Toward Spontaneous Speech Recognition and Understanding

Sadaoki Furui

Tokyo Institute of Technology
Department of Computer Science
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552 Japan
Tel/Fax: +81-3-5734-3480
furui@cs.titech.ac.jp
http://www.furui.cs.titech.ac.jp/

0205-03

Outline

- Fundamentals of automatic speech recognition
- Acoustic modeling
- Language modeling
- Database (corpus) and task evaluation
- Transcription and dialogue systems
- Spontaneous speech recognition
- Speech understanding
- Speech summarization

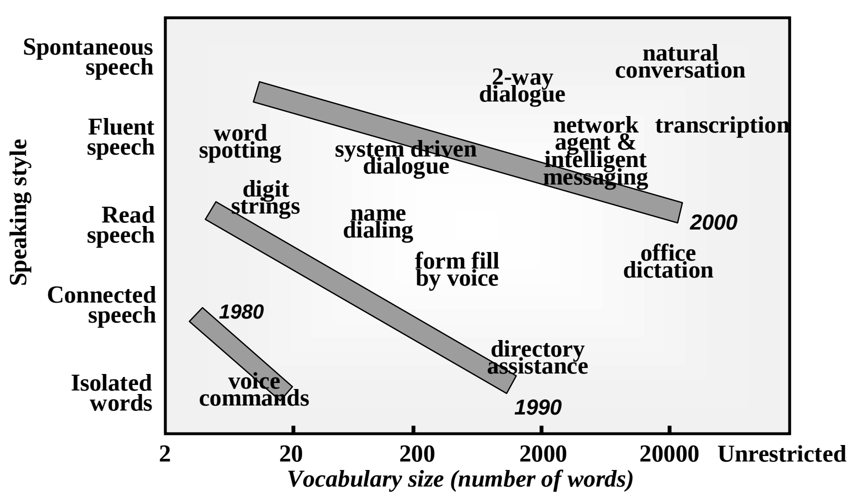
0205-03

Outline

- Fundamentals of automatic speech recognition
- Acoustic modeling
- Language modeling
- Database (corpus) and task evaluation
- Transcription and dialogue systems
- Spontaneous speech recognition
- Speech understanding
- Speech summarization

0010-11

Speech recognition technology



0201-01

Categorization of speech recognition tasks

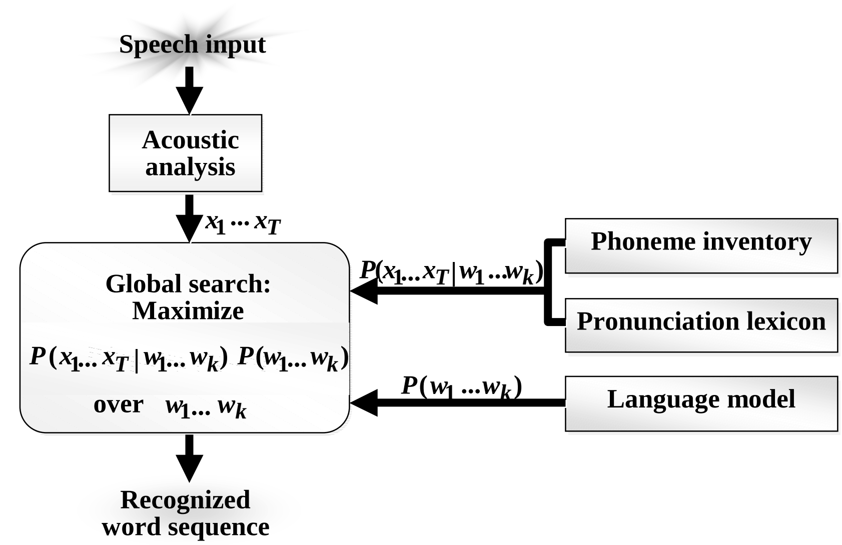
	Dialogue	Monologue
Human to human	(Category I) Switchboard, Call Home (Hub 5), meeting task	(Category II) Broadcasts news (Hub 4), lecture, presentation, voice mail
Human to machine	(Category III) ATIS, Communicator, information retrieval, reservation	(Category IV) Dictation

Major speech recognition applications

- **Conversational systems** for accessing information services
  - Robust conversation using wireless handheld/hands-free devices in the real mobile computing environment
  - Multimodal speech recognition technology
- Systems for **transcribing, understanding and summarizing** ubiquitous speech documents such as broadcast news, meetings, lectures, presentations and voicemails

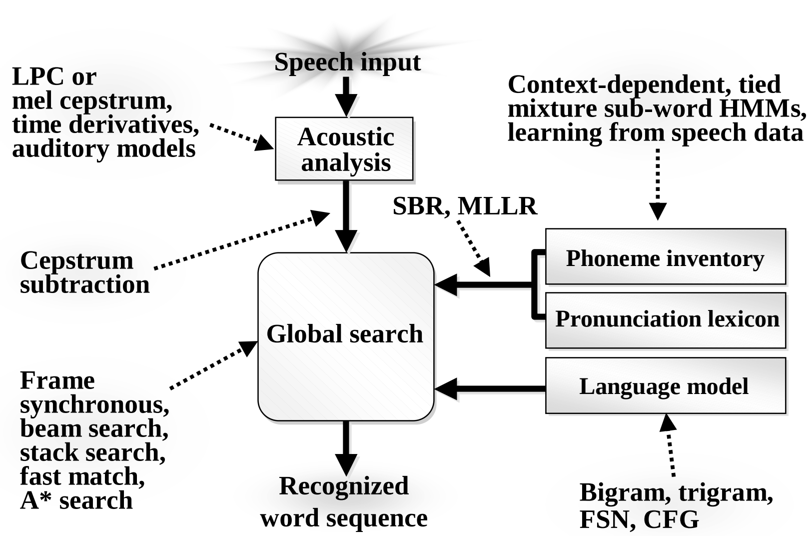
0010-12

Mechanism of state-of-the-art speech recognizers



0010-13

State-of-the-art algorithms in speech recognition

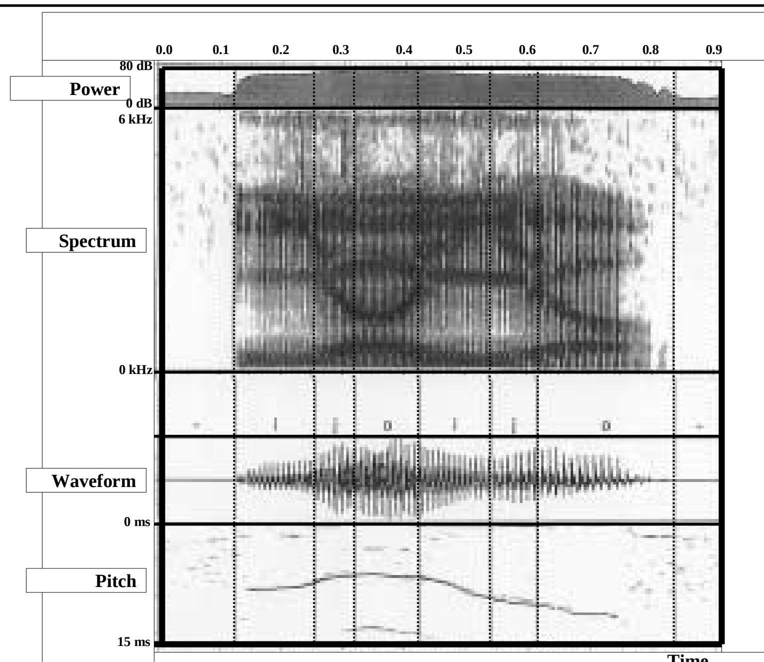


0205-03

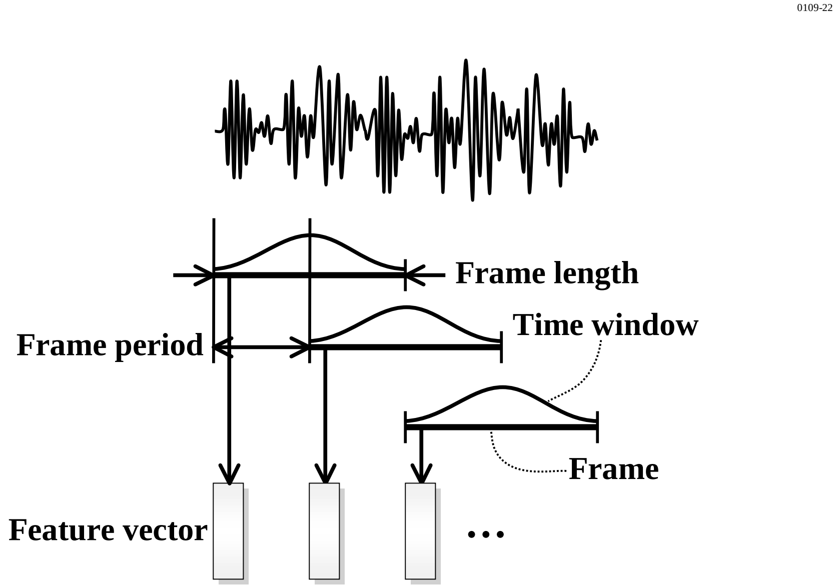
Outline

- Fundamentals of automatic speech recognition
- Acoustic modeling
- Language modeling
- Database (corpus) and task evaluation
- Transcription and dialogue systems
- Spontaneous speech recognition
- Speech understanding
- Speech summarization

