Proceedings of the LREC 2018 Special Speech Sessions

LREC 2018 Special Speech Sessions

Speech Resources Collection in Real-World Situations

PROCEEDINGS

Edited by

Yuichi Ishimoto and Kikuo Maekawa

9 May 2018

Proceedings of the LREC 2018 Special Speech Sessions
“Speech Resources Collection in Real-World Situations”

9 May 2018 — Miyazaki, Japan

Edited by Yuichi Ishimoto and Kikuo Maekawa

Acknowledgments: This session is supported by the Center for Corpus Development of the National Institute for Japanese Language and Linguistics (NINJAL).

NINJAL
National Institute for Japanese Language and Linguistics

Programme

	Session S1
14.35 – 14.55	Yuichi Ishimoto, Tomoko Ohsuga Spontaneous Speech Resources in Japan
14.55 – 15.15	Yoshiko Arimoto Challenges of Building an Authentic Emotional Speech Corpus of Spontaneous Japanese Dialog
15.15 – 15.35	Norihide Kitaoka, Yurie Iribe, Hiromitsu Nishizaki Construction of a Corpus of Elderly Japanese Speech for Analysis and Recognition
15.35 – 15.55	Nobuko Kibe, Tomoyo Otsuki, Kumiko Sato Intonational Variations at the End of Interrogative Sentences in Japanese Dialects: From the "Corpus of Japanese Dialects"
15.55 – 16.15	Hanae Koiso, Yasuharu Den, Yuriko Iseki, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda Construction of the Corpus of Everyday Japanese Conversation: An Interim Report
	Session S2
16.35 – 16.55	Mayumi Bono, Rui Sakaida, Ryosaku Makino, Ayami Joh Miraikan SC Corpus: A Trial for Data Collection in a Semi-open and Semi-Controlled Environment
16.55 – 17.15	Kotaro Funakoshi A Multimodal Multiparty Human-Robot Dialogue Corpus for Real World Interaction
17.15 – 17.35	Ryuichiro Higashinaka, Ryo Ishii, Narimune Matsumura, Tadashi Nunobiki, Atsushi Itoh, Ryuichi Inagawa, Junji Tomita Speech and Language Resources for the Development of Dialogue Systems and Problems Arising from their Deployment
17.35 – 17.55	General Discussion

Table of Contents

Spontaneous Speech Resources in Japan

Yuichi Ishimoto[†], **Tomoko Ohsuga**[‡]

[†]National Institute for Japanese Language and Linguistics
10–2 Midori-cho, Tachikawa, Tokyo 190–8561, Japan
yishi@ninjal.ac.jp

[‡]National Institute of Informatics
2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo 101–8430, Japan
osuga@nii.ac.jp

Abstract

In this paper, we introduce representative corpora of spontaneous speech, which have been provided publically in Japan. A large amount of spontaneous speech data is required for research on various themes in speech studies such as speech analysis, speech recognition systems, and natural language processing in recent years. However, it is difficult to collect spontaneous speech data, and few corpora of spontaneous speech are available. Considering the diversity of speech in real-world situations, the data remain insufficient. We show the characteristics of spontaneous Japanese speech corpora gathered and distributed by two organizations: the Speech Resources Consortium at the National Institute of Informatics, and the National Institute for Japanese Language and Linguistics. Then, we describe prospects for the development of spontaneous speech resources.

Keywords: Japanese corpus, spontaneous speech, natural conversation, corpus distribution

1. Introduction

Speech resources are necessary to promote speech research; therefore various speech corpora have been compiled. Initially, most of the corpora consisted of words and sentences read aloud such as numbers, greetings, place names, and phonetically balanced phrases because in the past, some providers usually collected them for use in constructing early speech recognition systems. Although prior data used to be effective, it is no longer sufficient for systems to show high performance in real-world situations.

Read-aloud speeches have different characteristics from those of the words and sentences that we utter in everyday conversations; consequently, the old system derived from speech data did not exhibit competent performance for real-life situations. Moreover, spontaneous utterances are more complex and have more disfluency than sentences prepared in advance.

Spontaneous speech data have thus been required by researchers; however, it takes much more time to record spontaneous speech than read-aloud speech. The recorder needs to prepare an environment in which the speaker makes spontaneous utterances, or to visit a place in which natural conversations occur. Few corpora of spontaneous Japanese speeches exist.

In this paper, we introduce several spontaneous Japanese speech corpora that are publically distributed and describe their characteristics. Then, we describe prospects for the development of spontaneous resources.

2. Spontaneous Japanese Speech Corpora

In this section, we introduce representative speech resources of spontaneous Japanese gathered and distributed by two organizations in Japan.

2.1. Corpora from NII-SRC

The Speech Resources Consortium at the National Institute of Informatics (NII-SRC) was established in 2006. It aims to collect speech resources from researchers who belong to universities, as well as companies that record speech sounds for various purposes, and to distribute them to researchers who need speech data suitable for their investigations. Although most researchers record speech for their purposes only and utilize it, they do not have the means or knowledge to distribute their data.
NII-SRC has distributed 43 corpora as of May 2018; Table 1 shows a list of them. As described in the introduction, corpora distributed earlier consist of read-aloud speeches, mainly because the providers aimed to apply words and sentences uttered fluently to fundamental research on speech. Subsequently, spontaneous speech was collected to apply speech information processing in an actual environment. Most of the earlier corpora for spontaneous speech were composed of role-play in different situations (such as navigation and shopping) because a question-and-response format was preferred for human-computer dialogue systems based on speech recognition technology. For example, RWCP-SP96 and RWCP-SP97 — the formal names of which are “RWCP Spoken Dialogue Corpus, 1996 edition” and “1997 edition” — contain face-to-face di-

Name	Launched	Contents	Style	Situation	Note
PASL-DSR	2006	Words, Sentences	Read-aloud	—	
UT-ML	2006	Words, Sentences	Read-aloud	—	
TMW	2006	Words	Read-aloud	—	
GSR-JD	2006	Words, Dialogue	Read-aloud, **Spontaneous**	**Natural**	Dialect
RWCP-SP96	2006	Dialogue	**Spontaneous**	Role-play	
RWCP-SP97	2006	Dialogue	**Spontaneous**	Role-play	
RWCP-SP99	2006	Monologue	Read-aloud	—	
RWCP-SP01	2006	Dialogue	**Spontaneous**	Role-play	
PASD	2006	Dialogue	**Spontaneous**	Role-play	
CIAIR-VCV	2006	Words, Sentences	Read-aloud	—	
CENSREC-1	2006	Words	Read-aloud	—	
CENSREC-1-C	2006	Words	Read-aloud	—	
CENSREC-2	2006	Words	Read-aloud	—	
CENSREC-3	2006	Words, Sentences	Read-aloud	—	
JNAS	2006	Sentences	Read-aloud	—	
FW03	2006	Words	Read-aloud	—	
RWCP-SSD	2007	Sentences, Non-speech	Read-aloud	—	
UME-ERJ	2007	Words, Sentences	Read-aloud	—	
UME-JRF	2007	Words, Sentences	Read-aloud	—	
RIKEN-DLG	2007	Monologue, Dialogue	**Spontaneous**	Role-play	
MapTask	2007	Dialogue	**Spontaneous**	**Natural**	Task-oriented
S-JNAS	2007	Sentences	Read-aloud	—	
ASJ-JIPDEC	2007	Sentences, Dialogue	Read-aloud, **Spontaneous**	Role-play	
FW07	2007	Words	Read-aloud	—	
CENSREC-4	2008	Words	Read-aloud	—	
UUDB	2008	Dialogue	**Spontaneous**	**Natural**	Task-oriented
ETL-WD	2008	Words	Read-aloud	—	
Tsuruoka91-92	2008	Words, Sentences	Read-aloud	—	
INFANT	2008	Dialogue	**Spontaneous**	**Natural**	
X-Ray	2010	Sentences	Read-aloud	—	
MULTEXT-J	2010	Monologue	Acted	—	
MULTEXT-C	2010	Monologue	Acted	—	
CENSREC-1-AV	2011	Words	Read-aloud	—	
Keio-ESD	2011	Words	Acted	—	
JVPD	2011	Words	Read-aloud	—	
TITML-IDN	2011	Sentences	Read-aloud	—	
TITML-ISL	2012	Sentences	Read-aloud	—	
AWA-LTR	2012	Words, Sentences	Read-aloud	—	
Aragusuku	2013	Words, Sentences	Read-aloud	—	
Oogami	2013	Words, Sentences	Read-aloud	—	
OGVC	2013	Dialogue	Acted, **Spontaneous**	**Natural**	
Chiba3Party	2014	Dialogue	**Spontaneous**	**Natural**	
JWC	2017	Words	Read-aloud	—	

Table 1: Corpora distributed by the NII-SRC (as of May 2018). Note: “Launched” means the first year of distribution by the NII-SRC, rather than the year in which the speech was recorded or distributed directly by the developers.

alogues involving two people: a professional and a customer who asks questions about purchasing a car and overseas travel plans. The Priority Areas “Spoken Dialogue” Simulated Spoken Dialogue Corpus (PASD) also contains conversations between a user and various systems (such as those that involve a secretary system, scheduling appointments, travel guides, and telephone shopping); two people simulate the user and the system.

Although the speakers in these dialogues played roles in simulated situations, they produced spontaneous utterances because they improvised what to say. These corpora have performed to some extent; however, they are still insufficient for general studies on spontaneous speech. The critical point of such investigations is not only to demonstrate the spontaneity of utterances, but also their naturalness and diversity; it is difficult to achieve these goals in role-playing situations.

Through these circumstances, various natural conversations have been collected as a new trend. As shown in Table 1, in recent years, natural situations have become more popular than role-playing[1]. We introduce five corpora, as follows.

The Chiba University Japanese Map Task Dialogue Corpus (MapTask) (Ichikawa et al., 2000) is a Japanese version derived from the Home Communications Research Centre (HCRC) Map Task Corpus, which was developed by a group at the University of Edinburgh, mainly for linguistic research (Thompson et al., 1993). It contains task-oriented dialogues using maps, with two participants involved: an instruction-giver who has a map with a route, and an instruction-follower who has a map without one. Although the participants had the roles of giver and follower, this was not role-play because they talked spontaneously in order to simulate how they naturally speak in everyday life.

The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB) (Mori et al., 2008) also consists of task-oriented dialogues. The dialogues were produced from "four-frame cartoon sorting tasks" (Mori et al., 2003) in which two participants have four cards extracted from a four-frame cartoon and they estimate the original order. The unique characteristic of this corpus is that it was designed to collect spontaneous and emotional utterances for studies on paralinguistic behavior.

The NTT Infant Speech Database (INFANT) (Amano et al., 2009) contains speech data uttered by five children (from three families) who are native Japanese speakers. The data were recorded for more than one hour per month since they were born until they were five years old. From this corpus, we can obtain the children's spontaneous utterances in daily conversations and the changes they experienced that are associated with growing up.

The Online Gaming Voice Chat Corpus with Emotional Labels (OGVC) (Arimoto et al., 2012) is a collection of natural and acted speeches used for emotional studies. The natural speech dialogues were recorded from voice chats that took place in an online game involving 2–3 players. The players expressed a lot of emotions because they were absorbed in playing the game. In addition, the speech that professional actors uttered in accordance with transcriptions of the natural speech is also included. For applications of emotional research, perceptual emotion labels and their intensity rates are appended to the utterances.

The Chiba Three-party Conversation Corpus (Chiba3Party) (Den and Enomoto, 2007) is a collection of casual conversations among three people of the same gender who are friendly with each other. The recording operator tried to avoid placing any restrictions on the content and progress of the conversations; thus, the conversations have a high degree of spontaneity. This corpus aims to contribute to descriptions and modeling of human interactions. Consequently, the transcriptions and morphological information based on conversation analysis are substantial.

Hence, recent corpora have been constructed that take diversity of speech into account.

2.2. CSJ

The National Institute for Japanese Language and Linguistics (NINJAL) is a comprehensive research organization. Collaborative efforts between NINJAL and the Communications Research Laboratory led to the development of a large-scale, spontaneous speech corpus called the Corpus of Spontaneous Japanese (CSJ). This corpus is useful for the investigation and modeling of spontaneous speech, as well as the study of speech recognition and summarization technology; NINJAL has publically distributed the CSJ since 2004 (Maekawa, 2003). The CSJ contains monologues consisting of academic presentations, simulated public speech, and dialogues (such as interviews with speakers and free-form conversations). The academic presentations were recorded live in nine different academic societies covering the fields of engineering, the social sciences, and the humanities. The public speeches are studio recordings of paid laypeople on everyday topics presented in front of a small audience.

One of the special features of the CSJ is that it is the largest spontaneous speech corpus in Japan. Its speech signals amount to about 660 hours and were uttered by around 1,400 different speakers. This quantity of data satisfies the construction of the language model for recognition of spontaneous speech, as well as applications to natural language processing studies on spontaneous speech. Furthermore, the wide range of speakers is useful for investigations on phonetic and linguistic variation caused by spontaneity.

Another unique quality of this corpus is its abundance of linguistic, phonetic, and prosodic labels aligned to the data. As for the linguistic labels, transcription texts were annotated using two types of part-of-speech systems, and differed regarding the length of morphological units that reflect the complex word boundaries of the Japanese language. In addition, transcription tags that were designed to represent fillers and disfluency particular to spontaneous speech were embedded in the transcriptions. As for the phonetic labels, phoneme labels considering phonetic events — such as the release of stop closure, the distinction between voiced affricates and fricatives, and the voicing of vowels — were assigned to the speech signals. Regarding the prosodic labels, X-JToBI labels (Maekawa et al., 2002) — which were extended from the J_ToBI scheme representing the intonational structure of Japanese — were appended to the transcriptions to represent prosodic variations observed in spontaneous speech. Although all of these labels have been adopted into

[1] The GSR(A) "Regional Differences in Spoken Japanese Dialects" Spoken Japanese Dialect Corpus (GSR-JD) aimed to record dialects in each region of Japan and compare them. Although the launch year of GSR-JD is older than that of other natural speech corpora, the spontaneity of the collected conversations is not the primary purpose.

only a subset of the CSJ (called the CSJ-Core) due to the high cost of labeling, there is no other corpus with as many types of labels as these.

The CSJ is useful for research on speech recognition, natural language processing, prosodics, linguistics, and the paralinguistics of spontaneous speech

3. Prospects

As described in Section 2., some spontaneous speech resources were developed. However, considering the diversity of speech in real-world situations, the data remain insufficient. For example, although INFANT provides utterances of children under six years old, utterances of children who are a little older, as well as elderly speakers, are necessary to represent the diversity caused by the growth and ages of speakers. The Chiba3Party provides casual conversations among three participants sitting face-to-face, but conversations in everyday life do not always happen this way. The CSJ is mostly limited to presentations; therefore, it is possible that the data do not represent general spontaneous speech. We believe that spontaneous speech resources should be developed by many researchers in various organizations to satisfy the diversity of utterances, because single organizations may produce biased data.

Currently, studies are investigating the following themes related to spontaneous Japanese speech:

- Emotional speech (Arimoto, 2018)
- Elderly speech (Kitaoka et al., 2018)
- Areal dialects (Kibe et al., 2018)
- Everyday conversations (Koiso et al., 2018)
- Multi-party interactions (Bono et al., 2018)
- Human-machine (i.e., robot and speech assistant systems) interactions (Funakoshi, 2018; Higashinaka et al., 2018)

The refereed papers provide details of each study.

4. Conclusion

We introduced representative speech resources of spontaneous Japanese that are publically distributed, and described the characteristics of each resource. In recent years, the amount of corpora containing spontaneous Japanese speech have increased; however, the quantity of speech resources is still insufficient to meet the demands of studies examining topics such as automatic speech recognition and natural language processing. We expect that more speech corpora that gather improvised utterances will gradually be developed to cover the diversity of spontaneous speech.

5. Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 17H00914, and a project of the Center for Corpus Development, NINJAL.

6. Bibliographical References

Amano, S., Kondo, T., Kato, K., and Nakatani, T. (2009). Development of Japanese infant speech database using longitudinal recordings from birth to five years old. In *2009 Oriental COCOSDA International Conference on Speech Database and Assessments*, pages 31–37, Aug.

Arimoto, Y., Kawatsu, H., Ohno, S., and Iida, H. (2012). Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment. *Acoustical Science and Technology*, 33(6):359–369.

Arimoto, Y. (2018). Challenges on building authentic emotional speech corpus of spontaneous Japanese dialog. In *Proceedings of LREC2018 Special Speech Sessions*, pages 6–13.

Bono, M., Sakaida, R., Makino, R., and Joh, A. (2018). Miraikan SC corpus: A trial for data collection in a semi-open and semi-controlled environment. In *Proceedings of LREC2018 Special Speech Sessions*, pages 30–34.

Den, Y. and Enomoto, M. (2007). A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation. In T. Nishida, editor, *Conversational informatics: An engineering approach*, pages 307–330. John Wiley & Sons.

Funakoshi, K. (2018). A multimodal multiparty human-robot dialogue corpus for real world interaction. In *Proceedings of LREC2018 Special Speech Sessions*, pages 35–39.

Higashinaka, R., Ishii, R., Matsumura, N., Nunobiki, T., Itoh, A., Inagawa, R., and Tomita, J. (2018). Speech and language resources for the development of dialogue systems and problems arising from their deployment. In *Proceedings of LREC2018 Special Speech Sessions*, pages 40–46.

Ichikawa, A., Horiuchi, Y., and Tutiya, S. (2000). The Japanese map task dialogue corpus. *Journal of the Phonetic Society of Japan*, 4(2):4–15.

Kibe, N., Otsuki, T., and Sato, K. (2018). Intonational variations at the end of interrogative sentences in Japanese dialects: From the "corpus of japanese dialects". In *Proceedings of LREC2018 Special Speech Sessions*, pages 21–28.

Kitaoka, N., Iribe, Y., and Nishizaki, H. (2018). Construction of a corpus of elderly Japanese speech for analysis and recognition. In *Proceedings of LREC2018 Special Speech Sessions*, pages 14–20.

Koiso, H., Den, Y., Iseki, Y., Kashino, W., Kawabata, Y., Nishikawa, K., Tanaka, Y., and Usuda, Y. (2018). Construction of the corpus of everyday Japanese conversation: An interim report. In *Proceedings of LREC2018 (in print)*.

Maekawa, K., Kikuchi, H., Igarashi, Y., and Venditti, J. (2002). X-JToBI: An extended J_ToBI for spontaneous speech. In *Proc. ICSLP2002*, pages 1545–1548.

Maekawa, K. (2003). Corpus of spontaneous Japanese

: its design and evaluation. *Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 7–12.

Mori, H., Kasuya, H., Nakamura, M., and Amanuma, M. (2003). Some considerations for designing spoken dialogue database from the viewpoint of paralinguistic information. *Acoustical Science and Technology*, 24(6):376–378.

Mori, H., Satake, T., Nakamura, M., and Kasuya, H. (2008). Uu database: A spoken dialogue corpus for studies on paralinguistic information in expressive conversation. In *Proceedings of the 11th International Conference on Text, Speech and Dialogue*, TSD '08, pages 427–434, Berlin, Heidelberg. Springer-Verlag.

Thompson, H. S., Anderson, A., Bard, E. G., Doherty-Sneddon, G., Newlands, A., and Sotillo, C. (1993). The HCRC map task corpus: Natural dialogue for speech recognition. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 25–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Challenges of Building an Authentic Emotional Speech Corpus of Spontaneous Japanese Dialog

Yoshiko Arimoto
Faculty of Science and Engineering, Teikyo University
1-1 Toyosato, Utsunomiya, Tochigi, Japan
ar@mac-lab.org

Abstract

This paper introduces the challenges involved in studying authentic emotional speech collected from spontaneous Japanese dialog. First, three key issues related to emotional speech corpora are presented: data type (acted or spontaneous), efficient collection of emotional speech, and appropriate emotion labeling. To address these issues, a data collection scheme was developed, and a labeling experiment was performed. First, a data collection scheme using an online game task was applied to efficiently collect speakers' authentic emotional expressions during their real-life conversations. Then, to elucidate appropriate emotion labels for emotional speech and to commonize the emotion labels among several corpora, the relationship between emotion categories and emotion dimensions, which are two major approaches to psychological emotional modeling, was demonstrated by conducting a cross-corpus emotion labeling experiment with two different Japanese dialogue corpora (the Online Gaming Voice Chat Corpus with Emotional Label (OGVC) and the Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB)). Finally, the results are presented, and the advantages and disadvantages of these approaches are discussed.

Keywords: emotional speech corpus, Japanese dialog speech, data collection, emotion labeling

1. Introduction

Emotional speech has been studied to elucidate its acoustic profiles and for applications in automatic emotion recognition and emotional speech synthesis. Various emotional speech corpora have been used for such studies. Emotional speech corpora can be classified into two types based on how the speech is produced: acted emotional speech corpora and authentic emotional speech corpora. Many of the studies on emotional speech have used acted emotional speech to investigate the acoustical correlation with emotion (Williams, 1972; Itoh, 1986; Kitahara, 1988; Banse and Scherer, 1996; Engberg et al., 1997). Such acted speech consists of idealized speech samples generated to match someone's conception of what an emotion should be like (Cowie, 2009), with well-designed prosodic and acoustic expression recorded in the noiseless environment of a soundproof room.

The contrast to acted emotional speech is authentic emotional speech. For practical applications such as automatic emotion recognition research and emotional or expressive speech synthesis, speech corpora containing authentic emotional speech samples evoked during real-life conversation are indispensable because such applications are designed for a real-world environment, not a laboratory setting. Several research groups began to study spontaneous emotional speech in the late 1990s (Ang et al., 2002; Arimoto et al., 2007). In that research, several attempts were made to record the expression of authentic emotions during spontaneous dialogs: dialogs between the AutoTutor system and students (Litman and Forbesriley, 2006), dialogs between a robotic pet and a child (Batliner et al., 2011), and interviews in which the speaker's emotions were controlled by the experimenter (Douglas-Cowie, 2003). Devillers and Vidrascu investigated real conversations during telephone calls with a call center (Devillers et al., 2006). In addition, several studies on authentic emotional speech have been performed with spontaneous materials (Campbell, 2004; Arimoto et al., 2008; Mori et al., 2011). Zeng et al. (Zeng et al., 2009) and Cowie (Cowie, 2009) have presented detailed reviews of the history of emotional speech corpora and suggestions for constructing an emotional speech corpus.

However, some issues arise with regard to the use of authentic emotional speech samples collected from spontaneous dialog. One issue is the data type: acted speech or spontaneous speech. Cowie (Cowie, 2009) demonstrated an example of the implications of this issue by means of a meta-analysis of automatic emotion recognition. The recognition rate using authentic emotional speech is lower than that using acted emotional speech. This report suggested that authentic emotional speech acoustically differs from acted emotional speech. Jürgens et al. supported this suggestion by identifying acoustic differences between authentic emotional speech and acted speech (Jürgens et al., 2011). Moreover, a method trained on acted speech, with deliberately and exaggeratedly expressed emotion, failed to generalize to authentic speech with subtle and complex emotional expression (Batliner et al., 2003; Zeng et al., 2009). Another critical issue noted with respect to spontaneous materials is the quantity of authentic emotional speech collected during spontaneous dialog. Cowie observed that even a large speech corpus contains few emotional samples (Cowie, 2009). Campbell recorded telephone conversations and labeled each recorded utterance with an observed emotion (Campbell, 2004). Although real-life conversations were successfully recorded, little of the speech displayed strong emotional content. Ang et al. (Ang et al., 2002) also obtained little emotional speech, although approximately 22,000 utterances were collected from a pseudodialog. Those studies suggested that methods of evoking emotion are necessary to efficiently collect authentic emotional speech from spontaneous dialog.

Another issue is emotion labeling for authentic emotional speech. In research on emotion recognition from speech, the use of multiple large-scale speech corpora with common emotion labels is needed to test the effectiveness of recognition. However, two different corpora typically cannot be used together because the emotion labels for each of the corpora are assigned based on their own criteria; there is no common shared labeling for both of them. A more crucial problem is that different emotion labeling schemes are adopted for different speech corpora. There are two primary types of emotion labels, each based on one of two different psychological emotion theories. One is emotion category theory, which claims that emotions are discrete internal states such as joy or sadness, such as Ekman's Big Six emotions (Ekman and Friesen, 1975) or Plutchik's eight primary emotions (Plutchik, 1980). The other is emotion dimension theory, which claims that emotion is a continuous internal state with several dimensions, such as pleasant–unpleasant and aroused–sleepy, as described by Russell's circumplex model (Russell, 1980), for example. When different emotional speech corpora are labeled with different emotion labels based on different labeling schemes, it is not possible to use both corpora in the same study. Even if two corpora are labeled with emotion labels of the same type, the emotion labels are not considered to be equivalent between the two corpora.

Although the emotion labels cannot be equivalent among multiple corpora, several researchers have examined emotion recognition and emotional speech synthesis with multiple corpora (Zong et al., 2016; Song et al., 2016; Schuller et al., 2012; Zhang et al., 2011; Schuller et al., 2010; Schuller et al., 2009). Schuller et al. used eight emotional speech corpora in their research (Schuller et al., 2012; Zhang et al., 2011; Schuller et al., 2010; Schuller et al., 2009). The emotion labels for each of the eight corpora varied: one used four emotion categories, another used two emotion dimensions, another used two different emotion categories, and so on. The various emotion labels were classified by the researchers into one of four quadrants of an orthogonal two-dimensional space (pleasant–unpleasant and aroused–sleepy) to obtain ground-truth labels for the speech samples. However, this approach to using multiple corpora does not guarantee the equivalency of the emotion labels among the corpora. Zong et al. used four corpora for emotion recognition research by selecting speech samples that were labeled with the same emotions across all four corpora. However, this method also does not guarantee the equivalency of the emotion labels across the corpora and allows the use of only a limited number of utterances from the corpora. Thus, a standardization of common emotion labels across emotional speech corpora is required.

