早稲田大学審査学位論文
博士（人間科学）

Automatic Estimation of Speaking Style in Speech Corpora

２０１４年７月

早稲田大学大学院　人間科学研究科
沈　睿
SHEN, Rui
研究指導教員：　菊池　英明　教授

Contents

FIGURES

TABLES

Chapter 1: INTRODUCTION

In this chapter, two main topics: Speech corpora and issue of speaking style are to be discussed in Section 1.2 and Section 1.3 respectively. However, the specific aspects of speaking style like definition and quantification are to be discussed afterward.

In the following section, we discuss speech corpora and several applications of speech corpora. Additionally, the two search systems of speech corpora that we constructed and developed are also to be introduced.

1.1 Speech Research

Speech Researches are never stopped. At the beginning, speech researches were mainly done in the field of linguistics, especially in phonetics. However, with the development in some related fields, we realized the importance of speech researches and various fields of researches benefit from results of speech researches. For instance, in the field of linguistic development, speech researches contribute greatly (J. Van de Weijer 1997; etc.). Moreover, the technologies on speech processing are being promoted in the field of speech technology for practical realization.

While in this thesis, our main interest is on those speech researches aiming at contributing to the field of foreign language education.

1.2 Speech Corpora

In speech researches, speech data are crucial. The lack of speech data has always been a serious issue to discuss. Detailed databases with various annotations, which enable high-level studies, are called 'corpora'.

To construct a speech corpus is not simple because it requires considerable manpower and financial resources. Because of the limit of processing and arranging skills, most early corpora are limited to text corpora. For instance, the Brown Corpus from the early 1960s and the Lancaster–Oslo/Bergen (LOB) Corpus from the 1970s are renowned as text corpora of modern written English. In addition to written languages, several spoken language corpora have been constructed. For instance, the London–Lund Corpus of spoken English was finished in 1975.

Since the 1980s and in order to manage various issues in the application of speech, some large-scale speech corpora began to be constructed in several countries. For instance, the Spoken English Corpus (SEC) in 1987 and the Corpus of Spontaneous Japanese (CSJ) in the early 2000s are well-known speech corpora. CSJ will be discussed further in this thesis.

1.2.1 Application of speech corpora

Recently, with the development of computer technologies, the construction of large-scale speech corpora has become easier, and the application of speech corpora has widened in many research fields, especially in the fields

of phonetics, linguistic development, speech technology and foreign language education, as we mentioned above.

1.2.1.1 Phonetics

As the origin field of speech researches, researchers in phonetics prefer speech corpora mostly. A lot of honourable works have been done by utilizing speech corpora (T. Cho 2004, etc.).

1.2.1.2 Linguistic development

Through speech researches, we might find out the process of linguistic development of human beings. Some large scales of speech corpora, especially those of infants have been constructed (E.E. Lyakso 2010; etc.). Researchers are making great progress with the help of such speech corpora.

1.2.1.3 Speech technology

One of the most well-known applications is broadcasting in railway stations, which are usually realized through Text-To-Speech (TTS) technology. The content of such broadcasts are prepared in advance, and some speech models are required as well. To construct a speech model, the analysis of quantities of raw speech data is required, which can be found frequently in speech corpora.

The technology of speaker recognition is in wide demand in various fields, like Human Agent Interaction (HAI), speech dialogue systems, user verification and so on. To achieve high accuracies in the field of speaker recognition, raw speech data is required in order to construct a speaker's model. Speech corpora provide plenty of raw speech data to make the technology possible.

Similar to speaker recognition, speech recognition is a technology that is in demand in various fields. Technologies such as voice search, voice input, and voice response are all based on speech recognition. To construct a robust speech recognition environment, raw speech data of speech corpora is necessary.

1.2.1.4 Foreign language education

As we mentioned before, the application of speech corpora in the field of foreign language education, is the main concern in this thesis.

Although the utilization of speech corpora in this field is just at a preliminary phase, the potential of speech corpora, especially as a part of Computer-assisted Language Learning (CALL) system has been revealed. Some previous studies that utilize speech corpora are to be introduced in detail later.

1.2.2 Need for retrieval from speech corpora

Speech corpora are increasingly constructed for different purposes in various research fields. With the added requirements of speech data, several organizations collecting and managing linguistic resources that include such data have been launched worldwide. The Linguistic Data Consortium (LDC), launched by the University of Pennsylvania, USA, and the European Language Resources Association (ELRA), which is active in Europe, are two leading organizations in this field. However, because of the extremely different resources and standardizations provided from these two organizations, it might be difficult for users to select resources suitable to

their intended purpose. In Japan, the Speech Resources Consortium administered by the National Institute of Informatics (NII-SRC), is also actively collecting and publishing speech corpora. With the increasing amount of corpora, users gain more freedom of choice; therefore, efficient search and selection support need to be developed urgently.

1.2.3 Corpus Search

To manage the problem described in Section 1.1.2, a visualization system of multiple speech corpora based on the Multi Dimensional Scaling (MDS) method was constructed in 2011. The system, called Corpus Search (Y. Ishimoto 2011), was implemented with a search function of speech corpora under the management of NII-SRC. Furthermore, the relationship between these speech corpora can be visualized in this system. The system was constructed in a Linux operating system and made available to the public as a web application.

In the Corpus Search system, the content of multiple corpora are described with several attributes and items (K. Yamakawa 2009). Similarities among corpora are condensed into two dimensions using MDS. The corpora spatial arrangement applied with MDS is output, and a two-dimensional map is displayed in the Corpus Search system (Figure 1.1).

As shown in Figure 1.1, in the Corpus Search system, corpus users can select attributes and weigh items according to their different utilization purposes so that the matched corpora are placed in the Corpus Search in an

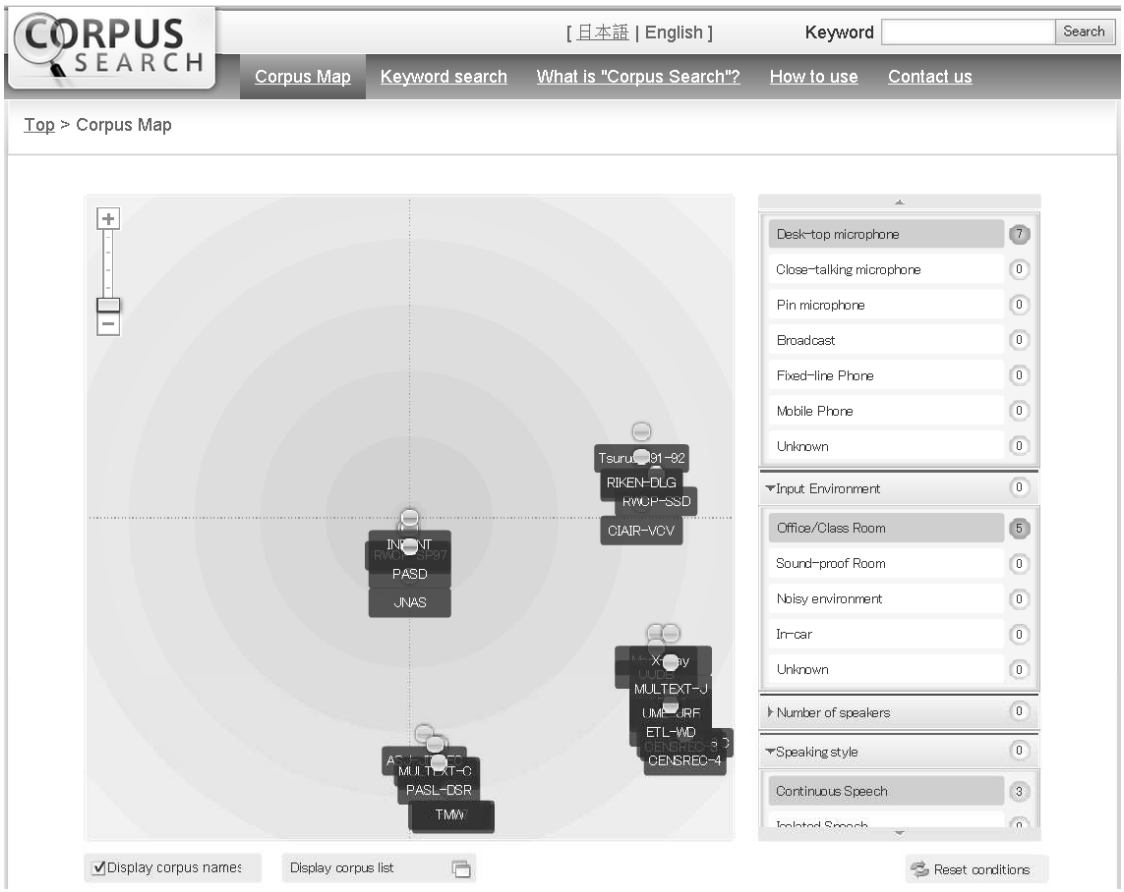


Figure 1.1: Corpus Search

arrangement that represents the relationship between these matched corpora.

However, despite the visualization, when comparing the similarities among speech corpora, 'to find the corpora I need' might be a priority for users with limited experience in speech corpora.

To evaluate the Corpus Search system, a questionnaire survey regarding its usability was conducted (Table 1.1). The result indicates that, compared to users who are familiar with speech corpora (more than five, in this case), users who know few corpora (less than five corpora, in this case) did not

Table 1.1: A questionnaire in a five-point scale (Y. Ishimoto 2011)

Questions	Average (Standard Deviation)	
	Less than 5 corpora known	More than 5 corpora known
1. Did it help?	*2.94*(1.06)	3.75(1.39)
2. Will you use it again?	*2.94*(1.06)	3.88(1.45)

consider that the system is helpful for their search purpose. Therefore, to expand the number of corpora users, a combination system of Corpus Search that is suitable for users with limited experience in speech corpora needs to be proposed and constructed.

1.2.4 Catalog-Search

With the goal of expanding the number of speech corpora users, we propose a combination system Catalog-Search with Corpus Search (R. Shen 2011).

To realize individualized corpora search, the current status survey of corpora use is first performed. Based on the proposal made in Ishimoto's work and referring to the results of the survey, ten attributes and 66 descendant corpus items (Table 1.2) are adopted. In the Catalog-Search system, all speech corpora are described with attributes and items similar to the method used in the Corpus Search system.

Presently, 26 speech corpora have been added and prepared in the

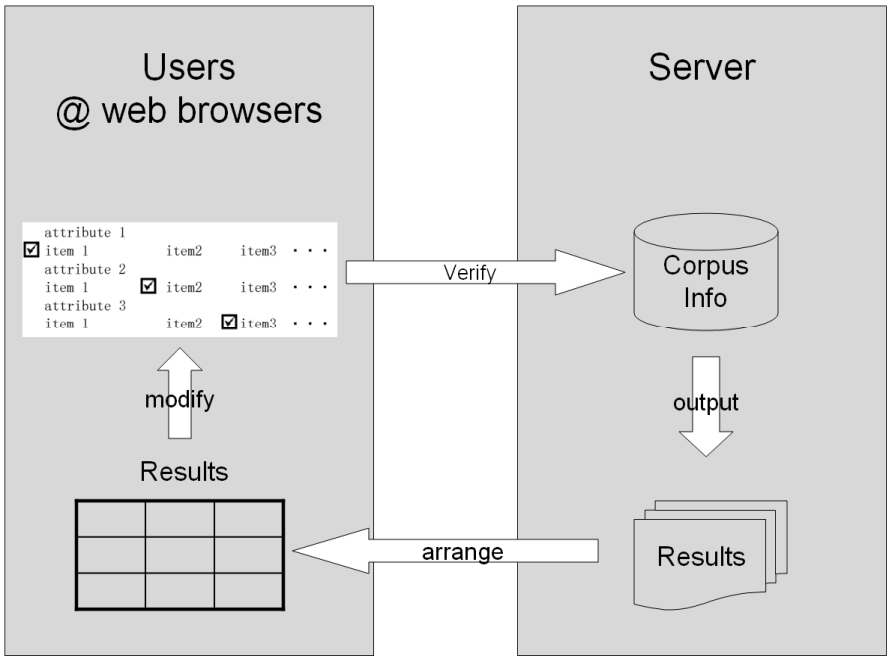


Figure 1.2: A system flow of Catalog-Search

Figure 1.3: Catalog-Search

Corpus	Sources	Environment	Speakers	Quantity	Style	Mode	Sampling rate	Data	Languages	Purpose
RWCP 音声対話データベース(96年版)	desktop microphone	office/classroom								
RWCP 音声対話データベース(97年版)	desktop microphone	office/classroom								
RWCP 会議音声データベース	desktop microphone	office/classroom								
重点領域研究「音声対話」対話音声コーパス	desktop microphone	office/classroom								
日本音響学会 新聞記事読み上げ音声コーパス	desktop microphone	office/classroom								
NTT 乳幼児音声データベース	desktop microphone	office/classroom								

Figure 1.4: An example of result tables in Catalog-Search

Table 1.2: Corpus attributes and items

Attributes	Items	Contents
Sources	7	Recording devices
Environment	4	Recording environment
Speakers	12	Numbers of speakers
Quantity	7	Quantity of data
Style	4	Speech style
Mode	5	Speech mode
Sampling Rate	3	Sampling rate
Data	9	Miscellaneous data
Languages	4	Languages of data
Purpose	11	Purpose of construction

Catalog-Search system. We also plan to add more speech corpora in the future.

1.2.4.1 System operation

Figure 1.2 shows the simple flow of the Catalog-Search system.

All user operations can be performed on web browsers. According to the utilization purpose, users should identify the attributes and items on which to focus; then, the filtering information is sent to the system server. Multiple identification in a single attribute is also possible (Figure 1.3).

Then, through verification with Corpus information on the system server, corpora that match the user filtering is arranged and listed in a table format and sent back to the user web browser.

Figure 1.4 shows an example of the resulting output. All matched corpora with detailed attributes and item information are listed on the user web browser, and by clicking the name of each corpus, the corpus information managed in NII-SRC is also displayed.

After reviewing the results, the users can modify their filtering using the results table on the same web page.

1.2.4.2 Application

As mentioned in the previous section, Catalog-Search is recommended as a combination system of Corpus Search; a combined use of these two systems might lead users to a more flexible and convenient search of speech corpora. Corpora that match attributes and item filtering can be found using the Catalog-Search system. Furthermore, the relationship between these matched corpora can be visualized using Corpus Search. Users already familiar with speech corpora might evaluate Corpus Search highly (Table 1.1); on the other hand, for users with limited experience in speech corpora, Catalog-Search might be more helpful.

1.2.4.3 Evaluation

Before making the Catalog-Search system available to the public, a survey was conducted to evaluate this system.

Ten college students (five males and five females between the ages of 20 and 29) participated in the survey. Before the survey, the participants were

asked to indicate the number of speech corpora with which they were familiar among the 26 corpora. The results indicated that almost all the participants had limited experience with the speech corpora (half of the participants indicated that they did not know any of the speech corpora, and only one participant claimed to know about more than five speech corpora). In this case, the participants can be considered similar to users who have limited experience in speech corpora.

