Japan Electronic Industry Development Association's

Common Speech Data Corpus

prepared for:

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

by:

Jonathan Hamaker, Richard J. Duncan, Joe Picone
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: hamaker@isip.msstate.edu

EXECUTIVE SUMMARY

The Japan Electronic Industry Development Association's Common Speech Data (JCSD) Corpus is an isolated phrase corpus consisting of 150 speakers (75 males/75 females) and almost 200,000 utterances. It represents an important milestone in Japanese speech recognition technology development. The JCSD Corpus was originally collected in 1986 in Japan in a nationwide project managed by Professor Shuichi Itahashi in coordination with the Japan Electronic Industry Association (JEIDA). Its importance to Japanese speech recognition technology development is, to some extent, comparable to Texas Instruments' famous 46-word speaker-dependent corpus. The JCSD Corpus was one of the first industry-standard and freely available corpora for the study of Japanese language speech recognition. Most of the competitive Japanese language speech recognition systems developed in Japan have been benchmarked on various subsets of this corpus. Hence, it is one of the most important standards of comparisons that exist for Japanese language systems.

As was typical of corpora developed in the mid-1980s, JCSD Corpus was collected in a quiet laboratory setting (fairly pristine recording conditions) using Sony PCM-F1 technology — a system that employed a 14 bit A/D converter that sampled data in stereo at 44.1 kHz and wrote the data to standard analog video cassette tapes in a proprietary digital format. This technology, though primitive by today's standards, provided a high quality audio recording capability —thereby making this corpus fairly uncontaminated by archaic analog impairments.

A summary of the size and content of the corpus is given below:

number of speakers	150 speakers
males	75
females	75
range of speaker age	10 yrs. to 70 yrs.
number of items per speaker	323 items
isolated digits	15
four digit sequences	35
city names	100
monosyllables	110
control words (set A)	13
control words (set B)	24
control words (set C)	26
number of repetitions per item	4 repetitions
total number of utterances	193,763 utterances (per channel)
sample frequency	16 kHz
sample type	16-bit linear
number of microphones	2 (dynamic and condenser)

Over 75 institutions in Japan have acquired this corpus and developed technology based on it. Some portions of it have become fairly famous and are quoted extensively in the literature — for example, the city name subset. Hence, the availability of this corpus will make it easier to port applications to Japanese, and to benchmark performance against state-of-the-art Japanese language technology.

TABLE OF CONTENTS

1. ABSTRACT

The Japan Electronic Industry Development Association's (JEIDA) Common Speech Data (JCSD) Corpus represents an important milestone in Japanese speech recognition technology development. The JCSD Corpus was one of the first widely available corpora to support research into Japanese language speech recognition. It is an isolated phrase corpus, consisting of three sets of control words, isolated digits, four digit sequences, city names, and monosyllables. Results on various subsets of the corpus, particularly the city names, have been extensively published. For historical reasons, this corpus constitutes a valuable addition to LDC's impressive collection of speech corpora. This document describes a computerized version of the corpus developed at the Institute for Signal and Information Processing (ISIP) at Mississippi State University for the Linguistic Data Consortium (LDC).

2. HISTORICAL BACKGROUND

The JCSD Corpus was originally collected in 1986 [1] in Japan in a nationwide project managed by Professor Shuichi Itahashi in coordination with JEIDA. Its importance to Japanese speech recognition technology development is, to some extent, comparable to Texas Instruments' famous 46-word speaker-dependent corpus [2]. The JCSD Corpus was one of the first industry-standard and freely available corpora for the study of Japanese language speech recognition. Most of the competitive Japanese language speech recognition systems developed in Japan have been benchmarked on various subsets of this corpus. Hence, it is one of the most important standards of comparisons that exist for Japanese language systems [3].

As was typical of corpora developed in the mid-1980s, JCSD Corpus was collected in a quiet laboratory setting (fairly pristine recording conditions) using Sony PCM-F1 technology — a system that employed a 14-bit A/D converter that sampled data in stereo at 44.1 kHz and wrote data to standard analog video cassette tapes in a proprietary digital format. This technology, though primitive by today's standards, provided a high quality audio recording capability — thereby making this corpus fairly uncontaminated by archaic analog impairments (such as the common phenomena of "print-through" that occurs in analog reel-to-reel tapes).

Recently, a DAT version of the corpus was made available to the general public for purchase [3]. Unfortunately, the DAT version, though it contained some segmentation information that can be used by special PC-based hardware and software, did not contain the type of segmentation useful for developing speech recognition technology. The DAT version was, in essence, simply a tape-to-tape copy of the original PCM-F1 material — the PCM-F1 tapes were converted to analog audio and re-digitized by a DAT machine. Also, the source format was recorded at a sample frequency, 44.1 kHz, much higher than what is needed by present-day research. Hence, preparation of the corpus in a format consistent with LDC's other products was essential.

2.1. Experimental Conditions

The corpus was originally recorded at 15 different sites in Japan. While attempts were made to keep the recording conditions fairly uniform across all sites, this is not what happened in practice. There is significant variation in the data from site to site, though overall the data can be classified as high signal to noise (SNR) data — comparable to a fairly quiet office environment. The corpus has been recorded on two channels. One channel contains data recorded with a standard dynamic microphone — a Sanken MU-2C microphone. The second channel contains data recorded simultaneously with a condenser microphone that presumably varied from site to site. If a single channel had to be selected for distribution, we would recommend the dynamic microphone channel. The dynamic microphone represents the best choice in that it is comparable to most of the microphone types used in research corpora and workstation speech applications (condenser microphones typically contain a battery and, hence, are not popular in workstation applications).

2.2. Speaker Demographics

There is a reasonable coverage of geographic region and age given the relatively small speaker population. Coverage by major geographic region (defined as the address at which the subject spent the most amount of time under the age of 12) is given below in Table 1:

Geographic Region	Combined (150)	Males (75)	Females (75)
Chubu	16	10	6
Chugoku	5	3	2
Hokkaido	1	1	0
Kanto	90	40	50
Kinki	16	10	6
Kyushu	9	5	4
Tohoku	10	6	4
Unknown	1	0	1

Table 1. An overview of the geographic distribution by region of the speaker population.

A more detailed analysis, showing the distribution within each geographic region, is given in Figure 1. Note that the Kanto region contains Tokyo. We expect a larger than normal concentration of speakers from that region. From Table 1, we see that 60% of the speakers in the corpus are from this region.

