Spontaneous speech resources in Japan

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.