Because this is a comparison evaluation of the Corpus Search and Catalog-Search systems, and in considering the sequence effect, the ten participants were divided into two groups with opposite sequences for the two systems.

The survey was conducted through a well-devised questionnaire. In the questionnaire, each participant was asked to finish six search tasks using both of the systems separately. Search tasks were set carefully so that participants can use the two systems extensively within the permitted time. To finish the search tasks, participants needed to use the systems to find corpora that match given conditions. For example, one of the search tasks was to search speech corpora collected in a noisy environment and constructed for speech recognition. To finish this task, the participants would have to identify the item 'Noisy Environment' in the attribute 'Environment', and the item 'Speech Recognition' in the attribute 'Purpose'. The matched corpora are considered the correct answer. After finishing all six search tasks, the participants were asked to rate the usability of the two systems using a five-point scale, and to provide their

Table 1.3: Results of rating evaluation

Questions	**Average（Standard Deviation）**	
	Catalog-Search	*Corpus Search*
a. User Interfaces		
1.fonts/size	***4.2***(0.75)	3.3(1.27)
2.buttons/windows	***3.5***(1.12)	3.3(0.78)
3.color	3.1(1.14)	4.4(0.66)
b. Usability		
1. operation	***3.9***(0.83)	2.6(1.11)
2.speed	***4.7***(0.64)	3.6(0.92)
3.result	***3.8***(0.87)	3.1(0.54)
c. Overall	*Catalog-Search*	
1. Did it help?	***4.2***(0.75)	
2. Will you use it again?	***4.2***(0.6)	

opinion freely at the end of the questionnaire.

The rating results from the questionnaire survey are listed in Table 1.3. Compared to the Corpus Search system, the Catalog-Search system obtained a higher rating for 'fonts/size' and 'buttons/windows' in 'User Interfaces', and 'operation', 'speed', and 'result' in 'Usability'. In the 'Overall' part, the Catalog-Search also obtained a satisfactory score (4.2 of 5). Based on the questionnaire, it seems that participants with limited experience in speech corpora granted the proposed Catalog-Search system a higher evaluation rate than the Corpus Search system.

The participants also provided their opinion about the Catalog-Search. For example, the participants indicated that in the 'User Interfaces' part, the border between two attributes should be clearer; they also indicated that

the search button at the bottom of the top page should be bigger. In the 'Usability' part, the participants were mostly concerned with the operation of the attributes and item filtering. For example, the participants indicated that the item 'Mixed-Language' should be added into the attribute 'Languages'; they also indicated that there should be more items in the attribute 'Data'. As was expected, because of different utilization purposes, corpus users have different views of speech corpora. To solve this problem, the attributes and items used to describe a speech corpus need to be standardized urgently (S. Itahashi 2010). Overall, a substantial amount of helpful advice was received for system improvements.

1.2.4.4 Summary

The Catalog-Search system is proposed and constructed as a Corpus Search combination system. In the Catalog-Search system, the attributes and items proposed to describe a speech corpus are modified and improved. To realize a more flexible and effective search and selection of speech corpora, a combination of the Corpus Search system and Catalog-Search system is proposed to not only match speech corpora with user search conditions and list the match results, but also to visualize the relationship among those speech corpora. It is considered that such a combination system can lead users to a more pleasant search experience between large quantities of speech corpora.

For future work in this study, modifications will be made to the system based on the opinions received from the participants about the proposed system mentioned in the previous paragraphs. Furthermore, the attributes

and items used to describe speech corpora need to be standardized. Moreover, the user interface and operation of the Catalog-Search should be improved based on the valuable opinion and advice provided by the evaluation participants. Finally, the combination of the two systems should be discussed in detail so that an overall evaluation can be conducted after making the systems available to the public.

1.2.5 Attributes to describe speech corpora

As mentioned in the previous section and as indicated by S. Itahashi, because of different utilization purposes, users have different views on speech corpora. To improve search usability, the attributes and items used to describe a speech corpus need to be standardized urgently (S. Itahashi 2010).

Conversely, the existing attributes and items remain limited. In the construction of the Corpus Search and Catalog-Search, most of the attributes and items used to describe the speech corpora are meta-information such as 'Sources', 'Environment', and 'Speakers'. To manage the diversity of various speech corpora, and to fulfil the extremely different needs for research, an increasing amount of attributes obtained from content analysis in speech corpora should be designed and introduced.

As mentioned before, K. Yamakawa organized those attributes and suggested new helpful attributes as well (K. Yamakawa 2009). For instance, the attribute of Sound/Noise (S/N) ratio was introduced. Obviously, an S/N ratio is known as a crucial attribute in a speech corpus. Several speech

corpora users consider this ratio to be an important parameter. By analysing the speech content in several speech corpora, it was discovered that the S/N ratio attribute is workable (K. Yamakawa 2009).

To improve user satisfaction with regard to the search of speech corpora for different research purposes, attributes such as S/N ratio obtained by analysing speech corpora content should be discussed and further developed.

1.3 Issue of speaking style

On the separate note, in this section, we discuss the main concern of this thesis, issue of speaking style.

The term 'speaking style' is commonly used in our daily life. According to the 1964 definition in the Uhlmann dictionary, the term 'speaking style' means 'the way of oral or written expression in its specific application' (A.M. Uhlmann 1964). Speaking style might be important in both spoken languages and written languages. For speech corpora that consist of spoken languages, speaking style should be considered an important attribute, without doubt.

The issue of speaking style is being discussed in various research fields. For instance, speaking style is considered an important skill in the field of sociolinguistics. In the field of speech and language engineering, especially in Text-To-Speech (TTS), the change of speaking style has been greatly discussed. Moreover, in the field of personality research, speaking style seems to be an important indication when describing personalities.

1.3.1 Speaking style in foreign language education

Notwithstanding the research fields mentioned in the previous section, we have more interest in the speaking style that concerns the field of foreign language education.

Speaking style has not been a focus in the field of foreign language education; however, speaking style is considered an important element in any language. S. Kori indicated that 'speech style', which is a term similar to speaking style, is an important issue that ought to be discussed when describing a specific language (S. Kori 2006). In E. Jorden's opinion, every language reflects stylistic differences, but the pervasiveness of the differences in Japanese is overwhelming. These opinions indicate that speaking style is attached to every language. Meanwhile, traditional strategies on foreign language education remain focused on grammar, vocabulary, syntax, and other fundamental linguistics elements. Teachers follow traditional strategies in their classrooms, and learners learn as the strategies suggest. As a result, most learners can grasp the basic parts of the target language, but when attempting to use the target language to practice speaking or writing, it appears that a gap occurs between the taught knowledge and the real language performance. H. Noyama referred to this issue using the term 'speech style', which includes the spoken languages (H. Noyama 2014). He indicated that in the real field of Japanese education, there is no systematic instruction on speech style, and learners have to realize and grasp the proper speech style through experimentation.

If learners have questions regarding speech style, their only option is to ask for help. According to Uhlmann's definition, the same problem can be assumed in written languages, which is usually discussed in terms of 'writing style'. This term has been discussed in several research fields through the analysis of text data, which will be discussed later in this thesis. Recently, some research on the support or correction of writing in foreign languages has been conducted, and the knowledge of 'writing style' is utilized. Thanks to the research results on text data and some advanced technologies on 'writing style', teachers and learners can substitute the basic elements of traditional strategies.

However, because of the processing difficulty of speech data, speaking style in speech data has not been discussed significantly. Information on speaking style is barely mentioned in manuals of speech corpora, which is the issue we discuss and consider in this thesis.

1.4 Structure of the thesis

The following structure of this thesis is as below.

In Chapter 2, the aim of this study is to be clarified. The main concern of this study, speaking style, is to be defined and the previous studies on speaking style are to be introduced in Chapter 3. Based on these previous studies, the general method of this study is to be explained in Chapter 4. In Chapter 5, details of the rating experiment are to be discussed and the result of this study, which is an estimation model of speaking style, is to be

introduced in Chapter 6. At the end, in Chapter 7, we conclude the study done by far and discuss about the possible further steps to enhance the perfection of this study.

Chapter 2: AIM

In this chapter, we discuss and clarify the goal of this thesis.

2.1 Speaking style as an attribute of speech corpora

As mentioned in the previous chapter, we consider that additional useful attributes should be introduced to improve user satisfaction in speech corpora. Similar to the S/N ratio attribute mentioned in the previous chapter, speaking style is a useful attribute to describe speech corpora, especially for users planning to consider the issue of speaking style in the foreign language field by utilizing the speech data in speech corpora.

The following two questions are problems that should be resolved.

1. To what extend can we attribute speaking style to speech corpora?

2. With the speaking style attributed to speech corpora, how can we consider the issue of speaking style in the field of foreign language education?

The answer to the first question depends on the answer to the second question. First, we need to clarify what it is that teachers and learners need to know and how they will utilize speech corpora. Based on this, we can

decide how to attribute speaking style in speech corpora (the first question).

2.2 Previous studies

To answer the question mentioned above, first we need to review some previous studies made in the field of foreign language education, especially on the utilization of language resources.

In this section, we introduce a workable system that utilizes computer technologies and a pilot corpus with a helpful tool that had been proved promising in the practical language teaching and pedagogic research as well.

2.2.1 Computer-assisted Language Learning

One of the trends in the field of foreign language education is using modern technologies such as Computer-assisted Language Learning (CALL). Much research has been conducted in CALL, and these systems have been found to be effective and efficient in actual teaching and learning.

There are several successful CALL systems available to the public. Here, we introduce a self-teaching CALL system for discriminating four tones in Chinese (Q. Sun 2012).

As is known in Chinese learning, four tones are always difficult to teach and learn. Chinese teachers usually require more time to explain repeatedly how the four tones should be pronounced, which is not very efficient. Q. Sun

and her colleagues developed a CALL system for the self-teaching of discriminating Chinese four tones, and released the system on the Internet for college students of Chinese learning. By selecting suitable uses among screening, practicing, and reviewing, and also among word lists of bi-syllabic combination of four tones with different degrees of difficulty, the required time for achieving the goal was reduced; furthermore, an evaluation experiment was conducted to prove the effectiveness of the system.

The method for learning languages through computer technologies allow learners to achieve certain goals without the aid of teachers, which shows the system's potential and is one of the directions in the near future for foreign language education. Discriminating four tones in Chinese learning is simply one application of the trend.

However, to realize additional high-level applications, corpora are indispensable to CALL systems. Not only learners, but also teachers and researchers can utilize corpora and CALL systems.

2.2.2 Chinese Learners' Speech Corpus

With the goal of providing raw data for learners, teachers and researchers identified and analysed mistakes and disfluencies made by learners; we used such data to construct the Chinese Learners' Speech Corpus (CLSC) and designed an XML browser for visualized operation, which is an example of the application of speech corpora (R. Shen 2009).

In essence, we designed the entire corpus according to the specifications of CSJ, which will be introduced in detail later in this thesis.

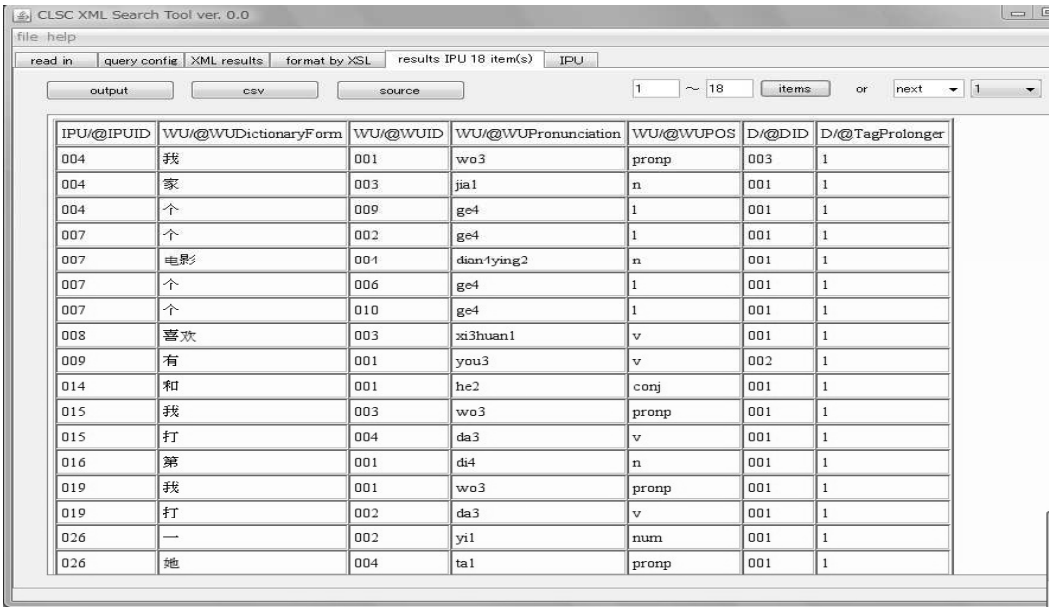
First, we conducted a survey based on a questionnaire at one of the institutions that teach Chinese as a foreign language in Shanghai, and obtained considerably useful information from some teachers about the types of annotation that are valuable to the research of language education, as well as to the teaching and learning of Chinese as a foreign language. Then, based on the questionnaire results, we collected a speech each from ten speakers that lasted approximately 100 minutes in total; four types of languages were involved. When transcribing the speech data, we made the usual annotations, such as 'word order', 'incompletion', 'falterer', 'repeat', 'mutter', 'meaning' and etc., to study mistakes. Then, we focused on the XML documentation, which is a feature of CSJ that has proven effective for eliminating conflict occurring among various types of information. The transcription data with morphological and integrated annotating information will be transformed into the XML documents.

However, the XML scheme in CSJ is not designed for a non-native corpus, which cannot be applied in the case of CLSC. Therefore, we modified the XML scheme to fit CLSC use. The detailed tags of the modified XML scheme are listed in Table 2.1. For instance, mistakes caused by word order can be retrieved by the 'TagOrder' tag, and disfluencies such as falterer can be found by the 'TagFalterer' tag.

Furthermore, we prepared an XML browser where detailed information and the annotations being processed in speech data can be visualized

Table 2.1: XML elements and attributes in CLSC

Element	Attributes
Talk	TalkID, SpeakerGender, SpeakerAge, SpeakerNationality, SpeakerMotherTongue, SpeakerLevel, CommentStrings
IPU	IPUID, IPUStart, IPUEnd,
WU	WUID, WUDictionaryForm, WUPronunciation, WUPOS, LineID
GD	TagOrder, TagIncomplete, TagMatch, TagParticle, TagDuplicate, TagLiang, TagMeaning, TagTone
D	TagFalterer, TagRepeat, TagProlonger, TagNative, TagMutter, TagTopicSwitch
SU	SUID, SUStart, SUEnd, SUSymbol



IPU/@IPUID	WU/@WUDictionaryForm	WU/@WUID	WU/@WUPronunciation	WU/@WUPOS	D/@DID	D/@TagProlonger
004	我	001	wo3	pronp	003	1
004	家	003	jia1	n	001	1
004	个	009	ge4	l	001	1
007	个	002	ge4	l	001	1
007	电影	001	dian4ying2	n	001	1
007	个	006	ge4	l	001	1
007	个	010	ge4	l	001	1
008	喜欢	003	xi3huan1	v	001	1
009	有	001	you3	v	002	1
014	和	001	he2	conj	001	1
015	我	003	wo3	pronp	001	1
015	打	004	da3	v	001	1
016	第	001	di4	n	001	1
019	我	001	wo3	pronp	001	1
019	打	002	da3	v	001	1
026	一	002	yi1	num	001	1
026	她	004	ta1	pronp	001	1

Figure 2.1: An example of results in XML browser

(Figure 2.1). Finally, we also present the analysis results of the disfluencies caused by prolonging and repetition in order to show the effectiveness of the CLSC corpus and the XML browser.

