

Recent Advances in Automatic Speech Summarization

Sadaoki Furui

**Department of Computer Science
Tokyo Institute of Technology**

Outline

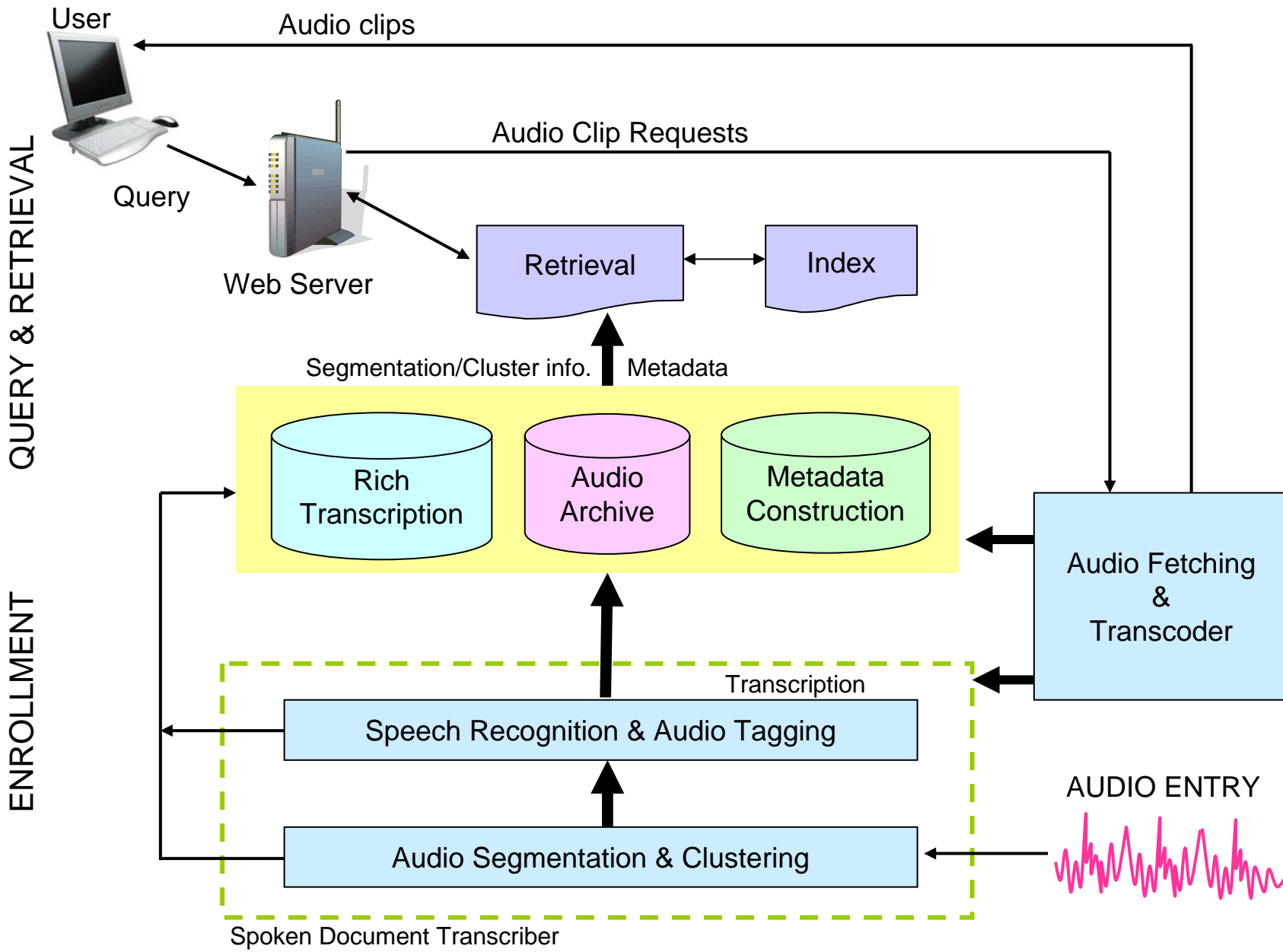
- Introduction
- Speech-to-text & speech-to-speech summarization
- Summarization methods
 - Sentence extraction-based methods
 - Sentence compaction-based methods
 - Combination of sentence extraction and sentence compaction
 - Sentence segmentation
- Evaluation schemes
 - Extrinsic and intrinsic evaluations
 - SumACCY
 - ROUGE
 - Experimental results
- Conclusions

Major speech recognition applications

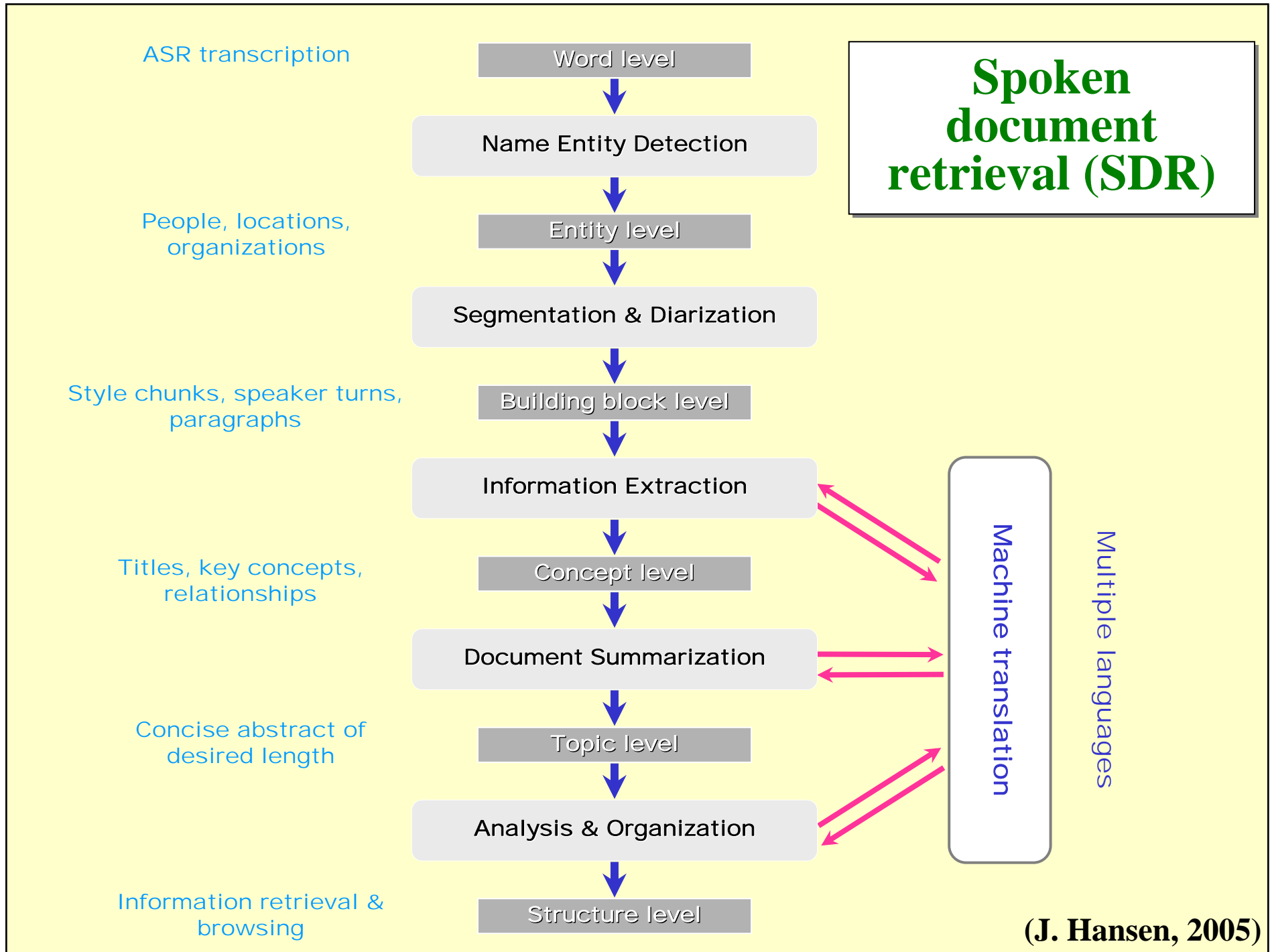
- **Conversational systems** for accessing information services
(e.g. automatic flight status or stock quote information systems)
- Systems for **transcribing, understanding and information extraction** from ubiquitous speech documents
(e.g. broadcast news, meetings, lectures, presentations and voicemails)



Spoken Document Retrieval (SDR)



Spoken document retrieval system at Univ. Colorado Boulder



Speech transcription and summarization for spoken document retrieval (SDR)

- Although **speech** is the most natural and effective method of communication between human beings, it is not easy to quickly review, retrieve and reuse speech documents if they are simply recorded as audio signal.
- Therefore, **transcribing speech** is expected to become a crucial capability for the coming IT era.
- **Speech summarization** which extracts important information and removes redundant and incorrect information is necessary for transcribing spontaneous speech.
- Efficient speech summarization saves time for **reviewing speech documents** and improves the **efficiency of document retrieval**.
- Summarization results can be presented by either **text or speech**.