This paper reports the author's attempts to confront the issues described above. First, an authentic emotional speech collection scheme was developed to confront the issue of the efficient collection of emotional speech. Then, the relationship between the two well-known types of emotion labels, i. e., emotion categories and emotion dimensions, was investigated in a cross-corpus emotion labeling experiment using two publicly available Japanese emotional speech corpora to confront the issue of standardized emotion labeling. Finally, the results of these studies are summarized in the conclusion section.

2. Collection of Authentic Emotional Speech

For the efficient collection of emotional speech, a collection scheme based on an online game task was applied, and the results were assessed in comparison with other emotional speech material. The content of this section is a rewrite of the research paper (Arimoto et al., 2012).

2.1. Recording

2.1.1. Task

To record authentic emotional expression during real-life conversations, massively multiplayer online role-playing games (MMORPGs), which are part of daily life for some Japanese university students, were adopted as tasks for our recording sessions. The effectiveness of games in evoking emotion has been proven in previous studies (Anderson and Bushman, 2001; van't Wout et al., 2006; Ravaja et al., 2008; Hazlett, 2006; Hazlett and Benedek, 2007; Tijs et al., 2008). The MMORPG used for each recording session depended on the group of players. The players in each group were allowed to select a game that more than one of them had actually played and enjoyed in their daily lives. The most popular online game was $Ragnarok\ Online$, which three groups played during recording. $Monster\ Hunter\ Frontier$ and $Red\ Stone$ were chosen by the other groups. All players were instructed to form a party and to participate together in quests (tasks in the game) while they were gaming.

To encourage the game players to talk with each other and to vocally express their emotions, an online voice chat system was adopted as a tool for communication among the players. Players of a MMORPG typically discuss their strategies for collaboratively achieving their goals in game events through a chat function provided by the MMORPG. To ensure that their emotional reactions would be reflected in their speech, the players were instructed to communicate through a voice chat system rather than the text chat function. Through the use of a voice chat system, it was expected that the players' emotional reactions to game events and expressive speech influenced by the players' internal emotional states would be observed.

2.1.2. Speakers

The speakers were 13 university students (9 males and 4 females, mean age 22 years ($SD = 1.17$)) with experience playing online games. They participated in our recording sessions as online game players. The players participated in each recording session as a group with one or two friends of the same gender. Six dialogs (five dyadic dialogs and one triad dialog) were recorded. The mean prior online gaming experience per player was 38 months ($SD = 14$), and the mean playing time per month was 33 hours ($SD = 35$).

2.1.3. Recording Environment

Figure 1 shows our recording environment. Each player in the group was located at a remote site on the campus of

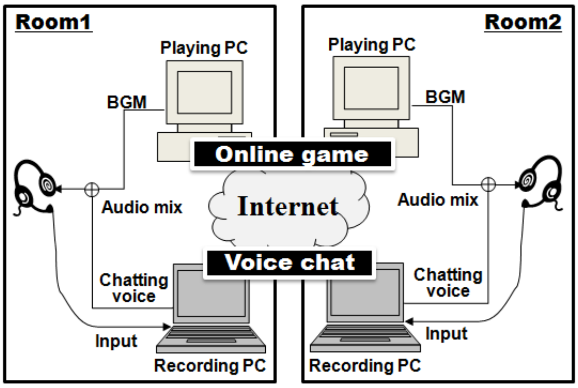


Figure 1: Recording environment.

Table 1: Number of utterances for each speaker.

Speaker	Utterances	Speaker	Utterances
01_MMK	816	04_MNN	934
01_MAD	740	04_MSY	938
02_MTN	884	05_MYH	464
02_MEM	736	05_MKK	539
02_MFM	557	06_FTY	712
03_FMA	561	06_FWA	781
03_FTY	452		
		Total	9114

Tokyo University of Technology and joined an online game together via the Internet. To make the recording environment as close as possible to the environments in which the players would usually play the game in their daily lives, a soundproof room was not used for recording. Each player sat on a chair in a classroom or on a tatami in a multipurpose space to play the game. The players put on headset microphones (Audio Technica ATH-30COM dynamic headsets) and talked with each other in a non-face-to-face environment via the Skype voice chat system. The dialogs among the players were recorded with a voice-recording system, Tapur for Skype. The speech was recorded separately at each recording site where each player was playing the game. Tapur recorded the local player's voice and a remote player's voice in different channels of a stereo sound file.

The recording time was approximately 1 hour for each group, and the total recording time was approximately 14 hours. The sound data were sampled at 48 kHz and digitized to 16 bits.

2.1.4. Segmentation and Transcription

The utterances in the recorded material were defined based on interpausal units (IPUs). Any continuous speech segment between pauses exceeding 400 ms was regarded as one utterance. The segmented utterances were orthographically transcribed into *kanji* (Chinese logograms) and *kana* (Japanese syllabograms). Jargon and special terms for online games, e. g., "bot" or "strage ("e su thi a: ru a ji" in reading)", and figures and counters were transcribed in *katakana* (angular Japanese syllabograms) as these words were heard. The following three transcription tags were prepared for laughs, coughs, and other purposes.

- {laughs},{coughs}
  Laughs, excluding utterances with laughing, and coughs.
- (?), (? (comment))
  An utterance that could not be transcribed due to noise or low sound volume.
- [comment:(comment)]
  Transcriber's comment.

Ultimately, the total number of utterances in our corpus was 9114. Table 1 shows the number of utterances for each speaker. In Table 1, the speakers are represented by speaker IDs.

2.2. Emotion Labeling

2.2.1. Speech Materials

For two speakers, 03_FMA and 02_MFM, 1009 utterances were not used in the analysis due to their low sound levels. Moreover, 1527 utterances with tags were also not used because these utterances could not be transcribed and their acoustic features could not be calculated. As a result, the total number of utterances used in the following analysis was 6578.

2.2.2. Procedure

The utterances were labeled with emotional categories in accordance with their perceived emotional information. After category labeling, the labeled utterances were rated for emotional intensity on the basis of how strongly the emotion was perceived from each utterance. Both the labelers and the raters were instructed to judge each utterance according to its acoustic characteristics, not its content.

Twenty-two labelers (14 males and 8 females) participated in the emotion labeling. Because the labeling of all 6587 utterances by each labeler would be costly and difficult, the number of utterances to be evaluated by each labeler was adjusted such that each utterance was labeled by three labelers. The labelers were instructed to choose one emotional state with which to label each utterance from ten alternatives: fear (FEA), surprise (SUR), sadness (SAD), disgust (DIS), anger (ANG), anticipation (ANT), joy (JOY), acceptance (ACC), a neutral state (NEU) with no emotion, or an utterance exhibiting an emotional state that is impossible to classify into any of the nine states above or subject to high noise or other disruption (OTH). The eight emotional states were selected with reference to the primary emotions of Plutchik's multidimensional model (Plutchik, 1980). Table 2 lists the ten emotional state classifications, their abbreviations, and their definitions. These ten definitions were presented to the labelers to give them a common understanding of each emotional state. The definitions were prepared by referring to a dictionary (Yamada et al., 2005). Each utterance was presented in a random order to each labeler to mitigate possible order effects.

Each utterance was rated for its emotional intensity by 18 raters (13 males and 5 females). Only utterances for which at least two of the three labelers agreed on one of the eight emotion labels were rated. The utterances were presented

Table 2: Abbreviations and definitions of emotional states.

State	Abbr.	Definition
Fear	FEA	Feelings of avoidance toward people or things that are harmful
Sadness	SAD	Feelings of sorrow for irrevocable consequences such as misfortune or loss
Disgust	DIS	Feelings of avoidance toward unacceptable states or acts
Anger	ANG	Feelings of irritation or annoyance with an unforgiven subject
Surprise	SUR	Feelings of being disturbed, caught off balance, or confused after experiencing unexpected events
Anticipation	ANT	Feelings of longing for a desirable eventuality or a favorable opportunity
Joy	JOY	Feelings of gladness and thankfulness indicating intense satisfaction with something
Acceptance	ACC	Feelings of active involvement in something fascinating or positive
Neutral	NEU	No feelings at all
Other	OTH	Impossible to classify into any of the nine states above, or utterances with noise, etc.

Table 3: Results of emotion labeling. The percentages were calculated by dividing the number of utterances corresponding to each emotional state by the total number of utterances. The total number of utterances was 6578.

State	Partial		Full	
	Utterances	Percent	Utterances	Percent
FEA	142	2.2	33	0.5
SAD	243	3.7	49	0.7
DIS	335	5.1	45	0.7
ANG	237	3.6	60	0.9
SUR	565	8.6	177	2.7
ANT	427	6.5	69	1.0
JOY	595	9.0	174	2.6
ACC	303	4.6	27	0.4
NEU	798	12.1	116	1.8
OTH	200	3.0	30	0.5
Total	3845	58.5	780	11.0

in a random order to each rater. The raters were instructed to rate the emotional intensity of each utterance on a five-point scale from 1 (weak) to 5 (strong).

2.3. Analysis

Two types of agreement among the three label evaluations were calculated: partial agreement (two out of three labelers agreed on one emotion) and full agreement (all three labelers agreed on one emotion). Moreover, the mean correlation coefficient among the 18 raters was calculated.

To assess the efficiency of our data collection scheme for authentic emotional speech, the number of labeled instances among our speech materials was compared with those of two other sets of speech materials. One of these consists of spontaneous pseudodialogs for angry speech classification (Ang et al., 2002), and the other is a speech database for paralinguistic information studies, the Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB) (Mori, 2008). The emotion labeling rate was calculated by dividing the number of emotion labels by the total number of labels, in accordance with (Ang et al., 2002). Note that the labeling schemes for the three sets of materials are not completely the same and that the calculation was performed for the sake of comparison among them. Each utterance in the angry speech material set (Ang et al., 2002) is labeled with one of 7 emotional state labels: neutral, annoyed, frustrated, tired, amused, other, or not applicable (containing no speech data from the user). Utterances with the annoyed, frustrated, tired, amused, and other labels were regarded as emotional utterances for the comparison. The utterances in the UUDB are not labeled with single emotional states. Instead, they are rated on a seven-point scale for each of six paralinguistic information values: pleasant–unpleasant, aroused–sleepy, dominant–submissive, credible–doubtful, interested–indifferent, and positive–negative. The utterances are all associated with six paralinguistic information values; hence, a nonemotional state is never assessed. To compare the emotion labeling rates between our speech materials and the UUDB, the UUDB utterances rated with scores from 3 to 5 (weak or none) for all 6 values were regarded as nonemotional utterances, and the rest were regarded as emotional utterances. A χ^2 test was conducted to compare the emotion labeling rates among the three speech material sets.

2.4. Results

Table 3 shows the numbers of utterances exhibiting the two types of interlabeler agreement. The number of utterances with partial agreement is 3,845, and the number of utterances with full agreement is 780. The partial and full agreement rates are 58.5% (chance level: 28%) and 11.0% (chance level: 1%), respectively.

The mean correlation coefficient among the 18 raters is 0.24 (range $= -0.01$ – 0.52). The range of correlation coefficients among the 18 raters is widely spread, indicating that the criteria used to rate emotional intensity were different among the raters.

Figure 2 shows the frequency of emotion labels in each set of speech materials. The χ^2 test revealed a significant difference among the three speech material sets ($\chi^2(2) = 27659.87$, $p < 0.001$). Our speech material set has a significantly higher emotion labeling rate than the other two ($p < 0.01$, indicated by asterisks in Fig. 2).

2.5. Discussion

Quite high agreement rates were obtained for both partial and full agreement. The partial and full agreement rates are 58.5% and 11.0%, respectively, which are much higher than the chance levels for partial and full agreement (28%

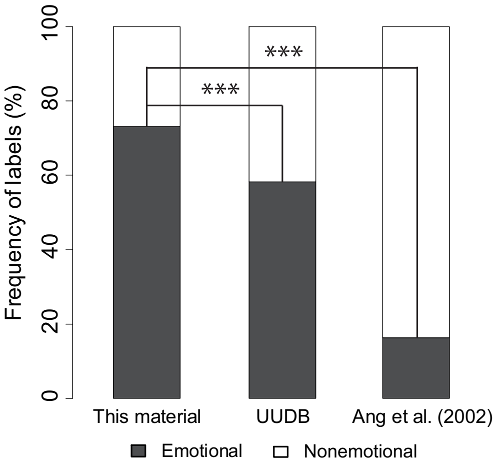


Figure 2: Frequencies of emotional labels.

and 1%, respectively). The results suggest that the labelers could perceive the same emotions from the recorded utterances. This implies that the emotional speech collected via the proposed approach is perceptually distinguishable for listeners.

The χ^2 test revealed a significant difference among the three sets of speech materials ($\chi^2(2) = 27659.87$, $p < 0.001$). Our speech material set has a significantly higher emotion labeling rate than the other two. The total number of labeling instances in our speech material set is 19,734 labels (6,578 utterances × three labelers). Among them, 14,414 labels are emotional labels corresponding to the eight types of emotional state; consequently, a very high percentage, 73.0%, of the total labeling instances have emotional labels. The total number of labeling instances in the speech material set of Ang et al. (Ang et al., 2002) is 49,553; these instances were judged by 2.62 mean labelers per utterance and include 4,904 emotional labels. The corresponding emotion labeling rate is thus quite low, 9.9%. The UUDB has 14,520 labels assigned by three labelers. Of these labels, 58.2% (8,446 labels) are emotional labels. These results imply that the proposed collection scheme can yield a relatively high percentage of emotional speech that is perceptually distinguishable by listeners.

The speech materials with emotion labels recorded via the proposed collection scheme are publicly available from the distributor, NII–SRC, as the Online Gaming Voice Chat Corpus with Emotional Label (OGVC) (Arimoto and Kawatsu, 2013).

3. Cross-corpus Emotion Labeling

To elucidate appropriate emotion labels for emotional speech and to standardize the emotion labels among several corpora, we investigated the relationship between two well-known types of emotion labels, i. e., emotion categories and emotion dimensions. Using two publicly available Japanese dialog speech corpora with emotion labels, we conducted cross-corpus emotion labeling to label the utterances in the two corpora with both emotion category labels and emotion dimension labels. The content of this section is a rewrite of the conference paper (Arimoto and Mori, 2017).

3.1. Speech Materials

Two publicly available Japanese dialog speech corpora were used for this research: the OGVC (Arimoto and Kawatsu, 2013) and the UUDB (Mori, 2008).

The UUDB is a collection of natural, spontaneous dialogs from Japanese college students. The participants engaged in a "four-frame cartoon sorting" task, in which four cards, each containing one frame extracted from a cartoon, are shuffled and each participant is given two cards out of the four and is asked to estimate their original order without looking at the remaining cards. The current release of the UUDB includes dialogs from seven pairs of college students (12 females and 2 males), comprising 4,840 utterances. An utterance is defined as a continuous speech segment bounded by either silence (> 400 ms) or slash unit boundaries. For all utterances, the perceived emotional states of the speakers are provided. The emotional states are annotated with the following six abstract dimensions:

- pleasant–unpleasant
- aroused–sleepy
- dominant–submissive
- credible–doubtful
- interested–indifferent
- positive–negative

The emotional state corresponding to each utterance is evaluated on a seven-point scale for each dimension. On the pleasant–unpleasant scale, for example, 1 corresponds to extremely unpleasant; 4, to neutral; and 7, to extremely pleasant. All 4,840 utterances were used in this experiment.

3.2. Procedure

The two corpora used in this study have different types of emotion labels; consequently, they cannot be used together for any research in their original forms. Therefore, in this experiment, the emotion labels included in the original corpora were discarded, and all utterances in both corpora were newly labeled with emotion categories and emotion dimensions to obtain common emotion labels across the two corpora.

Three qualified labelers, selected via a previously performed labeler screening process, performed the cross-corpus emotion labeling. The mean age of the three labelers was 22 years ($SD = 0.82$).

The emotion labeling frameworks for both emotion category labeling and emotion dimension labeling were the same as those used in the construction of the two original corpora. For emotion category labeling, the labelers were instructed to choose one of 10 categories (JOY, ACC, FEA, SUR, SAD, DIS, ANG, ANT, NEU, and OTH) for each utterance. The ground-truth label for each utterance was determined by majority vote among the labelers. For emotion dimension labeling, the labelers were instructed to rate each of the six emotion dimensions on a seven-point scale for each utterance. The ground-truth label for each emotion dimension for each utterance was defined as the mean score among the labelers. Each labeler performed both the emotion category and emotion dimension labeling tasks. The emotion dimension labeling task preceded the emotion category labeling task.

Table 4: The number of utterances in each emotion category.

Emotion	OGVC	UUDB	Total
JOY	438	259	697
ACC	623	1030	1653
FEA	282	94	376
SUR	313	120	433
SAD	488	331	819
DIS	970	406	1376
ANG	128	39	167
ANT	186	59	245
NEU	18	13	31
Total	3446	2351	5797

Each labeler evaluated a total of 11,418 utterances from the OGVC and the UUDB (6,578 from the OGVC and 4,840 from the UUDB). The 11,418 utterances were randomly separated into blocks. The cross-corpus emotion labeling was performed in 104 blocks for 11,418 utterances × 2 types of labeling (category and dimension).

3.3. Analysis

To assess the independence of each emotion category from the others in an n-dimensional emotional space, equivalence tests between two n-dimensional Gaussian mixture models (GMMs) were conducted. For each pair of emotion categories E_1 and E_2, the n-dimensional variables X_1 and X_2 belonging to each category were assumed to be generated from their corresponding GMMs. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ denote the subdatasets belonging to E_1 and E_2, respectively, and N_1 and N_2 denote the respective data sizes. The null hypothesis (H_0) and the alternative hypothesis (H_1) are as follows:

H_0: All instances of X_1 are generated from a GMM M_1, and all instances of X_2 are generated from a GMM M_2 that is identical to M_1.

H_1: All instances of X_1 are generated from a GMM M_1, and all instances of X_2 are generated from a GMM M_2 that differs from M_1.

The null hypothesis can be tested using a parametric bootstrap likelihood ratio test, in which the distribution of the difference of the deviances (-2 times the log likelihood ratio) between the null model (M_1 and M_2 are trained as identical models on random samples with a data size of $N_1 + N_2$) and the alternative model (M_1 and M_2 are trained separately on random samples with a data size of N_1 and random samples with a data size of N_2, respectively) is estimated via random sampling under H_0. If the difference of the deviances between the null model (identical GMMs trained on $\mathbf{x}_1 + \mathbf{x}_2$) and the alternative model (GMMs trained separately on $\mathbf{x}_1$ and $\mathbf{x}_2$) falls into the critical region ($\alpha = 5\%$), then the null hypothesis is rejected, and the two emotion categories are considered to be independently distributed in the n-dimensional emotional space. Such likelihood ratio tests were conducted for all combinations of the nine emotion categories.

3.4. Results

Table 4 shows the number of utterances in each emotional category identified as a result of the emotion category labeling process. The total number of utterances for which two out of the three labelers agreed on one emotion label is 5,797 (3,446 for the OGVC and 2,351 for the UUDB), corresponding to 51% of the total utterances subjected to cross-corpus labeling (52% of the OGVC utterances and 49% of the UUDB utterances). The emotions assigned to the highest numbers of utterances, in descending order, are ACC, DIS, JOY and SAD. Following emotion category labeling, these 5,797 utterances were used in the analysis of the mapping of the emotion categories to n-dimensional emotional spaces.

Table 5: Differences in deviances between emotion categories mapped to a three-dimensional emotional space.

	ACC	FEA	SUR	SAD	DIS	ANG	ANT	NEU
JOY	1187.3*	762.7*	680.1*	1406.7*	1660.8*	679.6*	178.0*	159.1*
ACC		501.8*	585.3*	1248.5*	1169.5*	765.3*	368.4*	367.2*
FEA			98.3*	342.9*	123.7*	215.2*	268.4*	41.7*
SUR				802.0*	482.7*	287.7*	253.2*	31.0
SAD					534.4*	603.8*	678.8*	38.2
DIS						108.6*	463.0*	11.6
ANG							361.6*	101.6*
ANT								99.8*

Figure 3 shows the distributions of the emotion categories in the two-dimensional emotional spaces of arousal vs. pleasantness, dominance vs. pleasantness, and dominance vs. arousal. Table 5 shows the differences in the deviances between the emotion categories when mapped to the corresponding three-dimensional emotional space. The asterisks in Table 5 indicate the combinations of emotion categories for which the hypothesis H_0 is rejected and the hypothesis H_1 is accepted ($p < 0.05$). For many combinations of emotion categories, H_0 is rejected; H_0 was not rejected in only three tests, namely, for NEU when testing with SUR, SAD, and DIS.

3.5. Discussion

In the pleasantness vs. arousal space shown in the left panel of Fig. 3, JOY (the solid red line in Fig. 3) is placed in the upper right quadrant, corresponding to high arousal and high pleasantness; SUR (dashed green line) corresponds to high arousal; SAD (solid green line) corresponds to low arousal; and ANG (solid blue line) lies in the upper left, corresponding to high arousal and low pleasantness. These distributions are similar to Russell's circumplex model (Russell, 1980). The results also show that NEU (solid purple line) lies near 4 on the pleasantness axis but between 2 and 4 on both the arousal and dominance axes. NEU is generally considered to be an emotionally neutral state, which should correspond to a score of 4 in any emotion dimension. However, our results imply that neutral utterances are neutral in the pleasantness dimension but are not necessarily neutral in the other dimensions.

The results of the likelihood ratio tests on the distributions of the emotion categories in the three-dimensional emotional space suggest that all pairs of emotion categories except NEU–SUR, NEU–SAD, and NEU–DIS exhibit significant differences between each other ($p < 0.05$). In other words, all emotion categories except NEU are independent of each other. This finding suggests that the information of the eight emotion categories (JOY, ACC, FEA, SUR, SAD, DIS, ANG, and ANT) is not lost even in the emotion di-

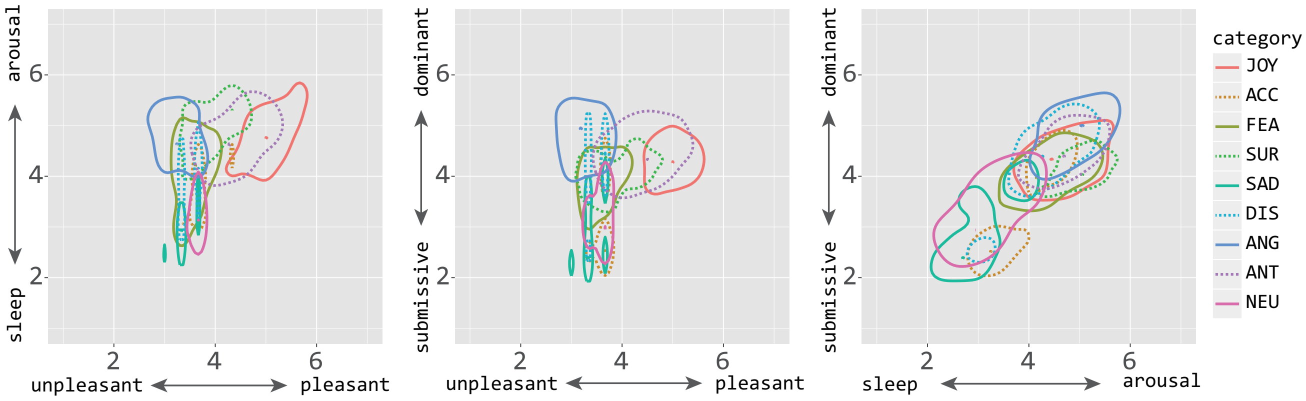


Figure 3: Distributions of emotion categories in two-dimensional emotional spaces.

mension representation.

4. Conclusions

For the efficient collection of emotional speech, a collection scheme based on an online game task and a voice chat system was developed, and its results were assessed by comparison with other emotional speech materials. A χ^2 test revealed that by using the proposed collection scheme, emotionally expressive speech can be efficiently collected.

To elucidate appropriate emotion labels for emotional speech and to commonize emotion labels among several corpora, we first studied the relationship between emotion categories and emotion dimensions. Using two Japanese dialog speech corpora with emotion labels, cross-corpus emotion labeling was conducted to label the utterances in the two corpora with both emotion category labels and emotion dimension labels. Then, likelihood ratio tests were conducted to assess the independence of each emotion category from the others in a three-dimensional emotional space.

The tests revealed that all pairs of emotion categories except neutral–surprise, neutral–sadness, and neutral–disgust exhibit significant differences between each other. Thus, all emotion categories except neutral are independent of each other in the dimensional emotional space.

These results suggest the surprising conclusion that the information of the eight emotion categories, including joy, acceptance, fear, surprise, sadness, disgust, anger, and anticipation, is not lost even in the emotion dimension representation. However, future research with other speech corpora in different languages may yield different results, because emotion perception heavily depends on language, culture and social norms. The universal standardization of emotion labeling can be accomplished only after examining the linguistic differences, cultural differences, and social differences that must be encompassed by standardized emotion labels.

5. Acknowledgments

This work was supported by a TATEISHI Science and Technology Foundation Grant for Research (A) and by JSPS KAKENHI Grant Number 17K00160. I would like to express my sincere appreciation to my collaborators, Dr. Hiromi Kawatsu from IBM and Prof. Hiroki Mori from Utsunomiya University.