2.3 A recommendation system for speech corpora

According to the studies mentioned in the previous sections, we can see the effectiveness of a visualizable CALL system in the field of foreign language education. To return to the two questions posed in Section 2.1, we can indicate that a CALL system that recommends speech corpora to learners, teachers, and researchers with speaking style visualization appears to be a proper choice for the goal of this thesis. The processing required to realize the system is summarized in the following three sections; the first of these sections, 'Automatic estimation of speaking style' (Section 2.3.1), is the main discussion of this thesis.

2.3.1 Automatic estimation of speaking style

The first step is to estimate the speaking style from the contents of speech corpora so that the speaking style attribute can be determined. Because of the significant amount of speech data to be estimated, we need to find out how to estimate speaking styles automatically. The method will be discussed in this thesis later.

2.3.2 Visualization of speaking style in speech corpora

After estimating the speaking style, to improve system usability, the second step is to make the estimated speaking style of speech corpora visualizable. We consider two reasons for this step: 1. The estimated speaking style might not be simple to understand for learners and teachers.

2. Visualization allows the speaking style comparison of several speech corpora to be more comprehensive. Y. Ishimoto's work (Y. Ishimoto 2011) also proved this consideration.

2.3.3 Recommendation of speech corpora

For the last step and based on the position of the estimated speaking style in each speech corpora, we can recommend speech corpora to users searching for proper speech corpora for different purposes. We consider foreign language learners, teachers, and researchers as the potential users of this system.

Chapter 3: SPEAKING STYLE

To reach the goal outlined in the previous chapter, the first issue we have to consider is how to define speaking style and the types of measurement that can be used to quantify speaking style. In this chapter, we discuss these issues.

In Section 3.1, some previous studies on speaking style (or similar issue) are to be reviewed. In Section 3.2, we discuss about how to define and quantify 'speaking style'.

3.1 Previous studies on speaking styles

Since the middle of the 20th century, there have been studies that mention the issue of speaking styles. In 1964, Uhlmann defined speech style as 'the way of oral or written expression in its specific application' (A.M. Uhlmann 1964); according to this definition, speaking style might be important for both spoken and written languages. Then in 1968, from the aspect of sociolinguistics, Joos indicated that speaking style can be defined according to 'casualness' in speeches (M. Joos 1968), and in 1972, Labov mentioned that speaking style changes when the degree of attention that a speaker grants to his or her speech changes (W. Labov 1972). Through factor analysis, Biber summarized six factors from several text data of different

registers using linguistic features (D. Biber 1988). During the 1990s, Delgado and Freitas indicated that announcer news reports and teacher speeches in classrooms, which can be considered 'professional speech', are a type of speaking style (M.R. Delgado 1991). The lack or presence of speech scripts should also be considered when defining speaking styles, in Cid and Corugedo's opinion (M. Cid 1991). Moreover, Llisterri indicated five issues of speaking style in speech studies, one of which is the definition of speaking style. The author indicated that, although the studies mentioned in this paragraph were considered, because of the different focus on different fields of studies, the definition of speaking style has still not been provided in explicit detail (J. Llisterri 1992).

However, in the field of speech technology, speaking style is considered in a different manner. Recently, some studies of speaking style changes have been conducted. For instance, M. Abe and his colleagues introduced a method for changing speaking style among novels, advertisements, and encyclopaedias in TTS systems by controlling several prosodic parameters and formant frequencies (M. Abe 1994). K. Hirose and his colleagues introduced a method for synthesizing calm speech and three other speaking styles from texts using the F0 Contour Generation model (K. Hirose 2004). Both of the studies focused on the prosodic aspect of speaking style, but the linguistic aspect, especially speech transcriptions, will be considered in this thesis.

In the field of natural language processing, many studies have been conducted on the Topic/Author identification of written languages by

'writing style', which can be considered as part of speaking styles. Recently, H. Koiso and her colleagues introduced an effective method for topic identification of texts using the features of part-of-speech and word classification (H. Koiso 2009). In their work, from the text of seven different genres (including five written languages and two speech transcriptions of spoken languages), 150 texts from each genre and a total of 1,050 texts were selected for discriminant analysis. Morphological analysis was also conducted using UniDic, a lexicon for Japanese morphological analysis. Using the rate feature of some of the part-of-speech and word classifications and with the examination of the leave-one-out method, a discriminant model was constructed and the identification rate of seven-text genre was 79.9%. T. Koyama and his colleagues conducted similarities evaluation among several different research fields based on morpheme occurrence patterns in the transcription of conference proceedings. They also indicated the feasibility of constructing a distance scale to reproduce such similarities (T. Koyama 2008). Mairesse worked on an automatic recognition of personalities using linguistic clues and effective features, and the accuracy of several machine learning methods has also been discussed and compared (F. Mairesse 2007).

As J. Llisterri indicated, although some definitions and measurements have been provided for different purposes, there still is no common method for defining or quantifying speaking styles.

To cope with this issue, M. Eskenazi made a promising proposal, which is discussed in the next section.

3.2 Defining speaking style

After reviewing the issues solved in communities such as sociolinguistics and psycholinguistics referring to about 40 papers, which have been studying speaking style for a considerable time, Eskenazi granted the term 'speaking style' a data-driven definition, and determined three compatible scales to capture the nature of speaking style: Intelligibility-Oriented, Familiarity, and Social strata. The abstract is as follows (M. Eskenazi 1993):

> *In order to better capture the nature of speaking styles, it is useful to define the dimensions along which styles may be located. We have chosen to define three axes, compatible with the above definition.*
>
> *Intelligibility-oriented*
>
> *The degree of clarity that the speaker intends his message to have is the first dimension. It goes from minimum effort to be clear, when the message can be heard and understood well, to much effort when the channel is in some way noisy, or the listener has a problem understanding. This takes into account characteristics of the environment as well as the listener, and the judgement here concerns elements of a more physical nature.*

Familiarity

The expression of style may change greatly according to the speaker's familiarity with the listener. Extremes may go from identical twins to talking to someone from another culture who has little knowledge of the speaker's language and culture. The judgement here is based on more abstract elements, and takes the image of the listener in the dialogue context into account.

Social strata

The degree of cultivation that the speaker and listener wish to accord their conversation (see Labov) is the third axis. This goes from a totally colloquial and/or 'lower class' tone to a 'highly cultivated' and/or 'upper class' tone. It takes into account the context in which the conversation is taking place as well as the backgrounds of the two interlocutors at an abstract level. It should be noted here, that in a less limited style space, the interaction of the speaker and his environment must be taken into account, either on this third axis, or on another. Since the literature has not, to our knowledge, yet dealt with this aspect, it is not included here.

Briefly, the scale of Intelligibility-Oriented (I) represents the degree of clarity that the speaker intends for his speech to have. This differs from

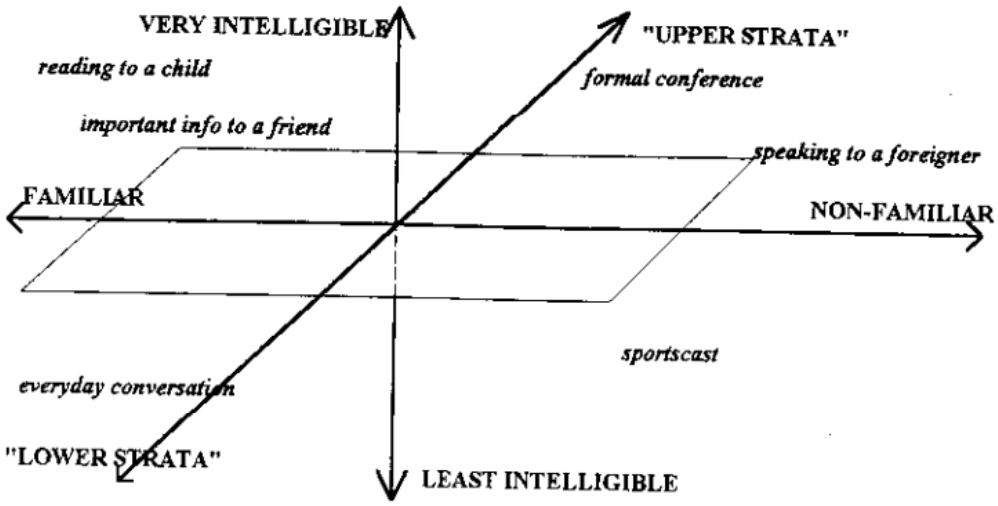


Figure 3.1: A three dimensional style space (M. Eskenazi 1993)

knowing that the listener can understand what the speaker says in a noisy background. It appears that the scale is more concerned with a physical nature. The scale of Familiarity (F), in a literal sense, means the degree of separation between the speaker and the listener. This scale might differ from identical twins to talking to a foreigner with limited knowledge of the speaker's language and culture. Occasionally, the dialogue context is also considered. The third scale, Social strata (C), appears to be more complicated. This scale represents the degree of cultivation that the speaker and listener intend to accord their dialogue. It differs from a completely colloquial (lower class) tone to a highly cultivated (upper class) tone. The context of the dialogue and the speaker and listener background need to be considered for this scale.

A figure of the three scales is used to explain the definition in Eskenazi's work (Figure 3.1). For instance, a sportscast does not usually attain a high cultural level; it is destined for a largely unseen, thus unfamiliar, public; and at the height of the action, provides less than maximum attention to intelligibility. Reading to a child finds a speaker who attempts to be a good cultural vehicle to an extremely familiar listener who is attempting to be well understood.

Chapter 4: METHODOLOGY

The main direction and method of the proposed system is discussed in this chapter.

To reach the goal indicated in Chapter 2, we consider that Eskenazi's proposal for defining speaking style appears to be suitable for this study for the following three reasons:

1. The definition must be fit to various wide range data. There are several types of data in speech corpora, such as readings, chatting, and speeches that consist of many partial units such as speakers and topics. These definitions or measurements that are constructed through a bottom-up method from limited types of speech data might not be able to cover certain speaking styles in various types of speech. Eskenazi's definition was summarized from many types of research conducted from different points of view that cover various wide range data; therefore, Eskenazi's definition appears proper for our aim of study. Certainly, the measurement constructed through a bottom-up method such as the one by Biber (D. Biber 1988) might be more credible; however, we focus on coverage first in this thesis.

2. Partial units of the speech rather than the entire speech should be estimated. To fulfil this need, categorised definitions such as 'news report', 'minute', or 'public speech' are not suitable because the changes or speaking

style shifts within speech data cannot be estimated in categories. However, Eskenazi's definition was summarized from many previous studies that considered various partial units in speech data, so it is not limited to large categories.

3. The definition shall be language-independent. Although every language reflects stylistic differences (E. Jorden 1987), a definition or a measurement constructed by the data of a certain language might not be adaptable to other languages. Eskenazi's definition was proposed based on the communication mechanism that does not depend on any particular language.

To summarize, Eskenazi's definition is considered suitable for estimating speaking styles in speech corpora, and it is adaptable to the proposed recommendation system.

The effective factors to be considered in order to realize the automatic estimation of speaking styles are acoustic factors (intonation, pause, etc.) and linguistic factors (morpheme, syntactic structure, etc.). We focus on linguistic factors mainly for two reasons. First, according to Eskenazi, although many of the factors that affect speaking style have been discussed, few studies have been conducted from the aspect of automatic speech processing (M. Eskenazi 1993). On the other hand, in the field of natural language processing, style discrimination and author estimation using linguistic features have been actively studied and some satisfactory results have been achieved (F. Mairesse 2007)(Koiso, 2009)(Koyama, 2008). Second, considering one of the applications of this study, which is to benefit learners for the perception and acquisition of speaking style, speech contents are

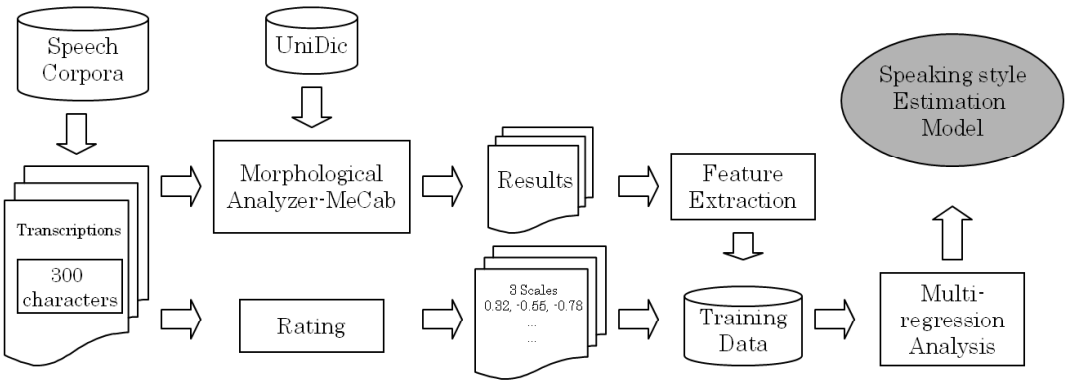


Figure 4.1: Construction of a speaking style estimation model

usually to be focused on firstly. In other words, as for the novice learners, it is more effective to improve speaking style acquisition through linguistic factors. Therefore, in this thesis, we attempt to focus on speech transcriptions and use the existing methods mentioned above to construct a speaking style estimation model to estimate speaking style in speech corpora automatically by referring to the three scales of speaking style proposed by Eskenazi.

The process for constructing the speaking style estimation model is shown in Figure 4.1.

