A Study on Statistical Methods for Automatic Speech Summarization

Chiori Hori

A thesis submitted for the degree of Doctor of Philosophy in the Department of Computer Science, Graduate School of Information Science and Engineering

Tokyo Institute of Technology

March 2002

Abstract

This dissertation proposes a new automatic speech summarization method through word extraction. In this method, a set of words maximizing a summarization score indicating an appropriateness of summarization is extracted from automatically transcribed speech. This extraction is performed according to a target compression ratio using a dynamic programming technique sentence by sentence. The extracted set of words is then connected to construct a summarization sentence. The summarization score consists of a word significance measure, a confidence measure, linguistic likelihood, and a word concatenation probability. The word concatenation score is determined by a dependency structure in the original speech given by Stochastic Dependency Context Free Grammar (SDCFG). This summarization process aims to maintain the original meaning as much as possible within a limited number of words.

In order to make abstracts, we then extend this method to summarization of a set of multiple utterances (sentences). This is done by adding a rule which restricts application of the score beyond the sentence boundaries. This summarization technique can preserve more words inside information rich utterances and shortening or even completely deleting less informative ones.

This paper proposes a summarization method for multiple utterances using a two-level Dynamic Programming (DP) technique in order to reduce the amount of calculation.

Japanese and English broadcast news speech which has been transcribed using a large-vocabulary continuous-speech recognition (LVCSR) system is summarized using our proposed method. In addition, lecture speech is also transcribed and summarized to make abstracts. and compared with manual summarization by human subjects. The manual summarization results are combined to build a word network. This word network is used to calculate the word accuracy of each automatic summarization result using the most

similar word string in the network. Experimental results show that the proposed method effectively extracts relatively important information by removing redundant and irrelevant information.

Acknowledgments

I'm deeply grateful to Prof. Sadaoki Furui for giving me the precious opportunity to study the challenging research topic, "Speech summarization", at TITECH in Japan. Without his invaluable discussion and exact guidance, this thesis would not have been possible. I also thank him for his support and encouragement for my doctoral course.

I would also like to thank other members of my thesis committee, Prof. Hozumi Tanaka, Prof. Takenobu Tokunaga, Prof. Toru Noguchi and Prof. Manabu Okumura (TITECH, Japan), for reviewing to this thesis, and giving me helpful remarks.

I'd like to appreciate Prof. Waibel in CMU for allowing to stay at the Interactive Systems Laboratories and use their English speech recognition system, JRTk. Also Ms. Céline Morel and Mr. Michael Bett for setting up my stay of the Interactive Systems Laboratories. Thanks also to Mr. Rob Malkin and Mr. Hua Yu for helping with the English news speech recognition. I'd like to thank Dr. Klaus Zechëner for having a very fruitful discussion about "Speech summarization" with me. Also I'd like to thank Dr. Yoshi Gotoh (Sheffield University) for his arrangement of generating the English correct answers for automatic summarization.

I'd like to express deep appreciate to Ms. Rie Akisawa, Mr. Ryuta Taguma and Dr. Kouji Iwano in the Furui laboratory for helping me do all the things necessary for my doctor course. Also thanks to Mr. Zhipeng Zhang who shared this unforgettable doctor course with me. I'd like to also thank Ms. Chrystabel Butler for her encouragement and checking my English for this thesis. Thanks to Dr. William Robert for discussions about my method in detail. Thanks also to the graduate and undergraduate students, in the Furui laboratory who have contributed to evaluate automatic summarization results. I'd like to also thank Mr. Atsushi Iwasaki who tested the performance of word significance measures.

Finally, this thesis would not have been possible without the economic and mental support of my husband, Dr. Takaaki Hori. He has been the one who helped me solve the many technical difficulties I encountered along the way. Without constant patient and support, I could not have completed this project.

This thesis is dedicated to my husband and our parents, especially to our mothers in heaven, Tokuko Hori and Fusako Yoshida.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech is the most natural form of communication used among humans. Speech communication has always been, and will continue to be, the most dominant media of information exchange due to its simplicity and efficiency. Speech can be propagated by telephony, radio, television and Internet and also can be saved as in the form of archives. Spoken language processing by computers requires various technologies from signal processing to acoustics, phonology, phonetics, syntax, semantics, pragmatics and discourse analysis.

Since the advent of research into automatic speech recognition technology four decades ago, substantial advances has been accelerated by the enhancement of computational power and storage capacity. As a result of these technical advances, various commercial products using speech recognition systems are now on the market. A brief review of the research leading up to the present helps to introduce some of the challenges faced by current efforts to advance speech recognition technology [1] [2] [3] [4] [5]. Only a few years ago, speech recognition was primarily associated with a limited number of applications: small-vocabulary isolated word recognition or phrases, mid-sized vocabulary domain specific spoken language systems, and dictation systems. For the past decade, large-vocabulary continuous speech recognition (LVCSR) has been one of the focal areas of research in speech recognition, serving as a test bed to evaluate models and algorithms. The core technology developed for LVCSR can be used for applications other than general dictation systems. Spoken language processing systems have been developed for a wide variety of applications ranging as follows:

- small-vocabulary

  - keyword recognition for telephone routing

- medium-size vocabulary

  - voice command and control systems in rudimentary dialogue systems for information retrieval and electronic commerce

- large-vocabulary speech dictation

  - spontaneous speech understanding
  - limited-domain speech translation

These systems have already had a large impact on society. A major advance has been the ability of today's laboratory systems to deal with non-homogeneous data, as is exemplified by broadcast data, which includes changing speakers, languages, backgrounds and topics. This capability has been enabled by advances in techniques for the following:

- robust signal processing
- improved training techniques, which can take advantage of very large audio and textual corpora
- algorithms for audio segmentation
- a probabilistic model using the hidden Markov model (HMM)
- unsupervised acoustic model adaptation
- efficient decoding with long-span language models
- ability to use much larger vocabularies than in the past
  (64000 words or more is common to reduce errors due to out-of-vocabulary words)

Additionally, the adoption of an assessment-driven development methodology, largely fostered by the U.S. DARPA efforts, have spurred these technical advancements. Fig. 1.1 shows the current research target for speech recognition tasks [6].

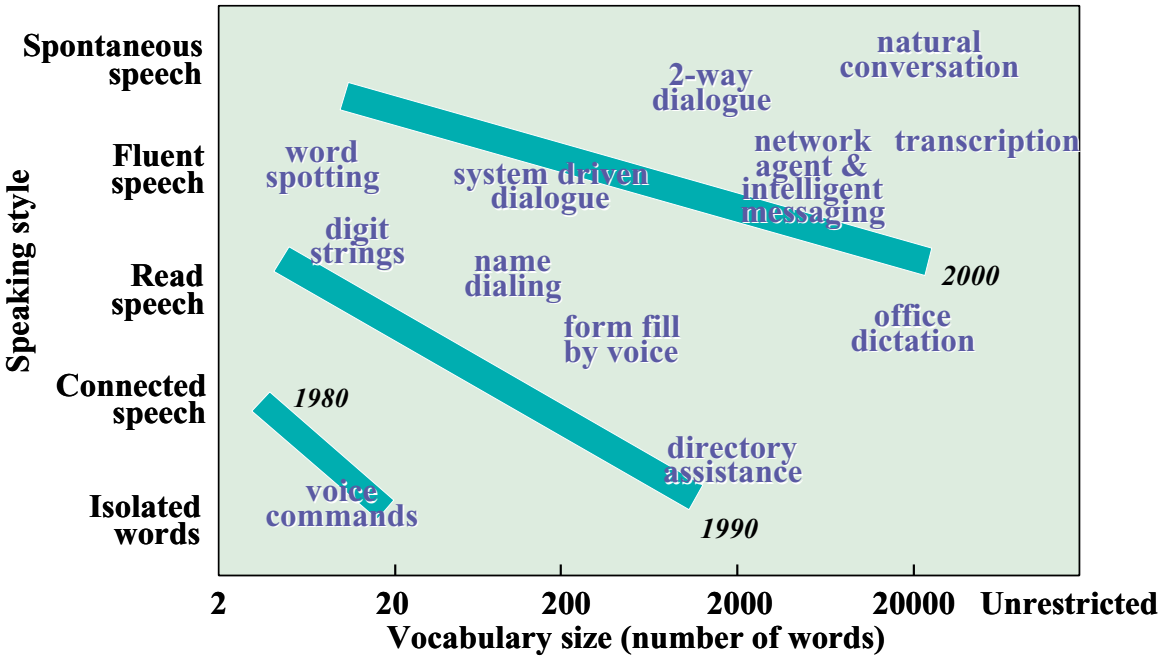


Figure 1.1: *The shift of speech recognition tasks in the U.S. DARPA projects.*

1.1 Application using Speech Recognition Technology

Recently, LVCSR technology has been making significant advancements. Major applications of the LVCSR systems in the near future will include automatic closed captioning broadcast TV programs and audio and video archives, meeting/conference summarization, unstructured multimedia data management, and natural, user-friendly interface with the computer. These applications using LVCSR systems can be broadly classified into two types: speech interface with machines and speech information management.

1.1.1 Speech interface

Human and machine communication via a speech interface realizes a more "friendly" use of computers without the necessity of typing. Speech recognition technology is used for systems such as a speech typewriter or voice commands to a machine. Human and machine dialogue systems using speech interface already on the market are in phone systems, i.e. air ticket reservation, information retrieval for shops or stock prices. The main challenge now of spoken language dialogue systems is to provide a more natural, user-friendly inter-

face with the computer. In the current dialogue systems, machines recognize given phrases (voice commands), or machines extract key words or phrases in users' speech representing information which is needed to achieve the task, such as location, time (key words or phrase spotting). However, it is not yet possible for speakers to speak to a machine in the same way as that they can speak to human operators, as the current technology cannot extract speakers' exact message via restricted voice commands and immature speech understanding technology based on key word spotting. In order for the technology to progress, recognition of natural speech is required.

1.1.2 Unstructured Information Management

With the remarkable progress in computer power, now an enormous amount of speech data, or multimedia data including speech, can be managed as an information database. The next step for speech recognition advancement is to create a system in which speech data is tagged (annotated) by text allowing for retrieval and extraction of information from databases. The extracted information can be represented by a part of original speech, a text which is transcribed speech by a recognition system, or synthesized speech.

Speech can be broadcasted with captions generated by speech recognition systems and simultaneously saved as speech and text (i.e. captions) archives in a database. Captions can be considered as an indexing of each word in whole speech. In the U.S., closed captions for broadcast news speech have been started. Since English is phonogram language, professional typists can transcribe speech in real time, and applications of speech recognition technologies have not been needed to make closed captions. On the other hand, since Japanese text is written with a mixture of three types of characters, Chinese characters (Kanji) and two types of Japanese characters (Hira-gana and Kata-kana), it is impossible even for professional typists to transcribe speech in real time. Therefore, an automatic closed captioning system using speech recognition technology is necessary. Currently an automatic transcription by a speech recognition system has been developed by NHK (Japan Broadcasting Corporation) is being directly used for an automatic closed captioning in a real system [7]. The multimedia database including indexes for information retrieval and extraction can be easily constructed using speech recognition systems. Recently one approach attempted to extract information from such a database by tracking

speech by query matching to indexes based on an automatic recognition result which had been synchronized with the speech data [8].

However, users attempting to retrieve information from such a speech database would, and do, prefer to access at abstracts rather than at the whole data, before they decide whether they are going to read or hear the entire data or not. Meeting/conference summarization will become useful relatively important information scattered about in the original speech can be extracted for an abstract. In order to compact information, meeting/conference summarization is actively being investigated [9] [10] .

In the closed captioning of broadcast news, in order to reduce of the number of words representing speech, speech summarization is also indispensable, because the number of words spoken by professional announcers sometimes exceeds the number of words that people can read and understand when all of them are presented on the TV screen in real time.

1.2 Problem Statement

Until now research into automatic speech recognition technology has focussed on extraction of linguistic information from speech signals based on each word. The goal of current recognition systems is to transcribe each word accurately.

Although state-of-the-art speech recognition technology can obtain high recognition accuracy for speech read from a written text or for similar types of pre-prepared language, the accuracy is quite poor for freely spoken spontaneous speech. Spontaneous speech is ill-formed and very different from written text. Even if a speech recognition system can transcribe it accurately, the transcribed speech usually includes redundant information such as disfluencies, filled pauses, repetitions, repairs and word fragments. In addition, irrelevant information included in a transcription caused by recognition errors is inevitable. Transcription results including such redundant and irrelevant information, cannot be directly used for indexing, or making abstracts and minutes.

Even in face to face human communication, such redundant and irrelevant information in speech can cause misunderstandings. However, a human's ability to guess the exact meaning of a speakers' messages based on world knowledge can alleviate such misunderstandings in face to face human communication. Therefore, in order to construct

practical applications of speech recognition in the real world, first and foremost, systems must understand speakers' message before processing tasks to achieve speakers' or systems' demand. Thus, to create truly useful technology for understanding spontaneous speech, practical applications using speech recognizers require a speech summarization process which removes redundant and irrelevant information, and at the same time extracts relatively important information depending on users' or systems' needs.

The computational power and storage capacity allows us to access freely archived multimedia data as a database and thus need for skimming technology from unstructured multimedia data will be required rapidly in the near future. In addition, the combination of multimedia data and the transcription of its' audio by a recognition system will make available databases. In order to build a system to freely access archived multimedia using large vocabulary continuous recognition (LVCSR) systems, a speech summarization technique for skimming information including both users' requirement information extraction and available information extraction technology will be required.

1.3 Speech Understanding

Speech summarization producing understandable sentences from original utterances can be considered as a kind of speech understanding. "Understanding" by a machine has been investigated in the Natural Language Processing (NLP) and Artificial Intelligence (AI) fields [11] [12]. In the current designs for computer systems understanding written sentences, knowledge is represented by symbols independent from linguistic notation and systemized to estimate meanings based on the meaning and context of the linguistic notation, and world knowledge. However, in the designs, the number of possible meaning hypotheses explode, so understanding via intermediate representation of world knowledge is not a realistic solution for understanding language by computer systems. Such an understanding method to estimate all of the factors of syntax, semantics and pragmatics for grammatically correct sentences based on empirical rules, cannot be applied to understand the meanings of the transcription of ill-formed speech obtained by a recognition system. On the other hand, the current Information Retrieval (IR) and Dialogue systems can almost output users' required answer based on the features of word distribution in text without such heuristic understanding methods. Although the understanding process

of computer systems using statistical models based on the frequency of words in text is different from humans' understanding, this technique is very simple and effective.

In this dissertation, a new method to distill information from transcription of speech by a recognition system and produce a summary in the form of understandable sentences will be proposed.

1.4 Automatic Summarization Technology

The purpose of summaries generated by humans is to give readers a short and coherent impression of the main idea of an article. Although, in general, humans do not just collect the important phrases from original articles, studies have shown that about 80% of the sentences in abstracts by humans were "close sentence matches" , i.e., they were "either extracted verbatim from the original or with minor modifications", [13]. A broad distinction is usually drawn between indicative summaries and informative summaries [14]. Indicative summaries are used to indicate what topics are addressed in the source text, and thus can be used to alert the user to the content. Informative summaries are intended to cover the concepts in the source text to an extent possible given the compression ratio of original text words to summary text words.

Techniques of automatically summarizing written text have been actively investigated in the field of natural language processing for 40 years. An automatic summarization system must understand text and generate indicative/informative summaries according to users' demands. One of the major techniques for summarizing written text is the process of extracting important sentences from original text. Current machine systems can only produce an extract of the text, i.e., to select a number of "most relevant sentences for users' requests" and present them to the users. However, the sentence extraction technique cannot be directly applied to speech summarization. A major difference between text summarization and speech summarization exists in the fact that transcribed speech is sometimes linguistically incorrect due to the spontaneity of human speech and recognition errors by the system. A new approach is needed to automatically summarize speech to cope with such problems.

1.4.1 Automatic written text summarization

This section describes a brief review of current text summarization techniques [15] [16] [17]. Techniques of automatically summarizing written text have been actively investigated in the field of natural language processing. Text summarization methods can be classified into extractive or non-extractive method. Extractive summaries consist of a passages sequence (phrases, sentences, paragraphs) extracted from an original text, that are then sequenced together to create the summary. Non-extractive methods are generating summary by implying a generative component.

One of the major techniques for summarizing written text is the process of extracting important sentences. Interest in producing simple indicative abstracts, i.e., "extracts as abstracts", arose as early as the 1950s. An important paper of these days is the one by (Luhn, 1958) who suggested to weight the sentences of a document as a function of high frequency words, disregarding the very high frequency common words [18]. The Abstract Generation System by Edmundson [19] implemented a system for automatically generating text abstracts which, additionally to the standard keyword method (i.e., frequency depending weights), also used the following three methods for determining the sentence weights, cue word, title and location of a sentence.

A major difference between text summarization and speech summarization exists in the fact that transcribed speech is sometimes linguistically incorrect due to the humans' spontaneity of speech and recognition errors by a speech recognition system. The sentence extraction techniques cannot avoid being affected by recognition errors. A new approach to automatically summarizing speech is needed to cope with such problems.

1.4.2 Automatic speech summarization

Automatic summarization techniques have been intensively studied in the field of natural language processing as described in Section 1.4.1. An rapid enhancement of computers including accelerated processing, huge capacity of storage and high communication speed allows us to deal with enormous multi-data. Now our information source is not only text but also speech, image and so on.

One approach attempted to extract information from such a database by tracking speech by query matching to indexes based on an automatic recognition result which

had been synchronized with the speech data [8]. However, users attempting to retrieve information from such a speech database, would and do prefer to access abstracts rather than the whole data, before they decide whether they are going to read or hear the entire data or not. Meeting/conference summarization will become useful if it can be developed to extract relatively important information scattered about in the original speech. In the natural language processing field, recently a sentence compression technique using both text and its abstract has been proposed [20].

However, this technique cannot be directly used for speech summarization. The current issue for automatic speech summarization is how to deal with recognition result including word errors. Handling word errors becomes on fundamental aspect for successfully summarizing transcribed speech. In addition, since most approaches extract information based on each word, approaches based on longer phrase, or compressed sentences are required for extracting messages in speech. This dissertation proposes and automatic speech summarization method to generate meaningful sentences from transcribed speech excluding word errors by a recognition system.

1.5 Outline of Dissertation

The topics covered in the this dissertation proceed as follows: Chapter 2 describes the automatic speech summarization system. Chapter 3 to Chapter 4 describe topic score based on term weighting and word concatenation score based on SDCFG (Stochastic Dependency Context Free Grammar), respectively. Chapter 5 proposes an evaluation method for automatic summarization. Chapter 6 describes the evaluation experiments conducted. Chapter 7 discusses conclusions and future research orientations.

Chapter 2

Automatic speech summarization system

2.1 Overview of automatic speech summarization system

Recently, large-vocabulary continuous-speech recognition (LVCSR) technology has made significant advancements. Real time systems can now achieve word accuracy of 90 % of more for speech dictated from newspapers. Currently various applications of LVCSR systems, such as automatic closed captioning [7], meeting/conference summarization [10], and indexing for information retrieval [8] are actively being investigated.

Transcribed speech usually includes not only redundant information such as disfluencies, filled pauses, repetitions, repairs, and word fragments, but also irrelevant information caused by recognition errors. Therefore, practical applications using LVCSR systems require a process of speech summarization which removes redundant and irrelevant information and extracts relatively important information depending on users' requirements, especially for spontaneous speech.

Techniques for automatically summarizing written text have been actively investigated in the field of natural language processing [15]. One of the major techniques for summarizing written text is the process of extracting important sentences. A major difference between text summarization and speech summarization exists is that transcribed speech is sometimes linguistically incorrect due to recognition errors and the fact that spontaneous speech is often linguistically incorrect. A new approach is needed to automatically summarize speech to cope with such problems.

Our goal is to build a system that extracts and presents information from spoken

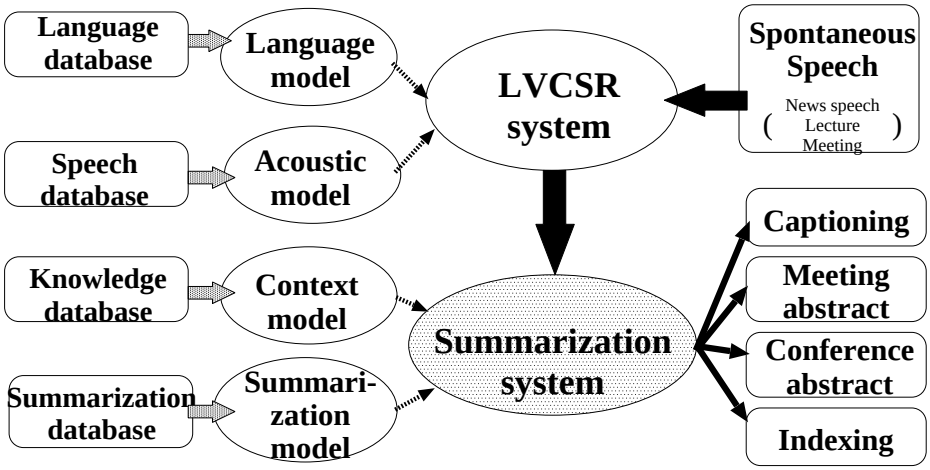


Figure 2.1: Automatic speech summarization system.

utterances according to users' desired amount of information. Fig.1 shows our proposed system. The output of the system can be either a simple set of keywords, a summarized sentence for each utterance, or summarization of an article consisting of multiple utterances. These outputs can be used for indexing, making closed captions and abstracts, etc. In the closed captioning of broadcast news, the number of words spoken by professional announcers sometimes exceeds the number of words that people can read and understand if all of them are presented on the TV screen. Therefore, reduction of the number of words in speech is indispensable. Meeting/conference summarization will be useful if it can extract relatively important information scattered about in the original speech.

In this paper, we first propose a new method of automatically summarizing each utterance. In this method, relatively important words are extracted removing redundant and irrelevant words according to a target compression ratio. The summarization method focuses on topic word extraction, weighting linguistically and semantically correct word concatenation [21][22], and acoustically as well as linguistically reliable parts of speech recognition results [23]. All of these features are represented as probabilistic scores. Summarization results obtained by this method simultaneously maintain topic words and keep a syntactic structure by properly weighting the scores.

We then extend this method to summarization of a set of multiple utterances (sentences) having consistent meanings. This is done by adding a rule which restricts applica-

tion of the score beyond the sentence boundaries. As a result, original sentences including many informative words are preserved, and those including less informative words are deleted or shortened. This summarization technique can be considered as a combination of the summarization method extracting important sentences developed in the field of natural language processing and the sentence-by-sentence summarization method. This multiple utterance summarization method should be especially useful for making lecture abstracts, meeting minutes, etc.

2.2 LVCSR system

In this section, a large vocabulary continuous speech recognition (LVCSR) system is described [24] [25]. A LVCSR system translates speech signals continuously spoken by a human into a corresponding word sequence. Figure 2.2 shows a typical LVCSR system. The system is roughly divided into two processing parts (speech analysis and decoder) and three knowledge sources (acoustic model, word lexicon, and language model).

In the speech analysis, the speech input S, a series of analog signals, is first transformed into a series of digital signals by analog-to-digital (AD) conversion. Then the features useful for speech recognition are extracted from the signals. A time series of the extracted feature, O, is sent to the next decoding part. In the decoder, the most likely word sequence $\hat{W}$ for the time series O is searched from among all the possible sequence of words in the lexicon. The recognition process is generally formulated as a stochastic process using Bayes' theorem.

$$\begin{aligned} P(\hat{W}|O) &= \max_{W} P(W|O) \\ &= \max_{W} \frac{P(O|W)P(W)}{P(O)} \end{aligned} \tag{2.1}$$

where the conditioned probabilities $P(O|W)$ is given by the acoustic model, and a priori probability $P(W)$ is given by the language model. $P(O)$ can be ignored because it's independent from W. Consequently, the decoding process is to solve the following problem.

$$\hat{W} = \operatorname*{argmax}_{W} P(O|W)P(W) \tag{2.2}$$

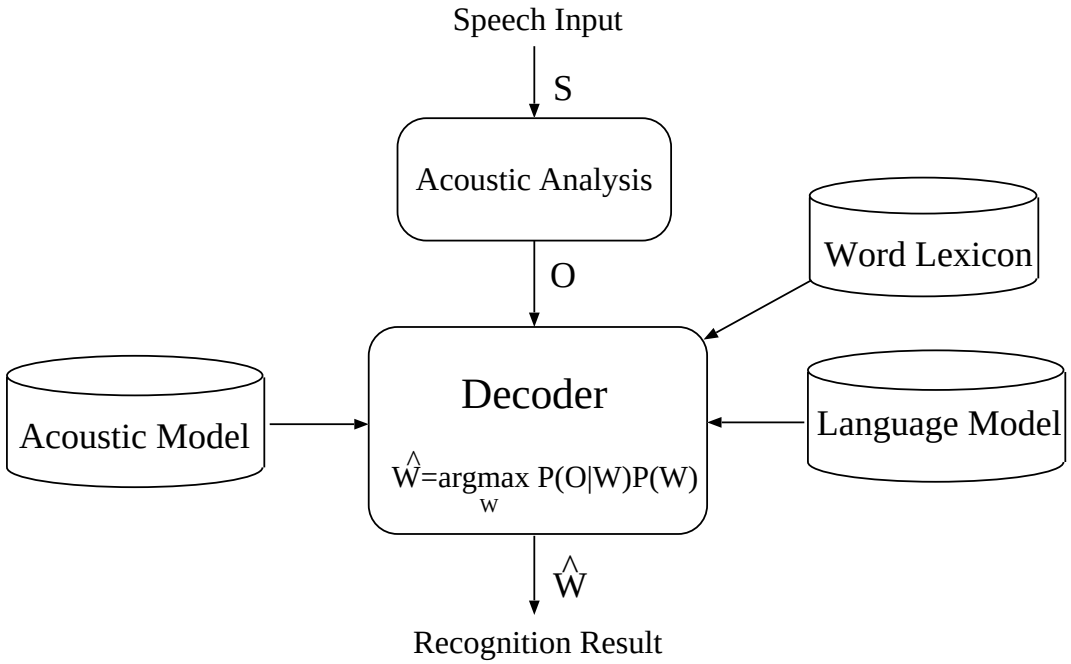


Figure 2.2: LVCSR system

2.2.1 Speech Analysis

AD conversion

The input analog signals are first sampled for digital processing. While the frequency range perceptible to human beings is between 20 and 20 000 Hz, the range including the sound of human speech is only from 50 to 7000 Hz. A sampling frequency of 14 000 Hz is therefore high enough for speech processing. A frequency lower than 14 000 Hz, however, is sometimes used because of the limitations of the transmission channel or to reduce computational costs.