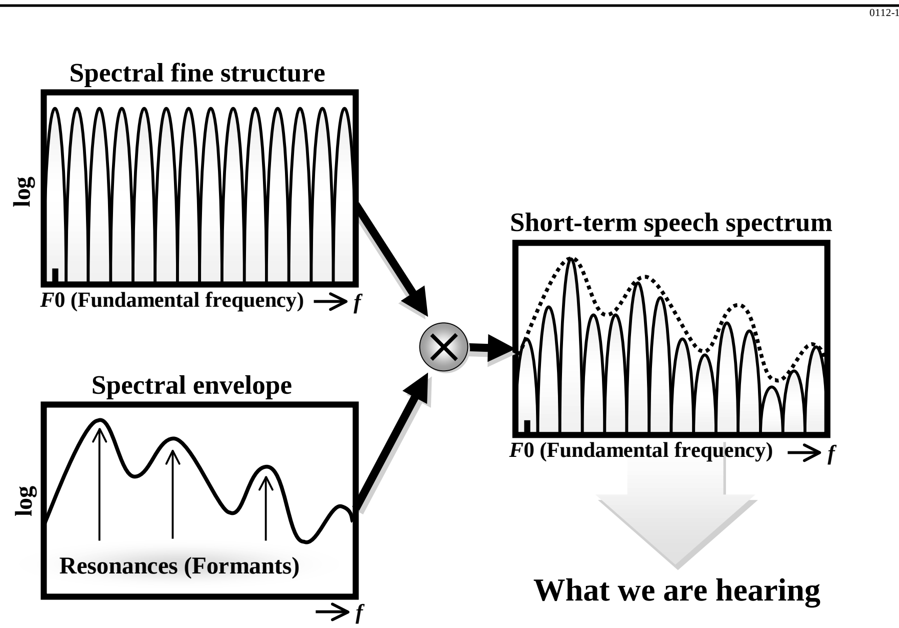
0104-05



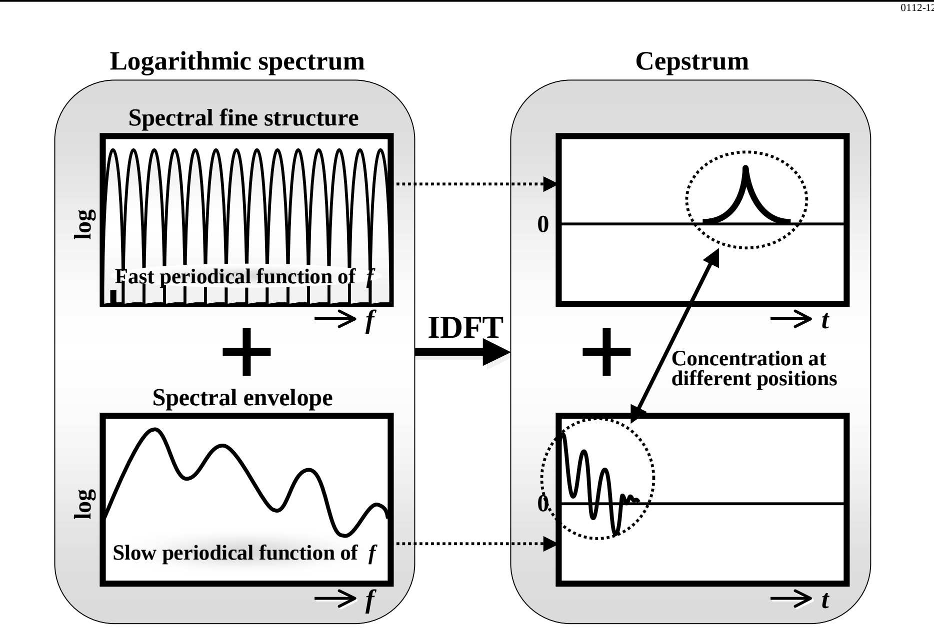
Digital sound spectrogram



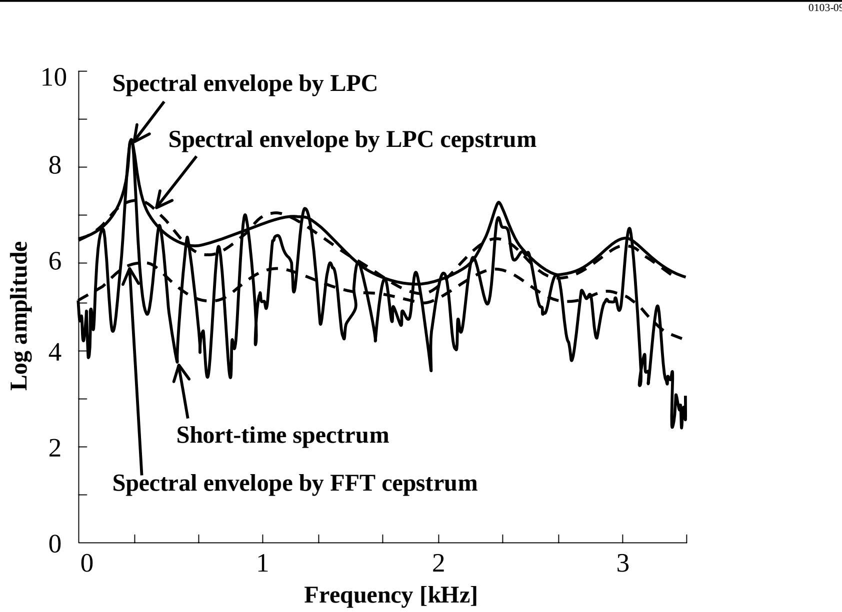
Feature vector (short-time spectrum) extraction from speech



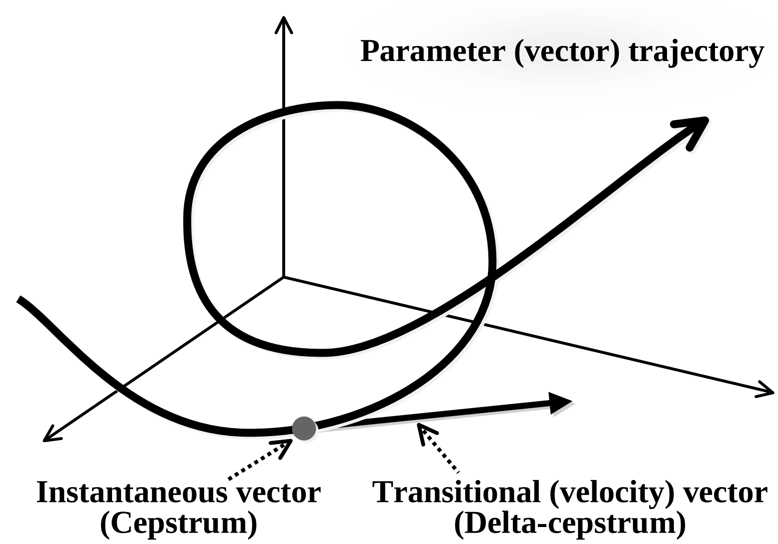
Spectral structure of speech



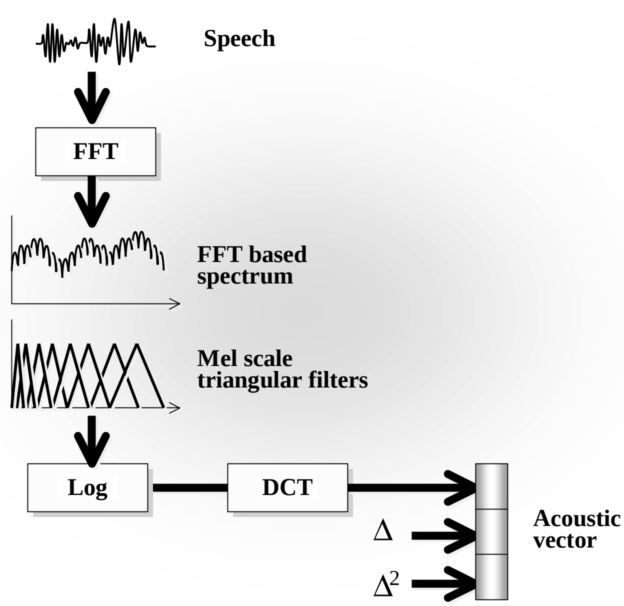
Relationship between logarithmic spectrum and cepstrum



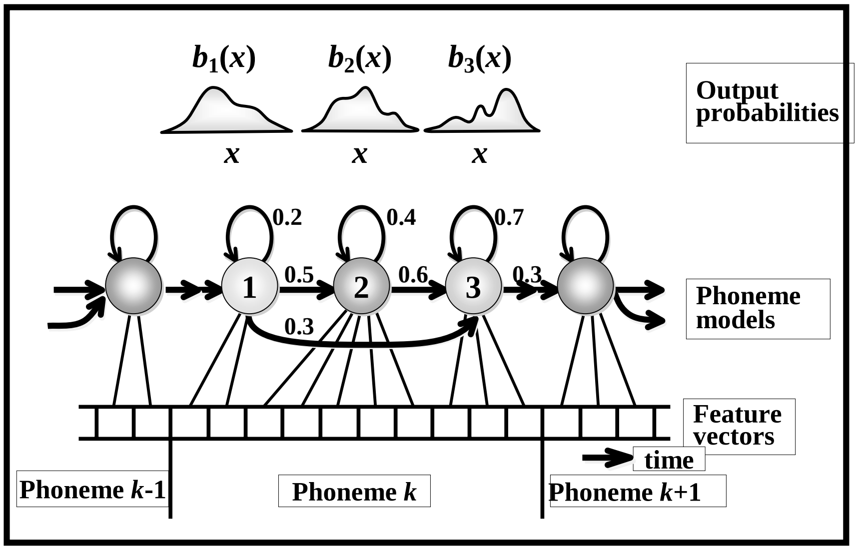
Comparison of spectral envelopes by LPC, LPC cepstrum, and FFT cepstrum methods



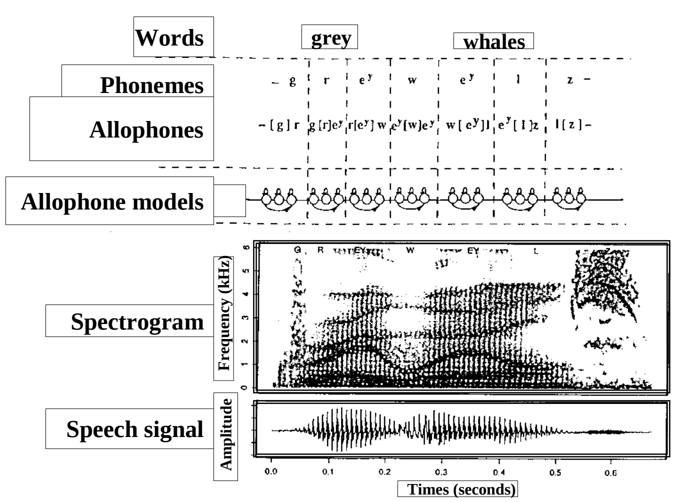
Cepstrum and delta-cepstrum coefficients



MFCC-based front-end processor



Structure of phoneme HMMs



Units of speech (after J.Makhoul & R. Schwartz)

0205-03

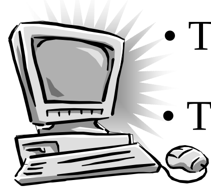
Outline

- Fundamentals of automatic speech recognition
- Acoustic modeling
- Language modeling
- Database (corpus) and task evaluation
- Transcription and dialogue systems
- Spontaneous speech recognition
- Speech understanding
- Speech summarization

0103-14

Language model is crucial !

