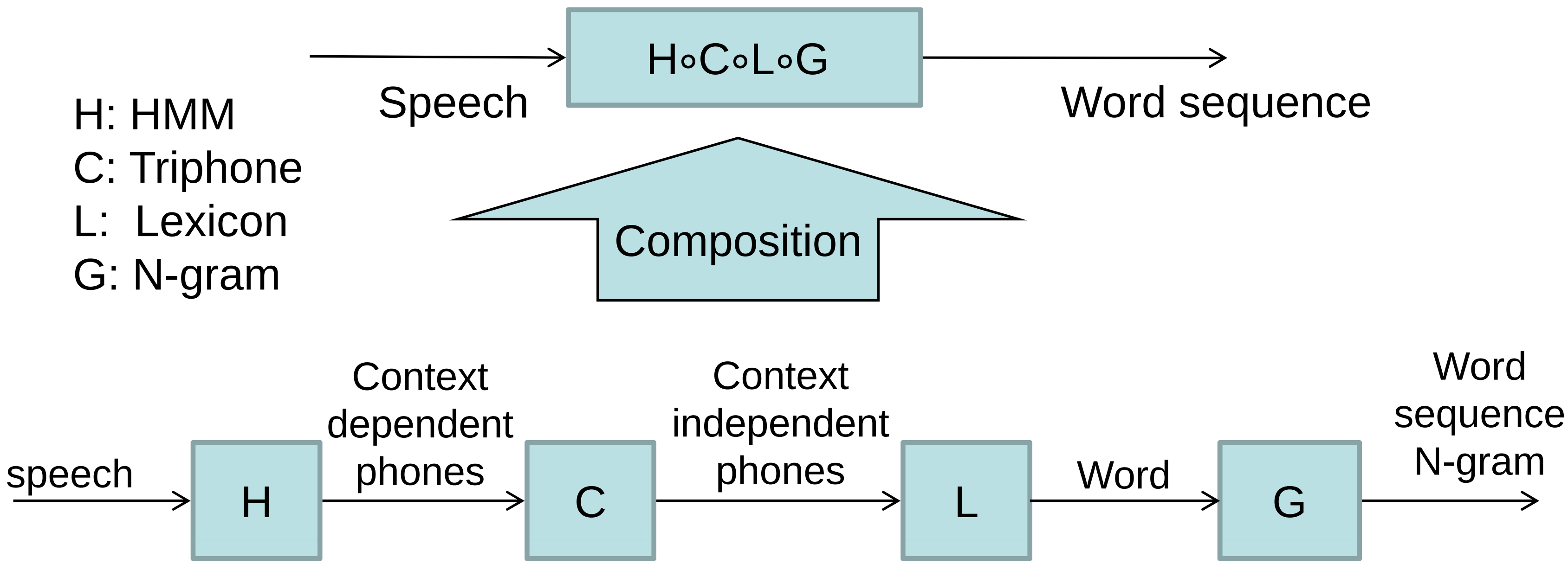




2000s (2)

- Development of WFST-based decoder and application
- Unsupervised cross-validation and aggregated adaptation methods

WFST (Weighted Finite State Transducer)-based “T^3 decoder”

