CPJD Corpus: Crowdsourced Parallel Speech Corpus of Japanese Dialects

Shinnosuke Takamichi, Hiroshi Saruwatari
Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
{shinnosuke_takamichi, hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp

Abstract

Public parallel corpora of dialects can accelerate related studies such as spoken language processing. Various corpora have been collected using a well-equipped recording environment, such as voice recording in an anechoic room. However, due to geographical and expense issues, it is impossible to use such a perfect recording environment for collecting all existing dialects. To address this problem, we used web-based recording and crowdsourcing platforms to construct a crowdsourced parallel speech corpus of Japanese dialects (CPJD corpus) including parallel text and speech data of 21 Japanese dialects. We recruited native dialect speakers on the crowdsourcing platform, and the hired speakers recorded their dialect speech using their personal computer or smartphone in their homes. This paper shows the results of the data collection and analyzes the audio data in terms of the signal-to-noise ratio and mispronunciations.

Keywords: crowdsourcing, Japanese dialect, speech, transcription

1. Introduction

A public corpus of low-resourced languages can accelerate related studies, such as natural language, spoken language, and speech signal processing. Thanks to recent improvements in machine learning techniques for these research areas (v. d. Oord et al., 2016; Hinton et al., 2012; Takamichi et al., 2017), the accuracies in the rich-resourced languages (such as English, Chinese, and Japanese) have become high, and attention is shifting to more challenging tasks, including language and speech processing of dialects. Dialect processing is one of the topics being actively targeted in machine translation (Salloum et al., 2014), speech recognition (Hirayama et al., 2014), speech perception (Jacewicz and Fox, 2017), and speech synthesis (Masmoudi et al., 2016). For studies on Japanese dialect, (Yoshino et al., 2016) collected a parallel database of Japanese dialects with voice recording in their well-equipped recording environment. However, geographical and expense issues make it impossible to use such a perfect recording environment for all existing Japanese dialects.

In this work, we collected parallel data of the common language and several dialects in Japanese using a web-based recording and crowdsourcing platform. We recruited native dialect speakers on the crowdsourcing platform, and the hired speakers of 21 dialects uttered 250 sentences for each. They also converted text in the common language into their dialect, and read the dialect text in the web-based recording platform available on personal computers and smartphones. Thus, the resultant resources consist of texts (one common language and 21 dialects), nine hours of speech data, and their geographic contexts (i.e., areas of the dialect). We analyzed the collected speech data from perspectives of the speech signal and spoken language processing. This paper shows the signal-to-noise ratio (SNR) results for the recording environments and the number of mispronunciations.

2. Existing resources of Japanese dialects and their problems

Several text/speech corpora of Japanese dialects have been compiled (Kubozomo, 2001 2008; National Institute for Japanese Language and Linguistics, 2016). For example, the database at National Institute for Japanese Language and Linguistics (National Institute for Japanese Language and Linguistics, 2016) includes text and speech data spoken by native dialect speakers. However, the recording setting is not suitable for spoken language processing. For instance, it is hard for current acoustic modeling techniques to model such spontaneous speech recorded in such a poor environment. Yoshino et al. (Yoshino et al., 2016) collected parallel data of Japanese dialects. They constructed parallel text between the common language and dialects in Japanese and had native dialect speakers read the dialect texts in their recording studio. Because the reading-style speech data was recorded in the well-equipped recording environment, their corpus is useful for spoken language processing research. In addition, because their corpus includes the parallel text of the common language and dialects, it is also useful for natural language processing research. However, dialects that can be collected in such a perfect environment are very limited and collecting many dialects is unrealistic because of geographical and expense issues. For example, travel expenses from the speaker's hometown (e.g., in the countryside where a rare dialect is spoken) to the recording studio have to be covered, and the cost significantly increases when the variety of dialects to record is large. Also, elderly people of countrysides is hard to come to the recording studio due to their physical burden.

Crowdsourcing can greatly reduce the time and money needed for building speech corpora (Hughes et al., 2010; Gutkin et al., 2016). The crowdsourced data is noisy compared to the data obtained in the perfect environment, but it is expected to solve above issues. In this work, using a web-based recording platform, we collected the speech data recorded in an indoor recording environment with comparably stationary audio noise. This is because the effects of such environments can be alleviated more easily than (National Institute for Japanese Language and Linguistics, 2016). Moreover, we collected small amounts of speech data of a variety of dialects rather than large amounts of speech data of one dialect. This is because there is a

strong relationship between the it is known that phonetic and prosodic properties of dialects are strongly related to the geographical relationship (Preston, 1999), and mixed dialect models that take into account this relationship will improve the accuracy of dialect modeling.