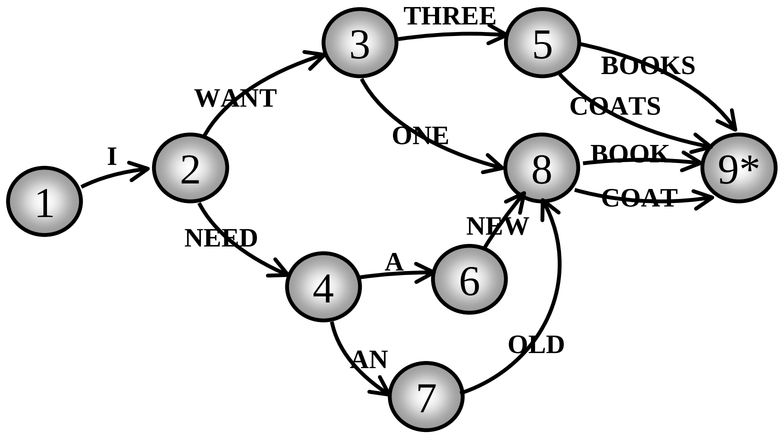
- Rudolph the red nose reindeer.
- Rudolph the Red knows rain, dear.
- Rudolph the Red Nose reigned here.



- This new display can recognize speech.
- This nudist play can wreck a nice beach.

0011-21

1. I	5. ONE	9. BOOKS	13. OLD
2. WANT	6. A	10. COAT	
3. NEED	7. AN	11. COATS	
4. THREE	8. BOOK	12. NEW	



An example of FSN (Finite State Network) grammar

0205-01

Syntactic language models

- **Rely on a formal grammar of a language.**
- **Syntactic sentence structures are defined by rules that can represent global constraints on word sequences.**
- **Mainly based on context-free grammars.**
- **Very difficult to extend to spontaneous speech.**

0205-02

Problems of context-free grammars

- **Over generation problem: not only generates correct sentences but also many incorrect sentences.**
- **Ambiguity problem: the number of syntactic ambiguities in one sentence becomes increasingly unmanageable with the number of phrases.**
- **Suitability for spontaneous speech is arguable.**

→Stochastic context-free grammars

0012-01

Statistical language modeling

Probability of the word sequence $w_1^k = w_1 w_2 ... w_k$:

$$P(w_1^k) = \prod_{i=1}^{k} P(w_i | w_1 w_2 \ldots w_{i-1}) = \prod_{i=1}^{k} P(w_i | w_1^{i-1})$$

$$P(w_i | w_1^{i-1}) = N(w_1^i) / N(w_1^{i-1})$$

where $N(w_1^i)$ is the number of occurrences of the string w_1^i in the given training corpus.

Approximation by Markov processes:

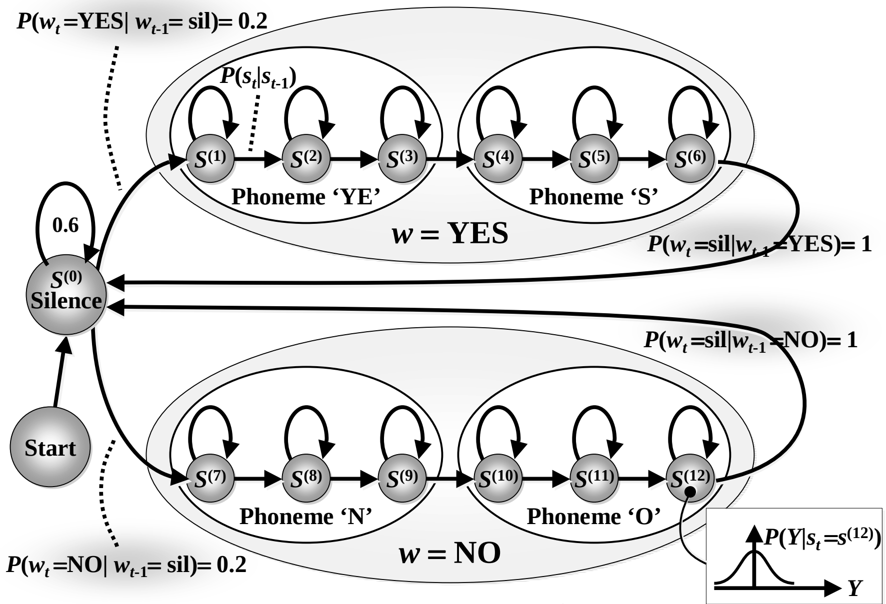
Bigram model $P(w_i | w_1^{i-1}) = P(w_i | w_{i-1})$

Trigram model $P(w_i | w_1^{i-1}) = P(w_i | w_{i-2} w_{i-1})$

Smoothing of trigram by the deleted interpolation method:

$$P(w_i | w_{i-2} w_{i-1}) = \lambda_1 P(w_i | w_{i-2} w_{i-1}) + \lambda_2 P(w_i | w_{i-1}) + \lambda_3 P(w_i)$$

0202-11



Complete Hidden Markov Model of a simple grammar

0205-03

Variations of *N*-gram language models

- Various smoothing techniques
- Word class language models
- *N*-pos (parts of speech) models
- Combination with a context-free grammar
- Extend *N*-grams to the processing of long-range dependencies
- Cache-based adaptive/dynamic language models

0201-08

Good-Turing estimate

For any *n*-gram that occurs *r* times, we should pretend that it occurs *r times as follows:**

$$r^* = (r+1)\,\frac{n_{r+1}}{n_r}$$

where n_r is the number of *n*-grams that occur exactly *r* times in the training data.

0108-16

Katz smoothing algorithm

Katz smoothing extends the intuitions of the Good-Turing estimate by adding the combination of higher-order models with lower-order models.

$$P_{Kats}(w_i|w_{i-1}) = \begin{cases} C(w_{i-1}w_i)/C(w_{i-1}) & \text{if } r>k \\ d_r C(w_{i-1}w_i)/C(w_{i-1}) & \text{if } k\geq r>0 \\ \alpha(w_{i-1})/P(w_i) & \text{if } r=0 \end{cases}$$

where $d_r = \dfrac{\frac{\mathrm{r}^*}{\mathrm{r}} - \frac{(k+1)\,n_{k+1}}{n_1}}{1 - \frac{(k+1)\,n_{k+1}}{n_1}}$ and $\alpha(w_{i-1}) = \dfrac{1-\sum_{w_i:r>0} P_{Kats}(w_i|w_{i-1})}{1-\sum_{w_i:r>0} P(w_i)}$

0205-03

Outline

- Fundamentals of automatic speech recognition
- Acoustic modeling
- Language modeling
- Database (corpus) and task evaluation
- Transcription and dialogue systems
- Spontaneous speech recognition
- Speech understanding
- Speech summarization

0205-04

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?**

0205-05

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.

0202-13

Main database characteristics (Becchetti & Ricotti)

Name	Quantities				No. units
	No.CD	No.hours	Giga-bytes	No. speakers	
TI Digits	3	~14	2	326	>2,500 numbers
TIMIT	1	5.3	0.65	630	6,300 sentences
NTIMIT	2	5.3	0.65	630	6,300 sentences
RM1	4	11.3	1.65	144	15,024 sentences
RM2	2	7.7	1.13	4	10,608 sentences
ATIS0	6	20.2	2.38	36	10,722 utterances
Switchboard (Credit Card)	1	3.8	0.23	69	35 dialogues
TI-46	1	5	0.58	16	19,136 isol. words
Road Rally	1	~10	~0.6	136	Dialogues/sentences
Switchboard (Complete)	30	250	15	550	2,500 dialogues
ATC	9	65	5.0	100	30,000 dialogues
Map Task	8	34	5.1	<256	128 dialogues
MARSEC	1	5.5	0.62	-	53 monologues
ATIS2	6	~37	~5	351	12,000 utterances
WSJ-CSR1	18	80	9.2	>124	38,000 utterances

0202-14

Further database characteristics (Becchetti & Ricotti)

Name	Transcription		Speech style	Recording environment	SR kHz	Sponsor
	Based on:	TA				
TI Digits	Word	No	Reading	QR	20	TI
TIMIT	Phone	Yes	Reading	QR	16	DARPA
NTIMIT	Phone	Yes	Reading	Tel	8	NYNEX
RM1	Sentence	No	Reading	QR	20	DARPA
RM2	Sentence	No	Reading	QR	20	DARPA
ATIS0	Sentence	No	Reading spon.	Ofc	16	DARPA
Switchboard (Credit Card)	Word	Yes	Conv. spon.	Tel	8	DARPA
TI-46	Word	No	Reading	QR	16	TI
Road Rally	Word	Yes	Reading spon.	Tel	8	DoD
Switchboard (Complete)	Word	Yes	Conv. spon.	Tel	8	DARPA
ATC	Sentence	Yes	Spon.	RF	8	DARPA
Map Task	Sentence	Yes	Conv. spon.	Ofc	20	HCRC
MARSEC	Phone	-	Spon.	Various	16	ESRC
ATIS2	Sentence	No	Spon.	Ofc	16	DARPA
WSJ-CSR1	Sentence	Yes	Reading	Ofc	16	DARPA

0112-13

Entropy: Amount of information (Task difficulty)

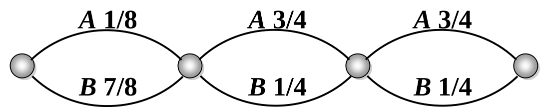
- Yes, No 1 bit (log 2)
- Digits
  - * 0 ~ 9 : 0.1 3.32 bits (log 0.1)
  - * $\begin{cases} 0 : 0.91 \\ 1 \sim 9 : 0.01 \end{cases}$ 0.722 bits

    (0.91 log 0.91 + 0.09 log 0.01)