6. Bibliographical References

Anderson, C. a. and Bushman, B. J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: a meta-analytic review of the scientific literature. *Psychological science*, 12(5):353–9, sep.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A. (2002). Prosody-Based Automatic Detection Of Annoyance And Frustration In Human-Computer Dialog. In *Proceedings of ICSLP 2002*, pages 2037–2040. in Proc. ICSLP 2002.

Arimoto, Y. and Mori, H. (2017). Emotion category mapping to emotional space by cross-corpus emotion labeling. In *Proceedings of Interspeech 2017*, pages 3276–3280.

Arimoto, Y., Ohno, S., and Iida, H. (2007). An Estimation Method of Degree of Speaker's Anger Emotion with Acoustic and Linguistic Features. *Journal of natural language processing*, 14(3):147–163. (in Japanese).

Arimoto, Y., Kawatsu, H., Ohno, S., and Iida, H. (2008). Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems. In *Proceedings of Interspeech 2008*, pages 322–325.

Arimoto, Y., Kawatsu, H., Ohno, S., and Iida, H. (2012). Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment. *Acoustical Science and Technology*, 33(6):359–369.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636.

Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. (2003). How to find trouble in communication. *Speech Communication*, 40(1-2):117–143, apr.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., and Kessous, L. (2011). Whodunnit - Searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech & Language*, 25(1):4–28, jan.

Campbell, N. (2004). Speech & Expression ; the Value of a Longitudinal Corpus The JST ESP corpus. In *LREC 2004*.

Cowie, R. (2009). Perceiving emotion: towards a realistic understanding of the task. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1535):3515–25, dec.

Devillers, L., Vidrascu, L., and Bp, L.-c. (2006). Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs. In *Interspeech 2006.*, pages 801–804.

Douglas-Cowie, E. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60, apr.

Ekman, P. and Friesen, W. V. (1975). *Unmasking the Face: A Guide to Recognizing Emotions From Facial Expressions*. Prentice Hall, New Jersey.

Engberg, I. S., Hansen, A. V., Andersen, O., and Dalsgaard, P. (1997). Design, Recording and Verification of a Danish Emotional Speech Database. In *Proceedings of Eurospeech 1997*, volume 4, pages 1695 – 1698.

Hazlett, R. and Benedek, J. (2007). Measuring emotional valence to understand the user's experience of software. *International Journal of Human-Computer Studies*, 65(4):306–314, apr.

Hazlett, R. L. (2006). Measuring emotional valence during interactive experiences. In *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, pages 1023–1026, New York, New York, USA, apr. ACM Press.

Itoh, K. (1986). A basic study on voice sound involving emotion. III. Non-stationary analysis of single vowel [e]. *The Japanese journal of ergonomics*, 22(4):211–217.

Jürgens, R., Hammerschmidt, K., and Fischer, J. (2011). Authentic and play-acted vocal emotion expressions reveal acoustic differences. *Frontiers in psychology*, 2(July):180, jan.

Kitahara, Y. (1988). Prosodic components of speech in the expression of emotions. *The Journal of the Acoustical Society of America*, 84(S1):S98–S99, nov.

Litman, D. and Forbesriley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5):559–590, may.

Mori, H., Satake, T., Nakamura, M., and Kasuya, H. (2011). Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication*, 53(1):36–50, aug.

Plutchik, R. (1980). *Emotions: A psychoevolutionary synthesis*. Harper & Row, New York.

Ravaja, N., Turpeinen, M., Saari, T., Puttonen, S., and Keltikangas-Järvinen, L. (2008). The psychophysiology of James Bond: phasic emotional responses to violent video game events. *Emotion (Washington, D.C.)*, 8(1):114–20, feb.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A. (2009). Acoustic emotion recognition: A benchmark comparison of performances. *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009*, pages 552–557.

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2010). Cross-Corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131.

Schuller, B., Zhang, Z., Weninger, F., and Burkhardt, F. (2012). Synthesized speech for model training in cross-corpus recognition of human emotion. *International Journal of Speech Technology*, 15(3):313–323.

Song, P., Zheng, W., Ou, S., Zhang, X., Jin, Y., Liu, J., and Yu, Y. (2016). Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. *Speech Communication*, 83:34–41.

Tijs, T., Brokken, D., and Ijsselsteijn, W. (2008). Creating an Emotionally Adaptive Game. In S M Stevens et al., editors, *Proceedings of the 7th International Conference on Entertainment Computing*, volume 5309 of *LNCS 5309*, pages 122–133. Springer-Verlag.

van't Wout, M., Kahn, R. S., Sanfey, A. G., and Aleman, A. (2006). Affective state and decision-making in the Ultimatum Game. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 169(4):564–8, mar.

Williams, C. E. (1972). Emotions and Speech: Some Acoustical Correlates. *The Journal of the Acoustical Society of America*, 52(4B):1238–1250, oct.

Yamada, T., Shibata, T., Kuramochi, Y., and Yamada, A. (2005). *Shin meikai kokugo jiten*. Sanseido, Tokyo, 6 edition. (in Japanese).

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58.

Zhang, Z., Weninger, F., Wöllmer, M., and Schuller, B. (2011). Unsupervised learning in cross-corpus acoustic emotion recognition. *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings*, pages 523–528.

Zong, Y., Zheng, W., Zhang, T., and Huang, X. (2016). Cross-Corpus Speech Emotion Recognition Based on Domain-Adaptive Least-Squares Regression. *IEEE Signal Processing Letters*, 23(5):585–589, may.

7. Language Resource References

Arimoto, Yoshiko and Kawatsu, Hiromi. (2013). *Online gaming voice chat corpus with emotional label (OGVC)*. Speech Resource Consortium, National Institute of Informatics, ISLRN 648-310-192-037-7.

Mori, Hiroki. (2008). *Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB)*. Speech Resource Consortium, National Institute of Informatics.

Construction of a Corpus of Elderly Japanese Speech for Analysis and Recognition

Norihide Kitaoka[1], Yurie Iribe[2] and Hiromitsu Nishizaki[3]
[1]Department of Computer Science, Tokushima University,
2-1 Minamijohsanjima, Tokushima, Japan
[2]School of Information Science and Technology, Aichi Prefectural University,
1522-3 Ibaragabasama, Nagakute-shi, Aichi, Japan
[3]Graduate School of Interdisciplinary Research, Faculty of Engineering, University of Yamanashi,
4-3-11 Takeda, Kofu-shi, Yamanashi, Japan
kitaoka@is.tokushima-u.ac.jp, iribe@ist.aichi-pu.ac.jp, hnishi@yamanashi.ac.jp

Abstract

We have constructed a new speech data corpus using the utterances of 100 elderly Japanese people, in order to improve the accuracy of automatic recognition of the speech of older people. Humanoid robots are being developed for use in elder care nursing facilities because interaction with such robots is expected to help clients maintain their cognitive abilities, as well as provide them with companionship. In order for these robots to interact with the elderly through spoken dialogue, a high performance speech recognition system for the speech of elderly people is needed. To develop such a system, we recorded speech uttered by 100 elderly Japanese who had an average age of 77.2, most of them living in nursing homes. Another corpus of elderly Japanese speech called S-JNAS (Seniors-Japanese Newspaper Article Sentences) has been developed previously, but the average age of the participants was 67.6. Since the target age for nursing home care is around 75, much higher than that of most of the S-JNAS samples, we felt a more representative corpus was needed. In this study we compare the performance of our new corpus with both the Japanese read speech corpus JNAS (Japanese Newspaper Article Speech), which consists of adult speech, and with the S-JNAS, the senior version of JNAS, by conducting speech recognition experiments. Data from the JNAS, S-JNAS and CSJ (Corpus of Spontaneous Japanese) was used as training data for the acoustic models, respectively. We then used our new corpus to adapt the acoustic models to elderly speech, but we were unable to achieve sufficient performance when attempting to recognize elderly speech. Based on our experimental results, we believe that development of a corpus of spontaneous elderly speech and/or special acoustic adaptation methods will likely be necessary to improve the recognition performance of dialog systems for the elderly.

Keywords: elderly speech corpus, nursing home care, speech corpus construction, speech recognition, companion robots

1. Introduction

Previous research suggests that elderly people have more difficulty using information and communication technology (ICT) than younger adults (Júdice, 2010). The main reasons for this are the complexity of existing user interfaces, lack of familiarity with ICT on the part of many elderly and the limited set of available interaction modalities, since this technology is mainly designed with younger users in mind. Hence, adapting the technology to better suit the needs of the elderly, for instance by increasing the choice of available interaction modalities, will help ensure that the elderly have access to these technologies. Previous research suggests that speech is the easiest and most natural modality for human-computer interaction (HCI) (Acartürk, 2015). Speech is also the preferred modality for interacting with mobile devices when users have permanent impairments such as arthritis, or when temporary limitations such as driving or carrying objects make it difficult to use other modalities such as touch.

It is hoped that ICT can be used to help maintain the health of the elderly. Daily verbal interaction helps them maintain their cognitive ability, reducing the risk of dementia, and may also ease loneliness. In a super-aging society such as Japan, where we face an acute shortage of care workers, spoken dialogue systems could play an important role. However, the speech recognizers which would need to be used in interfaces such as spoken dialogue systems do not currently work well for elderly users. A mismatch between the acoustic model and the acoustic characteristics of user speech is one factor which reduces speech recognition accuracy. Some studies have found differences in the acoustic characteristics of the speech of the elderly and that of younger people (Winkler, 2003). In particular, elderly speech frequently contains inarticulate speech, which occurs when the speaker does not fully open the mouth.

Additionally, acoustic models are often constructed using the speech of adults, excluding the aged. As a result, it has been reported that deterioration in speech recognition accuracy with elderly users is caused by mismatches between acoustic models and the acoustic characteristics of elderly speech (Anderson, 1999; Baba, 2001; Vipperla, 2008). Therefore, it is important to construct an acoustic model which takes into account the characteristics of elderly speech, in order to improve the speech recognition accuracy of speech applications designed for the elderly.

To address this problem, we have constructed a new speech data corpus using the utterances of 100 elderly Japanese people in three age categories; the young-old, old-old and oldest-old, in order to improve recognition of the speech of the elderly. In this study we compare the characteristics of our new corpus with those of two other speech databases which have been used to construct acoustic models in Japan

Figure 1: Recording speech in a Japanese nursing home for the elderly.

(JNAS and S-JNAS), and then experimentally evaluate our corpus by comparing speech recognition performance when using each of the three corpora. In order to evaluate the effect of using spontaneous speech data, we also used the Corpus of Spontaneous Japanese (CSJ), which consists of speech from presentations given at Japanese acoustics conferences, in our experiment. In addition, we compared using elderly read speech versus dialog speech as our test data. Since our end goal is a speech recognition system for the elderly that can recognize dialog speech, we performed several pilot experiments using these various sources of test data.

2. Data Collection

Between May 2014 and February 2015, we collected 9.2 hours (5,030 sentences) of read speech from 100 elderly Japanese subjects. During data collection we recorded read speech from elderly subjects at four nursing homes for the elderly and at one university. Figure 1 shows a typical recording scene at a nursing home. The number of elderly subjects recorded at nursing homes was 56, and their ages ranged from 66 to 98 (average age: 82). The number of elderly subjects recorded at the university was 44, and their ages ranged from 60 to 78 (average age: 71). Table 1 shows the number of speakers recorded at each location. In this section, we describe the collected speech in detail.

2.1 Speaker Selection

Although it is possible to observe differences between teenage speech, young adult speech and elderly speech at the acoustic/phonetic level, it has not been conclusively determined whether or not there are any clear-cut, age-related acoustic/phonetic differences in human speech. This is partly because the aging of the speech organs is influenced by factors such as the abuse or overuse of the vocal folds, smoking, alcohol consumption, psychological stress and tension. Furthermore, features which are often considered to be typical of elderly speech can be related to situational circumstances, such as lexical and grammatical factors which are associated with different sociolinguistic registers. While it might be impossible to precisely determine an exact age at which an individual's speech should be considered to be elderly, researchers usually regard 60-70 years of age as the minimum age range for elderly speech. Therefore, for our corpus we decided to collect speech from subjects aged 60 and over. Apart from age, literacy and basic technical comprehension requirements, we had no other criteria for selecting speakers. We did not, for example, aim at a specific ratio of female to male speakers or screen speakers for pronunciation, etc. The age and sex distribution of our speakers are shown in Table 2. All of the speakers lived in Aichi prefecture. Some of the subjects were suffering from dementia and more than half of the subjects were living in nursing homes. In the popular S-JNAS elderly Japanese speech database, there are only eight speakers over 80 years of age, and the overall age distributions is also biased, therefore we chose as many subjects over 80 as possible. As a result, we were able to include recorded speech from 39 individuals more than 80 years of age in our corpus. This speech data should prove valuable for acoustic modeling and elderly speech analysis.

2.2 Data Collection Procedure

Each speaker uttered about 50 ATR phoneme-balanced sentences, for a total of about 9.2 hours of recorded speech for all of the subjects combined. The utterances were recorded using a desktop microphone. We explained the recording procedure and provided the sentences to be read to each subject, printed in kana characters. The subjects then practiced reading the sentences. Rest breaks were provided during the recording process in consideration of the physical condition of the subjects. In addition, after each recording session a subjective mood evaluation survey was conducted with each subject by facility staff, and their likelihood of suffering from dementia was assessed using the HDS-R (Hasegawa's Dementia Scale-Revised).

Table 1: Number of speakers and their average age at each recording location.

Recording location	**Number of speakers**	**Average age**
Nursing home A	10	85.5
Nursing home B	10	82.8
Nursing home C	17	80.6
Nursing home D	19	81.4
Nagoya University	44	70.8

Table 2: Age and sex distributions of speakers in our corpus.

Age	**Male**	**Female**	**Total**
60-64	1	3	4
65-69	3	10	13
70-74	3	22	25
75-79	6	13	19
80-84	4	16	20
85-89	1	10	11
90-94	3	3	6
95-99	1	1	2
Total	22	78	100

2.3 Selection of Japanese Sentences

When choosing sentences for the speakers to read, the goal was to create a corpus of read speech suitable for training acoustic models, thus sentence selection and structure were based on the existing JNAS speech corpora, a typical corpus used for constructing Japanese acoustic models (Iso, 1988; Kurematsu, 1990). The JNAS database consists of sentences from newspaper articles and is divided into 155 text sets of about 100 sentences per set, with 16,176 sentences in total. In addition, it contains ATR phonetically balanced sentences divided into 10 text sets, with about 50 sentences per set and 503 sentences in total. The ATR phonetically balanced sentences included 402 two-phoneme sequences and 223 three-phoneme sequences (625 items in total). The phonetically balanced sentences were extracted from newspapers, journals, novels, letters and textbooks, etc., so that different phonetic environments occur at the same rate as much as possible. The sentences consist of Set A ～ Set I (50 sentences each) and Set J (53 sentences). We selected these JNAS ATR phonetically-balanced sentences as phrases for our speech corpus. The number of utterances of each sentence set in each corpus is shown in Table 3.

2.4 Database

The total duration of the recorded speech in our corpus is approximately 9.2 hours. Each speaker was recorded using a desktop microphone, and their speech was stored with a wav header. The speech waves were digitized at a sampling frequency of 16 kHz using 16 bit audio. The recorded speech data was then divided into sentence units, and a pause of about 300 ms was inserted before and after each sentence. The new corpus was transcribed and the transcription of the speech data was verified and edited manually by trained employees who listened to the recorded speech data. When necessary, the phonemes and words of the sentences were changed to correspond to what the speakers actually said. The database includes information about the speakers (age, gender, subjective assessment of mood and likelihood of dementia), recording location, text transcription (in Japanese), set (A~J) and sentence number as attribute information.

Table 3: Number of utterances in each sentence set.

Set Name (Number of sentences)	**JNAS**	**S-JNAS**	**New Corpus**
Set A (50)	1,600	2,950	500
Set B (50)	1,600	2,950	500
Set C (50)	1,600	2,950	500
Set D (50)	1,400	2,950	500
Set E (50)	1,600	3,000	500
Set F (50)	1,700	3,000	500
Set G (50)	1,600	3,100	500
Set H (50)	1,600	3,100	500
Set I (50)	1,400	3,050	500
Set J (53)	1,272	3,233	530
Total [Number of speakers]	15,372 [306]	30,283 [301]	5,030 [100]

Table 4: Age and sex distribution of JNAS speakers.

Age	**Male**	**Female**	**Total**
10-19	1	0	1
20-29	90	81	171
30-39	40	47	87
40-49	11	16	27
50-59	5	5	10
60+	5	3	8
Age unknown	1	1	2
Total	153	153	306

Table 5: Age and sex distribution of S-JNAS speakers.

Age	**Male**	**Female**	**Total**
60-64	47	52	99
65-69	49	46	95
70-74	39	35	74
75-79	11	14	25
80-84	4	2	6
85-89	1	0	1
90-94	0	1	1
95-99	0	0	0
Total	151	150	301

2.5 Emotion labels

Analysis of elderly speech is indispensable for the development of dialogue systems and dialogue interfaces for elderly people. In addition to the content being conveyed by speech, the speaker’s emotions are also part of the messages being conveyed, particularly during dialogues, so information about speaker emotion is increasingly being incorporated into dialogue control in dialogue systems. However, studies examining the emotional content of speech generally focus on the adult speech of subjects of relatively young ages, and little research has been done on the emotional content of the speech of the elderly. Therefore, our corpus includes emotion labels for the speech of our elderly speakers.

How to apply emotion labels to speech is an important task, and descriptive methods have largely been divided into models based on basic emotional perspectives and models based on dimensional perspectives of emotion. For our corpus, we adopted the eight basic emotions identified by Plutchik (Plutchik, 2001) which include the continuums of "joy - sadness", "acceptance - disgust", "fear - anger" and "surprise - anticipation", which we translated into Japanese. After making each speech recording, staffs with experience working with the elderly elicited information from the speakers about their emotional state. The pairs of adjectives were then used to label the emotional states of the speakers. Then, a five-point evaluation was performed on the speaker’s adjective pairs. For example: emotion label of Subject 1 is a very joy, a little acceptance, a weak fear and a little surprise.

2.6 Dementia testing using HDS-R

As Japanese society rapidly ages, the number of people with dementia in the country increases each year. In July 2008, the Ministry of Health, Labor and Welfare launched an "emergency project to raise the quality of medical care and improve daily life for citizens with dementia." Early diagnosis of dementia is listed as one of measures to be taken, so it has been studied through the analysis of physical movement, brain wave measurement, cognitive assessments, etc. However, there have been few studies on the relationship between speech and dementia. For this reason, we conducted a simple diagnostic test used to measure dementia risk with each speaker after recording their speech so that we could analyze possible characteristics of elderly speech which might be linked with dementia.

Two widely used cognitive assessments used to detect dementia are the revised Hasegawa’s Dementia Scale (HDS-R) (Imai, 1994) and the Mini-Mental State Examination (MMSE) (Fostein, 1975). The HDS-R consists of nine questions on age, time, place, memory, calculation, etc., and has an evaluation scale of 0 to 30 points. The MMSE examination consists of eleven questions such as time, place, memory, calculation, writing, drawing figures, etc., and also scores subjects on a scale of 0 to 30. Both tests indicate a tendency towards dementia if a subject’s score falls below a given threshold. The HDS-R only detects the presence or absence of a tendency towards dementia, while the MMSE ranks the severity of a subject’s dementia based on their score. For this study, we adopted the simpler HDS-R. Each subject who provided speech for our corpus was scored using the HDS-R at a later date. As a result, 11 out of the 100 original participants were judged to have a tendency towards dementia.

3. Japanese Speech Corpora

3.1 Comparison of Speech Corpora

Sets of phrases are selected for speech corpora so that the result is a collection of read speech suitable for training acoustic models for a wide variety of speech-driven applications, including dictation. Although elderly speech corpora for various languages exist (Cucchiarini, 2006; Hamalainen, 2012), in this paper we focus on Japanese corpora. The JNAS corpus is typically used to construct Japanese acoustic models for standard adult speech data. Here we compare our new, elderly speech database with the JNAS database and the S-JNAS elderly speech database. The JNAS database includes 306 speakers, with each speaker uttering about 50 ATR phoneme-balanced sentences, while the S-JNAS includes 301 speakers, with each speaker uttering two sets of ATR phoneme-balanced sentences (around 100 sentences). The age distributions of the speakers in the JNAS and S-JNAS corpora are shown in Tables 4 and 5, respectively, and the number of utterances of each sentence in each of the three corpora is shown in Table 3.

Table 6: Recording equipment.

Database	Microphone	Recorder
JNAS	Desktop microphone: Sony ECM530, etc. Headset microphone: Senhhizer HMD410 & HMD25-1	Not given
S-JNAS	Desktop microphone: Sony ECM530 Headset microphone: Senhhizer HMD25-1	DAT PCM-R500
New corpus	Desktop microphone: Audio-Technica AT9930	TASCAM DR-05 VERSION2

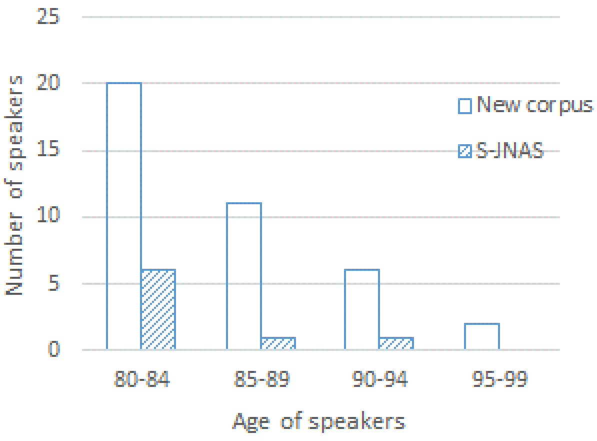


Figure 2: Age distributions of new corpus and S-JNAS.

The majority of the speakers in the JNAS corpus were from 20 to 39 years old, while in the S-JNAS corpus most of the speakers ranged from 60 to 69 years old, with an average age of 67.6 and equal numbers of male and female speakers.

The JNAS speech data was recorded at 39 facilities, mostly universities and research institutes. The S-JNAS speech data was recorded at two facilities in Nara Prefecture which were not elderly facilities. Table 6 shows the type of recording equipment used to record each corpus. The microphones and recorders used to record both the JNAS and S-JNAS varied depending on the recording facility. For our corpus, we used the same microphone and recorder with all of our subjects.

3.2 Comparison of Speech Duration

We conducted Voice Activity Detection (VAD) based on the volume level of all of the speech data in each corpus, and then calculated the average speaking rate for each corpus by dividing the total duration of the speech in each corpus by the total number of morae in each corpus. The results, presented in Table 7, reveal that the average speaking rate in the JNAS corpus corresponds to the average rate of speech of typical Japanese utterances (Han,

Table 7: Average speaking rate for each corpus.

Database	Total duration of speech [sec.]	Average speaking rate [mora/sec.]
JNAS	64,403.0	7.66
S-JNAS	176,824.4	5.44
New corpus	33,008.0	4.98

Table 8: Average speaking rates in new corpus for each age group.

Age	**Average speaking rate [mora/sec.]**	**Standard Deviation**
60-69	6.21	0.74
70-79	5.89	1.14
80-89	5.65	1.15
90-99	4.88	0.83

1994). Regarding the age distribution, speakers aged over 80 accounted for only 3% (8 persons) of the speakers in the S-JNAS corpus, while our new corpus includes 39 subjects over 80, which represents about 40% of the speakers. Figure 2 compares the number of subjects in each age group in the S-JNAS corpus and in our new corpus. The average age of speakers in the new corpus is approximately 10 years older than in the S-JNAS corpus, and the average speaking rate in the new corpus is also slower than in the S-JNAS corpus.

Table 8 shows the rate of speech for each age group in our corpus. We can see how the rate of change in speaking rates increases when we compare the speech of speakers 60 to 79 years old with the speech of speakers older than 80. Although differences in age-related changes between individuals are large, we can clearly see from these results that aging causes a decline in speaking rates, and that this change becomes especially noticeable when the speakers are over 90 years old.

In this paper we compared the average speaking rates of each corpus by calculating total speech time. However, although duration of one mora is almost constant in the Japanese language, the duration of speech changes slightly depending on the position of the phonemes (Ota, 2003). Therefore, in order to accurately calculate speech duration, it will be necessary to precisely calculate the duration of each mora. Moreover, there is a possibility that noise in the recorded speech affected the VAD process, so it may also be necessary to improve our VAD technique.

4. Speech Recognition Experiments

By examining the spectral features of speech, it becomes clear that there are various differences between the speech of older and younger adults besides speech duration, and that these differences are likely to affect speech recognition performance, the improvement of which is the goal of our research. First, we examine speech recognition performance by conducting speech recognition experiments using JNAS and S-JNAS utterances with acoustic models constructed using JNAS and S-JNAS data. We then use 30 utterances from our new corpus, consisting of acoustically balanced sentences read by elderly persons, as our test data. As a comparison, we also used 200 utterances from JNAS and 10,313 utterances from S-JNAS, which are newspaper article sentences, as test data. To train our acoustic models, we used training scripts based on the CSJ recipe in the Kaldi speech recognition toolkit (Povey, 2011). For our language model, we only used the sentences included in the training data, thus there was a very strong linguistic constraint.