First, in order to cope with the diversity of speaking styles, we choose several speech samples (including speech transcriptions) from speech corpora randomly. From those speech transcriptions, and from the middle part of each speech transcription, we extract approximately 300 characters as text stimuli to ensure the stability of perceptions of speaking style. Then, to collect the training data for the estimation model, participants are asked to rate the speaking style perceived in such text stimuli according to

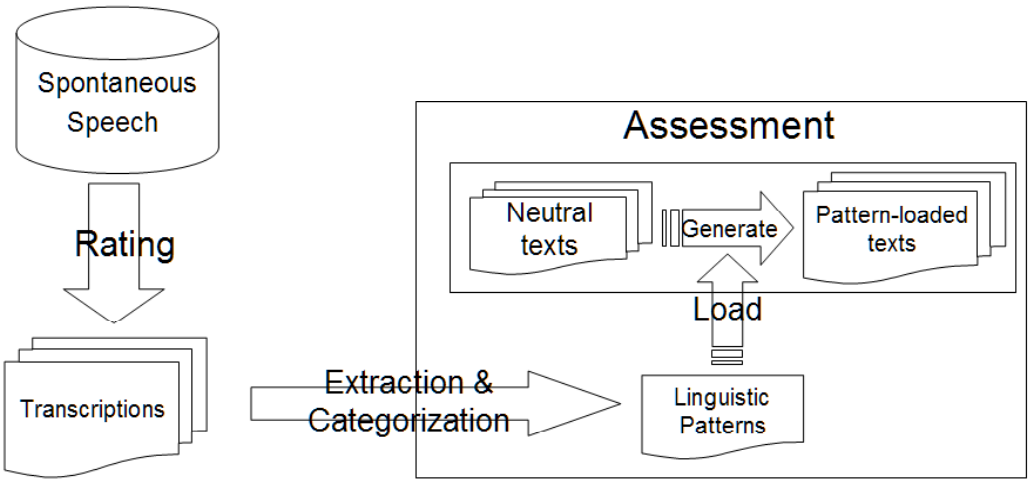


Figure 4.2: Procedure of the extraction of linguistic patterns

Eskenazi's three scales; the results are calculated as the score of the speaking style of each text. Morphological analysis using MeCab and UniDic are conducted to extract the part-of-speech, word classification, and linguistic patterns that are useful features for model construction. Last, we construct the estimation model by Multi-regression Analysis and using the proposed model, we can estimate the speaking style of any given speech transcription from any speech samples.

More specifically on the extraction of linguistic patterns (Figure 4.2), we first conduct an impression rating on spontaneous speech and select the transcriptions of the speech that is rated highly. Second, we extract and categorise linguistic patterns that help in perceiving speaking styles from the transcriptions. Third, we adapt those linguistic patterns into neutral texts to generate pattern-loaded texts, and using these texts, conduct an impression assessment to verify the effectiveness of these linguistic patterns.

Chapter 5: RATING

As mentioned in the previous chapter, in order to collect training data for model construction, we need to obtain ratings on speaking style. In this chapter, details of the conducted rating experiment are introduced, and the results are discussed.

5.1 Rater

A total of 22 college students majoring in information science participated in the rating experiment. None of the participants were involved in the preparation stage of this study.

5.2 Stimuli

For the rating experiment, we used speech transcriptions from various speech corpora as the text stimuli.

5.2.1 Corpora selection

Considering the cost of the rating experiment, and to cope with the diversity of speaking styles in speech corpora, we randomly chose ten speech transcriptions each from six categories of speech corpora (R. Shen 2012): CSJ1, CSJ2, FDC, MAPTASK, AUTO, and TRAVEL.

5.2.1.1 CSJ1 (K. Maekawa 2000)

For CSJ1, the transcriptions are selected from CSJ, which is a large-scale annotated speech corpus of spontaneous Japanese. CSJ is the outcome of the Japanese national priority area research project known as Spontaneous Speech: Corpus and Processing Technology (1999-2003) supported by the Ministry of Education, Culture, Sports, Science, and Technology. The entire CSJ contains approximately 650 hours of spontaneous speech that correspond to approximately 7,000k words with various useful annotations for research. CSJ is known as being of the highest level in both quantity and quality worldwide, and is widely used in the fields of speech processing, natural language processing, linguistics, phonetics, language education, dictionary compiling, etc. There are several speech sources in CSJ, but from CSJ1, we chose Academic Presentation Speech (APS) and Simulated Public Speaking (SPS). APS is live recordings of academic presentations in nine different academic societies, and SPS data consist of 'layman's speech' on everyday topics in front of a small, friendly audience. Both APS and SPS are monologues.

5.2.1.2 CSJ2

For CSJ2, the transcriptions are also selected from CSJ, but not from monologues, as is the case with CSJ1; the transcriptions in CSJ2 are from dialogues. The content of these dialogues is in regard to the speeches given in APS and SPS. Because of the diversity of speaking styles in CSJ, we decided to treat each speaking style as a different category.

5.2.1.3 FDC (S. Nakazato 2013)

For FDC, the transcriptions are selected from a Freshmen Dialogue

Corpus. In this corpus, the speakers are asked to chat about topics of their daily or school life. For the first recording, two speakers meet each other for the first time, or have low intimacy; for the second recording, two speakers have high intimacy. The purpose of the FDC construction is to provide raw speech data to investigate the relationship between the level of intimacy and the manner of speech.

5.2.1.4 AUTO (K. Miyazawa 2010)

For AUTO, the transcriptions are from dialogues between an interviewer and a driver or a navigator who has just finished an automobile-simulated driving task. The drivers or navigators are asked their thoughts and feelings during the simulation. The content of the dialogues is usually extremely limited verbally.

5.2.1.5 MAPTASK (Y. Horiuchi 1999)

For MAPTASK, the transcriptions are selected from the Chiba University Japanese Map Task Dialogue Corpus, also known as MAPTASK. MAPTASK contains task-oriented dialogues in which two speakers participate using maps: an instruction-giver who has a map with a route, and an instruction-follower who has a map without a route. The giver instructs the follower verbally to reconstruct the giver's route on the follower's map. We intentionally ignore the parameter 'acquaintance or not' and select speech transcriptions randomly. The speech in this category is task-oriented dialogues.

5.2.1.6 TRAVEL

For TRAVEL, free dialogues occur between two college students of the

same laboratory who are considered to be extremely familiar with each other. The topic is about 'to make a travel plan'. The transcriptions are selected as samples in TRAVEL.

5.2.2 Pre-processing

We randomly select ten speech samples from each category, as mentioned in the previous section; there are a total of 60 samples. However, almost all the samples in the speech corpus are longer than 10 minutes; therefore, we extract approximately 300 characters from the middle part of each speech transcription, which is considered sufficient to perceive speaking style. Moreover, in order to avoid distraction from the content of each transcription, we replace every noun (pronoun is not included) with a '○○' automatically (Figure 5.1) (A Java programme is used to make the conversion automatically). To manage different specifications in the transcription of different categories of speech corpora, we delete all the annotations, such as time information, fillers, falters, laughter, cough, and more, and maintain only characters and kanas with no punctuations.

5.3 Rating experiment

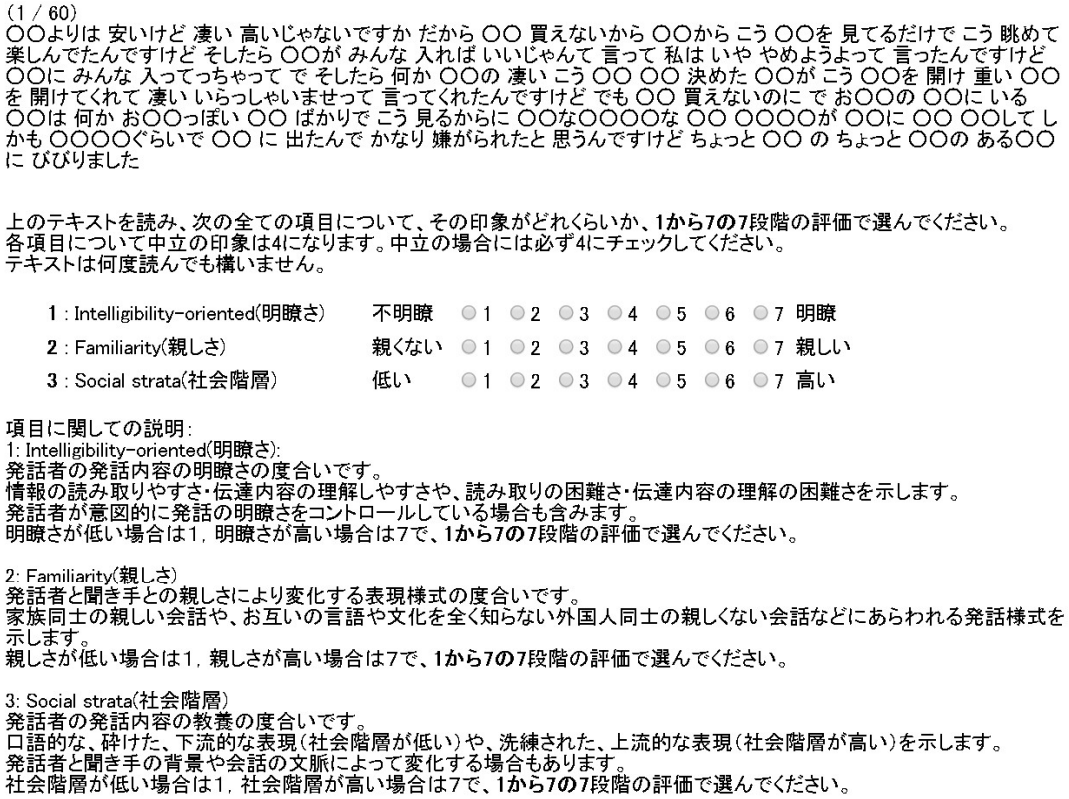
The rating experiment is conducted through a CGI on the web. All the participants are asked to rate for Eskenazi's three scales: Intelligibility-oriented, Familiarity, and soCial strata using a seven-point Semantic

Figure 5.1: Replacing nouns (pronouns not included) with "○○"

Differential scale method (SD method) after reading each transcription (one for least intelligible and seven for extremely intelligible, one for non-familiar and seven for familiar, one for lower strata and seven for upper strata). The text stimuli order for each participant is randomized. However, rereading is allowed and a time limit is not set.

However, before conducting the rating experiment, we decided to conduct a trial rating with three other participants in order to determine whether the explanation of the three scales can be well understood, and to confirm the rating stability. As a result, the Intraclass Correlation Coefficient (ICC (2,1)) (G.G. Koch 1982) of the three scales is 0.50, 0,79, and 0.82 for I, F, and C, respectively. According to Landis (J.R. Landis 1977), the ICC value for I is 'moderate', for F is 'substantial', and for C is 'almost perfect'; this shows the high credibility of the three participants in the trial rating.

Based on the results of the trial rating, we made minor corrections in the explanation of the three scales, and conducted the rating experiment. One of the stimuli sample is shown in Figure 5.2.


(1 / 60)
〇〇よりは 安いけど 凄い 高いじゃないですか だから 〇〇 買えないから 〇〇から こう 〇〇を 見てるだけで こう 眺めて 楽しんでたんですけど そしたら 〇〇が みんな 入れば いいじゃんて 言って 私は いや やめようよって 言ったんですけど 〇〇に みんな 入ってっちゃって で そしたら 何か 〇〇の 凄い こう 〇〇 〇〇 決めた 〇〇が こう 〇〇を 開け 重い 〇〇 を 開けてくれて 凄い いらっしゃいませって 言ってくれたんですけど でも 〇〇 買えないのに で お〇〇の 〇〇にいる 〇〇は 何か お〇〇っぽい 〇〇 ばかりで こう 見るからに 〇〇な〇〇〇〇な 〇〇 〇〇〇〇が 〇〇に 〇〇 〇〇して しかも 〇〇〇〇ぐらいで 〇〇 に 出たんで かなり 嫌がられたと 思うんですけど ちょっと 〇〇 の ちょっと 〇〇の ある〇〇 に びびりました

上のテキストを読み、次の全ての項目について、その印象がどれくらいか、**1から7の7**段階の評価で選んでください。
各項目について中立の印象は4になります。中立の場合には必ず4にチェックしてください。
テキストは何度読んでも構いません。

1 : Intelligibility-oriented(明瞭さ)	不明瞭	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	明瞭
2 : Familiarity(親しさ)	親くない	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	親しい
3 : Social strata(社会階層)	低い	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	高い

項目に関しての説明:
1: Intelligibility-oriented(明瞭さ):
発話者の発話内容の明瞭さの度合いです。
情報の読み取りやすさ・伝達内容の理解しやすさや、読み取りの困難さ・伝達内容の理解の困難さを示します。
発話者が意図的に発話の明瞭さをコントロールしている場合も含みます。
明瞭さが低い場合は1，明瞭さが高い場合は7で、**1から7の7**段階の評価で選んでください。

2: Familiarity(親しさ)
発話者と聞き手との親しさにより変化する表現様式の度合いです。
家族同士の親しい会話や、お互いの言語や文化を全く知らない外国人同士の親しくない会話などにあらわれる発話様式を示します。
親しさが低い場合は1，親しさが高い場合は7で、**1から7の7**段階の評価で選んでください。

3: Social strata(社会階層)
発話者の発話内容の教養の度合いです。
口語的な、砕けた、下流的な表現（社会階層が低い）や、洗練された、上流的な表現（社会階層が高い）を示します。
発話者と聞き手の背景や会話の文脈によって変化する場合もあります。
社会階層が低い場合は1，社会階層が高い場合は7で、**1から7の7**段階の評価で選んでください。

Figure 5.2: A sample of stimuli in the rating experiment

5.4 Rating results

In this section, we discuss about the results of the rating experiment.

Before the rating experiment, we consider I, F, and C as three independent scales of speaking styles. However, from the rating results of all 22 participants, we find that the correlation between I and F is -0.27 (weak), the correlation between I and C is 0.48 (moderate), and the correlation between F and C is -0.55 (moderate). The correlations indicate that the three scales might not be independent from each other. Eskenazi

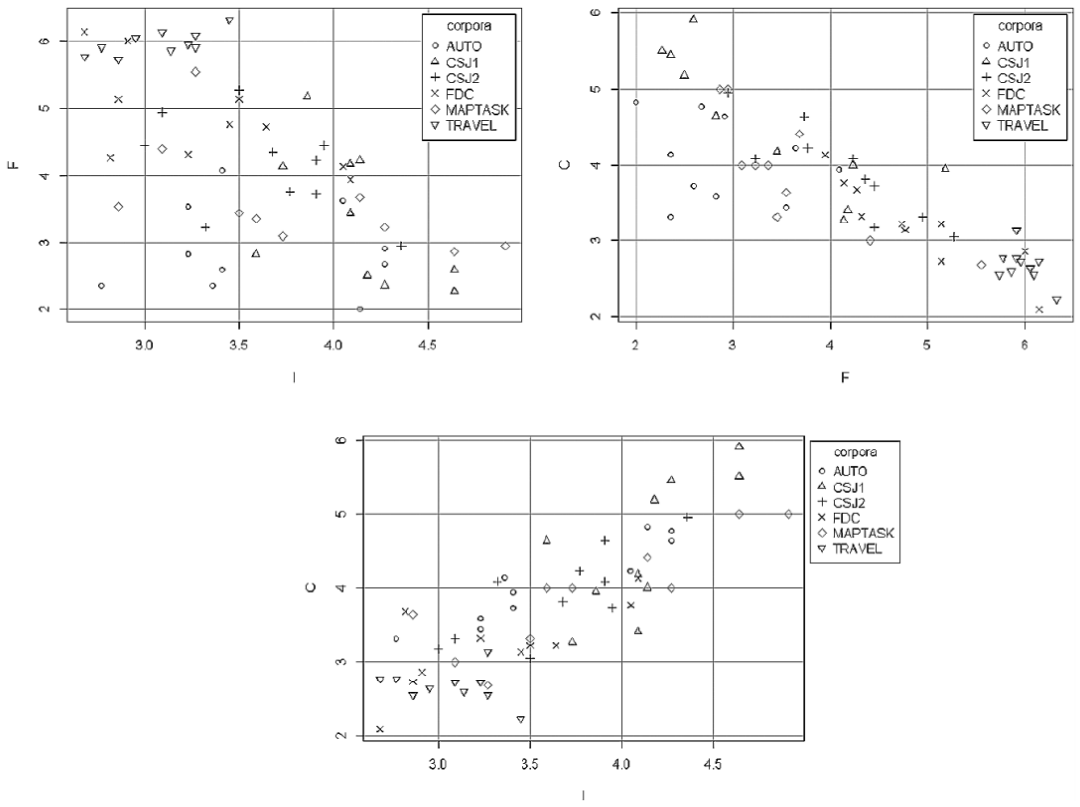


Figure 5.3: Distribution of three scales by pairs

did not discuss the correlations among the three scales particularly. However, it can be assumed that because all the transcriptions used in the rating experiment are in Japanese, and the manner of speaking in Japanese as well as the cultivation level of the speech content influence the relationship between Japanese speakers, the result is moderate correlations among the three scales. Because the goal of this study is to estimate and introduce speaking style as an attribute to describe and to search speech corpora, we consider that weak or moderate correlations would not harm the real application.