0112-14

Test-set perplexity

Vocabulary: *A, B*   Test sentence: *ABB*



Test sentence entropy: $\log 8 + \log 4 + \log 4 = 7\,\text{bit}$ **(*ABB*)**

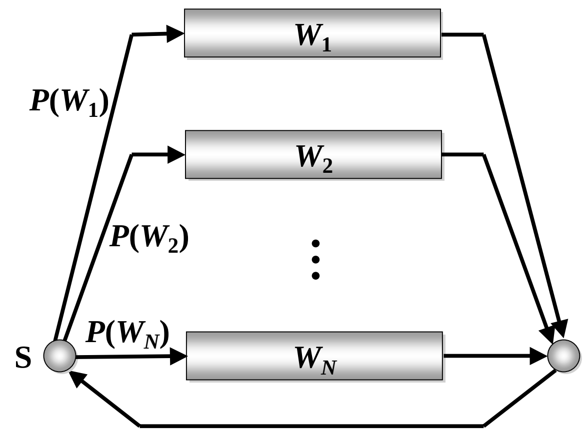
Entropy per word: 7 / 3 = 2.33 bits

Test-set perplexity (branching factor): $2^{2.33} = 5.01$

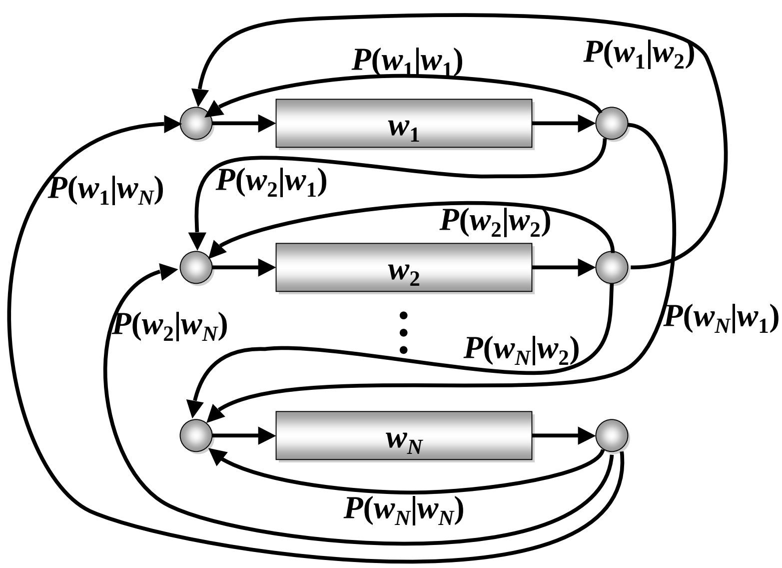
0205-03

Outline

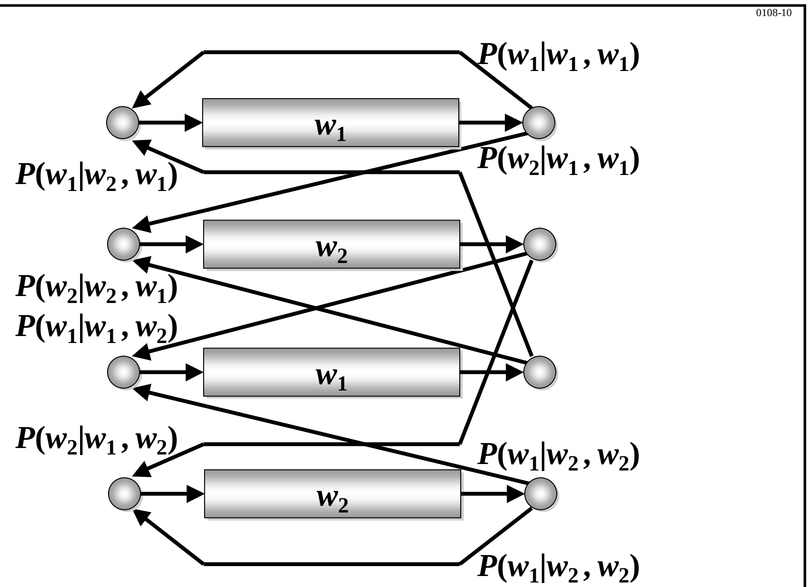
- Fundamentals of automatic speech recognition
- Acoustic modeling
- Language modeling
- Database (corpus) and task evaluation
- Transcription and dialogue systems
- Spontaneous speech recognition
- Speech understanding
- Speech summarization



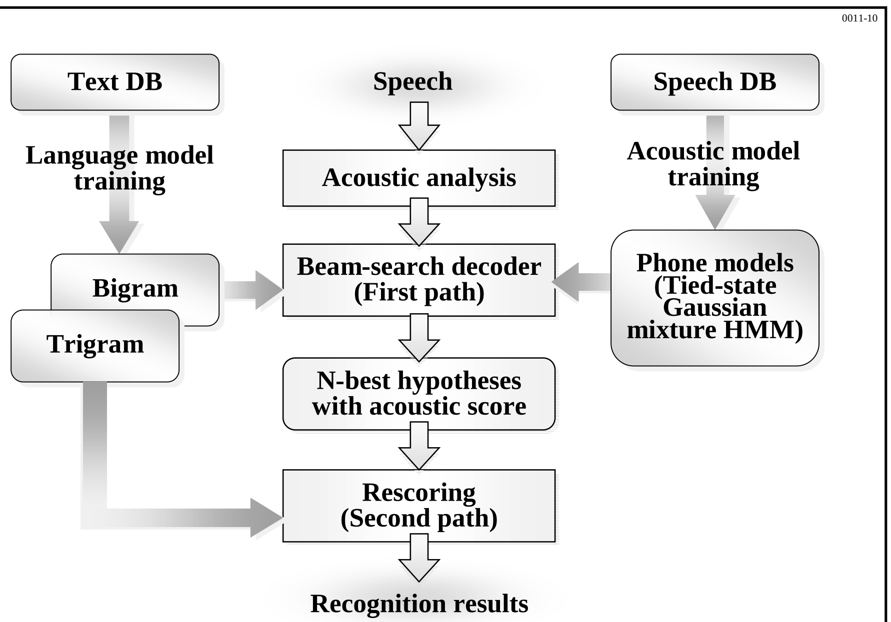
A unigram grammar network where the unigram probability is attached as the transition probability from starting state S to the first state of each word HMM.



A bigram grammar network where the bigram probability $P(w_j|w_i)$ is attached as the transition probability from word w_i to w_j.



A trigram grammar network where the trigram probability $P(w_k|w_i, w_j)$is attached to transition from grammar state w_i, w_j to the next word w_k. Illustrated here is a two-word vocabulary, so there are four grammar states in the trigram network.



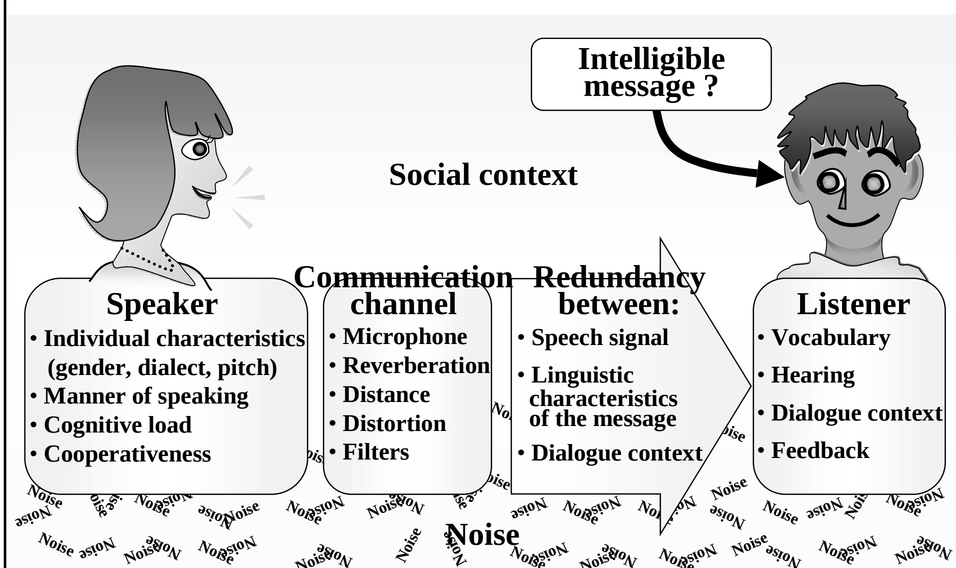
Two-pass search structure used in the Japanese broadcast-news transcription system

0205-03

Outline

- Fundamentals of automatic speech recognition
- Acoustic modeling
- Language modeling
- Database (corpus) and task evaluation
- Transcription and dialogue systems
- Spontaneous speech recognition
- Speech understanding
- Speech summarization

0204-03



The main factors influencing speech communication

The possible distortions in the communication and the redundancy of the speech signal interact, yielding to a message intelligible or unintelligible to a listener

Difficulties in (spontaneous) speech recognition

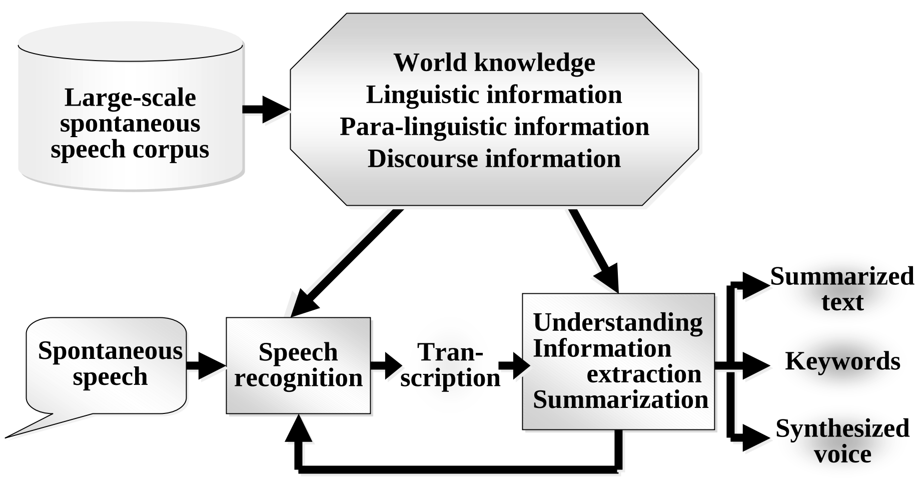
- **Lack of systematic understanding in variability**
  - **Structural or functional variability**
  - **Parametric variability**

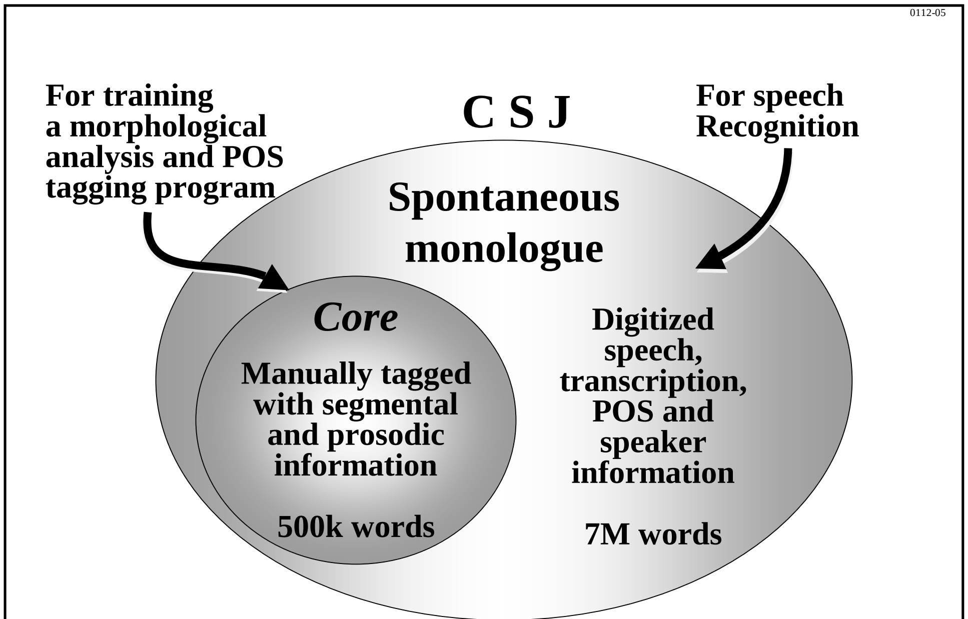
- **Lack of complete structural representations of (spontaneous) speech**