Recognition results are shown in Table 9. By comparing the various approaches, we can see that state-of-the-art Deep Neural Network - Hidden Markov Model (DNN-HMM) acoustic models clearly do improve recognition performance. When using JNAS and S-JNAS data for both training and evaluation, we can see from our results that matched acoustic models are very effective for obtaining good performance. By comparing the effect of using the JNAS versus the S-JNAS corpus for training, we can also see that use of the S-JNAS data improves recognition performance for our elderly target speakers. However, even when we use the S-JNAS corpus, recognition performance for the speech of our target speakers are still insufficient. We may be able to improve recognition performance by using our new corpus as training data, but the quantity of data is currently insufficient to train DNN-HMM acoustic models.

Table 9: WERs (%) when using various data sources during speech recognition testing. Note that the language model constraints were very strong, thus the WERs are much lower (better) than for conventional approaches.

Train	JNAS	JNAS	S-JNAS	JNAS	S-JNAS
Eval.	JNAS	S-JNAS	S-JNAS	Ours	Ours
GMM(1)	6.00	14.35	8.27	37.67	26.28
GMM(2)	2.79	6.79	4.05	20.47	15.35
DNN(1)	2.38	4.74	3.60	19.77	13.02
DNN(2)	2.34	4.92	3.41	23.72	13.49

GMM (1): GMM-HMM (LDA+MLLT)
GMM (2): GMM-HMM (LDA+MLLT+SAT+MMI+fMMI)
DNN (1): DNN-HMM (Cross-entropy optimization)
DNN (2): DNN-HMM (state-level Minimum Bayes Risk (sMBR) w/ lattice regeneration)

Table 10: WERs (%) for speech recognition experiments using utterances of elderly speakers and BCCWJ language model (w/o acoustic adaptation).

Acoustic models		**JNAS**	**S-JNAS**	**CSJ**
Test data	Read	55.50	39.45	59.17
	Dialog	64.64	67.52	45.28

Table 11: WERs (%) for speech recognition experiments using utterances of elderly speakers and BCCWJ language model (w/ acoustic adaptation).

Original acoustic models		**JNAS**	**S-JNAS**	**CSJ**
Test data	Read	44.61	38.30	34.06
	Dialog	67.98	70.67	47.83

Next, we investigated recognition performance under more realistic conditions using a more general language model with DNN-HMM acoustic models. We constructed a language model using the Balanced Corpus of Contemporary Written Japanese (BCCWJ), and in addition to JNAS and S-JNAS we also used the CSJ to train the acoustic models, in order to investigate the effect on performance when recognizing spontaneous speech.

Note that we also collected additional samples of speech from 39 elderly people (13 males and 26 females) in

Tokushima, under recording conditions similar to those used during our collection of data in Nagoya. We then included this data in the training data. As for test data, we used newly recorded utterances of five elderly persons reading newspaper articles aloud, and utterances extracted from dialogs between each of eight elderly persons and an interviewer.

Experimental results are shown in Table 10. From these results we can see that recognition of speech generated by elderly speakers is a very difficult task. Results when using the S-JNAS read speech data for training the acoustic model were much better than when using the JNAS read speech data, which is understandable, but even the use of S-JNAS data was not effective for the recognition of elderly dialog speech. In contrast, the CSJ trained acoustic models were more effective for dialog speech recognition, indicating that matching the speaking styles of training and test data is very important, even if there is a "generation gap" between the speakers.

We also used acoustic models adapted with transfer training, using JNAS, S-JNAS, and CSJ models, respectively, as the original models and then applying additional back-propagation to the models using the read speech of elderly people from our new corpus. Results are shown in Table 11. In the case of read speech, performance of all of the original acoustic models were improved, however this was not the case for dialog speech. The adaptation data was read speech, and thus it was very effective for improving performance with read speech, by "bridging the gap" of the various generational differences discussed in the previous sections. As for dialog speech, we could not achieve any improvements, even when the original acoustic models were trained using the CSJ data. These results indicate that mismatches between test data and training/adaptation data are highly detrimental to recognition accuracy.

Although we have not achieved acceptable performance in our attempt to develop a dialog recognition system for the elderly, we have learned valuable lessons, and we think that there may be two other ways to tackle this problem:

Collection of spontaneous elderly speech for model training.

Development of a new adaptation method which does not corrupt important characteristics of the original acoustic models, for example, corruption of the spontaneity of the CSJ trained models.

In our future work, we will use these two techniques to improve our method, but the first technique is very costly, thus we will mainly concentrate on the second technique.

5. Conclusions

In this study we attempted to improve recognition of the speech of elderly Japanese by spoken dialog systems through the creation a superior corpus of elderly Japanese speech. Experimental evaluation of our new corpus, using a variety of ASR systems and techniques, showed that it was not effective for improving elderly speech recognition performance. However, we did confirm that using elderly speech when developing speech recognition systems for the elderly is effective, and that matching the speaking styles of the training and test data for the language model (read speech vs. spontaneous speech vs. dialog speech) is also very important. In addition, we discovered that it is important not to corrupt the acoustic model during adaptation. We also confirmed that state-of-the-art DNN-HMM acoustic models improve speech recognition performance. Finally, we consider our process for developing a corpus of elderly speech to have been largely successful, as we were able to collect a sizeable corpus of high-quality read speech at a low cost, from a subset of the population that is relatively difficult to engage.

6. Acknowledgements

This work was partially supported by the Center of Innovation (COI) Program of the Japan Science and Technology Agency (JST), the Strategic Information and Communications R&D Promotion Program (SCOPE) of the Ministry of Internal Affairs and Communications of Japan, and JSPS KAKENHI Grant-in-Aid for Scientific Research (B) Grant Number 17H01977.

7. References

Acartürk, C., Freitas, J., Fal, M., Dias, M.S., (2015). Elderly Speech-Gaze Interaction: State of the Art and Challenges for Interaction Design, Universal Access in Human-Computer Interaction. Access to Today's Technologies, Volume 9175 of the series Lecture Notes in Computer Science, pp 3-12.

Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B., Hudson, R., (1999). Recognition of elderly speech and voice-driven document retrieval. In Proc. of ICASSP.

Baba, A., Yoshizawa, S., Yamada, M., Lee, A., Shikano, K., (2001). Elderly Acoustic Model for Large Vocabulary Continuous Speech Recognition. In Proc. of EUROSPEECH 2001.

Cucchiarini C., Van hamme, H., van Herwijnen, O., Smits, F., (2006). JASMIN-CGN: Extension of the spoken Dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality. In Proc. of International Conference on Language Resources and Evaluation, pp. 135-138

Fostein, M.F., Fostein, S.F., McHugh, P.R., (1975). MiniMental State: A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res, vol. 12, pp. 189-198.

Hamalainen, A., Pinto, F.M., Dias, M.S., Júdice, A., Freitas, J., Pires, C.G., Teixeira, V.D., Calado, A., Braga, D., (2012). The First European Portuguese Elderly Speech Corpus. In Proc. of IberSPEECH 2012.

Han, M.S., (1994). Acoustic manifestations of mora timing in Japanese, Journal of the Acoustical Society of America, vol. 96, no. 1, pp. 73-82.

Imai Y., Hasegawa K., (1994). The Revised Hasegawa's Dementia Scale (HDS-R): Evaluation of its usefulness as a screening test for dementia. Hong Kong Coll Psychiatr, vol. 4, pp. 20-24.

Iso, K., Watanabe, T., Kuwabara, H., (1988). Design of a Japanese Sentence List for a Speech Database, Preprints, Spring Meeting of Acoustic Society of Japan, Paper 2-2-19, pp. 89-90 (in Japanese).

Júdice, A., Freitas, J., Braga, D., Calado, A., Dias, M., Teixeira, A., Oliveira, C., (2010). Elderly Speech Collection for Speech Recognition Based on Crowd Sourcing. In Proc. of DSAI2010, pp. 103-110.

Kurematsu, A., Takeda, K., Sagisaka, S., Katagiri, S., Kuwabara, H., Shikano, K., (1990). ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis, Speech Communication, vol. 9, pp. 357-363.

Maekawa, K., Koiso, H., Furui, S., Isahara, H., (2000). Spontaneous speech corpus of Japanese, In Proc. of LREC2000, pp. 2013-2018.

Ota, M., Ladd, D.R., Tsuchiya, M., (2003). Effects of foot structure on mora duration in Japanese?, In Proc. of 15th International Congress of Phonetic Science (ICPhS-15), pp. 459-462.

Plutchik, R., (2001). The nature of emotions, American Scientist, 89(4), pp. 344-350.

Povey D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., and Schwarz, P., (2011). The Kaldi Speech Recognition Toolkit, Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

Vipperla, R., Renals, S., Frankel, J., (2008). Longitudinal study of ASR performance on ageing voices. In Proc. of Interspeech 2008.

Winkler, R., Brückl, M., Sendlmeier, W.,(2003). The aging voice: an acoustic, electroglottographic and perceptive analysis of male and female voices. In: Proc. of ICPhS 03, Barcelona, pp. 2869-2872.

Intonational Variations at the End of Interrogative Sentences in Japanese Dialects: From the “Corpus of Japanese Dialects”

Nobuko Kibe, Tomoyo Otsuki, Kumiko Sato

National Institute for Japanese Language and Linguistics (NINJAL)
{nkibe, otsukit, satok}@ninjal.ac.jp

Abstract

In general, it is said that interrogative sentences have a final rising intonation (Kori 2003). However, this rule is not true of some Japanese dialects. Kibe (2010, 2011,2013) classify sentence-final tones of interrogatives in Japanese dialects into four types: Type A as a rising tone (Tōkyō dialect), Type B as a falling tone (Hirosaki dialect, Kagoshima dialect), Type C as a rising/falling tone (Hiroshima dialect), and Type D as a gradual rising tone (Fukuoka dialect). Since the data in Kibe (2010, 2011 and 3013) were extracted from an existing nation-wide dialect survey where an elicitation task was employed, it is not clear whether how much such intonation patterns appear in a spontaneous speech in each region. This article examines sentence-final tones of interrogatives extracted from a natural discourse stored in the “Corpus of Japanese Dialects” (COJADS), which is currently in preparation for release by the National Institute for Japanese Language and Linguistics (NINJAL). The results revealed that the four types are attestable even in a natural discourse, and furthermore, we identified a dialect such as Hirosaki dialect which distinguishes interrogatives from declaratives by the pitch range in the final falling tone.

Keywords: rising tone, falling tone, wh-question, yes/no question, Mito dialect, Hirosaki dialect, Kagoshima dialect

1. Introduction

In general, it is said that interrogative sentences have a final rising intonation. Modern standard Japanese also uses a rising tone at the end of wh-questions and yes/no questions. However, this rule is not true of some Japanese dialects. For instance, in the Hirosaki dialect, spoken in Aomori Prefecture in the northwestern part of Japan, and the Kagoshima dialect, spoken in Kagoshima Prefecture in the southern part of Japan, a falling tone appears both in wh-questions and yes/no questions. In some dialects, like Matsumoto dialect, spoken in the Chūbu area, and the Hiroshima dialect, spoken in the Chūgoku area, a final rising tone is used for wh-questions but a falling tone is used for yes/no questions.

In the previous studies, the following four patterns were identified as sentence-final tones of interrogatives in Japanese dialects (Kibe 2010, 2011,2013).

Type A: Rising tone <Tōkyō dialect>

> Interrogative sentences are pronounced with a final rising tone, regardless of whether any interrogative word or interrogative final particle is used or not.

Type B1: Falling tone <Hirosaki dialect>

> Interrogative sentences are pronounced with a falling tone, regardless of whether any interrogative word or interrogative final particle is used or not.

Type B2: Falling tone – final particle <Kagoshima dialect>

> Interrogative sentences always contain an interrogative word form or interrogative final particle, and such sentences are pronounced with a final falling pitch.

Type C: Rising tone / falling tone < Hiroshima dialect>

> Interrogative sentences with an interrogative word form are pronounced with a final falling pitch, and sentences without an interrogative word form are pronounced with a final rising tone.

Type D: Gradual rising tone <Fukuoka dialect>

> Interrogative sentences are pronounced with a gradual rise starting from an interrogative word at the beginning to the end of the sentence.

This report provides an overview of rising and falling tones at the end of interrogatives in Japanese dialects. We do not discuss Ryukyuan languages since further research is needed.

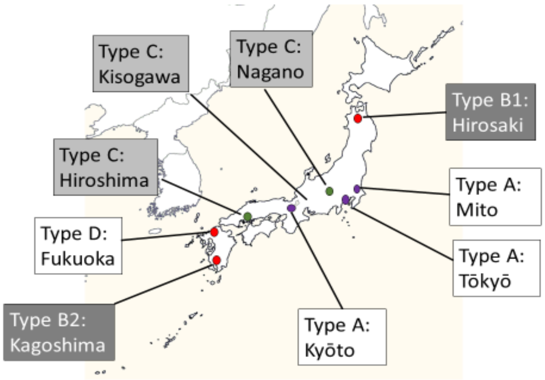


Figure 1: Sites where the data used in this report were collected

2. Data

The data of this article was obtained from recorded conversations in “*Kakuchi Hōgen Shūshū Kinkyū Chōsa*” (Urgent surveys to collect data of dialects throughout the nation), which was led by the Agency for Cultural Affairs, Governments of Japan, and “*Nihongo Onsei*” (Japanese Prosody, JP), which stores recordings of Japanese dialects as audio DVDs (Grant-in-Aid for Scientific Research from 1989 to 1992). The recording of “*Kakuchi Hōgen Shūshū Kinkyū Chōsa*” was conducted from 1977 to 1985 at 224 sites throughout the nation (A part of the data is published by *Zenkoku Hōgen Danwa Dētabēsu: Nihonno Furusato Kotoba Shūsei vol.1 ~ vol.20* (Speech Database of Japanese Dialects: Collection of Japanese Dialects)). The data is currently stored at NINJAL and became searchable using the corpus COJADS (Corpus of Japanese Dialects), which we used for data extraction.

3. Analysis on final tones in Japanese dialects

3.1 Tōkyō and Mito Dialect (Type A)

Tōkyō and Mito dialects have a final rising tone both in wh-questions and yes/no questions. Examples in the following sections are such interrogative sentences. (The shaded part is an interrogative word form or an interrogative final particle. ↑ represents a rising tone, ↓ represents a fallig tone.)

3.1.1 Tōkyō dialect

The following are examples of the pitch patterns in interrogatives in the Tōkyō dialect. The sentence in (1a) is a wh-question, and the sentence in (1b) is a yes/no question, both of which are usually pronounced with a rising tone.

(1a) *nani=ŋa hosii* ↑
what= ACC want
"What do you want?"

(1b) *nanika hosii* ↑
anything want
"Do you want anything?"

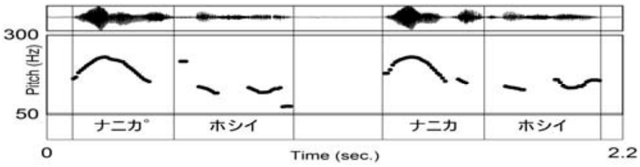


Figure2:

"What do you want?" "Do you want anything?"
(from JP)

If the former sentence is pronounced with a falling pitch, as in (2a), it can be interpreted as a question with a modal meaning such as reproach or antipathy (Moriyama 1989). For example, (2a) can convey impatience, implying "I bought this and that for you. What more do you want?" If the sentence in (1b) is pronounced with a falling pitch, as in (2b), it simply means "I (the speaker) want something."

(2a) *nani=ŋa hosii* ↓
what= ACC want
"What do you want?"(with reproach or antipathy)

(2b) *nanika hosii* ↓
anything want
"I want something."

It is often said that the final particle =*ka* used in the Tōkyō dialect conveys an interrogative meaning. In reality, however, the final particle =*ka* can have an interrogative meaning only when it appears with a rising tone at the end of a sentence. If a sentence with the final particle =*ka* is pronounced with a final falling pitch, the sentence conveys the speaker's acceptance of the fact stated in the sentence, as exemplified below:

(3a) *hanako=to kjo:to=e it-ta=ka*↑
Hanako=COMIT Kyōto=ALL go-PST=SFT.Q
"Did you go to Kyōto with Hanako? "

(3b) *hanako=to kjo:to=e it-ta=ka*↓
Hanako=COMIT Kyōto=ALL go-PST=SFP
"You did go to Kyōto with Hanako."

The sentence with a rising tone in (3a) is an interrogative sentence asking whether the addressee went to Kyōto with Hanako. Contrastively, the sentence with a final falling pitch in (3b) can be used in a case where the speaker mutters to himself/herself and reluctantly accepts the fact that the addressee went to Kyōto with Hanako. In the Tōkyō dialect, therefore, =*ka* is not necessarily used to convey an interrogative meaning. An interrogative sentence is made by adding a final rising intonation to =*ka*, rather than just by adding =*ka* at the end. Thus, in yes/no questions in the Tōkyō dialect, including sentences ending with =*ka*, an interrogative meaning can be conveyed only by a sentence-final rising tone.

The same applies to wh-questions. Having an interrogative word form in a wh-question clearly marks the sentence as interrogative. Therefore, such a sentence should not need to be pronounced with a final rising tone. (In English, a final falling pitch is used in wh-questions, like "What is this?") In the Tōkyō dialect, however, a final rising tone is generally used in wh-questions. Thus, in this dialect, a final rising tone serves as an interrogative marker. Conversely, a sentence that is not pronounced with a final rising pitch is regard as a non-interrogative sentence or an interrogative with a special meaning as seen in (2a).

3.1.2 Mito dialect

Interrogatives in Mito dialect are pronounced with a rising tone in the same way as the Tōkyō dialect. The following examples show the pitch patterns in Mito dialect. The sentence in (4a) is a wh-question, and (4b) is a yes/no question, both of which are usually pronounced with a rising tone.

(4a) *maa arenani=sa it-ta=ɴ=dak=ke* ↑
FILL that what=ACC good-PST=NMLZ=COP=SFP.Q
"Well, what is that good for?"

(4b) *are siɴkeetuu=ke*↑
that neuralgia=SFP.Q
"Is that neuralgia?"

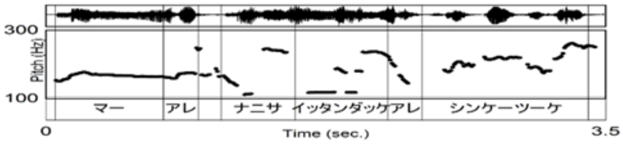


Figure 3:
"Well, what is that good for?" "Is that neuralgia?"
(from COJADS)

Tables 1 and 2 below show sentence-final pitch patterns of wh-questions and yes/no questions in Mito dialect in COJADS. "A" and "B" in the column "ID" represent individual speakers (A is male and B is female in this example. recording time: 38'50 ").

	final tone			total
	rising	falling	rising faling	
WH	3	0	1	4
Y/N	4	0	0	4

Table 1: Number of pitch patterns (Mito dialect)

ID Utterance	Question Type	Sentence-final tone
(1.1) 17A *naɴ=tyuu=ke* what=QUOT=SFP.Q "What do we call that?"	WH	↑
(1.2) 142A *naɴ=tuu=ke* what=QUOT=SFP.Q "What do we call that?"	WH	↑
(1.3) 244B *doɴna kusa=desu=ka* what grass=COP.HON=SFP.Q "What kind of grass is it? "	WH	↑
(1.4) 488B *naɴ=tuu=ke=ne* what=QUOT=SFP.Q=SFP "What do we call that?"	WH	↑↓
(1.5) 203A *siɴkeetuu=ke* neuralgia= SFP.Q "Is that neuralgia? "	Y/N	↑
(1.6) 263B *sore=wa tiŋai=masu=ka* that=TOP different=HON=SFP.Q "Is it wrong? "	Y/N	↑
(1.7) 414A *kikime=ŋa haee=ɴ=da nee= ge* effect=NOM fast=NMLZ=COP NEG=SFP.Q "Doesn't it work fast? "	Y/N	↑
(1.8) 1097A *siti+hati+neɴ=ni nak=ka=na* 7+8+years=DAT become=SFP.Q=SFP "Has it been 7 or 8 years? "	Y/N	↑↓

Table 2: Examples of interrogative sentences (Mito dialect)

Table 2 shows that interrogatives in Mito dialect always has the sentence-final particle *=ke* and the *=ke* is always pronounced with a rising pitch. In addition, other particles can follow *=ke* , such as *=ne* (1.4) and *=na* (1.8) which are used to address listeners. In this case, *=ne* or *=na* is accompanied with a falling tone.

3.2 Hirosaki Dialect (Type B1)

The Hirosaki dialect has a final falling tone both in wh-questions and yes/no questions. The following are examples of such interrogative sentences.

(5a) *doosute sono mada rosuya+zuɴ*
why FILL FILL Russian[NOM]

e-su-ta=ba ↓
be-HON-PST=SFP
"Why was there a Russian?"

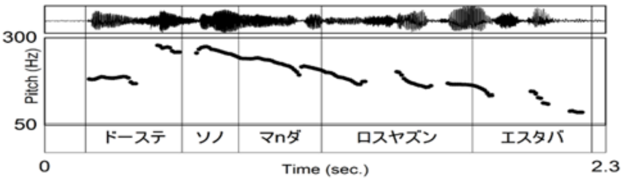


Figure 4: "Why was there a Russian?" (from COJADS)

(5b) *beɴzya=sa tumakawa*
clogs =ALL top.cover[NOM]

tude ae-su-ta=gaa ↓
attach.SEQ be-HON-PST=SFP.Q

"Does the beɴzya (a kind of Japanese clogs with iron blades) have a top cover?"

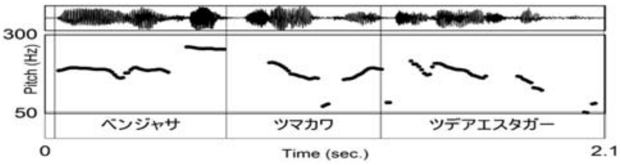


Figure 5: "Does the beɴzya have a top cover?" (from COJADS)

The sentence in (5a) is a wh-question, and (5b) is a yes/no question, both of which are usually pronounced with a falling tone. As discussed in Section 3.1, Tōkyō and Mito dialects always have a sentence-final rising tone in interrogatives; conversely, The Hirosaki dialect usually has a falling tone in interrogatives. Tables 3 and 4 show sentence-final pitch patterns of wh-questions and yes/no questions in the Hirosaki dialect in COJADS (A and B are male, and C is female. recording time: 36'26 ").

	final tone		total
	rising	falling	
WH	1	3	4
Y/N	1	5	6

Table 3: Number of pitch patterns (Hirosaki dialect)

ID Utterance	Question Type	Sentence-final tone
(2.1) 011C *doosute sono mada rosuya+zuɴ* why FILL FILL Russian[NOM] *e-su-ta=ba* be-HON-PST=SFP "Why was there a Russian?"	WH	↓
(2.2) 051B *doyate nomu=ɴ=zu* how drink=NMLZ=SFP "How do you drink it?"	WH	↓
(2.3) 036B *naɴ=te su-ta=moɴ=da=be* what=QUOT say-PST=NMLZ=COP=INFR "How would we say that?"	WH	↓

(2.4) 108C *nani* *kimono* *ki-tera* what clothes[ACC] wear-PROG "What kind of kimono do you wear?"	WH	↑
(2.5) 063C *neɴde* *ae-su-ta=ga* exist.NEG.SEQ be-HON-PST=SFP.Q "Isn't it there? "	Y/N	↓
(2.6) 076C sa*ɴbu=desu=ga* 3/8.intches=COP.HON=SFP.Q "Is it sa*ɴbu* (3/8 inches)?"	Y/N	↓
(2.7) 098C *beɴzya=sa* *tumakawa* clogs=ALL top.cover[NOM] *tude* *ae-su-ta=gaa* attach.SEQ be-HON-PST=SFP.Q "Does the beɴzya have the top cover?"	Y/N	↓
(2.8) 024A a*ɴta-dazu* *waga-ne=ga* 2-PL[NOM] know-NEG=SFP.Q "Don't you understand it?"	Y/N	↓
(2.9) 055C *tama-ko* *tusegu* *naru=* ball-DIM[NOM] small become= *ɴ=de* *heɴ=be* NMLZ=COP.SEQ NEG.HON=INFR "Is the ball not getting smaller?"	Y/N	↓
(2.10) 114C *paɴtu* *hai-deraa* underpants[ACC] wear-PROG "Are you wearing underpants?"	Y/N	↑

Table 4: Examples of interrogative sentences (Hirosaki dialect)

The sentence-final particle *=ga* (*gaa*) in examples (2.5) - (2.8) in Table 4 marks the sentences as interrogative. Table 4 indicates that, in the Hirosaki dialect, a wh-question contains a wh-phrase in a sentence, and a yes/no question is marked by the particle *=ga* (*gaa*)*,* both of which are generally accompanied with a falling tone.

However, the examples (2.9) and (2.10) in Table 4 are interrogative without containing a wh-phrase or the particle *=ga* (*gaa*) (The sentence-final particle *=be* in the example (2.9) indicates an assumption, and *-deraa* in the example (2.10) is a morpheme to express an aspect (progressive), and thus, neither of them does not mark a sentence as interrogative.) It suggests that these sentences are marked as interrogative by the final falling tone. (The example (2.10) is pronounced with a final rising tone. This sentence can be interpreted as a question with a modal meaning.)

In the Hirosaki dialect, a declarative sentence is pronounced with falling tone too. How is a yes/no question distinguished from a declarative sentence in the Hirosaki dialect where both are pronounced with a falling pitch? The following sentence-final pitch patterns are identified from examining examples in COJADS:

(6) There is a difference in the pitch range of falling between wh-questions, yes/no questions, and declarative sentences, whereby yes/no questions have the largest falling and declaratives have the smallest (Figure 6, 7, and 8).