Classification of speech summarization methods

Audience

- Generic summarization
- User-focused summarization
 - Query-focused summarization
 - Topic-focused summarization

Function

- Indicative summarization
- Informative summarization

Extracts vs. abstracts

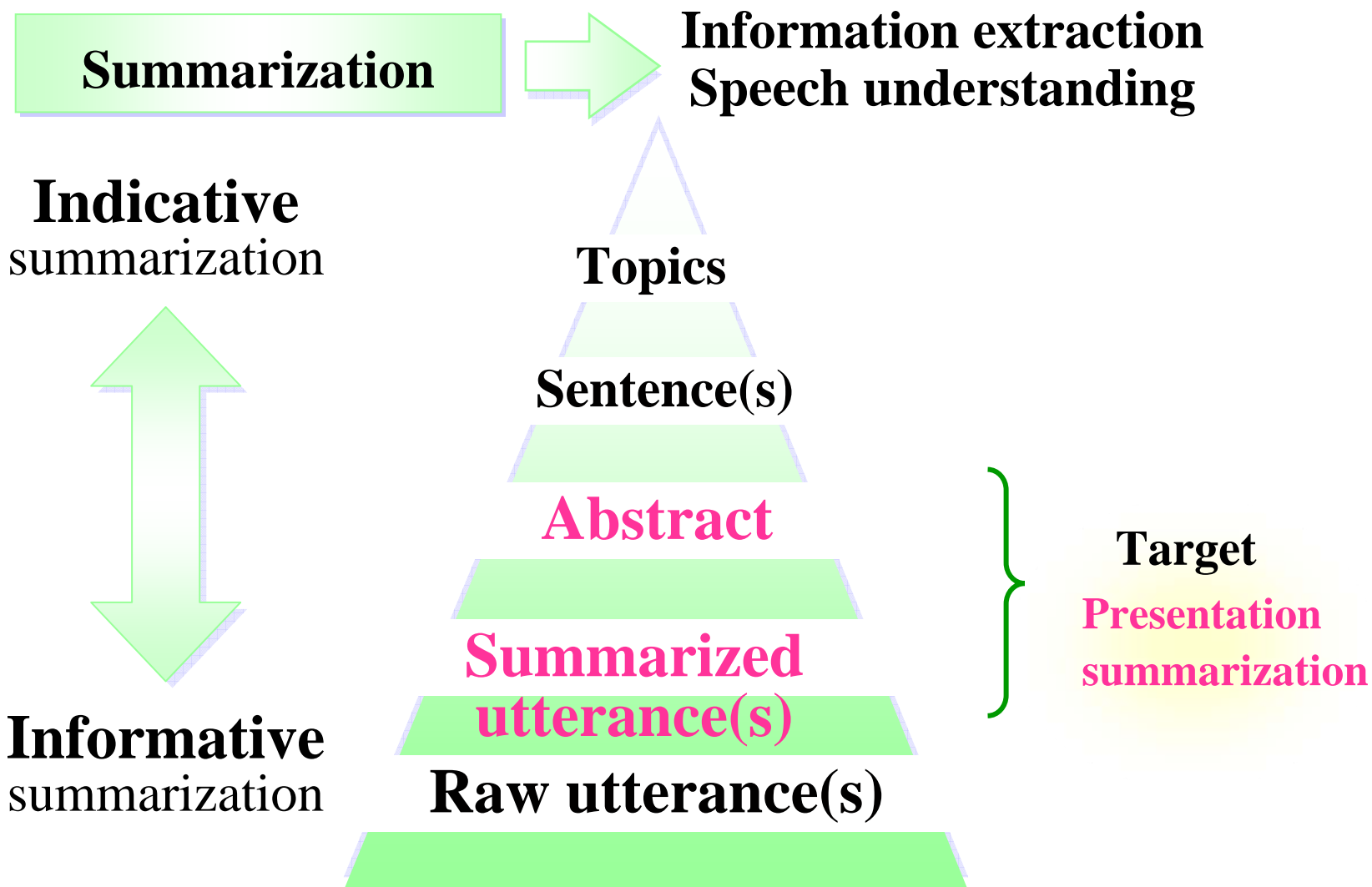
- Extract: consists wholly of portions from the source
- Abstract: contains material which is not present in the source

Output modality

- Speech-to-text summarization
- Speech-to-speech summarization

Single vs. multiple documents

Indicative vs. informative summarization



Fundamental problems with speech summarization

- Disfluencies, repetitions, word fragments, etc.
- Difficulties of sentence segmentation
- More spontaneous parts of speech (e.g. interviews in broadcast news) are less amenable to standard text summarization
- Speech recognition errors

Speech-to-text/speech summarization

Speech-to-text summarization:

- a) The documents can be easily looked through
- b) The part of the documents that is interesting for users can be easily extracted
- c) Information extraction and retrieval techniques can be easily applied to the documents

Speech-to-speech summarization:

- a) Wrong information due to speech recognition errors can be avoided
- b) Prosodic information such as the emotion of speakers that is conveyed only by speech can be presented

Speech-to-speech summarization

- Simply presenting **concatenated speech segments** that are extracted from original speech, or
- **Synthesizing summarized text** using a speech synthesizer.
 - Since state-of-the-art speech synthesizers still cannot produce completely natural speech, **the former method** can easily produce better quality summarizations, and it does not have the problem of synthesizing wrong messages due to speech recognition errors.
 - The major problem is how to avoid unnatural noisy sound caused by the concatenation.

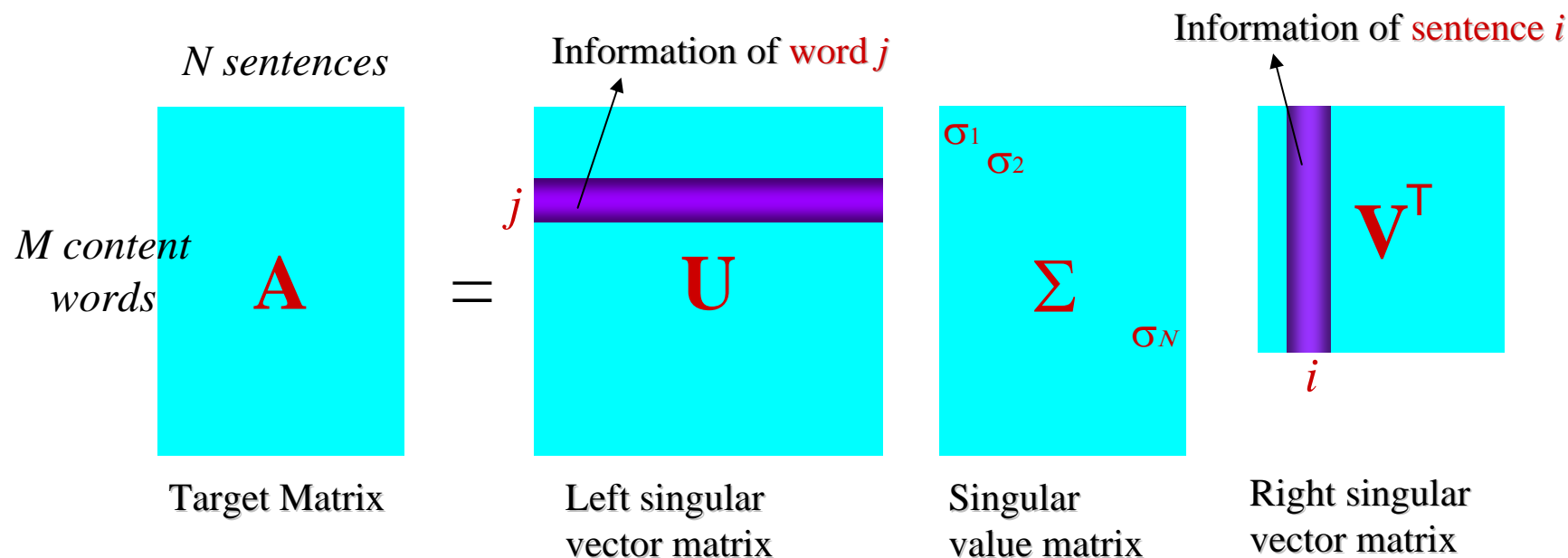
Speech-to-text summarization methods

- Sentence extraction-based methods
 - LSA-based methods
 - MMR-based methods
 - Feature-based methods
- Sentence compaction-based methods
- Combination of sentence extraction and sentence compaction

Speech-to-text summarization methods

- Sentence extraction-based methods
 - LSA-based methods
 - MMR-based methods
 - Feature-based methods
- Sentence compaction-based methods
- Combination of sentence extraction and sentence 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

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

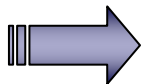
f_{mn} : Number of occurrences of a content word (m) in the sentence (n)

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

LSA-based sentence extraction - 1

- ◆ One of the summarization techniques using the SVD (Gong et al, 2001)
- ◆ Each singular vector represents a salient topic
- ◆ The singular vector with the largest corresponding singular value represents the topic that is the most salient in the presentation speech

$$\mathbf{V}^T = \begin{bmatrix} v_{11} & v_{21} & \cdots & v_{N1} \\ \vdots & \vdots & \vdots & \vdots \\ v_{1k} & v_{2k} & \cdots & v_{Nk} \\ \vdots & \vdots & \vdots & \vdots \\ v_{1N} & v_{2N} & \cdots & v_{NN} \end{bmatrix}$$



Choose a sentence having the largest index within the singular vector k

- ▶ The sentence best describes the topic represented by the singular vector

Extracted sentences best describe the topics represented by the singular vectors and are semantically different from each other.

Drawbacks to the LSA-based method - 1

- Dimensionality is tied to summary length and that good sentence candidates may not be chosen if they do not “win” in any dimension.
- When singular vectors are selected incrementally, as the number of vectors being selected increases, the chances that non-relevant topics get included in a summary also increases.