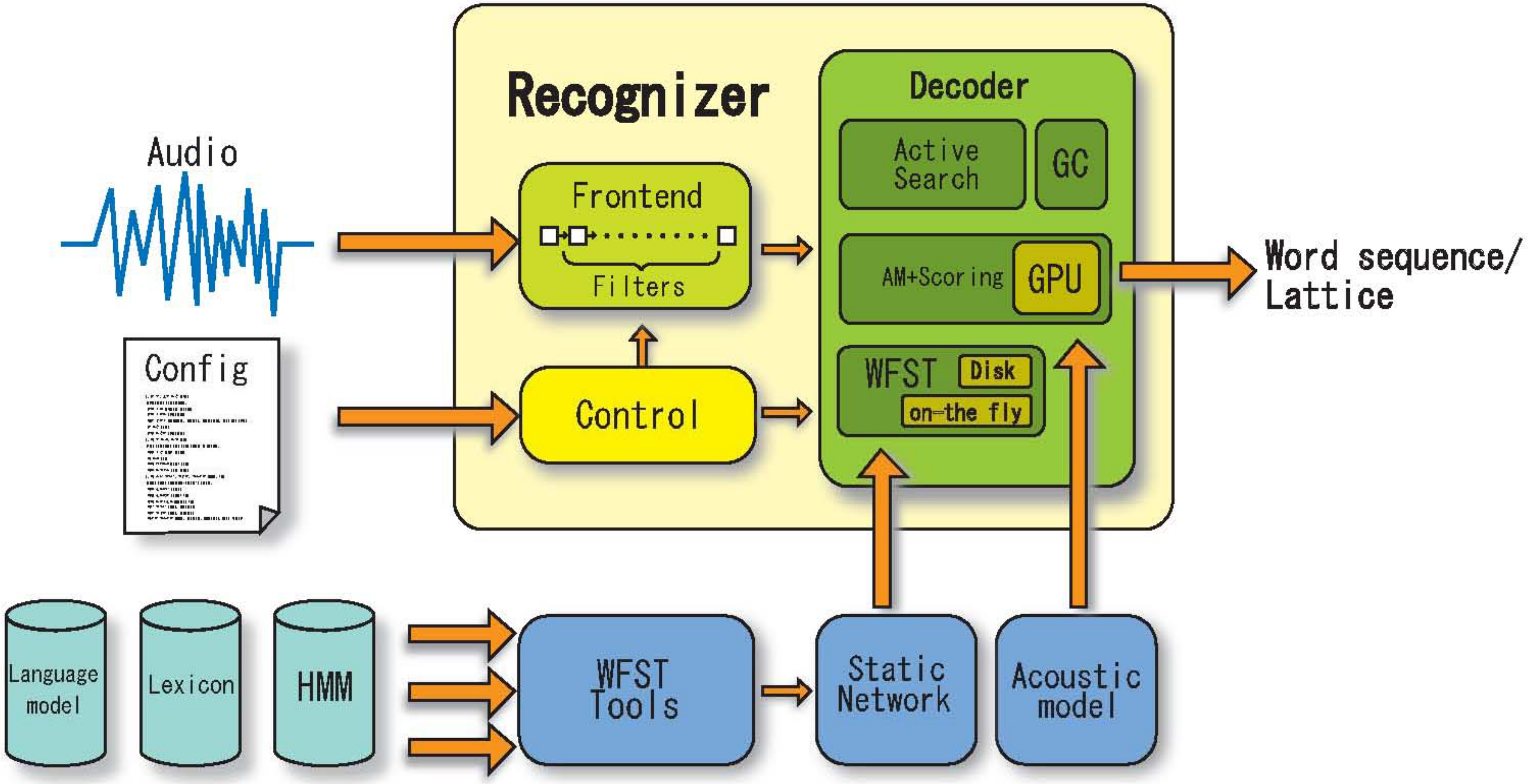


Problems:
- Large memory requirement
- Small flexibility
  - Difficult to change partial models