We also note that many of the regions are quite linguistically diverse (particularly Tokyo) and hence we don't expect a strong correlation between geographic region and dialect. One of the reasons the geographic coverage is reasonable is because the data collection sites were arranged at strategic locations around Japan. Since it is hard to assess dialect from a small sample of phrases, a dialect classification is not included in the corpus. However, informal listening indicates that the speaker population is far from homogenous, and fairly rich for such a small corpus.

The distribution of speaker age is shown below in Table 2. While the distribution at the edges is not well-sampled, the age range of 20 years to 60 years is well-represented in the corpus. A number of other interesting demographic and ambient factors are available in the corpus as well,

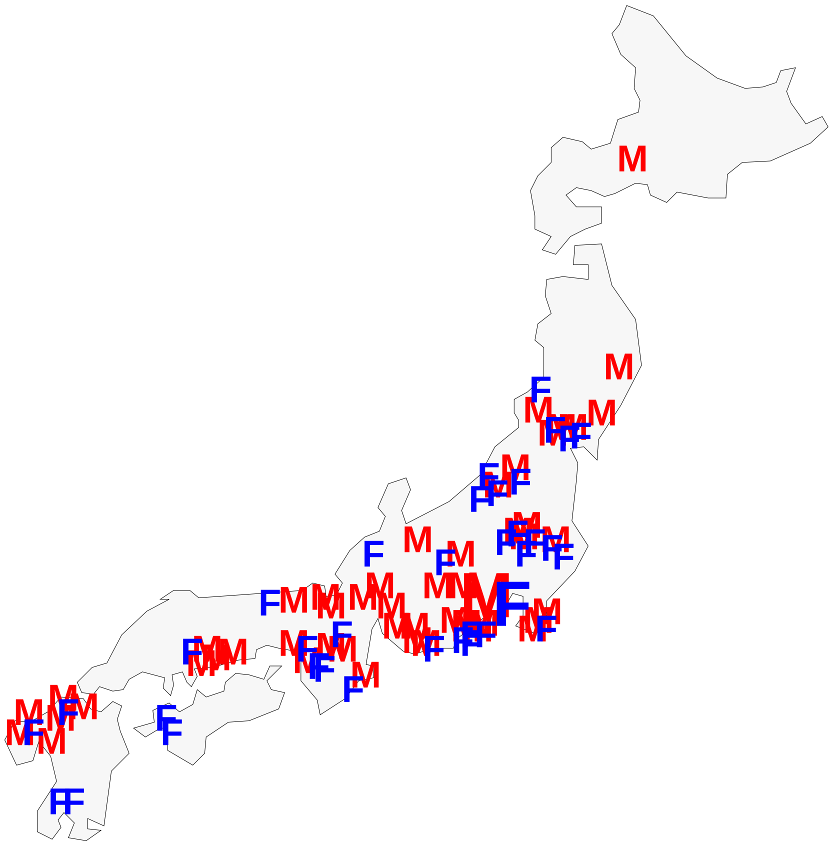

Figure 1. The geographic distribution of the speaker population is shown. The letter “M” (in red) denotes a male speaker, while the letter “F” (in blue) denotes a female speaker. These location is defined as the place in which a speaker spent the majority of time through age 12.

including height (loosely related to vocal tract length), a speaker’s present address, a brief description of the ambient recording environment, and the background noise level measured in dBA. All demographic information for the speaker population has been preserved in the corpus within each speech data file. This issue is discussed at length in Section 5.

Age	Combined (150)	Males (75)	Females (75)
10-19	1	0	1
20-29	50	25	25
30-39	40	20	20
40-49	32	15	17
50-59	22	11	11
60-69	5	4	1

Table 2. Distribution of speaker age in the JCSD Corpus.

2.3. Prompting Text

The JCSD Corpus is primarily an isolated word corpus. In addition, one segment of the corpus contains thirty-five four digit sequences that should be useful in supporting limited connected-digit recognition experimentation. A summary of the prompting material used in the corpus is shown below in Table 3.

Description	Number of items
Control Words:	
Banking Services	13
Word Processors	24
Home Electronic Equipment	26
Digits:	
Isolated Digits	15
Four Digit Sequences	35
City Names: a phonetically-rich subset of common Japanese city names	100
Monosyllables: all Japanese monosyllables plus several used to pronounce foreign words	110

Table 3. An overview of the prompting material used in the JCSD Corpus.

Each speaker was presented each item in the above lists four times — enabling the corpus to support some experimentation with speaker-dependent technology. Unfortunately, for all speakers and all repetitions within a given speaker, the items were presented in EXACTLY the same order. Hence, we expect researchers to encounter some undesirable session artifacts, such as artificial prosodic queues (for example, duration profiles).

A complete listing of all prompting text is given in Appendix A.

2.4. DAT Tapes

The JCSD Corpus described in this document was derived from a DAT copy provided by Professor Itahashi of the University of Tsukuba. This version was recently released [3] and is generally available directly from JEIDA. For archival purposes, we have included an inventory of the tapes, so that the Unix-formatted tapes can be easily cross-referenced back to the original data. A detailed tape inventory is given in Appendix B.

The DAT copy of the corpus consisted of 76 tapes totaling approximately 121 hours of material. The tapes were organized by speaker and session — a single tape contained multiple speakers of the same sex from the same segment of the corpus. This is also cross-referenced in Appendix B. Each session was delimited by an audio prompt describing the contents of the session. Two-hour tapes generally contained about 3500 valid utterances, while one-hour tapes contained about

1500 utterances. The overall quality of the DAT tapes was high — only occasionally did we encounter tracking problems or dropout problems. Tracking problems could be resolved by using another DAT machine, cleaning the drive, or simply restarting the tape. Dropouts and other such anomalies are documented in Section 5.2.

3. PREPARATION

Our process for preparing the corpus consisted of eight steps: digitization, segmentation, validation, certification, conversion to SPHERE files, verification of the SPHERE files, archival to tape, and certification of the tapes. The goal of this process was to listen to every piece of data twice (a large percentage of the data was reviewed three times), and to review generated files by automated computer processing three times. We believe this careful approach has reduced the defect rate to something very close to 0%. Each of these steps is described in detail below.