LSA-based method -2

LSA-based sentence extraction -2

- ◆ 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

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

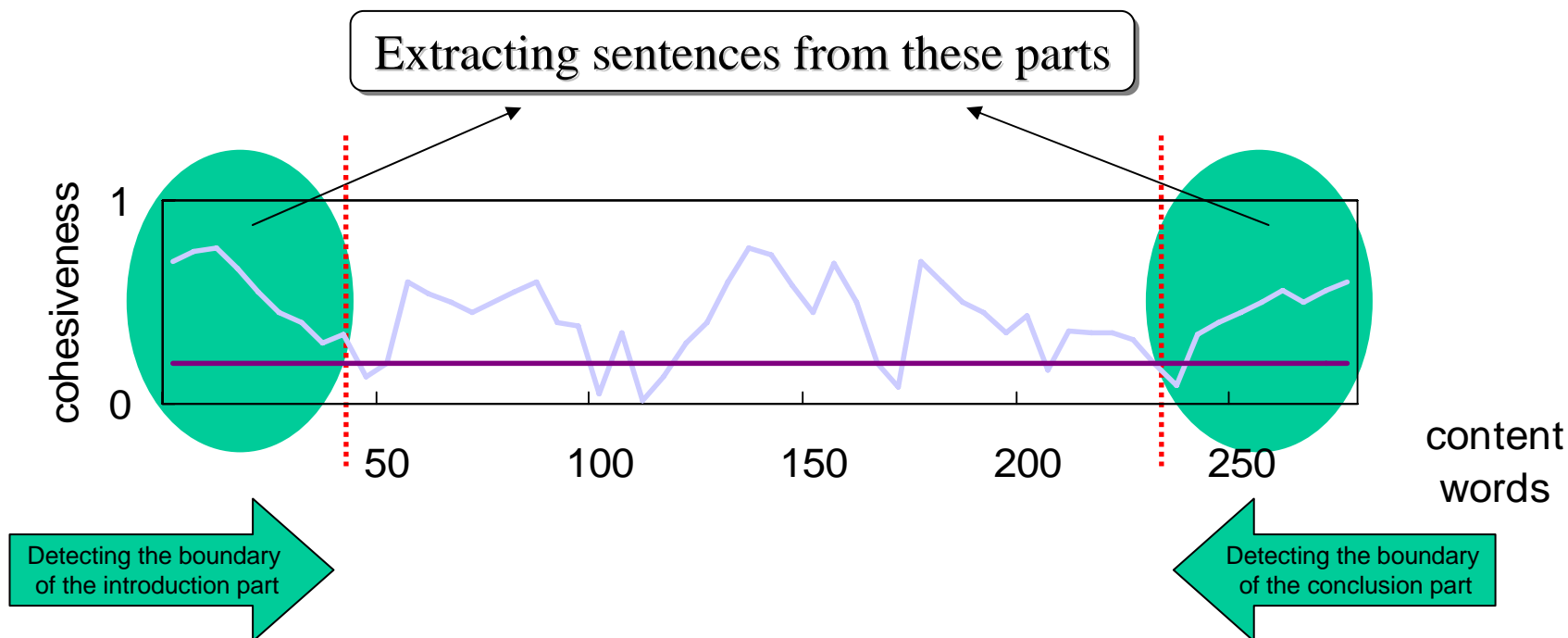
A fixed number of sentences having relatively large sentence scores in the reduced dimensional space are selected.

Sentence extraction from introduction and conclusion parts

Hypothesis : presentation speech consists of introduction, main subjects and conclusion parts

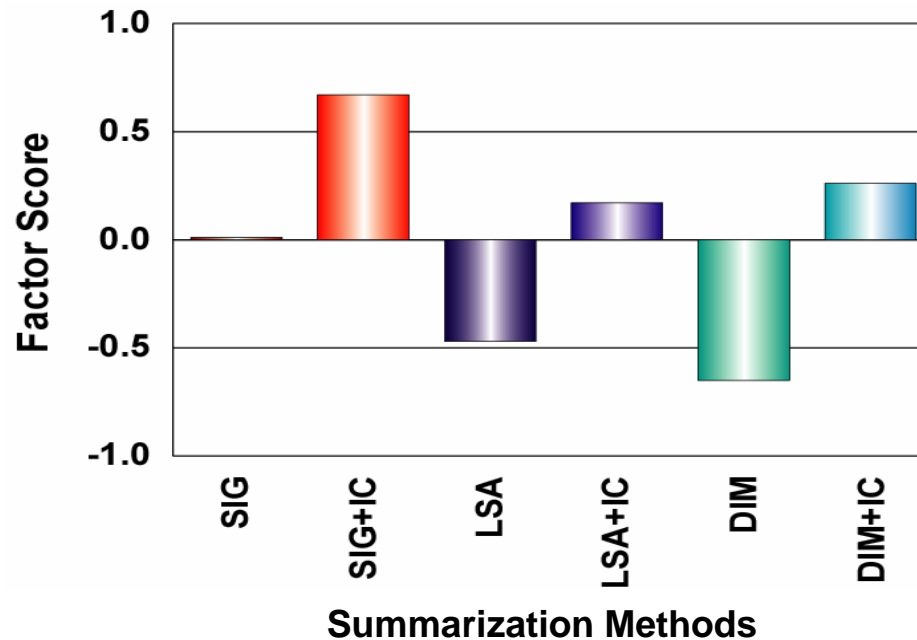
Under the condition of 10% summarization ratio

Human subjects tend to extract sentences from introduction and conclusion parts



Cohesiveness is measured by a cosine value between content word-frequency vectors consisting of a fixed number of content words

Subjective evaluation results represented by the normalized score



SIG: sentence extraction by a significant score (amount of information)
LSA: LSA-based method-1
DIM: LSA-based method-1
IC: beginning and ending period weighting

- 180 automatic summaries (30 presentations X 6 summarization methods) were evaluated by 12 human subjects in terms of **ease of understanding** and **appropriateness as summaries** in five levels.
- Converted into **factor scores** to normalize subjective differences.
- **IC** method significantly improves summarization performance.
- Difference between **SIG+IC** and **DIM+IC** is not significant.

MMR-based method

- Vector-space model of text retrieval
- Particularly applicable to query-based and multi-document summarization
- Chooses sentences via a weighted combination of their relevance to a query (or for generic summaries, their general relevance) and their redundancy with sentences that have already been extracted, both derived using cosine similarity
- MMR score for a given sentence S_i in the document:

$$Sc^{MMR}(i) = \lambda(Sim(S_i, D)) - (1 - \lambda)(Sim(S_i, Summ))$$

D : Average document vector

$Summ$: Average vector from the set of sentences already selected

λ : Trade off between relevance and redundancy (annealed)

Sim : Cosine similarity between documents

Feature-based method

- Textual features
 - Named entities (person, organization and place names)
 - Mean and maximum TF-IDF scores
 - LSA sentence score
 - Topic significance scores and term entropy obtained through PLSA
 - Confidence score
- Structural and discourse features
 - Structural features (sentence position, speaker type, etc.)
 - Discourse features (number of new nouns in each sentence, etc.)
- Prosodic features
 - F0, energy, and duration (mean, standard deviation, minimum, maximum, range, slope, etc.)
 - Speaking rate

An example of feature-based important sentence extraction method

Sentence with N words $W = w_1, w_2, \dots, w_N$

Sentence extraction score

$$S(W) = \frac{1}{N} \sum_{i=1}^N \left\{ L(w_i) \right.$$

$$+ \lambda_I I(w_i)$$

$$+ \lambda_C C(w_i) \left. \right\}$$

Linguistic score

Linguistic correctness
(Bigram/Trigram)

Significance (topic) score

Important information extraction
(Amount of information)

Confidence score

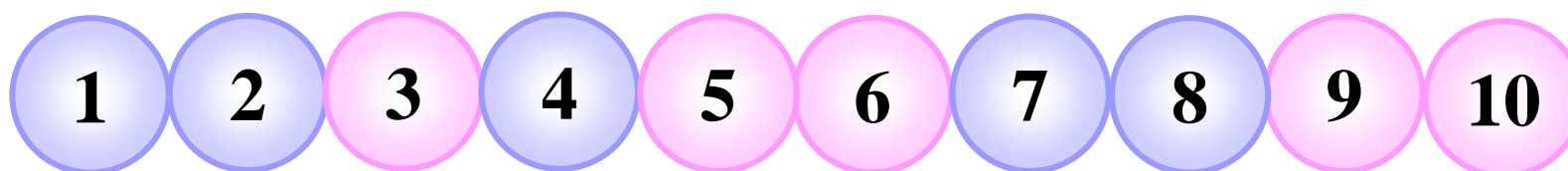
Recognition error exclusion
(Acoustic & linguistic reliability)

Speech-to-text summarization methods

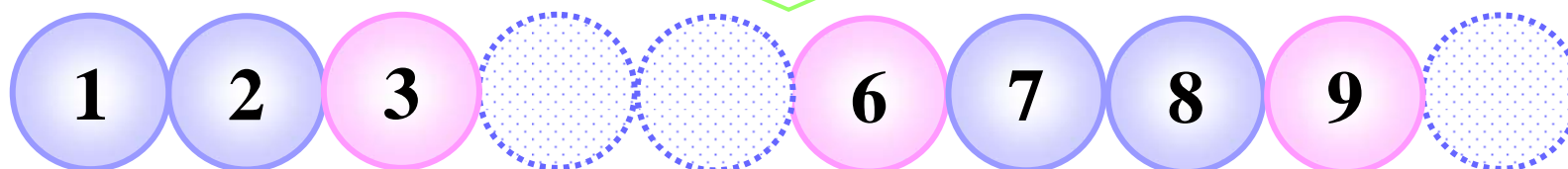
- Sentence extraction-based methods
 - LSA-based methods
 - MMR-based methods
 - Feature-based methods
- Sentence compaction-based methods
- Combination of sentence extraction and sentence compaction

Sentence compaction

Each transcribed utterance



A set of words is extracted



Summarized (compressed) sentence

Word extraction score

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

Score

$$S(V^M) = \sum_{m=1}^M \left\{ \begin{array}{l} L(v_m | \dots v_{m-1}) \\ + \lambda_I I(v_m) \\ + \lambda_C C(v_m) \\ + \lambda_T T_r(v_m) \end{array} \right.$$

Linguistic score

Linguistic correctness
(Bigram/Trigram)

Significance (topic) score

Important information extraction
(Amount of information)

