ASJ Continuous Speech Corpus

— Japanese Newspaper Article Sentences (JNAS) —

June 1997
(revised Nov. 2014)
Shuichi ITAHASHI

1 Outline

This corpus consists of 6 DVD-Rs. It contains speech recordings and their orthographic transcriptions of 306 speakers (153 males and females each) reading excerpts from the Mainichi Newspaper and the ATR 503 PB-Sentences. All utterances and sentences are in the Japanese language.

We prepared 155 text sets. Each set consists of about 100 sentences from the Mainichi Newspaper. As a general rule, each text set was read by one male and one female. Every speaker also read any subset of the ATR 503 PB-Sentences (about 50 sentences for each subset). That is, this corpus contains utterances of about 45,000 sentences as a whole with all speakers reading about 150 sentences each.

Each utterance was recorded with two microphones: a head-set microphone (all recoding sites used Sennheiser HMD410/HMD25-1 or the equivalent) and a desk-top microphone of different types at each site (Sanken, Sony, and so on). These two-microphone data were stored into separate files and have a parallel directory structure in the DVD-R directories; three of the discs (Vol.1 through Vol.3) contain the head-set-microphone data and the other (Vol.4 through Vol.6) the desk-top-microphone data.

The speech waves were sampled at 16 kHz and quantized into 16 bits. They are stored in the wav file format.

The corpus includes text sets for reading, orthographic transcriptions of the speech data and the bigram language models for the Mainichi Newspaper articles from which the prompting text was selected. These materials are contained in Vol.1 and Vol.4.

The Speech Database Committee of the Acoustical Society of Japan, established in July 1990, has discussed the design and creation of this corpus, which has been recorded in collaboration with 39 institutions. The recording and AD conversion characteristics, including low-pass filter characteristics, are not necessarily unified.

2 Sentences of the Mainichi Newspaper Articles

The Large Vocabulary Continuous Speech Database Working Group of the

Information Processing Society of Japan established in November, 1995, selected 155 text sets for reading, using articles of the Mainichi Newspaper issued during 1991-1994.

A bigram language model was estimated from the articles of 45 months with their morphological information taken from RWCP text corpus (RWC-DB-TEXT-95-1) which was automatically generated with a morphological analyzer. The CMU SLP toolkit was used for the estimation. Sentences in the articles for three months were classified into 30 categories based on the bigram model. Each category is characterized by the sentence length (2 types), the vocabulary size (5 types) and perplexity (3 types).

A statistically controlled text set consists of 90 sentences (SC-sentences) collected from the categories according to Table 1 and about 10 connected sentences taken from a few paragraphs. 150 text sets of controlled sentences (about 100 sentences each) were prepared. The details of the sentence id in each text set and its category are shown in Table 2.

Table 1. The number of SC-sentences collected from each category

	LENGTH=NORMAL			LENGTH=LONG		
	PERP=LOW	PERP=MID	PERP=HIGH	PERP=LOW	PERP=MID	PERP=HIGH
VOC=MID	2	6	2	1	3	1
VOC=MID+	2	6	2	1	3	1
VOC=LARGE	4	12	4	2	6	2
VOC=LARGE+	2	6	2	1	3	1
VOC=LARGE++	2	6	2	1	3	1

VOC=MID: 5k voc. without an unknown word
VOC=MID+: 5k voc. with one unknown word
LENGTH=NORMAL: 5-19 morphemes
LENGTH=LONG: 20-39 morphemes
PERP=LOW: 0 < perplexity < 40
PERP=MID: 40 <= perplexity < 85
PERP=HIGH: 85 <= perplexity < 400

VOC=LARGE: 20k voc. without an unknown word
VOC=LARGE+: 20k voc. with one unknown word
VOC=LARGE++: 20k voc. with two or more unknown words
LENGTH=NORMAL: 5-29 morphemes
LENGTH=LONG: 30-39 morphemes
PERP=LOW: 0 < perplexity < 70
PERP=MID: 70 <= perplexity < 130
PERP=HIGH: 130 <= perplexity < 400

Table 2. The details of the sentence id in each text set and its category

sentence-id (#)	vocab-class	length-class	perplexity-class
1- 2 (2)	MID	NORMAL	LOW
3- 8 (6)	MID	NORMAL	MID
9-10 (2)	MID	NORMAL	HIGH
11 (1)	MID	LONG	LOW
12-14 (3)	MID	LONG	MID
15 (1)	MID	LONG	HIGH
16-17 (2)	MID+	NORMAL	LOW
18-23 (6)	MID+	NORMAL	MID
24-25 (2)	MID+	NORMAL	HIGH
26 (1)	MID+	LONG	LOW
27-29 (3)	MID+	LONG	MID
30 (1)	MID+	LONG	HIGH
31-34 (4)	LARGE	NORMAL	LOW
35-46 (12)	LARGE	NORMAL	MID
47-50 (4)	LARGE	NORMAL	HIGH
51-52 (2)	LARGE	LONG	LOW
53-58 (6)	LARGE	LONG	MID
59-60 (2)	LARGE	LONG	HIGH
61-62 (2)	LARGE+	NORMAL	LOW
63-68 (6)	LARGE+	NORMAL	MID
69-70 (2)	LARGE+	NORMAL	HIGH
71 (1)	LARGE+	LONG	LOW
72-74 (3)	LARGE+	LONG	MID
75 (1)	LARGE+	LONG	HIGH
76-77 (2)	LARGE++	NORMAL	LOW
78-83 (6)	LARGE++	NORMAL	MID
84-85 (2)	LARGE++	NORMAL	HIGH
86 (1)	LARGE++	LONG	LOW
87-89 (3)	LARGE++	LONG	MID
90 (1)	LARGE++	LONG	HIGH
91- (10+)	(PARAGRAPH)		

Five other text sets are made up of connected sentences chosen from some paragraphs.