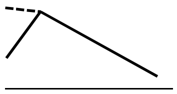


Figure 6: wh-question

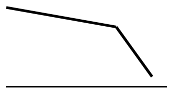


Figure 7: yes/no question

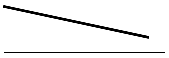


Figure 8: declarative

Table 5 shows the average pitch ranges as falling for each sentence type.

Sentence Type	Average pitch range as falling (Hz/s)
wh-question	152.696
yes/no question	170.098
declarative	**129.41**

Table 5: Difference in the degree of falling between sentence types (Hirosaki dialect)

As Table 5 shows, yes/no questions have a large pitch fall in comparison to declarative sentences. We conducted a t-test in order to verify whether the difference is statistically significant. For the purpose of normalizing the difference between f0 values of the sentence types across speakers, we calculated z-scores of f0 values for each speaker, and then calculated the difference between the highest and lowest (z-scored) f0 values. The result of the t-test confirmed that there is a statistical significant difference in the z-scored pitch range between yes/no questions and declaratives with a significance level of 5% (p = 0.019). This result indicates that the Hirosaki dialect distinguishes interrogatives from declaratives by the pitch range.

Although the data is not abundant and conditions of examples found in a natural discourse are difficult to control, the current data from the COJADS indicates that the Hirosaki dialect distinguishes interrogatives and declaratives by the pitch range. Previous studies have reported that dialects in the Tohoku region use a final falling pitch for interrogatives (Yamaura 2000, Kibe 2010), but this is the first study that reveals the pitch range of falling in the Hirosaki dialect has a distinctive feature. Further research is needed to identify where the falling pitch starts in interrogatives and conduct a quantitative analysis.

3.3 Kagoshima Dialect (Type B2)

In the Kagoshima dialect, both wh-questions and yes/no questions are pronounced with a falling pitch as shown in the following (Kibe 1997):

(7a) *nai=ga* *hoɕika=ka*↓
what= ACC want=SFP.Q
"What do you want? "

(7b) *naika* *hoɕi=ka*↓
anything want=SFP.Q
"Do you want anything? "

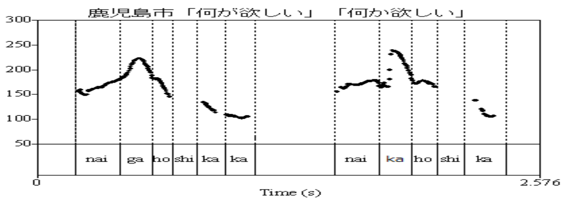


Figure 9:
"What do you want?" "Do you want anything?"
(from JP)

(8a) *dai=to* *kjo:to =i* *it-ta=ka:*↓
who=COMIT Kyōto=ALL go-PST=SFP.Q
"Who did you go to Kyōto with?"

(8b) *hanako=to* *kjo:to=i* *it-ta=to=ja*↓
Hanako=COMIT Kyōto=ALL go-PST=NMLZ=SFP.Q
"Did you go to Kyōto with Hanako?"

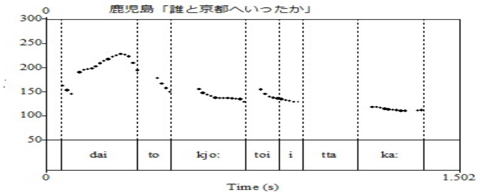


Figure 10: "Who did you go to Kyōto with?"
(from JP)

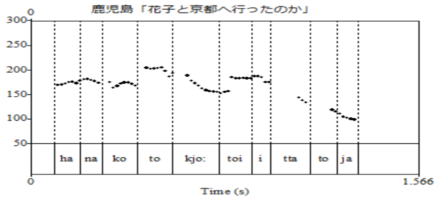


Figure 11: "Did you go to Kyōto with Hanako?"
(from JP)

The sentences in (7a) and (8a) are wh-questions, and (7b) and (8b) are yes/no questions. The sentence-final particle *=ka* in (7a), (7b), and (8a) and the particle *=ja* in (8b) mark the sentences as interrogative. These sentences are all pronounced with a final falling pitch.

Tables 6 and 7 are lists of all wh-questions and yes/no questions in the Kagoshima dialect in COJADS (A and B are male, and C is female. recording time: 34'29 "), which show that the dialect always has an interrogative sentence-final particle (i.e. *=ka* (*ga*)*,* *=ke* (*ge*), *=na*). A sentence without a sentence-final particle in interrogatives (e.g. *kjo:to*↑ "Kyōto? ", *i ku*↑ "You go") is unnatural in this dialect. The different particles reflect a type of relationship between a speaker and a listener; *=na* is used when a listener is older than a speaker, *=ke* (*ge*) is used when a listener is younger than a speaker, and *=ka* (*ga*) can be used for any relationship.

Tables 6 and 7 shows that most of the sentences are pronounced with a final falling pitch.

	final tone		total
	rising	falling	
WH	2	6	8
Y/N	0	7	7

Table 6: Number of pitch patterns (Kagoshima dialect)

ID Utterance	Question Type	Sentence-final tone
(3.1) 098A *naɴ+gak=ka* what+month=SFP.Q "What month?"	WH	↓
(3.2) 038A *taisyoo* Taisho(the era name) *naɴ+neɴ=dzyat-ta=ga* what+year=COP-PST=SFP.Q "What year was it in Taisho era?"	WH	↓
(3.3) 020A *naɴ+neɴ=no* what+year=GEN *koro=yat-ta=ga=nii* time=COP-PST=SFP.Q= SFP "Around which year was it?"	WH	↓
(3.4) 358B *naɴ=tyu=ga* what=QUOT=SFP.Q "What do we say?"	WH	↓
(3.5) 367A *dai=dzyat-ta=ga* who=COP-PST= SFP.Q "Who was it?"	WH	↓
(3.6) 013B *gena huu=dzyat-ta=ga* how appearance=COP-PST=SFP.Q "How was it?"	WH	↓
(3.7) 073C *naɴ+kwai=yay-taro=ga* what+frequency=COP-PST= SFP.Q "How many times is it?"	WH	↑

(3.8) 037B
do=yat-taro=ga
how=COP-PST. INFR=SFP.Q
"How was it?"
	WH	↑

(3.9) 52A *odzi=ŋa mugeme* uncle=NOM meeting *idat togoi=dzya nagat-ta=ga* go time= COP NEG-PST=SFP.Q "Has the uncle just left to pick up (someone)?"	Y/N	↓
(3.10) 58A *zuutto=zyar=a* always=COP.SEQ=TOP *se-ɴ= zyat-ta=ga* do-NEG=COP-PST=SFP.Q "Hasn't it been all the time?"	Y/N	↓
(3.11) 60C *geɴek=kara zuutto* active=ABL always *hito+tyure= zyar=a* one+chain=COP.SEQ=TOP	Y/N	↓

se-ɴ=dzyat-ta=ga do-NEG=COP-PST=SFP.Q “Has it been like this before retirin		
(3.12) 13B *X2=ŋa* X2 (person’s name) =GEN *odot=wa* brother=TOP *X4=dzyat-ta=ke* X4 (person’s name) =COP-PST=SFP.Q “Was the brother of X2 X4?”	Y/N	↓
(3.13) 25A *seɴsisya=wa hutai* war.deaths=TOP two.people *ot=to=na* be=NMLZ=SFP.Q “Were there two died in the war?”	Y/N	↓
(3.14) 54A *modot ki-ta=ge* return.SEQ come-PST=SFP.Q “Has (someone) come back? ”	Y/N	↓
(3.15) 87C *modot-te* return-SEQ *ki-ta=moɴ=dzya* come-PST=NMLZ=COP *naga-do=na* NEG-INFR=SFP.Q “Is it not a case that the person came back?”	Y/N	↓

Table 7:Examples of interrogative sentences (Kagoshima dialect)

3.4 Hiroshima Dialect (Type C)

Next, let’s look at another type of Japanese dialect, which has a different intonation pattern from those of Tōkyō, Hirosaki and Kagoshima dialects.

In the Hiroshima dialect, interrogative sentences that contain any interrogative word forms are pronounced with a final falling pitch, and those without interrogative word forms are pronounced with a final rising tone. Some examples are as follows.

(9a) *dare=to kjo:to=e it-ta=ɴ*↓
who=COMIT Kyōto=ALL go-PST=SFP
“Who did you got to Kyoko with? ”

(9b) *hanako=to kjo:to=e it-ta=ɴ*↑
Hanako=COMIT Kyōto=ALL go-PST=SFP
“Did you go to Kyōto with Hanako? ”

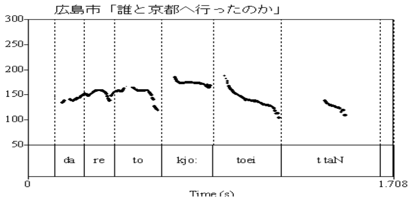


Figure 12: “Who did you got to Kyoko with? ” (from JV)

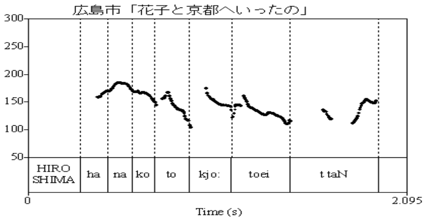


Figure 13: “Did you go to Kyōto with Hanako? ” (from JP)

Please note that the final particle =ɴ at the end of the sentences does not convey an interrogative meaning.

Tables 8 and 9 show sentence-final pitch patterns of wh-questions and yes/no questions in the Hiroshima dialect (A is male, B and C are female, recording time: 38’08 ").

	final tone			total
	rising	falling	rising falling	
WH	1	5	1	7
Y/N	2	6	0	8

Table 8: Number of pitch patterns(Hiroshima dialect)

ID Utterance	Questio n Type	Sentence-final tone
(4.1) 11A *naɴbo naɴ=de nezir-yaa* how.many what=INST twist-COND *ee=ɴ=zya* good=NMLZ=COP “How much should I twist it with what?”	WH	↓
(4.2) 2C *ar=yaa naɴ-gen* that=TOP how.many-CLF *aru=ɴ=desu=ka* be=NMLZ=COP.HON= SFP.Q “How big is it?”	WH	↓
(4.3) 156C *otoosaɴ nani=i suru=ɴ* father.VOC what=DAT do=NMLZ “What do you use it for, Dad?”	WH	↓
(4.4) 65C *hiti+hati yuu-tara* seven+eight say-COND *doɴto=na=ɴ=desu=ka* what=COP=NMLZ=COP.HON= SFP.Q “What is that *hiti-hati* like?”	WH	↓
(4.5) 109C *naɴde ee=no ee=no* why good=SFP good=SFP *yuu-te tataki-yot-ta=ɴ=* say-SEQ crap-PROG-PST=NMLZ= *desu=ka=no* COP.HON=SFP.Q=SFP “I wonder why people say ‘good, good’ when they crap their hands.”	WH	↓
(4.6) 203C *ar=yaa dokoo=no* that=TOP where=GEN *miɴgee=desu=ka* folk.craft=COP.HON= SFP.Q “Where do people make such folk crafts?”	WH	↑

(4.7) 16C *imaa nanyu=u yuu-te* now what=ACC say-SEQ *ee=desu=ka=ne* good=COP.HON=SFP.Q=SFP “What do we call it now? ”	WH	↓↑
(4.8) 53C *tama=a aru=ka* ball=TOP be=SFP.Q “Is there a ball?”	Y/N	↓
(4.9) 60A *ano-gurai=zya nai=ka* that-approximately=COP NEG=SFP.Q “Isn’t that about it?”	Y/N	↓
(4.10) 126B *moo nat-ta=ɴ=* already become-PST=NMLZ= *desu=ka* COP.HON=SFP.Q “Is it registered as a cultural heritage yet?’	Y/N	↓
(4.11) 163B *mai+baɴ mai+baɴ* evry+night every+night *aa yat-te tak-areru=* that.way do-SEQ boil-HON= *ɴ=desu=ka* NMLZ=COP.HON=SFP.Q “Does he boil the bamboos every night?’	Y/N	↓
(4.12) 222B *hurawaa=desu=ka* flower=COP.HON=SFP.Q “Is it Flower Festival? ”	Y/N	↓
(4.13) 240C *son-kurai=gurai=de* that-approximately=approximately=INST *deki=mahyoo=ka* make=HON.INFR=SFP.Q “Can we make it at the cost of approximately that much?”	Y/N	↓
(4.14) 59C *naɴka aru=kai=na* something be=SFP.Q=SFP “Is there anything else?”	Y/N	↑
(4.15) 72A iiya sore=zya=a nai-zya-ro no that=COP.SEQ=TOP NEG-COP-INFR “No, that’s not him, right?”	Y/N	↑

Table 9: Examples of interrogative sentences (Hiroshima dialect)

The sentence-final particle *=ka* marks the sentences as interrogative. Table 9 indicates that, in the Hiroshima dialect, interrogative sentences that contain any interrogative word forms are pronounced with a final falling pitch, and those without interrogative word forms are pronounced with a final rising tone.

3.5 Fukuoka Dialect (Type D)

Interrogative sentences in the Fukuoka dialect are quite different in intonation patterns from those in the other Japanese dialects mentioned earlier. In the Fukuoka dialect, wh-questions have a gradual rising intonation, with an interrogative word placed at the beginning of a sentence pronounced on a low tone and the tone gradually rising toward the end of the sentence. The accents of all the words contained in the sentence are canceled by this gradual rising intonation (Hayata1985, Kubo1990). The sentence in (10a) is a wh-question, and the sentence in (10b) is a yes/no question. Note that the final particle *=na* at the end of these sentences conveys an interrogative meaning.

(10a) ➚ *nani=ga hoshi-ka=na*➚
what= ACC want-NPST=SFP.Q
“What do you want?”

(10b) *naɴka hosi-ka=na*↓
anything want-NPST=SFP.Q
“Do you want anything?”

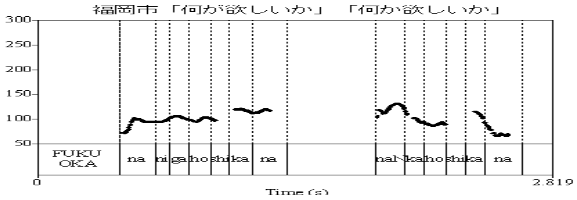


Figure 14:
“What do you want?” “Do you want anything?”
(from JP)

In the Fukuoka dialect, however, when a sentence without an interrogative final particle is pronounced with a final rising tone, the sentence conveys an interrogative meaning. Take a look at the following examples. (Note that the final particle *=to* at the end of the sentence (11b) does not mark the sentence as interrogative.)

(11a) mizu=ba nomu=na↓
Water= ACC drink=SFP.Q
“Do you drink water? ”

(11b) mizu=ba nomu=to↑
Water= ACC drink=SFP
“Do you drink water? ”

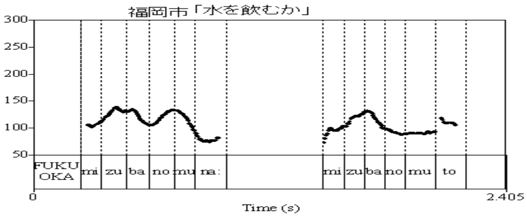


Figure 15: “Do you drink water? ”
(from JP)

The Fukuoka dialect has the following feature as well as the above-mentioned feature; interrogative sentences that contain interrogative sentence-final particles are pronounced with a falling tone, and those with no interrogative sentence-final particles are pronounced with a rising tone.

4. List of Glosses

ACC: accusative
ADD: additive
ADN: adnominal
ADV: adverbial
ABL: ablative
ALL: allative
CAUS: causative
CAUSL: causal
CLF : classifier
COND: conditional
COMIT: comitative
COP: copula
DAT: dative
DIM: diminutive
EVID: evidential
FILL: filler
FOC: focus
GEN: genitive

HON: honorific
INFR: inferential
INST: instrumental
LOC: locative
INST: instrumental
NEG: negative
NMLZ: nominalizer
NOM: nominative
OBLG: obligative
PROH: prohibitive
PROG: progressive
PST: past
PURP: purposive
Q: question/ interrogative
QUOT: quotative
SEQ: sequential
SFP: sentence final particle
VOC : vocative

5. Acknowledgment

The work reported in this article was supported by the NINJAL collaborative research project 'Endangered Languages and Dialects in Japan', and the JSPS Grant-in-Aid for Scientific Research (B) 25284087 (2015～2017), and (A) 16H01933 (2018～2021).

6. Bibliographical References

Hayata, Teruhiro (1985) *Hakata Hogen-no Akusento: Keitairon* [Accent in Hakata dialect: Morphology]. Fukuoka: Kyushu Daigaku Shuppankai.

Kibe, Nobuko (1997) Kagoshima hogen-no intoneishon [Intonation in Kagoshima dialect]. *Shohogen-no Akusento-to Intoneishon* [Accent and intonation in Japanese dialects]. Tōkyō : Sanseido.

Kibe, Nobuko (2010) Intoneishon-no chiikisa: Shitsumon-no intoneishon [Regional difference in intonation: Intonation of question]. *Hogen-no Hakken: Shirarezaru chiikisa-o shiru* [Finding dialects: Detecting unknown regional difference]. Tōkyō : Hituji Syobo.

Kibe, Nobuko (2011) Intonation at the end of interrogative sentences in Japanese dialects, Presentation at ICPP, Kyōto University, 13.Dec.2011.

Kibe, Nobuko (2013) *Soudattanda! Nihongo: Jadde hougen-na omoshitoka* [Aha! Japanese: Dialect is interesting]. Tōkyō : Iwanami Shoten.

Kori, Shiro (2003) Intoneishon [Intonation]. *Asakura Nihongo kouza 3*: *Onsei, On'in* [Asakura Japanese course 3: Phonetics and phonology]. Tōkyō : Asakura Shoten.

Kubo, Tomoyuki (1990) Fukuokashihougen-no toikaeshigimonbun-no akusentogenshou [Accent of echo questions in Fukuoka dialect]. *Nihongo Onsei*: *Kenkyu Houkoku* [Spoken Japanese: research report] 3.

Moriyama, Takuro (1989) Bun-no imi-to intoneishon [Meaning and intonation of a sentence]. *Kouza Nihongo-to Nihongo Kyouiku 1*: *Nihongogaku Gakusetsu* [Course of Japanese and Japanese language education 1: Overview of Japanese linguistics]. Tōkyō : Meiji Shoin.

Yamaura, Harutsugu. (2000) *Kesen-go daijiten*: Zyou [A dictionary of Kesen language: the first volume], Bunpouhen goihen [Grammar and vocabulary]. Sendai: Mumyosha.

7. Language Resource References

Corpus of Japanese Dialects (COJADS). In preparation for release. NINJAL.

Kakuchi Hōgen Shūshū Kinkyū Chōsa (Urgent surveys to collect data of dialects throughout the nation) (1977-1985). Speech database of Japanese dialects. Agency for Cultural Affairs, Governments of Japan.

Nihongo Onsei (Japanese Prosody, JP) (1989-1992). Recordings of Japanese dialects as audio DVDs. Grant-in-Aid for Scientific Research from 1989 to 1992.

Zenkoku Hōgen Danwa Dēta bēse: Nihonno Furusato Kotoba Shūsei (Speech Database of Japanese Dialects: Collection of Japanese Dialects) vol.1-vol.20 (2001-2002). Tōkyō : Kokusho Kankō-kai.

Construction of the Corpus of Everyday Japanese Conversation: An Interim Report

Hanae Koiso[†], Yasuharu Den[‡,†], Yuriko Iseki[†], Wakako Kashino[†], Yoshiko Kawabata[†], Ken'ya Nishikawa[†], Yayoi Tanaka[†], Yasuyuki Usuda[†]

[†]National Institute for Japanese Language and Linguistics
10–2 Midori-cho, Tachikawa, Tokyo 190–8561, Japan
koiso, iseki, waka, kawabata, nishikawa, yayoi, usuda@ninjal.ac.jp

[‡]Graduate School of Humanities, Chiba University
1–33 Yayoicho, Inage-ku, Chiba 263–8522, Japan
den@chiba-u.jp

Abstract

In 2016, we launched a new corpus project in which we are building a large-scale corpus of everyday Japanese conversation in a balanced manner, aiming at exploring characteristics of conversations in contemporary Japanese through multiple approaches. The corpus targets various kinds of naturally occurring conversations in daily situations, such as conversations during dinner with the family at home, meetings with colleagues at work, and conversations while driving. In this paper, we first introduce an overview of the corpus, including corpus size, conversation variations, recording methods, structure of the corpus, and annotations to be included in the corpus. Next, we report on the current stage of the development of the corpus and legal and ethical issues discussed so far. Then we present some results of the preliminary evaluation of the data being collected. We focus on whether or not the 94 hours of conversations collected so far vary in a balanced manner by reference to the survey results of everyday conversational behavior that we conducted previously to build an empirical foundation for the corpus design. We will publish the whole corpus in 2022, consisting of more than 200 hours of recordings.

Keywords: Corpus of everyday Japanese conversation, corpus design, legal and ethical issues, corpus evaluation

The content of this talk is identical to that of the paper 469 of the LREC main conference. The full PDF is available in the proceedings of the main conference, pp. 4259–4264.

Miraikan SC Corpus: A Trial for Data Collection in a Semi-open and Semi-Controlled Environment

Mayumi Bono[1&2], Rui Sakaida[1], Ryosaku Makino[3], Ayami Joh[4]
[1] National Institute of Informatics, [2] SOKENDAI (The Graduate University of Advanced Studies), [3] Waseda University,
[4]The University of Shiga Prefecture,
[1] 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 JAPAN, [2] Shonan Village, Hayama, Kanagawa 240-0193 JAPAN,
[3] 2-579-15 Mikajima, Tokorozawa-city, Saitama 359-1192 JAPAN,
[4] 2500, Hassaka-cho, Hikone-city, Shiga 522-8533 JAPAN
{bono, lui}@nii.ac.jp, rmakino@aoni.waseda.jp, ayami.joh@gmail.com

Abstract

This paper shows the concept and design of our Miraikan SC Corpus. A well-structured and well-prepared corpus would be useful to engineers for understanding the mechanism of speech production and the nature of social interaction with regard to informing the design of their systems. Applications of the corpus range from speech recognition and dialogue processing to human-agent interaction systems, among others. We started collecting audio-visual data using multiple video cameras and microphones in October 2012 at a science museum in Tokyo, Japan. In this paper, we describe the reason why we chose the museum as a research field for data collection, how we audio-video-recorded the interactions, and how we dealt with personal information in the data set, such as participants' names, jobs, and places of residence.

Keywords: Miraikan SC Corpus, science communication, data collection, publication, personal information

1. Introduction

This paper shows the concept and design of our Miraikan SC Corpus. A well-structured and well-prepared corpus would be useful to engineers for understanding the mechanisms of speech production and the nature of social interaction with regard to informing the design of their systems. Applications of the corpus range from speech recognition and dialogue processing to human-agent interaction systems, among others. We started collecting audio-visual data using multiple video cameras and microphones in October 2012 at a science museum in Tokyo, Japan. In this paper, we describe the reason why we chose the museum as a research field for data collection, how we audio-video-recorded the interactions, and how we dealt with personal information in the data set, such as participants' names, jobs, and places of residence.

2. Concept of Miraikan SC Corpus

The Miraikan Science Communication Corpus (hereafter, Miraikan SC Corpus) is a multimodal interaction corpus composed of numerous face-to-face conversations between science communicators (SCs) and visitors. It was collected at the National Museum of Emerging Science and Innovation (hereafter, Miraikan) in Odaiba, Tokyo, Japan.

2.1 What is Miraikan?

Miraikan is a science museum that was built with the purpose of sharing knowledge, gained from science and technology research, with the public. It has several permanent exhibitions that share current science and technology advances from the following broad bases: human beings, space, innovation, and the information society. The permanent exhibitions are separated into three zones, all of which were produced under the supervision of scientists and engineers working at the forefront of their respective fields. Moreover, it provides three or four special exhibitions per year and some interactive events in which scientists and visitors have the opportunity to communicate directly with each other.

2.2 Who are Science Communicators?

Some science museums in Japan, including Miraikan and the National Museum of Nature and Science in Japan, have trained SCs to share their knowledge of science with people visiting the museum. In 2009, Miraikan started a program to develop SCs. Their official website explains what SCs do as follows:

> “Miraikan trains Science Communicators internally and externally through a unique human development system based on practical science communication activities. At Miraikan, Science Communicators are appointed on a fixed-term system for a maximum of five years. During their terms as Science Communicators, they engage in science communication activities, including providing explanations of exhibits on the exhibition floor, and planning events and exhibitions. At the end of their terms, armed with their experience as Science Communicators, these knowledgeable personnel take up jobs in research institutes, universities, science museums and other museums, corporations, and educational institutions. They also provide training for

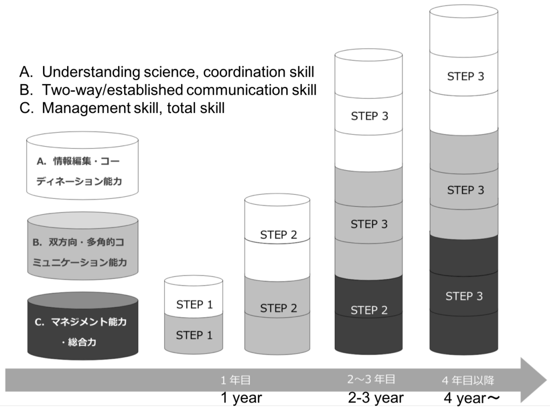

A. Understanding science, coordination skill
B. Two-way/established communication skill
C. Management skill, total skill
A. 情報編集・コーディネーション能力
B. 双方向・多角的コミュニケーション能力
C. マネジメント能力・総合力
STEP 1
STEP 1
STEP 2
STEP 2
STEP 3
STEP 3
STEP 2
STEP 3
STEP 3
STEP 3
1年目
1 year
2～3年目
2-3 year
4年目以降
4 year～

Figure 1: SCs' skills and steps in five years (Miraikan official website)

people who require knowhow in the area of science communication." [1]

Furthermore, they published a report about the purpose and results of the SC program on June 2017, in Japanese (Fig.1)[2].