Confidence score

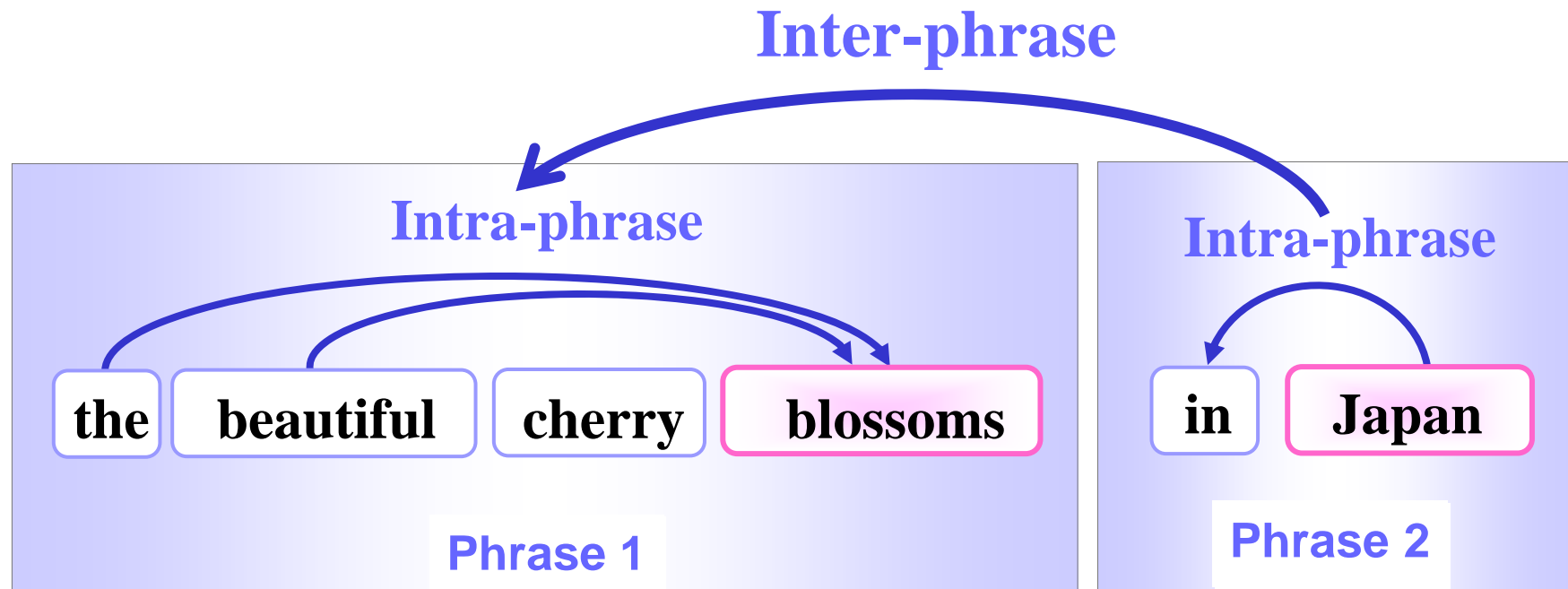
Recognition error exclusion
(Acoustic & linguistic reliability)

Word concatenation score

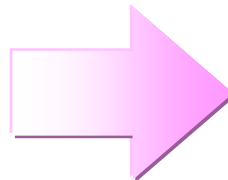
Semantic correctness
(Word dependency probability)

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**

Word concatenation score based on SDCFG

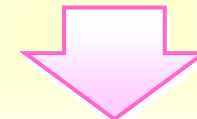
Word dependency probability

If the dependency structure between words is **deterministic**,



0 or 1

If the dependency structure between words is **ambiguous**,

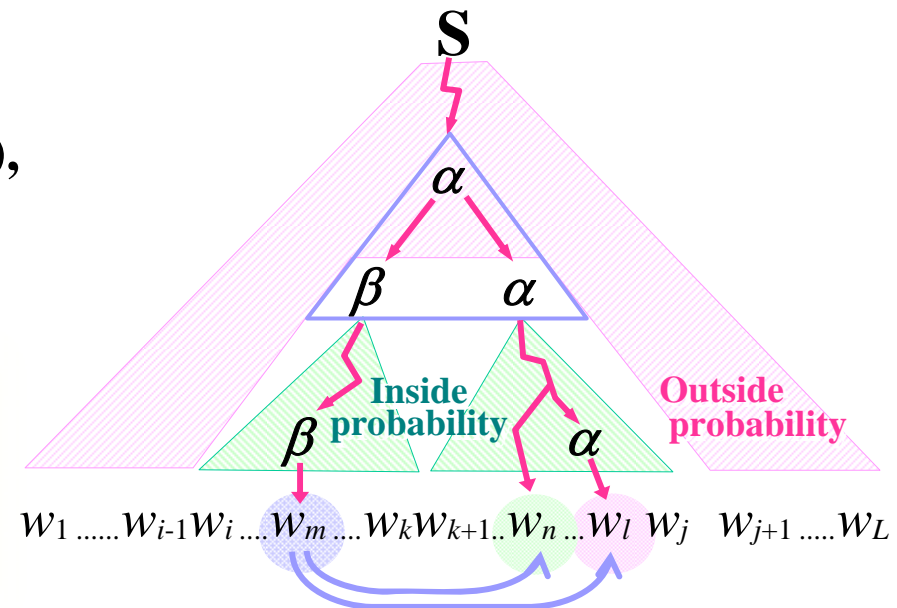


SDCFG
(Stochastic DCFG)

The dependency probability between w_m and w_l , $d(w_m, w_l, i, k, j)$, is calculated using Inside-Outside probability based on SDCFG.

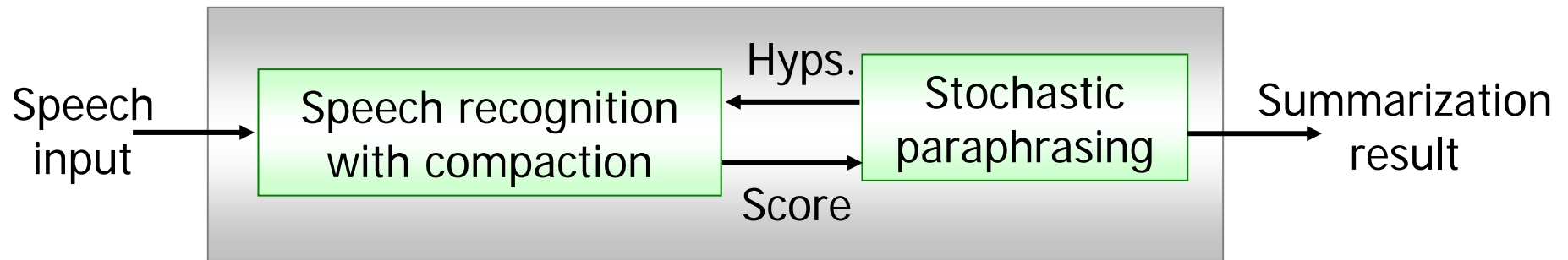
$$T_r(w_m, w_n)$$

$$= \log \sum_{i=1}^m \sum_{k=m}^{n-1} \sum_{j=n}^L \sum_{l=n}^j d(w_m, w_l, i, k, j)$$



S : Initial symbol, α, β : Non-terminal symbol, w : Word

Integration of ASR and sentence compaction by WFST (Hori et al., NTT)

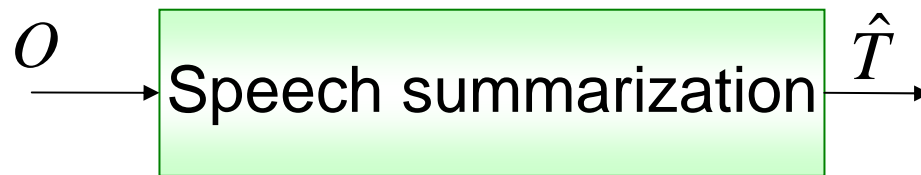
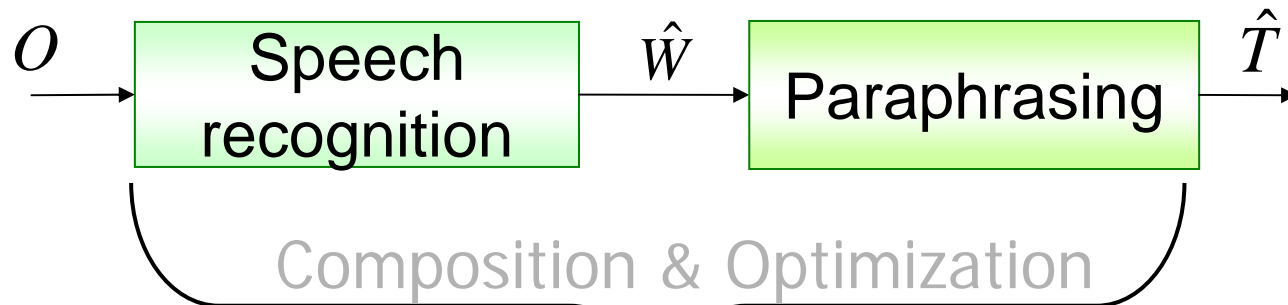


- **Speech Recognition with compaction**
 - Transcribe speech signal & extract important phrases excluding recognition errors
- **Paraphrasing**
 - Translate spoken language into written language

➡ *Weighted Finite-State Transducer*

Integration of speech recognition and paraphrasing

$$\hat{W} = \arg \max_W P(O|W)P(W) \quad \hat{T} = \arg \max_T P(\hat{W}|T)P(T)$$