Short-term spectrum

Speech research has proved that *pitch* and *formant* are important features for human perception. In the simplest speech production model, the vocal code generates excited signals (which correspond the pitch) and the vocal tract (whose response corresponds the formant) filters the signals. Although it may appear that speech recognition could

be realized by simulating the human speech production system, this is not possible with today's technology for the following two reasons. First, it is difficult to observe the real-time movement of the vocal organization. Second, it is not always possible to estimate the pitch and formant frequency from the acoustic signals observed.

Instead, the short-term spectrum is used for speech recognition in most cases. Here it is assumed that speech signals are stationary during a short period of 10-100 ms. Then the spectrum for this period is computed using the fast Fourier transform (FFT).

Cepstrum

The *cepstrum* of a signal is defined as a Fourier transform of the logarithm of the signal's power spectrum. It is especially useful when input signals are the superposition of excitation signals and linear filters. Speech signals are one such superposition: the excitation signals generated from the vocal code are filtered by the vocal tract response.

Let $y(n)$ be the speech signal at time n, $v(n)$ be the excited signal from the vocal code, and $h(n)$ be the vocal tract response. Then

$$y(n) = v(n) * h(n), \tag{2.3}$$

$$Y(e^{jw}) = V(e^{jw}) * H(e^{jw}), \tag{2.4}$$

$$\log|Y(e^{jw})| = \log|V(e^{jw})| + \log|H(e^{jw})|, \tag{2.5}$$

where $Y(\cdot)$, $V(\cdot)$, $H(\cdot)$ are the Fourier transforms of $y(n)$, $v(n)$, and $h(n)$.

Then the cepstrum $c(k)$ is

$$c(k) = v(k) + h(k). \tag{2.6}$$

Although the dimension of cepstrum is the same as that of time, the term *quefrency* is used as the dimension of cepstrum. In the cepstrum, the component $v(k)$ of the vocal tract response is dominant in the lower quefrency, while the component $h(k)$ of the vocal code has a strong peak at high quefrencies, which corresponds to a pitch frequency and its harmonics.

The influence of the pitch is efficiently removed by *liftering* (analogous to filtering in the spectral domain) the components $v(k)$ in the quefrency domain. The components representing the vocal tract response are used as the features for speech recognition.

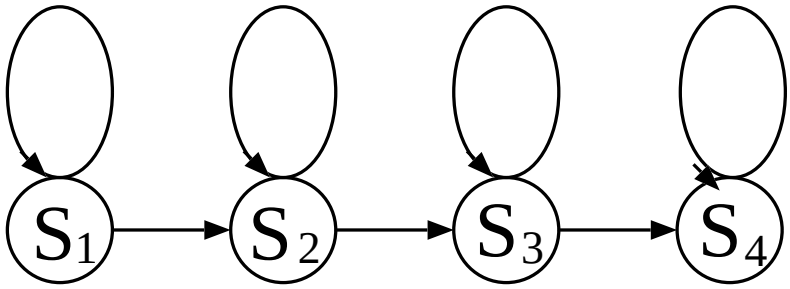


Figure 2.3: Left-to-right HMM.

Dynamic features

Dynamic features of the spectrum play an important role in human speech perception, and the delta cepstrum [26] developed to take these features into consideration is defined as follows:

$$\Delta c_n(t) = \frac{\sum_{k=-K}^{K} k c_n(t+k)}{\sum_{k=-K}^{K} k^2}. \tag{2.7}$$

This delta cepstrum and the second derivative feature, delta-delta cepstrum, are often used with cepstrum in many speech recognition systems and have been found to be effective.

2.2.2 Acoustic Models

Acoustic models based on *hidden Markov models* (HMMs) have recently been used widely for speech recognition(e.g., [27]). In this approach, the speech signals are characterized as outputs from Markov sources. In the recognition of words, for example, an HMM is assigned to each word in the lexicon, and for each utterance the recognized word selected is the one whose HMM is most likely to produce that utterance.

HMMs are classified into three types according to the form of the output *probability density function* (pdf) in each state: the discrete HMMs, in which the output pdf is discrete; the continuous density HMMs, in which the output pdf is continuous; and the semi-continuous HMMs, which is the combination of the discrete HMMs and the continuous-density HMMs.

The example of an HMM (a left-to-right HMM) is shown in Figure 2.3. For the simplicity of explanation, only the discrete HMMs (DHMM) is referred here. Let $\mathbf{O} =$

$(\mathbf{o}_1, \ldots, \mathbf{o}_t, \ldots, \mathbf{o}_T)$ be an observation sequence. Let S be the number of states, and $A = \{a_{ij}\}$ be a set of transition probability distributions, in which a_{ij} is the probability of transition from state i to state j; let $B = \{b_i(\mathbf{o}_t)\}$ be a set of the output probability distributions, where $b_i(\mathbf{o}_t)$ is the output probability of the feature vector $\mathbf{o}_t$ at state i; let $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M\}$ be the set of symbols, where M is the number of symbols; and let q_t be the state at time t; let $\mathbf{q} = \{q_1, \ldots, q_T\}$ be a state sequence in which q_t is the state at time t. Then

$$b_i(\mathbf{o}_t) = b_i(k) = P(\mathbf{o_t} = \mathbf{v_k}, |q_t = i), \qquad i = 1, \ldots, S, \tag{2.8}$$

where S is the number of states in the HMM. The initial probability distribution $\pi = \{\pi_i\}$ is also defined, where π_i is the probability of being state i at time 1:

$$\pi_i = P(q_1 = i), \qquad i = 1, \ldots, S. \tag{2.9}$$

The parameter set $\lambda = (A, B, \pi)$ is the complete parameter set for the model. From now on, the focus is on a left-to-right HMM, in which a number of states form a sequence and from each state only transitions to itself and to the next state on the right are allowed.

Recognition using HMMs

This subsection shows how the probability of an observation sequence for a given HMM is calculated. Let $\mathbf{O} = (\mathbf{o}_1, \ldots, \mathbf{o}_T)$ be the observation sequence, $\lambda = (A, B, \pi)$ be the parameter set of the HMM, and let $P(\mathbf{O}|\lambda)$ be the probability of $\mathbf{O}$, given the model λ. In principle, $P(\mathbf{O}|\lambda)$ is obtained by adding up the probabilities of all the possible state transitions, each of which can be expressed as a path in a two-dimensional plane (see Figure 2.4).

There are two algorithms that can be used to calculate the probability: the Forward algorithm and the Viterbi algorithm. In the Forward algorithm the forward probability $\alpha_t(i)$ is defined as:

$$\alpha_t(i) = P(\mathbf{o}_1, \ldots, \mathbf{o}_t, q_t = i|\lambda). \tag{2.10}$$

The forward probability is calculated as follows:

$$\alpha_1(i) \quad = \quad \pi_i b_i(\mathbf{o}_1), \qquad 1 \leq i \leq S, \tag{2.11}$$

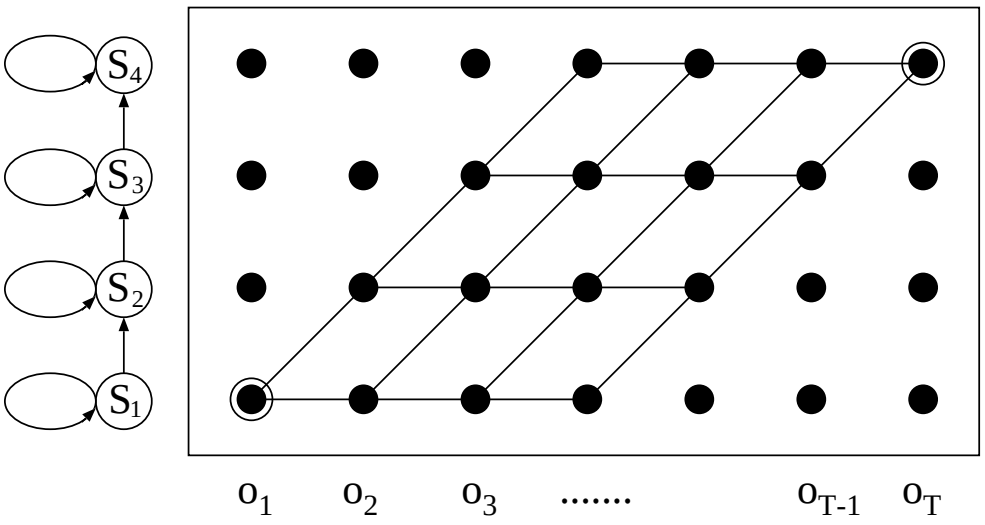


Figure 2.4: Forward algorithm.

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{S} \alpha_t(i)a_{ij}\right) b_j(\mathbf{o}_{t+1}), \quad 1 \leq t \leq T-1, 1 \leq j \leq S, \tag{2.12}$$

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{S} \alpha_T(i). \tag{2.13}$$

This recursive process finally yields the probability $P(\mathbf{O}|\lambda)$. In the Viterbi algorithm (see Figure 2.5) a maximization process is used instead of the summing procedure used in the Forward algorithm:

$$\delta_1(i) = \pi_i b_i(\mathbf{o}_1), \quad 1 \leq i \leq S, \tag{2.14}$$

$$\delta_t(j) = \max_{1 \leq i \leq S} (\delta_{t-1}(i)a_{ij}) b_j(\mathbf{o}_t), \quad 2 \leq t \leq T, 1 \leq j \leq S, \tag{2.15}$$

$$P^*(\mathbf{O}|\lambda) = \max_{1 \leq i \leq S} \delta_T(i). \tag{2.16}$$

Strictly speaking, the probability $P^*(\mathbf{O}|\lambda)$ obtained by the Viterbi algorithm is only an approximation of the probability obtained by the Forward algorithm. The Viterbi algorithm is often used, however, and it has been proved that the recognition accuracy obtained with the Viterbi algorithm is not significantly different from that obtained with the Forward algorithm. In the system, the Viterbi algorithm is used in the recognition process.

2.2.3 Language Models

The statistical model that gives $P(W)$ for each word sequence W is called a *language model* and its parameters are estimated from a very large linguistic corpus.

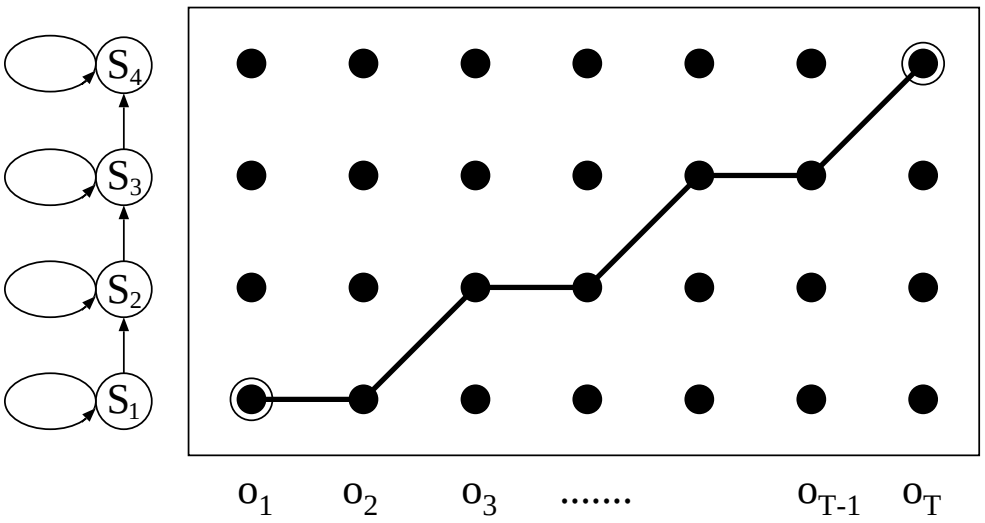


Figure 2.5: Viterbi algorithm.

Let $W = w_1 w_2 \dots w_L$ be a word sequence in which L is the number of words in W. Then the probability $P(W)$ can be written as:

$$\begin{aligned} P(W) &= P(w_1 w_2 \dots w_L) \\ &= P(w_1)P(w_2|w_1)\dots P(w_L|w_{L-1}, \dots, w_2, w_1). \end{aligned} \tag{2.17}$$

The probabilities for all the possible word sequences are almost impossible to estimate from a limited amount of data. Thus, the assumption most often used is that the probability of each word is affected only by the $(N-1)$ preceding words:

$$P(w_i|w_1, w_2, \dots, w_{i-1}) \simeq P_N(w_i|w_{i-1}, w_{i-2}, \dots, w_{i-N+1}). \tag{2.18}$$

This model is called an N-gram model. Let $C(w_i, w_{i+1}, \dots, w_{i+N-1})$ be the number of occurrences of the word sequence $w_i w_{i-1} \dots w_{i-N+1}$ in the training corpus. Then, the probability $P_N(w_i)$ for each word w_i is calculated as follows:

$$P_N(w_i|w_{i-1}, \dots, w_{i-N+1}) = \frac{C(w_i, w_{i-1}, \dots, w_{i-N+1})}{C(w_{i-1}, \dots, w_{i-N+1})}. \tag{2.19}$$

Usually, a *bigram* in which $N = 2$, or a *trigram* in which $N = 3$ is used.

2.3 Summarization of each sentence utterance

2.3.1 An Approach of Speech Summarization

Our proposed method to summarize speech, sentence by sentence, extracts a set of words maximizing a summarization score from an automatically transcribed sentence according to a target compression ratio. The summarization score indicates goodness of a summarized sentence, and it consists of a word significance score I as well as a confidence score C of each word in the original sentence, a linguistic score L of the word string in the summarized sentence [21] [23], and a word concatenation score Tr. The word concatenation score indicates a word concatenation probability determined by a dependency structure in the original sentence given by SDCFG [22]. The total score is maximized using a dynamic programming (DP) technique [21] [23].

This method is effective in reducing the number of words by removing redundant and irrelevant information without losing relatively important information.

Given a transcription result consisting of N words, $W = w_1, w_2, \ldots, w_N$, the summarization is performed by extracting a set of $M(M < N)$ words, $V = v_1, v_2, \ldots, v_M$, which maximizes the summarization score given by eq. (2.20).

$$S(V) = \sum_{m=1}^{M} \{L(v_m|\ldots v_{m-1}) + \lambda_I I(v_m) + \lambda_C C(v_m) + \lambda_T Tr(v_{m-1}, v_m)\} \quad (2.20)$$

where λ_I, λ_C and λ_T are weighting factors for balancing among I,L, C and Tr.

2.3.2 Word significance score

The word significance score I indicates relative significance of each word in the original sentence [21]. The amount of information based on the frequency of each word given by eq.(2.21) is used as the word significance score for topic words.

$$I(w_i) = f_i \log \frac{F_A}{F_i} \quad (2.21)$$

where,

w_i : a word in the transcribed speech
f_i : number of occurrences of w_i in the transcribed article
F_i : number of occurrences of w_i in all the training news articles
F_A : summation of all F_i in all the training news articles($= \sum_i F_i$)

This equation means that the significance is higher when the word is more frequent in the target article and is relatively rarer in all the training data. Therefore it is inferred to be similar to a $tf * idf$ (term-frequency * inverse-document-frequency) broadly used in the field of information retrieval. However the above score gave better performance than that of the $tf * idf$ in a preliminary experiment described in Chapter 3.

To improve the summarization, it is effective to limit words to which the significance scores are applied. In an article, there exist topic words associated with a topic of the article. Such words are usually content words (nouns, verbs, etc). If the score is applied to all words, summarization results are greatly influenced by frequent non-topic words such as function words (particles, prepositions, etc). Therefore a set of words characterizing topics, $TopicWord$, and a flat score for the words except for $TopicWord$, $const$, are introduced, and the equation is rewritten as eq. (2.22). And furthermore to reduce the repetition of words in the summarized sentence, the flat score is also given to each reappearing topic words.

$$I(w_i) = \begin{cases} f_i \log \frac{F_A}{F_i} & if \;\; w_i \in TopicWord, \\ & \quad \text{and } w_i \text{ is first appearing in the article.} \\ const & \text{otherwise} \end{cases} \tag{2.22}$$

Where $const$ is set to relatively low value. We choose only nouns including words that mean verbal sense as $TopicWord$ in Japanese, while we choose nouns and verbs for English.

2.3.3 Linguistic score

The linguistic score $L(v_m| \ldots v_{m-1})$ indicates the appropriateness of the word strings in a summarized sentence. and is measured by a bigram probability $P(v_m|v_{m-1})$ or trigram probability $P(v_m|\ v_{m-2}v_{m-1})$ [21]. In contrast with the word significance score which focuses on topic words, the linguistic score is helpful to extract other words necessary to construct a readable sentence.

2.3.4 Confidence score

The confidence score $C\ (v_m)$ is incorporated to weight acoustically as well as linguistically reliable hypotheses [23]. This score is aimed at extracting high-confidence words (clearly spoken words) and excluding low-confidence words (recognition errors). Specifically, a posterior probability of each transcribed word, that is the ratio of a word hypothesis

probability to that of all other hypotheses, is calculated using a word graph obtained by a recognizer and used as a confidence measure [28] [29].

Word-graph-based posterior probabilities

A word graph consisting of nodes and links from a beginning node S to an end node T in time course is shown in Fig.2.6.

Nodes represent time boundaries between possible word hypotheses and links connecting these nodes represent word hypotheses. Each link is given acoustic log likelihood and linguistic log likelihood of a word hypothesis.

The posterior probability of a word hypothesis $w_{k,l}$ is given by

$$C(w_{k,l}) = \log \frac{\alpha_k P_{ac}(w_{k,l}) P_{lg}(w_{k,l}) \beta_l}{\mathcal{G}} \quad , \tag{2.23}$$

where,

k,l	: node number in a word graph ($k < l$)
$w_{k,l}$	: word hypothesis occurred between node k and node l
$C(w_{k,l})$	: log of the posterior probability of $w_{k,l}$
α_k	: forward probability from the beginning node S to node k
β_l	: backward probability from node l to the end node T
$P_{ac}(w_{k,l})$	: acoustic likelihood of $w_{k,l}$
$P_{lg}(w_{k,l})$	: linguistic likelihood of $w_{k,l}$
$\mathcal{G}$	: forward probability from the beginning node S to the end node $T(= \alpha_T)$

2.3.5 Word concatenation score

Suppose "the beautiful cherry blossoms in Japan" is summarized as "the beautiful Japan". The latter phrase is grammatically correct but a semantically incorrect summarization. Since the above linguistic score is not powerful enough to alleviate such a problem, a word concatenation score $T(v_{m-1}, v_m)$ is incorporated to give a penalty for a concatenation between words with no dependency in the original sentence. Every language has its own dependency structure, and in Section 2.3.5, a basic computation of the word concatenation score independent of the type of language is described. In the following section, this computation is adjusted to process the dependency structure specific to the Japanese language.

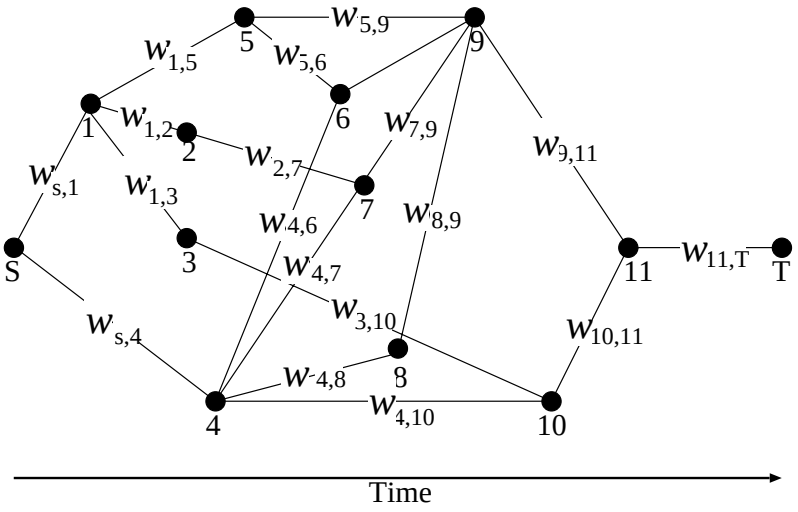


Figure 2.6: An example of word graph.

2.3.5.1 Basic approach

Dependency structure

An example of the dependency structure represented by a dependency grammar is shown as the curved arrows in Figure 2.7. In a dependency grammar, one word is designated as the

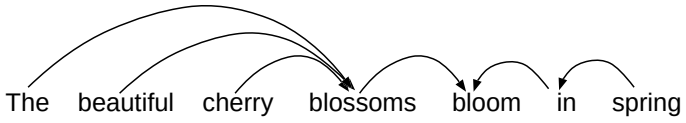


Figure 2.7: *An example of dependency structure.*

head of a sentence, and all other words are either a dependent of that word, or dependent on some other word which connects to the head word through a sequence of dependencies [30]. The word at the beginning of an arrow is named the "modifier" and the word at the end of the arrow is named the "head" respectively. For instance, the dependency grammar of English consists of both "right-headed" dependency indicated by right arrows and "left-headed" dependency indicated by left arrows as shown in Figure 2.7. These dependencies can be represented by a phrase structure grammar, DCFG (Dependency Context Free

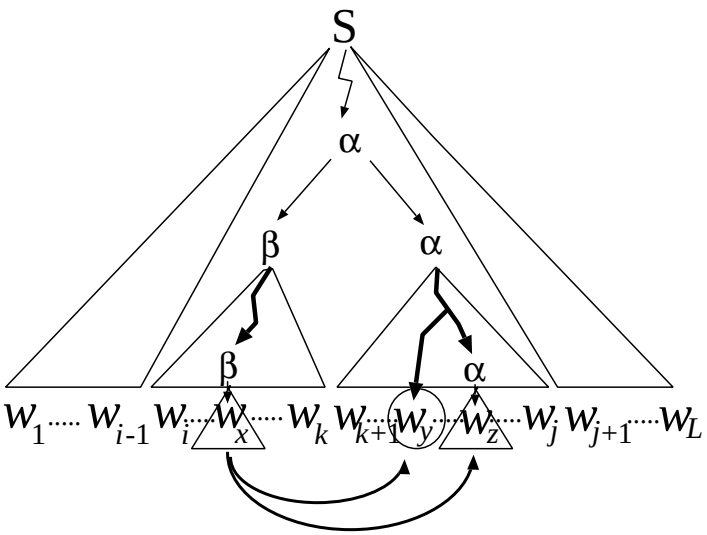


Figure 2.8: *A phrase structure tree based on a word-based dependency structure.*

Grammar), using the following rewrite rules based on Chomsky normal form.

$$\alpha \rightarrow \beta\alpha \quad \texttt{(right-headed)}$$

$$\alpha \rightarrow \alpha\beta \quad \texttt{(left-headed)}$$

$$\alpha \rightarrow w$$

where α and β are nonterminal symbols and w is a terminal symbol (word). Figure 2.8 illustrates an example of a phrase structure tree based on a word-based dependency structure for a sentence which consists of L words, $w_1, \ldots, w_L$. The w_x modifies w_z, when a sentence is derived from the initial symbol S and the following requirements are fulfilled: 1) the rule of $\alpha \rightarrow \beta\alpha$ is applied; 2) $w_i \ldots w_k$ is derived from β; 3) w_x is derived from β; 4) $w_{k+1} \ldots w_j$ is derived from α and 5) w_z is derived from α.

Dependency probability

Since dependencies between words are usually ambiguous, whether dependencies exist or not between words must be estimated by a dependency probability that one word is modified by others. In this study, the dependency probability is calculated as a posterior probability estimated by the Inside-Outside probabilities [31] based on SDCFG.