3.1. Digitization

Our goal was to deliver the corpus at a sample frequency of 16 kHz. This sample frequency is a *defacto* standard today in the speech research community for high quality corpora. Such a choice provides minimal loss of information and affords a reasonable reduction in disk space requirements. It is not trivial to convert from 44.1 kHz sample frequencies in that this frequency does not easily subdivide into other commonly used frequencies. (In fact, with DAT technology, 48 kHz is a much better choice. Rather than develop custom software for this step, and use homegrown filters, we decided to use the built-in capability of our DAT-based audio systems. This allows the experimental setup to be easily replicated and implicitly documents the conversion process.

The first processing step required was to upload speech data from its original source, DAT, to our Unix disks, and to convert the sample frequency from 44.1 kHz to 16 kHz. A Townshend Computer Tools (TCT) DAT-Link Digital Audio Interface (*http://www.tct.com*) was employed for this step. This unit interfaces a DAT machine to a Sun workstation via the SCSI bus, and performs sample frequency conversion in real-time on a dedicated AT&T DSP processor. This unit has become somewhat of an industry standard in the speech research community, and is uniformly regarded as producing high quality speech. It uses the standard linear-phase FIR filter approach to signal resampling, and operates on stereo data up to sample frequencies of 48 kHz. We used the DAT-Link's standard filters, which are linear phase FIR filters that deliver SNRs in excess of our minimum requirement of 80 dB.

The DAT-Link offers superb frequency response characteristics for a sample frequency conversion from 44.1 kHz to 16 kHz. The frequency response of the downsample filter has a 3 dB point above 14 kHz, is more than 90 dB down at the half-sample frequency of 8 kHz, and has less than 0.25 dB ripple in the passband. We prefer such anti-aliasing filters for corpora involving high quality speech. We have found the DAT-Link downsample filters to be more than adequate for a wide variety of speech processing applications including speech recognition. This unit is widely used in the ARPA and DoD communities, and more importantly, by LDC as well. In addition to doing the necessary signal processing, it provides highly reliable real-time audio, and has reliable error reporting in the event there are problems in the digitization process.

3.2. Segmentation

It is unfortunate that the JCSD Corpus was not segmented in a conventional fashion. While it contains segmentation information that can be used by some custom PC-based hardware developed in Japan, we cannot access any of this segmentation information from our standard Unix tools. For example, the DAT tapes were delimited by program numbers — something the DAT machines can recognized. Individual utterances, however, were not delimited. (Also, tape dropouts and discontinuities caused program counters to be unreliable.)

Hence, each DAT had to be segmented during the digitization process. Our approach was to perform this step off-line so that problems could be easily resolved. Our process involved digitizing an entire DAT tape, postprocessing the data through a segmentation algorithm, reviewing the results, and then automatically generating files from the segmentation information. An energy-based algorithm [5] was used to perform the segmentation. In cases where the automatic algorithm failed, problems were corrected manually using some interactive tools developed for the project.

By and large, the segmentation of the data was reasonably successful once we tuned the algorithm to the specific data set. We had the most problems with the four digit sequences. These were not spoken with an acceptable amount of silence between phrases (due to the way the data was collected — this could have been prevented with computer prompting). Hence, inter-word gaps were often longer than inter-utterance gaps. A significant percentage of the utterances had to be corrected by hand. Other tapes exhibited similar problems with inter-utterance gaps. We typically had to set the signal detector to accept a gap as small as 0.3 seconds to reliably segment the data, and this often caused problems with polysyllabic phrases.

Once the endpoints were determined, we padded each utterance with approximately 0.25 secs of leading and trailing silence. This will allow technology developers a chance to get a more realistic measure of their algorithms robustness to channel impairments (a major problem in fielding speech technology). There are many examples of common mouth noises and other such artifacts in the corpus, and we have expended extra effort to mark these. It is possible, for example, to build statistically-trained models based on these markings (a popular approach in Hidden Markov Model-based technology). Since the corpus does contain some amount of nonstationary background noise, there is some value to retaining this data. Also, mouth artifacts are abundant, so algorithms developed from this data will have to deal with them in some intelligent manner (we see this as a positive aspect of the corpus).

3.3. Validation

Since the utterances were recorded in DAT in a fixed order, the task of validation, defined as the step in which we add an orthographic transcription to the data, was fairly straightforward. The data was validated using a simple tool that supports audio playback of a file, text entry, mouse-based editing of the orthographic transcription, and form-based editing of the auxiliary information identifying the utterance. The TCT DAT-Link was also used for validation — we consider the use of high-quality audio essential to performing accurate validation efficiently (especially for noting anomalous behavior). Orthographic transcriptions were performed using an

ASCII encoding of a hiragana system developed and described in [3]. Since this is an isolated phrase corpus, transcription in kanji was deemed to be an unnecessary additional burden — the kanji equivalents of the hiragana can be easily provided via a table lookup. With this streamlined approach, our validators were able to reach a peak speed of approximately 600 utterances per hour and maintain excellent accuracy and consistency.

There were two main issues involving validation. First, there was the issue of transcription conventions for marking of non-speech sounds. We based our work on other corpora [4] providing similar information, and tried to minimize the number of unique makers were used. These are shown in Table 4. Our general criteria was that if a sound was clearly audible, it should be marked. This decision was supplemented by a waveform display of the utterance. Our validators were trained to use both visual and audio cues in determining whether a non-speech sound should be marked. A significant percentage of utterances contain some amount of non-speech sounds. For example, for the isolated digits component of the corpus, 12% of the utterances contain at least one non-speech marker.

Non-speech Orthographic Items	
{breath noise}	{sigh}
{mouth noise}	{sneeze}
{throat clear}	{sniff}
{cough}	{whistle}
{paper rustle}	{non-speech noise}

Table 4. A listing of the non-speech orthographic items used in the JCSD Corpus. Note that the last item, *{non-speech}*, was only used in the rare case that none of the other existing items applied. This list is a subset of that used in other ARPA/LDC corpora [4]

Second, a convention for anomalous, or alternate, pronunciations had to be established. For the digit data, we were particularly interested in flagging uncommon pronunciations, because these have been traditionally used to optimize recognition performance. For digits, the number of variants are small and easily predicted. Hence, we decided to incorporate the set of words shown in Table 5. This system consists of standard orthography, plus the use of brackets to denote variant pronunciations in which a phone was missing or significantly reduced (for example, it is common in Japanese to drop the "i" in "hachi"). Though transcription at this level is generally expensive, transcribing data in this manner for the digit portion of the corpus did not significantly add to the cost of the project.