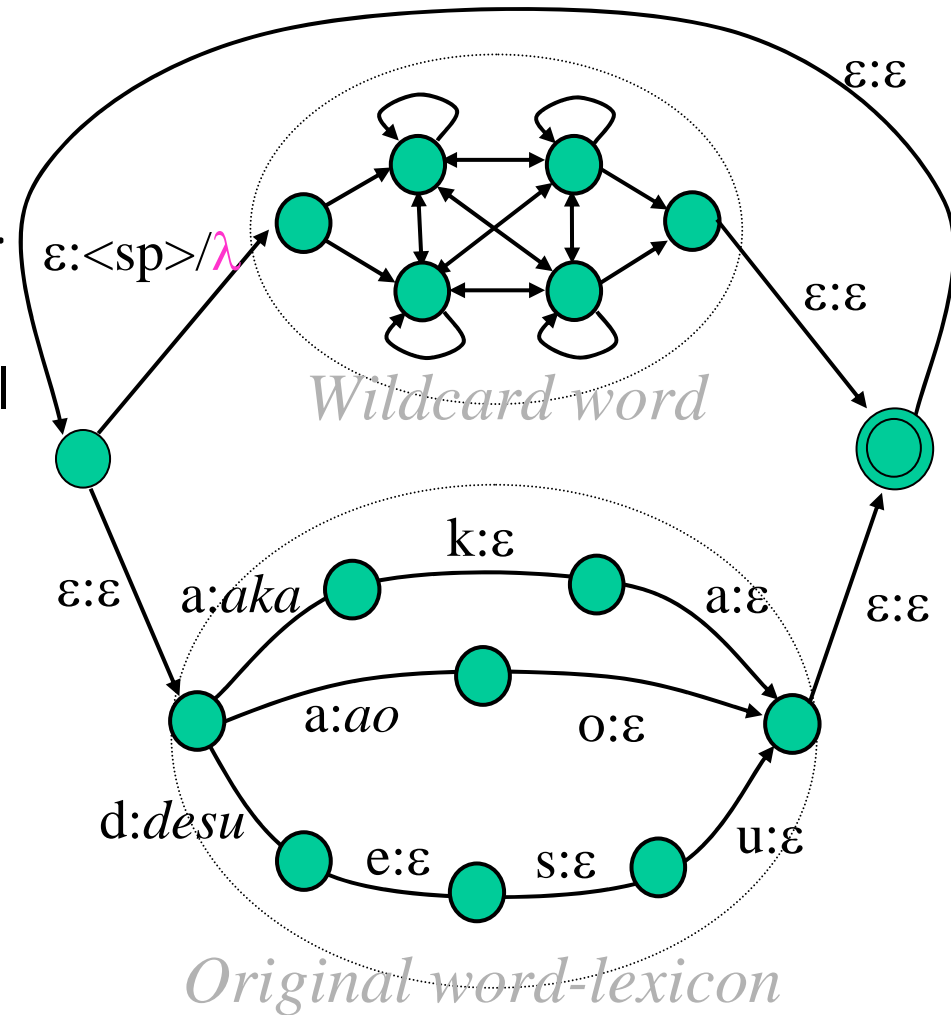
O : Feature vector seq.
 W : Source word seq.
 T : Target word seq.

$$\hat{T} = \arg \max_T \max_W P(O|W)P(W|T)P(T)$$

Extended Lexicon WFST for sentence compaction

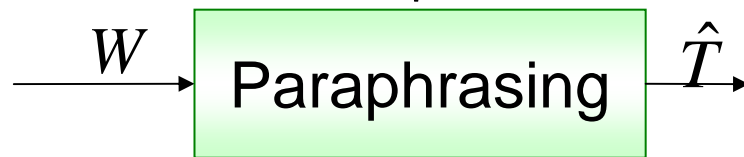
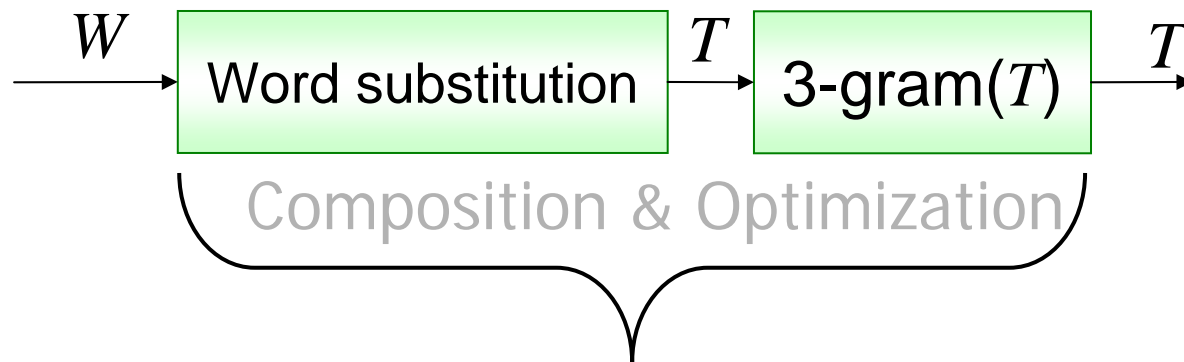
Wildcard word:

- Accept an arbitrary phone seq. weighted by phone 2-grams
- Output an inter-phrase symbol (<sp>)
- Control summarization ratio by the **penetration weight**



WFST for paraphrasing

Translation of spoken language into written language



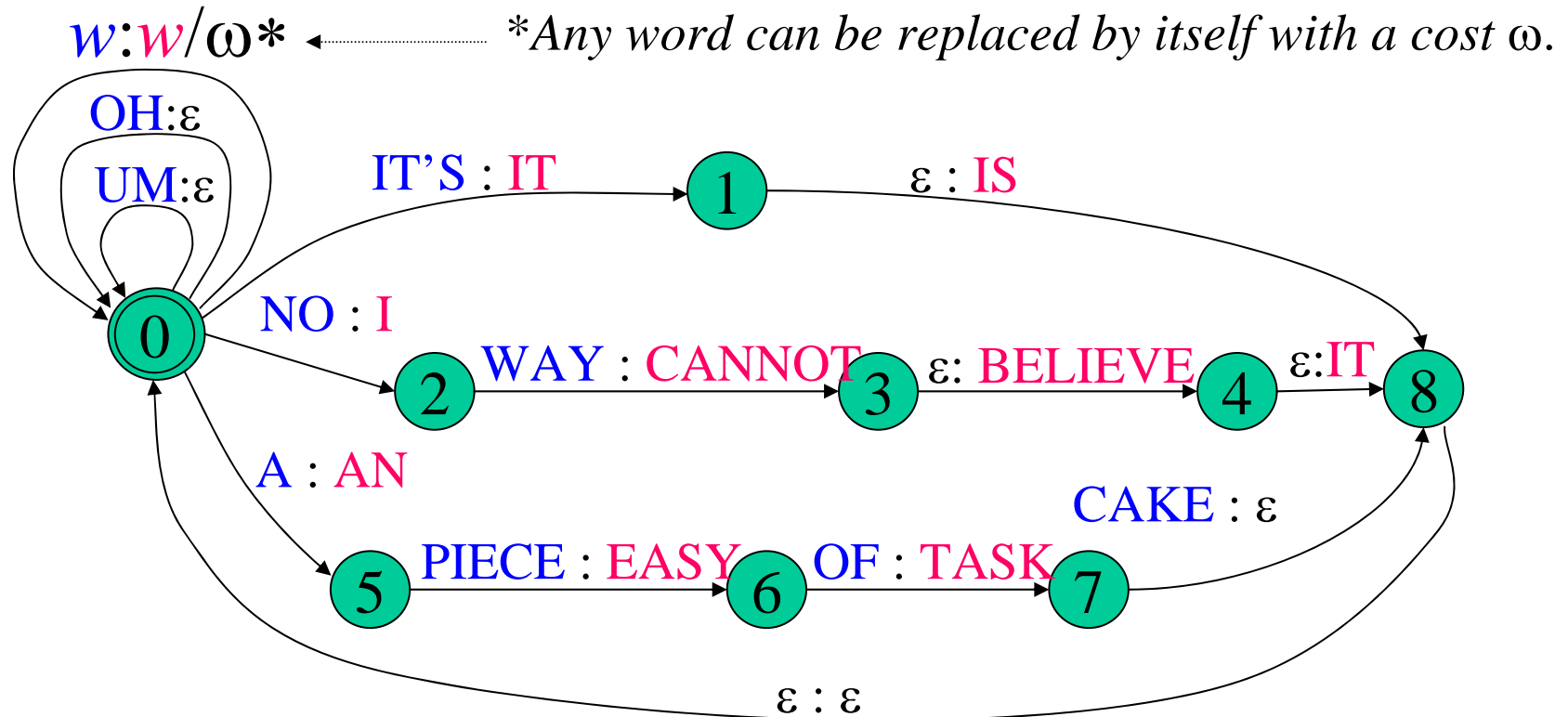
$$\hat{T} = \underset{T}{\operatorname{argmax}} P(W|T)P(T)$$
$$\approx \underset{T}{\operatorname{argmax}} \delta(W, T)P(T)$$

W : Source word seq.

T : Target word seq.