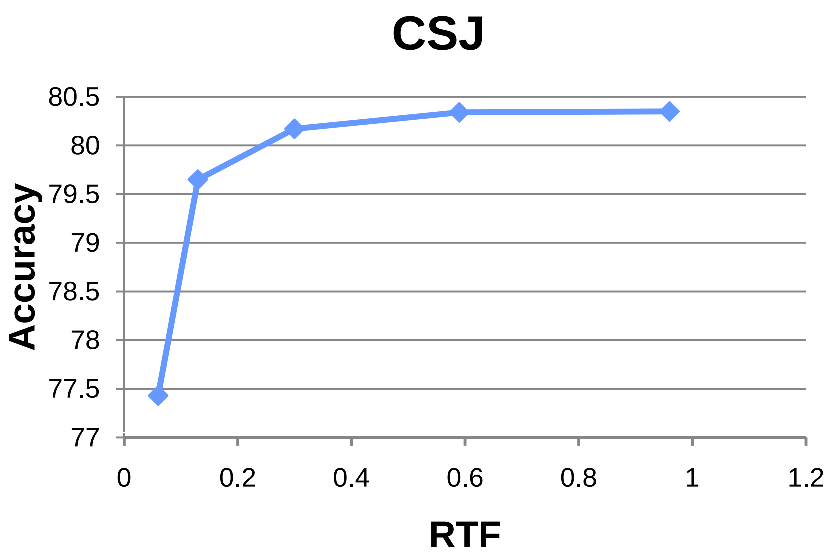
On-the-fly composition

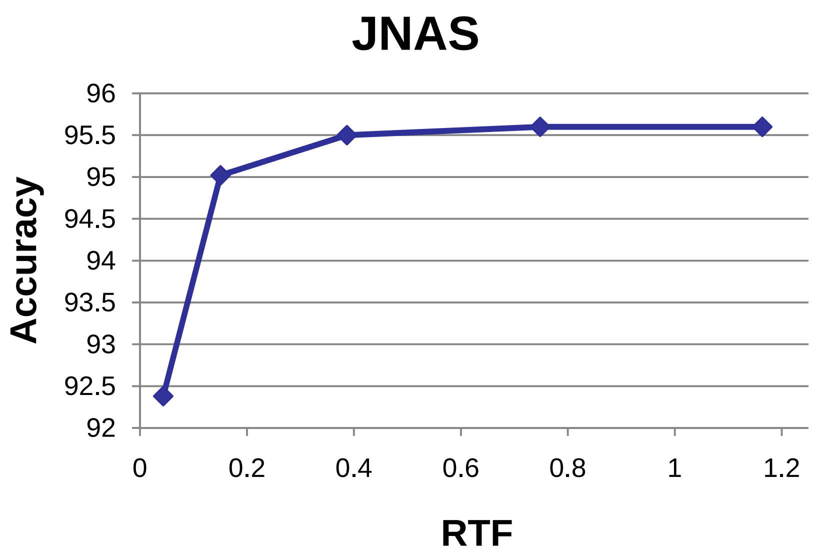
Parallel decoding

Structure of the T^3 decoder



T^3 decoder performance





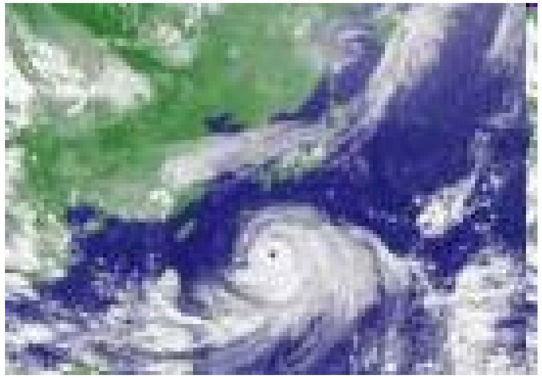
- Spontaneous speech
- "Corpus of Spontaneous Japanese (CSJ)"
- Test set of 10 lectures
- 128 Gaussians per mixture
- 65K word vocabulary

- Read speech
- "Japanese Newspaper Article Sentences (JNAS)"
- Test set of 200 utterances
- 16 Gaussians per mixture
- 465k word vocabulary

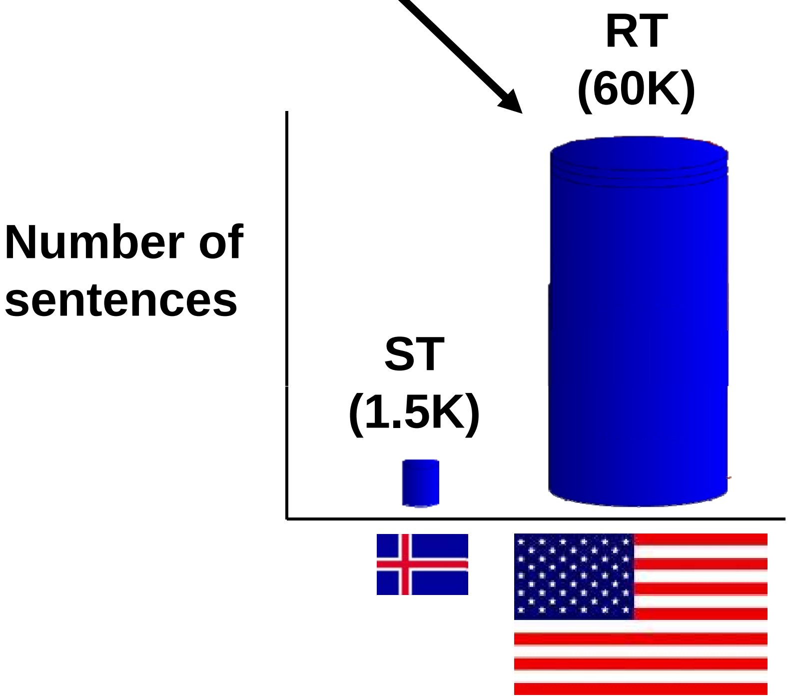
Icelandic speech recognition using an English corpus

- The Jupiter corpus

  (a weather information corpus developed by MIT) was used as the English rich corpus



Weather information domain



RT: Rich text
ST: Sparse text

Icelandic speech recognition using English LM (English output)

Traditional format

$$\widetilde{T} = \underset{T}{\operatorname{argmax}} \max_{W} P(O|W) P(W|T)^{\lambda_{ST}} P(T)^{\lambda_{RT}}$$