Miraikan expects SCs to develop three skills each year. The first is the skill of obtaining, understanding, and organizing scientific information as a body of knowledge in their minds. Next is the skill of organizing two-way communication and establishing communication with visitors on the exhibition floors and event spaces. The last skill involves the management and total coordination of SC activities, including teaching and mentoring young SCs.

There is a three-step training process: step 1 is training under senior SCs; step 2 is learning to become an independent decision-maker; and step 3 consists of advanced tasks and mentoring young SCs. This training process has been developing many SCs so far (Table 1).

Table 1: Number of employed persons (2009–2016) (Miraikan official website).

2009	2010	2011	2012	2013	2014	2015	2016	Sum
6	10	10	17	12	13	8	16	92

2.3 Entering the Field

As the Miraikan official website says, almost all of the SCs at Miraikan have five-year employment contracts. They were selected from various academic fields, and they hold masters or doctoral degrees. The SC who explained this arrangement to us also works under this contract, and he had a long academic career before coming to Miraikan. During these five years, SCs learn communication skills aimed at stimulating the public's interest in science and entertaining them. They meet visitors directly on the exhibition floor.

When SCs describe their activities at Miraikan to outsiders, they always say that there is a clear boundary between science communicators and interpreters or curators in a museum. Even if they start by explaining an exhibition to visitors, expert SCs are able to transform this exchange from a one-way explanation to an effective dialogue.

SCs at Miraikan mainly engage in three kinds of activities inside the museum: (1) they explain specific research themes using model kits of objects, such as the brain in a booth where they sit; (2) they sometimes present mini-lectures in a lecture hall, explaining specific topics based on their own experiences and skills; and (3) each of them has time with which to interact with visitors on the exhibition floor, during a pre-determined 1-hour shift.

Starting in October 2012, we conducted a field study at Miraikan in Japan. In April 2012, before we started our fieldwork, one of the science communicators, who had three years' experience as a SC at that time, gave us the opportunity to observe and examine his daily activities at Miraikan. He had been trying to establish criteria with which to evaluate the skills of science communicators and to build an education system for better communication between SCs and visitors (Bono et al. 2014).

2.4 Participants

In general, the term limit of five years at Miraikan is too short to establish and develop scientific knowledge and communication skills on the exhibition floor, especially because the SCs are from diverse academic backgrounds. In addition, they need to plan other activities and events using their specific background in their work places. It is clear that they do not have the time to concentrate too much on brushing up their scientific knowledge and communication skills, to interact with visitors.

Therefore, we decided to make an audio-visual catalog of expert SCs to share methods of interacting with visitors on the exhibition floor at the official kick-off meeting in October 2013. We selected fourteen expert SCs to participate in our recordings and after starting the recordings, other SC recommended one more high-skilled SC, then the total number of SCs for the catalog became fifteen. Half of them had already finished their contracts with Miraikan. We invited them to Miraikan for the recordings. The aim of our project and how we intended to manage the data were explained to all of the SCs and visitors who participated in the recordings, and they granted us permission not only to use the data for our own work but also to publish the video clips, transcripts, and annotation data as a multimodal corpus. This data collection and data publication were allowed by the ethics committee of the National Institute of Informatics.

2.5 Settings and Recording

To construct the Miraikan SC Corpus, we asked Miraikan to allow us to audio-video-record routine conversations between SCs and visitors in the "Spread of Space" and "Challenge the Universe with a Giant Telescope" exhibitions. The reason why we selected these exhibitions was because they offered ideal route for SCs to explain points to visitors; furthermore, it was easy for us and the cameraman to anticipate where SCs and visitors would start their interaction. Additionally, there were high walls, making it possible to avoid filming other visitors who had not agreed to be filmed. The space used for the recordings was separate from other spaces of the exhibition floor.

The shape of the surrounding space resembled a Japanese fan (Fig.2). In many cases, SCs started to explain the exhibition with the planet models from the right edge of the fan. We prepared a rope partition to make a boundary between inside and outside the exhibition and set up a table and chairs to enable visitors to sign the agreement form outside of the exhibition. aJust in case, one of team members was always standing there to receive understandings from other visitors. Next, almost all of the SCs walked through to the Subaru telescope model located at the center of the exhibition. Finally, if visitors had sufficient time, SCs were able to add further explanations of the exhibition using posters and models behind the Subaru telescope. Subsequently, visitors crossed the rope partition placed at the other corner. Afterwards, visitors answered a questionnaire using the table and chairs outside the exhibition.

The recording instruments included six video cameras[3], seven microphones, two lights, and five Kinects (Figure 2) (Trejo et al., 2018). Five video cameras were fixed around the separate area, while a professional camera operator

[1] http://www.miraikan.jst.go.jp/en/aboutus/approach/

[2] http://www.miraikan.jst.go.jp/aboutus/docs/sc_report_201706.pdf

[3] SONY HDR XR550V/PMW EX1R

recorded the front view of the participants with a mobile video camera. Similarly, four shotgun microphones[4] were fixed in position, and several pin microphones[5] were attached to the participants' chests (Figure 3). After the recording, all the recorded sound was mixed, with the noise removed.

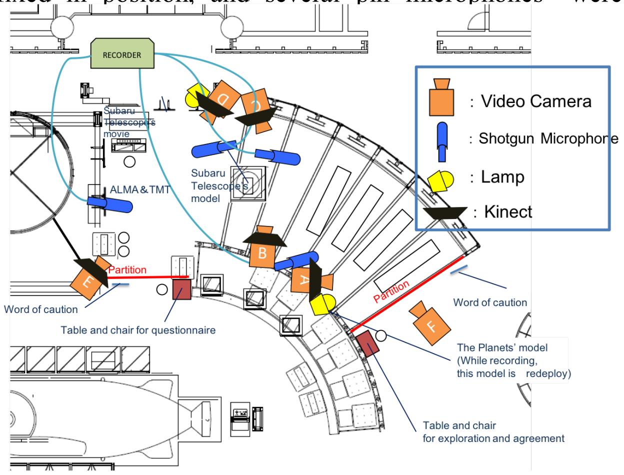


Figure 2: Layout of devices.

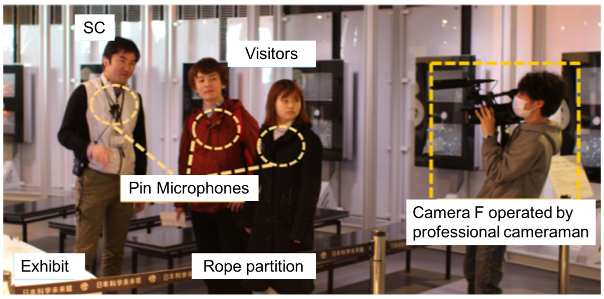


Figure 3: Image of microphones settings and filming by camera F (human operated).

3. Data

3.1 Amount of Data

The video recordings were gathered over 10 days in February and March 2014, for about 1 hour per day. On each day, we asked two SCs to talk separately with three groups of visitors, as naturally as possible. Although we did not tell the SCs which exhibit to explain or the route to the next exhibit, they usually followed the same route and explained the same series of exhibits. Most of the SCs started the talk with an explanation of the model of the solar system and invited the visitors to the next exhibit, a model of the Subaru Telescope. Each group consisted of from 1 to 5 people. The average group had 2.26 visitors.

As a result, 15 expert SCs and 79 visitors (44 males, 35 females) participated in the recordings. The participants ranged from 5 to 66 years old, and the average age was 31.26 (*SD* = 15.78). Thirty-five sessions were recorded in total. The total time was 8 hours and 20 minutes. The average length of the sessions was about 14'30". The shortest interaction was around 4 minutes. The longest interaction was around 20 minutes. All of the video and audio data were merged into one file for each session (Figure 4).

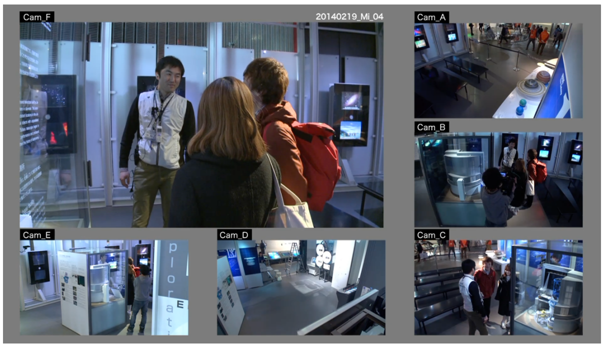


Figure 4: Screenshot of merged six video clips.

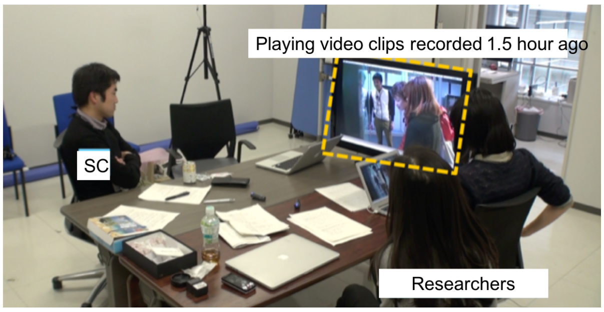


Figure 5: Image of reflection meeting with SCs in backstage.

Almost 1.5 hours after the recording, we asked the SCs to have a reflective session, in which researchers (including the authors) and SCs discussed specific video clips backstage for an hour. These sessions were also filmed by one or two video cameras (Figure 5).

3.2 Dealing with Personal Information

One of appealing points of this corpus is that it includes not only speech and text but also video clips to share with the academic community. However, video clips and images always included personal information, especially a participant's face.

Video files

In cases in which video clips included other people who had not agreed to be filmed, we used pixelation to cover their faces. In Figure 6, the upper image includes a person, marked with a yellow dotted square. Because she had not agreed to participate in the filming, we made her unidentifiable through pixelation. On the other hand, the person on the right who is wearing a black top and jeans is a member of this project; thus, no pixelation was needed. The exhibit composed of mirrors sometimes captured human figures; thus, we utilized pixelation, as in the dotted square on the right in the lower image in Figure 6.

Audio files

SCs always explained the exhibit aligning with the story lines predesigned. However, some SCs asked visitors for their name or place of residence, hometown, etc. to

[4] SENNHEISER MKH 416-P48U3/AKG C414-XLII

[5] SONY UWP-V1

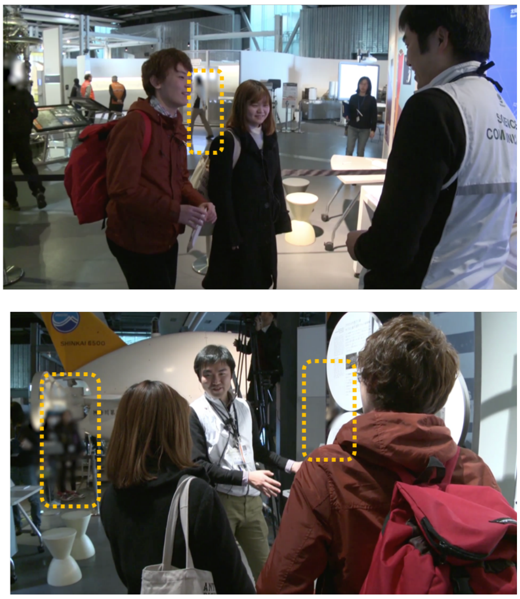


Figure 6: Images of masked video clips.

stimulate the visitors' interest and knowledge related to the exhibition as below:

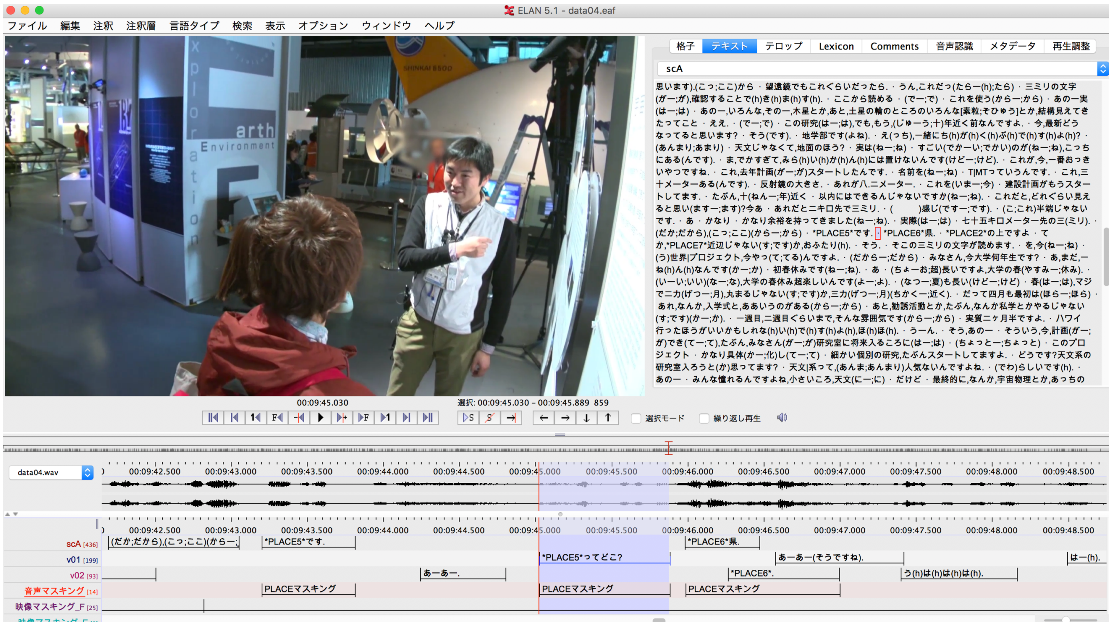
SC: Are you from Tokyo?
V01: umm, no, from *PLACE1*.
SC: Wow, from *PLACE1*, now it's heavily snowing.
V01: Yeah, <laugh>
SC: Where were you born?
V02: *PLACE2*
SC: *PLACE2*, it doesn't have such heavy snow, does it?
V01: No snow.
V02: Yeah, no snow.
SC: Do you know *PLACE3* science museum in *PLACE1*?
V01: uumm, no, I'm sorry.
SC: That museum is…
(1.0)
SC: I'm from *PLACE4*, there is no snow there either.
(1.0)
SC: This (Subaru telescope) is on the top of a mountain in Hawaii.

PLACES 1, 2, and 4 are the names of hometowns for SCs and visitors, which is personal information for them. And PLACE 3 is the name of the science museum in PLACE 1 (V01's hometown). In this case, the SC tried to elicit visitors' interest and uncover their knowledge related to science. If visitors had known the science museum, the SC may have skipped some basic explanations about this exhibition.

We replaced this personal information with a beep sound on the audio file. Furthermore, we replaced place names with a place ID, such as *PLACE1*, *PLACE2* on annotation and transcript files. Our team kept the specific names for certain information- and knowledge-sensitive analyses, such as those involving epistemics in Conversation Analysis (CA) and embodied interaction analysis (Sakaida et al. 2018). We prepared NAME-IDs and JOB-IDs as well. Moreover, we checked the nature of the content in all files. As the results, two files contained politically incorrect topics, so we prepared the other 33 files for publication, with the exception of those two.

4. Publishing Miraikan SC Corpus

We would like to explain our vision for publishing the corpus. We are planning to publish masked video and audio files of interactions on the exhibition floor (except for the reflection meetings backstage), masked transcripts in Japanese on ELAN (.eaf, .pfsx) (Figure 7), and manuals in the near feature. The Miraikan SC Corpus was created in a semi-open and semi-controlled environment, which entails that visitors had an intrinsic motivation to come to the



思います).(こっ;ここ)から ・ 望遠鏡でもこれぐらいだったら. ・ うん,これだっ(たらー(h);たら) ・ 三ミリの文字(がー;が),確認することで(h)き(h)ま(h)す(h). ・ ここから読める ・ (でー;で) ・ これを使う(からー;から) ・ あのー実(はー;は) ・ あのー,いろんな,そのー,木星とか,あと,土星の輪のところのいろんな[素粒;そひゅう]とか,結構見えてきたってこと ・ ええ. ・ (でー;で) ・ この研究(はー;は),でも,もう,(じゅーう;十)年近く前なんですよ. ・ 今,最新どうなってると思います? ・ そう(です). ・ 地学部です(よね). ・ え(っち),一緒にち(h)が(h)く(h)ぶ(h)で(h)す(h)よ(h)? ・ (あんまり;あまり) ・ 天文じゃなくて,地面のほう? ・ 実は(ねー;ね) ・ すごい(でかーい;でかい)のが(ねー;ね),こっちにある(んです). ・ ま,でかすぎて,みら(h)い(h)か(h)ん(h)には置けないんです(けどー;けど). ・ これが,今,一番おっきいやつですね. ・ これ,去年計画(がー;が)スタートしたんです. ・ 名前を(ねー;ね) ・ T|MTっていうんです. ・ これ,三十メーターある(んです). ・ 反射鏡の大きさ. ・ あれが八.二メーター. ・ これを(いまー;今) ・ 建設計画がもうスタートしてます. ・ たぶん,十(ねんー;年)近く ・ 以内にはできるんじゃないですか(ねー;ね). ・ これだと,どれぐらい見えると思い(ますー;ます)?今あ ・ あれだと二キロ先で三ミリ. ・ (　　　)感じ(ですー;です). ・ (こ;これ)半端じゃないです. ・ あ ・ かなり ・ かなり余裕を持ってきました(ねー;ね). ・ 実際(はー;は) ・ 七十五キロメーター先の三(ミリ). ・ (だか;だから),(こっ;ここ)(からー;から) ・ *PLACE5*です. ・ *PLACE6*県. ・ *PLACE2*の上ですよ ・ てか,*PLACE7*近辺じゃない(す;です)か,おふたり(h). ・ そう. ・ そこの三ミリの文字が読めます. ・ を,今(ねー;ね) ・ (う)世界|プロジェクト,今やっ(て;てる)んですよ. ・ (だからー;だから) ・ みなさん,今大学何年生です? ・ あ,まだ,一ね(h)ん(h)なんです(かー;か) ・ 初春休みです(ねー;ね). ・ あ ・ (ちょーお;超)長いですよ,大学の春(やすみー;休み). ・ (いーい;いい)(なー;な),大学の春休み超楽しいんです(よー;よ). ・ (なつー;夏)も長い(けどー;けど) ・ 春(はー;は),マジで二カ(げつー;月),丸まるじゃない(す;です)か,三カ(げつー;月)(ちかくー;近く). ・ だって四月も最初は(ほらー;ほら) ・ あれ,なんか,入学式と,ああいうのがある(からー;から) ・ あと,勧誘活動とか,たぶん,なんか私学とかやるじゃない(す;です)(かー;か). ・ 一週目,二週目ぐらいまで,そんな雰囲気です(からー;から) ・ 実質二ヶ月半ですよ. ・ ハワイ行ったほうがいいかもしれな(h)い(h)で(h)す(h)よ(h),ほ(h)ほ(h). ・ うーん. ・ そう,あのー ・ そういう,今,計画(がー;が)でき(てー;て),たぶん,みなさん(がー;が)研究室に将来入るころに(はー;は) ・ (ちょっとー;ちょっと) ・ このプロジェクト ・ かなり具体(かー;化)し(てー;て) ・ 細かい個別の研究,たぶんスタートしてますよ. ・ どうです?天文系の研究室入ろうと(か)思ってます? ・ 天文|系って,(あんま;あんまり)人気ないんですよね. ・ (でわ)らしいです(h). ・ あのー ・ みんな憧れるんですよね,小さいころ,天文(にー;に) ・ だけど ・ 最終的に,なんか,宇宙物理とか,あっちの

scA: (だか;だから),(こっ;ここ)(からー; | *PLACE5*です. | *PLACE6*県.
v01: *PLACE5*ってどこ? | あーあー(そうですね). | はー(h).
v02: あーあー. | *PLACE6*. | う(h)は(h)は(h)は(h).
音声マスキング: PLACEマスキング | PLACEマスキング | PLACEマスキング

Figure 7: Image of published files on ELAN

Miraikan to engage with the latest results in science and technology. SCs interacted with visitors as part of their daily work, which was filmed.

Previous studies of dialogue corpus research were oriented toward recording high-quality audio data collected in experimental rooms (e.g., the HCRC Map Task Corpus). Using this method, it is difficult to collect naturally-occurring conversation, and it is not a self-motivated interaction. CA analysts have been collecting naturally-occurring conversations for a long time. However, they tend to not release their data, as it contains too much personal information. Furthermore, CA analysts sometimes use this information as ethnographic background in their analyses. Of course, the theoretical backgrounds and research purposes of corpus researchers and CA analysts are very different. However, we assume that some parts of their research interests overlap. In this regard, we tried to create platforms to discuss future directions for interaction studies examining shared and common data sets. Combining two different research fields, such as corpus studies and CA studies, and publishing the Miraikan SC Corpus constitutes a challenge. Furthermore, if some informatics engineers of dialogue and AI systems prove interested in our project, we hope that our efforts may contribute to changing the paradigm of dialogue and interaction studies.

5. Bibliographical References

Bono, M., Ogata, H., Takanashi, K. and Joh, A. (2014). The Practice of Showing 'Who I am': A Multimodal Analysis of Encounters between Science Communicator and Visitors at Science Museum. Organized Session: Brightening Life Style up with Technologies. HCI International 2014, pp. 22-27.

Sakaida, R., Makino, R. & Bono, M. (2018) Preliminary Analysis of Embodied Interactions between Science Communicators and Visitors Based on a Multimodal Corpus of Japanese Conversations in a Science Museum, The 11th edition of the Language Resources and Evaluation Conference (LREC).

Trejo, K., Angulo, C., Satoh, S. & Bono, M. (2018) Towards robots reasoning about group behavior of museum visitors: Leader detection and group tracking. Journal of Ambient Intelligence and Smart Environments, Vol. 10, No. 1, pp. 3-19.

A Multimodal Multiparty Human-Robot Dialogue Corpus for Real World Interaction

Kotaro Funakoshi
Graduate School of Informatics, Kyoto University, funakoshi.k@i.kyoto-u.ac.jp
Honda Research Institute Japan Co., Ltd., funakoshi@jp.honda-ri.com

Abstract

We have developed the MPR multimodal dialogue corpus and describe research activities using the corpus aimed for enabling multiparty human-robot verbal communication in real-world settings. While aiming for that as the final goal, the immediate focus of our project and the corpus is non-verbal communication, especially social signal processing by machines as the foundation of human-machine verbal communication. The MPR corpus stores annotated audio-visual recordings of dialogues between one robot and one or multiple (up to tree) participants. The annotations include speech segment, addressee of speech, transcript, interaction state, and, dialogue act types. Our research on multiparty dialogue management, boredom recognition, response obligation recognition, surprise detection and repair detection using the corpus is briefly introduced, and an analysis on repair in multiuser situations is presented. It exhibits richer repair behaviors and demands more sophisticated repair handling by machines.

Keywords: human-robot interaction, verbal and non-verbal communication, social signal processing

1. Introduction

Although most conventional (spoken) dialogue system research has assumed one-to-one conversations between a user and a machine, one-to-many situations between multiple users and a servicer (a machine) are also common and even predominant in the real world. Therefore, there has been much recent research on how to handle such situations (Bennewitz et al., 2005; Bohus and Horvitz, 2009; Al Moubayed et al., 2012; Matsuyama et al., 2015).

Machines assuming one-to-one conversation can get by with just being reactive to input speech and indifferent to the user's status. However, those assuming one-to-many situations must be more proactive and attentive to users' statuses in order to provide meaningful interactions. For example, a conversational system must identify the addressee of a speaking user, i.e., the system or the other human participant (Nakano et al., 2014). Handling social signals (Vinciarelli et al., 2009) and non-verbal information thus has greater significance in multiparty dialogues.

In light of this background, we designed HALOGEN, a software framework for enabling multimodal human-robot interaction (Funakoshi and Nakano, 2017) (shown in Figure 1). HALOGEN primarily handles non-verbal information (mostly audio and visual) from sensor inputs. The information is handled by multiple sub-modules (detectors/recognizers) and integrated by the core module, which oversees not only the final information integration but also mutual communication with the dialogue manager such as (Nakano et al., 2011). It also manages user profile information collected from both the non-verbal and verbal information. Social information such as name, gender, and occupation can be obtained or confirmed through dialogue, and the dialogue manager passes such information to HALOGEN. The dialogue manager can query HALOGEN about both user profiles and user statuses to achieve better verbal communication. HALOGEN can use the confirmed information from the dialogue manager in order to refine user status estimation.