$\delta(W, T)$: Word substitution model

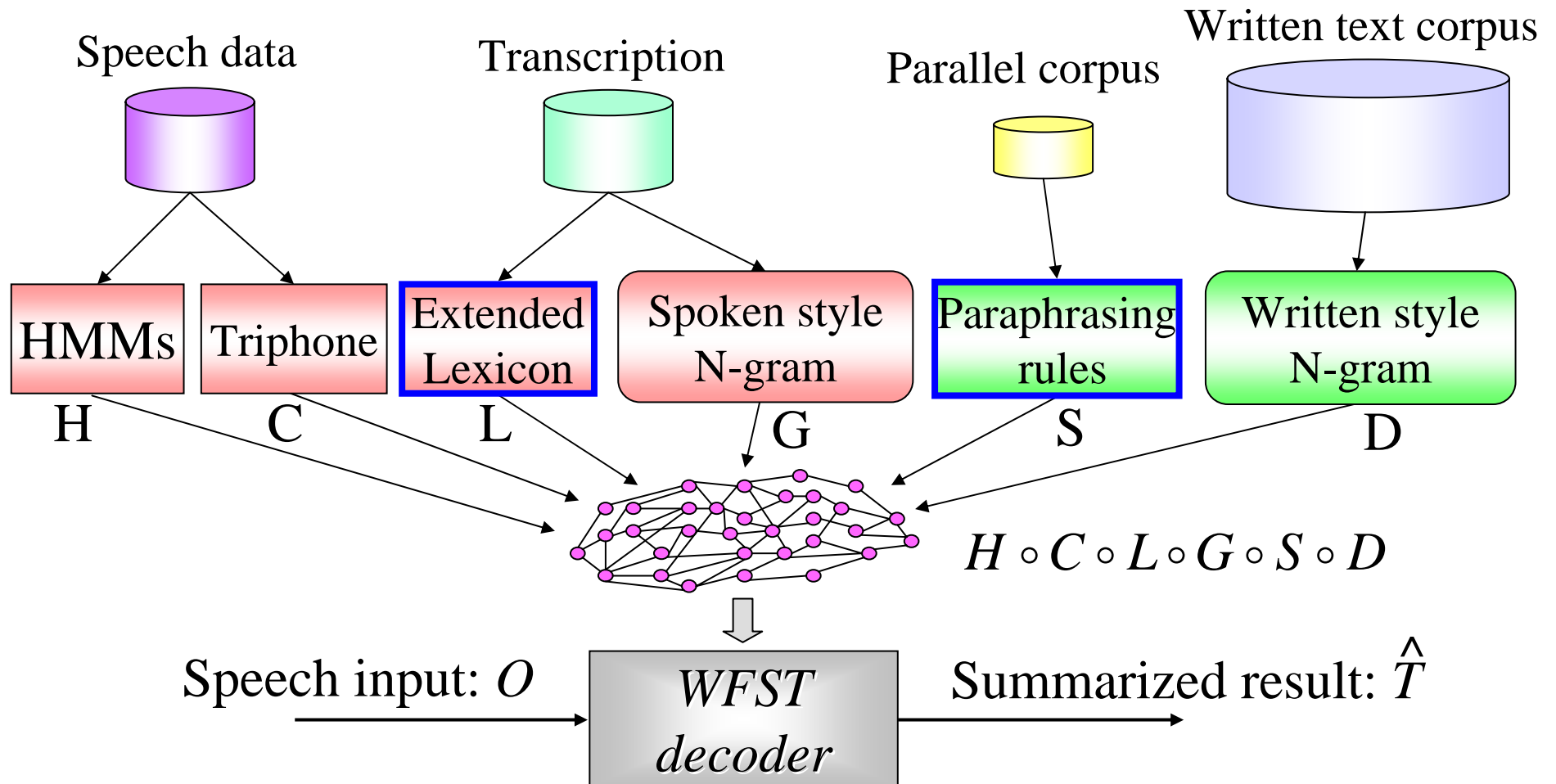
WFST for paraphrasing



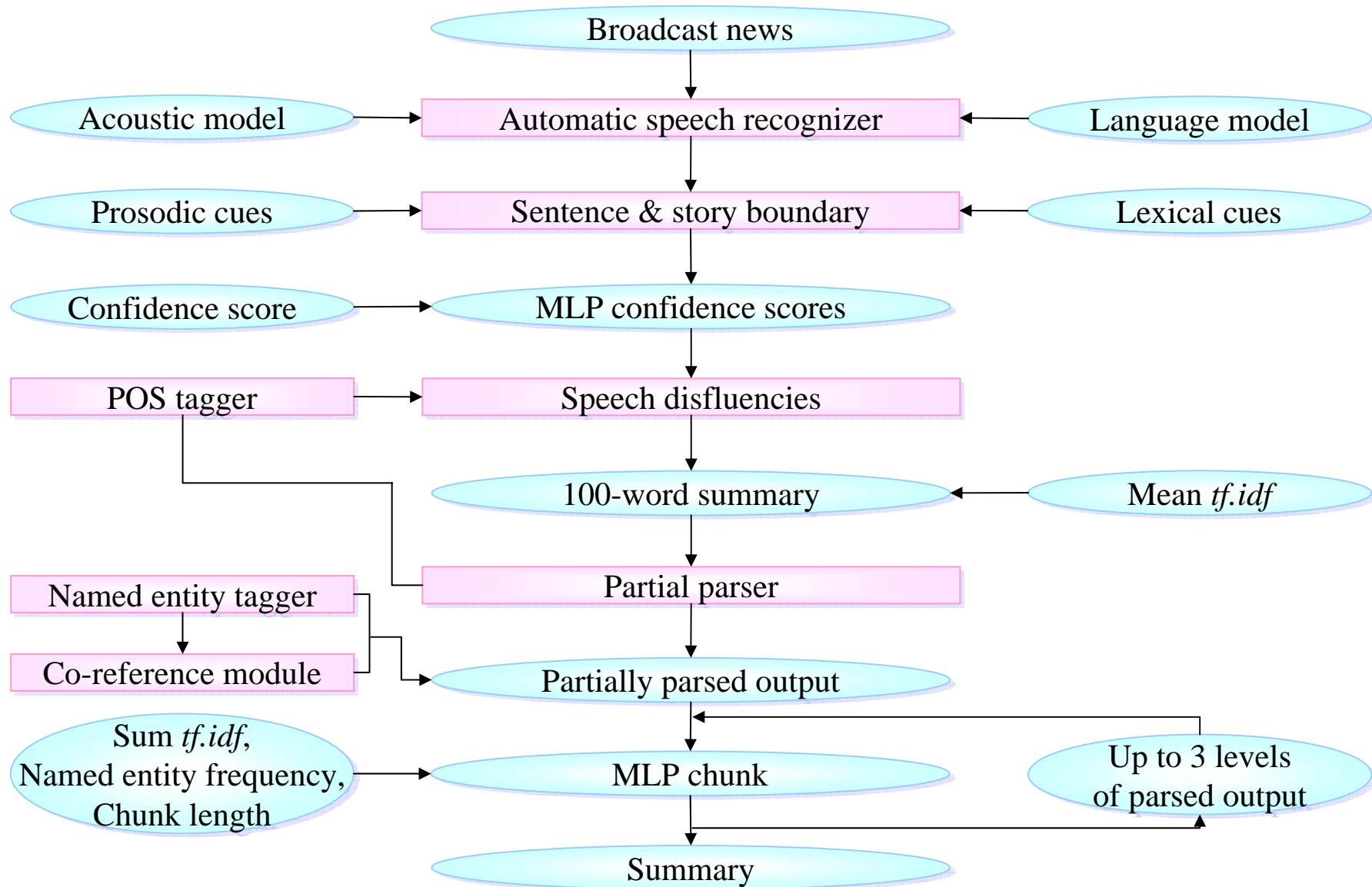
Ex. *OH, NO WAY. IT'S A PIECE OF CAKE.*

⇒ *I CANNOT BELIEVE IT. IT IS AN EASY TASK.*

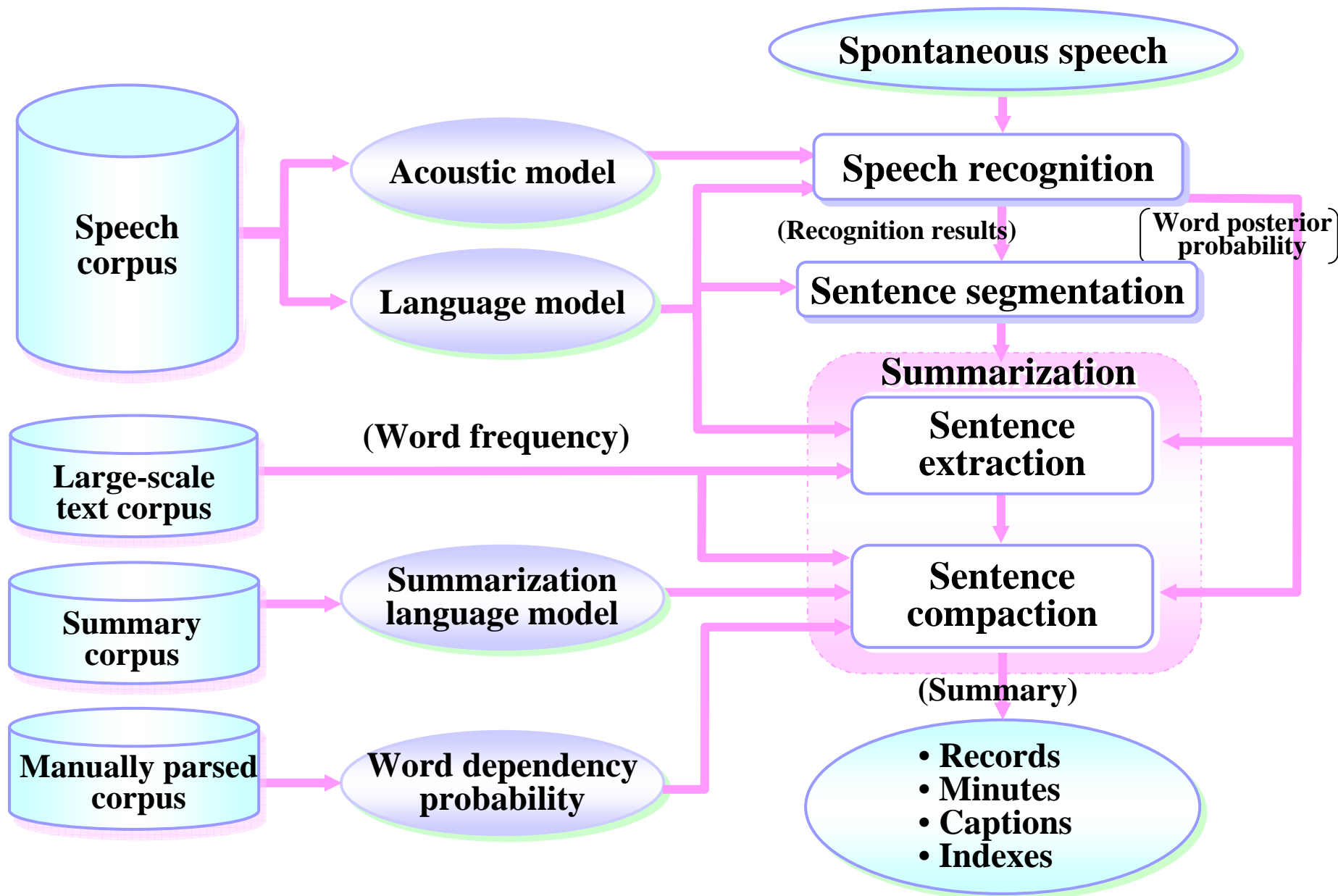
Speech summarization using WFST



A multi-stage compaction approach to broadcast news summarization (by Kolluru)