WFST format

Icelandic speech $\longrightarrow$ $H \circ C \circ L \circ G_{ST} \circ Tr \circ G_{RT}$ $\longrightarrow$ **English Text**

- $P(O|W)$: Icelandic acoustic model
- $P(W|T)$: English to Icelandic translation model
- $P(T)$: English language model

Icelandic speech recognition using English LM (Icelandic output)

Traditional format

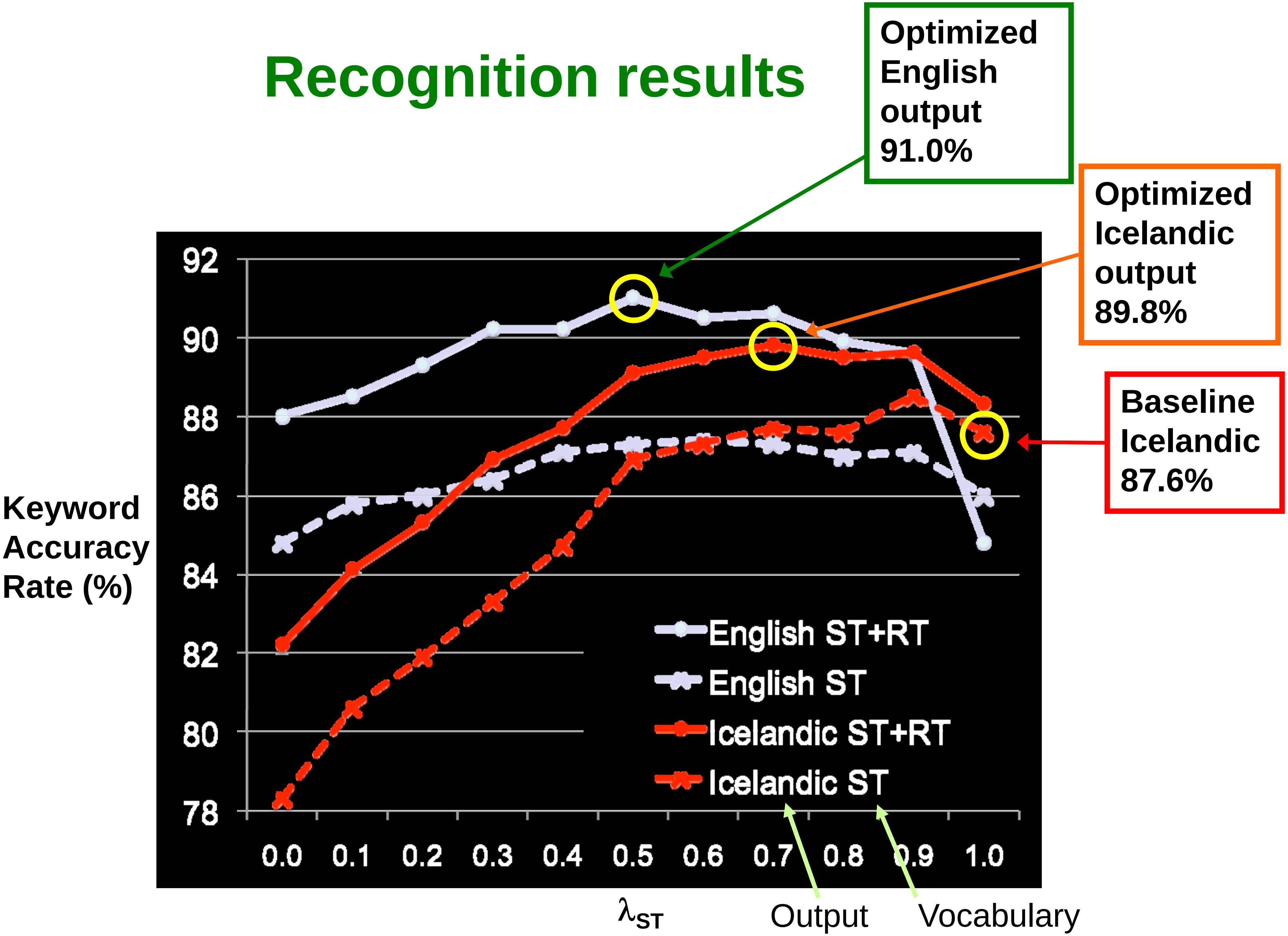
$$\widetilde{W} = \operatorname*{argmax}_{W} \max_{T} P(O|W)P(W|T)^{\lambda_{ST}} P(T)^{\lambda_{RT}}$$