The probability that the w_x and w_z relationship has a "right-headed" dependency

structure is calculated as a product of the probabilities of the above-mentioned steps from 1) to 5). On the other hand, the "left-headed" dependency probability is calculated as the product of the probabilities when the rule of $\alpha \rightarrow \alpha\beta$ is applied. Since English has both right and left dependencies, the dependency probability is defined as the sum of the "right-headed" and "left-headed" dependency probabilities. If a language has only "right-headed" dependency, the "right-headed" dependency probability is used for the dependency probability. For simplicity, the dependency probabilities between w_x and w_z is denoted by $d(w_x, w_z, i, k, j)$, where i, k are the indices of the initial and final words derived from β, and j is the index of the final word derived from α. The dependency probability $d(w_x, w_z, i, k, j)$ with both right-headed and left-headed dependency is calculated as follows:

$$
\begin{aligned}
& d(w_m, w_l, i, k, j) \\
& = \Bigg\{ \sum_{\alpha\beta} f(i, j|\alpha) P(\alpha \rightarrow \beta\alpha) h_m(i, k|\beta) h_l(k+1, j|\alpha) \\
& \qquad + \sum_{\alpha\beta:\alpha\neq\beta} f(i, j|\alpha) P(\alpha \rightarrow \alpha\beta) h_m(i, k|\alpha) h_l(k+1, j|\beta) \Bigg\} \qquad (2.24)
\end{aligned}
$$

where P is a rewrite probability and f is an outside probability given by eq. (4.7) in Section 4. h is a head-dependent inside probability that w_n is a head of a word string derived from α, which is defined as follows:

$$
\begin{aligned}
& h_n(i, j|\alpha) \\
& = \begin{cases}
\displaystyle\sum_{\beta} \left\{ \sum_{k=i}^{n-1} P(\alpha \rightarrow \beta\alpha) e(i, k|\beta) h_n(k+1, j|\alpha) + \sum_{k=n}^{j-1} P(\alpha \rightarrow \alpha\beta) h_n(i, k|\alpha) e(k+1, j|\beta) \right\} & \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad if \quad i < j & \\
P(\alpha \rightarrow w_n) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad if \quad i = j = n & \\
0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad otherwise &
\end{cases} \qquad (2.25)
\end{aligned}
$$

where e is the inside probability given by eq. (4.6) in Section 4.

Word concatenation probability

In a summarized sentence generated from the example in Figure 2.7, "beautiful" can be directly connected with "blossoms" and also with "cherry" which modifies "blossoms". In general, as shown in Figure 2.8, a modifier derived from β can be directly connected with a head derived from α in a summarized sentence. In addition, the modifier can be also

connected with each word which modifies the head. The word concatenation probability between w_x and w_y is defined as a sum of the dependency probabilities between w_x and w_y, and between w_x and each of the $w_{y+1} \ldots w_z$. Using the dependency probabilities $d(w_x, w_y, i, k, j)$, the word concatenation score is calculated as a logarithmic value of the word concatenation probability given by:

$$T(w_x, w_y) = \log \sum_{i=1}^{x} \sum_{k=x}^{y-1} \sum_{j=y}^{L} \sum_{z=y}^{j} d(w_x, w_z, i, k, j). \tag{2.26}$$

2.3.5.2 Word concatenation score for Japanese

Japanese has a different dependency structure from English. In order to efficiently summarize Japanese speech, the word concatenation score must be converted for the dependency structure of Japanese. Japanese sentences are divided into phrase-like units (*bunsetsu*) as exemplified in Figure 2.9. We denote the phrase-like unit *bunsetsu* by 'phrase'. Since each

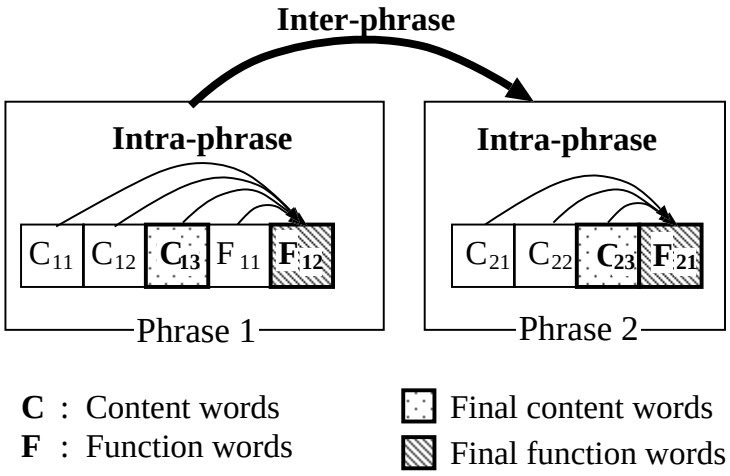


Figure 2.9: Japanese phrase-based dependency structure.

content word always starts a new phrase, it is easy to convert a sentence into a phrase sequence. According to the modification rules for Japanese, a content word modifies function words following it, and forms one phrase. Each phrase is made up of a content word followed by zero or more function words, and each word modifies succeeding words within the phrase.

Japanese sentences have only "right-headed" dependency indicated by right arrows in Figure 2.9. In addition, word dependency structures in each phrase are deterministic and can be represented by the regular grammar. The dependency structures of Japanese sentences can be represented by *inter-phrase* and *intra-phrase* dependencies. The dependency structures between phrases (*inter-phrase* dependency) can be represented as follows:

$$\alpha \rightarrow \beta\alpha$$
$$\alpha \rightarrow P$$

where P is a phrase. On the other hand, the dependency structures between words in each phrase (*intra-phrase* dependency) can be represented as follows:

$$\alpha \rightarrow \beta w$$
$$\alpha \rightarrow w$$

where w is a word. A word concatenation probability between words within a phrase of the original sentence is calculated using *intra-phrase word concatenation probability* based on a rule described below. Word concatenation probability between words in different phrases is calculated using *inter-phrase word concatenation probability* based on a phrase-based SDCFG.

Intra-phrase word concatenation probability

Since a dependency structure between words within a phrase is deterministic in Japanese, *intra-phrase word concatenation probability* is set to 0 or 1 by the *intra-phrase word concatenation rule* consisting of the following 4 rules.

1. A phrase boundary can be connected to any content words in the succeeding phrase.
2. The final content word or the final function word in a phrase can be connected to the succeeding phrase boundary.
3. Each word in a phrase can be connected to the next word in the same phrase.
4. A phrase boundary can be connected to any following phrase boundaries.

Figure 2.10 illustrates word concatenations allowed in a summarized sentence based on the *intra-phrase word concatenation rule* for a sentence consisting of 2 phrases in Figure 2.9.

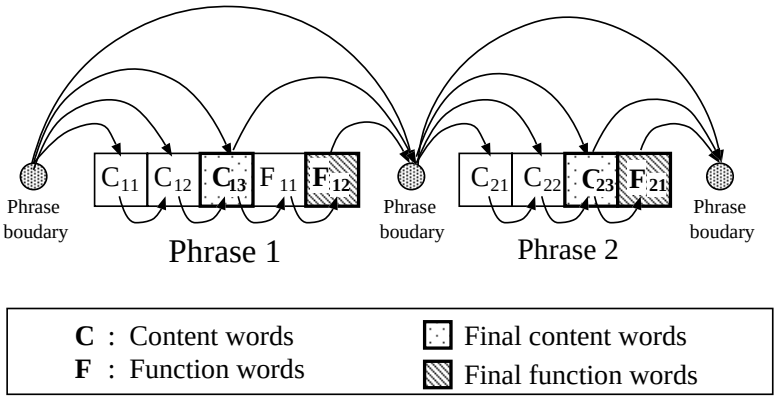


Figure 2.10: Intra-phrase rule.

The arrows toward the right direction indicate possible concatenations between words within a phrase in a summarized sentence. Word concatenation probabilities between words within a phrase in the original sentence satisfying the *intra-phrase word concatenation rule* in Figure 2.10 are set to 1, and probabilities between words without satisfying the rule are set to 0. Summarizing a sentence based on the *intra-phrase word concatenation rule* is exemplified using "phrase 1" in Figure 2.10. The summarization process is one of the following types of word extractions.

1. No word is extracted from a phrase.

2. Only the final content word is extracted.
   $\{C_{13}\}$

3. Content word sequences including the final content words are extracted.
   $\{C_{12}C_{13}\}$, $\{C_{11}C_{12}C_{13}\}$

4. The final content word or content word sequence are attached to all function words.
   $\{C_{13}F_{11}F_{12}\}$, $\{C_{12}C_{13}F_{11}F_{12}\}$, $\{C_{11}C_{12}C_{13}F_{11}F_{12}\}$

Inter-phrase word concatenation probability

A word concatenation probability between words in different phrases is determined by a dependency structure between phrases. Since dependency between phrases is ambiguous, an *inter-phrase word concatenation probability* is calculated as a probability (phrase

dependency probability) that one phrase is modified by others based on a phrase-based SDCFG [22].

The dependency probability between phrases is represented using the dependency probability between words described in 2.3.5.1. Suppose a sentence consists of M phrases, $P_1, \ldots, P_M$, the phrase dependency probabilities between P_x and P_z $(1 \leq x \leq z \leq M)$ is defined as $d_p(P_x, P_z, m, l, n)$ by converting a word dependency probability as shown in Figure 2.8 in 2.3.5.1, where M, m, l and n in $d_p(P_x, P_z, m, l, n)$ correspond to L, i, k and j in $d(w_x, w_z, i, k, j)$ respectively.

Using the phrase dependency probabilities $d_p(P_x, P_z, m, l, n)$, the word concatenation score $T_p(P_x, P_y)$ between words in different phrases is calculated by:

$$T_p(P_x, P_y) = \log \sum_{m=1}^{x} \sum_{l=x}^{y-1} \sum_{n=y}^{M} \sum_{z=y}^{n} d_p(P_x, P_z, m, l, n). \tag{2.27}$$

Since Japanese sentences can be represented only by the rule of $\alpha \rightarrow \beta\alpha$, the final phrase P_l, in a phrase string $P_m, \ldots, P_l$ derived from β, is always derived from the same nonterminal symbol β. The final phrase P_n, in a phrase strings $P_{l+1}, \ldots, P_n$ derived from α, is also derived from the same nonterminal symbol α. As shown in Figure 2.11, the phrase dependency structure is simpler than the general word dependency structure illustrated in Figure 2.8. Therefore, applying only $\alpha \rightarrow \beta\alpha$ results in $l = x$ and $z = n$. The word concatenation score $T_p(P_x, P_y)$ given by eq. (2.27) is simplified as follows:

$$T_p(P_x, P_y) = \log \sum_{m=1}^{x} \sum_{n=y}^{M} d_p(P_x, P_y, m, x, n) \tag{2.28}$$

Here, $d_p(P_x, P_y, m, x, n)$ is calculated as a posterior probability estimated using the Inside-Outside probability [31] based on a phrase-based SDCFG described in Section 4:

$$d(P_x, P_y, m, x, n) = \sum_{\alpha,\beta} g(\alpha \rightarrow \beta\alpha; m, x, n) \tag{2.29}$$

Computation of word concatenation score for Japanese

Suppose w_x and w_y belong to $P_{ph(w_x)}$ and $P_{ph(w_y)}$ respectively, where $ph(w)$ denotes an index of a phrase including a word w. A word concatenation score of w_x and w_y within a phrase $(ph(w_x) = ph(w_y))$ is calculated using the *intra-phrase word concatenation rule*

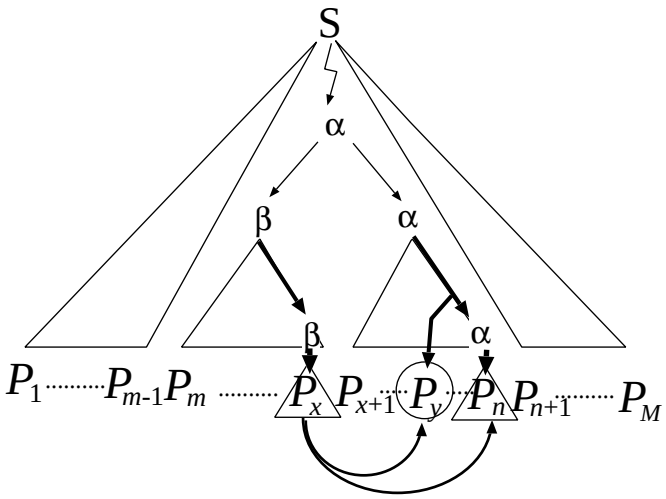


Figure 2.11: A phrase structure tree based on a phrased-based dependency structure.

$(R(w_x, w_y) = 0, 1)$. On the other hand, the word concatenation score when w_x and w_y occur in different phrases $(ph(w_x) < ph(w_y))$ is calculated using a dependency probability between $P_{ph(w_x)}$ and $P_{ph(w_y)}$ based on phrase-based SDCFG. The word concatenation score $T(w_x, w_y)$ is calculated as a logarithmic value of the word concatenation probability as follows:

$$T(w_x, w_y) = \begin{cases} T_p(P_{ph(w_x)}, P_{ph(w_y)}) & if\ ph(w_k) < ph(w_l) \\ \log R(w_k, w_l) & if\ ph(w_k) = ph(w_l) \end{cases} \tag{2.30}$$

2.3.5.3 SDCFG

SDCFG is constructed using a manually parsed corpus. Parameters of SDCFG are estimated using the Inside-Outside algorithm as described in Section 4. In our SDCFG [22], only the number of non-terminal symbols is determined and all possible phrase trees are considered. The rules consisting of all combinations of non-terminal symbols are applied to each rewriting symbol in a phrase tree. In this method, the non-terminal symbol is not given a specific function such as a noun phrase function, and the function of non-terminal symbols are automatically learned from data. Probabilities for frequently used rules become bigger, and those for rarely used rules become smaller. Since words in the

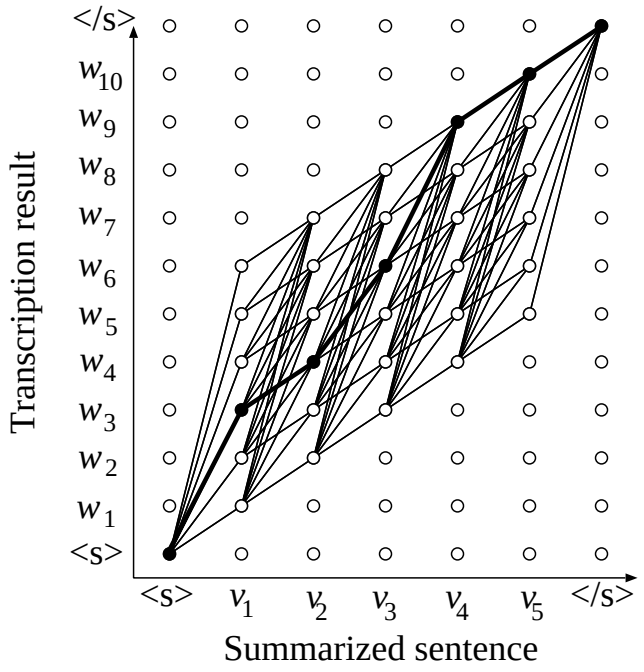


Figure 2.12: An example of DP alignment for speech summarization.

learning data for SDCFG are tagged with POS (part-of-speech), the dependency probability of words excluded in the learning data can be calculated based on their POS. Even if the transcription results obtained by a speech recognizer are ill-formed, the dependency structure can be robustly estimated by the SDCFG.

2.3.6 Dynamic programming for automatic summarization

Given a transcription result consisting of N words, $W = w_1, w_2, \ldots, w_N$, the summarization is performed by extracting a set of $M (M < N)$ words, $V = v_1, v_2, \ldots, v_M$, which maximizes the summarization score given by eq. (2.20) using a Dynamic Programming (DP) technique. An example of the two-dimensional space for performing the dynamic programming process is shown in Figure 2.12. The vertical axis indicates the transcription result consisting of 10 words, and the horizontal axis indicates the summarized sentence having 5 words. All possible sets of 5 words extracted from the 10 words are indicated by the paths from the bottom-left corner to the top-right corner.

The algorithm is as follows:

1. Definition of symbols and variables

$\langle s \rangle$	: beginning symbol of a sentence
$\langle /s \rangle$	: ending symbol of a sentence
$P(w_n \mid w_k w_l)$	: linguistic score
$I(w_n)$	: word significance score
$C(w_n)$	: confidence score
$T(w_l, w_n)$	: word concatenation score
$s(k,l,n)$	: summarization score of each word $s(k,l,n) = \log P(w_n \mid w_k w_l) + \lambda_I I(w_n) + \lambda_C C(w_n) + \lambda_T T(w_l, w_n)$
$g(m,l,n)$	: summarization score of a sub-sentence $\langle s \rangle, \ldots, w_l, w_n$, consisting of m words, beginning from <s>, and ending w_l,w_n $(0 \leq l < n \leq N)$
$B(m,l,n)$	: back pointer

2. Initialization

Summarization score is calculated for each sub-sentence hypothesis consisting of one word. $-\infty$ is given for each word which is never selected as the first word in the summarization sentence consisting of M words.

$$g(1,0,n) = \begin{cases} \log P(w_n \mid \langle s \rangle) + \lambda_I I(w_n) + \lambda_C C(w_n) & if\ 1 \leq n \leq (N-M+1) \\ -\infty & otherwise \end{cases}$$

3. The DP process

A dynamic programming recursion is applied for each pair of the last two words (w_l,w_n) of each sub-sentence hypothesis consisting of m words.

for $m = 2$ to M
  for $n = m$ to $N - m + 1$
    for $l = m - 1$ to $n - 1$

$$g(m,l,n) = \max_{k<l} \{ g(m-1,k,l) + s(k,l,n) \}$$
$$B(m,l,n) = \operatorname*{argmax}_{k<l} \{ g(m-1,k,l) + s(k,l,n) \}$$

4. Select the optimal path

The best complete hypothesis consisting of M words is decided by selecting the last two words ($w_{\hat{l}}$,$w_{\hat{n}}$).

$$S(V) = \max_{\substack{N-M<n\leq N \\ N-M-1<l\leq N-1}} g(M,l,n) + \log P(\langle /s \rangle \mid w_l w_n)$$

$$(\hat{l}, \hat{n}) = \operatorname*{argmax}_{\substack{N-M<n\leq N \\ N-M-1<l\leq N-1}} g(M,l,n) + \log P(\langle /s \rangle \mid w_l w_n)$$

5. Backtracking

   We can get the word sequence $V = v_1 \dots v_M$ of the best summarization result by backtracking the back pointers retained in 3.

for $m = M$ to 1
$v_m = w_{\hat{n}}$
$l' = B(m, \hat{l}, \hat{n})$
$\hat{n} = \hat{l}$
$\hat{l} = l'$

2.4 An approach of Multiple Utterances Summarization

2.4.1 Summarization of multiple utterances using DP technique

Our proposed automatic speech summarization technique for each sentence can be extended to summarize a set of multiple utterances (sentences) having consistent meanings by combining a rule which gives restrictions at sentence boundaries. As a result, original sentences including many informative words are preserved, and sentences including few informative words are deleted or shortened.

Given a transcription result consisting of J utterances, $S_1, \dots, S_J$ ($S_j = w_{j1}, w_{j2}, \dots, w_{jN_j}$) the summarization is performed by extracting a set of $M (M < \sum_j N_j)$ words, $V = v_1, v_2, \dots, v_M$ which maximizes the summarization score given by eq. (2.20). The algorithm is as follows:

1. Definition of symbols and variables
   - $s_j(k, l, n)$: summarization score of each word
     $s_j(k, l, n) = \log P(w_{jn}|w_{jk}w_{jl}) + \lambda_I I(w_{jn}) + \lambda_C C(w_{jn}) + \lambda_T T(w_{jl}, w_{jn})$
   - $g_j(m, l, n)$: local optimal score of $\langle s \rangle, w_{11}, \dots, w_{jl}, w_{jn}$ consisting of m words beginning with $\langle s \rangle$ of the sentence S_1 and ending with w_{jl},w_{jn} in the sentence S_j $(0 \leq l < n \leq N_j)$
   - $G_j(m)$: local optimal score at the end of the sentence, consisting of m words beginning with $\langle s \rangle$ of the sentence 1 and ending with $\langle /s \rangle$ in the sentence j
   - $b_j(m, l, n)$: back pointer
   - $B_j(m)$: back pointer of the end of a sentence

2. Initialization

$$G_0(m) = \begin{cases} 0 & m = 0 \\ -\infty & otherwise \end{cases}$$
$$B_0(m) = \phi$$

3. The DP process

Dynamic programming recursion is applied and the summarization score is summed up through sentences $S_1 \ldots S_J$.

for $j = 1$ to J

calculation for the beginning of a sentence : the summarization score is calculated as the score up

to the preceding sentence, $G_{j-1}(m-1)$, plus the score for the first one word selected from the

current sentence.

$$g_j(m, 0, n) = \begin{cases} G_{j-1}(m-1) + \log P(w_{jn}|\langle s\rangle) + \lambda_I I(w_{jn}) + \lambda_C C(w_{jn}) & if\ 1 \le n \le N_j \\ -\infty & otherwise \end{cases}$$
$$b_j(m, 0, n) = \phi$$

calculation for the inside of a sentence : DP recursion is applied for each sentence in the same

manner as that of sentence-by-sentence summarization described in Section 2.3.6:

for $m = j \times 2$ to N_j
  for $n = 2$ to N_j
    for $l = 1$ to $n - 1$

$$g_j(m, l, n) = \max_{0 \le k < l} \{g_j(m-1, k, l) + s_j(k, l, n)\}$$
$$b_j(m, l, n) = \operatorname*{argmax}_{0 \le k < l} \{g_j(m-1, k, l) + s_j(k, l, n)\}$$

calculation for the end of a sentence : the score of the local best hypothesis up to the end of S_j is

calculated:

$$G_j(m) = \max_{\substack{0 < n \le N_j \\ 0 \le l \le N_j - 1}} g_j(m, l, n) + \log P(\langle /s\rangle | w_{jl} w_{jn})$$
$$(\hat{l}, \hat{n}) = \operatorname*{argmax}_{\substack{0 < n \le N_j \\ 0 \le l \le N_j - 1}} g_j(m, l, n) + \log P(\langle /s\rangle | w_{jl} w_{jn})$$
$$B_j(m) = (\hat{l}, \hat{n})$$

4. Backtracking

   We can get the word sequence $V = v_1 \dots v_M$ of the best summarization result for the multiple utterances by backtracking the back pointers retained within each sentence and at the end of each sentence, where:

   $j = J$
   $m = M$
   while $m > 0$
     $v_m = w_{\hat{n}}$
     $l' = b_j(m, \hat{l}, \hat{n})$
     $\hat{n} = \hat{l}$
     if $l' \neq \phi$ then
       $\hat{l} = l'$
       $m = m - 1$
     else
       $v_{m-1} = \langle /s \rangle$
       $v_{m-2} = \langle s \rangle$
       $(\hat{l}, \hat{n}) = B_{j-1}(m-2)$
       $m = m - 3$
       $j = j - 1$

Figure 2.13 illustrates the DP process for summarizing multiple utterances. This summarization technique can be considered as a combination of the summarization method developed in the field of natural language processing which extracts important sentences, and our sentence-by-sentence summarization method.

2.4.2 Summarization of multiple utterances using two-level DP

However, the amount of calculation for selecting the best combination among all possible combinations of words in the multiple utterances increases as the number of words in the original utterances increases. In order to alleviate this problem, we have proposed a new method in which each utterance is summarized according to all possible summarization ratio and then the best combination of summarized sentences for each utterance is determined according to a target compression ratio using a two-level DP technique. Figure 2.14 illustrates the two-level DP technique for summarizing multiple utterances.

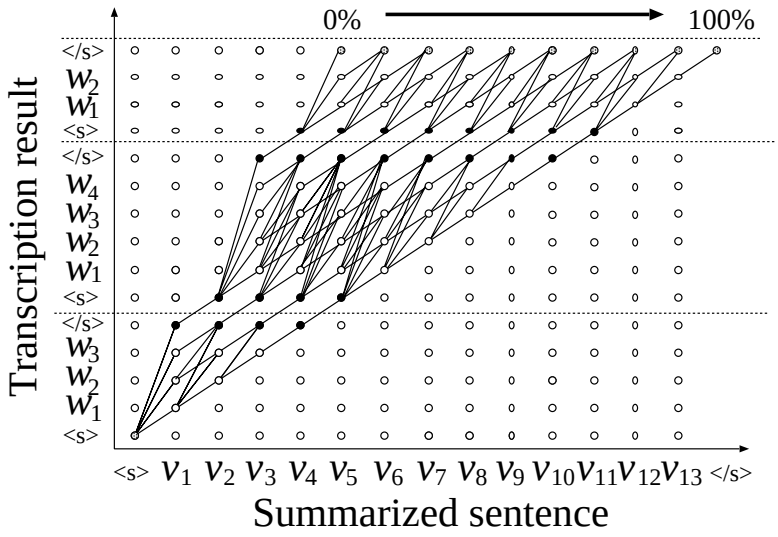
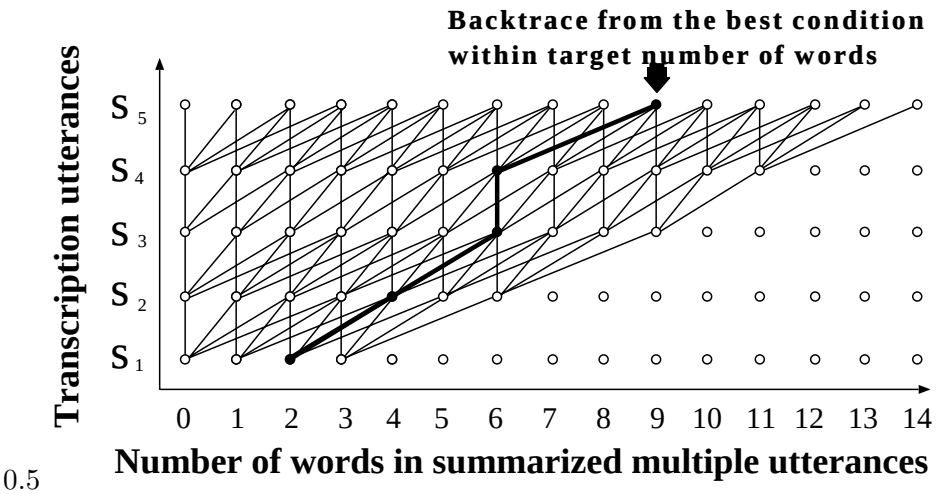


Figure 2.13: An example of DP process for summarization of multiple utterances.