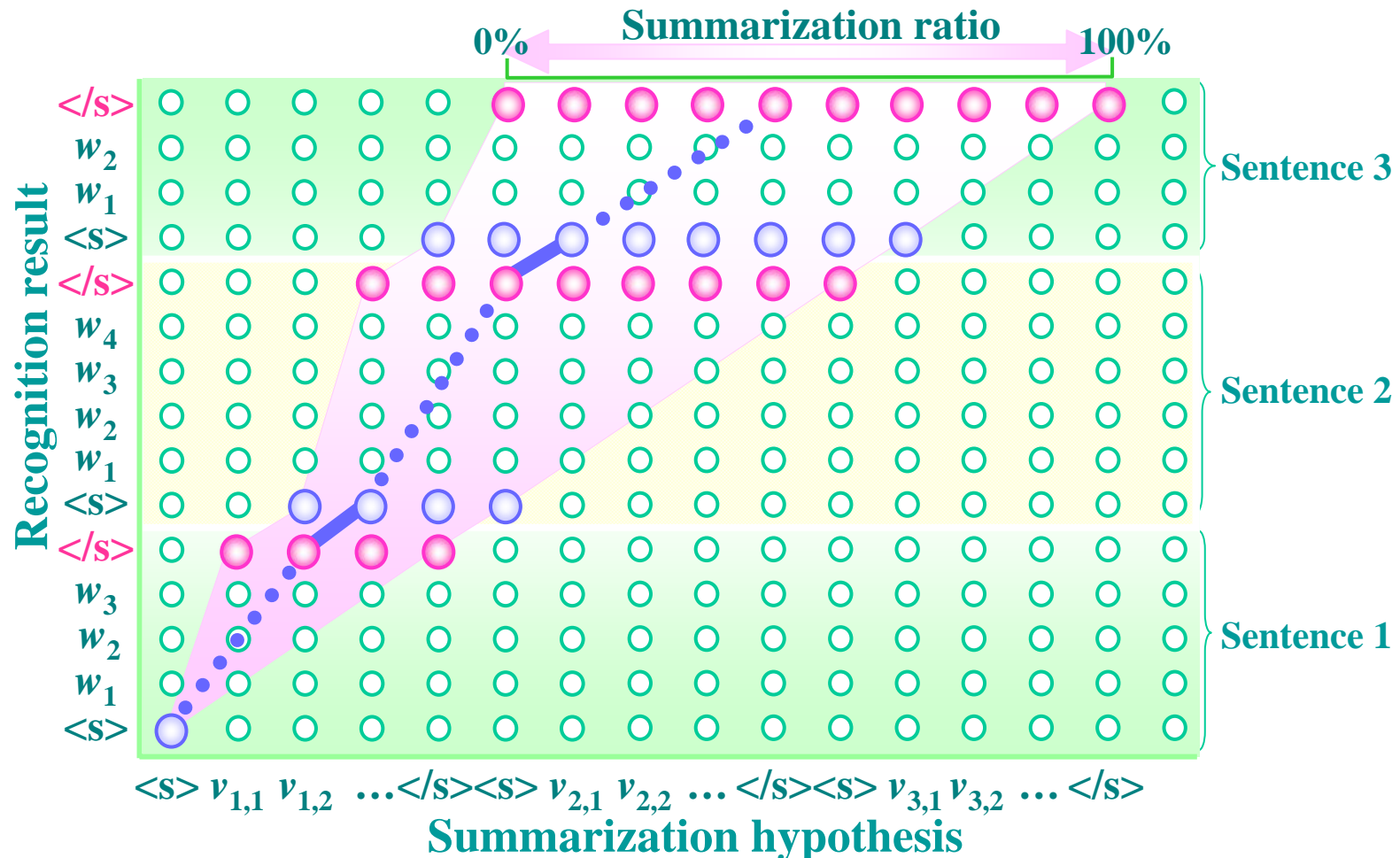


Combination of sentence extraction and compaction



2-stage dynamic programming for summarizing multiple sentences

- * Initial and terminal symbols cannot be skipped.
- * Word concatenation score is not applied to the sentence boundaries.



Sentence segmentation

- Speech recognition results have no punctuation or proper segmentation.
- **Readability and usability** of transcripts can be significantly improved by segmenting text into logical units such as sentences.
- Segmentation has a significant effect on the further processing of speech, such as **information extraction, topic detection and summarization.**
- **Prosodic and N-gram features** have been employed.
- Due to **poor grammatical structure, unclear definition of sentences, disfluencies, and incorrectly recognized words**, sentence segmentation of speech is still difficult.

Evaluation schemes


- **Quality of a summary** depends on how it is used, how readable an individual finds, and what information an individual thinks should be included.
- **Extrinsic evaluation:** assessed in a task-based setting; e.g. information browsing and access interface (ideal, but time-consuming and expensive)
- **Intrinsic evaluation:** assessed in a task-independent setting (normally employed)
- **Subjective evaluation:** too costly
- **Objective evaluation:** essential (using manual summaries, which vary according to human subjects, as targets)

Objective evaluation methods

- **Summarization accuracy** using a network merging manual summaries (SumACCY) (Hori et al., 2001)
- Summarization accuracy weighted by the majority of manual summaries (WSumACCY) (Hori et al., 2003)
- Summarization accuracy using individual manual summary (SumACCY-E) (Hirohata et al., 2004)
- **N-gram precision** (Hori et al., 2000)
- **Number of overlapping n-grams** (ROUGE-N) (Lin et al., 2003)
- **Sentence recall/precision** (Kitade et al., 2004)

Summarization accuracy

SumACCY

- ◆ Variations of manual summarization results are **merged into a word network**
- ◆ The word network is considered to approximately express all possible correct summarization covering subjective variations
- ◆ Word accuracy of automatic summarization is calculated as the summarization accuracy **using the word network**
- ◆ The variations are too large at 10% summarization ratio compared to 50%  Inappropriate summaries

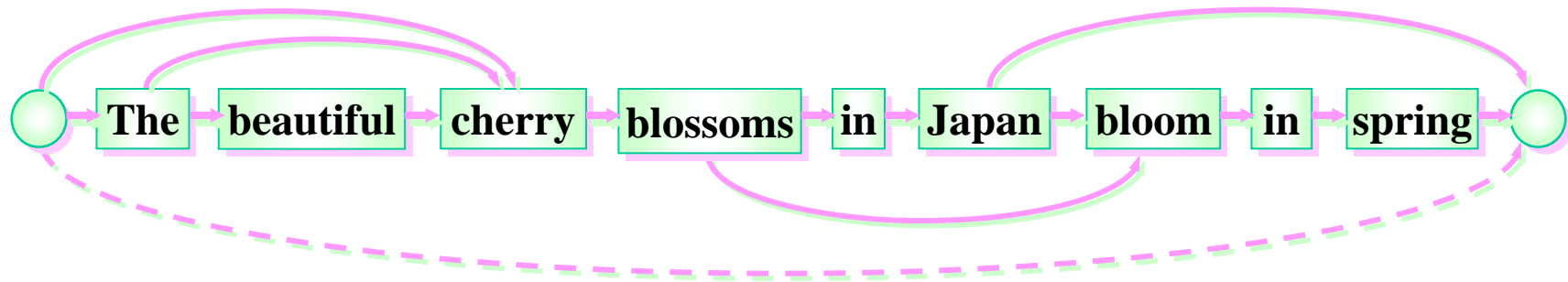
SumACCY-E

- ◆ Word accuracy of automatic summarization is calculated **using manual summaries individually** (not using a network)
 - ▶ SumACCY-E/max : Largest score of the word accuracy
 - ▶ SumACCY-E/ave : Average score of the word accuracy

Summarization accuracy (SumACCY)

Human summaries are merged into a single network.

- The network approximately covers **all possible correct summaries** including subjective variations.