WFST format

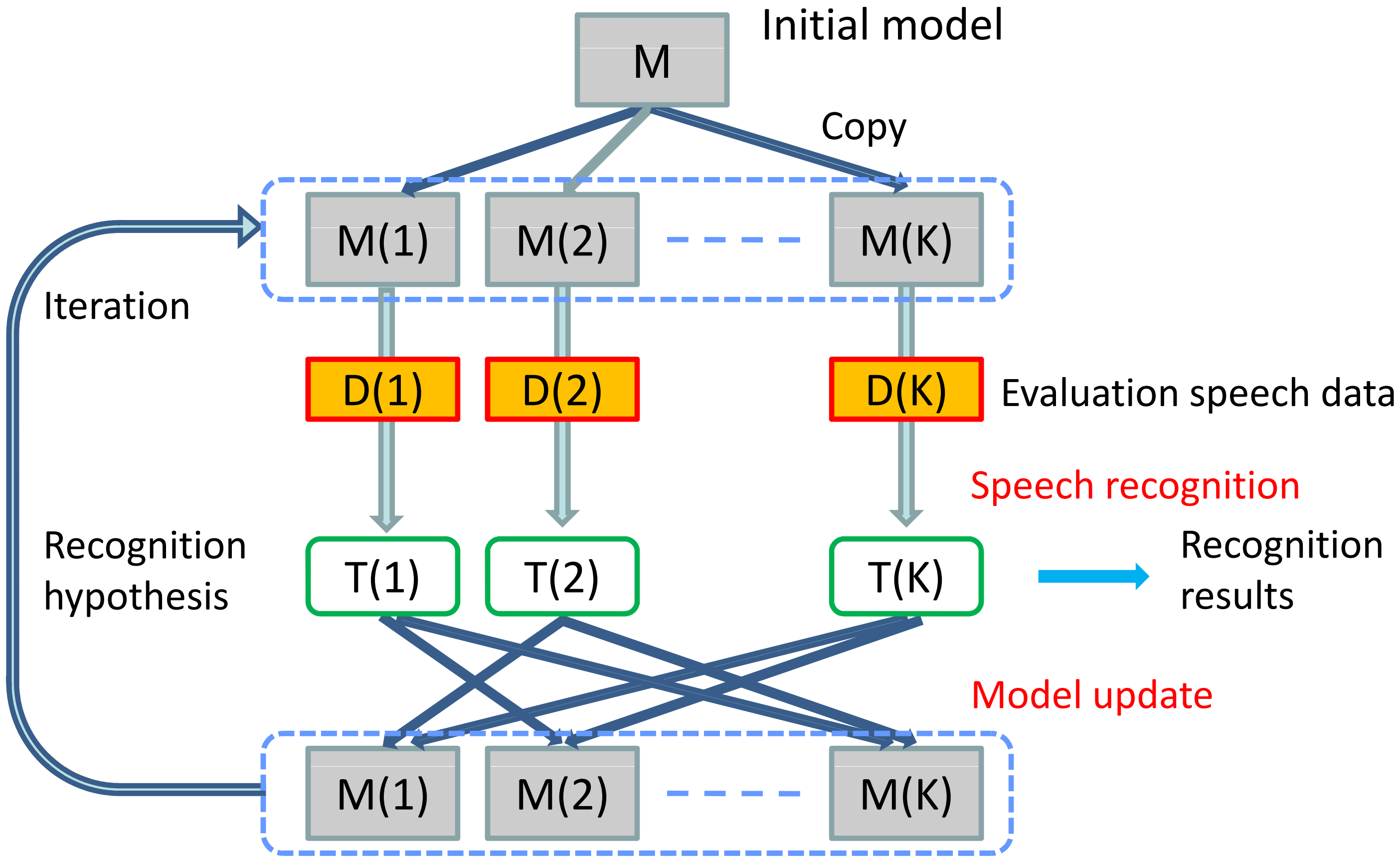
Icelandic speech $\longrightarrow$ $H \circ C \circ L \circ G_{ST} \circ \pi(Tr \circ G_{RT})$ $\longrightarrow$ **Icelandic Text**

- *P*(*O*|*W*): Icelandic acoustic model
- *P*(*W*|*T*): English to Icelandic translation model
- *P*(*T*): English language model