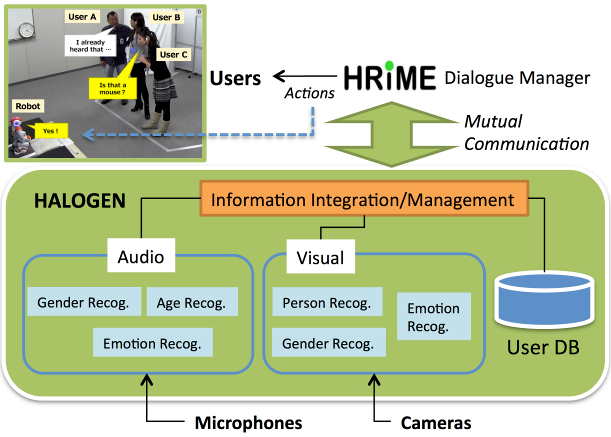


Figure 1: HALOGEN framework.

To implement and evaluate HALOGEN, we have collected sessions of human-robot interaction between a robot and multiple users. The data was collected in a situation between the public and laboratory settings described in (Al Moubayed et al., 2012), although the participants were somewhat controlled in the laboratory by the experimenter. Sections 2. and 3. of this paper describe the recording of the audio-visual data and the annotation, respectively, which form the core of the MPR (Multi-Party Robot) corpus. In Section 4., we introduce several research topics using the corpus. We conclude in Section 5. with a brief summary.

2. Data Recording

We recorded human-robot interaction data twice, once in 2012 and once in 2016. These two recordings were done in mostly similar settings but with a few differences in terms of participants, recording environment, and interaction design. We refer to them as MPR2012 and MPR2016.

The two main differences between the two editions are (1) the specifications of the visual recording devices and (2) the robot operations (full manual operation only in 2012; both full manual and full automatic operations in 2016).

2.1. MPR2012

We recruited 30 trios of individuals through a research-support agency. The three participants in each trio were friends or family and ranged in age from their 20s to 60s. The genders of the participants were balanced.

2.1.1. Situation and recording settings

Each trio participated in two 25-minute interaction sessions with a robot in which they repeatedly engaged in a conversational game. They were instructed that the autonomous robot was under development and that they should be tolerant of errors. After the sessions, they were informed that the robot was controlled by a human operator.

The robot spoke English only, as it was explained that the robot was designed for English learning purposes. The participants were allowed to speak either English or Japanese. The interaction setting is shown Figure 2, and the upper-left picture in Figure 1 shows a recording scene in this setting. The participants started the sessions from the waiting spaces. They came into the interaction field and went back to one of the two waiting spaces in accordance with the instructions from the director (experimenter), who stayed outside the laboratory. The interaction field was indicated by lines so that they would stay in the proper shooting area. The instructions included *participating in the game, observing the game, passing through the field*, and *returning to one of the waiting spaces*. Each participant in a session stayed in the field for about 15 minutes in total.

In the waiting spaces, the participants stayed quiet while listening to music provided through headphones so that they could not sense what was going on in the field. Throughout each session, they were equipped with a transceiver and an earphone. The instructions from the director were sent only to the earphone of the relevant participant. The idea here was to create information imbalance among the human participants that would result in frequent communications among them.

The operator watched the situation and the participants by means of a NAO robot[1] with a camera installed in its head and a static birds-eye-view camcorder, which was also used to record the sessions for annotation. The operator controlled the direction of the robot's head and hand gestures according to the interaction. The possible utterances of the robot were fixed and prepared as buttons in the GUI interface for the operator. The sessions were recorded with Microsoft Kinect v1 and four omni-directional distant microphones behind the robot.[2]

2.1.2. Interaction games

Each trio engaged in the 20 Questions game for their first session. In this game, the robot as game master secretly chooses a target object *tiger, sushi, cell phone, etc.* The participants can ask the robot about an attribute of the object as a yes-no question or make a guess up to 20 times. If they make a correct guess within 20 tries, they win; otherwise the robot wins. Although they were instructed about the game before starting the session, the robot explained the rules of the game after initial greetings and engagement phases (inviting participants to the game). When one game ended, the robot immediately started another game.

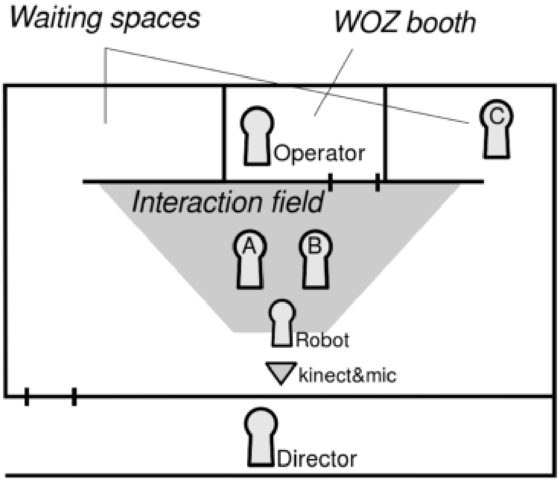


Figure 2: Data recording settings used for MPR2012.

In their second session, participants played a gesture mimicking game. First, each participant was taught gestures corresponding to various English words. When more than one participant was ready in the field, the robot started a mimicking speed competition in which it said a word and participants had to make the corresponding gesture as fast as they could. The robot judged the answers and declared as the winner whoever made the correct gesture fastest.

2.2. MPR2016

The participant population and statistics are nearly identical to MPR2012, with gender-balanced 90 participants divided into 30 trios.

2.2.1. Situation and recording settings

Figure 3 shows a data collection scene in MPR2016, where the setting was almost equal to that of MPR2012 except that the waiting spaces and the operator booth were not behind the interaction field. In this edition, we adopted a prototype spoken dialogue system for the second sessions. A manually operated robot was used for the first sessions, the same as MPR2012.

The Kinect device was upgraded to version 2. This brought us (1) higher image resolution (HD), (2) better skeletal tracking performance, and (3) synchronized recording of video and audio in one Kinect data file. The other equipment used in the recording was the same.

The instructions given to the participants and the director's role were also almost the same as in MPR2012. This time, however, the participants were informed in advance that the robot was operated by a human in the first session and by a system in the second, as the difference in interaction quality between a human operator and the system was significant. The director also had another task, namely, to insert attention-drawing events into the recording sessions in order to collect preliminary data for the surprise detection

[1] http://www.aldebaran.com/en/cool-robots/nao

[2] Kinect v1 could not record audio by itself.

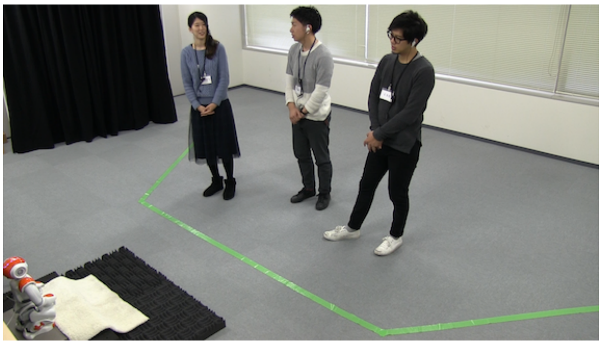


Figure 3: Data recording scene in MPR2016.

research (discussed in Section 4.4.)[3]
For the first 10 trios, the robot spoke English, the same as it did for MPR2012. For the remaining 20 trios, the dialogue system was brushed up based on the experiences and data from the first 10 trios, and it was also modified to speak Japanese, as some participants could not quite catch it when the robot suddenly uttered an irrelevant or irrational message in English to induce surprised reactions.

2.2.2. Interaction games

The game used for the first sessions was the same as in MPR2012, i.e., 20 Questions. In the second sessions we used a simplified version of 20 Questions, as it was hard for the system to correctly answer participants' arbitrary yes-no questions with regard to the chosen object. In the simplified version, participants could only make a guess or ask for a hint.

3. Annotation

3.1. MPR2012

The recorded data was annotated using ELAN (Brugman and Russel, 2004) with regard to the following information for each participant.

Participation status One of the following labels was placed over the timeline whenever a given participant was visible: *participating* (maintaining interaction with the robot), *observing* (staying in the field without interacting with the robot), and *passing* (leaving from the field or transiting behind the field).

Gaze Based on the head directions, the targets of attention were labeled as the combination of the three participants and the robot, That is, each label was a combination of the names of the participants (A, B, and C) and the robot (NAO). If the gaze at the moment could not be estimated, an *invalid* label was given.

Utterance segment and addressee The speech segments of the participants and the robot were annotated by segmenting speech with pauses over 400 ms or sentential boundaries. Each speech segment was labeled with its addressee(s) in the same way as gaze was. Otherwise, it was labeled as *laugher* or *monologue*.

Transcript Each speech segment was transcribed. These transcripts contain a mixture of English and Japanese.

Dialogue act Each speech segment of 39 sessions out of 60 (9 first sessions and 30 second sessions) was labeled with one of the dialogue act types based on DIT++ (Bunt, 2009). The dialogue types contain some domain-specific labels such as *Quiz-Challenge* (making a guess) and *Quiz-Judge* (evaluating the guess).

3.2. MPR2016

The recorded data was annotated using ELAN with regard to participation status, utterance segment and addressee, and transcript in the same manner as MPR2012. Gaze and dialogue act have not been annotated thus far.

4. Research Using the MPR Corpus

In this section, our studies using the MPR corpus are introduced.

4.1. Multiparty Dialogue Management

Dialogue management entails tracking the the current dialogue state based on the given context, i.e., present situation and past discourse, and deciding the system's next action, that is, *what to speak*. In addition to *what to speak*, a dialogue system in a multiparty situation has to consider seriously *who to speak to* and *when to speak*.
Using the dialogue act annotation in MPR2012 as a basis, (Kennington et al., 2014) proposed and preliminarily evaluated a model for multiparty dialogue management that manages a multiparty situation as a bundle of one-to-one dialogues while suppressing conflicting or redundant actions at a pre-output action manager. The action manager is also responsible for *who to speak to* and *when to speak*. By this means, the model can flexibly handle an arbitrary number of participants.

4.2. Boredom Detection

To achieve a long-term relationship between users and a dialogue system, it is important to ensure that users maintain a willingness to continue using the system (Funakoshi et al., 2010). Detecting boredom in users is a key technology to maintain such willingness or motivation. (Shibasaki et al., 2017) annotated the first session's data of MPR2012 with regard to boredom based on the intuitive sense of two annotators, and proposed a detection model using the body motion of participants, the relationship between their face directions and standing positions, and the information obtained from participation statuses.

4.3. Response Obligation Estimation

Response obligation is whether the system has to respond to input sound or not. Even when a speech input from a user is directed to the system, it does not always mean that the system has to respond to it immediately. Moreover, in a multiparty situation, the system should not respond to a speech input that is directed from one user to another user.

[3] These events were mostly sounds such as breaking glass, the meow of a cat, a stamping sound made by a participant in the waiting space (he/she was instructed to do so by the director through the transceiver), etc. Although these sounds were made to draw the attentions of the participants, they were not nearly as loud as a sudden warning message made by the NAO robot when it was overheated.

(Sugiyama et al., 2015) proposed a response obligation estimation method using non-verbal information and evaluated it with MPR2012. The proposed method handles the response obligation estimation problem as the composition of noise rejection, addressee identification, end-of-turn detection, and speaker intention recognition.

The proposed method does not use speech recognition results so as to ensure domain versatility. This feature, however, and the diversity in user behaviors, limits the estimation performance, especially against unseen people in the model training data. To overcome this, currently we are working on error recovery from failures of response obligation estimation, and online adaptation to new users. In the next two subsections, two key elements of the error recovery are discussed.

4.4. Surprise Detection

Failures of response obligation estimation can happen in two ways. One is the false-positive case: the system wrongly responds to an irrelevant input sound. In response to this case, users often exhibit surprised reactions. It seems that such a reaction mostly appears as a sudden movement in the head or body, as a repair request such as "Huh?" (Dingemanse et al., 2013), or as an expression of confusion in the face or voice, typically as confused laughter.

To build an effective surprise detection method, we tried to artificially increase surprised reactions of participants in MPR2016. This data is to be examined in future.

4.5. Repair Detection

The other type of failure is the false-negative case: the system ignores a user's speech input that it should have responded to. In response to this case, users often try repair, i.e., repeating or rephrasing the previous utterance.

We are currently developing a repair detection method based on previous approaches such as (Cevik et al., 2008). However, all the previous approaches assume one-to-one clean communication. In a real multiparty situation, the problem is not so simple. Given an input sound, it is not obvious for the system to decide the target sound to be compared with the input for repair detection because occationaly noises, monologues, or conversation with other participants are interjected between a repairing utterance and the utterance to be repaired (in our case, the ignored utterance).

Here, we discuss an analysis of repair activities on occasions of false-negative errors in response obligation estimation. Ten of the second sessions of MPR2016 were used. This data includes 2,032 utterances from the robot to participants, 2,506 from from participants to the robot, and 934 from a participant to other participant(s). Of the 2,032 robot-directed utterances, 613 are ignored.

Table 1 lists the distribution of user behaviors after a speech directed to the robot was ignored. In 31% of all cases, another participant talked first, and in 46%, the ignored participant made some action. In 22%, the participants did nothing until the robot made a prompting message.

After the robot's ignoring, in the 192 cases, another participant spoke 81 repairs instead of the ignored person. Table 2 shows the breakdown of these 81 cases along with the 88 cases where the repair was done by the ignored person. Rephrasing accounts for almost half of the cases. This indicates we should prepare for both repetition and rephrasing. Repetition detection based on Dynamic Time Warping between two speech sounds is expected to be robust against speech recognition errors. However, its performance would be degraded when the speakers are different (as in the DP case in Table 2). It is important to note that a repair utterance may not come immediately after the utterance to be repaired. Indeed, we found that three out of 11 repetition cases in one session contained interjections of one or two irrelevant utterances. We have to build a smarter repair handling that can manage all these issues.

Table 1: Participant responses after robot's ignoring.

Response pattern	Count	Ratio
Another participant talks	192	31%
Shifting to another topic	150	24%
Repairing by repeating or rephrasing	88	14%
Talking to another participant	47	8%
Waiting until robot speaks	136	22%
Total	613	100%

Table 2: Breakdown of repair behaviors by same participant (SP) and by different participant (DP).

Repair behavior	SP	DP	Sum	Ratio
Rephrasing into different words	49	34	83	49%
Repeating the original words	16	33	49	29%
Repeating an extended expression	13	9	22	13%
Repeating a part of the original	10	5	15	9%
Total	88	81	169	100%

Repair is considered a universal part of language use (Levinson, 2016), but handling repair in spoken dialogue systems is currently quite limited. As discussed above, it seems most previous approaches to repairing are oriented to verbal aspects. It is essential now that non-verbal approaches be studied, too.

5. Concluding Remarks

To expand the area in which dialogue systems and conversational machines can function, it is important to make systems capable of handling multiparty situations, where multimodal processing of non-verbal information or social signals is a key component.

We have designed and implemented the HALOGEN framework for multimodal multiparty interaction, and collected roughly 50 hours of audio-visual data on one-to-many human-robot interactions with 180 participants. The data is annotated with speech segment, addressee, transcript, etc. and has been used in several of the studies introduced in this paper. The corpus is not public but is available in research collaboration with Honda Research Institute Japan Co., Ltd.

6. Bibliographical References

Al Moubayed, S., Beskow, J., Granström, B., Gustafson, J., Mirnig, N., Skantze, G., and Tscheligi, M. (2012). Furhat goes to robotville: A large-scale multiparty human-robot interaction data collection in a public space. In *Proc. LREC workshop on multimodal corpora for machine learning*.

Bennewitz, M., Faber, F., Joho, D., Schreiber, M., and Behnke, S. (2005). Towards a humanoid museum guide robot that interacts with multiple persons. In *Proc. Humanoids*, pages 418–424.

Bohus, D. and Horvitz, E. (2009). Models for multiparty engagement in open-world dialog. In *Proc. SIGDIAL*, pages 225–234.

Brugman, H. and Russel, A. (2004). Annotating multimedia/ multi-modal resources with elan. In *Proc. LREC*.

Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. In *Proc. EDAML*.

Cevik, M., Weng, F., and Lee, C.-H. (2008). Detection of repetitions in spontaneous speech in dialogue sessions. In *Proc. Interspeech*, pages 471–474.

Dingemanse, M., Torreira, F., and Enfield, N. J. (2013). Is "huh?" a universal word? conversational infrastructure and the convergent evolution of linguistic items. *PLoS ONE*, 8(11).

Funakoshi, K. and Nakano, M. (2017). Online evaluation of response obligation estimation on the halogen multimodal interaction framework. In *Proc. HRI (LBR)*.

Funakoshi, K., Nakano, M., Kobayashi, K., Komatsu, T., and Yamada, S. (2010). Non-humanlike spoken dialogue: A design perspective. In *Proc. SIGDIAL*, pages 176–184.

Kennington, C., Funakoshi, K., Nakano, M., and Takahashi, Y. (2014). Probabilistic multiparty dialogue management for a game master robot. In *Proc. HRI (LBR)*.

Levinson, S. C. (2016). Turn-taking in human communication origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1).

Matsuyama, Y., Akiba, I., Fujie, S., and Kobayashi, T. (2015). Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. *Computer Speech & Language*, 33(1):1–24.

Nakano, M., Hasegawa, Y., Funakoshi, K., Takeuchi, J., Torii, T., Nakadai, K., Kanda, N., Komatani, K., Okuno, H. G., and Tsujino, H. (2011). A multi-expert model for dialogue and behavior control of conversational robots and agents. *Knowledge-Based Systems*, 24:248–256.

Nakano, Y., Baba, N., HUANG, H.-H., and Hayashi, Y. (2014). A multiparty conversation system with an addressee identification mechanism based on nonverbal information. *Transactions of the Japanese Society for Artificial Intelligence*, 29(1):69–79.

Shibasaki, Y., Funakoshi, K., and Shinoda, K. (2017). Boredom recognition based on users' spontaneous behaviors in multiparty human-robot interactions. In *Proc. MMM*.

Sugiyama, T., Funakoshi, K., Nakano, M., and Komatani, K. (2015). Estimating response obligation in multi-party human-robot dialogues. In *Proc. Humanoids*.

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27:1743–1759.

Speech and Language Resources for the Development of Dialogue Systems and Problems Arising from their Deployment

Ryuichiro Higashinaka[1], **Ryo Ishii**[1], **Narimune Matsumura**[1]
Tadashi Nunobiki[1], **Atsushi Itoh**[2], **Ryuichi Inagawa**[2], **Junji Tomita**[1]
[1]NTT Corporation, [2]NTT Data Corporation
higashinaka.ryuichiro@lab.ntt.co.jp, ishii.ryo@lab.ntt.co.jp, matsumura.narimune@lab.ntt.co.jp
nunobiki.tadashi@lab.ntt.co.jp, inagawar@nttdata.co.jp, itouats@nttdata.co.jp, tomita.junji@lab.ntt.co.jp

Abstract

This paper introduces the dialogue systems (chat-oriented and argumentative dialogue systems) we have been developing at NTT together with the speech and language resources we used for building them. We also describe our field trials for deploying dialogue systems on actual premises, i.e., shops and banks. We found that the primary problem with dialogue systems is timing, which led to our current focus on multi-modal processing. We describe our multi-modal corpus as well as our recent research on multi-modal processing.

Keywords: Chat-oriented dialogue system, argumentative dialogue system, deployment of dialogue systems, multi-modal processing

1. Introduction

We are seeing an emergence of dialogue systems in our daily lives. Many task-oriented dialogue systems, such as Siri, Cortana, and Alexa, have been in use in our daily lives, and there have been a number of non-task oriented ones for social and entertainment purposes (Onishi and Yoshimura, 2014; Vinyals and Le, 2015; Shang et al., 2016; Higashinaka et al., 2017a).

NTT has been working on dialogue systems for decades, and, in terms of research, we are now specifically focusing on chat-oriented dialogue systems. This is because chat is an important part in human-machine communication. According to the survey done by the National Institute for Japanese Language and Linguistics, more than 60% of our conversations can be classified as chat (Koiso et al., 2016). This means, if we do not equip dialogue systems with chat capability, they will not be able join our conversations most of the time, which makes it difficult for such systems to become our "partners". In addition to the survey, it has also been pointed out that we tend to chat with systems, even though users are explicitly informed that the systems are task-oriented (Takeuchi et al., 2007). This means that, even for task-oriented dialogue systems, chat capability is necessary for them to be useful.

We first introduce our chat-oriented dialogue system that we are developing. Since the system has to handle open-domain utterances from users, it needs to have an abundance of knowledge, requiring a number of resources. We describe the speech and language resources we created to develop our chat-oriented dialogue system. In addition to our chat-oriented dialogue system, we describe our recent work on an argumentative dialogue system that can have discussions with people. The aim of creating this system is to investigate ways to make users more engaged in conversation; topics tend to transit from one to the other in chat, whereas discussion requires more attention on a certain discussion topic, making argumentation an ideal research subject. Second, apart from our research prototypes, we have also been conducting trials of dialogue systems with actual users, placing systems on premises, such as shops and banks. This paper presents two case studies of such trials. Finally, we describe our recent work on multi-modal processing because in our research and also from deployment experience, we found that timing is by far the key problem with current dialogue systems.

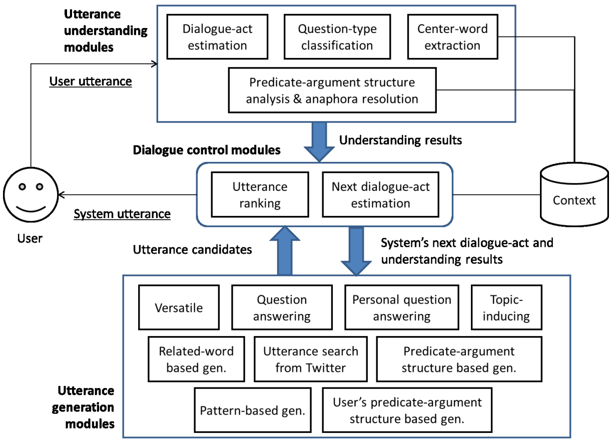

Utterance understanding modules
Dialogue-act estimation
Question-type classification
Center-word extraction
User utterance
Predicate-argument structure analysis & anaphora resolution
Understanding results
Dialogue control modules
Utterance ranking
Next dialogue-act estimation
Context
System utterance
User
Utterance candidates
System's next dialogue-act and understanding results
Versatile
Question answering
Personal question answering
Topic-inducing
Related-word based gen.
Utterance search from Twitter
Predicate-argument structure based gen.
Utterance generation modules
Pattern-based gen.
User's predicate-argument structure based gen.

Figure 1: System architecture of our chat-oriented dialogue system (see (Higashinaka et al., 2014) for details)

In Section 2, we describe our chat-oriented and argumentative dialogue systems. In Section 3, we describe our deployment of dialogue systems, covering two case studies. In Section 4, we describe our multi-modal corpus and our research regarding multi-modal processing. We summarize the paper and mention future work in Section 5.

2. Dialogue systems and resources

2.1. Chat-oriented dialogue system

Figure 1 shows the architecture of our chat-oriented dialogue system. The system has multiple modules, which can be classified into three blocks: utterance understanding, dialogue control, and utterance generation.

The system works in the following steps: given an input user utterance, utterance-understanding modules analyze the utterance, estimate a dialogue act and question type, extract center words (foci/topics in an utterance), and predicate argument structures (PASs). Dialogue-control modules receive the utterance-understanding results and determine the next dialogue act of the system. The utterance-understanding results and dialogue act of the system are fed to the utterance-generation modules to generate utterance candidates, which are finally ranked by the ranking module in the dialogue control. Finally, the top-rank utterance is selected to be output to the user (see (Higashinaka et al., 2014) for details of these modules).

Since we focus on open-domain conversation, we created a number of language resources for handling a variety of topics. For dialogue-act estimation, center-word extraction, and PAS analysis, we created training data to realize such functions with machine-learning methods. Specifically, on top of the chat dialogue data we collected, we carried out multiple annotations; namely, dialogue-act annotation, center-word annotation, and PAS annotation. We also carried out discourse relation annotation using the relations in the Penn Discourse Tree Bank (PDTB) (Miltsakaki et al., 2004). To generate a variety of system utterances, we created large-scale response rules in Artificial Intelligence Markup Language (AIML) (Wallace, 2009), which are used in the pattern-based generation module in Figure 1. We describe these resources below.

2.1.1. Chat dialogue corpus and its annotations

We use our chat-dialogue corpus as a base corpus. We collected 3,680 chat dialogues between two human users using a messenger interface. The total number of utterances is about 134K, and the number of users is 95. More than 1.2M words are included in the corpus. The length of a dialogue is about 36 utterances on average with about 9 words per utterance.

We sampled 20K utterances and carried out center-word annotation, in which noun phrases (NPs) denoting the foci/topics are annotated in utterances. We carried out dialogue-act annotation on all utterances in our chat-dialogue corpus. We used the dialogue-act taxonomy in (Meguro et al., 2010). The dialogue-act tag covers diverse utterances, making it suitable for open-domain conversation. There are 33 dialogue acts in the tag set. For PAS annotation, we sampled about 300 dialogues and annotated them with PASs; for each predicate, we mainly annotated `ga` (nominative), `wo` (accusative), and `ni` (dative) cases as well as several optional cases. We also carried out co-reference annotation, including zero-anaphora annotation (Imamura et al., 2014). Finally, for all utterances in the corpus, we carried out PDTB-style discourse-relation annotation. This chat-dialogue corpus is, as far as we know, by far the most well-annotated chat-dialogue corpus in Japanese. The annotations have been used to train models for center-word extraction, dialogue-act estimation, PAS extraction (including anaphora resolution), and discourse-relation detection. Discourse-relation detection has been found effective for ranking utterance candidates (Otsuka et al., 2017).

