THE WEST POINT KOREAN SPEECH CORPUS

The Center For Technology Enhanced Language Learning
United States Military Academy
Department Of Foreign Languages
745 Brewerton Road
West Point, NY 10996
Email: john.morgan@usma.edu
Phone: 845-938-5329
Fax: 845-938-3585

March 23, 2004

The West Point Korean Speech Corpus is a data base of digital recordings of spoken Korean. Corpus design and data collection were carried out by staff and faculty of the Department of Foreign Languages (DFL) and Center for Technology Enhanced Language Learning (CTELL), located at the United States Military Academy (USMA), West Point, New York. The corpus was designed to develop speech recognition systems that would be used by the US government for speech-recognition enhanced language learning courseware .

The prompt scripts were created from 20,000 distinct sentences, along with a subset of prompts designed to elicit free response answers to questions for use in domain-specific speech to speech translation systems. Each speaker attempted to record 100 utterances. Three data collection scripts were designed by Ms. Jennifer Son, a native speaker of Korean under contract with the Department of Foreign Languages.

The scripts are in the file transcripts.txt. Each line of this file has two fields separated by a tab, the first denoting the name of the waveform file, and the second the transcription of the utterence.

The non-native data comes from informants from the 501st Military Intelligence Brigade who read from a subset of simplified sentences from the original text corpus.

In addition there is a small subcorpus of English spoken by native Korean speakers. These informants read sentences extracted from a local english speaking newspaper.

all the data was collected between September 20 and October 04 2002 at 2 different sites.

Native Korean speech was collected at the Korean Military Academy and the Korean Language Institute at Yonsei University in Seoul, Republic of Korea. Army linguists from the 501st Military Intelligence Brigade contributed to the non-native speech corpus.

Speech data was collected using Pentium 850 mHz laptop computers running Windows XP. Recordings were captured at a sampling rate of 16 bit @ 22050 Hz using a Shure M10 microphone. The recording script presented a visual display in Korean (Hangul) of the sentence to be recorded. The informant pressed a key and spoke the sentence. The informant's recording was played back for review, and the utterance was re-recorded, if necessary. A member of the data collection team was on hand during the recording session to verify recordings and provide technical assistance due to malfunctioning equipment. Several different versions of the recording script were used, dependent on whether the informant was a native or non-native speaker of Korean.