Thus, a typical transcription for an utterance might look something like this:

{mouth noise} ichi [shich] san [hach] {breath noise}

which indicates that the phrase "ichi shichi san hachi" was preceded and followed by the indicated non-speech noise, and contained alternate pronunciations for "shichi" and "hachi." In order that we not constrain potential users of the corpus into this system, the original prompting text, in this case "ichi shichi san hachi," was also stored in each speech file. Further, to ignore this additional

Orthographic Items Denoting Alternate Pronunciations For Digits	
[dei]	[ich]
[dok]	[doku]
[rok]	{sniff}
[shich]	[hach]

Table 5. A listing of the orthographic items corresponding to alternate pronunciations for data containing digit sequences (isolated and four digit sequences). For example, "[hach]" denotes a pronunciation of "hachi" in which the final vowel was omitted.

information, one can simply replace bracketed items with their non-bracketed equivalents, or use the prompting text as the orthographic transcription. Since the delimiters for non-speech sounds and alternate pronunciations are mutually exclusive, it is easy to mix and match this information as needed. Since SPHERE headers (described in the next section) are stored in an ASCII format, it is easy to use standard Unix tools to filter this information.

In Table 6, we present some statistics for the transcriptions of the isolated digits to provide a feel for the extent to which such nonstandard items are present in the corpus. This table provides a glimpse into the nature of the non-speech items as well. Breath noises and mouth artifacts dominate the non-speech markings. We can see that a bulk of the corpus is fairly clean, indicative of the rather controlled conditions under with which it was collected.

3.4. Certification

Once validation was completed, every file was passed through a second pass of review denoted certification. The output of the validation program is a set of files organized by speaker, session, item, etc. A second pass of review was applied to the data in which a different person (the project manager, who did not serve as a validator) listened to a file while viewing the transcription. The purpose of this step was to verify that the assignment of speaker numbers and such were correct, and to identify missing files, anomalous files that needed to be rechecked against the DATs, and resolve any other problems identified in the validation stage. The task of certification involved the following steps: software checks of data (file sizes, file integrity, etc.), human checks of data (listen to each file), and corrections of problems (most often, the original DAT recording was reviewed). This process was repeated until all outstanding issues were resolved.

There were several useful computer-automated checks that were performed using a handful of simple utilities. These utilities counted the number of files, checked each file for consistency between channels, and compared each file against each other to make sure no overlaps had occurred (a problem that arose because speakers were out of order on the original DATs, and validators sometimes transcribed the speaker and repetition numbers incorrectly). These programs (described in detail in Section 4) flag any files which deviated from the expected results. The expected results consisted of a correct number of files per speaker per repetition, identical file sizes but different sampled data for each channel, and uniqueness across all speakers and all repetitions.

Description	Frequency (17,390 items)	
"clean data": nominal pronunciation/no non-speech markers	15,364	(88.4%)
with a non-speech marker and/or an alternate pronunciation	2,026	(11.6%)
with an alternate pronunciation	1,161	(6.7%)
distribution:		
only an alternate pronunciation	1,121	(96.6%)
both an alternate pronunciation and a non-speech marker	40	(3.4%)
with a non-speech marker:	905	(5.2%)
distribution:		
only a non-speech marker	865	(95.6%)
both an alternate pronunciation and a non-speech marker	40	(4.4%)
non-speech markers	979	(5.2%)
distribution:		
{mouth noise}	543	(55.5%)
{breath noise}	344	(35.2%)
{paper rustle}	48	(4.9%)
{non-speech noise}	20	(2.0%)
{throat clear}	7	(0.7%)
{background noise}	7	(0.7%)
{cough}	6	(0.6%)
{sniff}	2	(0.2%)
{mouth_noise}	2	(0.2%)

Table 6. Distributions of nonstandard orthographic items for the isolated digit subset of the JCSD Corpus. The vast majority of the data is fairly clean, with approximately 10% showing some type of nonstandard behavior. The statistics for other segments of the corpus are comparable.

Once the computer-based checks had been performed to verify the data, and we had generated the necessary information to fix the majority of the problems, the project manager carried out such tasks as resegmentation of the data (often done manually), recording missing data, revalidation, etc. This step in the certification process often required manual listening to the validated data. This step was performed as an independent check on the data, using the *verify_data* utility. With this tool, we were able to listen to an utterance and simultaneously view its transcription. This was probably the single-most important step in the process because it provided a check on the validators' work. Common problems found in this phase were incorrectly transcribed utterances and incorrectly segmented data.

At the end of the certification stage, every utterance in the corpus had been reviewed at least twice, and the remaining problems with the corpus could generally be traced back to the source DAT data. There were three types of problems found on the original DAT tapes: a single utterance was truncated (typically the end of the utterance, occasionally the beginning of an utterance at the start of a new repetition of the data), a single utterance or group of utterances were missing on the tape, or an utterance was mispronounced on the tape. A record of each of these anomalies was maintained and is found in Section 5.2.

3.5. Conversion to SPHERE

The output of each of the previous steps was a non-SPHERE formatted file. using a simple utility we developed, the last file creation step consisted of creating SPHERE files from our raw files. This step also deposited files into the final corpus directory structure using the official filenaming scheme (described in Section 5.1). A typical SPHERE header is shown in Table 7. The range of values that each of these fields can take is fully described in Section 5.1. Note that compression was not used in storing the sampled data files. Also, anticipating that users would prefer to use only one channel of the data at a time, as has been our experience with previous corpora involving multiple microphones (for example, TIMIT), we decided to explicitly separate the data in the corpus by channel. Hence, each sampled data file contains only one channel of data.

All attempts have made to conform our construction of the SPHERE header to conventions used in other LDC corpora. Since this is an isolated phrase corpus, the header is reasonably self-contained. Since the SPHERE header simply consists of the first 1024 bytes of the file, it is also a relatively easy matter to remove the header and identify the remaining speech data (another reason we prefer to avoid the use of exotic compression schemes).

3.6. Verification

Once the files were converted into SPHERE format, two final checks were run on the data. The final SPHERE files were counted to make sure that the correct number of files were created for each speaker, channel, and repetition. Next, a utility that collects statistics on the items appearing in the orthographic transcription and prompting text fields of the SPHERE file was run. Anomalous utterances were checked against the errata to make sure everything was properly accounted for. This information is presented in detail in Section 5.2.