Recognition results



Unsupervised cross-validation (CV) adaptation

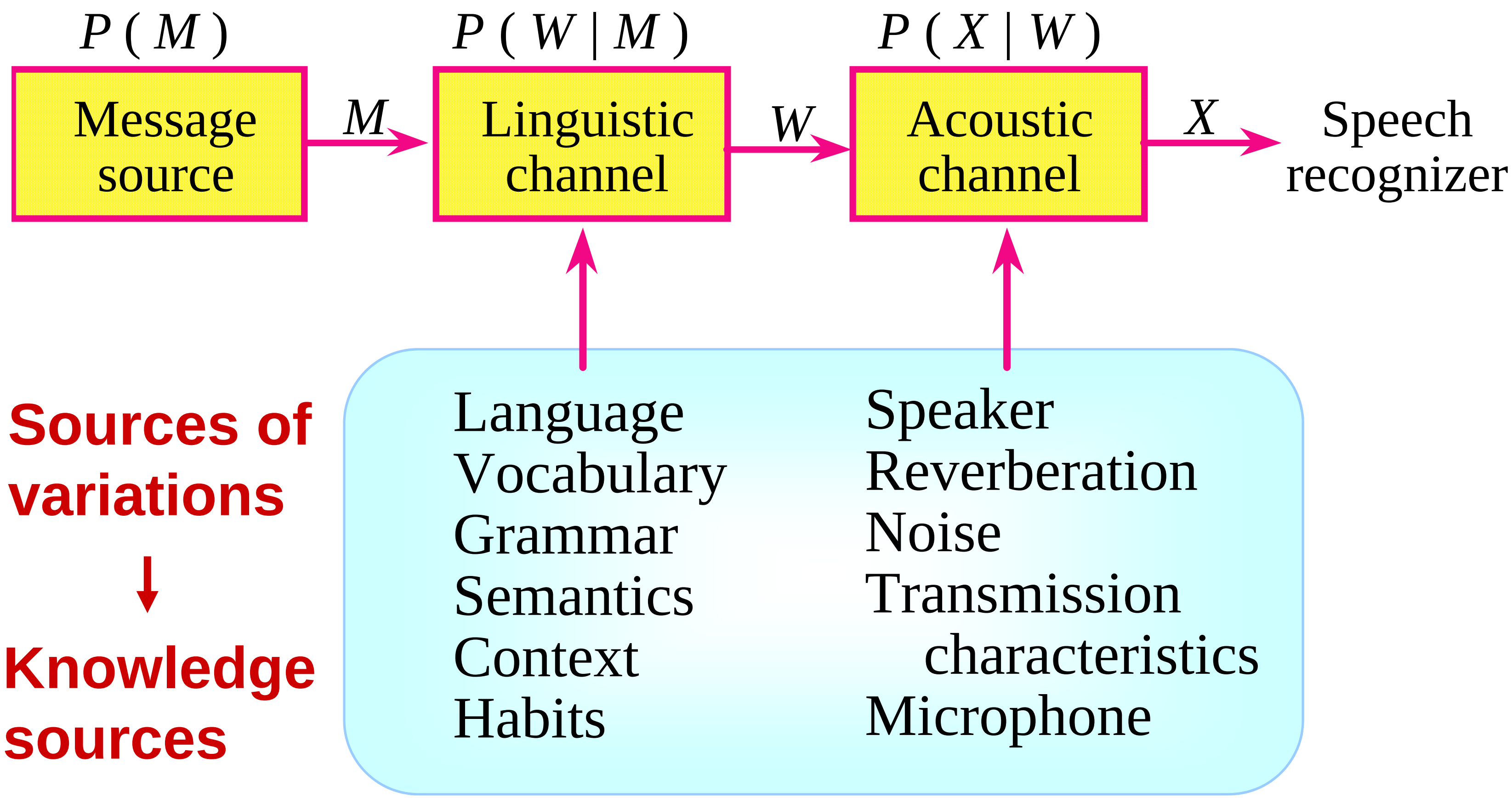


● Reducing the influence of recognition errors by separating the data used for the decoding step and the model update step

Future

- Increasing flexibility and robustness against various sources of variations
- Spoken language comprehension