Our proposed automatic speech summarization technique for each sentence has recently been extended to summarize a set of multiple utterances (sentences) [36]. A set of words maximizing the summarization score is extracted from multiple utterances under some restrictions applied at the sentence boundaries. These restrictions realizes the summarization of multiple utterances by handling them as a single long utterance. This results in preserving more words inside information rich utterances and shortening or even completely deleting less informative ones. However, the amount of calculation for selecting the best combination among all possible combinations of words in the multiple utterances increases as the number of words in the original utterances increases. In order to alleviate this problem, we have proposed a new method in which each utterance is summarized according to all possible summarization ratio and then the best combination of summarized sentences for each utterance is determined according to a target compression ratio using a two-level DP technique. Figure 2.14 illustrates the two-level DP technique for summarizing multiple utterances.

1. Definition of symbols and variables



0.5

Figure 2.14: *An example of DP process for summarization of multiple utterances.*

$s_n(l)$: summarization score of a sentence consisting of l words that summarized from sentence S_n
$0 \leq l \leq L_n$, $1 \leq n \leq N$

2. Initialization

$$\begin{aligned} g(1,l) &= s_1(l) \\ B(1,l) &= l \qquad\qquad where \quad 0 \leq l \leq L_1 \\ M &= L_1 \end{aligned}$$

3. DP process
for $n = 2$ to N
$M = M + L_n$
for $m = 0$ to M

$$g(n,m) = \max_{m-L_n \leq l \leq m, l \geq 0} \{g(n-1,l) + s_n(m-l)\}$$

$$B(n,m) = \operatorname*{argmax}_{m-L_n \leq l \leq m, l \geq 0} \{g(n-1,l) + s_n(m-l)\}$$

4. Traceback
for $n = N$ to 1
$l_n = M - B(n,M)$
$M = B(n,M)$
for $n = 1$ to N
Output $S_n(l_n)$

Chapter 3

Topic Score based on Term Weighting

As discussed in Section 2.3.2, a score to detect topic words from the transcribed news speech is applied to the word significance score for informative words.

Topic Detection

Technology for automatically detecting words representing topics of utterances is expected to be applied in many different ways, such as making broadcast news databases accessible with keywords.

Measure for Topic words

The measures based on the frequency of each word given by eq.(3.1) – eq.(3.7) is used as the word significance score for each noun.

$$TopicScore(w_i) = g_i \cdot \log \frac{G_A}{G_i} \tag{3.1}$$

$$TopicScore(w_i) = \frac{g_i}{\log G_i} \tag{3.2}$$

$$TopicScore(w_i) = \frac{g_i}{\log (g_A \cdot G_i)} \tag{3.3}$$

$$TopicScore(w_i) = \frac{g_i}{g_A \cdot G_i} \tag{3.4}$$

$$TopicScore(w_i) = \frac{g_i^2}{g_A \cdot G_i} \tag{3.5}$$

$$TopicScore(w_i) = \hat{g}_i - \hat{G}_i \tag{3.6}$$

$$TopicScore(w_i) = \frac{\hat{g}_i - \hat{G}_i}{\sqrt{\hat{G}_i}} \tag{3.7}$$

where,

w_i : a noun in the transcribed speech ($i = 1, 2, \cdots, N$)
g_i : number of occurrences of w_i in the transcribed article
G_i : number of occurrences of w_i in all the training news articles
G_A : summation of all G_i in all the training news articles($= \sum_i G_i$)

3.1 Evaluation Experiment

Twenty-nine broadcast news articles comprising 142 utterances by 15 male speakers (8 anchors and 7 others) were used for evaluation. Each news article has 2-l4 utterances (5 utterances on the average). Correct topic words were given by three subjects. Ten phrases, which correspond to 24.4 words, were given on the average for each news article by each subject.

Two correct topic word sets were constructed for each news. A "AND" set was made from topic words given by all subjects in common (10.4 words on the average for each news article), and an "OR" set was made from topic words given by at least one subject (35.7 words on the average for each news article). Supplementary experiments were also conducted by giving correct texts instead of transcription results as input.

Results using either the "AND" set or "OR" set as the correct topic word set averaged over the 29 news articles are shown below.

Training Data Set

Nikkei newspaper articles published over a five-year period (879,681 articles, 97400k words) [32] were morpholized by ChaSen [33] and used for calculating the topic score.

3.2 Evaluation Method

The words with higher $TopicScore(w_i)$ are selected as topic words according to the given number of words. In order to test the performance of topic words extraction, the following measures are evaluated using Recall and Precision given by eq. 3.8 and eq. 3.8 respectively.

$$Recall = \frac{C}{T} \cdot 100 \quad (\%)$$

$$Precision = \frac{C}{H} \cdot 100 \quad (\%)$$

where

C : the number of correctly detected topic words
T : the total number of correct topic words
H : the total number of detected topic words

Usually there is a trade-off between Recall and Precision. Precision decreases as Recall increase and Recall increases as Precision decrease. Both Recall and Precision are expected to be higher.

3.2.1 Topic Words Extraction from Transcribed Japanese News Speech

Five, ten, or fifteen topic words are respectively extracted from the "OR"set based on the topic score given by eq.(3.1) and eq. (3.2) – eq.(3.7). The evaluated result using precision is shown in table 3.1.

Table 3.1: The performance of the topic score ("OR" set, Precision[%]).

The number of extracted words	5		10		15	
Topic score	TRANS	REC	TRANS	REC	TRANS	REC
(3.1)	88.3	82.8	82.1	73.4	79.5	66.2
(3.3)	85.5	77.9	78.6	70.0	78.2	64.1
(3.6)	84.8	77.2	79.0	70.3	78.4	65.3
(3.2)	82.1	76.6	76.9	68.6	75.9	65.5
(3.5)	65.5	55.9	67.9	59.0	68.3	58.4
(3.7)	69.6	55.9	70.4	59.0	70.7	58.4
(3.4)	57.2	49.0	63.8	52.4	64.6	53.8
(3.8)	43.4	29.7	45.2	34.1	46.0	34.0

TRS: manual transcription, REC: recognition result

The topic score given by eq. (3.1) performs the best. The value of $\log G_A/G_i$ in eq. (3.1) means the total amount of information provided by w_i. Since G_i is normalized, the score is not directly affected by the size of the training data set. The topic score given by eq. (3.2) and eq. (3.3) also works well due to the application of the logarithmic value. On the other hand, the performance given by eq. (3.7) is worse because of highly weighting $\hat{G}_i$. In order to compare the score for Information Retrieval given by eq. (3.8) was evaluated.

$$TopicScore(w_i) = g_i \cdot \log \frac{M}{F_i} \tag{3.8}$$

where

M : number of articles in all data
F_i : number of articles that noun w_i appears in all data

The result using $tf \cdot idf$ measure is worse because of highly weighting $\log M/F_i$.

In order to compare with the topic extraction for recognition result, topic words are extracted from the manual transcriptions using eq. (3.1), eq. (3.2), eq. (3.3) and eq. (3.6). Figure 3.1 and fig. 3.2 show the Recall-Precision for the recognition result and the manual transcription result respectively.

The errors for extraction of topic words from the recognition results is 1.5 times of those for the manual transcription.

3.2.2 Topic Word Detection using News Manuscript

The topic scores were calculated using manuscripts which were prepared for anchor persons for five year period. In order to compare the performance of the topic scores between using the newspaper text and the news manuscripts, the 1030M words (almost the same size of the manuscripts) in the newspaper text 3.2 were used for calculation.

Table 3.2: Comparison between training sets using the newspaper text and the news manuscript (OR, Precision[%])

Number of extracted words	5		10		15	
Training data	TRANS	REC	TRANS	REC	TRANS	REC
Nikkei newspaper text	89.0	82.1	82.4	73.8	77.9	66.4
News speech	88.3	84.8	83.1	70.0	79.3	66.0

TRS: the manual transcription, REC: the recognition result

There is no significant difference between the performance using the newspaper text and the news manuscripts. The evaluation results using Recall-Precision are shown in fig. 3.3 and fig. 3.4

3.3 Summary

The evaluation result to detect topic words of Japanese broadcast news speech using the topic scores given by eq.(3.1) – eq.(3.7). based on the newspaper text shows that eq. (3.1) is the best. Therefore, topic scores are calculated based on eq. (3.1) in this thesis.

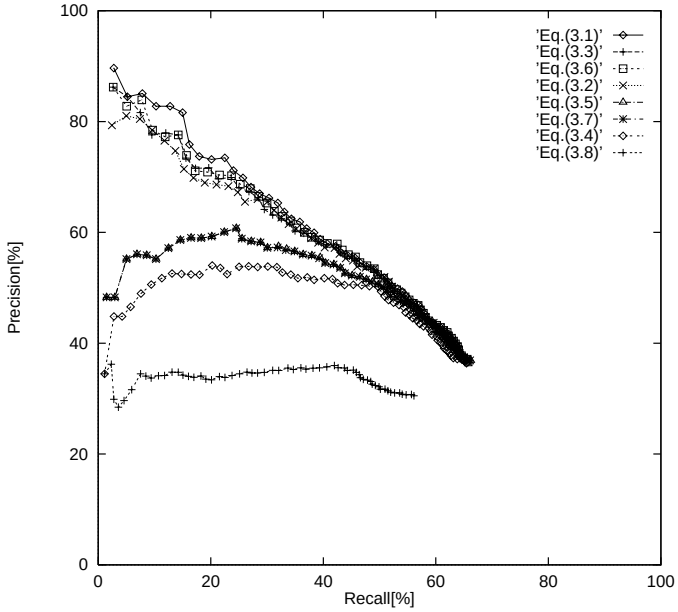


Figure 3.1: The performance for the recognition result("OR" set)

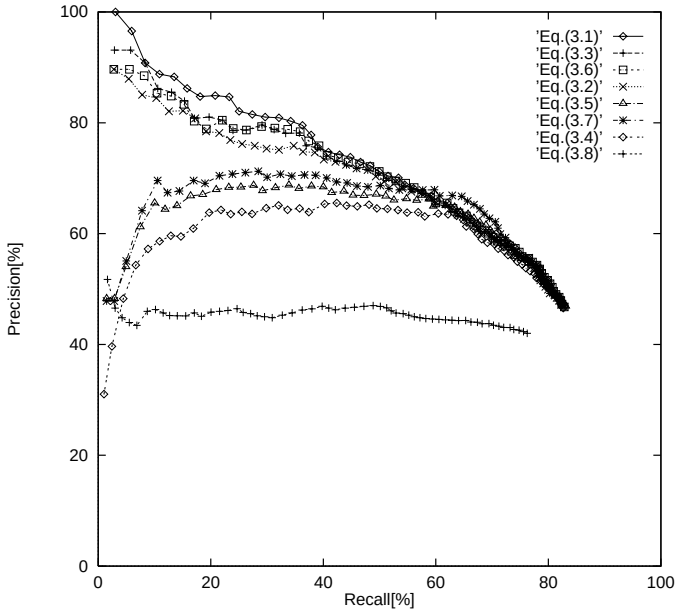


Figure 3.2: The performance for the transcription result ("OR" set)

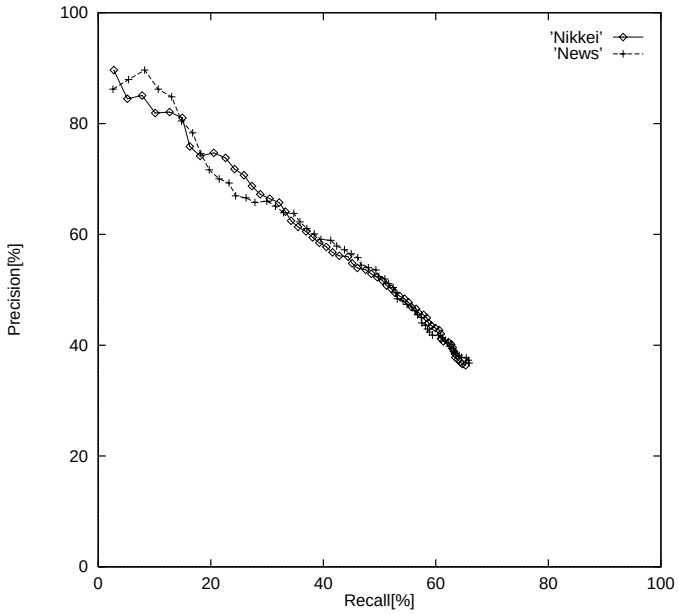


Figure 3.3: Topic extraction from the recognition results("OR" set)

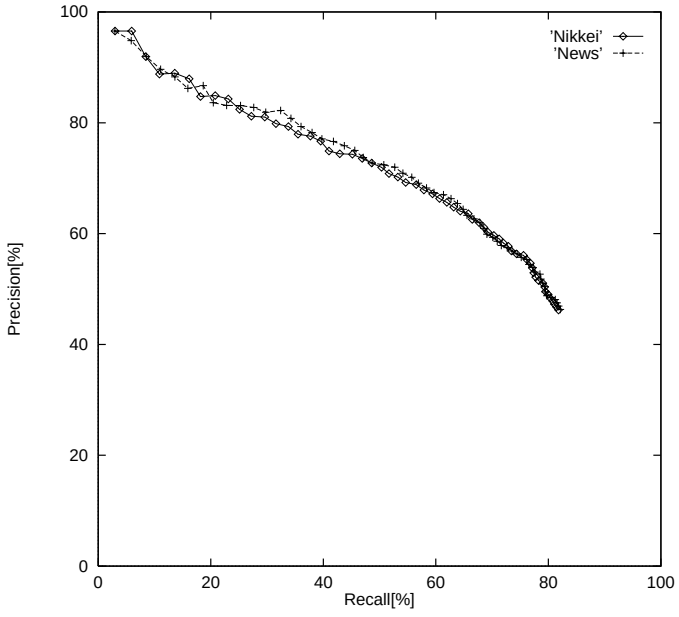


Figure 3.4: Topic extraction from the transcription results("OR" set)

Chapter 4

SDCFG for Word Concatenation Score

This chapter proposes a Stochastic Dependency Context Free Grammar (SDCFG) that is applied to a word concatenation score. In Section 2.3.5, it has already been shown that the word concatenation score can be derived from dependency probabilities between words in a sentence to be summarized based on the SDCFG. This chapter here describes how to estimate parameters of the SDCFG which are used for dependency probabilities.

First, formal language theory and grammars that express the dependency structure based on the NLP techniques are indicated. The syntactic structure of formally written sentences can be captured by deterministic grammars such as Context Free Grammars (CFGs), which is a rule-based approach. On the other hand, since speech is a peculiar language which consists of many informal ill-formed sentences, a stochastic approach to estimate the dependency structure is necessary. In this study, the stochastic approach based on CFG, i.e., Stochastic Dependency Context Free Grammar (SDCFG), is proposed for speech summarization. Since languages have their own dependency structure, an approach to calculate parameters of basic SDCFG and other approaches specific to English and Japanese are described.

4.1 Formal Language Theory

In order to capture the syntactic structure of a sentence, approaches have attempted to represent the grammar of language. Chomsky's formal language theory gives us a basic concept to solve this problem. The grammar is a formal specification of the permissible structures for language. By using parsing techniques, we can analyze the structure of the sentence, and check if it's compliant with the grammar.

The most common way of representing the grammatical structure of a sentence, "Bob

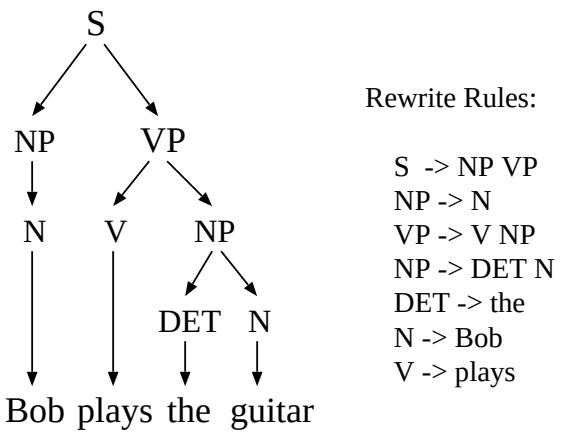


Figure 4.1: A tree representation of a sentence and its corresponding grammar.

plays the guitar," is by using a tree, as illustrated in Figure 4.1. The node labeled S is the parent node of the node labeled NP and VP for noun phrase and verb phrase, respectively. The VP node is the parent node of node V for the verb, and the node NP. Each leaf is associated with the word in the sentence to be analyzed. To construct such a tree for a sentence, we have to know the structure of the language so that a set of rewrite rules can be used to describe what tree structures are allowable. These rules, as illustrated in Figure 4.1, determine that a certain symbol may be expanded in the tree by a sequence of symbols.

In Chomsky's formal language theory, a grammar is defined as $G = (\mathcal{N}, \mathcal{T}, P, S)$, where $\mathcal{N}$ and $\mathcal{T}$ are finite sets of non-terminals and terminals respectively. $\mathcal{N}$ contains all the non-terminal symbols. In the example in Figure 4.1, S,NP,VP,V,N,and DET are included in $\mathcal{N}$. The terminal set $\mathcal{T}$ contains *Bob,plays,the*,and *guitar*. P is a finite set of production (rewrite) rules, $S \rightarrow NP\ VP$, $NP \rightarrow N$, and so on. In general, S is a special non-terminal, called the start symbol.

The language to be analyzed is essentially a string of terminal symbols, such as "*Bob plays the guitar.*" It is produced by applying production rules sequentially to the start symbol. The rewrite rule is of the form $A \rightarrow B$, where A and B are arbitrary strings of grammar symbols $\mathcal{N}$ and $\mathcal{T}$, and the A must not be empty. In formal language theory, four major languages groups and their associated grammars are hierarchically structured. They are referred to as the Chomsky hierarchy as defined in Table 4.1.

Table 4.1: Chomsky hierarchy

Types	**Constraints**
Phrase structure grammar	$A \rightarrow B$. This is the most regular grammar
Context-sensitive grammar	Phrase structure grammar that satisfies $\|A\| < \|B\|$ where $\|.\|$ indicates the length of the string.
Context-free grammar(CFG)	$\alpha \rightarrow A$ where α is a non-terminal. This type of rules can be written as Chomsky normal form: $\alpha \rightarrow w$ and $\alpha \rightarrow \beta\gamma$ where α,β,and γ are non- terminals and w is a terminal.
Regular grammar	The rewrite rule is expressed as: $\alpha \rightarrow w$ and $\alpha \rightarrow w\beta$.

4.2 Dependency Grammar

As shown in the previous section, the dominant tradition within modern linguistics and NLP has been to use phrase structure trees to represent the structure of sentences. However, an alternative, and much older, tradition is to describe linguistic structure in terms of dependencies between words. Such a framework is referred to as a dependency grammar (DG). In a dependency grammar, one word is the head of a sentence, and all other words are either a dependent of that word, or else dependent on some other word which connects to the head word through a sequence of dependencies. Dependencies are usually shown as curved arrows, as for the example in Figure 4.2.

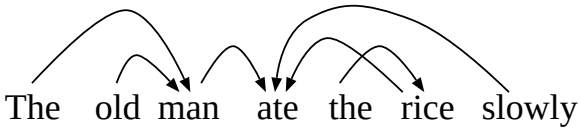


Figure 4.2: A sentence representation based on the dependency grammar.

4.3 Dependency Context Free Grammar

The dependency structure of the sentence can be written as phrase structure models. This type of grammar is referred to Dependency Context Free Grammar (DCFG). The example of the DCFG-based tree representation is illustrated in Figure 4.3. The rewrite rules are

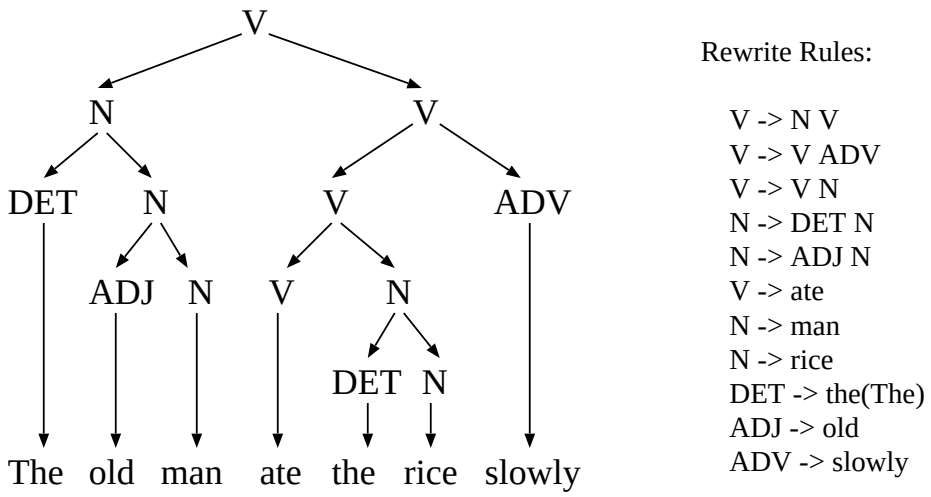


Figure 4.3: A sentence representation based on the dependency context free grammar.

divided based on Chomsky normal form into the following types of rules.

$$\alpha \rightarrow \beta\alpha$$
$$\alpha \rightarrow \alpha\beta$$
$$\alpha \rightarrow w,$$

where the first two types correspond to right-headed (forward) and left-headed (backward) dependencies respectively.

As you can see, the rules in the figure can not force strict constraints word-by-word, because each word is first rewritten as N, V, ADJ, ADV, or DET, each of which represents a corresponding part-of-speech (POS). Therefore, the constraints of DCFG look poorer than those of the general dependency grammar. However, the DCFG can be as isomorphic as the dependency grammar by increasing the number of non-terminals. Furthermore, the DCFG is a more flexible grammar because the balance between the coverage and the strictness of the grammar can also be controlled by the number of non-terminal symbols.

4.4 Stochastic Approach for Phrase Structure Grammar

Whereas rule-based approaches to model the syntactic structure of sentences have been described in the previous sections, this section presents stochastic approaches for speech are presented.

4.4.1 Stochastic Context Free Grammar

The CFG can be augmented with probability for each rewrite rule as follows:

$$\alpha \to \theta, \qquad P(\alpha \to \theta) \in [0, 1]$$
$$\alpha \in \mathcal{N}, \qquad \theta \in (\mathcal{N} \cup \mathcal{T})^*$$

$$\sum_{\theta} P(\alpha \to \theta) = 1$$

This grammar is called a stochastic context free grammar (SCFG), or a probabilistic context free grammar (PCFG). The derivative probability of a sentence, $w_1, w_2, \ldots, w_L (w_i \in \mathcal{T})$, is written as:

$$P(S \to w_1 w_2 \ldots w_L)$$

It can be obtained as a probability that the sentence is produced by recursively applying rewrite rules from S.

The advantages of the SCFGs lie in their ability to more accurately capture the embedded usage structure of spoken language to minimize syntactic ambiguity. The use of probability becomes increasingly important to discriminate among many competing choices when the number of rules is large.

If we have many parsed sentences as a training corpus, the parameters of the SCFG (the probability of each rewrite rule) can be estimated as ML estimates by counting the number of times that each rule is used in the corpus. However, since it is not easy to obtain such corpora, the Inside-Outside algorithm [31] [34] is used. The Inside-Outside algorithm is an EM-based algorithm that makes it possible to estimate the parameters of SCFG from any corpus without parsed sentences.

In the Inside-Outside algorithm, each rewrite rule has to be expressed as a Chomsky normal form:

$$\begin{aligned} \alpha &\to \beta\gamma \\ \alpha &\to w \end{aligned}$$

$$\alpha, \beta, \gamma \in \mathcal{N}, \quad w \in \mathcal{T}$$

The production probability of the rules, $\alpha \to \beta\gamma$ and $\alpha \to w$, are defined as $P(\alpha \to \beta\gamma)$ and $P(\alpha \to w)$ respectively. The algorithm for estimating parameters of the SCFG is described below. Figure 4.5 indicates the estimation steps.