To verify the conformation between the features of the six corpus

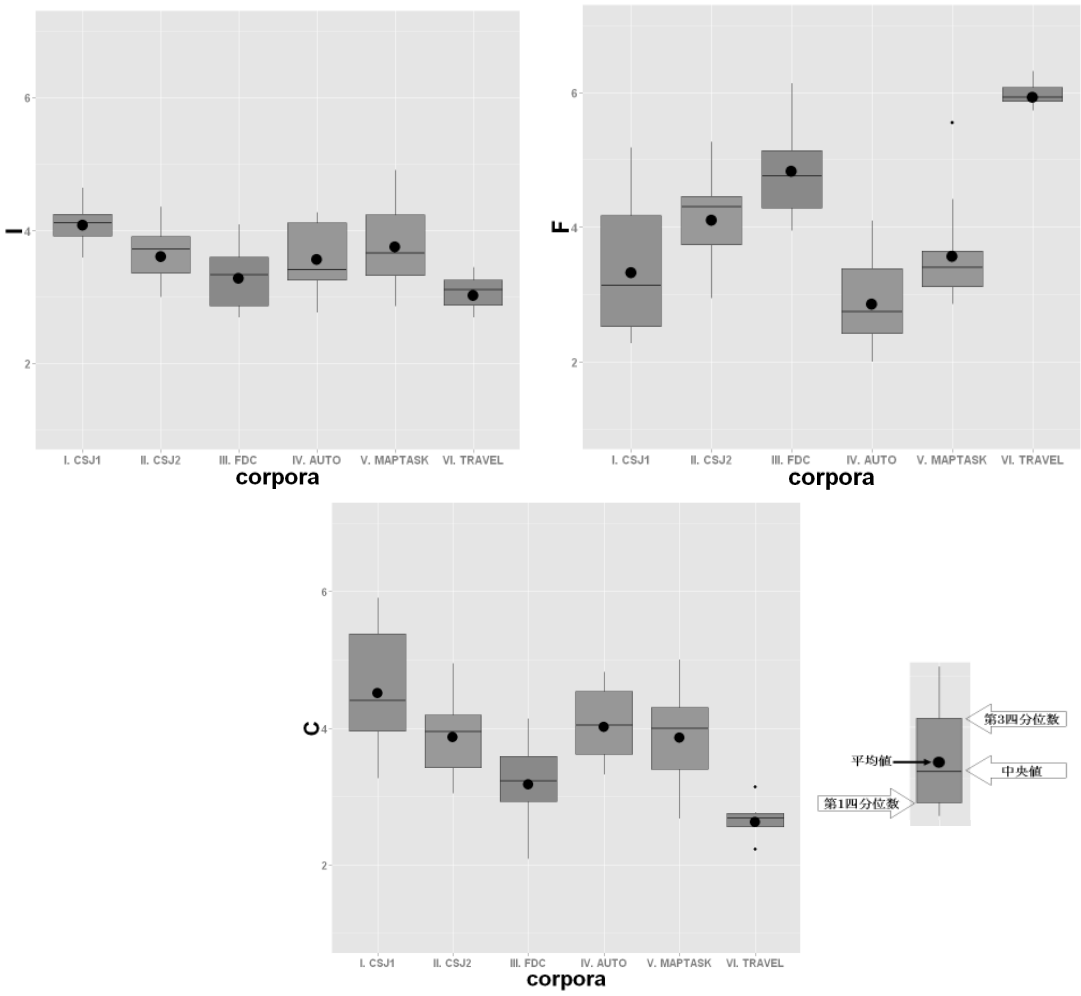


Figure 5.4: Boxplots of rating average

categories and the rating results, we observed the distribution of the 60 text stimuli (the average of all 22 participant ratings) on the dimension of the three scales. Figure 5.3 shows the distribution of the scales by pairs, and Figure 5.4 shows the average rating in a boxplot of each scale. For instance, the category CSJ1 distributes at the highest I (significant at the 0.05 level, with the exception of MAPTASK) and the average I of CSJ1 appears at the top of Figure 5.4. The reason might be that all the speech in CSJ1 is public speech, so speakers intentionally make themselves more intelligible. CSJ2

distributes a higher F and a lower C than CSJ1 (significant at the 0.10 level). This is because, unlike the monologues in CSJ1, the speech in CSJ2 is interview dialogues. FDC distributes at the low C (significant at the 0.05 level) because of the free dialogues between two college students. AUTO distributes at the low F (significant at the 0.05 level, with the exception of CSJ1) because of the verbal limitation in the dialogues. Compared to the other categories, text stimuli from TRAVEL distribute at the low I and high F (significant at the 0.05 level, with the exception of FDC). This is because the speech in TRAVEL is provided by two speakers from the same laboratory; therefore, the speakers should be extremely familiar with each other, and the topic is in regard to travel plans made in a daily life situation. According to Figure 5.3, it is clear that the results of the rating experiment reflect the features of the six categories; furthermore, the diversity of speaking style is ensured. However, the ratings of both the high I and the high F are not found in the results, and a few sparse octants remain, which indicates that additional speech corpora should be added to verify the effectiveness of our method.

To confirm the credibility of the rating results, we calculate the ICC of the three scales. ICC(2,1) (the credibility between raters) of I, F, and C are 0.11, 0.53, and 0.35, respectively, and ICC(2, k) (the credibility of the rating averages) of I, F, and C are 0.72, 0.96, and 0.92, respectively. Landis indicated that the range 0.0-0.20 means 'slight', 0.21-0.40 is 'fair', 0.41-0.60 is 'moderate', 0.61-0.80 is 'substantial', and 0.81-1.00 is 'almost perfect' (J.R. Landis 1977). Accordingly, of all the three scales, F and C are acceptable for

ICC(2,1), and I, F, and C are more than 'substantial' for ICC(2,k), which means that, because of the sufficiency of the raters, the rating averages are all sufficiently credible to be utilized in the model constructions described in Chapter 6. However, the rating difference in the I scale is significantly large, which is not preferable in the real application. The I scale might need to be discussed further in the future.

Chapter 6: MODEL

In this chapter, based on the rating results we discussed in the previous chapter, we start to construct the estimation model as proposed.

In Section 6.1, we discuss about the effective features to represent speaking style in speech corpora. By quantify features in speech transcriptions, we introduce the construction of the estimation model of speaking style in Section 6.2.

6.1 Features

What kinds of features in transcriptions are helpful to perceive speaking style from speech? That is the issue we discuss in the following parts of this section.

First, we need to review some workable studies on features in texts, especially those that are relevant to speaking style (or some similar issue) perception.

6.1.1 Previous studies

Some research has been performed on effective features in the description or discrimination of style.

H. Koiso and her colleagues introduced an effective method of topic identification of texts using the features of part-of-speech and word classification (H. Koiso 2009). In their work and from seven different text genres (including five written languages and two speech transcriptions of spoken languages), 150 texts each, and a total of 1,050 texts were selected for discriminant analysis. Using the rate features of some of the part-of-speech and word classification and with the examination of the leave-one-out method, a discriminant model was constructed and the identification rate of the seven-text genre was 79.9%. T. Koyama and his colleagues made similarities evaluation among several different research fields based on morpheme occurrence patterns in the transcriptions of conference proceedings. They also indicated the feasibility of constructing a distance scale to reproduce those similarities (T. Koyama 2008). Mairesse worked on an automatic recognition of personalities using linguistic clues and effective features, and the accuracy of several machine learning methods has also been discussed and compared (F. Mairesse 2007). S. Kinsui advocated that some particular manners of speaking (which is another concept similar to speaking style) made by speakers can cause listeners to recall certain characters; such manners of speaking are called 'role language' or '*yakuwarigo*' (M. Teshigawara 2011). He listed linguistic features used by imaginary characters in novels, dramas, anime, translations, etc. as examples.

Based on the studies mentioned in the previous paragraph, we decided to use part-of-speech, word classification, and linguistic patterns as features in

the model construction.

6.1.2 Part-of-speech and word classification

To summarize the information on part-of-speech and word classification, we conduct morphological analysis on all the 60 text stimuli using MeCab and UniDic. Then, we calculate ten rates of part-of-speech (Auxiliary, Verb, Adverb, Pronoun, Adnominal, Conjunction, Particle, Adjective, Interjection, and Prefix) and word classification (function words) in each category. The details are shown in Figure 6.1.

As can be seen, the rates for part-of-speech and word classification are related to the features of different categories of speech corpora. For instance, the pronoun (pron) rates in those categories that consist of more free speech, such as CSJ2, FDC, and TRAVEL, are significantly higher than those categories that consist of less free speech, such as AUTO and MAPTASK (at the 0.05 level). The auxiliary (aux) rate in those categories that consist of speech that is limited verbally, such as AUTO, is significantly higher than the others (at the 0.05 level).

To summarize, features such as the rate of part-of-speech and word classification appear to be extremely promising in helping with model construction.

6.1.3 Morphological patterns

In this section, we discuss about the process of extracting and verifying morphological patterns in texts that had been proved effective in style

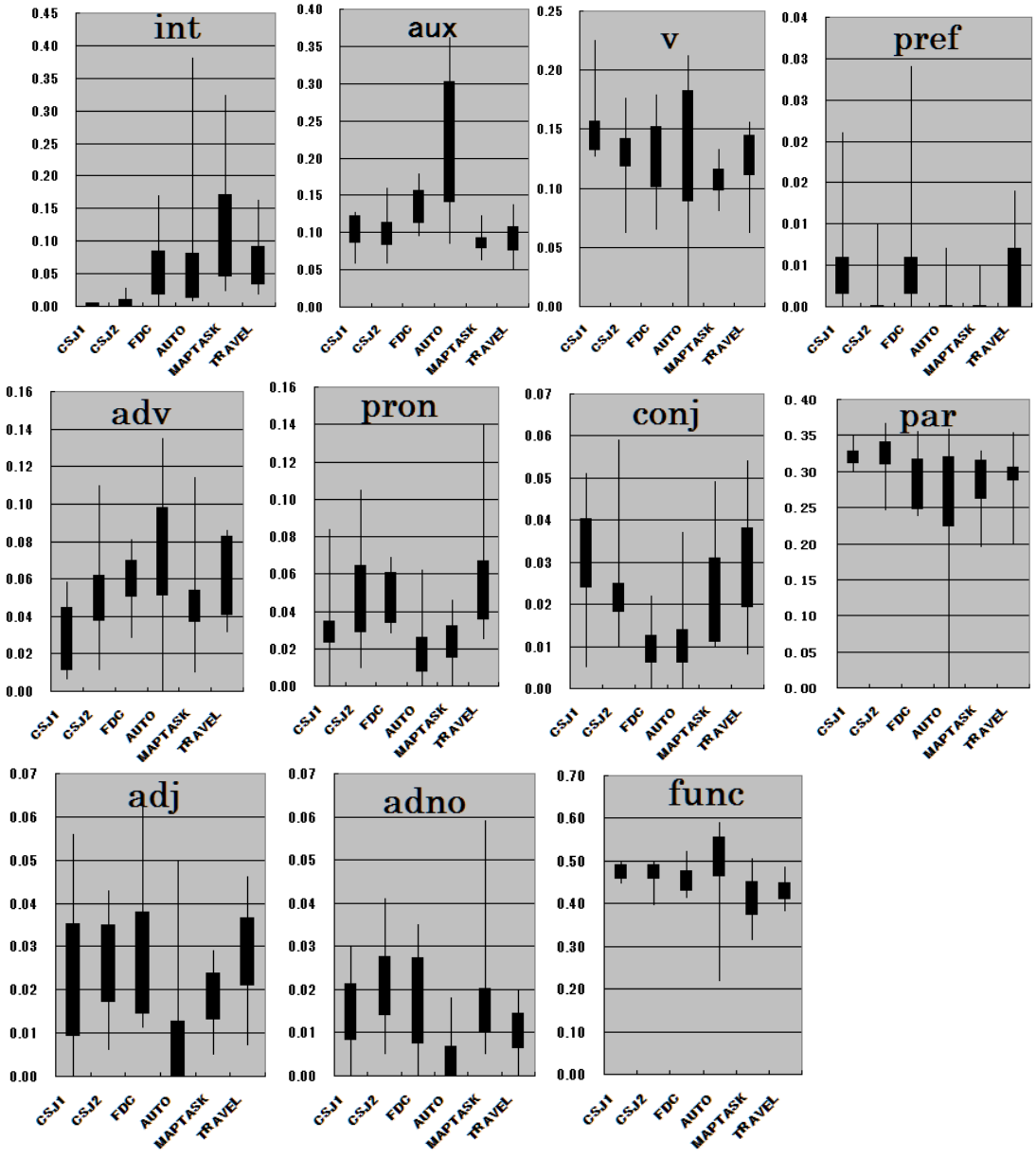


Figure 6.1: Rates of part-of-speech and word classification in the 6 categories of speech corpora

perception.

First, we conduct an impression rating on spontaneous speech and select the transcriptions of those speeches that are highly rated. Second, we extract and categorise linguistic patterns that help in perceiving speaking

styles from the transcriptions. Third, we adapt those linguistic patterns into neutral texts to generate pattern-loaded texts, and using these texts, conduct an impression assessment to verify the effectiveness of those linguistic patterns.

6.1.3.1 Impression rating of spontaneous speech

To prepare for the patterns extraction, we first conducted an impression rating experiment.

A total of 19 individuals participated in the experiment. Most of the participants were college students and research workers.

The speech samples we used are all from CSJ (K. Maekawa 2000). CSJ is a large-scale annotated corpus of spontaneous Japanese that contains approximately 650 hours of spontaneous speech corresponding to approximately 7,000k words. We chose SPS (Simulated Public Speaking) in CSJ as speech samples. SPS data consist of approximately 10–12 min of 'layman's speech' on everyday topics in front of a small, friendly audience. Prepared texts were not allowed, but the speakers were encouraged to create outlines of their talks. Because each speech data sample in CSJ is approximately 10–12 min in length, which is considered too long for impression formation, we chose 30 samples with balanced speaker properties and selected the middle part (approximately one minute) for use. The silent interval was set at fifteen seconds after each speech sample so that participants could have sufficient time to finish and verify their ratings.