- **Lack of data for understanding non-structural variability**

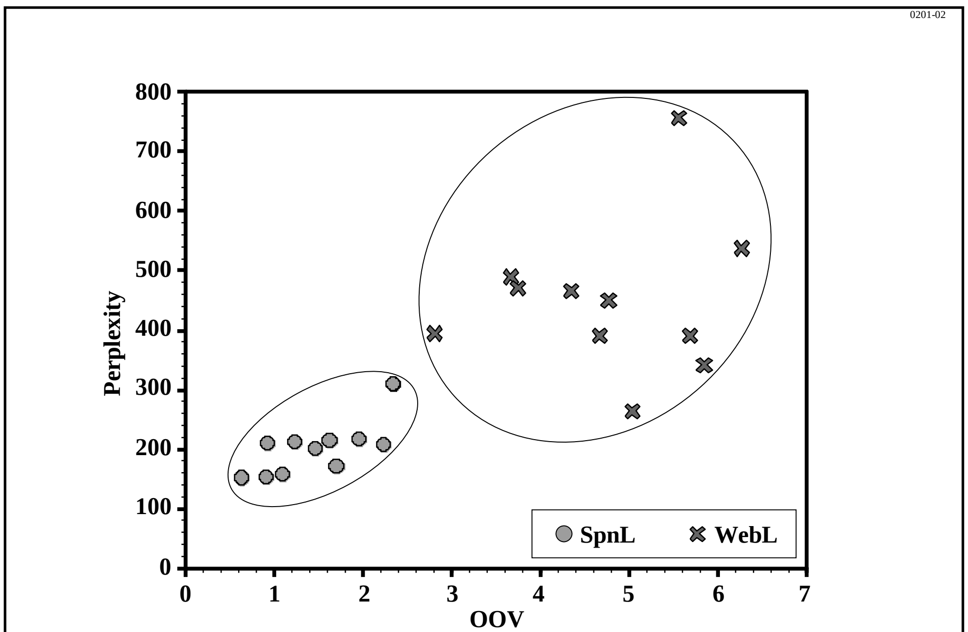
0010-24

Overview of the Science and Technology Agency Priority Program "Spontaneous Speech: Corpus and Processing Technology"

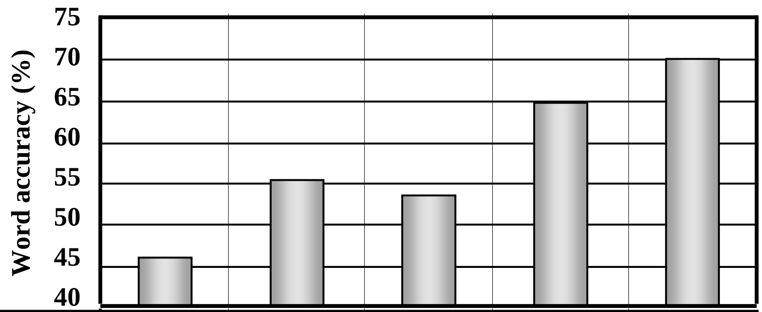




Overall design of
the Corpus of Spontaneous Japanese (CSJ)



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



Acous. model	RdA	SpnA	RdA	SpnA	
Ling. model	SpnA				
Speaker adapt.	WebL	WebL	SpnL	SpnL	SpnL
	w/o	w/o	w/o	w/o	with

Word accuracy for each combination of models

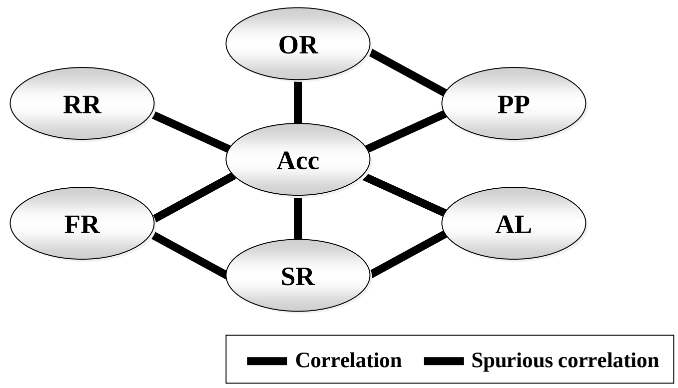
0204-13

Mean and standard deviation for each attribute of presentation speech

	Acc	AL	SR	PP	OR	FR	RR
Mean	68.6	-53.1	15.0	224	2.09	8.59	1.56
Standard deviation	7.5	2.2	1.2	61	1.18	3.67	0.72

Acc: word accuracy (%), AL: averaged acoustic frame likelihood, SR: speaking rate (number of phonemes/sec), PP: word perplexity, OR: out of vocabulary rate, FR: filled pause rate (%), RR: repair rate (%)

0201-05



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

Summary of correlation between various attributes

0202-03

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

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

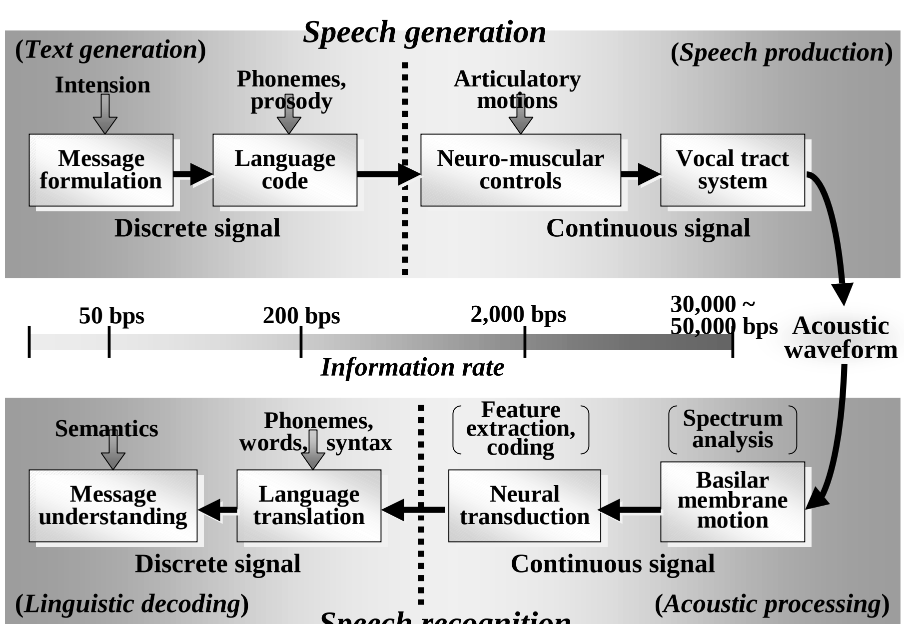
Acc: word accuracy, SR: speaking rate,
PP: word perplexity, OR: out of vocabulary rate,
FR: filled pause rate, RR: repair rate

0205-03

Outline

- Fundamentals of automatic speech recognition
- Acoustic modeling
- Language modeling
- Database (corpus) and task evaluation
- Transcription and dialogue systems
- Spontaneous speech recognition
- Speech understanding
- Speech summarization

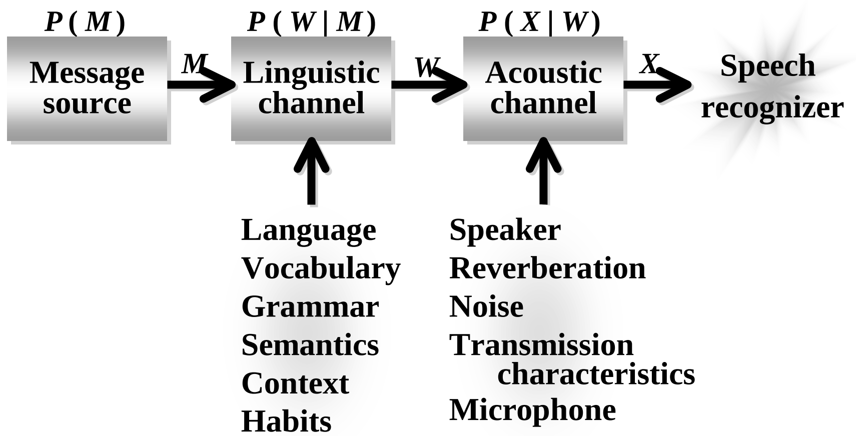
0011-11



Human speech generation and recognition process

0010-22

A communication - theoretic view of speech generation & recognition



9903-4

Message-driven speech recognition

Maximization of *a posteriori* probability,

$$\max_{M} P(M \mid X) = \max_{M} \sum_{W} P(M \mid W) P(W \mid X) \tag{1}$$

Using Bayes' rule, it can be expressed as

$$\max_{M} P(M \mid X) = \max_{M} \sum_{W} P(X \mid W) P(W \mid M) P(M) / P(X) \tag{2}$$

For simplicity, we can approximate the equation as

$$\max_{M} P(M \mid X) \approx \max_{M,W} P(X \mid W) P(W \mid M) P(M) / P(X) \tag{3}$$

$P(W|M)$: hidden Markov models, $P(M)$: uniform probability for all M.

We assume that $P(W|M)$ can be expressed as follows.

$$P(W \mid M) \approx P(W)^{1-\lambda} P(W \mid M)^{\lambda} \tag{4}$$

where λ, $0 \le \lambda \le 1$, is a weighting factor.