3.7. Archival to Tape and Tape Certification

Standard Unix tape tools were used to create the final corpus archive. A set of four 120 meter 8mm tapes totaling almost 20 Gbytes of data were delivered to LDC. ISIP also maintains two independent copies of these tapes. These tapes were created using uncompressed speech files (shorten, which is built into the SPHERE software was not used). We believe this is best for ease of portability. We did, however, use the compression capability of our Exabyte 10e tape stacker so that we could minimize the number of tapes. Hence, a typical tape command consisted of:

```
tar cvf /dev/rmt/1cn control_words_a
```

This command creates files on the tape with a root node of control_words_a, making it easy to untar and relocate the corpus. Given the size of the corpus, on most computers it will undoubtedly span multiple file systems. Hence, the tapes were mastered in a way that makes it easy to restore the data to different disks. The device “1cn” corresponds to using tape device no. 1 with the highest compression option and the no rewind option (tapes contain multiple tar files). These tapes were mastered on a Sun Sparcstation 5 running Solaris 2.4. We used GNU tar, version 1.11.2, to make tapes. The set of tapes delivered to LDC have been listed in their entirety, to make sure the contents of each tape were correct and readable. This data resides in INFO/tape_listings found at the root node of the corpus on the first tape (tape no. 1).

SPHERE Info Type Name/Value Pair	Comment
NIST_1A	SPHERE supplied
1024	SPHERE supplied
sample_min -i -9336	minimum sample value
sample_max -i 11542	maximum sample value
sample_count -i 15039	number of speech samples in the file
sample_n_bytes -i 2	speech samples are two-byte integers
sample_sig_bits -i 16	16-bit speech samples
channel_count -i 1	one channel of data
speaker_number -s5 f1027	speaker number
speaker_id -s7 IBM1301	original JEIDA speaker ID
recording_site -s3 IBM	recording site
database_id -s24 JEIDA Common Speech Data	corpus name
database_version -s3 1.0	corpus version
recording_environment -s12 Meeting room	recording environment
speaker_session_number -s4 B-03	session number
speaker_age_category -s5 30-39	age category at the time of recording
speaker_height -s5 157cm	height of the speaker
speaker_original_address -s13 Sado, Niigata	city and prefecture during childhood
speaker_present_address -s12 Komae, Tokyo	present city and prefecture
ambient_noise_level -s5 47dBA	noise level as measured in dBA
speaking_mode -s4 read	speaking mode
sample_rate -i 16000	sample frequency in Hz
orthographic_transcription -s4 ido-	ortho trans. (in this case, a control word)
prompting_text -s4 ido-	original prompting text
speaker_sex -s6 female	speaker sex
session_utterance_number -s4 b007	utterance type/number
microphone -s12 Sanken MU-2C	microphone used during recording
sample_byte_format -s2 10	samples are linearly encoded
sample_coding -s3 pcm	samples are linearly encoded
end_head	SPHERE supplied

Table 7. An example of the SPHERE header used in the JCSD Corpus.

4. SOFTWARE TOOLS

In addition to the corpus, we have included all software and documentation in the distribution. In this section, we briefly describe the major software tools we have developed for this project. Several of these, particularly the validation tool, are easily modified to support the development of new corpora. This software is located in INFO/tools at the root node of the corpus (and on the first tape in the four tape set). Underneath the tools directory, there is a directory *src* that contains source code, and *bin* which contains Solaris 2.4-compiled binaries.

4.1. Digitization

Digitization of the data on DAT tape was performed using the narecord program developed by Townshend Computer Tools (we used netaudio v2.27). This program reads stereo data in real-time from a DAT at a specified sample frequency and writes it to a file in a specified file format. The digitized data was recorded at a sample frequency of 16000 Hz, and stored in the raw data format. Tape were digitized in one or two-hour durations using the following command:

```
narecord -u isip03:0 -s 16000 -t 115200000
```

This "-u" option denotes the machine and audio device number, the "-s" option denotes the output sample frequency, and the "-t" option denotes the number of samples to record (in this case, one hour of data). The data is recorded in stereo with this command, so that both channels of the data, representing the two microphones used in the corpus, are recorded simultaneously.

4.2. Segmentation

The segmentation procedure consists of two programs: signal_detector and excise_signal. The *signal_detector* program is a flexible data-driven program that reads its parameters from a parameter file. The algorithm used is a standard energy-based adaptive-thresholding algorithm [5] used extensively in the speech research community. An example of one of the actual files used to segment the corpus is shown in Figure 2. The most important parameter in the context of this project was the *minimum_utterance_separation*, which controls the ability of the system to spot phrase boundaries versus internal pauses for polysyllabic phrases (typically inter-word pauses). The signal detector program takes a raw data file as input, and outputs start and stop times of utterances to an ASCII log file.

The second step is to run the speech data file and the log file through the *excise_signal* program. This program splits the data into a one-channel per file format, and creates the sampled data files in a generic directory structure — files are sequentially numbered and organized into subsets of 100 files per directory. At this point, the data is ready for validation.

4.3. Validation

The validation phase of this project was done using a GUI written in Tcl. This program was written to automate (as much as possible) the tasks of validation: audition, transcription, and filename creation. The GUI is mouse-driven and requires essentially no typing. This greatly decreases the time needed for validation and allows the human validator to concentrate on the task of properly transcribing the data and labeling problem utterances. A snapshot of the validation tool is shown in Figure 3(a). The validation tool was designed to make efficient use of screen real estate and computer resources, so that validators could work from small 15" black and white monitors served from a modest Sun Sparcstation (two validators can work from a Sparcstation 5 with 32M of memory and a 50 MHz processor).

The upper left of the screen contains parameter options which are used to input the initial conditions of validation, including speaker number and sex, utterance number and repetition, and output directory. Once the initial conditions are set, the validation tool automatically increments

```
# file: endpointer_00.params
#
# this file contains the parameters used to endpoint speech
# utterances recorded under near studio-quality conditions.
# it was originally developed to digitize the JEIDA CSD Corpus.
#
# this file has been "optimized" for short isolated word utterances,
# such as the isolated digits. the JEIDA data is packed quite closely,
# as little at 0.3 secs separates some utterances. so some parameters,
# such as minimum_utterance_separation, have been set very small.
#

# data format parameters
#
number_of_channels              = 2 channels
sample_size                     = 2 bytes
channel_to_be_processed         = 0 channel

# signal processing parameters
#
sample_frequency                = 16000.000 Hz
frame_duration                  =     0.020 sec
window_duration                 =     0.030 sec
preemphasis                     =     0.950 units

# signal level-related energy parameters
#
nominal_signal_level            = -35.00 dB
signal_adaptation_delta         =  15.00 dB
signal_adaptation_constant      =   0.50 units

# noise level-related energy parameters
#
nominal_noise_level             = -60.00 dB
noise_adaptation_delta          =  15.00 dB
noise_adaptation_constant       =   0.75 units
noise_floor                     = -70.00 dB

# utterance-related parameters
#
utterance_delta                 =  6.000 dB
minimum_utterance_duration      =  0.050 sec
minimum_utterance_separation    =  0.300 sec
maximum_utterance_duration      =  99.999 sec

# debug information
#
debug_level                     = 0 level
```

Figure 2. An example of the signal_detector parameter file that provides for easy control of key parameters, including the algorithm.