Following the 'Psychological Scale for the Impression Rating of Monologue' (K. Yamasumi 2005), we chose skilfulness, likes, fastness,

activeness, seriousness, couth, uniqueness, cuteness, healing, expressiveness, and urbanity as the target types for impressions in this experiment. The explanation listed in Table 6.1 in Japanese was provided to participants before the experiment in order to avoid misunderstandings.

We divided the 19 participants into four groups and conducted the experiment separately for each group. The participants were asked to rate the 11 types of impressions listed in the previous paragraph while playing the speech samples, and all the speech samples were played only once. Before the experiment, we prepared another three samples for rehearsal. The 30 samples were played in random order; meanwhile, participants were asked to rate their impressions on questionnaire sheets. Questions were set using a five-point scale (one is negative, three is neutral, and five is positive). As mentioned in Section 2.3, the explanation for each type of impression was also shown on the questionnaire sheets so that participants could refer to it occasionally during the experiment. Each run of the experiment was divided into first and second halves, each of which contained 15 samples in order to reduce to some extent the fatigue effect. We also asked participants not to be influenced by the content of the speech samples so that the influence from the topics could be reduced.

Given the rating results, we subjectively focused on three typical types of impressions: seriousness, cuteness, and expressiveness for pattern extraction. Finally, we selected all the speech transcripts that had an

Table 6.1: Explanation for target types of impressions

Types	Explanation
Skillfulness	Smooth/influential/spontaneous/slick
Likes	Pleasant/good feeling/fond/approachable
Fastness	Fast/speedy/bustling/restless
Activeness	Vigorous/loud/positive/energetic
Seriousness	Both speaking style and character
Couth	Both speaking style and character
Uniqueness	Both speaking style and character
Cuteness	Both speaking style and character
Healing	Speaking style
Expressiveness	Both speaking style and character
Urbanity	Speaking style

average rating of more than 3.4 points (on a five-point scale, the upper 40% is considered positive) for each of the three target types of impressions mentioned in the previous paragraphs. With these high rating speech transcripts, we could proceed to the next operation.

6.1.3.2 Pattern extraction and categorisation

After obtaining the results of the impression rating experiment described in the previous section, we conducted an operation of pattern extraction and categorisation in order to determine the relationship between impression formation and speech transcripts.

First, we asked certain participants to highlight the particular patterns or expressions that help to form impressions so that we could refer to them for extraction and categorisation. There were seven participants in the

experiment. Three of the participants were college students, and the other four were research workers. All of them had participated in the rating experiment mentioned in the previous paragraphs.

First, the participants were asked to read the speech transcripts and to indicate the clues or expressions that helped them form the aforementioned three types of impressions. For instance, if the adverb 'absolutely' occurred in the transcripts, and the participants felt that it helped them form the impression of Expressiveness, they were asked to highlight the word 'absolutely' in a different colour. Of course, these patterns might occur not only at the word level, but also at the phrase or utterance level. There was no time limit, and the participants were allowed to review the transcripts freely. Occasionally, one transcript was read several times to obtain different types of impressions.

Then, we collected the seven participant responses and arranged them manually in a single chart to summarize the results. Because we required the most number of patterns as possible, we focused on the patterns that were highlighted by half or more of the participants, but did not neglect minor results. Fortunately, we were able to extract many effective patterns (Figure 6.2).

Figure 6.2 clearly shows the patterns that attracted the most attention. The first seven columns contain the responses of the seven participants. Columns 8 to 11 show data from the extraction process and comments. We manually analysed the position of these patterns in the entire utterances (start, middle, or end of a segment). We also noted the linguistic level

(morphology, pragmatics, etc.), and morpheme (part-of-speech, conjugation, etc.). In the last column, we entered our comments and summary of the patterns that might be helpful for the next experiment.

Finally, after completing pattern extraction and analysis and in order to prepare for the next experiment, we categorised the result patterns.

In Table 6.2, we categorised the result patterns by part-of-speech. Some particular phrases or utterances were categorised as either 'in other words' or 'insertion'. The cells highlighted in green show the shared parts among types of impressions.

6.1.3.3 Impression assessment of pattern-loaded texts

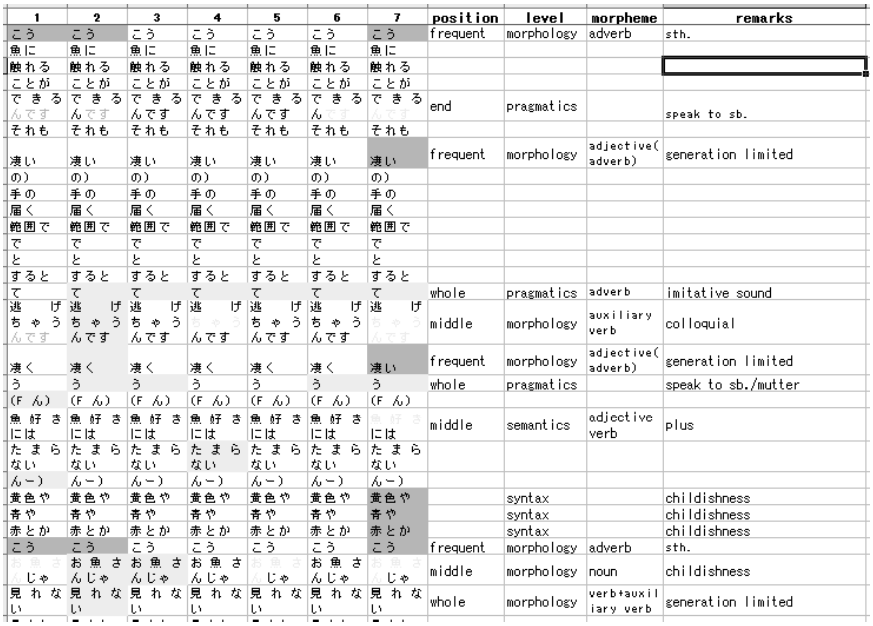
After obtaining the results of the rating rankings described in the previous section, we conducted the impression assessment of pattern-loaded texts. The purpose of this experiment is to verify the effectiveness of the patterns we extracted and categorised by assessment of the pattern-loaded texts.

There were 11 participants in the experiment. They were divided into two groups. All of the participants were research workers, and none participated in the experiment and the work described in the previous section.

We used neutral and pattern-loaded texts for assessment. Both text types were generated from origin texts. We attempted mainly to search the feasibility of the approach using patterns; text in conditions ideal for assessment is required first. Therefore, the spontaneity of the pattern-loaded texts was not confirmed in advance. However, as Figure 6.3 shows, compared with the neutral text, the pattern-loaded text does not appear

Table 6.2: Samples of extracted pattern categorisation

	Cuteness	Seriousness	Expressiveness	Types of Impressions
In other words	このように＞こう konoyohni＞koh	その時＞当時 sonotoki＞tohji	そういう＞そんな sohiu＞sonna	
Adverb	凄い sugoi	とりあえず toriaezu	絶対 zettai	
Conjunction(phrase)	･･･や･･･とか ･･･ya･･･toka	～て＞～まして ～te＞～mashite	そしたら(頻発) soshitara(freq)	
Categories				



1	2	3	4	5	6	7	position	level	morpheme	remarks
こう	こう	こう	こう	こう	こう	こう	frequent	morphology	adverb	sth.
魚に	魚に	魚に	魚に	魚に	魚に	魚に				
触れる	触れる	触れる	触れる	触れる	触れる	触れる				
ことが	ことが	ことが	ことが	ことが	ことが	ことが				
できる	できる	できる	できる	できる	できる	できる	end	pragmatics		speak to sb.
んです	んです	んです	んです	んです	んです	んです				
それも	それも	それも	それも	それも	それも	それも				
凄い	凄い	凄い	凄い	凄い	凄い	凄い	frequent	morphology	adjective(adverb)	generation limited
の)	の)	の)	の)	の)	の)	の)				
手の	手の	手の	手の	手の	手の	手の				
届く	届く	届く	届く	届く	届く	届く				
範囲で	範囲で	範囲で	範囲で	範囲で	範囲で	範囲で				
で	で	で	で	で	で	で				
と	と	と	と	と	と	と				
すると	すると	すると	すると	すると	すると	すると				
て	て	て	て	て	て	て	whole	pragmatics	adverb	imitative sound
進げ	進げ	進げ	進げ	進げ	進げ	進げ				
ちゃう	ちゃう	ちゃう	ちゃう	ちゃう	ちゃう	ちゃう	middle	morphology	auxiliary verb	colloquial
んです	んです	んです	んです	んです	んです	んです				
凄く	凄く	凄く	凄く	凄く	凄く	凄い	frequent	morphology	adjective(adverb)	generation limited
う	う	う	う	う	う	う	whole	pragmatics		speak to sb./mutter
(F ん)	(F ん)	(F ん)	(F ん)	(F ん)	(F ん)	(F ん)				
魚好さ	魚好さ	魚好さ	魚好さ	魚好さ	魚好さ	魚好さ	middle	semantics	adjective verb	plus
には	には	には	には	には	には	には				
たまら	たまら	たまら	たまら	たまら	たまら	たまら				
ない	ない	ない	ない	ない	ない	ない				
んー)	んー)	んー)	んー)	んー)	んー)	んー)				
黄色や	黄色や	黄色や	黄色や	黄色や	黄色や	黄色や		syntax		childishness
青や	青や	青や	青や	青や	青や	青や		syntax		childishness
赤とか	赤とか	赤とか	赤とか	赤とか	赤とか	赤とか		syntax		childishness
こう	こう	こう	こう	こう	こう	こう	frequent	morphology	adverb	sth.
お魚さ	お魚さ	お魚さ	お魚さ	お魚さ	お魚さ	お魚さ	middle	morphology	noun	childishness
んじゃ	んじゃ	んじゃ	んじゃ	んじゃ	んじゃ	んじゃ				
見れな	見れな	見れな	見れな	見れな	見れな	見れな	whole	morphology	verb+auxiliary verb	generation limited
い	い	い	い	い	い	い				

Figure 6.2: A sample of pattern extraction of 'cuteness' from one transcription

regular, but it is acceptable. The parts in italics in the pattern-loaded texts indicate the patterns that should form the impression of Cuteness. We chose some places of interest from guides of Yokohama city as the origin texts; moreover, these origin texts were used as TTS scripts for car navigation systems. The guides contained approximately 11 short texts describing each

place of interest in Yokohoma city. Most of the text is approximately 3–5 sentences long, and consists of approximately 150 words in Japanese. A neutral text is defined as one in which ideally, no particular type of impression will be formed (especially any of the three types of impressions mentioned in the previous section). We transformed all the origin texts into neutral text by changing words, phrases, or sentences in such a way as to form impressions into neutral. In addition, we used pattern-loaded texts in the experiment. Based on the neutral texts, we loaded the most possible number of extracted and categorised patterns. For example, despite the cultural and linguistic differences between Japanese and English, two English sentences can be discussed:

I. You are so thoughtful.

II. You are so sweet.

As can be determined, sentence I and sentence II are the same in meaning. However, it appears that by changing words, sentence II is more expressive than sentence I.

We conducted this type of pattern-loading work to prepare for the text material. Because there were three types of impressions to be verified, we created 33 (3 × 11 texts) pattern-loaded texts in total. Neutral texts were multi-loaded with the patterns we extracted in this experiment. Because the goal of this thesis is simply to create a preliminary proposal first, we intentionally ignored the relationship among the extracted patterns. To test

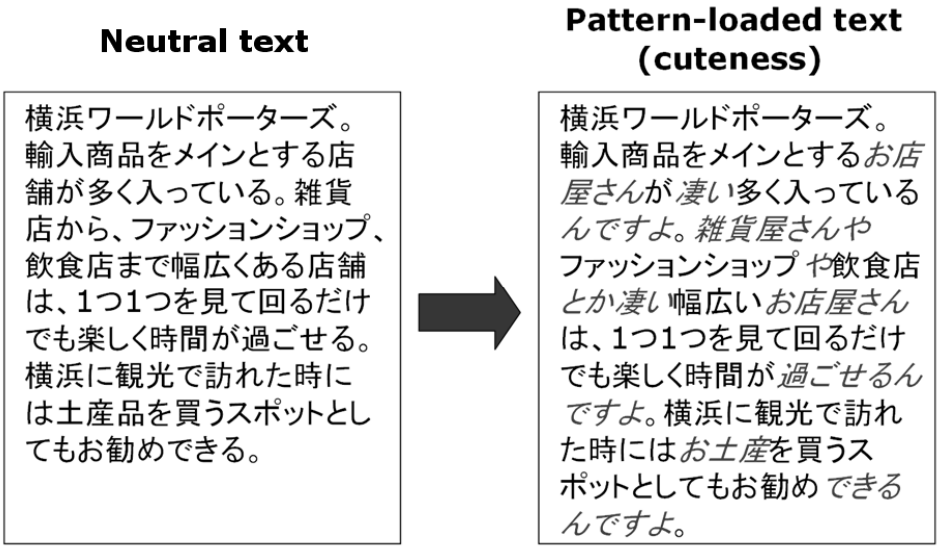


Figure 6.3: A sample of neutral text and pattern-loaded text
(With translation below)

Translation:

Yokohama World Porters.

There are many shops of imports like variety stores, fashion shops and restaurants. You will have a good time to look around. Getting souvenirs and gifts of Yokohama here is strongly recommended.

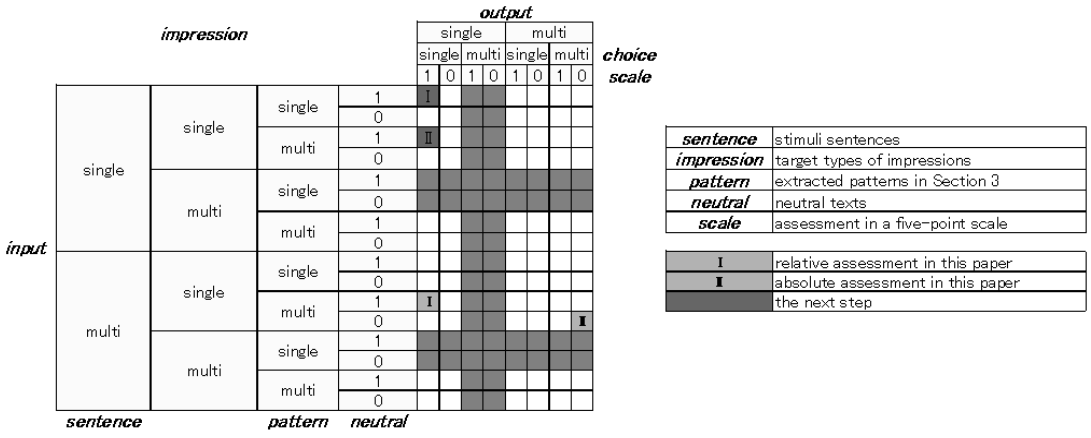


Figure 6.4: Parameters control in the assessment experiment

and quantify the impression formation effect of every pattern, the parameters in this experiment were reconsidered. We discuss the parameters in detail in Figure 6.4. Parts I and II in deep grey represent the relative and absolute assessment in this thesis, respectively. For the next step of this study and to quantify the contribution ratio of every pattern, the light grey part I might be appropriate. In addition, through the light grey part II, we might be able to clarify the interaction and the synergistic effect among these patterns. Moreover, the remaining combinations might be helpful for other goals.