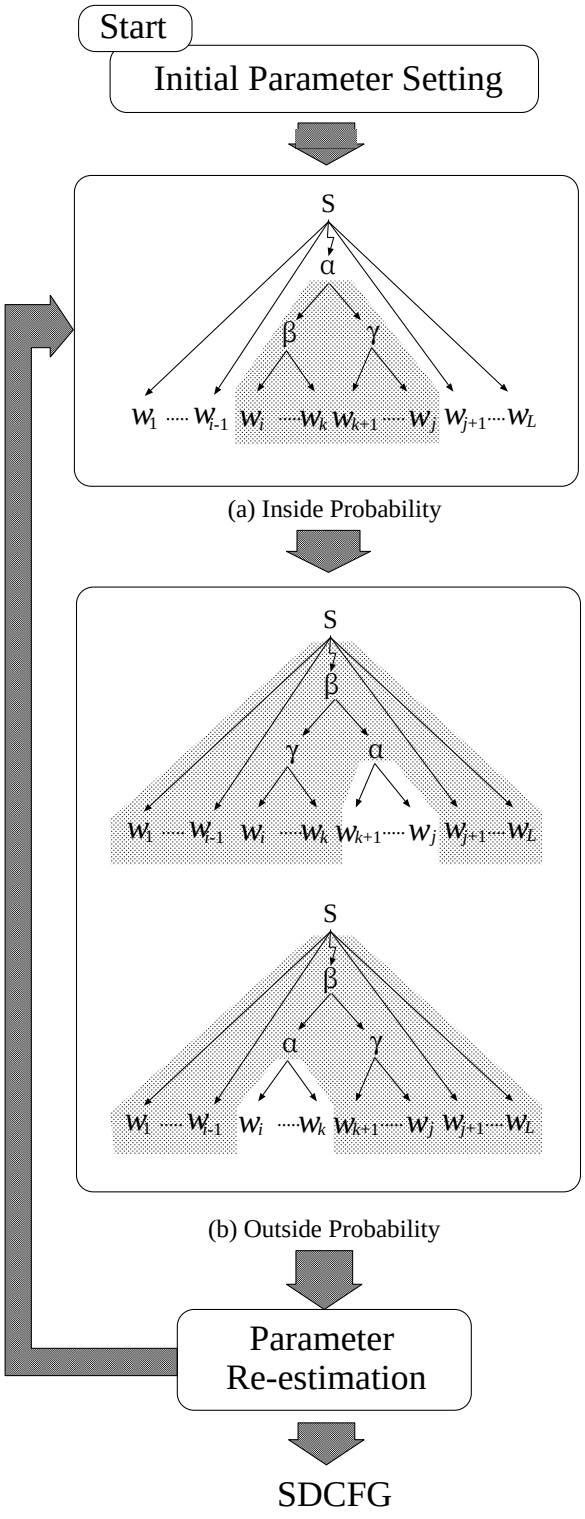


Figure 4.4: Estimation algorithm of SCFG

1. Initialization

   $P(\alpha \to \beta\gamma)$ is given a flat probability and $P(\alpha \to w)$ is given random values.

2. Calculation of the inside probability

   Given a sentence consisting of L words, $w_1, w_2, w_3, \ldots, w_L$ the inside probability is defined as the probability that $w_i \ldots w_j$ is derived from α, illustrated in Figure 4.4(a). The inside probability $e(i, j|\alpha)$ is calculated as follows:

$$
\begin{aligned}
& e(i,j|\alpha) \\
&= P(\alpha \to w_i \ldots w_j) \\
&= \begin{cases} \displaystyle\sum_{k=i}^{j-1} \sum_{\beta\gamma} P(\alpha \to \beta\gamma) e(i,k|\beta) e(k+1,j|\gamma) & if \quad i < j \\ P(\alpha \to w_i) & if \quad i = j \end{cases}
\end{aligned}
\tag{4.1}
$$

3. Calculation of the outside probability

   The outside probability $f(i, j|\alpha)$, illustrated in Figure 4.4(b), is calculated as follows:

$$
\begin{aligned}
& f(i,j|\alpha) \\
&= P(S \to w_1 \ldots w_{i-1} \alpha w_{j+1} \ldots w_L) \\
&= \sum_{k=1}^{i-1} \sum_{\beta\gamma} P(\beta \to \gamma\alpha) e(k, i-1|\gamma) f(k,j|\beta) \\
&\quad + \sum_{k=j+1}^{L} \sum_{\beta\gamma} P(\beta \to \alpha\gamma) e(j+1,k|\gamma) f(i,k|\beta)
\end{aligned}
\tag{4.2}
$$

4. Re-estimation of parameters

   When $w_1, w_2, \ldots w_L$ is derived from S, the probability of applying the rule of $\alpha \to \beta\gamma$ in any location in the derived tree is defined as follows:

$$
\hat{P}(\alpha \to \beta\gamma) = \frac{\displaystyle\sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \sum_{k=i}^{j-1} g(i,k,j;\alpha \to \beta\gamma)}{e(1,L|S)}
\tag{4.3}
$$

where $g(i,k,j;\alpha \rightarrow \beta\gamma)$ is calculated using a production probability that $\beta\gamma$ is derived from α, the inside probability that $w_i \ldots w_k$ is derived from β and $w_{k+1} \ldots w_j$ is derived from γ, respectively, and the outside probability that words other than a word string derived from α are derived from S.

The $g(i,k,j;\alpha \rightarrow \beta\gamma)$ is given as follows:

$$g(i,k,j;\alpha \rightarrow \beta\gamma) = e(i,k|\beta)e(k+1,j|\gamma)P(\alpha \rightarrow \beta\gamma)f(i,j|\alpha) \tag{4.4}$$

While the production probability of w is given as follows:

$$\hat{P}(\alpha \rightarrow w) = \frac{\sum_{i;w_i=w} P(\alpha \rightarrow w)f(i,i|\alpha)}{e(1,L|S)} \tag{4.5}$$

The estimated probabilities given by eqs. (4.3) and (4.5) obtained by the steps 2 to 3 are used for the initial probabilities and then estimated again.

However, the computation of the Inside-Outside algorithm is very expensive. Its computational complexity is $O(N^3L^3)$, where N is the number of non-terminals and L is the number of words in a given sentence, respectively. When the number of non-terminals gets larger, the learning process becomes slower rapidly. If a grammar for NLP is applied to rewrite rules, the number of non-terminals should be more than several hundred, which makes it impossible to estimate probabilities using the Inside-Outside algorithm. In this study, a stochastic approach based CFGs without using grammatical rules of NLP is proposed [22]. This SCFG is calculated using a manually parsed corpus based on the re-estimation algorithm described above. In this approach, only the number of non-terminal symbols is determined and all possible phrase trees are considered. The rules consisting of all combinations of non-terminal symbols are applied to each rewriting symbol in a phrase tree. In addition, the non-terminal symbol is not given a specific function such as a noun phrase function. The role of each non-terminal can be directly derived from text corpora according to the maximum likelihood (ML) criterion. Production probabilities for frequently used rules become bigger, and those for rarely used rules become smaller. It is not necessary to previously define any grammatical rules except for the number of non-terminals.

4.4.2 Stochastic Dependency Context Free Grammar

In this section, in order to estimate dependency structures, Stochastic Dependency Context Free Grammars are proposed.

SDCFG for English

In languages having both right and left dependencies such as English, the rewrite rules of that two nonterminal symbols are derived from a nonterminal symbol is modified as follows based on the dependency grammar:

$$\alpha \rightarrow \beta\alpha$$
$$\alpha \rightarrow \alpha\beta$$
$$\alpha \rightarrow w$$

Whereas the computational complexity in the SCFG is $O(N^3L^3)$, that of SDCFG is $O(N^2L^3)$ by this modification. Where N and L mean the number of non-terminals and the number of words in a sentence, respectively.

Re-estimation of Parameters in SDCFG for English

Suppose a sentence consists of L words,

$$S \rightarrow w_1 \dots w_i \dots w_L$$

Where,

L : number of words in a sentence
w_i : i-th word in a sentence

The rewrite probabilities of $\alpha \rightarrow \beta\alpha$ and $\alpha \rightarrow w$ are denoted by $P(\alpha \rightarrow \beta\alpha)$, and $P(\alpha \rightarrow w)$, respectively. The algorithm for estimating parameters of the SDCFG is described below. Figure 4.5 indicates the estimation steps.

1. Initialization
   $P(\alpha \rightarrow \beta\alpha)$ and $P(\alpha \rightarrow \alpha\beta)$ are given a flat probability, and $P(\alpha \rightarrow w)$ is given random values.

2. Calculation of the inside probability
   The inside probability illustrated in Figure 4.5(a) is calculated as follows:

$$\begin{aligned} e(i,j|\alpha) &= P(\alpha \rightarrow w_i \dots w_j) \\ &= \begin{cases} \sum_{k=i}^{j-1} \left\{ \sum_{\beta} P(\alpha \rightarrow \beta\alpha) e(i,k|\beta) e(k+1,j|\alpha) \right. \\ \qquad \left. + \sum_{\beta:\alpha\neq\beta} P(\alpha \rightarrow \alpha\beta) e(i,k|\alpha) e(k+1,j|\beta) \right\} \\ \qquad\qquad if \quad i < j \\ P(\alpha \rightarrow w_i) \qquad\qquad if \quad i = j \end{cases} \end{aligned} \tag{4.6}$$

3. Calculation of the outside probability

   The outside probability illustrated in Figure 4.5(b) is calculated as follows:

$$
\begin{aligned}
f(i,j|\alpha) &= P(w_1 \dots w_{i-1} \alpha w_{j+1} \dots w_L) \\
&= \sum_{k=1}^{i-1} \left\{ \sum_{\beta} P(\alpha \to \beta\alpha) e(k, i-1|\beta) f(k, j|\alpha) \right. \\
&\qquad \left. + \sum_{\beta:\alpha\neq\beta} P(\beta \to \beta\alpha) e(k, i-1|\beta) f(k, j|\alpha) \right\} \\
&\quad + \sum_{k=j+1}^{L} \left\{ \sum_{\beta} P(\beta \to \alpha\beta) e(j+1, k|\beta) f(i, k|\alpha) \right. \\
&\qquad \left. + \sum_{\beta:\alpha\neq\beta} P(\alpha \to \alpha\beta) e(j+1, k|\beta) f(i, k|\alpha) \right\}
\end{aligned}
\tag{4.7}
$$

4. Estimation of parameters

   The parameters are re-estimated using the probabilities obtained by the steps 2 to 3, in the same way as parameters of SCFG are re-estimated as described in 4.4.1.

$$
\hat{P}(\alpha \to \beta\alpha) = \frac{\sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \sum_{k=i}^{j-1} g(i,k,j;\alpha \to \beta\alpha)}{e(1, L|S)} \tag{4.8}
$$

$$
\hat{P}(\alpha \to w_c) = \frac{\sum_{i=1}^{L} P(\alpha \to w) f(i,j|\alpha)}{e(1, L|S)} \tag{4.9}
$$

   where

$$
g(i,k,j;\alpha \to \beta\alpha) = e(i,k|\beta) e(k+1, j|\alpha) P(\alpha \to \beta\alpha) f(i,j|\alpha) \tag{4.10}
$$

$$
g(i,k,j;\alpha \to \alpha\beta) = e(i,k|\alpha) e(k+1, j|\beta) P(\alpha \to \alpha\beta) f(i,j|\alpha) \tag{4.11}
$$

5. The steps from 2 to 4 are iterated until the parameters are saturated.

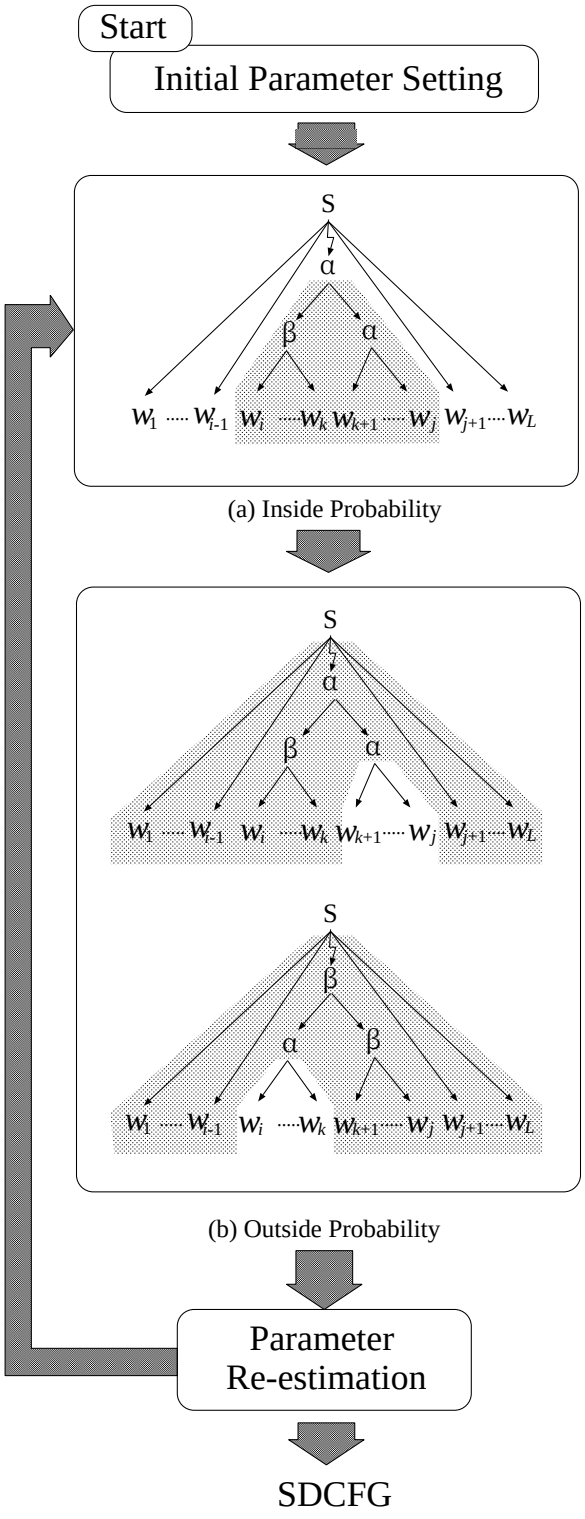


Figure 4.5: Estimation algorithm of SDCFG

SDCFG for Japanese

Japanese language has only right-headed dependency where a modifier precedes the head. In order to adjust the rewrite rules specific to the Japanese language, the following SDCFGs are proposed [22].

1. word-based SDCFG
   The rules of that two nonterminal symbols are derived a nonterminal symbol is modified as follows based on the dependency grammar:

$$\alpha \rightarrow \beta\alpha$$
$$\alpha \rightarrow w$$

   An amount of calculation is reduced from $O(N^3L^3)$ to $O(N^2L^3)$ by applying the rule of $\alpha \rightarrow \beta\alpha$. The SCFG using the dependency structure between words is call K-SCFG hereafter.

2. phrase-based SCFG
   In stead of words, phrases are derived from the nonterminal symbols and the regular grammar is applied for words in each phrase as follows:

$$\begin{aligned} \text{(inter-phrase)} \quad & \alpha \rightarrow \beta\gamma \\ \text{(intra-phrases)} \quad & \alpha \rightarrow w_c \\ & \alpha \rightarrow \beta w_f \end{aligned}$$

   where w_c is content words and w_f is function words. Given M phrases in a sentence, an amount of calculation can reduce $O(N^3L^3)$ to $O(N^3M^3)$. As Japanese phrases contain approximately two words on average, the training time is expected to be reduced to about $(\frac{1}{2})^3$.

3. phrase-based SDCFG
   The dependency rule $\alpha \rightarrow \beta\alpha$ is applied between phrases as follows:

$$\begin{aligned} \text{(inter-phrase)} \quad & \alpha \rightarrow \beta\alpha \\ \text{(intra-phrase)} \quad & \alpha \rightarrow w_c \\ & \alpha \rightarrow \beta w_f \end{aligned}$$

   The phrase based SDCFG (called PK-SCFG hereafter) can reduce an amount of calculation from $O(N^3L^3)$ to $O(N^2M^3)$. The computation amount is reduced by

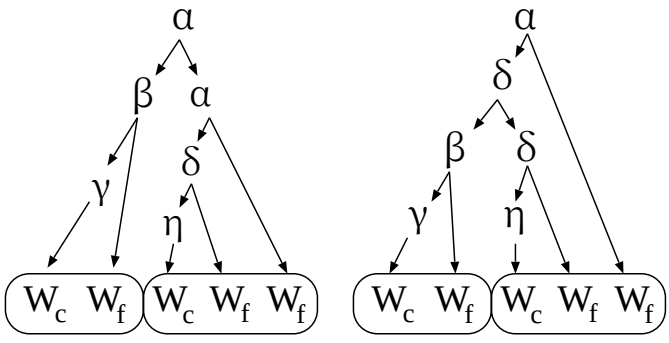
$\frac{1}{N}(\frac{1}{2})^3$. In the phrase-based SCFG (P-SCFG, PK-SCFG), phrase boundaries are automatically determined using POS information. In this method, a phrase consists of a content word and zero or more function words following it. Table 4.2 shows the specification of each model.

Table 4.2: Compared SCFGs

			word based		phrase based
Restricted grammar	not used	name	SCFG		P-SCFG
		rule type	$\alpha \to \beta\gamma$ $\alpha \to w$		$\alpha \to \beta\gamma$ $\alpha \to w_c$ $\alpha \to \beta w_f$
		amount of calculation	$O(N^3L^3)$		$O(N^3M^3 + N^2MI)$
	used	name	K-SCFG	K-SCFG2	PK-SCFG
		rule type	$\alpha \to \beta\alpha$ $\alpha \to w$	$\alpha \to \beta\alpha$ $\alpha \to w_c$ $\alpha \to \beta w_f$	$\alpha \to \beta\alpha$ $\alpha \to w_c$ $\alpha \to \beta w_f$
		amount of calculation	$O(N^2L^3)$		$O(N^2M^3 + N^2MI)$

N : number of nonterminals
L : number of words in a sentence
M : number of phrases in a sentence ($\approx \frac{1}{2}L$)
I : number of words in a phrase

In order to compare with PK-SCFG, K-SCFG2 consisting of $\alpha \to \beta\alpha$ and a regular grammar $\alpha \to \beta w$ is tested. The difference between PK-SCFG and K-SCFG2 is illustrated in Figure 4.6.



(a) Both 'K-SCFG2' and 'PK-SCFG'

(b) 'K-SCFG2' only

Figure 4.6: Examples of derivations of PK-SCFG and K-SCFG2

In this study, in order to estimate a dependency structure for speech summarization, a phrase-based SDCFG is used for the word concatenation score for Japanese.

Parameter Re-estimation in Phrase-based SDCFG

Parameters of a phrase-based SDCFG are estimated from a manual parsed corpus using the Inside-Outside algorithm. Since words in the corpus are tagged with POS, phrase boundaries are automatically detected based on the POS. Each phrase is made up of a content word followed by zero or more function words. In this study, content words include nouns, adjectives, verbs and adverbs, and the remaining words are included as function words. Suppose a sentence consists of M phrases,

$$S \rightarrow P_1 \dots P_m \dots P_M$$

P_m is defined as follows:

$$P_m = w_{mc} w_{mf,1} w_{mf,2} \dots w_{mf,Km}$$

Where,

M	: number of phrases in a sentence
w_{mc}	: content word of the m-the phrase
$w_{mf,i}$	: i-th function word on m-th phrase
K_m	: number of functions words in m-th phrase

Rewrite probabilities of $\alpha \rightarrow \beta\alpha$, $\alpha \rightarrow w_c$, $\alpha \rightarrow \beta w_f$ are denoted by $P(\alpha \rightarrow \beta\alpha)$, $P(\alpha \rightarrow w_c)$, $P(\alpha \rightarrow \beta w_f)$, respectively. The algorithm for estimating parameters of the phrase-based SDCFG is described below. Figure 4.7 indicates the estimation steps.

1. Initialization
   $P(\alpha \rightarrow \beta\alpha)$ is given a flat probability and $P(\alpha \rightarrow w_c)$, $P(\alpha \rightarrow \beta w_f)$ are given random values.

2. Calculation for intra-phrase forward probability
   The probability of deriving $w_{mc}\ w_{mf,1} \dots w_{mf,i}$ from α in the m-th phrase is calculated by the forward probability illustrated in Figure 4.7(a):

$$\begin{aligned} h(m,i,\alpha) &= P(\alpha \rightarrow w_{mc} w_{mf,1} \dots w_{mf,i}) \\ &= \begin{cases} P(\alpha \rightarrow w_{mc}) & if \quad i = 0 \\ \sum_{\beta} h(m, i-1, \beta) P(\alpha \rightarrow \beta w_{mf,i}) & if \quad i > 0 \end{cases} \end{aligned} \tag{4.12}$$

3. Calculation of the inter-phrase inside probability

   The inter-phrase inside probability illustrated in Figure 4.7(b) is calculated using the intra-phrase forward probability:

$$\begin{aligned} e(m,n|\alpha) &= P(\alpha \rightarrow P_m \ldots P_n) \\ &= \begin{cases} h(m, K_m, \alpha) & if \quad m = n \\ \sum_{l=m}^{n-1} \sum_{\beta} P(\alpha \rightarrow \beta\alpha) e(m,l|\beta) e(l+1,n|\alpha) & if \quad m < n \end{cases} \end{aligned} \tag{4.13}$$

4. Calculation of the inter-phrase outside probability

   The inter-phrase outside probability illustrated in Figure 4.7(c) is calculated using the inter-phrase inside probability:

$$\begin{aligned} f(m,n|\alpha) &= P(S \rightarrow P_1 \ldots P_{m-1} \alpha P_{n+1} \ldots P_M) \\ &= \sum_{l=1}^{m-1} \sum_{\beta} P(\alpha \rightarrow \beta\alpha) e(l, m-1|\beta) f(l,n|\alpha) \\ &\quad + \sum_{l=n+1}^{M} \sum_{\beta} P(\beta \rightarrow \alpha\beta) e(n+1,l|\beta) f(m,l|\beta) \end{aligned} \tag{4.14}$$

5. Calculation of the intra-phrase backward probability

   The intra-phrase backward probability illustrated in Figure 4.7(d) is calculated as follows using the inter-phrase outside probability:

$$\begin{aligned} r(m,i,\alpha) &= P(S \rightarrow P_1 \ldots P_{m-1} \alpha w_{mf,i+1} \ldots w_{mf,K_m} P_{m+1} \ldots P_M) \\ &= \begin{cases} f(m,m,\alpha) & if \quad i = K_m \\ \sum_{\beta} P(\beta \rightarrow \alpha w_{mf,i+1}) r(m, i+1, \beta) & if \quad i < K_m \end{cases} \end{aligned} \tag{4.15}$$

6. Estimation of parameters

   The parameters are re-estimated using the probabilities obtained by the steps 2 to 5.

$$\hat{a}(\alpha \rightarrow \beta\alpha) = \frac{\sum_{m=1}^{M-1} \sum_{n=m+1}^{M} \sum_{l=m}^{n-1} g(m,l,n;\alpha \rightarrow \beta\alpha)}{e(1,M|S)} \tag{4.16}$$

$$\hat{b}(\alpha \to w_c) = \frac{\sum_{m=1;w_{mc}=w_c}^{M} P(\alpha \to w_c) r(m,0,\alpha)}{e(1,M|S)} \tag{4.17}$$

$$\hat{c}(\alpha \to \beta w_f) = \frac{\sum_{m=1}^{M} \sum_{i=1;w_{mf,i}=w_f}^{K_m} h(m,i-1,\beta) P(\alpha \to \beta w_f) r(m,i,\alpha)}{e(1,M|S)} \tag{4.18}$$

where

$$g(m,l,n;\alpha \to \beta\alpha) = e(m,l|\beta) e(l+1,n|\alpha) P(\alpha \to \beta\alpha) f(m,n|\alpha) \tag{4.19}$$

7. The steps from 2 to 6 are iterated until the parameters are saturated.

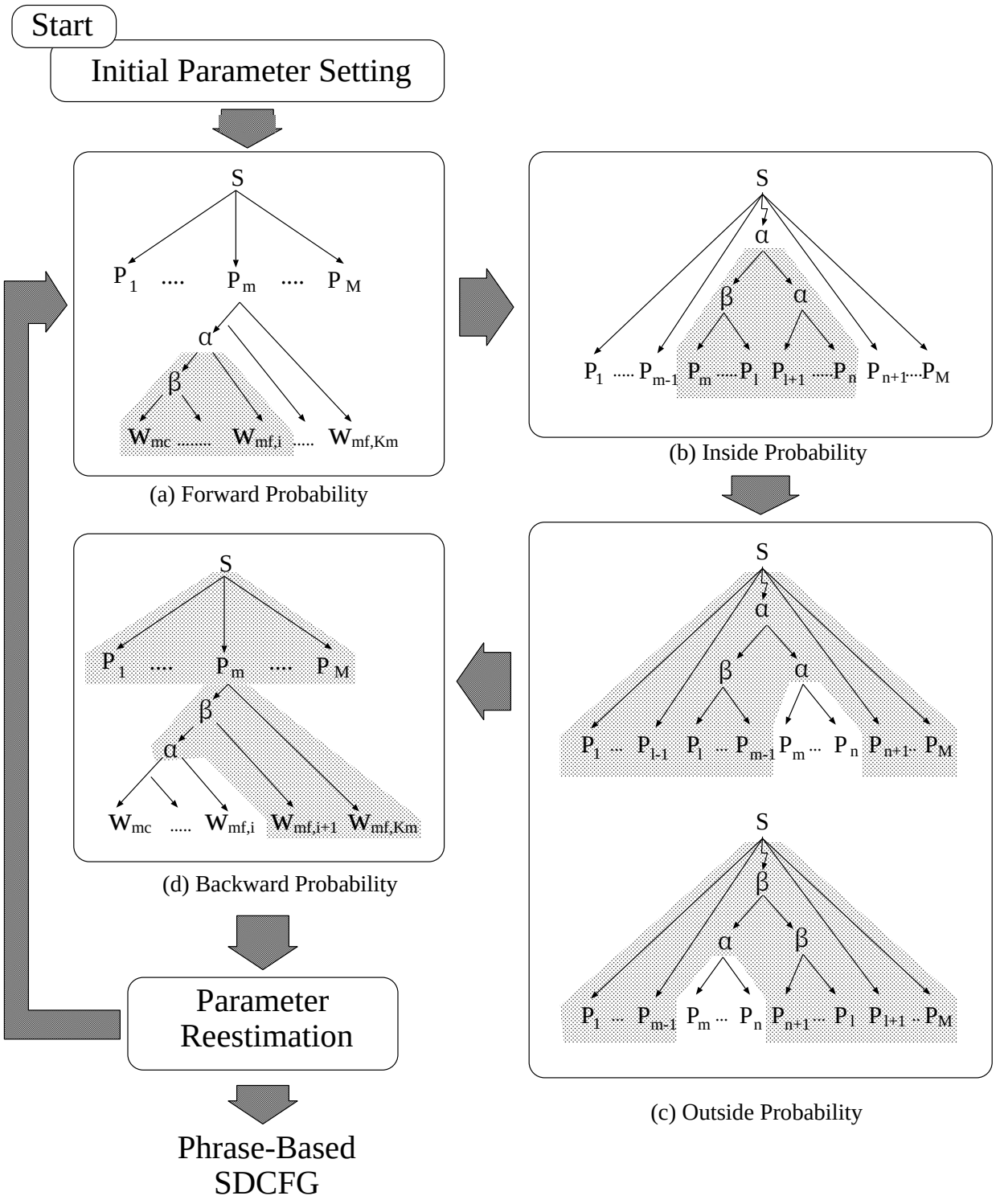


Figure 4.7: Estimation algorithm of phrase-based SDCFG

4.5 Application of SDCFG for a LVCSR system

SCFG is a very expressive language model because it allows the expression of not only local constraints like the N-gram model, but also the expression of global constraints over a whole sentence. However, in order to estimate parameters of a SCFG, use of the Inside-Outside algorithm is required, which needs a significant amount of computation, proportional to the cube of the number of the non-terminal symbols and the length of the input sequences. Because of the enormous calculation amount, SCFG has rarely been used for speech recognition.