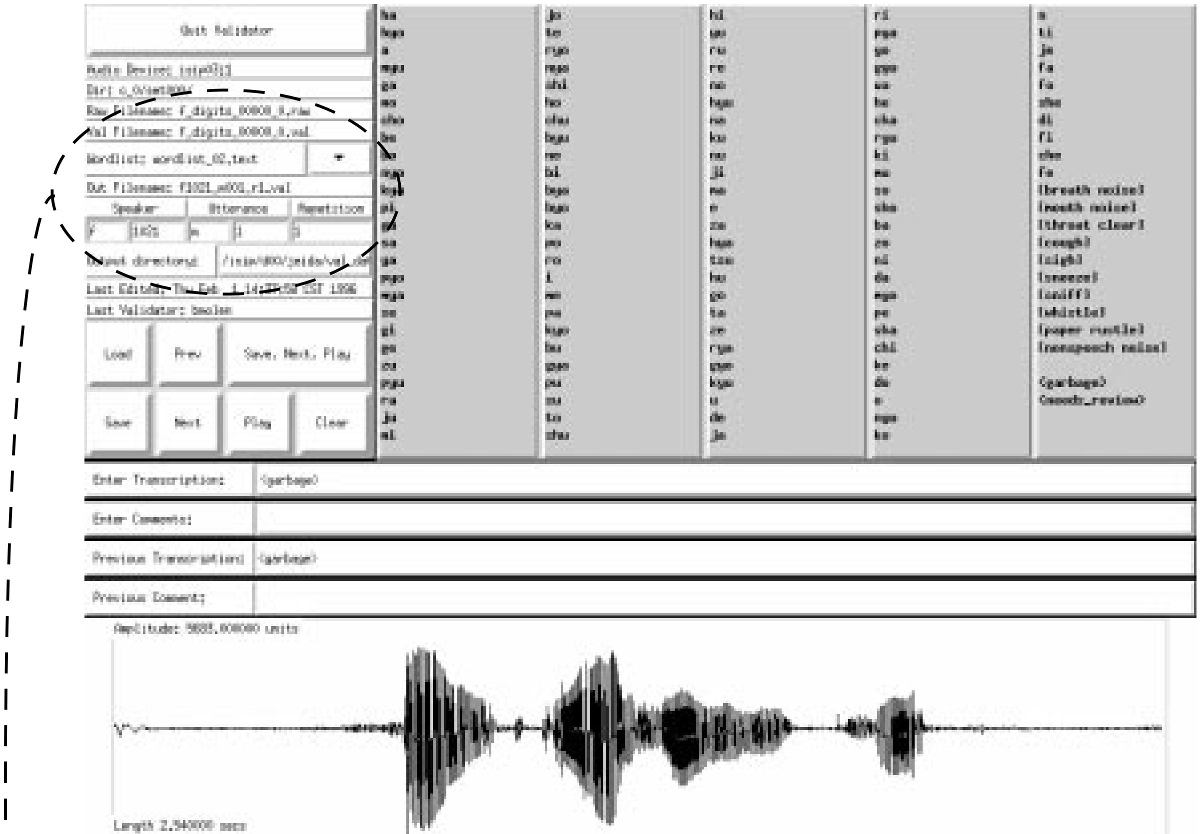

Figure 3(a). Version 2.0 of the validation tool developed for the JEIDA project is shown. Mouse functions are combined with buttons and menus to provide a high-speed validation capability.

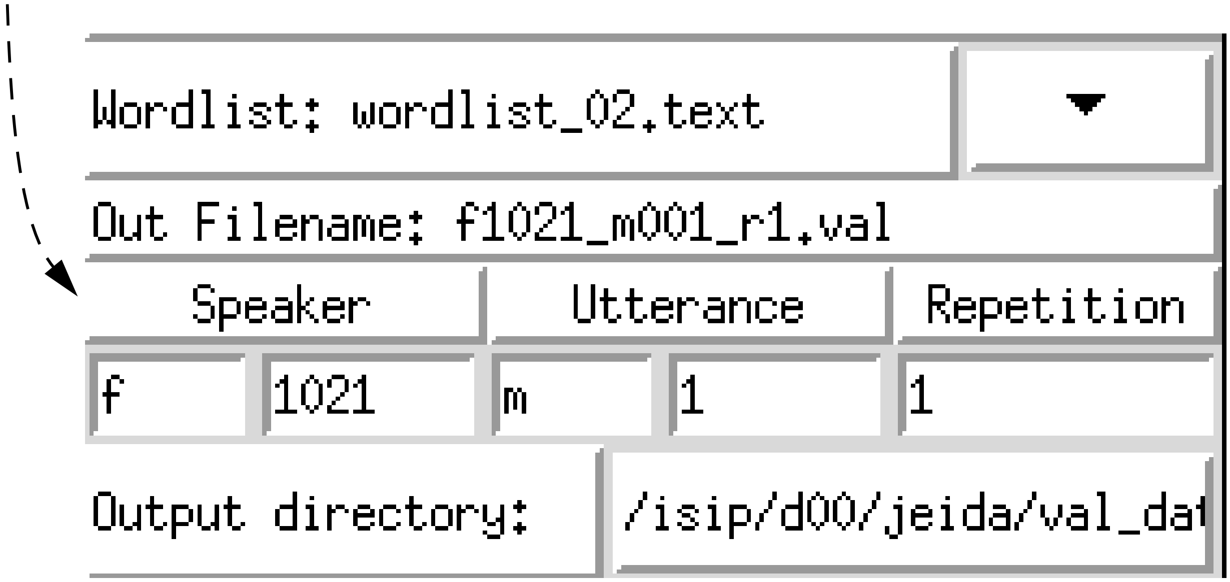

Figure 3(b). New features in version 2.0 include the ability to have the tool output final corpus files directly. Speaker number, utterance number, repetition number, and the output directory are automatically updated after each valid speech file, and only need to be set at the beginning of a tape, or after a change of speaker (the latter is a protective measure).

all utterances, repetitions, and speaker numbers in the sequence for an entire speaker's data (including multiple repetitions). This allows the validator, theoretically, to set the initial conditions and then validate an entire set (four digits, isolated digits, etc.) of the corpus without ever using the keyboard — an important consideration for maximizing throughput.