We divided the 11 participants into two groups and presented the texts in different order. The participants were asked to read the present text and to answer the questionnaires afterward. To obtain both absolute and relative assessments, the experiment was divided into two parts: first, the absolute assessment was conducted; second, the relative assessment was performed. In the first part, we presented the participants with nine (three types of impressions × three different texts) pattern-loaded texts and set three target types of impressions: seriousness, cuteness, and expressiveness with another five non-target types of impressions (urbanity, activeness, couth, likes, and healing) for multiple choice. We obtained the absolute assessment from the results of the first part. In the second part, we presented the participants with both the neutral texts and the pattern-loaded texts (three types of impressions × three different texts) for comparison. The questions used a five-point scale (for a particular type of impression, three means that the neutral text and the pattern-loaded text generate the same impression,

one means that the neutral text impresses more than the pattern-loaded text, and five means the opposite). We obtained the relative assessment from the results of the second part. Based on the relative assessment, we do not know whether stimuli from the text exceeded the participant's threshold of the expected type of impression. At a minimum, we know that, compared to the neutral text, the pattern-loaded text did generate the expected type of impression, which also means that the patterns loaded in the text are effective in the impression formation.

We devised a method for text presentation for the following reason: the main purpose of this thesis is to testify and clarify the TEXT effect for impression formation; therefore, we attempted to avoid possible disturbances from the TTS voice and focused on text only. To construct a circumstance in which human beings receive and respond to audio stimuli as expected, first, we displayed text word for word instead of simultaneously, which is similar to playing audio stimuli. The display speed was fully discussed and properly adjusted. Second, when the display was complete, the entire text disappeared so that rereading was prevented.

Through the absolute assessment and the relative assessment, we obtained two satisfactory results. Table 6.3 shows the results of the absolute assessment. The resulting numbers are arranged in a confusion matrix. The overlapping parts indicate the number of times that the participants (11 participants × three texts for each impression in total) chose the expected type of impression. The other numbers represent the remaining five types of impressions mentioned in the previous paragraphs (Table 6.3). Table 6.4

Table 6.3: Confusion Matrix of the results in the absolute assessment

		Predicted		
		Expressiveness	Cuteness	Seriousness
Actual	Expressiveness	17	5	1
	Cuteness	3	16	0
	Seriousness	5	8	29
	others	16	9	7

Table 6.4: Precision, Recall and F-value of the absolute assessment

	Expressiveness	Cuteness	Seriousness
Precision	0.52	0.48	0.88
Recall	0.74	0.84	0.69
F-value	0.61	0.62	0.77

shows the precision, recall, and F-value of the absolute assessment. Considering that text-based linguistic pattern control simply contributes partly to the characterization control, and given the interference from content and the variation of the threshold of human participants, the results of the absolute assessment are satisfactory.

Figure 6.5 shows the results of the relative assessment. In Figure 6.5, the numbers indicate the average ratings of all 11 participants; the nine bars represent the nine sample texts used in the relative assessment. It is clear that for 'Expressiveness' and 'Cuteness', the results are extremely satisfactory. The results for 'Seriousness' are not as satisfactory as those for the other two types of impressions, but only one result is below three (neutral). Perhaps the neutral texts used for 'Seriousness' should be discussed more. Overall, we had extremely satisfactory results in the

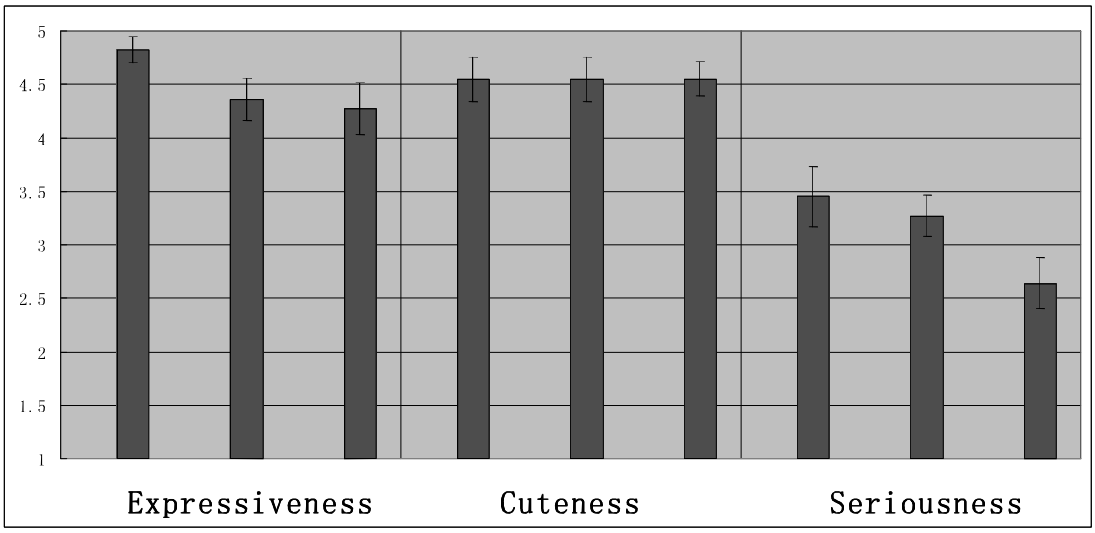


Figure 6.5: Results of the relative assessment (each bar represents one single text).

relative assessment.

6.1.3.4 Summary

From speech transcriptions in CSJ, we extracted 43 morphological patterns of linguistic patterns that are considered effective for perceiving speaking styles. In this thesis, we cross-verified the 43 patterns with 60 text stimuli and as a result, 23 matched patterns are used as features to construct the estimation model of the speaking style (Table 6.5).

6.2 Construction

In this study, we chose to use multiple-regression analysis to construct the estimation model. The main reason is that, considering the application of this estimation model, users are supposed to grasp those linguistic patterns that contribute to speaking style perception more easily. Moreover,

Table 6.5: Morphological Patterns
(The notation of morphological patterns: occurrence[lemma](POS))
(“.” is the wildcard and “A|B” means either A or B.)

No.	Morphological Pattern	Sample
1	to[to](par).[iu](v)	toiu
2	to[to](par)shi[suru](v)te[te](par)	toshite
3	yo[yo](par)ne[ne](par)	yone
4	.[keredo](par)mo[mo](par)	keredomo
5	to[to](par)ka[ka](par)	toka
6	.[teru](aux)te[te](par)	ittete
7	tte[tte](par).[iu](v)	tteiu
8	nanka[nanka](par)	nanka
9	ne[ne](par)	ne
10	yo[yo](par)	yo
11	.[zenzen](adv)	zenzen
12	.[yahari](adv)	yappari
13	.[sugoi](adj)	sugoku
14	.[kekkou](adv)	kekkou
15	.[desu](aux) | .[masu](aux)	desu
16	de[de](conj)	de
17	.[konna](adno) | .[sonna](adno) | .[anna](adno)	konna
18	.[chau](aux)	icchau
19	.[toku](aux)	ittoku
20	kocchi[kocchi](pron) | socchi[socchi](pron) | acchi[acchi](pron)	kocchi
21	jya[de](par)	jya
22	politeness	gozaimasu
23	onomatopoeia	hyutto

the estimated partial regression coefficient also shows the degree of contribution to speaking style perception so that it might be handier for users to modify their intended speaking style.

With the features mentioned in the previous section as explanatory variables (34 in total), and the average rating results from the rating

Table 6.6: Coefficient of determination (R^2 /adjusted R^2)

Features Set	Intelligibility-oriented(I)	Familiarity(F)	soCial strata(C)	Remarks
All features	0.76/0.67	0.93/0.91	0.84/0.79	closed
	0.54/0.37	**0.85**/0.81	**0.73**/0.66	leave-one-out
	0.36/0.13	0.80/0.74	0.62/0.52	leave-one-out(exclusive)
POS only	0.24/0.13	0.32/0.24	0.36/0.31	leave-one-out
	0.14/0.02	-0.08/-0.21	0.23/0.17	leave-one-out(exclusive)
Morph Patterns only	0.36/0.23	0.76/0.67	0.55/0.45	leave-one-out
	0.29/0.14	0.62/0.49	0.44/0.32	leave-one-out(exclusive)

Table 6.7: Details of the estimation model of speaking style
(Signif. codes: 0 "***", 0.001 "**", 0.01 "*", 0.05 ",", 0.1 " ")
(The notation of morphological patterns: occurrence[lemma](POS))
(". " is the wildcard and "A|B" means either A or B.)

	Intelligibility-oriented(I)		Familiarity(F)		soCial strata(C)	
	explanatory variable	Estimate	explanatory variable	Estimate	explanatory variable	Estimate
1	.[keredo](par)mo[mo](par)**	35.10	.[keredo](par)mo[mo](par),	-25.09	.[keredo](par)mo[mo](par)*	34.06
2	yo[yo](par)ne[ne](par),	19.01	.[desu](aux)| .[masu](aux)***	-19.92	yo[yo](par)***	-21.21
3	tte[tte](par).[iu](v),	-15.16	yo[yo](par)***	19.79	.[chau](aux),	-13.86
4	to[to](par).[iu](v)	12.79	.[chau](aux)*	17.66	.[desu](aux)| .[masu](aux)***	11.34
5	jya[de](par)**	12.33	.[kekkou](adv),	-16.13	adno**	-11.31
6	adno**	-11.30	ne[ne](par)***	14.19	to[to](par)ka[ka](par)*	-9.44
7	.[kekkou](adv)	-11.16	pref,	12.26	.[teru](aux)te[te](par)*	-7.87
8	yo[yo](par),	-9.21	adno*	10.63	aux***	-7.04
9	.[desu](aux) | .[masu](aux)***	7.65	to[to](par)ka[ka](par)*	9.06	ne[ne](par)*	-6.1
10	ne[ne](par)*	-7.44	aux***	6.69	func***	4.89
11	.[teru](aux)te[te](par),	-6.02	func***	-5.79	adj**	-2.84
12	aux***	-4.63	to[to](par).[iu](v)	-5.46	int	-1.51
13	func***	3.67	pron,	4.54	(Intercept)*	-1.78
14	adv,	3.13	adj*	2.69		
15	v,	2.24	(Intercept)***	2.72		
16	par	1.09				
17	(Intercept)***	-3.09				

experiment as the objective variable, we conducted a multi-regression analysis (stepwise method) to construct the estimation model. There are three sub-models that represent each of the three scales of speaking style. We also conducted cross-validation (leave-one-out) to verify the reliability of the training data. The coefficient of determination (R^2) (D.C. Montgomery 1982) is listed in Table 6.6, and the model details are listed in Table 6.7. In

Table 6.6, we also show R^2 by training the estimation model with different feature sets, such as using all features, using part-of-speech only, and using morphological patterns only. As a result, the R^2 of using all the features is the highest of all ('All features'-'leave-one-out', I: 0.54, F: 0.85, C: 0.73), which proves the effectiveness of our method. Moreover, to observe the possibility of feature unbalance among the six categories, we withheld the text stimuli from the same categories in cross-validation (leave-one-out). As a result, the R^2 are 0.36 (I), 0.80 (F), and 0.62 (C) ('All features'-'leave-one-out (exclusive)'). With the exception of I, the results of F and C are extremely satisfactory. In Table 6.7, the contributive explanatory variables of each sub-model (three scales) are listed in descending order (absolute value). All the sub-models are statistically significant at a 0.01 level.

We also compared the models using different feature sets. Details are listed in Table 6.6. Obviously, using both of the features of part-of-speech (word classification) and morphological patterns, the coefficient of determination, which also indicates the accuracy of the speaking style estimation, is higher than using only one type of feature.

More specifically, with regard to efficient explanatory variables such as '.[desu](aux)|.[masu](aux)', which work in the F model, the closer the speaker and the listener are, the less 'desu/masu' is used in Japanese, as expected. '.[keredo](par)mo[mo](par)' is the most effective variable in the C model, which indicates that the more the variable is used, the upper the social strata appears to be.

The results described in this section prove that our proposal for automatic

estimation of speaking style is effective, and by adapting the estimation model to any speech transcription, the speaking style in speech corpora can be estimated using three scales.

Chapter 7: CONCLUSIONS

The goal of this thesis is to answer the following two questions.

1. To what extend can we attribute speaking style to speech corpora?

2. With the speaking style attributed to speech corpora, how can we consider the issue of speaking style in the field of foreign language education?

One of the possible methods is to construct a CALL system of recommending speech corpora to learners, teachers, and researchers with speaking style visualization; this appears to be a proper choice for the goal of this thesis.

To realize the goal, three steps can be considered.

1. Automatic estimation of speaking style.
2. Visualization of speaking style in speech corpora.
3. Recommendation of speech corpora.

In this thesis, we discussed mainly the first step, which is to estimate speaking style from the content of speech corpora so that the attribute of speaking style can be determined. Because of the large amount of speech

data to be estimated, we needed to determine how to estimate speaking style automatically.

Then, we attempted to construct an automatic estimation model. Given the satisfactory research results, we focused on speech transcriptions and used part-of-speech, word classification, and morphological patterns, which are indicated to be effective in the field of natural language processing, to construct the estimation model of speaking style by referring to the three scales of speaking style proposed by Eskenazi. We constructed the estimation model by Multi-regression Analysis. The coefficients of determination of the three scales were 0.54, 0.85, and 0.73 respectively. The results of Familiarity (F) and soCial strata (C) were satisfactory and indicated the effectiveness of our method. However, the result of the Intelligibility-oriented (I) scale was the lowest of the three scales. We consider that this might be because of a lack of effective features of linguistic factors in speech transcriptions for the Intelligibility-oriented (I) scale. Therefore, for future work, other features will be discussed to improve the model.

There are still two remaining steps in our study.

To utilize the estimation model of speaking style in the field of foreign language education,

1. It is necessary to visualize the speaking style of both any given transcriptions by users, which are the learners, teachers, and researchers in the field of foreign language education here, and the prepared speech

corpora.

2. Based on the positional relationships of the visualized speaking style among the given transcriptions by users and the prepared speech corpora, it is necessary to make a proper recommendation based on those prepared speech corpora.

With this system, as for learners, they are supposed to confirm and correct their intended speaking style by inputting the contents (which mean the speech transcriptions here). As for teachers, they are supposed to be able to reach speech resources with similar speaking style as teaching materials on conversational lessons or for the purpose of textbook compilation. As for researchers, process of perception and acquisition of speaking style might be observed by investigating user logs.

Undoubtedly, if the CALL system incorporating the proposed speaking style estimation model is completed, learners, teachers, and even researchers will benefit from the utilization of the system.

ACKNOWLEDGEMENT

I want to express my deepest gratitude to my supervisor Prof. H. KIKUCHI, who has been guiding my study for more than seven years. Without his supportive advice and proper leading, I would never have been able to finish this dissertation.

