ASJ Continuous Speech Corpus for Research

The speech corpora are divided into two types. One is a collection of isolated sentences which are phonetically balanced. This corpus is composed of 503 Sentences selected by ATR Interpreting Telephony Research Laboratories and NEC Corporation. Three CD-ROMs were produced.

Another is a corpus which constitutes a story. For this purpose, played conversational speech samples were first recorded onto DAT. They were transcribed into orthographic transcription, which were then rewritten in order to get neater sentences; we deleted various types of insertions including hesitation, interjection and dialectal expression. Some phrases were added if necessary, otherwise we tried to keep flavor of original conversations. Next step is to read and record the rewritten sentences as a story with one speaker. The scale of this corpus is almost the same size as that of ATR PB-Sentences. This corpus was also recorded onto three CD-ROMs.

The speech waves were digitized with 16 kHz sampling frequency and 16 bit quantization. The recording and AD conversion characteristics, including low-pass filter characteristics (flat up to 6 kHz and -40 dB at 8 kHz recommended) are not necessarily unified. The creation of the CD-ROM was carried out by the Japan Information Processing Development Center (JIPDEC). Details are described below.

1. ASJ Continuous Speech Corpus (1): PB Sentences

1.1 Outline

This CD-ROM contains the text and corresponding speech wave of ATR 503 PB-Sentences spoken by 64 speakers, including 30 males and 34 females, making about 9,600 sentences in all. The corpus has been recorded in collaboration with 15 institutions. Speech data were checked by listening and the checklist is stored in a file. Several data were re-recorded after the check, on account of faltering or mispronunciation. In such cases, the original recordings are stored in another directory as supplementary data.

1.2 ATR 503 PB Sentences

An information entropy was calculated based on clusters of two phonemes (120 CV's, 227 VC's and 55 VV's, making 402 clusters in al1) 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. Original 10,196 sample sentences were extracted at random from newspapers, magazines, novels, letters, text books, etc. Of these, 503 sentences were chosen in such a way that its information entropy becomes as large as possible, or phonetically balanced. They were sorted so that each set of 50 sentences was also phonetically balanced.

1.3 Speakers

ATR 503 PB-Sentences are composed of 10 subsets. Subsets of A, B, C, … , I contain 50 sentences and the subset J contains 53 sentences, The sentences in set A were spoken by all speakers, while other sets were spoken by twelve speakers (six males and

six females) as a rule, The details are given in the attached table in a CD-ROM file.

2 ASJ Continuous Speech Corpus (2): Guide Task Sentences

2.1 Outline

This CD-ROM contains a continuous speech (read speech) corpus, a text corpus of played dialogues, and a report. The corpus has been recorded in collaboration with 17 institutions. They are composed of 16 dialogues including 5 sets of geographical guides, 7 sets of tourist guides, 2 sets of music concert guides, one set of passport inquiries and one set of ski tour information. The speech data were checked by listening. The check list is stored in a file. Several data were re-recorded after the check to correct errors due to mispronunciations.

2.2 Various Guide Task Continuous Speech Corpus

This corpus consists of continuous speech of ASJ Various Guide Task 1027 Sentences which were read by 36 speakers, including 18 males and 18 females, making 12,474 sentences in all. The ASJ Various Guide Task Sentences are referred to as ASJ Simulated Dialogue Sentences in CD-ROM's vol.4 and vol. 5. The texts were obtained by eliminating interjections and erroneous expressions from the original transcriptions of played dialogues between two speakers to whom a task of the dialogue had been given in advance. Finally, each dialogue text was read by one speaker.

There are two versions of sentences for two sets of the geographical guides, the Kanazawa tourist guide and the music concert guide. In the first version, deletions and particle ellipses were complemented and inverted expressions were corrected to provide polite expressions. In the second version, deletions and particle ellipses were not complemented and inverted expressions were not corrected.

2.3 Speakers

ASJ Various Guide Task Sentences are composed of 16 subsets k, 1, m,..., z, each containing 30 to 141 sentences. Each set was spoken by 10 to 18 speakers and each speaker spoke 5 to 9 sets. Eight speakers (4 males and 4 females) of the total 36 speakers in the ASJ Simulated Dialogue corpus are the same as those who spoke ATR 503 PB Sentences corpus (Vol. 1 to 3); other speakers in (Vol. 4 to 6) are all different from those in (Vol.1 to 3). Details are given in a CD-ROM file.

2.4 Text Corpus of Simulated Dialogue

Played dialogues about various guide tasks are recorded and transcribed in collaboration with 12 institutions. A task had been given in advance to the two speakers for each dialogue. The transcribed texts of 87 dialogues are included in CD-ROM vol. 6. The number of sentences in each dialogue is from 40 to 270. The Various Guide Task Sentences were obtained by eliminating interjections and erroneous expressions from the original transcriptions of the played dialogues, considering readability, ill-formedness, and politeness,

6.2.5 Report

The documents related with the corpora in vols.1-6 are included in CD-ROM vol. 6. A

short survey on continuous speech corpora, issues on the played dialogues and the transcribed texts (definition of tasks, transcription conventions, modification of transcribed texts into the read texts, etc.), the specifications of the continuous speech corpus are described.