On the upper right of Figure 3(a) is the word list for a particular set of data. This word list is loaded into the program via an external file specified in the validation tool parameter file. The word list contains all of the possible "legal" utterances that would be encountered. All transcription and error labeling for a particular utterance are selected from this list via the mouse. This is a point-and-click interface, such that the validator can click on the desired transcription with the left mouse button and then click the middle mouse button to advance to the next utterance. Clicking the middle mouse button saves the output files to the proper directory, advances to the next utterance and plays it. The program was heavily tailored to the specific task of JEIDA validation. Though tcl is a good language to write such applications, the program's throughput is somewhat limited by the speed of tcl. For future projects, we plan to upgrade this program using a combination of perl and tk.

At the bottom of the screen is a waveform display corresponding to the utterance. This is accomplished using a mixture of C programs and tcl plotting. The central component is the utility *plot_signal*. This utility gives a plot of signal magnitude vs. time for the utterance duration. A comparable utility, *plot_endpoints*, is included that plots the signal and its endpoints as determined by the segmenter. Placing the *plot_signal* screen in the validator allows the validator to easily detect "illegal" data such as background noises, tape anomalies, and unusually low signal levels. The validator can then flag these instances for later review using the "needs review" transcription item.

4.4. Certification

Certification involved four distinct processes: counting files, checking raw files for consistency between channels, checking raw files for size, consistency and overwrites, and most importantly, listening to all data a second time. Several simple scripts were developed to facilitate this process.

Counting files was done using the *count_files* program. This shell script uses standard Unix commands such as *ls*, *wc*, and *awk* to determine the number and types of validation files that are produced during the validation phase. It outputs the results by speaker and repetition allowing the user to easily determine the location of the missing utterance. This program is used primarily to find missing files in the validated data or missing utterances on the DAT.

The next program used to certify validated data was the *check_channels* program. The *check_channels* program compares each file corresponding to a different channel of the same utterance to insure that the files are the same size, yet hold different data. This C++ program takes as input a list of the channel 0 files needing review. The size of the channel files are compared by counting the number of samples; files with differing sizes are flagged. The program then checks the first 256 samples of each channel to make sure they are not identical. This step serves two purposes: to find incorrectly validated utterances and to find utterances with large segments of zero data. The automatic segmentation process sometimes included adjacent zero-value data on the DAT due to the way the tapes were mastered as concatenations of other tapes.

The *check_files* utility is a C++ program used to do 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 for further review and correction). By tagging the files that are too long, this program finds many instances where two utterances have been segmented into the same file or where there are long runs of zero amplitude signals in the raw files. This program also finds the instances where the validator has mistakenly saved a single utterance to multiple files due to an incorrect speaker or session number setting.

Once the above utilities perform their tasks, the *verify_data* utility is used to perform a second pass at listening to the data. This shell script plays an utterance and lists the proposed transcription of that utterance. The utterances to be verified are given as input from the command line. This step in the certification process is essential as a redundancy check on the data. This program allows the user to fix instances where the validator has incorrectly labeled an utterance, where an utterance is missing or defective, and where background noises are either unlabeled or incorrectly labeled.

4.5. Conversion to SPHERE Files

This phase of the project required implementation of a *make_sphere* program. The help page for the program is shown in Figure 4. 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 - and a destination directory. The end result is an ASCII header such as the one shown in Table 7.

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

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
 destination                    a directory in which to place the sphere files

