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 October 1, 1995 to January 31, 1996

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
413 Simrall, Hardy Rd.
Mississippi State, Mississippi 39762
Tel: 601-325-3149
Fax: 601-325-3149
email: picone@isip.msstate.edu

EXECUTIVE SUMMARY

The third phase of this project resulted in considerable progress in three areas: improvement of the overall efficiency of the file conversion process, distribution of the first set of sample SPHERE files, and validation of a large portion of the corpus. To date, we have completed preparation of 40% of the corpus (isolated digits, four digit strings, and monosyllables).

The validation tool was modified to support creation of output files directly. This allows validators to listen to the data, assign speaker, repetition, utterance, and transcription information in a single pass, and output copies of the data in the final filename format to an archival directory. Careful optimization of the GUI has allowed this to be done with no noticeable impact on performance. Validators still process approximately 500 utterances per hour (and are waiting on the tcl-based software most of the time). This saves significant time in the second pass verification step in which the final distribution files are certified.

A program to convert raw data files to sphere files was completed. This program loads all header information by combining session information (such as the speaker identity and recording location), validation information (e.g., the utterance transcription) into the SPHERE header. A sample of isolated digit and four digits utterances was made available on our ftp site (ftp://isip.msstate.edu/pub/ldc).

After the first pass of validation, a second pass is performed in which all data is verified by our senior corpus engineer. This involves listening to every utterance a second time, and verifying it matches the transcription. Several utilities were created to facilitate this. This second pass allows us to correct a number of problems, including missing utterances and missing right channel data (several speakers were recorded with a single microphone). All missing files and problematic files are documented in on-line documentation.

Progress in this phase of the project was slowed by the Christmas break and some staffing problems. We lost two validators due to academic and scheduling problems. These validators have been replaced. However, since this occurred close to the holiday break, the amount of validation performed in the month of December was unexpectedly small. We are now operating at full capacity to catch up (our two audio systems on which this work is performed are continuously busy from most of the day during weekdays). The anticipated delivery date for the corpus is mid-May. We should complete this first pass validation by mid-April, and the second pass verification by early May.

1. IMPROVEMENT OF THE VALIDATION TOOL

Our previous strategy for preparing the corpus consisted of three steps: validate the data "in place," rename the files based on the results of validation, and to verify the results using the same validation program. This led to problems when assigning the speaker number and repetition number in situations where speaker numbers were out of order, or data was missing. We were spending too much time on the verification pass correcting data and computing final output filenames. Hence, we decided the best solution was to make the validators assign this information at the time of validation.

A new version of the validation tool was implemented to accomplish this. The GUI is shown in Fig. 1(a). A magnified portion of the GUI depicting the new features of the tool is shown in Fig. 1(b). To better understand this, let's recall the filename convention being used in the corpus:

/isip/d00/jeida/val_data/iso_digits/m0001/m0001_i001_r4_c1.sphere

The information represented in this name is as follows:

/isip/d00/jeida/val_data/iso_digits/	output directory
m0001/	speaker directory
m	"m" denotes a male speaker
0001_	speaker number
i	"i" denotes isolated digits
001_	the utterance number
r4_	repetition number
c1	channel no. 1 (right channel)
.sphere	a SPHERE formatted file

Recall that the JEIDA corpus has audio introductions to each session on each tape (one of the things that makes segmentation more difficult). The validators enter this information based on what they hear on the tape. The validation tool is also told, via a parameter file, how many utterances are in each repetition. The validation tool then updates each one of these fields appropriately as a valid utterance is transcribed (utterances marked as "garbage" are automatically skipped), and moves on to the next repetition for the speaker automatically.

In order to prevent the accumulation of errors, we force the validators to manually change the speaker number when they begin a new speaker. This typically happens seven to eight times a tape.

2. SPHERE CONVERSION

In this phase of the project, we also completed the final version of the SPHERE conversion program (pending any feedback from LDC). The help page for the program is shown in Fig. 2. This is a simple program that stuffs the header of each file with a mixture of speaker, session, and utterance information. It takes as input three files: a session file containing information shared by all files in the session (the speaker's demographics, recording conditions, etc.), a list of the validation filenames (containing utterance-specific information such as the transcription), and a list of the raw speech data files.