2.1.2. Large-scale response rules in AIML

We created large-scale response rules in AIML. We first created an initial rule set then revised it in the following manner. First, one text analyst created 149,300 rules by referring to our dialogue resources, mainly our chat-dialogue corpus described above. Then, an external judge subjectively evaluated the quality of the rules by inputting sampled utterances into a system loaded with the rules, and only when more than 90% of the responses were above average (over 6 points out of 10) was the rule-creation terminated. Then, we revised this rule set by using online evaluation where one external judge chatted for two turns with the system and evaluated the interactions subjectively. The rule-revision process terminated only when the judge was satisfied (same criterion as above) 90% of the time within 100 interactions. We ran eight iterations of this procedure to finalize the revised rule set. The entire revision process took approximately three months. At the end, the rule set contained 333,295 rules (categories in AIML) (see (Higashinaka et al., 2015) for details of this rule-creation process).

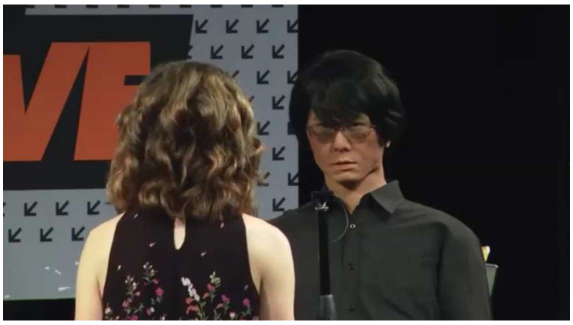


Figure 2: Geminoid HI-4 with our chat-oriented dialogue system at SXSW 2016. ©2015-2016 SXSW, LLC. This research was conducted in collaboration with Ishiguro laboratory of Osaka University.

2.1.3. Performances

We created two dialogue-system prototypes based on our architecture. One is Matsukoroid[1], which is an android robot that looks exactly like the famous TV personality Matsuko Deluxe in Japan. We incorporated our chat-oriented dialogue engine into the robot and let Matsuko Deluxe and his android chat with each other. This interaction was aired on Japanese television. The other is another android called Geminoid HI-4 (See Figure 2). We performed a live demonstration at South by South West (SXSW) in 2016. This system was an English port of our Japanese system; the overall architecture was the same with English data we newly created.

2.2. Argumentative dialogue system

Our chat-oriented dialogue system can maintain conversation by tracking center-words and by responding with large-scale rules as well as knowledge mined from the web. However, we also found that the content of a dialogue is rather superficial because the topics transit from one to the other, not going deeper into a topic. This sometimes made the dialogue less engaging for users.

[1] https://naturaleight.co.jp/matsukoroid/

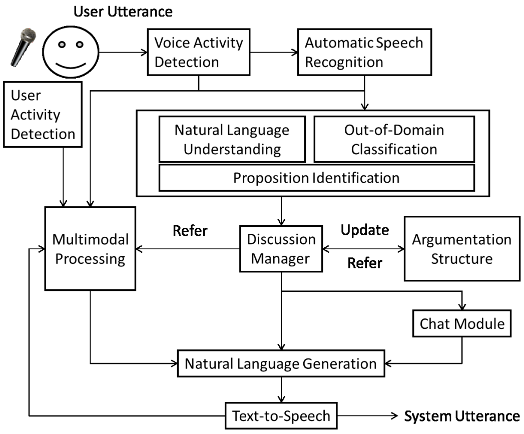


Figure 3: System architecture of our argumentative dialogue system (see (Higashinaka et al., 2017b) for details)

As our next step towards more engaging dialogue, we have been focusing on argumentative dialogue systems, in which users can engage in a discussion on a certain topic. Although much work has been done in argumentation mining (Lippi and Torroni, 2016), there has been less research on automated dialogue systems that can participate in discussion with users. We created large-scale "argumentation structures" as the knowledge of a system to conduct discussion. Our system uses such structures to generate supporting/non-supporting utterances as well as to keep track of the discussion.

Figure 3 shows the architecture of our argumentative dialogue system. At the core of the system is the argumentation structure, which is updated during the discussion, and from which the system's premises are generated.

2.2.1. Argumentation structures

We use a simplified version of the argumentation model described in (Gordon et al., 2007; Walton, 2013). The model has a graph structure, and nodes represent premises and edges represent support/non-support relationships between nodes. Each node has a natural language statement representing the content of its premise. We manually created several large-scale argumentation structures with each structure having more than 2,000 nodes. Each structure has two parts represented by main-issue nodes that enable the system to have opposing stances. Below the main-issue nodes, there are what we call viewpoints nodes that represent conversational topics. Under each viewpoint node, there are premise nodes that represent statements regarding each topic (see (Sakai et al., 2018) for details of our argumentation structures).

2.2.2. Performances

We integrated our argumentative system with an android and conducted a live demonstration at SXSW 2017 (See Figure 4). In our demonstration, two robots having opposite stances on a topic (e.g., which is the better living environment, east or west coast?) and three humans participated in a discussion[2]. Although there was some difficulty in controlling such multi-party conversation, since the argumentation structure was keeping track of the discussion and was updated appropriately on the basis of the utterances of the participants, we managed to conduct a reasonable demonstration.

Figure 4: Two androids discussing with three humans at SXSW 2017. ©2016-2017 SXSW, LLC. This research was conducted in collaboration with Ishiguro laboratory of Osaka University.

2.3. Problems with our current systems

In our efforts in building chat-oriented and argumentative dialogue systems, we encountered the following difficulties.

- Since our systems are working on the text level, it was difficult to distinguish nuances in speech. For example, we expect question marks at the end of an utterance for a question in text, but it is not present in speech. Such para-linguistic information should be incorporated when considering the integration of text-based systems with androids that work on speech.
- We had difficulty in turn-taking, especially in detecting whether the user was willing to start speaking and whether the user had finished speaking. This is related to the first issue; we need to use much richer information about multi-modality for better interaction.
- Emotion is an important issue in chat-oriented dialogue systems. Our system was not aware of user emotion, but we encountered cases in which users were not willing to continue with the current conversational topic. In such cases, it will be necessary to detect the emotion of users and change the current topic appropriately.
- In our argumentative dialogue system, it was rather difficult for humans to continue the discussion smoothly, even though we had large-scale argumentation structures. We believe this is mainly due to the difference in mental models between the system and humans. We need to find ways for humans and a system to have common conceptions and build common ground (Clark et al., 1991) so that discussion participants can build arguments on what has been discussed.

We are currently working with teams investigating paralinguistics and multi-modality to cope with the issues related to turn-taking and emotion. We are also considering

[2] https://www.youtube.com/watch?v=EpgBqjViyZE

ways for the system to disclose its personality, including its way of thinking, so that a common ground can be built and smooth discussion can be carried out.

3. Deployment of dialogue systems

Alongside our research, we have also been conducting field trials of dialogue systems, i.e., deployment of dialogue systems in the wild. We describe two case studies we conducted in Japan. The systems deployed are simple scenario-based systems so that it would be easy to customize them to make them fit actual environments and modify behaviors when necessary. In both cases, thousands of users used the deployed systems. We now describe the details of the field trials and their findings.

3.1. Case study 1

The first trial was conducted with NTT East Corporation and the Tokyo Chamber of Commerce. We placed Sota communication robots [3] on six different premises in Shinjuku, Tokyo, e.g., a fruit parlor, food company, book store, and department store. The robots were installed so that they could give guidance regarding the premises to their customers. The dialogue system is fully scenario-based. When the robot senses a customer with a human sensor, it addresses the customer and makes a greeting (opening phase). Then, the robot asks him/her if he/she had anything to ask about the premise. The system has a touch display to show the information asked by the customer (guidance phase). At the end of the interaction, the robot asked the customers for their level of satisfaction through a questionnaire and says good-bye to the user (closing phase). Figure 5 shows Sota on premises in the field trial.

For a period of four weeks, Sota attracted over 9,000 customers, out of which, about 4600 underwent the opening-phase of the dialogue (roughly one minute of interaction). About 4250 of these customers listened to the guidance from the robot, and about 1800 participated in the questionnaire at the closing phase. The averaged interaction time with the robot was just about one minute. Figure 6 shows the percentages of a three-scale evaluation (good, okay, bad evaluations) of the system through the questionnaire. When they reached the end of a dialogue, it seemed that many of the customers were satisfied with the system.

We asked the store owners/managers (N=14) about how they agreed with the following questionnaire items on a four-point Likert scale. The last question was answered with specific monetary values. Figures 7 and 8 show the results of the following questionnaire items:

Cost reduction The system contributed to the reduction in the cost (e.g., personnel expenses).

Sales increase The system contributed to an increase in sales.

PR effect The system had a positive PR effect.

Satisfaction It was a good idea to install the robot on my premise.

Future use I want to continue having the robot on my premise.

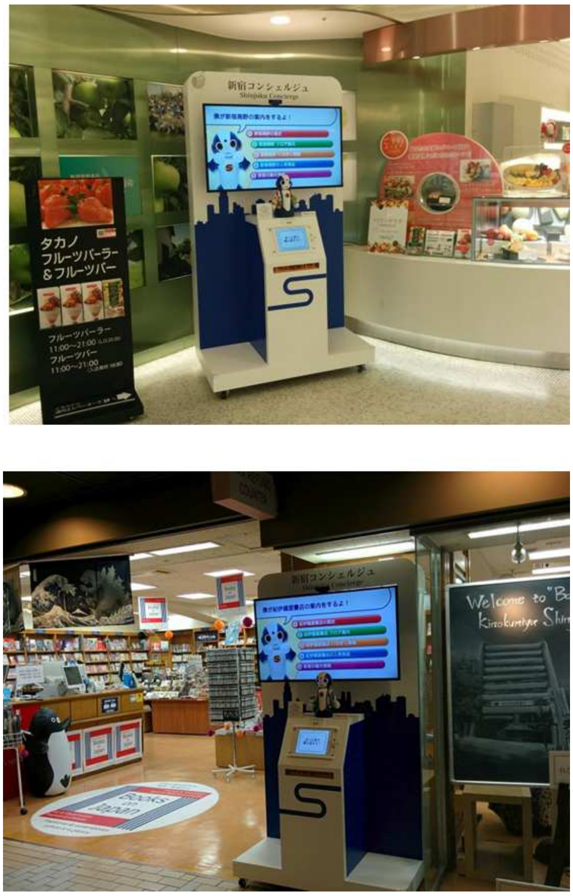


Figure 5: Sota on premises in Shinjuku, Japan

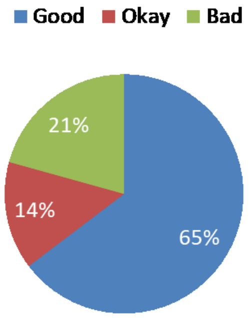


Figure 6: Questionnaire results from customers

Affordable cost How much can you afford to pay per month to have a robot on your premise? (for this item, N=13)

It can be seen that the store owners/managers were rather negative regarding the robot's effect on cost reduction and sales increase, although they felt it certainly had a positive PR effect. Overall, they were positive about having the robot on their premises and wanted to continue using it. One very interesting result was affordability. Most said they could only pay less than 30,000 yen (about 280 USD) per month, which is low compared to the cost of development, deployment, and maintenance.

[3] https://sota.vstone.co.jp/home/

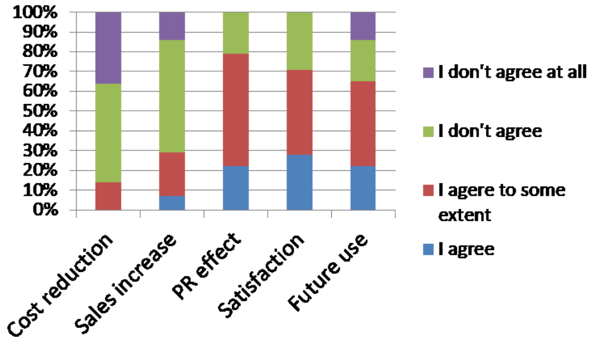


Figure 7: Questionnaire results from store owners/managers

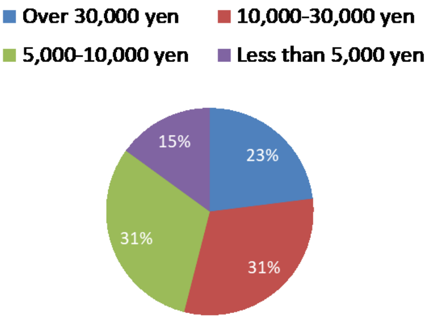


Figure 8: Questionnaire results regarding affordability (the price that owners/managers can afford for having the robot)

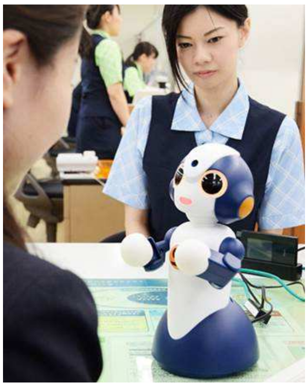


Figure 9: Sota at a regional bank

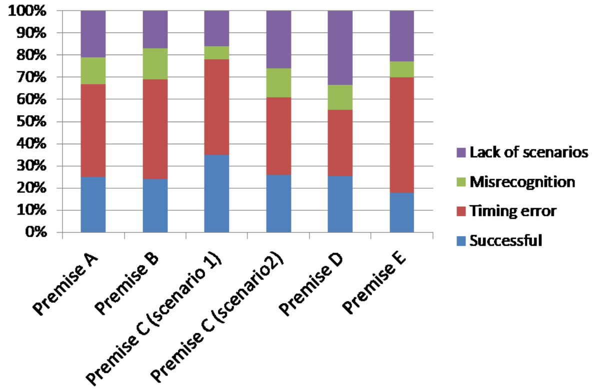


Figure 10: Interaction summary between Sota and customers

We encountered the following problems from this trial:

- The responsiveness of the robot should be improved. Speech-recognition accuracy is also a problem in actual noisy environments.
- The system has to cope with multiple languages of foreign customers. It is also necessary to cope with multiple customers at a time.
- The system needs to cope with nuances and emotional utterances.
- In addition to the information of the premises, the system was sometimes requested to provide information about neighboring areas and should cope with such requests.
- The system had limited information about the premises; it was necessary to show more detailed information when requested.

We learned many lessons from this trial. Although the system does not help from the sales point of view, the system was regarded to have some positive PR effect. Technically, the basic capability of the system needs to be improved, especially regarding responsiveness.

3.2. Case study 2

We conducted another trial involving several regional banks in Japan. This trial was conducted by NTT Data Corporation and several regional banks in Japan. Sota was installed on premises and interacted with customers to provide information. The robot could answer questions about housing loans, education loans, and other products by using a scenario. During a period of about four months, Sota interacted with over 8,000 customers, out of which several thousand engaged in verbal interaction with Sota. Figure 9 shows Sota interacting with a customer.

Figure 10 shows the summary of dialogues on five different premises, showing the percentage of successful (requested information was successfully provided to customer) and unsuccessful dialogues. For unsuccessful cases, the breakdown of the reasons (timing error, misrecognition, and lack of scenarios) are shown. It can be seen that the interactions were not very successful; about one fourth were successful. When we look at the breakdown of errors, we see that most of the errors were due to timing; the system could not respond to customers appropriately because it could not talk/listen to the customers at the right moment; when we listened to the recorded voices, we found that many were fragmented, with many initial parts stripped. This indicates that the customers started speaking, although the system was not ready for speech recognition. Compared to the timing issue, speech-recognition error was not a serious problem, although we should have prepared more scenarios to cope with more questions.

We encountered the following problems from this trial:

- The timing of the robot was the most serious issue. The customers were not aware of the robot's capability and interacted with the robot based on their sense of timing. The customers also had difficulty figuring out what they could do with the robot. It is necessary to explicitly state their functions, and if possible, the robot should act more proactively to provide information.
- The scenarios should be improved; it is necessary to add words/phrases and questions that were not included in the scenarios on a daily basis.

In this trial, we learned that timing was a primary issue with current dialogue systems when they are deployed on actual premises; this is in line with case study 1 in which we had an issue with responsiveness. In our research prototypes, we also had difficulty regarding timing when our chat-oriented/argumentative dialogue systems were built into androids. For deploying systems in the wild, timing has to be the primary concern.

4. Towards better timing

We have started working on multi-modal processing for better timing.

To make an utterance at an appropriate timing, it is necessary to estimate the end of an utterance of a user, queue of turn-taking from the user, and how long after the previous utterance to start speaking. The key is not only language information but also various nonverbal behaviors. For example, it is known that nonverbal behaviors, such as eye-gaze, head movement, breathing motion, and mouth movement, are useful in estimating the timing of turn-taking and appropriate utterances (Ishii et al., 2016a; Ishii et al., 2015; Ishii et al., 2016b; Ishii et al., 2016c; Ishii et al., 2017).

To estimate the appropriate timing more accurately, it is necessary to focus on more diverse nonverbal behaviors. In addition, there are many individual differences in nonverbal behavior depending on personality. There has not been sufficient research on the relationship between such nonverbal behavior and personal characteristics. To clarify the relationship between the proper timing of utterance and various and detailed nonverbal behaviors and to deal with personal characteristics and nonverbal behaviors, we are working on building a multi-modal corpus including various nonverbal behaviors and personal characteristics.

To construct a Japanese-conversation corpus including verbal and nonverbal behaviors in dialogue, we recorded 24 face-to-face two-person conversations (12 groups of two different people). The participants were Japanese males and females in their 20s to 50s who had never met before. They sat facing each other (Figure 11).

To acquire data of various dialogue scenes, three dialog scenes, i.e., discussion, chat, and story-telling, were recorded. In the story-telling scene, the participants had not seen the conversational content. Before the dialogue, they watched a famous popular cartoon animation called "Tom & Jerry" in which the characters do not speak. In each dialogue, one participant explained the content of the animation to the conversational partner. In each group, one session of discussion and chat and two sessions of story-telling were carried out.

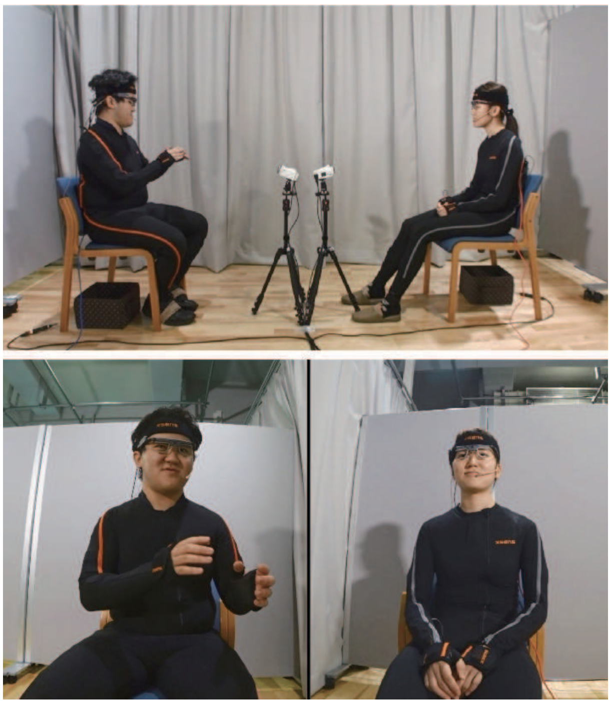


Figure 11: Two participants having dialogue

We recorded the participants' voices with a pin microphone attached to the chest and videoed the entire discussion. We also took bust (chest, shoulders, and head) shots of each participant (recorded at 30 Hz). In each dialogue, the data on the utterances and nodding behaviors of the person explaining the animation were collected in the first half of the ten-minute period (480 minutes in total) as follows.

- Utterances: We built an utterance unit using the inter-pausal unit (IPU) (Koiso et al., 1998). The utterance interval was manually extracted from the speech wave. A portion of an utterance followed by 200 ms of silence was used as the unit of one utterance.
- Gaze: The participants wore a glass-type eye tracker (Tobii Glass2). The gaze target of the participants and the pupil diameter were measured at 30 Hz.
- Body motion: The participants' body movements, such as hand gestures, upper body, and leg movements, were measured with a motion capture device (Xsens MVN) at 240 Hz.
- Personal trait: We obtained Big Five personality scores of the participants through subjective evaluation from the participants and a third party.

All verbal and nonverbal behavior data were integrated at 30 Hz for display using the ELAN viewer (Wittenburg et al., 2006). This viewer enabled us to annotate the multi-modal data frame-by-frame and observe the data intuitively.

In the future, we will clarify the relationship between the proper timing of an utterance and various and detailed nonverbal behaviors. We also want to deal with personal characteristics.

5. Summary and future work

We presented our research on chat-oriented and argumentative dialogue systems. We also described two case studies, one on various premises in Tokyo and the other in regional banks; we found that it is still a premature phase for systems to reduce cost or increase sales, but it seems that they have a positive PR effect. The current main problem of dialogue systems, in research and deployment alike, is timing. To this end, we started to work on multi-modal processing so that a system and users can interact more smoothly.

6. Acknowledgments

We would like to thank NTT East Corporation and Tokyo Chamber of Commerce for sharing the results of field trials. We also thank the members of NTT Data Corporation and the banks who participated in the trial for providing valuable data.

7. Bibliographical References

Clark, H. H., Brennan, S. E., et al. (1991). Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149.

Gordon, T. F., Prakken, H., and Walton, D. (2007). The carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10):875–896.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y. (2014). Towards an open-domain conversational system fully based on natural language processing. In *Proc. COLING*, pages 928–939.

Higashinaka, R., Meguro, T., Sugiyama, H., Makino, T., and Matsuo, Y. (2015). On the difficulty of improving hand-crafted rules in chat-oriented dialogue systems. In *Proc. APSIPA*, pages 1014–1018.

Higashinaka, R., Funakoshi, K., Inaba, M., Tsunomori, Y., Takahashi, T., and Kaji, N. (2017a). Overview of dialogue breakdown detection challenge 3. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Higashinaka, R., Sakai, K., Sugiyama, H., Narimatsu, H., Arimoto, T., Fukutomi, T., Matsui, K., Ijima, Y., Ito, H., Araki, S., Yoshikawa, Y., Ishiguro, H., and Matsuo1, Y. (2017b). Argumentative dialogue system based on argumentation structures. In *Proc. SemDial*, pages 154–155.

Imamura, K., Higashinaka, R., and Izumi, T. (2014). Predicate-argument structure analysis with zero-anaphora resolution for dialogue systems. In *Proc. COLING*, pages 806–815.

Ishii, R., Kumano, S., and Otsuka, K. (2015). Predicting next speaker based on head movement in multi-party meetings. In *Proc. ICASSP*, pages 2319–2323.

Ishii, R., Kumano, S., and Otsuka, K. (2016a). Analyzing mouth-opening transition pattern for predicting next speaker in multi-party meetings. In *Proc. ICMI*, pages 209–216.

Ishii, R., Otsuka, K., Kumano, S., and Yamamoto, J. (2016b). Predicting of who will be the next speaker and when using gaze behavior in multiparty meetings. *The ACM Transactions on Interactive Intelligent Systems*, 6(1):4.

Ishii, R., Otsuka, K., Kumano, S., and Yamamoto, J. (2016c). Using respiration to predict who will speak next and when in multiparty meetings. *The ACM Transactions on Interactive Intelligent Systems*, 6(2):20.

Ishii, R., Kumano, S., and Otsuka, K. (2017). Prediction of next-utterance timing using head movement in multiparty meetings. In *Proc. HAI*, pages 181–187.

Koiso, H., Horiuchi, Y., Tutiya, S., and Akira Ichikawa, a. Y. D. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech*, 41:295–321.

Koiso, H., Tsuchiya, T., Watanabe, R., Yokomori, D., Aizawa, M., and Den, Y. (2016). Survey of conversational behavior: Towards the design of a balanced corpus of everyday japanese conversation. In *Proc. LREC*.

Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):10:1–10:25.

Meguro, T., Higashinaka, R., Minami, Y., and Dohsaka, K. (2010). Controlling listening-oriented dialogue using partially observable Markov decision processes. In *Proc. COLING*, pages 761–769.

Miltsakaki, E., Prasad, R., Joshi, A. K., and Webber, B. L. (2004). The Penn Discourse Treebank. In *Proc. LREC*.

Onishi, K. and Yoshimura, T. (2014). Casual conversation technology achieving natural dialog with computers. *NTT DOCOMO Technical Jouranl*, 15(4):16–21.

Otsuka, A., Hirano, T., Miyazaki, C., Higashinaka, R., Makino, T., and Matsuo, Y. (2017). Utterance selection using discourse relation filter for chat-oriented dialogue systems. In *Dialogues with Social Robots*, pages 355–365. Springer.

Sakai, K., Inago, A., Higashinaka, R., Yoshikawa, Y., Ishiguro, H., and Tomita, J. (2018). Creating large-scale argumentation structures for dialogue systems. In *proc. LREC*.

Shang, L., Sakai, T., Lu, Z., Li, H., Higashinaka, R., and Miyao, Y. (2016). Overview of the NTCIR-12 short text conversation task. In *Proc. NTCIR*.

Takeuchi, S., Cincarek, T., Kawanami, H., Saruwatari, H., and Shikano, K. (2007). Construction and optimization of a question and answer database for a real-environment speech-oriented guidance system. In *Proc. Oriental COCOSDA*, pages 149–154.

Vinyals, O. and Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Wallace, R. S. (2009). The anatomy of alice. In *Parsing the Turing Test*, pages 181–210. Springer.

Walton, D. (2013). *Methods of argumentation*. Cambridge University Press.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN a professional framework for multimodality research. In *Proc. LREC*.