The corpus includes two kinds of text sets for reading using newspaper articles. One is the Japanese orthographic text with ruby, used as a prompting text for reading, in the TeX format and its compiled PDF format. The other is the Kanji, Kana, Romaji and Kanji with ruby text.

3 ATR 503 PB-Sentences

These PB-sentences were chosen by ATR Interpreting Telephony Research Laboratories. Entropy was calculated based on the clusters of two phonemes (120 CV's, 227 VC's and 55 VV's, making 402 clusters in all) and three phonemes (69 CVC's where C is an unvoiced consonant, 18 CVC's where C is a nasal consonant and 136 VCV's where C is a semivowel, making 223 clusters in all) on the assumption that they occur independently. 10,196 original sample sentences were extracted at random from newspapers, magazines, novels, letters, text books, etc. Of these, 503 PB sentences were chosen to maximize the entropy. They were sorted so that each set of 50 sentences also be phonetically balanced.

This corpus includes the plain text files of Kanji, Kana and Romaji format.

4 Orthographic transcription

The corpus contains the plain text files of Kanji, Kana and Romaji format that represents pronunciation of each utterance. Each sentence was modified to correctly transcribe recorded utterances according to the check reports from the recording sites.

5 Update Information

28 Nov. 2014	Speech files converted to wav format, flawed speech files resolved, the ruby of text sets for reading and transcriptions corrected, test-set lists added and all documents revised.
10 March 2009	The NIST SPHERE headers removed from speech files.
17 March 2008	Bug information added (only in Japanese).
22 March 2006	Compressed speech files extracted and converted to DVD-Rs.

6 Copyright information

Copyright holders of each part of this corpus are the following.

DATA	COPYRIGHT HOLDER
Speech wave data	The Acoustical Society of Japan (1997)
Newspaper article	The Mainichi Newspapers (1991-1994)
Morphological data and pronunciation of newspaper article	Real World Computing Partnership (1996)
ATR 503 PB-Sentences	ATR Interpreting Telephony Research Laboratories (1988)

Sources of each part of this corpus are the following.

DATA	SOURCES
Newspaper article	CD-Mainichi 1991-1994 (Nichigai Associates, Inc.)
Morphological data and pronunciation of newspaper article	RWC Text Database Ver.1 RWC-DB-TEXT-95-1 (Real World Computing Partnership)

[Notice]

(1) If you use only the data other than speech waves, you need to obtain the copyright holder's permission.

(2) A sentence extracted from newspaper articles does not retain its original sense in the context.

ASJ Continuous Speech Corpus

--- Japanese Newspaper Article Sentences (JNAS) ---

Editor: Speech Database Committee, Acoustical Society of Japan

Publisher: Acoustical Society of Japan

Nakaura 5th-Bldg., 2-18-20 Sotokanda, Chiyoda-ku, Tokyo 101-0021, Japan

Assisted by:

Chiba University
Doshisha University
Kyoto Institute of Technology
Kyoto University
Nagoya University
Nara Institute of Science and Technology
Osaka University
Ryukoku University
Shinshu University
Sizuoka University
Teikyo University of Science & Technology
Tohoku University
Toyohashi University of Technology
University of Electro-Communications
University of Tokyo
University of Tsukuba
Waseda University
Yamagata University
Yamanashi University

Electrotechnical Laboratory

ATR Interpreting Telecommunications Research Laboratories
Canon Inc.
Fujitsu Laboratories Ltd.
Hitachi, Ltd.
Kokusai Denshin Denwa Co., Ltd.
Matsushita Research Institute Tokyo, Inc.
Meidensha Corporation
Mitsubishi Electric Corporation
NEC Corporation
NTT Basic Research Laboratories
NTT Data Corporation
NTT Human Interface Laboratories
- Furui Research Laboratory
- Speech and Acoustics Laboratory
Oki Electric Industry Co., Ltd.
Ricoh Co., Ltd.
Sanyo Electric Co., Ltd.
Sharp Corporation
Sony Corporation
Toshiba Corporation

(DVD version) Produced by: Media Drive Corporation

(Revised version) Produced by: Speech Resources Consortium (NII-SRC)

Acknowledgements:

The prompting texts and the bigram language models of the Mainichi Newspaper article sentences were prepared by the Large Vocabulary Continuous Speech Database Working Group, Special Interest Group of Spoken Language Processing, Information Processing Society of Japan.

We used the NIST SPHERE package for attaching a header to wave files. The NIST SPHERE package was implemented by the Spoken Natural Language Processing Group, National Institute of Standards and Technology, U.S.A.

We would like to thank all the groups and people above.