The end result is an ASCII header shown in Fig. 3. We have included as much information as possible from the JEIDA corpus documents we have received from Professor Itahashi, yet

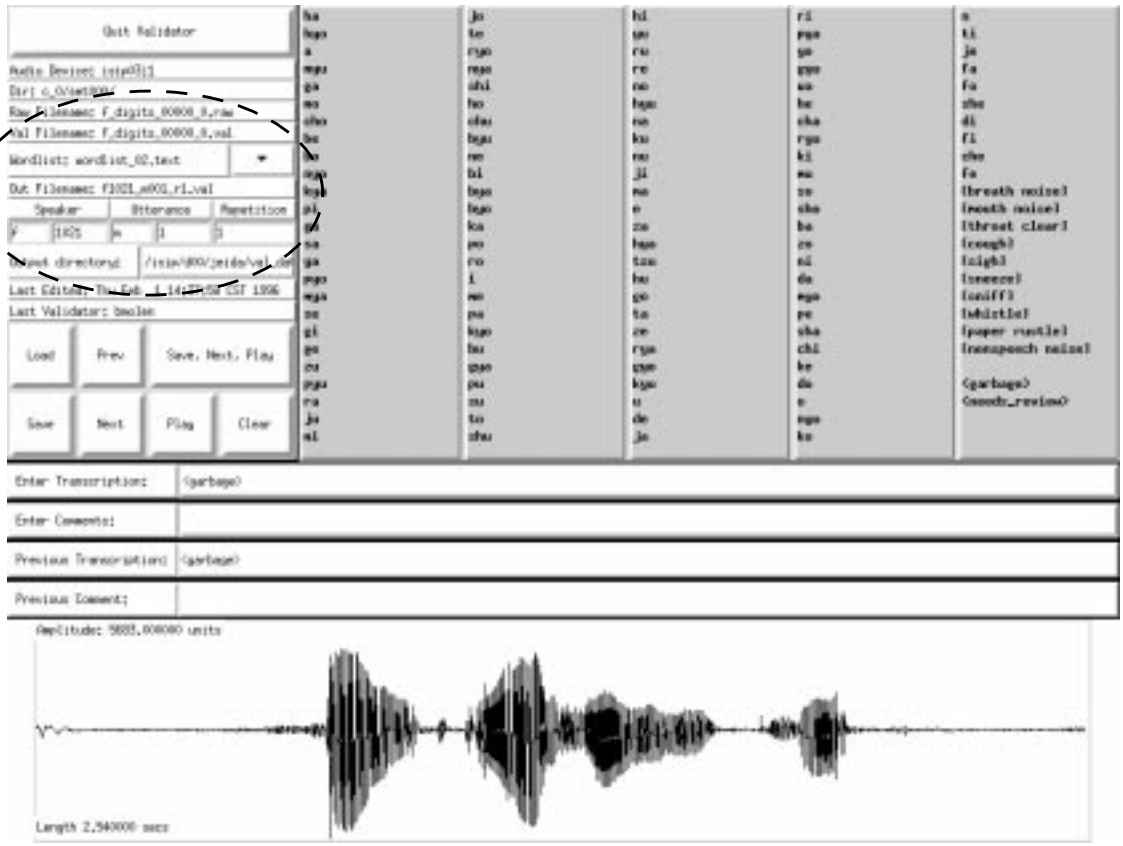

Figure 1(a). Version 2.0 of the validation tool developed for the JEIDA project is shown. Mouse functions are combined with buttons and menus to provide a high-speed validation capability.