In contrast with SCFG, the proposed SDCFG model can decrease the amount of computation of the Inside-Outside algorithm. Application of dependency grammar to SCFG results in computations proportional to the square of the number of the non-terminal symbols. In addition, application of the phrase based grammar results in an amount of computation proportional to 1/8 of the computation using the dependency grammar.

In Appendix A, the phrase-based SDCFG is tested for its performance in a speech recognition system in comparison with other types of SCFGs.

Chapter 5

Evaluation Method for Automatic Summarization

In our previous experiments the summarization results were evaluated according to the performance of extracting "important words" selected by human subjects from the manual transcriptions and maintaining the meaning of the original speech [21]. However, evaluation by humans for all automatic speech summarization results would be extremely cost inefficient. In this thesis, in order to evaluate the automatically summarized sentences, correctly transcribed speech is manually summarized by human subjects according to the target summarization ratio and used as correct targets. In consideration of subjective variations, the precision of extracted keywords and that of each word string with a certain length in the manual summarizations by human subjects were evaluated. In addition, the manual summarization results are merged into a word network and the most similar word string in the network is compared with an automatic summarization result based on a summarization accuracy.

5.1 Precision of Keywords and Word Strings

5.1.1 Precision of Keywords

The precision of extracted keywords corresponds to coverage of the core information. This measure is calculated as the mean of word significance values defined as percentages of subjects who have selected words as keywords. The precision of keywords R of the summarization $V = v_1, v_2, \ldots, v_M$ is given as follows.

$$R = \frac{\sum_{m=1}^{M} \frac{c(v_m)}{a}}{M} \quad , \tag{5.1}$$

a : number of subjects to make manual summarizations
M : total number of words in a summarized sentence
v_m : m-th word in a summarized sentence
$c(v_m)$: number of subjects extracting v_m

5.1.2 Precision of Word Strings

To evaluate linguistic correctness and maintenance of the original meanings of the utterance, the precision of each word string with a certain length in the automatically summarized sentences is defined as how many such word strings are included in at least one of the manual summarizations by human subjects.

The extraction ratio WS_D of each word strings consisting of D words in a summarized sentence $V = v_1, v_2, \ldots, v_M$ is given by

$$WS_D = \frac{\sum_{m=D}^{M} \delta(v_{m-D+1}, \ldots, v_{m-1}, v_m)}{M - D + 1} \quad , \tag{5.2}$$

where

$$\delta(u_D) = \begin{cases} 1 & if \quad u_D \in U_D \\ 0 & if \quad u_D \notin U_D \end{cases} \tag{5.3}$$

u_D : each word string consisting of D words
U_D : a set of word strings consisting of D words in all manual summarizations

Note that words occurring in different locations in an original sentence are considered to be different words even though they are the same words. When D is 1, WS_D indicates the precision of each word and when D is the length of a summarized sentence M, WS_D indicates the precision of the summarized sentence itself.

5.1.3 Test Evaluation Performance

5.1.3.1 Structure of the broadcast news transcription system

Acoustic models

The acoustic model in the sharable software repository for Japanese large vocabulary continuous speech recognition by IPA was used [35]. The feature vector extracted from speech consists of 12 MFCCs (mel-frequency cepstrum coefficients), the delta of their features, and the delta of normalized logarithmic power (derivatives). The total number of parameters in each vector is 25. MFCCs were normalized using the CMS (cepstral mean

subtraction) method. The phone models used were shared-state triphone HMMs designed using tree-based clustering. The total number of states was 1012, and the number of Gaussian mixture components per state was 8. They were trained using speech by 100 speakers reading newspaper.

Language models

Broadcast-news manuscripts recorded from August 1992 to May 1996, comprising of approximately 500k sentences consisting of 22M words, were used for constructing language models. The vocabulary size was 20k words.

Decoder

We used a word-graph-based 2-pass decoder for transcription. In the first pass, a frame-synchronous beam search was performed using the above-mentioned HMMs and a bigram language model. A word graph was generated as a result of the first pass. In the second pass, the word graph was rescored using a trigram language model to derive the final transcription, which was then used for summarization.

5.1.3.2 Evaluation Experiments

Evaluation data

Japanese broadcast news speech on TV in 1996 was used as a test set to evaluate our proposed method. The set consisted of 419 utterances by a female anchor speaker, and was manually segmented into sentences. The out-of-vocabulary (OOV) rate for the 20k word vocabulary is 2.5% and the perplexity for the test set was 54.5. Fifty utterances with word recognition accuracy above 90%, which was the average rate over the fifty utterances, were selected and used for the evaluation. The summarization ratio, the ratio of the number of characters in the summarized sentences to that in the original sentences, was set to 20, 40, 60, 70 and 80%.

Language models for summarized sentences

A trigram language model for summarization was built using text from Mainichi newspaper published from 1996 to 1998, comprising of approximately 5.1M sentences consisting of 87M words. We did this because we considered newspaper text to be generally more compact and simpler than broadcast news text, and therefore more appropriate for building language models for summarization than broadcast news text. In our previous experiments the automatically summarized sentences using word trigrams constructed from newspaper text were much better than those constructed from broadcast-news manuscripts.

Evaluation results

Automatically summarized utterances without the confidence measure incorporated (REC), that with the confidence measure (CM), and the summarized transcriptions by humans (TRS) were evaluated. To set a goal for the automatic summarized sentences, each manual summarization by 25 human subjects (SUB) were evaluated based on the manual summarizations by 24 human subjects except for the evaluated summarization itself. To insure that our method was sound, we considered randomly generated summarizations according to the summarization ratio (RDM) to compare the precisions with those achieved by our proposed methods.

The results evaluated by precision of keywords is shown in Table 5.1. The better results of CM than that of REC shows the precision of keywords of the summarized speech improved significantly by using the confidence score. The difference between TRS and CM is that some important words in TRS were not included in CM.

Table 5.1: Evaluation results by precision of keywords

Target ratio	20%	40%	60%	70%	80%
RDM	0.17	0.35	0.54	0.66	0.75
REC	0.20	0.38	0.58	0.69	0.77
CM	0.27	0.41	0.57	0.68	0.75
TRS	0.31	0.43	0.60	0.71	0.78
SUB	0.45	0.56	0.68	0.75	0.80

The evaluation results by precision of word strings for the condition of 70% summarization ratio is shown in Fig.5.1. The concatenation of words is more constrained when word strings were evaluated on longer word strings. Therefore, the word string precisions of all types of summarizations decrease gradually with the length of word strings. The automatic summarization of TRS, REC and CM can maintain the correct word concatenation more often than RDM. The rapid decrease of RDM precision indicates the word concatenations of RDM are grammatically and semantically incorrect. However the results of the automatic summarizations cannot reach the performance level of the goal of SUB. The same types of results were also shown by all summarizations using the various summarization ratios.

In order to evaluate the efficiency of the confidence score for the recognition results with lower accuracy, 10 utterances with relatively lower accuracy in the test set (called "difficult test set") were separately evaluated. The word string precisions in the condition of 20% summarization ratio is shown in Fig.5.2. The left figure shows all the test set and

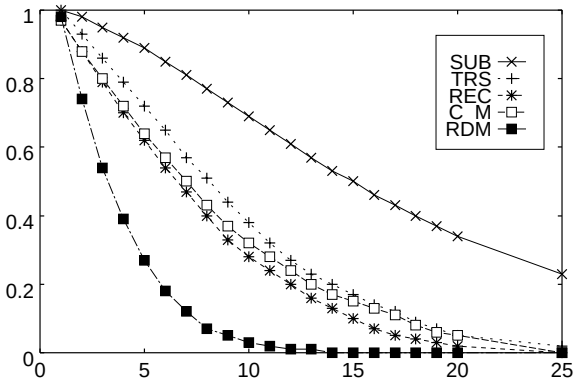


Figure 5.1: Precision of word strings vs. length of a word string at 70% summarization ratio.

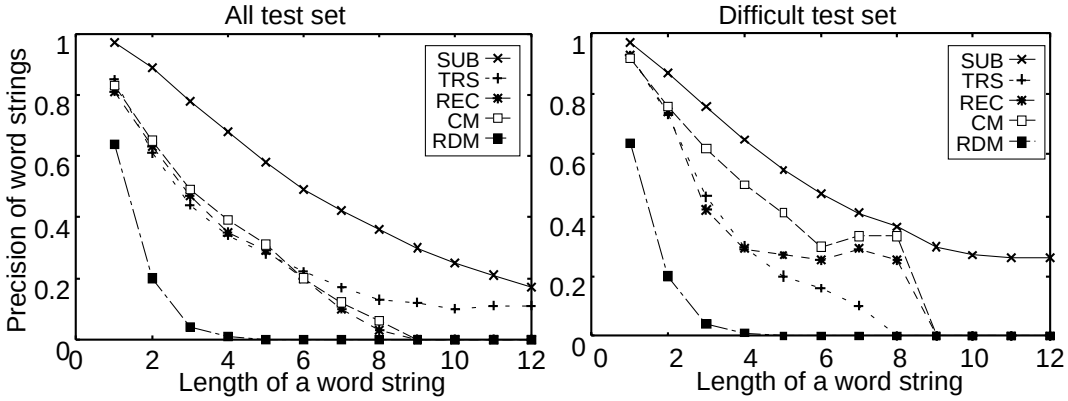


Figure 5.2: Precision of word strings vs. length of a word string at 20% summarization ratio.

the right one shows difficult test set results. The results show that the confidence measure can improve the automatic summarization especially when the speech recognition rate is relatively low.

Figure 5.3 shows the relationships between the precision of a 3-word string and the precision of keywords. TRS and REC indicates summarization results of manual transcribed speech and automatically transcribed speech, respectively. MAI and NHK show summarization results generated using language models using corpora of Mainichi newspaper text and manuscripts for broadcasting news of NHK, respectively. The linguistic score calculated using the newspaper text can generate much better summarization results.

In order to evaluate the relation between precision of word strings and subjective evalu-

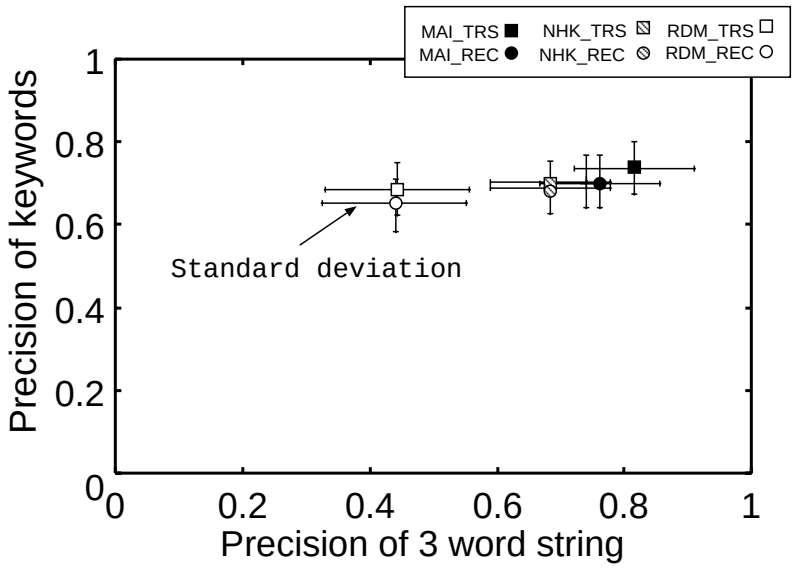


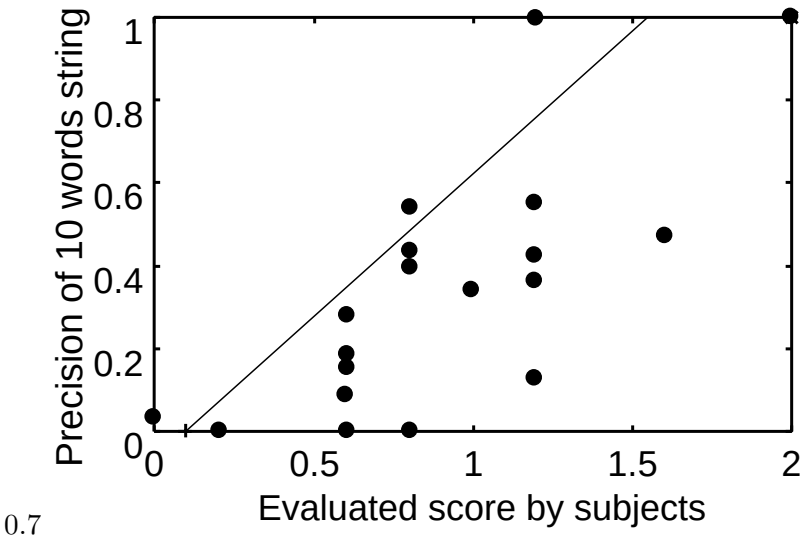
Figure 5.3: Precision of 3 word string vs. precision of keywords

ation for the automatic summarization results, 10 human subjects categorize the automatic speech summarization into 3 groups, that of same, inclusive and different meanings. Each automatic summarization result is given a score as follows:

Group	Score
SAME	2
INCLUSIVE	1
DIFFERENT	0

The score averaged among subjects is defined as the subjective evaluation score. Figure 5.4 shows the correlation of the precision of 10 word string with subjective evaluation scores by humans. The correlation coefficient is 0.68, and thus higher word string precision indicates higher availability to maintain original meanings.

However, the quality of an automatic summarization result as a "sentence" cannot be directly evaluated using this precision of word strings. A measure to evaluate how much automatic summarized sentences maintain original meanings is needed.



0.7

Figure 5.4: Correlation between precision word strings and subjective evaluation score.

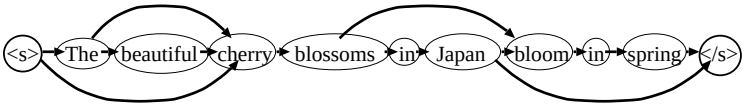
5.2 Summarization Accuracy Based on a Word Network of Manual Summarization

In order to automatically evaluate summarized sentences in terms of maintenance of original meanings, if must be compared to a correct answer for a summarized sentence. Since correct answers generated by human subjects vary subject to subject, the manual summarization results are merged into a word network, which approximately expresses all possible correct summarizations including subjective variations. A word string extracted from the word network that is the most similar to the automatic summarization result is considered as a correct target for the automatic summarization. The accuracy, comparing the summarized sentence with the target word string, is used as a measure of linguistic correctness and maintenance of original meaning of the utterance. A "summarization accuracy" given by eq.(5.4) is calculated using the word network [36].

$$Accuracy = \frac{Len - Sub - Ins - Del}{Len} \times 100 \quad [\%] \tag{5.4}$$

where,

Sub : number of substitution errors
Ins : number of insertion errors
Del : number of deletion errors
Len : number of words in the most similar word string in the network



Automatic summarization	<s> Chill (Substition) _____ (Deletion) bloom in spring </s>
The most similar word string to the automatic summarization in the net work	<s> Cherry blossoms bloom in spring </s>
Word Accuracy	5-(1+0+1)/5*100 = 60%

Figure 5.5: An example of a word network and calculation of the summarization accuracy.

Figure 5.5 shows an example of a summarization accuracy calculation using a word network. In this example, “cherry ” is misrecognized as “chill” by a recognition system and extracted into a summarized sentence.

Chapter 6

Evaluation Experiments

6.1 Evaluation Experiment for Japanese Broadcast News Speech

6.1.1 Evaluation data

Japanese news speech broadcast on TV in 1996 was used as a test set to evaluate our proposed method. The set consisted of 419 utterances by a female anchor speaker, and was manually segmented into sentences. The out-of-vocabulary (OOV) rate for the 20k word vocabulary was 2.5% and the perplexity for the test set was 54.5. Fifty utterances with word recognition accuracy above 90%, which was the average rate over the 50 utterances, were selected and used for the evaluation. The summarization ratio, the ratio of the number of characters in the summarized sentences to that in the original sentences, was set to 40, 60, 70 and 80%.

In addition, 5 news articles, consisting of approximately 5 sentences each, were summarized using the summarization technique for multiple utterances at 30% summarization ratio.

6.1.2 Structure of Broadcast News Transcription System

Acoustic Models

The feature vector extracted from speech consists of 16 cepstral coefficients, normalized logarithmic power, and their delta features (derivatives). The total number of parameters in each vector was 34. Cepstral coefficients were normalized using the CMS (cepstral mean subtraction) method. The acoustic models used were shared-state triphone HMMs designed using tree-based clustering. The total number of states was 2,106, and the number of Gaussian mixture components per state was 4. They were trained using phonetically-balanced sentences and dialogues read by 53 speakers (approximately 20 hours in total),

sentences which were completely different from that of the broadcast news task. All of the speakers were male, and so the HMMs were gender-dependent models. The total number of training utterances was 13,270 and the total length of the training data was approximately 20 hours.

Language Models

Broadcast-news manuscripts recorded from August 1992 to May 1996, comprising of approximately 500k sentences consisting of 22M words, were used for constructing language models. The vocabulary size was 20k words. To calculate word n-gram language models, we segmented the broadcast-news manuscripts into words by using a morphological analyzer since Japanese sentences are written without spaces between words.

Decoder

We used a word-graph-based 2-pass decoder for transcription. In the first pass, a frame-synchronous beam search was performed using the above-mentioned HMMs and a bigram language model. A word graph was generated as a result of the first pass. In the second pass, the word graph was rescored using a trigram language model.

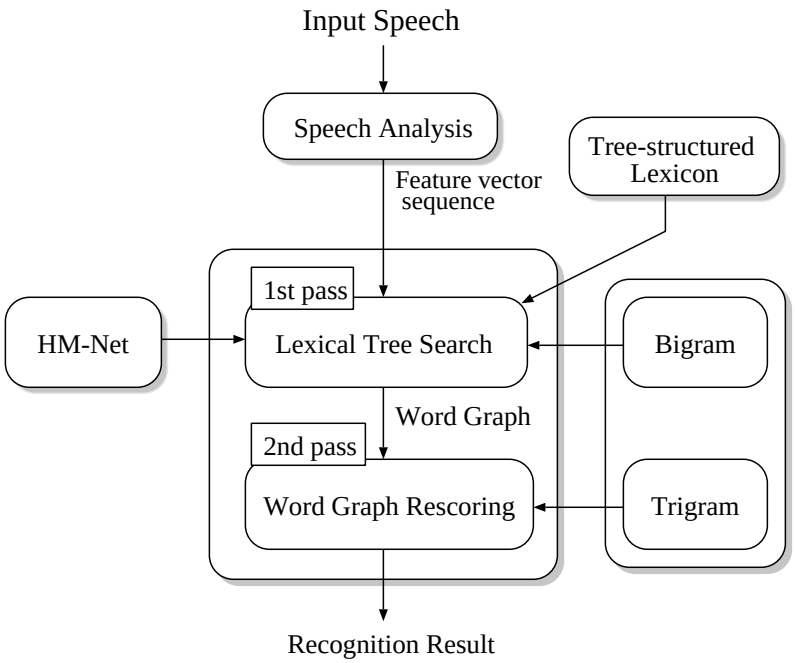


Figure 6.1: Large Vocabulary Continuous Speech Recognition System.

6.1.3 Training data for summarization models

Word significance model

The same broadcast-news manuscripts used for building a language model for the speech recognition system was used for calculating the word significance measure for summarization.

Language model

A trigram language model for summarization was built using text from Mainichi newspaper published from 1996 to 1998, comprising of 5.1M sentences with 87M words. We consider newspaper text to be generally more compact and simpler than broadcast news text and therefore more appropriate for building language models for summarization. Our previous experiments confirmed that the automatically summarized sentences using word trigram based on newspaper text were much better than those by broadcast-news manuscripts [21].

SDCFG

SDCFG for word concatenation score was built using text from the manually parsed corpus of Mainichi newspaper published from 1996 to 1998, comprising of approximately 4M sentences with 68M words. The number of non-terminal symbols was 100.

6.1.4 Evaluation results

Table 6.1: Summarization results for manual and automatic transcription.

書き起こし	ジュネーブで開かれている地球温暖化対策の国際会議で日本政府は西暦二千年以降先進各国がＧＤＰ 国内総生産に応じた二酸化炭素の排出削減に努めるという新たな国際目標を日本としては今回初めて提案することを決めました
要約率 80%	ジュネーブで開かれている地球温暖化対策の国際会議で日本政府は＿＿＿＿＿＿＿＿先進各国が＿＿＿＿ 国内総生産に＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿＿＿＿目標を日本としては今回初めて提案することを決めました
要約率 70%	ジュネーブで開かれている地球温暖化対策の国際会議で日本政府は＿＿＿＿＿＿＿先進各国が＿＿＿＿ ＿＿＿＿＿＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿＿＿＿目標を日本として＿＿＿＿＿＿＿提案することを決めました
要約率 60%	ジュネーブで開かれている地球温暖化対策の国際会議で日本政府は＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿ ＿＿＿＿＿＿＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿＿＿＿目標を日本として＿＿＿＿＿＿＿提案することを決めました
要約率 40%	＿＿＿＿＿＿＿＿＿＿＿＿地球温暖化対策＿＿＿＿会議で日本政府は＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿ ＿＿＿＿＿＿＿＿＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿＿＿＿目標を＿＿＿＿＿＿＿＿＿＿＿＿提案することを決め＿＿＿
要約率 20%	＿＿ ＿＿＿＿＿＿＿＿＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿＿＿＿目標を＿＿＿＿＿＿＿＿＿＿＿＿提案すること＿＿＿＿＿＿
音声認識結果	＜年＞で開かれている＜月いう＞温暖化対策の国際会議で日本政府は西暦二千年以降先進各国がＧＤＰ 国内総生産に応じた二酸化炭素の排出削減に努めるという新たな国際目標を日本としては今回初めて提案することを決めました
要約率 80%	＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿温暖化対策の国際会議で日本政府は＿＿＿＿＿＿＿＿先進各国が＿＿＿＿ ＿＿＿＿＿＿＿＿＿＿＿二酸化炭素の排出削減に努めるという新たな国際目標を日本としては今回初めて提案することを決めました
要約率 70%	＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿温暖化対策の国際会議で日本政府は＿＿＿＿＿＿＿＿先進各国が＿＿＿＿ ＿＿＿＿＿＿＿＿＿＿＿二酸化炭素の排出削減に努めるという＿＿＿＿＿＿＿＿日本としては今回初めて提案することを決めました
要約率 60%	＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿温暖化対策の国際会議で日本政府は＿＿＿＿＿＿＿＿先進各国が＿＿＿＿ ＿＿＿＿＿＿＿＿＿＿＿二酸化炭素の排出削減に努めるという＿＿＿＿＿＿＿目標＿日本として＿＿＿＿＿＿＿提案することを決めました
要約率 40%	＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿温暖化対策の国際会議で日本＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿ ＿＿＿＿＿＿＿＿＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿＿＿＿目標を＿＿＿＿＿＿＿＿＿＿＿＿提案することを決めました
要約率 20%	＿＿ ＿＿＿＿＿＿＿＿＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿＿＿＿目標＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

＿は削除された領域，＜＞は認識誤りを表す

Evaluation results using a word network

Tables 6.2 and 6.3 respectively show the types of summarization of manual transcription (TRS) and automatic transcription (REC) investigated in this paper. In these tables the symbols of I, L, C and T indicate the utilization of word significance score, linguistic score, confidence score and word concatenation score for summarization respectively.

In the summarization of REC, conditions with ($I\_L\_C\_T$) and without (I_L_T) the word confidence score were compared. Conditions with (I_L_T, $I\_L\_C\_T$) and without (I_L,I_L_C) the word concatenation score were compared in summarization for both TRS and REC.