S. ITAHASHI, emeritus professor of University of Tsukuba, led a great project that helped me finish a part of my dissertation. I am deeply grateful to Prof. S. ITAHASHI, Prof. T. MORIMORO, and Prof. T. MATSUI for providing comments, suggestions, and encouragement, which are incalculably valuable throughout the writing of this dissertation.

Mr. K. OHTA and Mr. T. MITAMURA from Mobility Services Laboratory, Nissan Research Centre, Nissan Motor Co. Ltd, led a productive collaborative research that helped me finish a part of my study. Without their guidance, I could not have been able to manage such a large-scale experiment.

I also want to thank the members of the H. KIKUCHI laboratory, who are always supportive and cheered me up during my study life.

Finally, yet importantly, I would like to thank my parents for the unspoken love and kindness to me through the good times and bad times. I also want to thank Leo (a 3-year-old boy poodle who is one of my family members), whose innocent face encouraged me from time to time.

REFERENCES

M. Abe and H. Mizuno, "Speaking style Conversion by Changing Prosodic Parameters and Formant Frequencies." ICSLP 94, Yokohama, Japan, pp.1455-1458, 1994.

D. Biber, Variation across Speech and Writing. Cambridge University Press, 1988.

T. Cho, "Prosodically Conditioned Strengthening and Vowel-to-vowel Coarticulation in English." Journal of Phonetics, Vol.32, pp.141-176, 2004.

M. Cid and S.G. Fernandez Corugedo, "The construction of a corpus of spoken Spanish: Phonetic and phonological parameters." Proc. of the ESCA Workshop, Barcelona, Catalonia, Spain, pp. 17-1 - 17-5, 1991.

M.R. Delgado and M.J. Freitas, "Temporal structures of speech: Reading news on TV." Proc. of the ESCA Workshop, Barcelona, Catalonia, Spain, pp. 19-1-19-5, 1991.

ELRA- [European Language Resources Association]
http://www.elra.info/

M. Eskenazi, "Changing speech styles, speakers' strategies in read speech and careful and casual spontaneous speech." Proc. of the International Conference on Spoken Language Processing, Banff, 1992.

M. Eskenazi, “Trends in Speaking style Research.” Keynote speech, Proc. of Eurospeech’93, Berlin, 1993.

M. Gregory and S. Carroll, “Language and Situation.” Language varieties and their social contexts, London: Routledge & Kegan Paul, 1978.

K. Hirose, K. Sato and N. Minematsu, “Corpus-based Synthesis of Fundamental Frequency Contours With Various Speaking Styles From Text using F0 Contour Generation Process Model.” 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA, USA, pp.161-166, 2004.

J. Hirschberg, “A Corpus-Based Approach to the Study of Speaking style.” in Festschrift in Honor of Gosta Bruce, ed. M. Horne, Kluwer, 2000.

Y. Horiuchi, Y. Nakano, H. Koiso, M. Ishizaki, H. Suzuki, M. Okada, M. Naka, S. Tutiya and A. Ichikawa, “The Design and Statistical Characterization of the Japanese Map Task Dialogue Corpus [in Japanese].” Journal of Japanese Society for Artificial Intelligence, Vol.14-2, pp.261-272, 1999.

Y. Ishimoto, S. Itahashi, K. Yamakawa, R. Shen, H. Kikuchi and T. Matsui, “Improvement of the visualization system of multiple speech corpora.” J. Acoust. Soc. Jpn, 2011.

S. Itahashi, K. Yamakawa, T. Matsui and Y. Ishimoto, “A proposal for standardizing catalogue specifications of speech corpora.” Proc. of Oriental COCOSDA Workshop 2010, 2010.

Y. Iwano, Y. Sugita, M. Matsunaga and K. Shirai, "Difference in Face-to-face and Telephone Dialogues: Analysis of the Role of Head Movement [in Japanese]." Information Processing Society of Japan, SIG Notes, Vol. 15-29, pp.105-112, 1997.

M. Joos, "The isolation of styles." In FISHMAN, J.A. (ED) Readings in the Sociology of Language, The Hague: Mouton, pp.185-191, 1968.

E. Jorden and M. Noda, Japanese the Spoken Language. New Haven & London: Yale University Press, 1987.

G.G. Koch, "Intraclass Correlation Coefficient." In Samuel Kotz and Norman L. Johnson. Encyclopedia of Statistical Sciences, 4, pp. 213-217, 1982.

H. Kikuchi, R. Shen, K. Yamakawa, S. Itahashi and T. Matsui, "Construction of the visualization system of multiple speech corpora." J. Acoust. Soc. Jpn, 3-P-33, 441-442, 2009.

H. Koiso, T. Ogiso and S. Miyauchi, "A Corpus-based Stylistic Comparison on Various Genres: Focusing on Short-Unit Word." (in Japanese, the title translated by the author), Proc. of the 15th Annual conference of the association for Natural Language Processing, Vol.15, pp.594-597, 2009.

S. Kori, "What Are the Major Speech Styles in Japanese?" Journal of Phonetic Society of Japan, Vol.10, No.3, pp.52-68, 2006.

T. Koyama and K. Takeuchi, “An Evaluation of Document Set Similarity Based on Morpheme occurrence Patterns.” IPSJ SIG Technical Report, NL-188 (8), pp.51-56, 2008.

H. Kruschke, “Simulation of Speaking styles with Adapted Prosody.” TSD2001, pp.278-284, 2001.

W. Labov, “The isolation of contextual styles.” In Sociolinguistic Patterns, Oxford: Basil Blackwell, 1978, pp.70-109, 1972.

J.R. Landis and G.G. Koch, “The Measurement of Observer Agreement for Categorical Data.” Biometrics, 33, pp. 159-174, 1977.

LDC- [Linguistic Data Consortium]
http://www.ldc.upenn.edu/

J. LLISTERRI, “Speaking style in speech research.” ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language, Dublin, Ireland, pp.15-17, 1992.

E.E. Lyakso, O.V. Frolova, A.V. Kurazhova and J.S. Gaikova, “Russian Infants and Children’s Sounds and Speech Corpuses for Language Acquisition Studies.” Proc. of INTERSPEECH 2010, pp. 1878-1881, 2010.

K. Maekawa, T. Kagomiya, H. Koiso, H. Ogura and H. Kikuchi, “Design of the Corpus of Spontaneous Japanese (<Feature Articles>Trends in Database for Phonetic Research) [in Japanese].” Journal of the Phonetic Society of Japan, Vol.4-2, pp.51-61, 2000.

F. Mairesse, M.A. Walker, M.R. Mehl and R.K. Moore, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text." Journal of Artificial Intelligence Research, 30, pp. 457-500, 2007.

MeCab, "MeCab: Yet Another Part-of-Speech and Morphological Analyzer." http://mecab.sourceforge.net/

K. Miyazawa, T. Kagetani, R. Shen, H. Kikuchi, Y. Ogawa, C. Hata, K. Ohta, H. Hozumi, and T. Mitamura, "Examination on mechanism which a driver accepts a robot navigator's suggestion in driving environment", Proc. Of Human-Agent Interaction Symposium 2009, 1A-5, 2009.

D.C. Montgomery and E.A. Peck, Introduction to Linear Regression Analysis. 2nd edition, John Wiley & Sons, New York, 1982.

S. Nakazato, Y. Oshiro and H. Kikuchi, "The Relationship between the Level of Intimacy and Manner of Speech" IEICE Technical Report, Vol.112, No.483, pp.109-114, 2013.

G. Neubig, S. Mori, and T. Kawahara, "A WFST-based Log-linear Framework for Speaking-style Transformation." Proc. of INTERSPEECH 2009, Brighton, UK, pp.1495-1498, 2009.

NII-SRC- [National Institute of Informatics– Speech Resources Consortium] http://research.nii.ac.jp/src/eng/

H. Noyama and K. Imamura, "The Formal Features on Speech Styles of Japanese Learners: Based on a Longitudinal Research on Resident Foreigners and the OPI Data in the KY Corpus." (in Japanese, the title translated by the author), Proc. of ICPLJ8, pp.86-89, 2014.

R. Shen and K. Hideaki, “Construction of the Speech Corpus Retrieval System: Corpus Search & Catalog-Search.” Proc. of Oriental-COCOSDA 2011, pp.76-80, 2011.

R. Shen and H. Kikuchi, “Ratings of Speaking Style in Speech Corpora - Focus on Transcriptions.” Proc. of Oriental-COCOSDA 2012, pp.274-278, 2012.

R. Shen, H. Kikuchi, K. OHTA and T. MITAMURA, “Towards the Text-level Characterization Based on Speech Generation.” Journal of Information Society of Japan, Vol.53, No.4, pp.1269-1276, 2012.

Q. Sun, S. Hiki and K. Sunaoka, “A Computer-assisted Instruction System for Self-teaching of Discriminating Chinese Four Tones.” The Journal of Modernization of Chinese Language Education, Vol.1, No.1, pp.52-59, 2012.

M. Teshigawara and S. Kinsui, “Modern Japanese ‘Role Language’ (Yakuwarigo): fictionalized orality in Japanese literature and popular culture.” Sociolinguistic Studies, Vol.5, No.1, pp.37-58, 2011.

A.M. Uhlmann, “Meyers Neues Lexikon.” VEB Bibliograhisches Institut Leipzig, ausgabe in acht bänden edition, 1964.

UniDic

http://www.tokuteicorpus.jp/dist/

J. Van de Weijer, “Language Input to a Prelingual Infant.” In A. Sorace, C. Heycock, & R. Shillcock (Eds.), Proc. of the GALA '97 conference on language acquisition, Edinburgh University Press, pp. 290-293, 1997.

K. Yamakawa, T. Matsui, and S. Itahashi, “Visualization of Various Speech Corpora by Multidimensional Scaling.” Proc. of Oriental COCOSDA Workshop 2007, Hanoi, Vietnam, pp.38-43, 2007.

K. Yamakawa, T. Matsui and S. Itahashi, “MDS-based Visualization Method for Multiple Speech Corpora.” Proc. of Interspeech 2008, Brisbane, Australia, 2008.

K. Yamakawa, H. Kikuchi, T. Matsui and S. Itahashi, “Utilization of Acoustical Feature in Visualization of Multiple Speech Corpora.” Proc. of Oriental COCOSDA 2009, Beijing, China, pp. 147-151, 2009.

K. Yamasumi, T. Kagomiya, Y. Maki, and K. Maekawa, “Psychological Scale for the Impression Rating of Monologue.” Journal of Acoustical Society of Japan, Vol.62, No.5, pp.303-311, 2005 (In Japanese).

ACHIEVEMENT

Dissertation

1. Developing Feasibility of the Chinese Learners' Speech Corpus Referring to the Corpus of Spontaneous Japanese (CSJ) 2008 Dissertation of Master's degree, Faculty of Human Sciences, Waseda University.

Journal

1. 宮澤幸希，影谷卓也，沈睿，菊池英明，小川義人，端千尋，太田克己，保泉秀明，三田村健：2010　自動車運転環境下におけるユーザーの受諾行動を促すシステム提案の検討．人工知能学会誌，25 巻 6 号，723-732 頁．
2. 沈睿，菊池英明，太田克己，三田村健：2012　音声生成を前提としたテキストレベルでのキャラクタ付与．情報処理学会論文誌「インタラクションの理解および基盤・応用技術」，53 巻 4 号，1269-1276 頁．
3. Raymond SHEN, Kazuko SUNAOKA: 2012 Construction and Application of a Text Corpus of Newspaper Articles about Disasters in Chinese Education. The Journal of Modernization of Chinese Language Education, Vol.1, No.2, pp.43-52.
4. 沈睿，菊池英明：2014　音声言語コーパスにおける speaking style の自動推定—転記テキストに着目して一．言語処理学会論文誌，21 巻 3 号．(印刷中)

Else

1. Raymond SHEN, KIKUCHI Hideaki: 2008 Developing feasibility of the Chinese Learners' Speech Corpus referring to the Corpus of Spontaneous Japanese (CSJ). In Proceedings of Oriental-COCOSDA 2008, pp.93-96.
2. 菊池英明，沈睿，山川仁子，松井知子，板橋秀一：2009　音声言語コーパスの類似性可視化システムの構築．日本音響学会秋季研究発表会講演論文集，441-442 頁．
3. Raymond SHEN, KIKUCHI Hideaki: 2009 Development and Construction of the XML Browser for the Chinese Learners' Speech Corpus. In Proceedings of Oriental-COCOSDA 2009, pp.199-202.

4. 宮澤幸希，影谷卓也，沈睿，菊池英明，小川義人，端千尋，太田克己，保泉秀明，三田村健：2009　自動車運転環境においてロボットナビゲーターの提案をドライバーが受諾するメカニズムの検討．HAI シンポジウム 2009, 1A-5.
5. Raymond SHEN, KIKUCHI Hideaki, OHTA Katsumi, MITAMURA Takeshi: 2010 Feasibility of the Characterisation Control by Text-based Speech Style. Oriental-COCOSDA 2010, No.18.
6. 沈睿，菊池英明：2011　音声言語コーパス目録検索システム Catalog-Search の構築および応用の検討．言語処理学会年次大会予稿集，721-724 頁．
7. 石本祐一，板橋秀一，山川仁子，沈睿，菊池英明，松井知子：2011　音声コーパスの類似性可視化システムの改良．日本音響学会春季研究発表会講演論文集，435-436 頁．
8. Raymond SHEN, Hideaki Kikuchi, Katsumi Ohta, Takeshi Mitamura: 2011 Towards the Characterization Control in Personalized Vehicles by Text-based Style Change. In Proceedings of NCMMSC2011, pp.14.
9. Raymond SHEN, KIKUCHI Hideaki: 2011 Construction of the Speech Corpus Retrieval System: Corpus Search & Catalog-Search. In Proceedings of Oriental-COCOSDA 2011, pp.76-80.
10. 砂岡和子，Raymond SHEN：2012　災害報道文の特徴語抽出．言語処理学会年次大会，A4-4.
11. Raymond SHEN, Kazuko SUNAOKA: 2012 Construction and Analysis of a Disasters News Reports Corpus. THE 8TH INTERNATIONAL CONFERENCE ON NEW TECHNOLOGIES IN TEACHING AND LEARNING CHINESE.
12. Raymond SHEN, Hideaki KIKUCHI: 2012 Ratings of Speaking Style in Speech Corpora - Focus on Transcriptions. In Proceedings Oriental-COCOSDA 2012, pp.274-278.
13. 菊池英明，宮島崇浩，沈睿：2013　多様な音声表現コーパスにおける句末音調のクラスタリング．第 3 回コーパス日本語学ワークショップ予稿集，23-28 頁．
14. 沈睿，菊池英明：2013　音声言語コーパスにおける speaking style の評定と分布—転記情報に着目して—．第 3 回コーパス日本語学ワークショップ予稿集，359-362 頁．
15. 沈睿，菊池英明：2013　音声言語コーパスにおける speaking style の多様性—転記テキストに着目して—．言語処理学会年次大会予稿集，620-623 頁．
16. Raymond SHEN, Hideaki KIKUCHI: 2014 Estimation of Speaking Style in Speech Corpora: Focusing on speech transcriptions. The 9th edition of the Language Resources and Evaluation Conference, O27-616.