status report for

Preparation of the JEIDA Japanese Common Speech Data Corpus

Contract No. MDA972-92-J-1016
ISIP Project No. 03-95

for the period of May 15, 1995 to June 30, 1995

submitted to:

Linguistic Data Consortium

441 Williams Hall
University of Pennsylvania
Philadelphia, PA 19104-6305

submitted by:

Joseph Picone, Ph.D., Associate Professor

Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University
Box 9571
216 Simrall, Hardy Rd.
Mississippi State, Mississippi 39762
Tel: 601-325-3149
Fax: 601-325-3149
email: picone@isip.msstate.edu

EXECUTIVE SUMMARY

This phase of our work on this contract was largely dedicated to the development of infrastructure required for the project. Many of these activities are common to another ISIP project for LDC — "A Sun Sparcstation-Based Data Collection Platform," LDC Subagreement 5-24431-C. For details on infrastructure development, we refer to the status report for that project dated June 30, 1995. In this report, we will focus on a few aspects of this activity specific to the JEIDA corpus.

Two high-quality networked audio systems (the Townshend Computer Tools DAT-Link+) have been installed under Solaris 2.4 on a dedicated audio server, a Sparc SLC. Several minor software and configuration issues related to this class of machine, and to Solaris 2.4 were resolved. Having a dedicated machine for audio lessens the likelihood of real-time data transfer errors. In addition, three NCD X terminals have been deployed that provide us with slightly inferior, but adequate audio quality. Each system provides easy access to the audio system through a command line and C programming interface. These systems will be used to validate JEIDA data, and give us an increased capacity for validation and certification of the data.

The latest version of the Japanese language extensions to X have been installed, including mule-19.28. The good news is that we now have emacs and mule running from the same basic version of emacs, so the user interfaces and extensions are similar. In the past, mule was one version behind emacs, which made installation and maintenance a headache.

The JEIDA corpus has been delivered to ISIP in the form of a set of 74 DATs: 8 tapes containing isolated digits, 10 tapes containing 4-digit sequences, 18 tapes containing city names, 8 tapes containing control words — set A, 10 tapes containing control words — set B, 4 tapes containing control words — set C, 16 tapes containing monosyllables. The recording ambient environment is extremely clean by today's standards, which will make automated processing of the data straightforward.

We have initiated construction of software to automatically segment and digitize the data. Three publicly available utterance detection algorithms are being integrated into the system using a common software framework. These algorithms range from simple to complex. In the event that we have problems with simple utterance detection algorithms, we will be prepared to employ more sophisticated algorithms. We have also acquired NIST's SPHERE software and are beginning to integrate it into our tools to generate the distribution version of the corpus.

We will release a sample of validated data as part of our next status report, which will be provided on August 31, 1995.

1. HARDWARE DEPLOYMENT

In addition to the delays encountered in acquiring the Sun computers (documented in the status report dated June 30 for the LDC project "A Sun Sparcstation-Based Data Collection Platform," LDC Subagreement 5-24431-C), we encountered unanticipated delays in obtaining the Sony DAT machines. These were back-ordered for most dealers in North America, so alternate sources could not be easily found. We finally received our units on June 5, 1995. Needless to say, these units were critical to this project.

Two high-quality networked audio systems (the Townshend Computer Tools DAT-Link+) have been installed under Solaris 2.4 on a dedicated audio server, a Sparc SLC. Several minor software and configuration issues related to this class of machine, and to Solaris 2.4 were resolved. Having a dedicated machine for audio lessens the likelihood of real-time data transfer errors, and will give us a more robust environment for processing the tapes. In addition, three NCD X terminals have been deployed that provide us with slightly inferior, but adequate audio quality. Each system provides easy access to the audio system through a command line and C programming interface. These systems will be used to validate JEIDA data, and give us an increased capacity for validation and certification of the data.

2. GENERAL SOFTWARE INSTALLATION

The latest version of the Japanese language extensions to X have been installed under Solaris 2.4. These amount to three packages: mule, kterm, and Wnn. Fortunately, emacs and mule are now built from the same version of emacs (version 19), so the support and maintenance burden is minimized. Rumors are that GNU is working on folding the mule features into emacs and releasing a multilingual emacs for the next major upgrade of emacs (version 20). This convergence is long overdue and will improve the level of integration of Japanese language processing into mainstream Unix.

3. CORPUS PREPARATION SOFTWARE

We have initiated construction of software to automatically segment and digitize the corpus. This includes a program that, in real-time, reads data from the data in stereo, locates utterances on a user-selected channel using energy-based segmentation, separates the segmented data into two files corresponding to each channel, and writes the data to a unique pair of filenames (per utterance). This software will be used to automatically upload and segment the JIEDA tapes.

Three publicly available utterance detection algorithms are being integrated into the system using a common software framework. These algorithms range from simple to complex. We believe the basic algorithms we currently have in-house will be sufficient. In the event that they are not, we will quickly shift to a more complex utterance detection algorithm that gives us more flexibility in optimizing its performance.

We have also acquired NIST's SPHERE software and are beginning to train our programmers on the use of these headers, and to integrate it into our tools to generate the distribution version of the corpus. At some point, we need to reach convergence on what information should be stored in the SPHERE header. A proposal will be included in our next status report.

4. NEAR-TERM PLANS

Over the next two months of the contract, we expect to accomplish two major tasks:

- complete development of all data processing software;
- complete a pilot phase of data processing in which we upload, verify, and distribute a small portion of data.

Our goal will be to distribute the first set of validated files to LDC along with the next status report, and to elicit feedback. Once we cross this milestone, the remainder of the project will simply involve improving the throughput and accuracy of the validation process.