A communication - theoretic view of speech generation & recognition



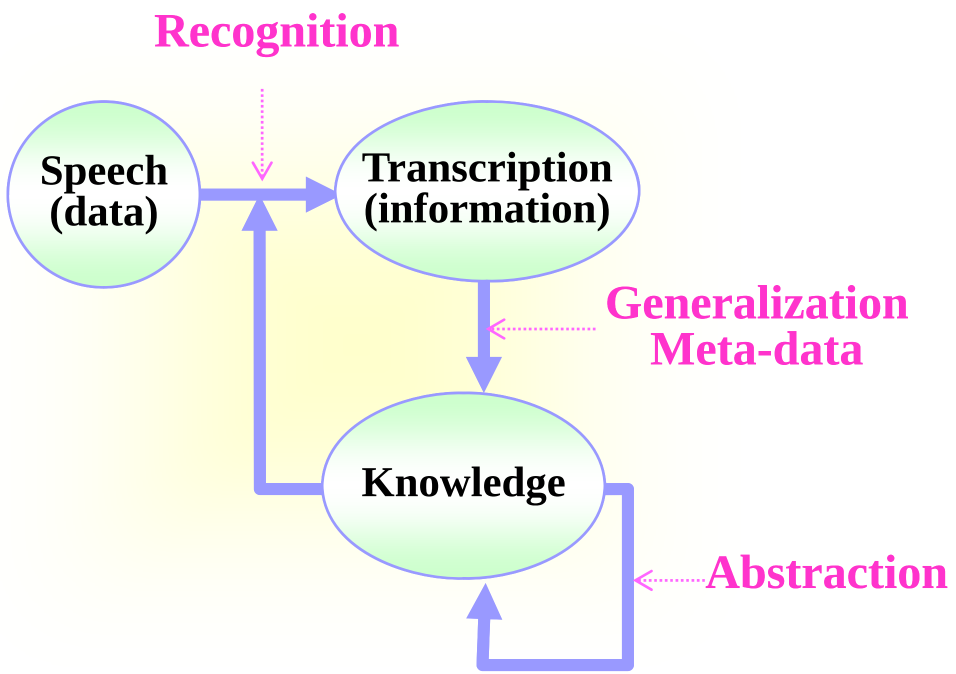
Knowledge sources for speech recognition

Human speech recognition is a matching process whereby an audio signal is matched to existing knowledge (comprehension).

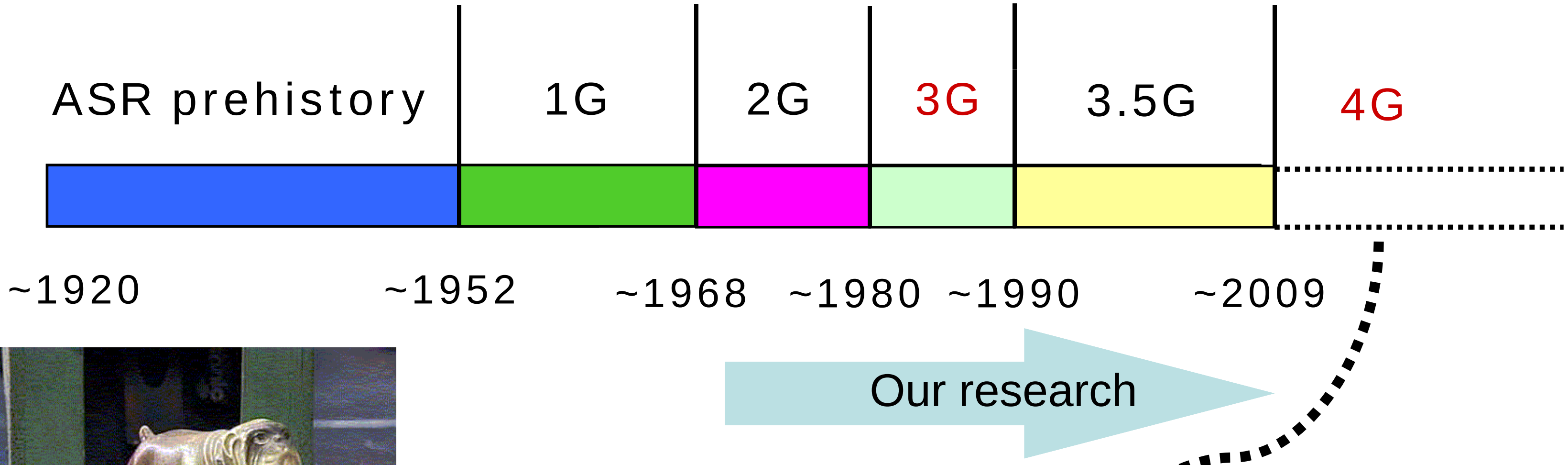
- Knowledge (Meta-data)
  - Domain and topics
  - Context
  - Semantics
  - Speakers
  - Environment, etc.