9809-1

Word co-occurrence score

***P*(*W*|*M*)** is represented by word co-occurrences of nouns;

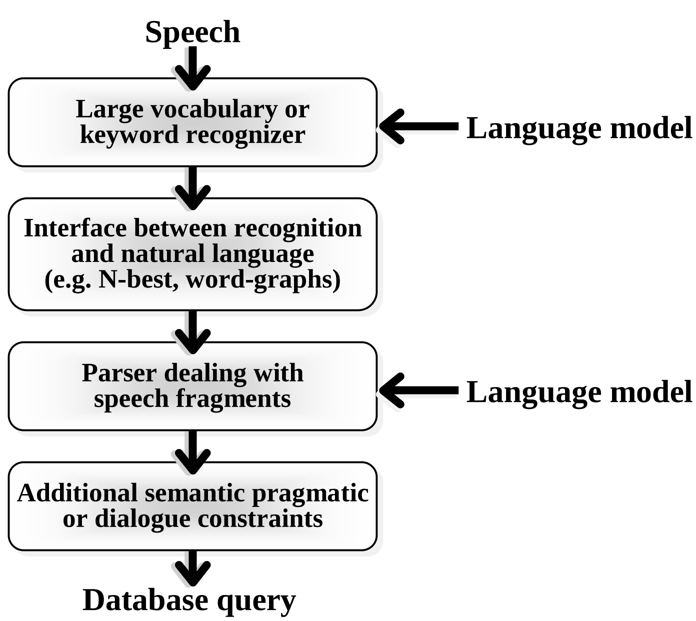
$$CoScore(w_i, w_j) = \log \frac{p(w_i, w_j)}{(p(w_i)\,p(w_j))^{1/2}}$$

$p(w_i,w_j)$: probability of observing words w_i and w_j in the same news article

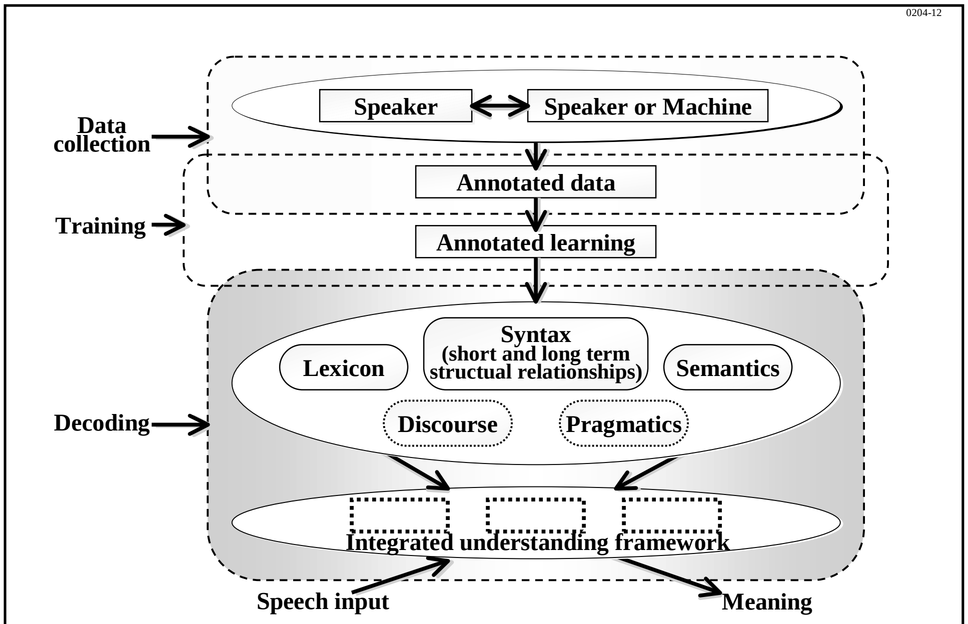
$p(w_i)$, $p(w_j)$: probabilities of observing word w_i and w_j in all the articles

A square root term was employed to compensate the probabilities of the words with very low frequency.

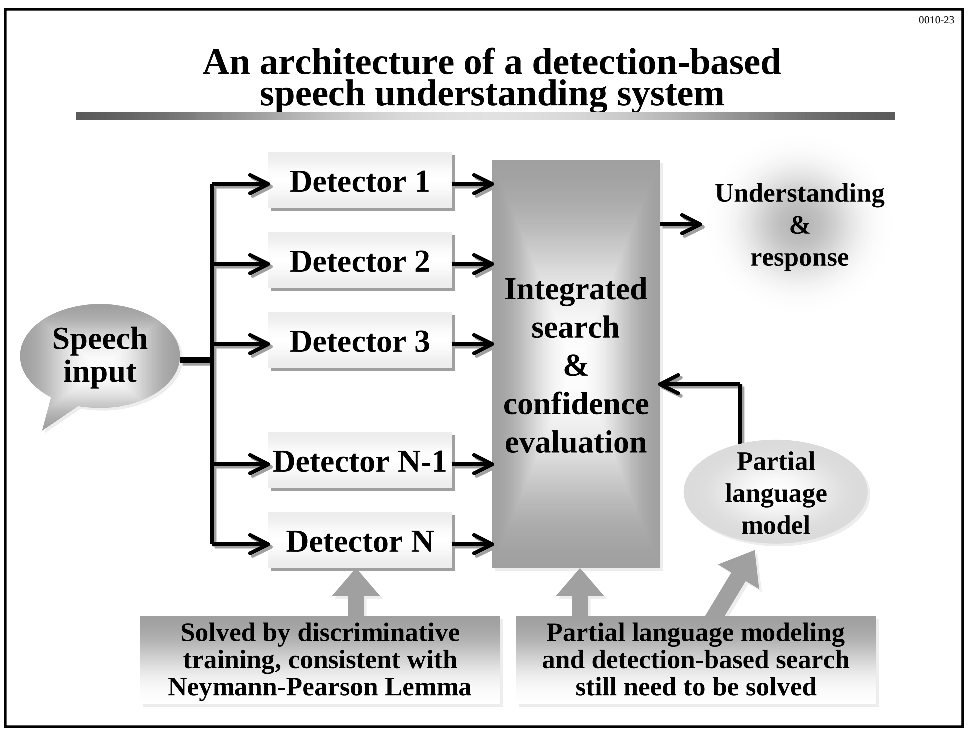
0204-11



Generic block diagram for spontaneous speech understanding



Generic semi-automatic language acquisition for speech understanding



0205-03

Outline

- Fundamentals of automatic speech recognition
- Acoustic modeling
- Language modeling
- Database (corpus) and task evaluation
- Transcription and dialogue systems
- Spontaneous speech recognition
- Speech understanding
- Speech summarization

0205-06

Sayings

- **The shortest complete description is the best understanding – *Ockham***
- **If I had more time I could write a shorter letter – *B. Pascal***
- **Make everything as simple as possible – *A. Einstein***

0205-07

From speech recognition to summarization

LVCSR (**L**arge **V**ocabulary **C**ontinuous **S**peech **R**ecognition) **systems** can transcribe **read speech** with 90% word accuracy or higher.

Current target

LVCSR systems for **spontaneous speech** recognition to generate closed captions, abstracts, etc.

- **Spontaneous speech features**
  - **filled pauses, disfluency, repetition, deletion, repair, etc.**
- **Outputs from LVCSR systems include recognition errors**

Automatic speech summarization

Important information extraction

0205-08

Summarization levels

Summarization → **Information extraction Understanding speech**

Indicative summarization

Topics

Sentence(s)

Abstract

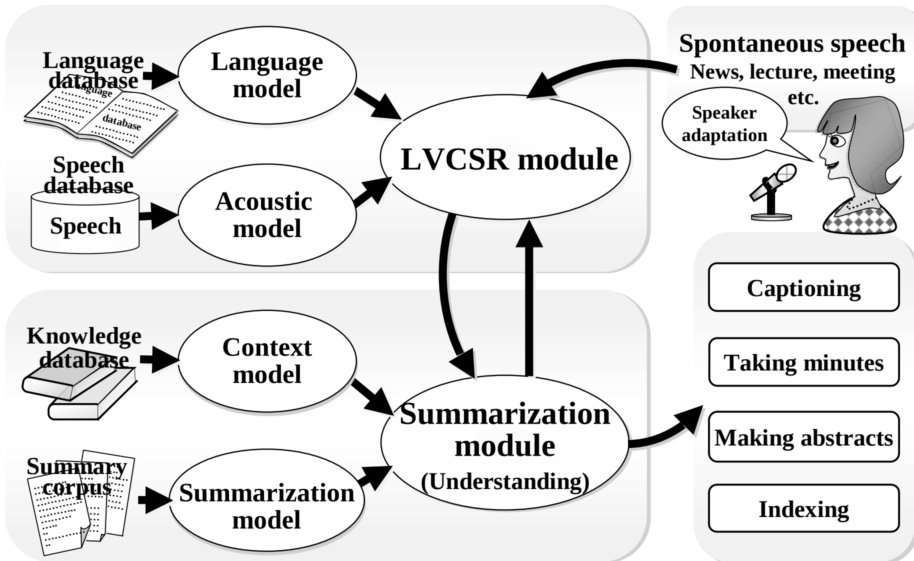
Summarized utterance(s)

Raw utterance(s)

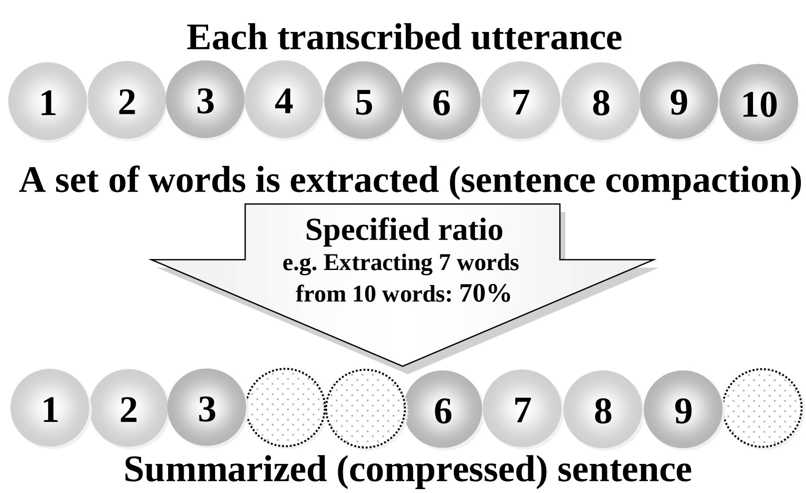
Informative summarization

Target
Closed captioning
Lecture summarization

Automatic speech summarization system



Approach to speech summarization utterance by utterance



0205-11

Target of summarized speech

Maintaining original meanings of speech as much as possible

Information extraction	**Significant (topic) words**
Linguistic correctness	**Linguistic likelihood**
Semantic correctness	**Dependency structure between words**
Word error exclusion	**Acoustic and linguistic reliability**

0205-12

Summarization score

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

Summarization score

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

- $L(v_m | \cdots v_{m-1})$ → **Linguistic score** — Trigram
- $\lambda_I I(v_m)$ → **Significance (topic) score** — Amount of information
- $\lambda_C C(v_m)$ → **Confidence score** — Acoustic & linguistic reliability
- $\lambda_T T_r(v_m)$ → **Word concatenation score** — Word dependency probability

0205-13

Linguistic score

Linguistic likelihood of word strings (bigram/trigram) in a summarized sentence

$$\log P\,(v_m \mid v_{m-2}\, v_{m-1})$$

Linguistic score is trained using a summarization corpus.