To set an upper limit for automatic summarization, manual summarization by human subjects from manual transcription (SUB_TRS) was performed. The results were evaluated using all other manual summarization results as a measure of correct summarization. In addition, as the upper bound of automatic speech summarization for transcription including speech recognition errors, manual summarization of automatically transcribed utterances at 70% summarization ratio was also evaluated (SUB_REC). To insure that our method was sound, we made randomly generated summarization sentences (RDM) according to the summarization ratio and compared them with those obtained by our proposed methods.

Table 6.2: Summarization types of manual transcription.

Target	Manual Transcription(TRS)			
Symbol	RDM	I_L	I_L_T	SUB_TRS
Manual summarization				○
Significance score (I)		○	○	
Linguistic score (L)		○	○	
Word concatenation score (T)			○	
Random word selection	○			

Table 6.3: Summarization types of automatic transcription.

Target	Automatic Transcription(REC)				
Symbol	RDM	I_L_C	$I\_L\_C\_T$	I_L_T	SUB_REC
Manual summarization					○
Significance score (I)		○	○	○	
Linguistic score (L)		○	○	○	
Confidence score (C)		○	○		
Word concatenation score (T)			○	○	
Random word selection	○				

Table 6.4: Number of word errors and summarized sentences including word errors.

	RDM	I_L_C
recognition result	69 word errors (50 utterances)	
80%	36 (17)	12 (8)
70%	31 (16)	5 (5)
60%	25 (15)	3 (3)
40%	18 (13)	2 (2)
20%	8 (7)	3 (3)

the number in () represents the number of sentences including word errors

Table 6.5: An example of evaluation results based on a manual summarization word network.

音声認識結果	＜年＞で開かれている＜月いう＞温暖化対策の国際会議で日本政府は西暦二千年以降先進各国がＧＤＰ 国内総生産に応じた二酸化炭素の排出削減に努めるという新たな国際目標を日本としては今回初めて提案することを決めました
要約率 80%	地球 温暖 化 対策 の 国際 会議 で 日本 政府 は ＧＤＰ に 応じた 二酸化炭素 の 排出 削減 に 努める と いう 新たな 国際 目標 を 日本 と して は 今回 初めて 提案 する こと を 決め ました
	DEL 温暖 化 対策 の 国際 会議 で 日本 政府 は <先進> <各国> <が> 二酸化炭素 の 排出 削減 に 努める と いう 新たな 国際 目標 を 日本 と して は 今回 初めて 提案 する こと を 決め ました
要約率 70%	温暖 化 対策 の 国際 会議 で 日本 政府 は ＧＤＰ に 応じた 二酸化炭素 の 排出 削減 に 努める と いう 目標 を 日本 と して は 今回 初めて 提案 する こと を 決め ました
	温暖 化 対策 の 国際 会議 で 日本 政府 は <先進> <各国> <が> 二酸化炭素 の 排出 削減 に 努める と いう DEL DEL 日本 と して は 今回 初めて 提案 する こと を 決め まし た
要約率 60%	温暖 化 対策 の 国際 会議 で 日本 政府 は 先進 各国 が二酸化炭素 の 排出 削減 に 努める と いう 目標 INS INS INS 提案 する こと を 決め ました
	温暖 化 対策 の 国際 会議 で 日本 政府 は 先進 各国 が二酸化炭素 の 排出 削減 に 努める と いう 目標 日本 と して 提案 する こと を 決め ました
要約率 40%	温暖 化 対策 の 国際 会議 で 日本 政府 二酸化炭素 の 排出 削減 INS を 提案 する こと を 決め ました
	温暖 化 対策 の 国際 会議 で 日本 DEL 二酸化炭素 の 排出 削減 目標 を 提案 する こと を 決め ました
要約率 20%	二酸化炭素 の 排出 削減 目標 提案
	二酸化炭素 の 排出 削減 目標 <DEL>

上段：正解要約文，下段：自動要約文，<>は置換，INS は挿入，DEL は脱落を表す

Summarization of each utterance

Figures 6.3 through 6.6 show summarization accuracy of both manual transcription (TRS) and automatic transcription (REC) at 40%, 60%, 70% and 80% summarization ratios. These results show that our proposed automatic speech summarization technique is significantly more effective than RDM. The method using the word concatenation score (I_L_T, $I\_L\_C\_T$) can reduce meaning alteration compared with the method without using the word concatenation score (I_L, I_L_C). The better result using the word concatenation score ($I\_L\_C\_T$) compared with that without using the word concatenation score (I_L_T) shows that the summarization accuracy is significantly improved by the confidence score.

The performance of automatic summarization of automatic transcription (REC) is comparable with that of manual transcription (TRS) under all conditions of summarization ratio. Although automatic summarization cannot achieve the performance of the manual summarization of manual transcription (SUB_TRS), it can achieve performance comparable to the manual summarization of the recognition result (SUB_REC).

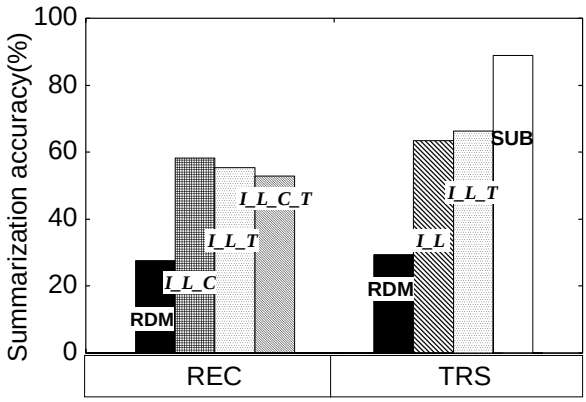


Figure 6.2: Each utterance summarization result at 20% summarization ratio. REC: summarization of recognition results, TRS: summarization of manual transcription, RDM: random word selection, C: confidence score, I: significance score, L: linguistic score, I_C,L_C,I_L: combination of 2 scores, I_L_C: combination of all scores, SUB: subjective summarization.

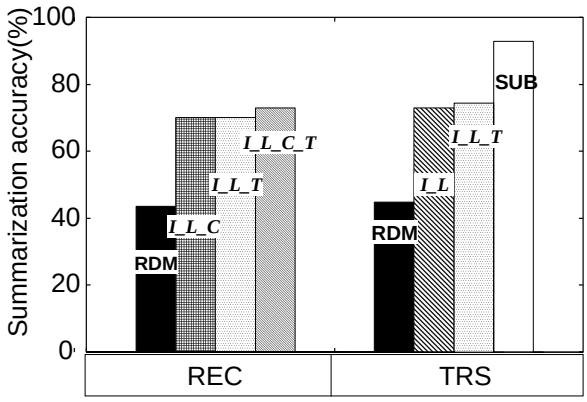


Figure 6.3: Each utterance summarization result at 40% summarization ratio.

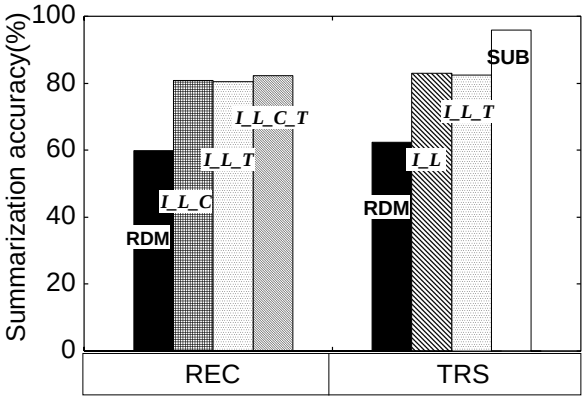


Figure 6.4: Each utterance summarization result at 60% summarization ratio.

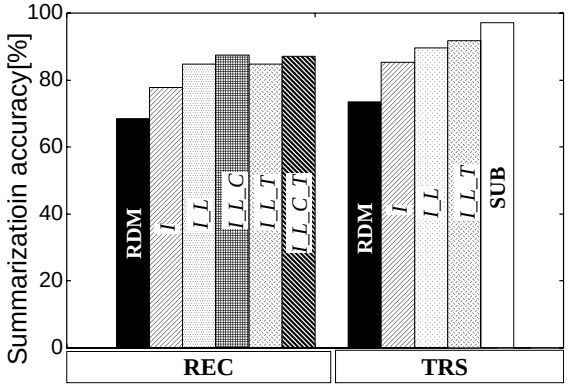


Figure 6.5: Each utterance summarization result at 70% summarization ratio.

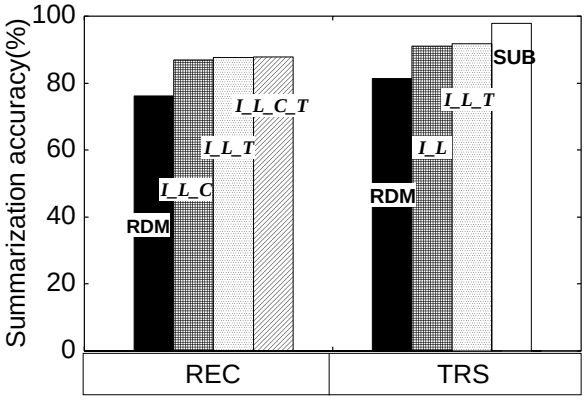


Figure 6.6: Each utterance summarization result at 80% summarization ratio.

Summarization of multiple utterances

Figure 6.7 shows the summarization accuracy of summarizing articles having multiple sentences at 30% summarization ratio. These results show that our proposed automatic speech summarization technique is effective for the summarization of multiple utterances.

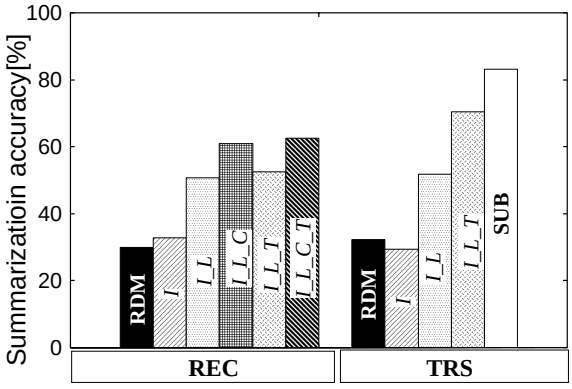


Figure 6.7: Article summarization results at 30% summarization ratio. REC: summarization of recognition results, TRS: summarization of manual transcription, RDM: random word selection, C: confidence score, I: significance score, L: linguistic score, I_C,L_C,I_L: combination of 2 scores, I_L_C: combination of all scores, SUB: subjective summarization.

6.2 Evaluation Experiment for English Broadcast News Speech

6.2.1 Evaluation data

English TV broadcast news utterances (CNN news) recorded in 1996 given by NIST as a test set of Topic Detection and Tracking (TDT) were tagged by Brilltagger [37] and used to evaluate our proposed method. Five news articles consisting of 25 utterances in average were transcribed by the JANUS [38] speech recognition system. The multiple utterance summarization was performed for each of the five news articles at 40% and 70% summarization ratio. 50 utterances arbitrarily chosen from the five news articles were used for the sentence by sentence summarization with the summarization ratios of 40% and 70%. Mean word recognition accuracies of the utterances used for the multiple utterance summarization and those for sentence by sentence summarization were 78.4% and 81.4%, respectively. In order to build word networks of manual summarization results, 17 native English speakers generated manual summarization by removing or extracting words.

6.2.2 Structure of Broadcast News Transcription System

English broadcast news speech was transcribed by the JRTk (Janus Speech Recognition Toolkit) [38] with the following conditions.

Feature extraction

Sounds were digitized with 16kHz sampling and 16bit quantization. Feature vectors had 13 elements consisting of MFCC. Vocal Tract Length Normalization (VTLN) and cluster-based cepstral mean normalization were used to compensate for speaker and channel. Linear Discriminant Analysis (LDA) was applied to reduce feature dimensions in each segment consisting of 7 frames to 42.

Acoustic model

A pentphone model with 6000 distributions sharing 2000 codebooks were used. There were about 105k Gaussians in the system. The training data was comprised of 66 hours of Broadcast News(BN).

Language model

Bigram and trigram were built using BN corpus. Its vocabulary size was 40k.

Decoder

A word-graph-based 3-pass decoder which was composed with JRTk was used for transcription. In the first pass, frame-synchronous beam search was performed using a tree-based lexicon, the above-mentioned HMMs and a bigram model to generate a word graph. In the second pass, frame-synchronous beam search was performed again using a flat lexicon hypothesized in the word graph by the first pass and a trigram model. In the third pass, the word graph was minimized and rescored using the trigram language model.

6.2.3 Training data for summarization models

A word significance model, a trigram language model and SDCFG were constructed using roughly 35M words (10681 sentences) of the Wall Street Journal corpus and the Brown corpus in Penn Treebank[39].

6.2.4 Evaluation results

Manual transcription (TRS) and automatic transcription (REC) were both summarized. Table 6.6 shows an example of evaluation results based on a manual summarization word network. Figure 6.8 shows summarization accuracies of utterance summarization at 40% and 70% summarization ratio and Fig. 6.9 shows those of summarizing articles having multiple sentences at 40% and 70% summarization ratio. In these figures, I, L, C and T

Table 6.6: *An example of evaluation results based on a manual summarization word network. upper: a set of words extracted from the correct summarization network which is the most similar to automatic summarization, lower: automatic summarization of recognition result.*

Recognition result	VICE PRESIDENT AL GORE SAYS THE GOVERNMENT HAS A PLAN TO AVOID THE INEVITABLE PROSPECT OF INCREASED AIRPLANE CRASHES AND FATALITY IS
70% summarization	VICE PRESIDENT AL GORE SAYS THE GOVERNMENT HAS A PLAN TO AVOID THE INCREASED AIRPLANE CRASHES
	VICE PRESIDENT AL GORE SAYS THE GOVERNMENT HAS A PLAN TO AVOID <DEL> INCREASED AIRPLANE CRASHES
40% summarization	<INS> THE GOVERNMENT HAS A PLAN TO AVOID THE INCREASED AIRPLANE CRASHES
	GORE THE GOVERNMENT HAS A PLAN TO AVOID THE INCREASED AIRPLANE CRASHES

_: recognition error, <>: substitution, <INS>: insertion, <DEL>: deletion

indicate that the word significance score, the linguistic score, the confidence score and the word concatenation score are used, respectively.

In the summarization of REC, conditions with and without the word confidence score, ($I\_L\_C\_T$) and (I_L_T), were compared. In summarization for both TRS and REC, conditions with and without the word concatenation score, (I_L_T, $I\_L\_C\_T$) and (I_L,I_L_C), were compared.

The averaged summarization accuracies of each manual summarizations (SUB) was considered to be the upper limit of the automatic summarization accuracy. To ensure that our method is sound, we made randomly generated summarized sentences (RDM) according to the summarization ratio and compared them with those obtained by our proposed methods.

These results show that our proposed automatic speech summarization technique is significantly more effective than RDM. Using the word concatenation score (I_L_T, $I\_L\_C\_T$) also reduce the meaning alteration compared to not using it (I_L, I_L_C). The result obtained when using the word confidence score ($I\_L\_C\_T$) compared with those not using it

Table 6.7: Number of word errors and summarized sentences including word errors.

	each utterance summarization		multiple utterance summarization	
recognition result	180 word errors (45)		326 word errors (94)	
summarization result	40%	70%	40%	70%
I	42 (27)	111 (40)	99 (56)	199 (71)
I_L	44 (28)	87 (37)	86 (53)	166 (69)
I_L_C	23 (15)	49 (22)	34 (28)	82 (47)
I_L_T	46 (27)	84 (37)	90 (56)	173 (69)
$I\_L\_C\_T$	22 (13)	51 (24)	25 (17)	80 (47)
RDM	82 (30)	87 (21)	89 (45)	169 (65)

():number of sentences including recognition errors

(I_L_T) shows that the summarization accuracy is improved by the confidence score.

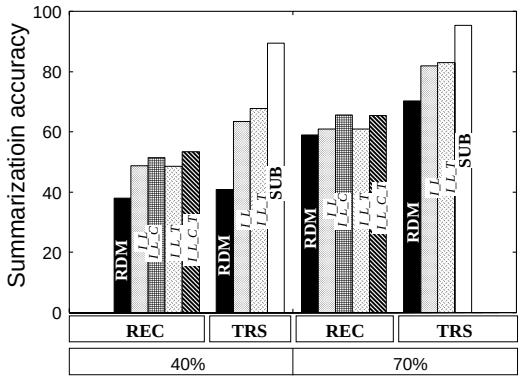


Figure 6.8: *Each utterance summarizations at 40% and 70% summarization ratio. REC: summarization of recognition results, TRS: summarization of manual transcription, RDM: random word selection, C: confidence score, I: significance score, L: linguistic score, I_C,L_C,I_L: combination of 2 scores, I_L_C: combination of all scores, SUB: subjective summarization.*

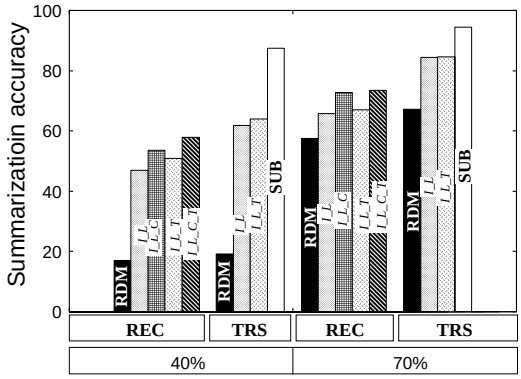


Figure 6.9: *Article summarizations at 30% and 70%summarization ratio. C: confidence score, I: significance score, L: linguistic score, I_C,L_C,I_L: combination of 2 scores, I_L_C: combination of all scores.*

6.2.5 Conclusions

Each utterance and a whole news consisting of multiple utterances of English broadcast news speech were summarized by our automatic speech summarization method based on the follwing scores: word significance score, linguistic likelihood, word confidence measure and word concatenation probability. Experimental results show that our proposed

method can effectively extract relatively important information and remove redundant and irrelevant information from Englsih news speech as well as Japanese one.

In contrast with the confidence score which has been incorporated into the summarization score to exclude word errors by a recognizer, the liguistic score is effective to reduce out of context words extraction both from recognition errors and human disfluencies. In summarizing Japanese news speech, the confidence measure could improve the summarizing performance by excluding incontext word errors. In English, the confidence measure can not only exclude word errors but also help extracting clearly pronouced important words. This results in the use of the confidence measure causing a lager increace in the summarization accuracy for English rather than Japanese.

6.3 Evaluation Experiment for Lecture Speech

Lecture speech spoken for roughly 12 minutes in an academic conference was recognized and summarized automatically. In order to apply the models for summarization which have been constructed using written text, an ad-hoc filter was used to convert the style of spontaneous speech to the style of written text.

6.3.1 Structure of Lecture speech Transcription System

Acoustic Models

The feature vector extracted from speech consisted of 16 cepstral coefficients, normalized logarithmic power, and their delta features(derivatives). The total number of parameters in each vector was 34. Cepstral coefficients were normalized using the CMS (cepstral mean subtraction) method. The acoustic models used were shared-state triphone HMMs designed using tree-based clustering. The total number of states was 2,106, and the number of Gaussian mixture components per state was 4. They were trained using phonetically-balanced sentences and dialogues read by 53 speakers (approximately 20 hours in total). The sentences used were completely different from the broadcast news task. All of the speakers were male, and so the HMMs were gender-dependent models. The total number of training utterances was 13,270 and the total length of the training data was approximately 59 hours.

6.3.2 Training data for summarization models

Word significance model

The word significance model was constructed using the manual transcription of lecture speech consisting of 1.5M words, 60 proceedings other than the test set, broadcast-news

manuscripts recorded from August 1992 to May 1996, comprised of approximately 500k sentences consisting of 22M words and Mainichi newspaper published from 1996 to 1998, comprised of 5.1M sentences with 87M words.

Summarization linguistic model

A trigram for summarization was constructed using the manual transcription of lecture speech consisting of 1.5M words converted the style of spontaneous speech to the style of written text by an ad-hoc filter. 60 proceedings other than the test set was also used for the construction of the trigram.

SDCFG

SDCFG for a word concatenation score was built using text from a manually parsed corpus of Mainichi newspaper published from 1996 to 1998, comprised of approximately 4M sentences with 68M words. The number of non-terminal symbols was 100.

6.3.3 Evaluation data

Manual transcription (TRS) and automatic transcription (REC) were both automatically summarized. In order to build a word network of manual summarization results, 9 human subjects generated manual summarization by removing or extracting words. The averaged summarization accuracies of each manual summarizations (SUB) was considered to be the upper limit of the automatic summarization accuracy. To ensure that our method was sound, we made randomly generated summarized sentences (RDM) according to the summarization ratio and compared them with those obtained by our proposed methods.

6.3.4 Evaluation results

Table 6.8 shows some of the automatic speech summarization results. Figures 6.10 and 6.11 show summarization accuracy of both manual transcription (TRS) and automatic transcription (REC) at 50%, 80% summarization ratios. In these figures, I, L, C and T indicate that the word significance score, the linguistic score, the confidence score and the word concatenation score are used, respectively.

These results show that our proposed automatic speech summarization technique is significantly more effective than RDM. The word linguistic score L is more efficient than the significance score I, and the combination of these two scores I_L can further increase the summarization accuracy at 80% summarization ratio. When L is worse than I, the combination between I and L results in decreasing word accuracy. The confidence score C can slightly increase the performance only at 50% summarization ratio. The word

concatenation score T cannot contribute significantly to the performance of automatic summarization. The word concatenation score T based on the newspaper text cannot represent the feature of lecture speech and thus the contribution by T is slight.

The summarization accuracy of manual transcription (TRS) is significantly higher than that of automatic recognition result (REC) at 80% summarization ratio. The manual summarization of the automatic recognition result (REC) by one human subject at 50% results in the same performance as the automatic summarization. Even if humans manually summarize the transcribed speech, the accurate summarization sentences cannot be generated because of the recognition errors. Therefore, the higher speech recognition accuracy is particularly crucial for the speech summarization.

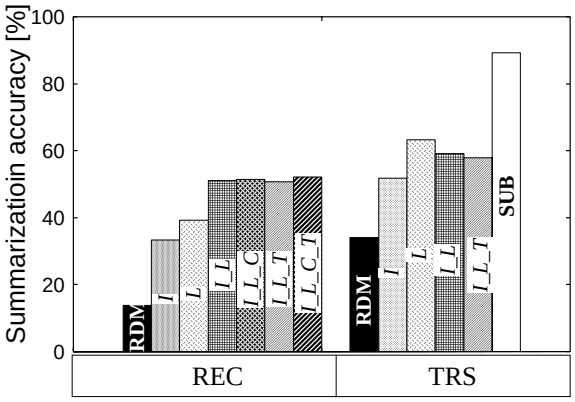


Figure 6.10: *Lecture summarizations at 50% summarization ratio. REC: summarization of recognition results, TRS: summarization of manual transcription, RDM: random word selection, C: confidence score, I: significance score, L: linguistic score, I_C,L_C,I_L: combination of 2 scores, I_L_C: combination of all scores, SUB: subjective summarization.*

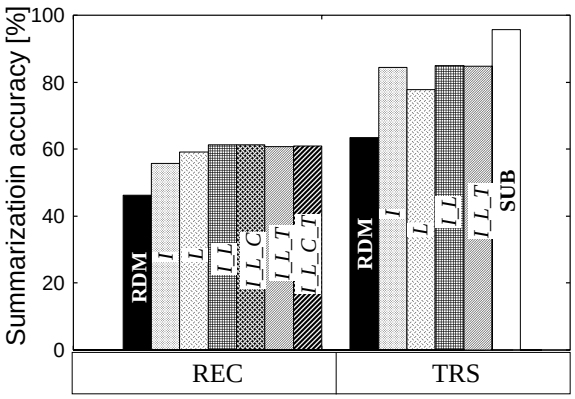


Figure 6.11: *Lecture summarizations at 80% summarization ratio. C: confidence score, I: significance score, L: linguistic score, I_C,L_C,I_L: combination of 2 scores, I_L_C: combination of all scores.*

Table 6.8: The result of automatic speech summarization for manual transcription and automatic recognition result.