- Systematization of various related knowledge is crucial

- How to incorporate knowledge sources into the statistical ASR framework



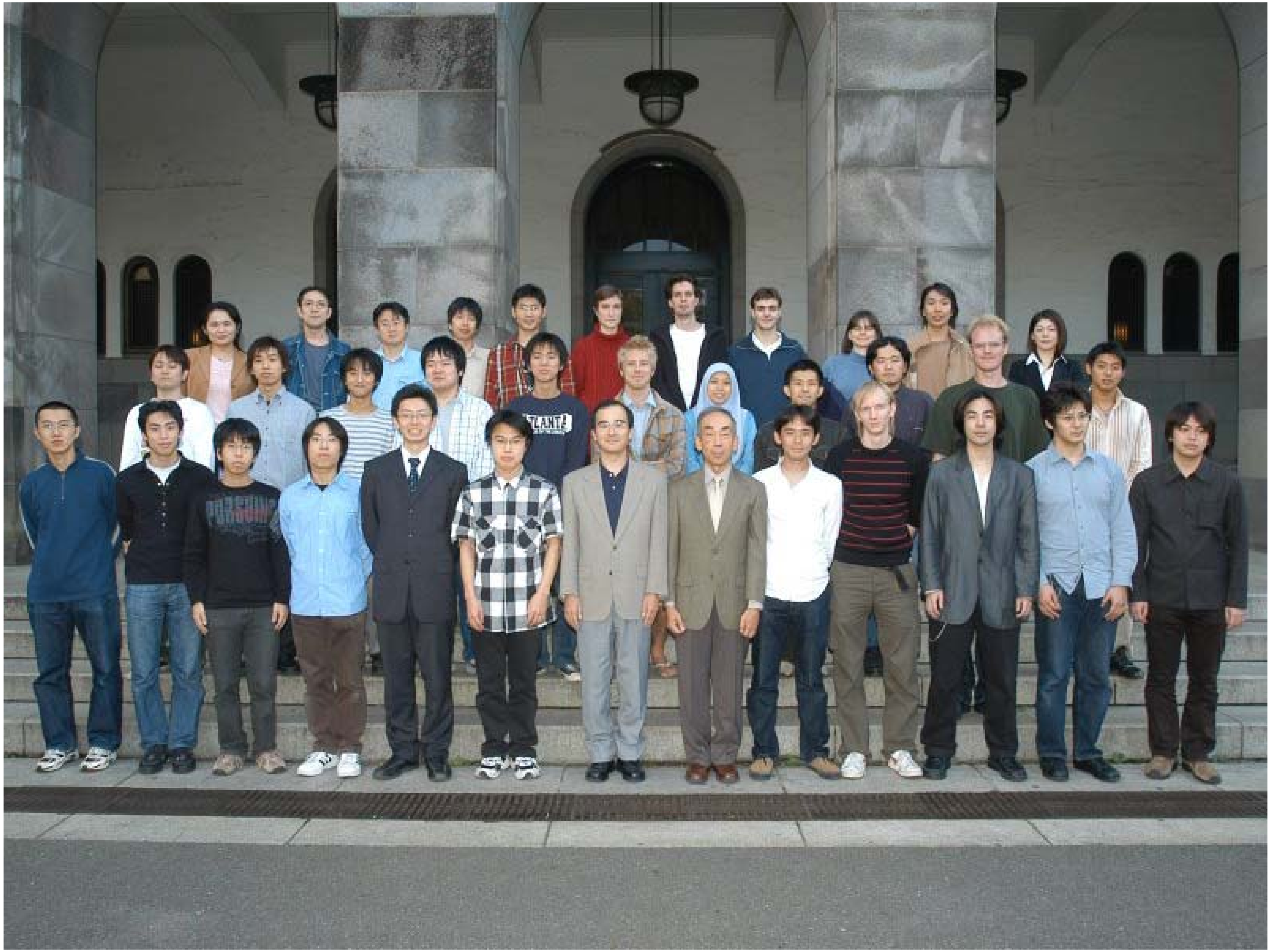
Generations of ASR technology



Extended knowledge processing
"Speech and Intelligence"

Future works

- Grand challenge-1: flexibility and robustness against various acoustic as well as linguistic variations
- Grand challenge-2: spoken language comprehension
- A much greater understanding of the human speech process is required before automatic speech recognition systems can approach human performance.
- Significant advances will come from extended knowledge processing in the framework of statistical pattern recognition.



Thanks to all our present and past colleagues and students at NTT Labs and Tokyo Tech!