0205-14

Word significance score

Amount of information

$$f_i \log \frac{F_A}{F_i}$$

f_i : **Number of occurrences of** u_i **in the transcribed speech**
u_i : **Topic word in the transcribed speech**
F_i : **Number of occurrences of** u_i **in all the training articles**
F_A : **Summation of all** F_i **over all the training articles**

$$(F_A = \sum_i F_i)$$

- **Significance scores of words other than topic words and reappearing topic words are fixed.**

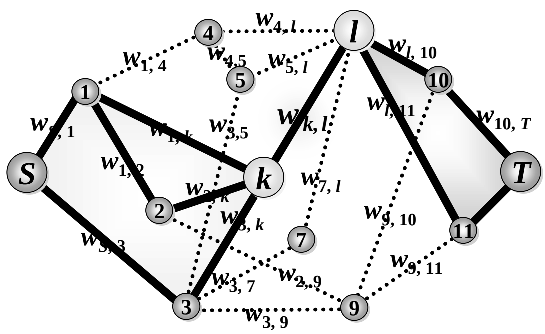
0205-15

Confidence score

Acoustic and linguistic reliability of a word hypothesis

Posterior probability

$$C(w_{k,l}) = \log \frac{\boxed{\alpha_k}\ \boxed{P_{ac}(w_{k,l})\ P_{lg}(w_{k,l})}\ \boxed{\beta_l}}{P_G}$$



$C(w_{k,l})$: Log posterior probability of $w_{k,l}$

k,l : Node index in a graph

$w_{k,l}$: Word hypothesis between node k and node l

α: Forward probability from the beginning node S to node k

β: Backward probability from node l to the end node T

P_{ac} : Acoustic likelihood of $w_{k,l}$

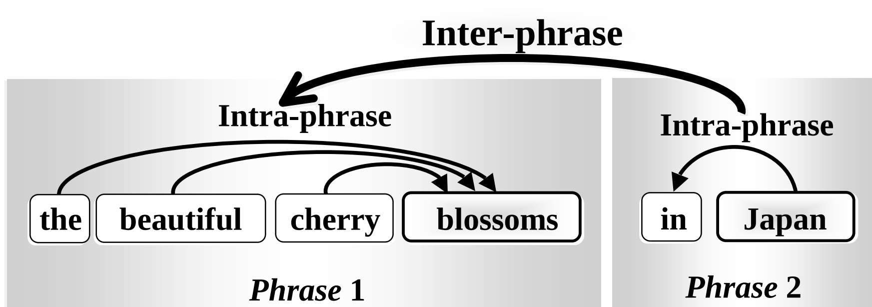
P_{lg} : Linguistic likelihood of $w_{k,l}$

P_G : Forward probability from the beginning node S to the end node T

0205-16

Word concatenation score

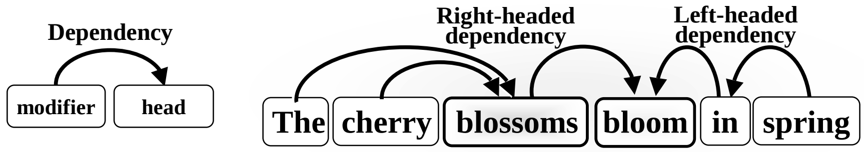
A penalty for word concatenation with no dependency in the original sentence



"the beautiful Japan" ⇨ **Grammatically correct but incorrect as a summary**

Dependency structure

Dependency Grammar

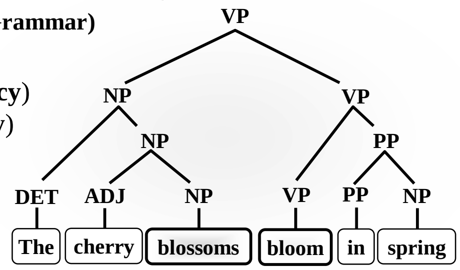


Phrase structure grammar for dependency

DCFG (Dependency Context Free Grammar)

$\alpha \rightarrow \beta\alpha$ **(Right-headed dependency)**
$\alpha \rightarrow \alpha\beta$ **(Left-headed dependency)**
$\alpha \rightarrow w$

α, β: **Non-terminal symbols,**
w: **Terminal symbols**



Word concatenation score based on SDCFG

Word dependency probability

If the dependency structure between words is deterministic,

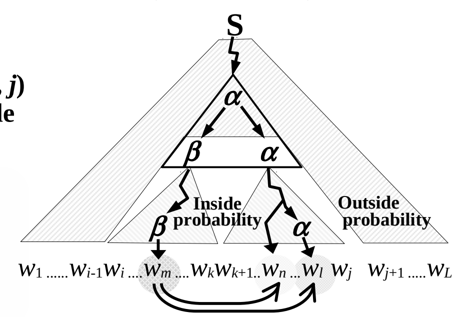
0 or 1

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.

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

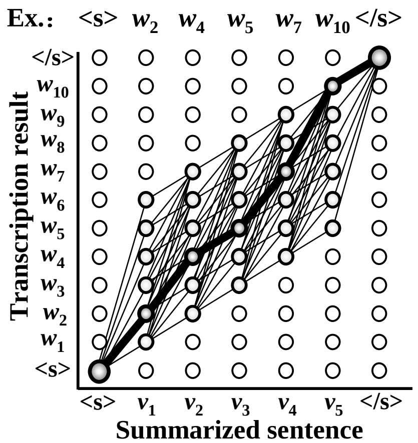


S: Initial symbol, α, β: Non-terminal symbol, w: Word

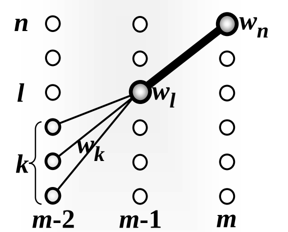
0205-19

Dynamic programming for summarizing each utterance

Selecting a set of words maximizing the summarization score



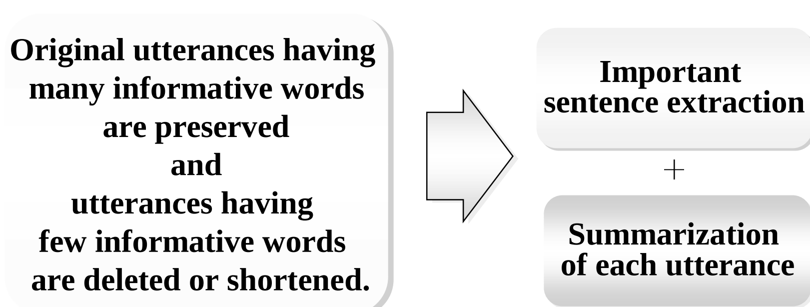
$$g(m, l, n) = \max_{k<l}\{g(m-1, k, l) + \log P(w_n|w_k w_l) + \lambda_I I(w_n) + \lambda_C C(w_n) + \lambda_T T_r(w_l, w_n)\}$$



0205-20

Summarization of multiple utterances

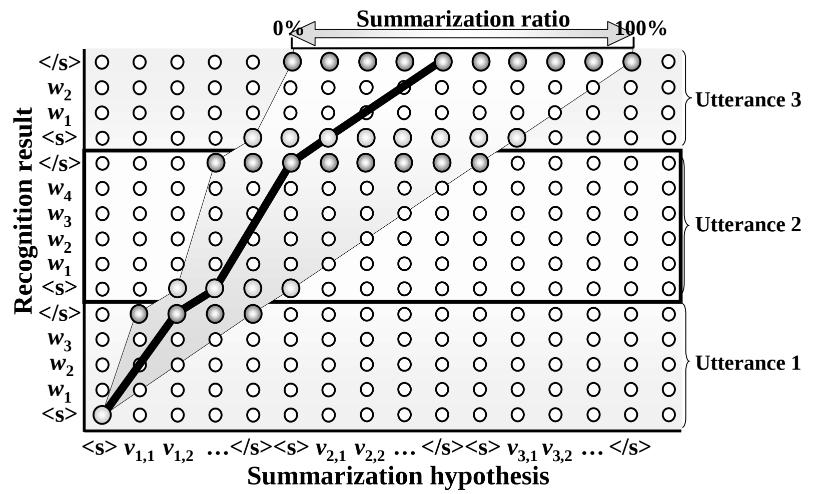
The method of summarizing each utterance is extended to summarize a set of multiple utterances by adding a rule giving a restriction at utterance boundaries.



0205-21

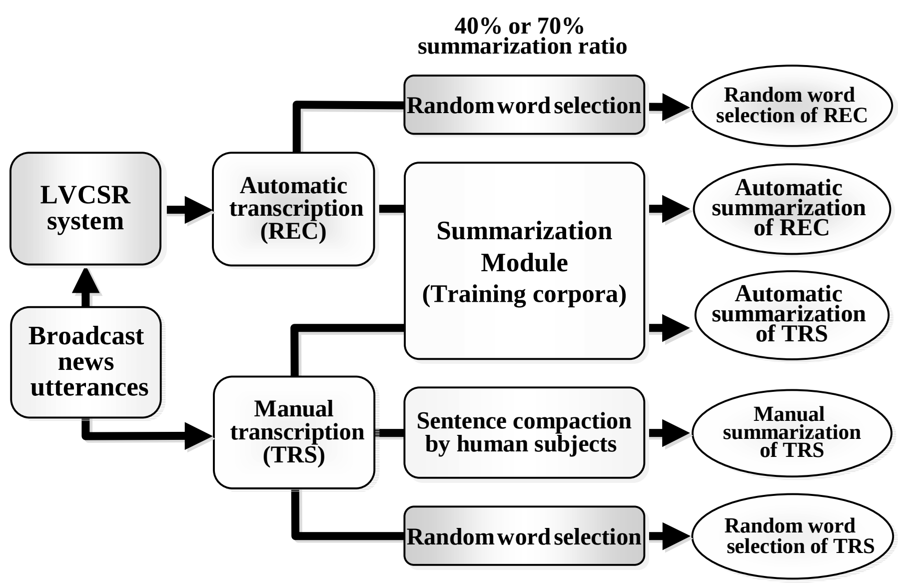
Dynamic programming for summarizing multiple utterances

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



0205-22

Evaluation experiments

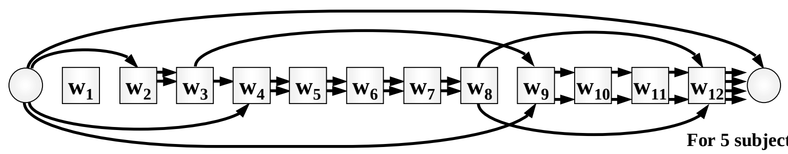


0205-23

Word network of manual summarization results for evaluation

Manual summarization results are merged into a network.