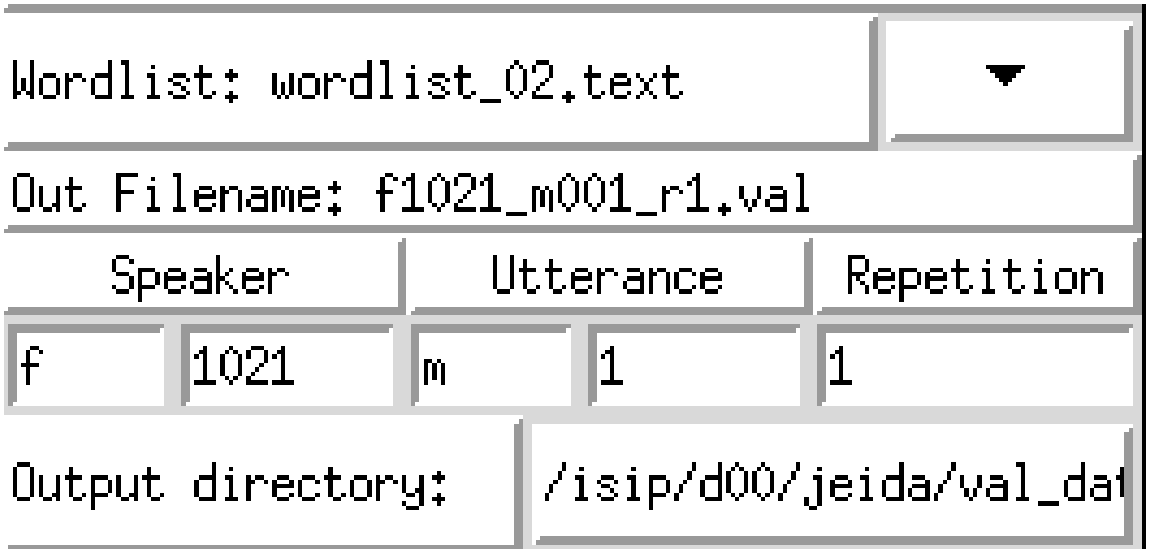

Figure 1(b). New features in version 2.0 include the ability to have the tool output final corpus files directly. Speaker number, utterance number, repetition number, and the output directory are automatically updated after each valid speech file, and only need to be set at the beginning of a tape, or after a change of speaker (the latter is a protective measure).

```
isip01_[2]: make_sphere -help
name: make_sphere
synopsis: make_sphere  <session_file>  <val_file_list>  <raw_file_list>
example: make_sphere  SESSION_INFO_FILE.text  val_list.text  raw_list.text

options:
  -help:                    display this help message

arguments: file names
  session_file:             file of session information in field = value format
  val_file_list:            file containing a list of JEIDA validation files
  raw_file_list:            file containing a list of RAW format audio files

  *note: every line of val_file_list should correspond to raw_file_list

man page: none
```

Figure 2. An overview of the make_sphere program.

```
NIST_1A
   1024
sample_min -i -4191
sample_max -i 3292
sample_count -i 19680
sample_n_bytes -i 2
sample_sig_bits -i 16
channel_count -i 1
speaker_number -s5 f1001
speaker_id -s7 CAN1201
recording_site -s3 CAN
database_id -s24 JEIDA Common Speech Data
database_version -s3 1.0
recording_environment -s15 soundproof room
speaker_session_number -s5 ID-01
speaker_age_category -s5 10-19
speaker_height -s6 181 cm
speaker_original_address -s8 Kanagawa
speaker_present_address -s1 -
ambient_noise_level -s5 19dBA
speaking_mode -s4 read
sample_rate -i 16000
orthographic_transcription -s4 zero
prompting_text -s4 zero
speaker_sex -s6 female
session_utterance_number -s4 i001
microphone -s12 Sanken MU-2C
sample_byte_format -s2 10
sample_coding -s3 pcm
end_head
```

Figure 3. The resulting header contained in the sphere file.

managed to stay within the 1024 character limit imposed by the NIST software. We believe this is a sufficiently detailed description of the corpus for most speech research applications.

A sample set of SPHERE files have been made available on ISIP's ftp server. This can be reached either by http or ftp:

ftp:	ftp://isip.msstate.edu/pub/ldc/jeida_sample.tar.gz
http:	http://www.isip.msstatee.edu (click on the ftp button)

This is a Unix tar file compressed using the GNU gzip program. It can be unpacked using the following sequence of commands:

```
gunzip jeida_sample.tar.gz
tar xvf jeida_sample.tar
```

This file contains the complete set of isolated digits and four digits strings for speaker f1001:

```
4_digits/f1001/f1001_d001_r1_c0.sphere
4_digits/f1001/f1001_d001_r1_c1.sphere
...
iso_digits/f1001/f1001_i001_r1_c0.sphere
iso_digits/f1001/f1001_i001_r1_c1.sphere
...
```

Unless we hear otherwise, we will assume the header formats for these files are acceptable.