書き起こし文
えー パラ言語情報ということなんですが あ 簡単に最初にえー 復習をしておきたいと思います ま あのー こうやって あっ 話しておりますとそれは勿論 あの 言語的情報を伝えるということが一つの重要な目的 ん なんでありますが 同時にパラ言語情報そして非言語情報が伝わっておりますま この三分法は藤崎先生によるものでして えー パラ言語情報というのは 要は あの 意図的に制御できる話者がちゃんとコントロールして出してるんだけども言語情報と違って連続的に変化するから カテゴライズすることがやや難しいそういった情報 で ありますで そういったものが音声生成過程の中で畳み込まれて まー 一次元の音声波として実際は出ている訳ですで 従来我々の研究といたしましてはこの部分を まー 音響的な手法によって 分析していた訳ですがきょうの発表の眼目は少し音声生成過程を遡りましてえー 調音運動の次元 で えー 従来の結果と一致するようなか違いが 観測されるかどうかということを見ようそういうのがきょうの発表の目的眼目でございます で 従来どのような成果が得られまし得られてきているかということを簡単にまたこれも復習になりますが えー これは今回用いました後程説明いたします笹田がという 文をえー 中立感心落胆そして疑いプラス反問というようなパラ言語情報を指定して えー ある話者が発音し分けたもの で ございます えー 具体的にどういうのかって言いますとえー 私が今ここで話者私ですんでやってみますとえー わ中立が笹田が感心が笹田が落胆が笹田が えー 疑いが笹田がというような形でありますえ 御覧のように えー F〇にも非常に大きな違いが観察されますしから デュレーションにも えー 非常に大きな違いが観察されます
書き起こし文自動要約結果
パラ言語情報ということなのだが最初にしたい やっているとそれは勿論言語的情報を伝えるということが一つ重要な目的なのだがパラ言 語情報非言語情報がいる この法は藤崎ものだがパラ言語情報というのは制御できる話者がコントロールしているの だが言語情報と違って連続的に変化するためカテゴライズすることがやや難しい情報である 音声生成過程の中で畳み込まれて一次元の音声実際は出ている 従来研究としてはこの部分を音響的な分析していたがきょうの発表の眼目は音声生成過程を 調音運動の結果と一致するような違いが観測されるかどうかということそれがきょうの発表の目的眼目である 従どのような成果が得られているかということを簡単にこれも復習になる これは後程説明する笹田がという文を中立感心落胆疑い反問というパラ言語情報を指定して話者が分けたものである 具体的にどういうものかと言うと私が今ここで話者私やって中立笹田感心笹田が落胆が笹田疑いが笹田いう形である F〇も非常に違いが観察されるデュレーションにも非常に違いが観察される
音声認識結果
えーパラ言語情報ということなんですが簡単に最初にえー復しゅうをしておきたいと思います まあのー声で話しておりますとそれは勿論あの言語的情報を伝えるということが一つの重要な目的がなりますが同時に パラ言語情報そして非言語情報が伝わったまず まこの三文ご藤崎先生によるものでしてえーは言語情報というのは今日はあのー意図的にそういうできる 話者がちゃんとコントロールした出してるんだけども言語情報と違って連続的に変化する肩ぐらいいーすることが や難しいそういったと答えます そういったものが温泉先生家庭の中で畳込まれてまー一次元の音素えーとして実際をやってる訳です で従来我々の研究で関しましてはこの部分をま音響的な情報によって分析していた訳ですが 表の発表の科目は少し音素えーせ下がってを遡るいましてえー調音運動の次元でえー従来のった結果と一致するような違いが 観測されるかどうかということを見ようそういうま表の発表の目的はもがございます で従来同様なそれがが得られました得られてきているかということを簡単にまたこれも復習になりますがえーこれは今回用いました あのちょうど説明いたします 刺さだかという文をえー注意する感心落胆そして疑いの三文とようなパラ言語情報してしてえある話者が発音してはけたものでございます えー具体的にどういうのかと言いますとえー私が今この話者が進んでいますとえー文字列がさ三番が関心がすその泊まったのは 落胆がそのそうだからえー疑いが三歳台がを有用かつあります えー御覧のようにえーFゼロにも非常に大きな違いが観察されますしがでしょうんでもえー非常に大きな違いが観察されます
音声認識結果の自動要約結果
パラ言語情報というなのが簡単に最初に復習 声でいるとそれは勿論言語的情報を伝えるということが一つの重要な目的がなるが同時にパラ言語情報非言語情報 この三文藤崎先生によるものは言語情報というのは今日は意図的に 話者がコントロールした言語情報と連続的に変化する肩ぐらいすることが難しいそういったと答える それらが温泉先生家庭の中で畳一次元の音素実際をして 従来我々の研究でしましてはこの部分を音響的な情報によって分析していた表の発表の科目は音素調音運動の次元で 従来の結果と一致するような違いが観測されるかどうかということを表の発表の目的はある 従来それがが得られているかということを簡単にまたこれも復習になる がこれは今回説明するだかという文を注意する感心落胆疑いの三文とようなパラ言語情報ある話者が発音してはけたものである 具体的にどういうのかと言うと私が今話者が進のでいると文字列が三番が関心がその泊まったのは落胆疑いが有用かつある 〇に非常に大きな違いが観察されるしが非常に大きな違いが観察される これは組織的な違い話者を越えて観察

Chapter 7

Conclusion

7.1 Problem Statement

Since LVCSR technology has made significance advancements, in the near future LVCSR systems can be expected to be applied to produce automatic closed captioning for broadcast TV programs, meeting/conference summarization, unstructured multimedia data management and natural, user-friendly interface with the computer. However, irrelevant information caused by recognition errors in transcription results is inevitable. Even if a speech recognition system could transcribe the utterance accurately, transcribed speech usually includes redundant information such as disfluencies, filled pauses, repetitions, repairs and word fragments. The transcription result including such redundant and irrelevant information cannot be directly used for making captions, indexing or making abstracts and minutes. Additionally, the speech interface to a machine using the transcription result needs to extract information according to the users' demands from speech. The speech recognition applications including speech interface and unstructured multimedia information management require understanding, the speaker's message and then extracting information to fulfill the user's or systems' demand. Therefore, especially for spontaneous speech, practical applications using speech recognizers require a speech summarization process which removes redundant and irrelevant information and extracts relatively important information depending on the users' or systems' requirements.

7.2 Contribution of the thesis

This dissertation has proposed a new approach for automatic speech summarization through word extraction. In this method, a set of words maximizing a summarization score indicating an appropriateness of summarization is extracted from automatically transcribed speech. Extraction is performed according to a target compression ratio using a DP

technique sentence by sentence. The extracted set of words is then used to construct a summarization sentence. A word significance measure, a confidence measure, linguistic likelihood, and a word concatenation probability are incorporated into the summarization score. The word concatenation score is determined by a dependency structure in the original speech given by SDCFG. This summarization process aims to maintain the original meaning as much as possible within a limited number of words. The contributions of this thesis can be summarized as follows:

- **sentence compaction technique depending on the amount of users' demand**

  Word extraction in the automatic summarization process is performed according to a target compression ratio (summarization ratio) using a DP technique sentence by sentence. The summarization ratio is given by users' or systems' derived summary length based on the number of words or characters. The extracted set of words is then used to construct a summarization sentence.

- **focusing on topic words**

  In order to generate summarization focusing on informative words (topic words) in transcribed speech, a word significance measure is incorporated into the summarization score. The summarization process extracts important information based on the significance measure.

- **generating readable summarization**

  Since spontaneous speech including redundant information caused by disfluencies is ill-formed and very different from written text, a technique to generate readable and understandable summaries is required. A linguistic score based on linguistic likelihood produces word strings where grammatically correctness is maintained as much as possible in a summarized sentence consisting of extracted words from the transcription.

- **special compact expressions for sentence compaction**

  In order to extract as much information as possible, the expression of summarized sentences need to be more compact under a lower summarization ratio. The special type of compact expressions found in headlines of newspaper text, such as skipping particles in Japanese, and E-prime in English respectively, can be used for summarization. A linguistic score constructed from using newspaper corpus can generate these compact expression for summarization of speech.

- **distilling technique by handling recognition errors**
  In order to alleviate the meaning alteration of summarized sentences by recognition errors, a confidence score based on the acoustic and linguistic reliabilities of recognition results is incorporated into the summarization score.

- **maintenance meanings based on a word dependency structure**
  The word concatenation score determined by a dependency structure in the original speech has been incorporated to maintain the original meanings. The word concatenation probability given by SDCFG is effective in maintaining the original dependency structure and meaning.

- **summarization approach for multiple utterances**
  In order to make abstracts, the proposed sentence summarization technique has been extended to summarization of multiple utterances. A set of words maximizing the summarization score is extracted from overall speech under some restrictions applied at the sentence boundaries. These restrictions realize the summarization of multiple utterances by handling them as a single long utterance. This process was used with the expectation that it would succeed in preserving more words inside information rich utterances, and shortening, or even completely deleting, less informative ones.

- **reduction of the amount of calculation for making abstracts**
  The amount of calculation for selecting the best combination among all available combinations of words in the multiple utterances increases as the number of words in the original utterances increases. A two-level Dynamic Programming (DP) technique is applied to a summarization method for multiple utterances in order to reduce the amount of calculation.

- **evaluation experiments**
  Each utterance and multiple utterances of Japanese broadcast news speech were summarized by the proposed method. Since this proposed method is based on a statistical approach, it can be applied to any language. English broadcast news speech was also automatically summarized. In order to apply the proposed method to English, the model of estimating word concatenation probabilities based on a dependency structure in the original speech given by a SDCFG was modified. Furthermore, a Japanese lecture speech was summarized to make academic conferences/meetings abstracts. This transcription of lecture speech was converted to the style of written text using an ad-hoc filter.

- **objective evaluation method including subjective variation**
  In order to evaluate automatic speech summarization, an objective evaluation method using summarization accuracy in comparison with manual summarization by human subjects is proposed. Manual summarization results are combined to build a word network, and word accuracy of each automatic summarization result is calculated compared to the most similar word string in the network. Since the network can approximately represent all possible correct summarization including subjective variations, the most similar word string in the network is assumed to be a correct answer for the automatic summarization result.

- **evaluation result**
  Experimental results show that the proposed method can effectively extract relatively important information and remove redundant and irrelevant information. A confidence score giving a penalty for acoustically, as well as linguistically unreliable words, reduces the meaning alteration of summarization caused by recognition errors. In contrast with the confidence score, the linguistic score effectively reduces out-of-context-word extraction both from recognition errors and human disfluencies. A word concatenation score giving a penalty for a concatenation between words with no dependency in the original sentence also reduces the meaning alteration of summarization.

7.3 Future research directions

Since the automatic summarization method proposed in this thesis remains on the level of distilling information through word extraction, further research is required to improve for speech summarization and understanding.

- **making a fine linguistic model based on summarized a corpus**
  The perplexity of the n-gram for the linguistic score is very high for manual summarization by subjects. A linguistic model that characterizes a feature of compact expression in a summarization is required.

- **estimation of word concatenation probability based on spontaneous speech**
  In this thesis, the word concatenation probability based on SDCFG was calculated using a manually parsed newspaper text. In order to estimate the dependency structure of spontaneous speech, the word concatenation probability must be calculated using a manually parsed corpus of spontaneous speech.

- **model to convert spontaneous speech style to written text style for summarizing spontaneous speech**

  Since spontaneous speech is ill-formed, spontaneous speech cannot be directly summarized using the linguistic score and the word concatenation score calculated from grammatically correct written text. In this dissertation, the conversion from spontaneous speech to written text was realized by using an ad-hoc filter. In order to generate a finer summarization, the construction of models to convert spontaneous speech style to written text style for spontaneous speech is necessary.

- **query biased summarization**

  A useful application of the summarization technique is to be able to extract information based on a users' query. Moreover, the summarization ratio could be automatically determined using calculation based on word collocation and text coherence.

- **summarization including reordering words**

  Since spontaneous speech includes repairs, a more highly refined summarization process needs to include reordering the words in transcriptions.

- **summarization from many hypotheses on a word graph**

  Taking into consideration many hypotheses on a word graph obtained by recognizers is expected to lead to more accurate summarizations.

- **intelligent summarization**

  In order to make finer abstracts for whole speech, the system must be able to recognize as useful information aspects of discourse structure such as follows;

  - coherence and cohesion
  - anaphoric relation
  - synonymy, polysemy, metaphors and metonymy

  However, the discourse structure must be estimated based on statistical models.

- **contribution of summarization score for word accuracy of a recognition system**

  Since the summarization of speech is a representation of users' messages extracted from speech, the word accuracy of recognizers could be improved by using feed back parameters in a summarization system.

Bibliography

[1] L. R. Rabiner and B. -H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[2] F. Jelinek, *Statistical Methods for Speech Recognition*, The MIT Press, 1997.

[3] B. -H. Juang and S. Furui, "Scanning the Issue," *Proceedings of the IEEE Transaction*, Vol. 88, No. 8, pp.1139-1140, August 2000.

[4] J. Gauvain, "Large-Vocabulary Continuous Speech Recognition: Advances and applications" *Proceedings of the IEEE Transaction*, Vol. 88, No. 8, pp. 1181–1200, August 2000.

[5] X. Huang, A. Acero and H. Hon, *Spoken Language Processing*, Prentice Hall, 2001.

[6] S. Furui et al., "Pattern Recognition in Speech and Language Processing," *CRC Press*, 2002 (will be published).

[7] T.Imai, A. Kobayashi, S. Sato, H. Tanaka and A. Ando, "Progressive 2-pass Decoder for Real-Time Broadcast News Captioning," Proc. ICSLP2000, Vol.I, pp. 246–249, Beijing, 2000.

[8] R. Valenza, T. Robinson, M. Hickey and R. Tucker, "Summarization of Spoken Audio through Information Extraction," Proc. ESCA Workshop on Accessing Information in Spoken Audio, 1999.

[9] Z.Klaus, "Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains," Proc. SIGIR2001, the 24th ACM-SIGIR International Conference on Research and Development in Information Retrieval, New Orleans(2001).

[10] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki and T. Ohdaira, "Toward the Realization of Spontaneous Speech Recognition — Introduction of a Japanese Priority Program and Preliminary Results —," Proc. ICSLP2000, Vol.III, pp. 518–521, Beijing, 2000.

[11] J. Allen, "Natural Language Understanding," The Benjamin/Cummings Publishing Company, 1988.

[12] T. Tokunaga, "Volume 5: Information Retrieval and Natural Language Processing," University of Tokyo Press, 1999 (in Japanese).

[13] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," In Proceedings of the 18th ACM-SIGIR Conference, pp. 68–73, 1995.

[14] T. Hand, "A Proposal for Task-based Evaluation of Text Summarization Systems," In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, pp. 31–38, 1997.

[15] I. Mani and M. Maubury, *Advances in Automatic Text Summarization*, The MIT Press, 1999.

[16] K. Zechëner "A Literature Survey on Information Extraction and Text Summarization," Paper for Directed Reading, 1996.

[17] M. Okumura, "A Study on Text Summarization Using a Large Electronic Dictionary," *Technical report*, 2000.

[18] H. P. Luhn, "The automatic creation of literature abstracts," *In IBM Journal of Research and Development*, pp. 159–165, 1958.

[19] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM* 16(2),pp. 264–285, 1969.

[20] K. Knight and D. Marcu, "Statistics-Based Summarization — Step One: Sentence Compression," *Proc. National Conference on Artificial Intelligence* (AAAI),

[21] C. Hori and S. Furui, "Automatic speech summarization based on word significance and linguistic likelihood", *Proc. ICASSP-2000*, Istanbul, pp.1579–1582 (2000).

[22] A. Ito, C. Hori, M. Katoh and M. Kohda, "Language Modeling by Stochastic Dependency Grammar for Japanese Speech Recognition", *Proc. ICSLP-2000*, Beijing, pp.I-246–249 (2000).

[23] C. Hori and S. Furui, "Improvements in Automatic Speech Summarization and Evaluation Methods", *Proc. ICSLP-2000*, Beijing, pp.IV-326–329 (2000).

[24] T. Hori, "A study on large vocabulary continuous speech recognition system," *Doctoral thesis submitted to Yamagata University*, 1999 (in Japanese).

[25] K. Shinoda, "A study on robustness against data insufficiency for speech recognition," *Doctoral thesis submitted to Tokyo institute of technology*, 2001.

[26] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 34, pp. 52–59, 1986.

[27] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[28] T. Kemp and T. Schaaf, "Estimating confidence using word lattices", Proc. 5th EUROSPEECH, Rhodes, Vol.2, pp.827–830, 1997.

[29] V. Valtchev, J.J. Odel, P.C. Woodland and S.J. Young, "MMIE training of large vocabulary recognition systems," Speech Communication Vol.22, pp.303–314, 1997.

[30] Christopher D. Manning and Hinrich Schütze, "Foundations of Statistical Natural Language Processing," *The MIT Press*, 2000.

[31] K. Lari, S. J. Young, "The estimation of stochastic context free grammars using the Inside-Outside algorithm," *Computer Speech and Language*, Vol4, pp.35-56, 1990.

[32] "Nihon Keizai Shimbun CD-ROM version 1990 version - 1994 version", *Nihon Keizai Shimbun, Inc.*,1994-1995.

[33] ChaSen (Japanese Morphological Analysis System),
http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

[34] K.Lari and S.J.Young, "Application of stochastic context free grammars using the Inside-Outside algorithm," *Computer Speech and Language*, 5, pp.237–257, 1991.

[35] T. Kawahara et al., "Sharable software repository for Japanese large vocabulary continuous speech recognition", Proc. ICSLP-1998, pp.3257–3260, 1998.

[36] C. Hori et al., "Advances in Automatic Speech Summarization," *Proc. EUROSPEECH2001*, vol.III, pp.1771–1774, Aalborg, 2001.

[37] http://www.cs.jhu.edu/~brill/

[38] Alex Waibel, Hua Yu, Martin Westphal, Hagen Soltau, Tanja Schultz, Thomas Schaaf, Yue Pan, Florian Metze, and Micheal Bett, "Advances in Meeting Recognition," Proc. HLT2001, pp.11–13, San Diego, 2001.

[39] http://www.cis.upenn.edu/~treebank/

[40] T. Hori, N. Oka, M. Katoh, A. Ito, and M. Kohda, "A Study on a Phoneme-graph-based Hypothesis Restriction for Large Vocabulary Continuous Speech Recognition," *Trans. IPS Japan*, vol. 40, no. 4, pp. 1365–1373, 1999.

Appendix A

Performance of SDCFG for Speech Recognition

The phrase-based SDCFG (Stochastic Dependency Context Free Grammar) described in Chapter 4 has been tested for its performance as a speech recognition system [22]. The SDCFG model constructed using the EDR corpus is compared with other types of SCFG shown in table 4.2, in terms of perplexity and computation requirements. In addition, a large scale SCFG and SDCFG using Mainichi news corpus is constructed and compared with a trigram model for recognition of a 5,000-word Japanese newspaper reading task. The conditions of the experiment in are described in table A.1.

Table A.1: Condition in evaluating SCFGs

Number of nonterminals	20	
Vocabulary size number of terminals)	3032 (Number of distinct words occurred more than twice in the corpus)	
Corpus	EDR corpus (Japanese corpus from newspapers and magazines)	
	Training text	Evaluation text
number of utterances	2000	100
number of words	53910	2782
UNK ratio	10.3%	22.0%

The perplexity for each model is shown in Figure A.1. Given a sentence consisting of $w_1 \dots w_L$, the perplexity of SDCFG, PK-SCFG, is calculated as follows:

$$Perplexity_{pk-scfg} = \log 2^{H_{PK-SCFG}} \tag{A.1}$$

where $H_{PK-SCFG}$ is an entropy of PK-SCFG.

$$H_{PK-SCFG} = \frac{\log P_{PK-SCFG}(w_1 \dots w_L)}{L} \tag{A.2}$$

where $P_{PK-SCFG}(w_1 \ldots w_L)$ is a production probability of a sentence consisting of $w_1 \ldots w_L$.

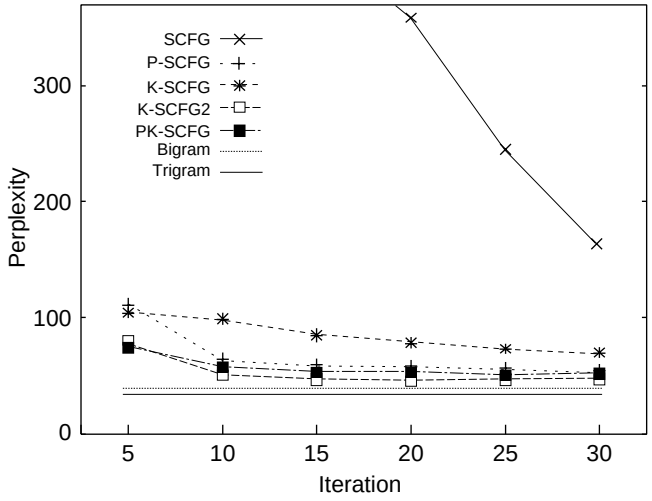


Figure A.1: Comparison of various types of SCFGs

Since the perplexities of the four enhanced models were lower than the original SCFG model, these models are considered to be more accurate language models. The perplexity of PK-SCFG was as good as that of K-SCFG2, which was the best model among five SCFGs. Figure A.2 shows training times of each model. The enhanced models were trained much faster than the original SCFG. The dependency grammar made the training process twenty times faster, and phrase boundary information made it eight times faster. As a result of these experiments, the expressive ability of the phrase-based SDCFG, PK-SCFG, is approximately similar to the bigram and the trigram models without the enormous amount of computation.

In addition, the performance of the PK-SCFG model as a language model for a LVCSR system was tested here. The task domain was read speech from Japanese newspaper articles from the Mainichi Shimbun. The vocabulary size was 5000. The test set consisted of 100 sentences without any OOV words. The training set was 46301 sentences chosen from the editions of Mainichi Shimbun from January to September of 1994, containing no OOV words. The number of nonterminals was set to 100 and 120. The initial values of the models were set in two steps. In the first step, all words in the training sentences were replaced with its category name, and SCFGS were trained using those category name sequences.

In the second step, output probability of a word was estimated as follows:

$$P'(\alpha \rightarrow w_c) = P(\alpha \rightarrow P(w_c))P(w_c|C(w_c)) \tag{A.3}$$

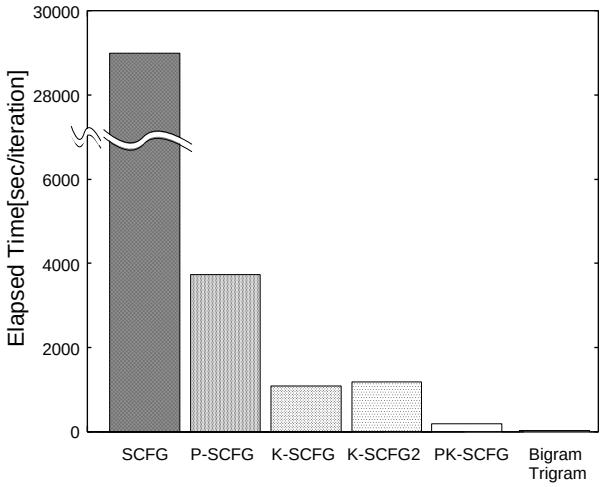


Figure A.2: Elapsed time for estimating parameters of SCFGs

$$P'(\alpha \rightarrow \beta w_f) = P(\alpha \rightarrow \beta P(w_f))P(w_f|C(w_f)) \tag{A.4}$$

where $C(w)$ denotes the category name of w. Using these initial values, the models were trained again using the original training sets. Acoustic models in these experiments were HM-Nets with state clustering [40] which had 2000 states of 16 Gaussian mixture. In this experiment, the 100-best candidates were generated from input speech using the bigram, then these candidates were rescored using the trigram and PK-SCFG. The total score of a candidate W for input speech O was calculated as follows:

$$\begin{aligned} Score(W|O) &= w_{pkscfg} \log P_{pkscfg}(W) + w_{ng} \log P_{ng}(W) \\ &\quad + acoust(O|W) + len(W) \times penalty \end{aligned} \tag{A.5}$$

where n was the length of the candidate, S_a was an acoustic score, P_1 and P_2 were probabilities from SCFG and n-gram respectively, w_1 and w_2 were language model weights of each model and p was an insertion penalty. The optimum values of those parameters are shown in Table A.2.

Figure A.3 shows perplexity of each model calculated upon test sentences. Perplexities of SCFGS are higher than those of the bigram and the trigram. Figure A.4 shows word error rates obtained through rescoring. This result shows that SDCFG is a good language model for a LVCSR system, as good as bigrams and trigrams. In addition, in comparison with only trigram or bigram applications, the combination of a trigram and SDCFG achieves further improvement of WER.

Table A.2: Optimum weight of linguistic score and insertion penalty

	w_{ng}	w_{pkscfg}	$penalty$
Bigram	16	0	-20
Trigram	17	0	-20
PK-SCFG 100	0	19	-16
PK-SCFG 120	0	18	-20
Trigram+PK-SCFG 100	14	10	0
Trigram+PK-SCFG 120	6	16	-12

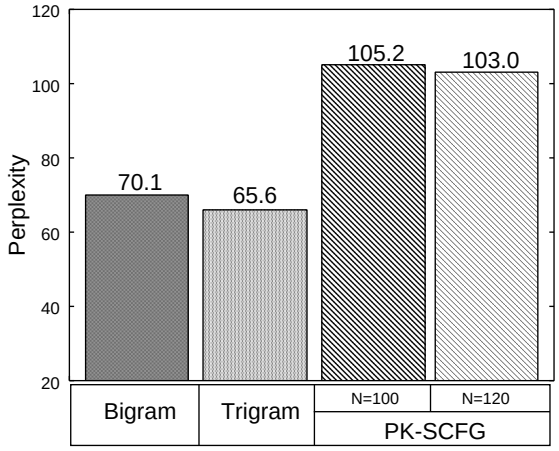


Figure A.3: Perplexity for correct sentences

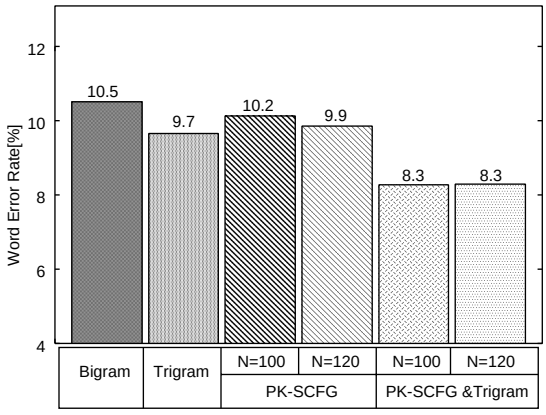


Figure A.4: Word error rate in speech recognition