3. Collection of the CPJD corpus

This section explains how we collected the CPJD corpus. First, we prepared sentences of the common language in Japanese so that the native dialect speakers could convert it to their dialects. Then, the speakers were recruited on the crowdsourcing platform, and the hired speakers converted the sentences and read the dialect sentences. Finally, we constructed the corpus comprising dialect sentences, their read speech, the name of the dialect, and the geographic contexts of the dialect.

3.1. Sentences of the common language

We selected sentences of the common language so that the native dialect speakers could convert then into their dialects. These sentences were randomly selected from a blog category of the BCCWJ (Balanced Corpus of Contemporary Written Japanese) corpus (Maekawa, 2014) and the KNB (Kyoto University and NTT Blog) corpus (Hashimoto et al., 2011). We expect the blog sentences are comparably informal and suitable for speaking informal dialects. Sentences including geographic words, e.g., a place name, were removed. In addition, words that are unsuitable for current daily life were changed into suitable ones, e.g., "cell phone" was changed to "smartphone."

3.2. Web-based recording platform

We prepared a web-based speech recording platform available on personal computers and smartphones. Using *Recorder.js*[1], we prepared buttons to start or stop an audio recording. The platform also has a spectrum and waveform plots so that speakers can check their voice. The spectrum is produced in real time during a recording, and the waveform is produced after the recording. Automatic mispronunciation detection, voice activity detection, noise detection, and voice gain control function were not implemented.

3.3. Recruiting native dialect speakers

We hired native dialect speakers to collect text and audio data. Before hiring them, we asked recruited speakers to tell us the name of their dialect and its prefecture on the crowdsourcing platform, and we hire some of the speakers so that many areas of the country would be covered. The hired speakers first converted sentences of the common language into their dialect and then read the converted sentences on the web-based recording platform. The conversion and recording were done in their indoor recording environments. The speakers were instructed not to include extraneous sounds like coughing in their recordings, and we let the speakers convert honorific expressions in the common language text into plain expressions to make it easier for them to convey their dialect.

[1] `https://github.com/mattdiamond/Recorderjs`

Table 1: List of collected dialects. The numbers correspond to the prefecture index in Fig. 1. Each dialect was spoken by just one native speaker (except Nara-ben). (The suffix "ben" means dialect. For example, "Hokkaido-ben" is a dialect of Hokkaido.)

Area	Dialect
Hokkaido	Hokkaido-ben (1)
Tohoku	Tsugaru-ben (2), Akita-ben (3), Iwaki-ben (4)
Kanto	Saitama-ben (5)
Chubu	Kanazawa-ben (6), Tohshuu-ben (7), Hukui-ben (8)
Kinki	Kyoto-ben (9), Kyo-kotoba (9), Nara-ben (10), Osaka-ben (11)
Chugoku	Okayama-ben (12), Izumo-ben (13), Hiroshima-ben (14)
Shikoku	Tosa-ben (15), Awa-ben (16), Iyo-ben (17)
Kyushu	Fukushima-ben (18), Miyazaki-ben (19), Morokata-ben (19, 20)
Okinawa	None

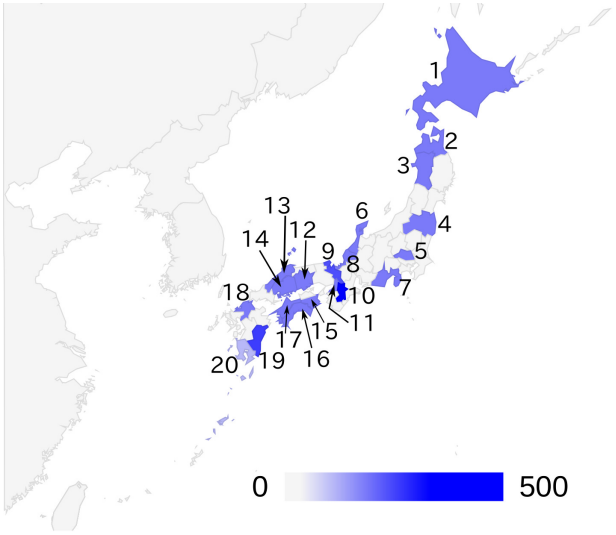

Figure 1: The number of collected utterances for each prefecture in Japan. The map chart was drawn using Google chart API (`https://developers.google.com/chart/`), but we moved the position of the legend and removed some areas outside of Japan for a clear illustration. The prefecture boundary is not always the same as the boundary of the dialect's area.

3.4. Data correction

After data collection, we manually identified misrecorded or mispronounced (e.g., hesitation or filler) data and manually added commas between breath groups.

Table 2: Examples of parallel text data. For comparison, the text translated into English is also listed. As described in Section 3.4, we corrected the text data after data collection, but the texts shown here was not corrected.