- **The network approximately expresses all possible correct summarization including subjective variations.**



Summarization accuracy is defined as the word accuracy based on the word string, extracted from the word network, that is most similar to the automatic summarization result.

Summarization accuracy = {Len-(Sub+Ins+Del) }/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

0205-25

Examples of automatic summarization for manually transcribed CNN news - 1

- **Transcription:**
  It's sulfur, and as Ed Garsten reports in today's edition of tech trends, the petroleum industry is proposing a cleanup.
- **Automatic summarization (30-40% summarization ratio) :**
  sulfur Ed Garsten reports tech petroleum __ proposing cleanup.
- **The most similar word string in the manual summarization network:**
  Ed Garsten reports tech trends industry proposing cleanup.
- **Automatic summarization (50-70% summarization ratio):**
  sulfur Ed Garsten reports in today's edition tech trends petroleum industry is proposing cleanup.
- **The most similar word string in the manual summarization network :**
  Sulfur, Garsten reports in today's tech trends the industry is proposing cleanup.

0205-26

Examples of automatic summarization for manually transcribed CNN news - 2

- **Transcription:**
  We are dealing with something of such a massive uh size and potential impact, um that a lot of people wisely are saying hands off.
- **Automatic summarization (20-40% summarization ratio) :**
  We're dealing something _ impact lot of people saying hands __.
- **The most similar word string in the manual summarization network :**
  We're dealing something such impact lot of people saying hands off.
- **Automatic summarization (50-70% summarization ratio) :**
  We're dealing with something of a size and impact, a lot of people wisely are saying hands __.
- **The most similar word string in the manual summarization network :**
  We're dealing with something of such size and impact, a lot of people wisely are saying hands off.

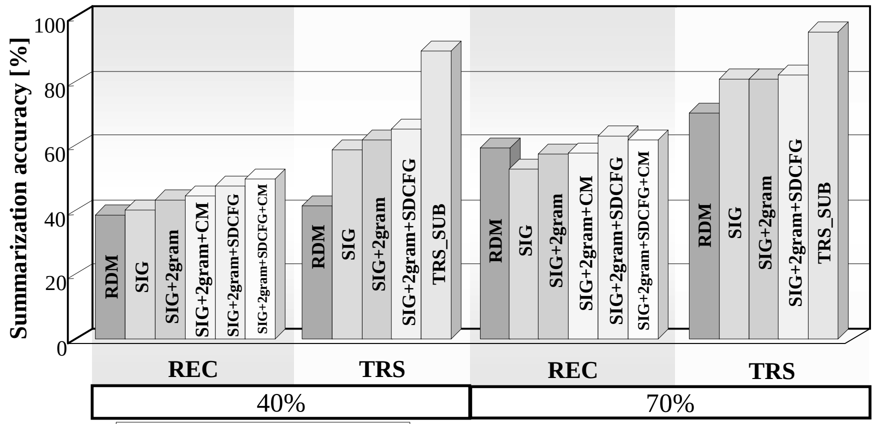
0205-27

Examples of automatic summarization for recognized CNN news (80% recognition accuracy)

- **Recognition result:**
  Vice president Al Gore says the government has a plan to avoid the inevitable prospect of increased airplane crashes and fatality is
- **Automatic summarization (40% summarization ratio) :**
  Gore the government has a plan to avoid the increased airplane crashes
- **The most similar word string in the word network:**
  <INS> ***the government has a plan to avoid the increased airplane crashes***
- **Automatic summarization (70% summarization ratio) :**
  Vice president Al Gore says the government has a plan to avoid **<DEL>** ***increased airplane crashes***
- **The most similar word string in the word network:**
  Vice president Al Gore says the government has a plan to avoid the increased airplane crashes

0205-28

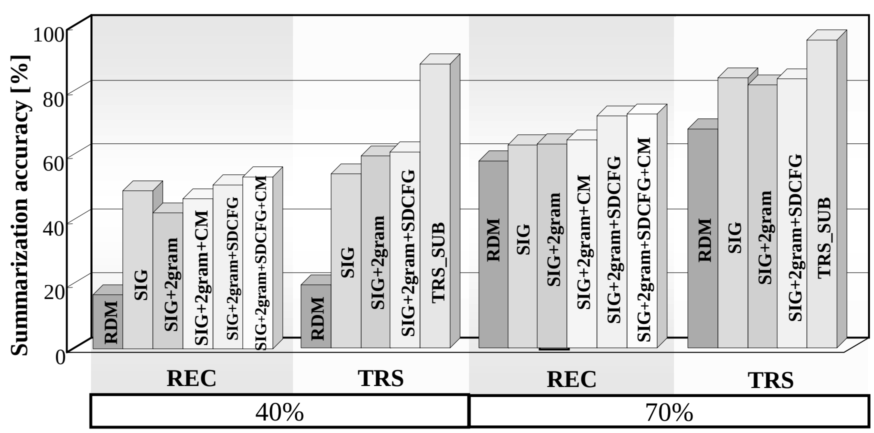
English news speech summarization (Each utterance summarization)



RDM: Random word selection
SIG: Word significance score
2gram: Linguistic score
CM: Confidence score
SDCFG: Word concatenation score
TRS_SUB: Manual summarization of manual transcription

0205-29

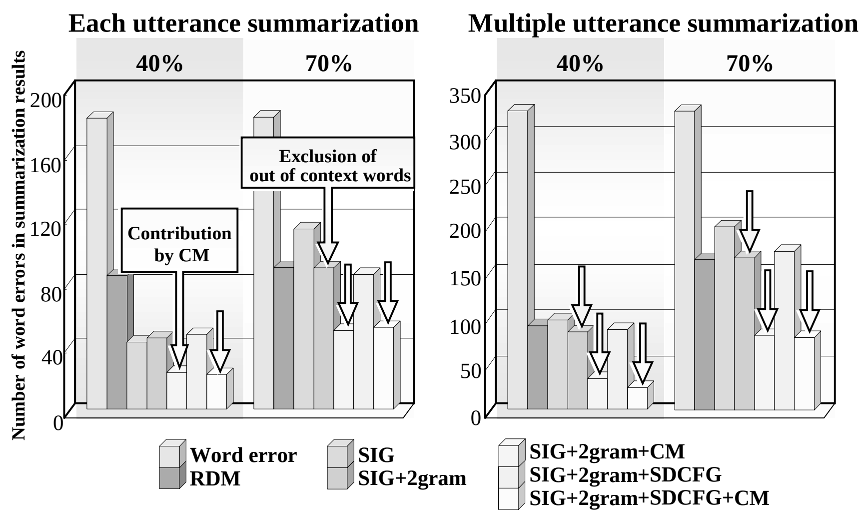
English news speech summarization (Multiple utterance summarization)



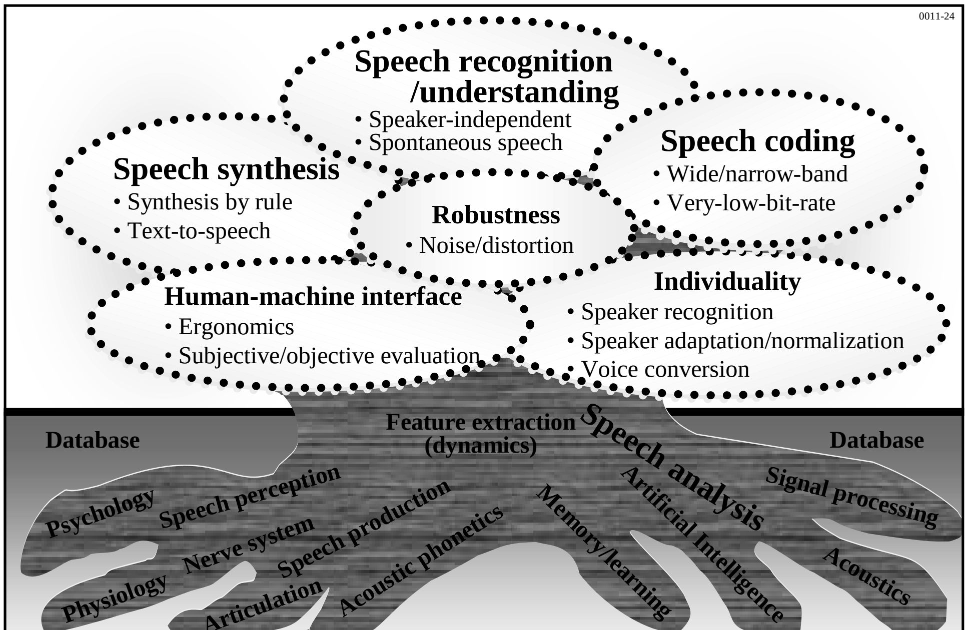
RDM: Random word selection
SIG: Word significance score
2gram: Linguistic score
CM: Confidence score
SDCFG: Word concatenation score
TRS_SUB: Manual summarization of manual transcription

0205-32

Recognition error reduction



0011-24



Speech information processing "tree", consisting of present and future speech information processing technologies supported by scientific and technological areas serving as the foundations of speech research.

Summary

- ***Speech recognition technology* has made significant progress with many potential applications.**
- **How to model and recognize *spontaneous speech* is one of the most important issues.**
- **Construction of a large-scale *spontaneous speech corpus* is crucial.**
- **Paradigm shift from recognition to *understanding* is needed.**
- ***Speech summarization* is attractive as information extraction and speech understanding.**

References

- C. Becchetti and L. P. Ricotti: *Speech Recognition*, John Wiley & Sons, Ltd., New York, 2000
- S. Furui: *Digital Speech Processing, Synthesis, and Recognition, Second Edition*, Signal Processing and Communications Series, Marcel Dekker, New York, 2000
- D. Gibbon, I. Mertins and R. K. Moore (Eds.): *Handbook of Multimodal and Spoken Dialogue Systems*, Kluwer Academic Publishers, Boston, 2000
- B.-H. Juang and S. Furui: "Automatic recognition and understanding of spoken language processing – A first step toward natural human-machine communication," Proc. IEEE, 88, 8, pp. 1142-1165, 2000
- C. Hori, S. Furui, R. Malkin, H. Yu and A. Waibel: "Automatic summarization of English broadcast news speech," Proc. Human Language Technology 2002, San Diego, pp. 228-233, 2002
- X.-D. Huang, A. Acero and H.-W. Hon: *Spoken Language Processing*, Prentice Hall PTR, New Jersey, 2001
- S. Young and G. Bloothooft (Eds.): *Corpus-based Methods in language and Speech Processing*, Kluwer Academic Publishers, Boston, 1997