3. VERIFICATION

A most important step in the creation of this corpus is a second pass in which our senior corpus engineer double checks the data. This steps involves processing the data through a listening utility:

```
isip01_[2]: verify_data f1001/*.val
f1001/f1001_i001_r1.val:
        transcription = "zero"
        playing f1001/f1001_i001_r1_c0.raw
        playing f1001/f1001_i001_r1_c1.raw

f1001/f1001_i001_r2.val:
        transcription = "zero"
        playing f1001/f1001_i001_r2_c0.raw
        playing f1001/f1001_i001_r2_c1.raw
```

The transcriptions are reviewed while listening to the audio data to verify the transcriptions. Speaker and session information is also reviewed. Missing files are checked and documented in cases where they cannot be recovered from the original tapes. In the case of the four digit strings, this was very time-consuming, as several tapes had to be virtually manually segmented. Numerous utterances were merged into a single file. We spent several weeks (almost as much time as validation of the data required) sorting out these files. The monosyllable data was much better behaved and proceeded much faster as a result.

Once this step is complete, the data is processed through a battery of programs to again check its consistency. Here is a summary of the procedure:

```
sd $JEIDA/val_data/monosyllables
ls m000[1-9]/*.val m0010/*.val > x_val.list
ls m000[1-9]/*_c0.raw m0010/*_c0.raw > x_c0.list
ls m000[1-9]/*.raw m0010/*.raw > x_raw.list

(1) count_files m000[1-9] m0010
(2) check_channels x_c0.list
(3) check_files x_raw.list 64000
```

Count_files performs a simple check making sure the appropriate number of files exist for each speaker and repetition. Check_channels makes sure there are two versions of each file corresponding to the two microphones used in the corpus. (For example, these two files should be the same size.) Finally, check_files does some rudimentary checks on the integrity of the data (files that are too long in duration, too low in amplitude, or identical to other files in the corpus are flagged as problematic).

This verification step requires about 25% of the time required to validate the data. Hence, we have three engineers allocated to validation, and our senior engineer dedicated to verification. This allows us to keep data rolling from DAT to disk to 8mm tape in an orderly fashion so that our disks do not fill, yet validators have enough data to keep moving.

4. SUMMARY OF PROGRESS

Our overall progress on the project thus far has been somewhat behind our original aggressive schedule. Table 1 contains of a summary of where we currently stand.

Type of Data	Num. Words	Num. Utterances	Validation Time (hours)	Status
Isolated Digits	10	6,000	12	Done
Four Digit Strings	35	21,000	42	Done
Monosyllables	110	66,000	132	Validated
City Names	100	60,000	120	
Control Words A	13	7,800	16	
Control Words B	24	14,400	29	
Control Words C	26	15,600	32	
Total	318	190,800	383	

Table 1. An overview of the progress to date on the JEIDA project.

As with most corpus projects, most of the time is spent in preparation. We spent a large amount of time debugging various aspects of the process. This was particularly true with the four-digit strings, in which we encountered numerous segmentation problems. In contrast, the monosyllables are being completed in four weeks — precisely the amount of time budgeted. We have observed only one segmentation error in the 66,000 utterances — and this was due to the speaker not pausing more than 0.3 secs between utterances.

According to our current production rate, the first pass of validation should be complete by mid-April. The verification phase that follows should be completed approximately one week after that (because verification by design only lags validation by one or two tapes). Hence, we should be able to deliver the corpus in May as originally planned.

Our current estimates are that the corpus will require somewhere between 10 Gbytes and 12 Gbytes of disk space for both channels of data. Monosyllables alone will require about 5 Gbytes of space. Fortunately, this will fit on only TWO 160m 8mm tapes — quite amazing.

5. NEAR-TERM PLANS

We have currently processed 40% of the corpus. Over the next three months, we expect to complete the following:

- digitization of the remainder of the data;
- validation of all remaining data: city names, control words A, B, and C (in this order);
- verification of the city names data;

We don’t expect any additional changes in staffing this semester, so our completion of the above items should proceed smoothly.