Dialect	Texts	Phonemes
Common	できるだけスマートフォンひとつで身の回りのこと全て片付けようとしているようだ.	d e k i r u d a k e s u m a: t o f o N h i t o ts u d e m i n o m a w a r i n o k o t o s u b e t e k a t a z u k e y o u t o sh i t e i r u y o u d a
Miyazaki-ben (19)	なるだけスマートフォンひとつで身んまわりんこと全部片付けようとしちょるみたいやね.	n a r u d a k e s u m a: t o f o N h i t o ts u d e m i N m a w a r i N k o t o z e N b u k a t a z u k e y o u t o sh i ch o r u m i t a i y a n e
Tsugaru-ben (2)	でぎるだげスマートフォンばりで身の回りのことまるっととっけるんた.	d e g i r u d a g e s u m a: t o f o N b a r i d e m i n o m a w a r i n o k o t o m a r u q t o t o q k e r u N t a
Cf. Translation	It seems like everything around here is done as much as possible on a smartphone.	-
Common	これからこの機能が加わったからといって特別ハッピーなわけでもない.	k o r e k a r a k o n o k i n o u g a k u w a w a q t a k a r a t o i q t e t o k u b e ts u h a q p i: n a w a k e d e m o n a i
Kyo-kotoba (9)	これからこの機能が加わったからゆうて特別ハッピーなわけでもあらへん.	k o r e k a r a k o n o k i n o u g a k u w a w a q t a k a r a y u u t e t o k u b e ts u h a q p i: n a w a k e d e m o a r a h e N
Awa-ben (16)	これからほの機能が加わったからといって特別ハッピーなわけやないし.	k o r e k a r a h o n o k i n o u g a k u w a w a q t a k a r a t o i q t e t o k u b e ts u h a q p i: n a w a k e y a n a i sh i
Cf. Translation	I would not be especially happy even if this function were added in the future.	-

4. Analysis of the CPJD corpus

4.1. Corpus specification and examples

To recruit the native dialect speakers, we used the crowdsourcing service Lancers[2], which is one of the biggest crowdsourcing platforms in Japan. To cover a large number of areas, we asked speakers to name their dialect and home prefecture, and hired speakers who can speak dialects we had not yet collected. The recruiting period ran for five days during April and May 2017. Hired speakers were paid approximately $45 for their work. The number of common language sentences to be converted and read was 250. The audio sampling rate was 44.1 or 48 kHz. The number of hired speakers was 22, nine male and 13 female speakers. Gender labels were added by our annotator after recording. The number of prefectures where dialects were collected was 20. The average duration of recorded speech for each speaker was 24 minutes and 36 seconds (including non-speech regions). The total duration was approximately nine hours. Compared to (Yoshino et al., 2016) (3.5 hours of speech data, eight prefectures), the CPJD corpus contains large amounts of speech data and covers more prefectures than previous corpora do.

Table 1 lists the collected dialects, and Fig. 1 maps their areas. We can see that a variety of areas and dialects were collected in this work. Table 2 shows examples of collected sentences and transcribed phonemes. The meanings of sentences are the same; however, there are small differences in minor word choices.

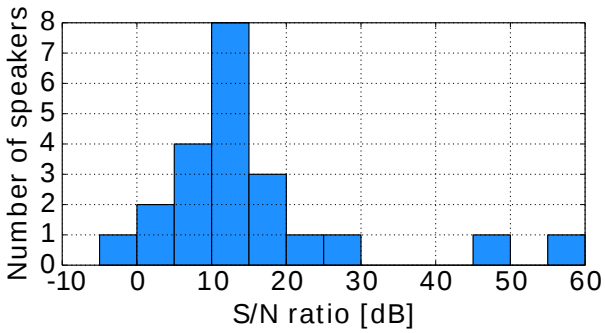

Figure 2: Histogram of SNRs. The values are averaged for each speaker.

Table 3: Statistics of SNR [dB]. The histogram is shown in Fig. 2.

Worst	Best	Mean	Median
-2.1	57.8	15.1	12.6

4.2. Analysis

We analyzed the collected audio data. First, for use in speech signal processing research, we calculated SNR (signal-to-noise ratio) of the speech data using the decision-directed method (Plapous et al., 2006). The high SNR means that speech was recorded in the clean (well-equipped) environment. The histogram of SNRs for each speaker is shown in Fig. 2, and the SNR statistics are summarized in Table 3. We can see that two speakers had a

[2] http://www.lancers.jp

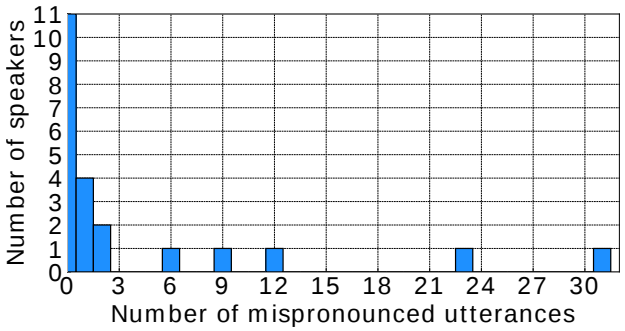

Figure 3: Histogram of mispronounced utterances. The values are averaged for each speaker. Automatic mispronounciation detection was not used for the recordings. Mispronounce were manually labeled after the recordings.

good recording environment (45-through-60 dB), but most of the speakers environments were less than ideal.