SumACCY is defined as word accuracy based on a word string, extracted from the word network, that is closest to the automatic summarization result.

$$\text{SumACCY} = \{ \text{Len} - (\text{Sub} + \text{Ins} + \text{Del}) \} / \text{Len} * 100 [\%]$$

Len: number of words in the most similar word string in the network

Sub: number of substitution errors

Ins: number of insertion errors

Del: number of deletion errors

ROUGE-N

- ◆ ROUGE-N : **N-grams recall** between an automatic summary and a set of manual summaries
- ◆ N-grams: 1-grams, 2-grams and 3-grams
- ◆ ROUGE-N is computed as follows:

$$\text{ROUGE-N} = \frac{\sum_{S \in S_H} \sum_{g_n \in S} C_m(g_n)}{\sum_{S \in S_H} \sum_{g_n \in S} C(g_n)}$$

S_H : A set of manual summaries

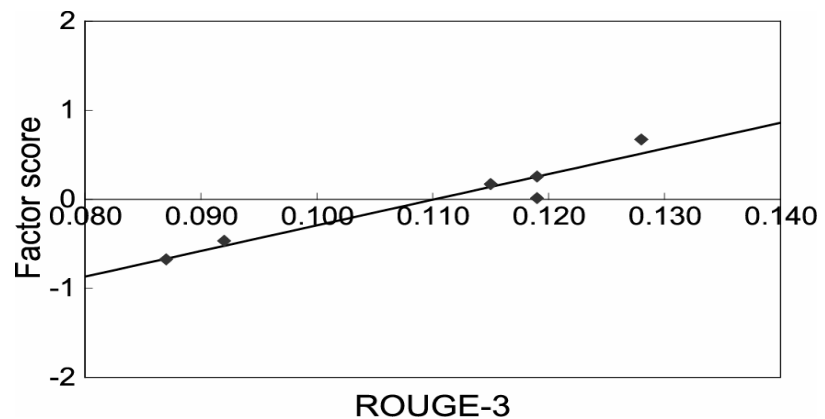
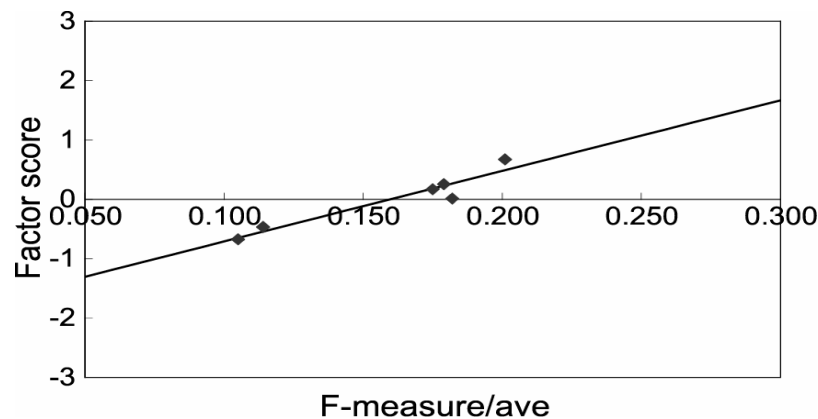
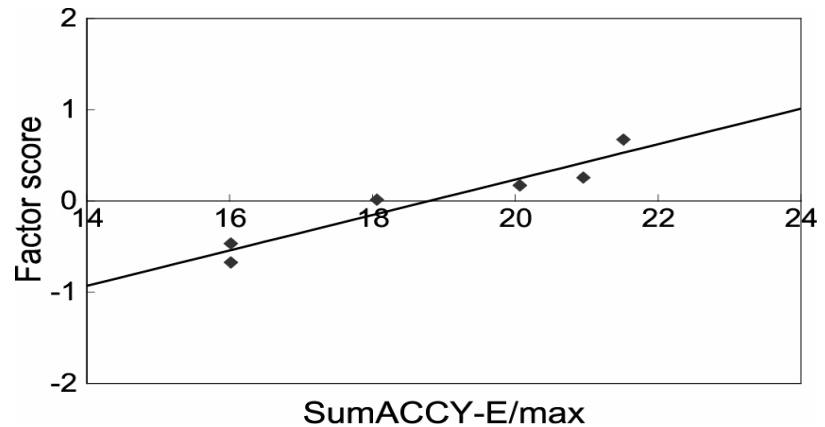
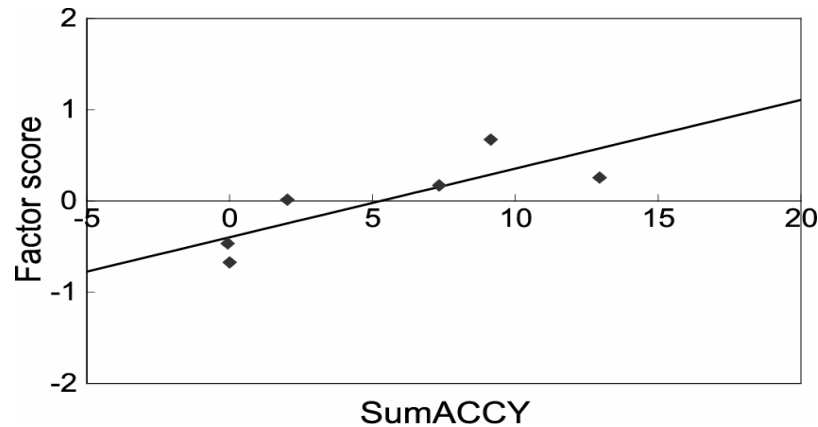
S : Individual summary

g_n : N-gram

$C(g_n)$: Number of g_n in the manual summary

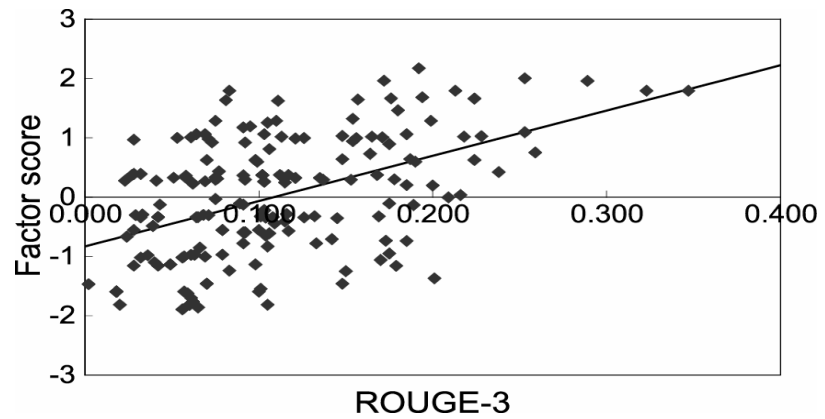
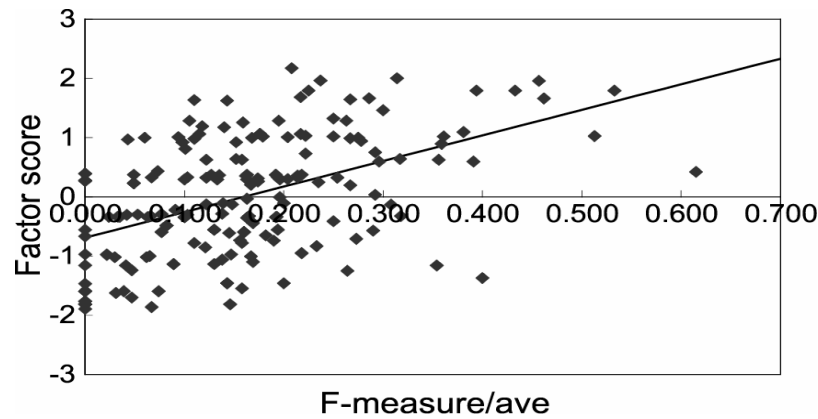
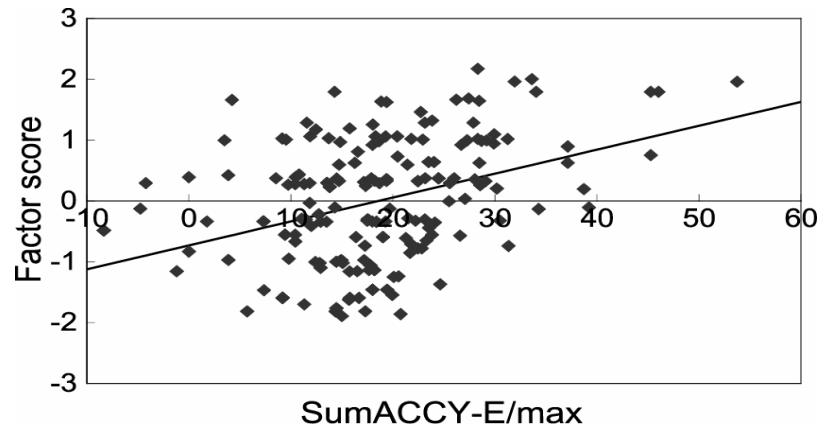
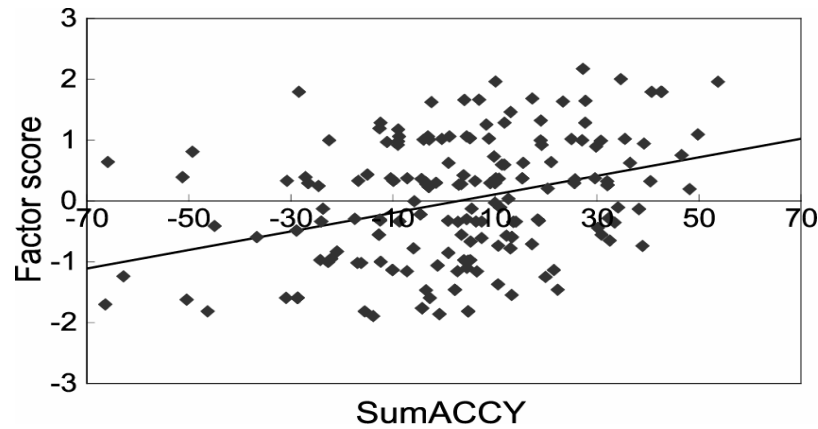
$C_m(g_n)$: Number of co-occurrences of g_n in the manual summary and the automatic summary

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)



Conclusions

- Although various **automatic speech summarization techniques** have been proposed and tested, their performance is still much worse than that of manual summarization.
- In order to build really useful speech summarization systems applicable to real applications, we definitely need **more efficient and speech-focused techniques, including sentence (utterance) segmentation methods**.
- It remains to be determined through further experiments by researchers using various corpora whether or not the **objective evaluation measures** that have been proposed correlate well with human judgments. There still exists large room for improvement in the objective measures.
- It is also necessary to **evaluate summaries extrinsically within the context of applications**, instead of only using intrinsic evaluation methods.