Next, for use in spoken language processing research, we calculated the number of mispronounced utterances. We manually detected mispronunciations where the text and speech were mismatched. Fig. 3 shows the histogram of the number of mispronounced utterances for each speaker. Half of the speakers made no mispronunciations, but a few speakers made more than 20 mistakes within their 250 utterances. This suggests the need to improve the recording platform.

5. Conclusion

We constructed a corpus of parallel data of Japanese dialects, called the CPJD corpus, using web-based recording and crowdsourcing platforms. The corpus contains sentences of one common language and 21 dialects in Japanese, nine hours of speech data read by dialect speakers, and their geographic contexts. We analyzed the collected speech data in terms of signal-to-noise ratio and the number of mispronounced utterances. In future work, we will build speech synthesis systems using the CPJD corpus and do the preparation to make the corpus available to the public.

Acknowledgment

Part of this work was supported by the SECOM Science and Technology Foundation.

6. Bibliographical References

Gutkin, A., Ha, L., Jansche, M., Pipatsrisawat, K., and Sproat, R. (2016). TTS for low resource languages: A Bangla synthesizer. In *Proc. LREC*, pages 2005–2010, Paris, France.

Hinton, G., Deng, L., Yu, D., Dahl, G., r. Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine of IEEE*, 29(6):82–97.

Hirayama, N., Yoshino, K., Itoyama, K., Mori, S., and Okuno, H. G. (2014). Automatic speech recognition for mixed dialect utterances by mixing dialect language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32(2):373–382.

Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P., and LeBeau, M. (2010). Building transcribed speech corpora quickly and cheaply for many languages. In *Proc. INTERSPEECH*, pages 1914–1917, Chiba, Japan, Sep.

Jacewicz, E. and Fox, R. A. (2017). Dialect perception by older children. In *Proc. INTERSPEECH*, pages 354–358, Stockholm, Sweden, Aug.

Masmoudi, A., Ellouze, M., Bougares, F., Esètve, Y., and Belguith, L. (2016). Conditional random fields for the tunisian dialect grapheme-to-phoneme conversion. In *Proc. INTERSPEECH*, pages 1457–1461, San Francisco, U.S.A., Sep.

Plapous, C., Marro, C., and Scalart, P. (2006). Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2098–2108.

Preston, D. R. (1999). *Handbook of perceptual dialectology*, volume 1. John Benjamins Publishing.

Salloum, W., Elfardy, H., Alamir-Salloum, L., Habash, N., and Diab, M. (2014). Sentence level dialect identification for machine translation system selection. In *Proc. ACL*, volume 2 (Short Papers), pages 772–778, Baltimore, Maryland.

Takamichi, S., Tomoki, K., and Saruwatari, H. (2017). Sampling-based speech parameter generation using moment-matching network. In *Proc. INTERSPEECH*, pages 3961–3965, Stockholm, Sweden, Aug.

v. d. Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. abs/1609.03499.

Yoshino, K., Hirayama, N., Mori, S., Takahashi, F., Itoyama, K., and Okuno, H. G. (2016). Parallel speech corpora of japanese dialects. In *Proc. LREC*, pages pp. 4652–4657, Portoroz, Slovenia, May.

7. Language Resource References

C. Hashimoto and S. Kurohashi and D. Kawahara and K. Shinzato and M. Nagata. (2011). *Construction of a Blog Corpus with Syntactic, Anaphoric, and Sentiment Annotations* `http://nlp.ist.i.kyoto-u.ac.jp/kuntt/#ga739fe2`. Journal of Natural Language Processing (in Japanese).

H. Kubozomo. (2001-2008). *Accent Database of Koshikijima Japanese,* `http://koshikijima.ninjal.ac.jp/`.

K. Maekawa. (2014). *Balanced corpus of contemporary written Japanese* `http://pj.ninjal.ac.jp/corpuscenter/bccwj/`. Proceedings of the 6th Workshop on Asian Language Resources.

National Institute for Japanese Language and Linguistics. (2016). *Database of Spoken Dialects all over Japan: Collection of Japanese Dialects* `http://pj.ninjal.ac.jp/publication/catalogue/hogendanwa_db/outline.htm` *(in Japanese).*