 *note: every line of val_file_list should correspond to raw_file_list
man page: none
```

Figure 4. An overview of the make_sphere program that converts raw files to SPHERE-formatted files.

4.6. Verification

Verification of the sphere data is the last “line of defense” in spotting validation and data errors. Verification involves recounting the SPHERE files and checking the transcriptions in the SPHERE files for consistency. The program used to check for transcription consistency is *check_sphere*. This shell script takes as input the directory where the SPHERE files are held and the parameter keyword to search for in the SPHERE files. It finds that keyword and counts every unique instance of its value. The program then tallies the number of times that value is found and outputs this distribution. The user should expect that each value would appear in the corpus a given number of times. Comparing the expected output and actual output gives the user the ability to find errors which may have slipped through the validation/certification process.

This collection of relatively simple programs has proven to be extremely valuable in maintaining a high quality corpus. Each step was carefully designed to provide useful information about the corpus, and serve as a redundant check. The *check_sphere* program, for example, though one of the last programs run, caught several errors in each segment of the corpus. ISIP will make every effort to maintain, support, and extend this software to support LDC’s long-term mission of providing high quality corpora cost-effectively.

5. CORPUS DESCRIPTION[1]

The Japanese Common Speech Data (JCSD) Corpus consists of a set of four 8mm tapes containing 20 Gbytes of data across 375,000 files. In this section, we summarize relevant properties of the corpus. Refer to Appendix D for information on how to read the tapes.

5.1. Overview

The corpus has been prepared on Unix tar-formatted 8mm tapes. The contents of each tape are summarized below in Table 8. The corpus has been subdivided so that it can be easily restored to reasonably small file partitions. Channel 0 and 1 have been separated in anticipation that most users will prefer to deal with only one microphone at a time.

An overview of the file and directory structure is given in Figure 5. The logical organization of the corpus begins with a subdivision by content, followed by a subdivision by channel, followed by a subdivision by sex, followed by a subdivision by speaker. This makes it easy to restore particular subsets of the corpus. The organization has been kept as symmetric as possible to facilitate regular expression/wildcard searching of the corpus. Speech data are contained in seven directories (lowercase), while documentation and readme files are presented in uppercase at the highest level of the directory hierarchy.

Our file naming convention for the corpus is described in Table 9. Other useful statistics about the size of the corpus are given in Table 10. Since this corpus is quite large, it is useful to have a detailed breakdown of the sizes of various components of the corpus. This is given below in

1. This section is meant to serve as a self-contained description of the corpus. Some of the information is redundant with other sections of this text.

Tape No.	No. Tar Files	Size (Gbytes)	Contents
1	5	0.8 1.6 1.5 0.8 0.1 Total: 4.8	Ctrl Words A Ctrl Words B Ctrl Words C Isolated Digits Documentation/Source
2	1	Total: 3.1	Four Digits
3	2	3.3 3.1 Total: 6.4	City Names: Channel 0 City Names: Channel 1
4	2	2.9 2.7 Total: 5.6	Monosyllables: Channel 0 Monosyllables: Channel 1

Table 8. A summary of the 4 tape set comprising the JCSD Corpus.

Filename: **m0001_i001_r4_c1.sphere**	
Substring	Description
m	speaker sex (m/f)
0001	speaker number: f1001 — f1075: females m0001 — m0077: males
i	utterance type: a: control words a b: control words b c: control words c i: isolated digits d: four digit sequences n: city_names m: monosyllables
001	utterance id (see Appendix A)
r4	repetition no. 4 (out of four repetitions)
c1	channel no. 1
sphere	a SPHERE formatted file

Table 9. The JCSD file naming convention.

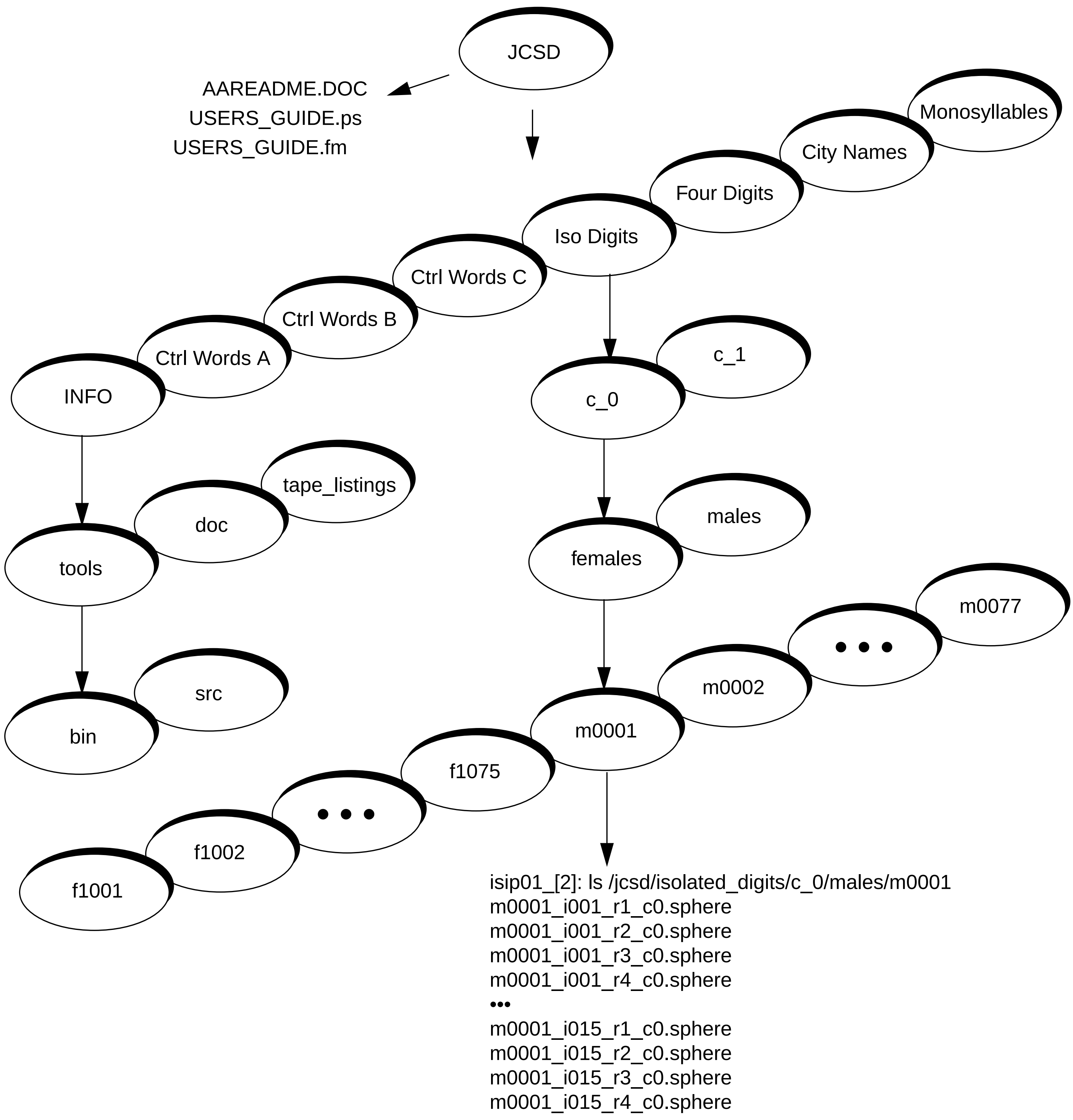

Figure 5. An overview of the filesystem used to organize the JCSD Corpus.

number of speakers	150 speakers
males	75
females	75
range of speaker age	10 yrs. to 70 yrs.
number of items per speaker	323 items
isolated digits	15
four digit sequences	35
city names	100
monosyllables	110
control words (set A)	13
control words (set B)	24
control words (set C)	26
number of repetitions per item	4 repetitions
total number of utterances	193,763 utterances (per channel)
sample frequency	16 kHz
sample type	16-bit linear
number of microphones	2 (dynamic and condenser)

Table 10. Key dimensions of the JCSD corpus.

Table 11. From this table, it is clear that there are some missing data files. This is extensively documented in Section 5.2. It is also clear that careful structuring of the corpus is required so that the number of files per directory does not become unmanageable (a very real issue under Unix). If nothing else, the JCSD Corpus is large and comprehensive, making it one of the most extensive Japanese corpora available today.

5.2. Errata

A complete listing of the number of files for each subset of the corpus for each speaker is given in Appendix C. In this section, we document all missing utterances. There is also a directory titled *INFO/doc/errata* included in the corpus distribution. This directory contains files that we consider too anomalous to be part of the corpus proper (for example, incorrectly spoken utterances in which the contents do not match the prompting text). A small number of speaker's data fell into this category: f1034, f1035, f1052, f1069, m0027, and m0035. There are six utterances, one per speaker, included in this directory. We include this data mainly for archival purposes. We don't expect these files to be useful for technology development. The file organization for the errata directory mirrors the one used in the corpus.

In Table 12 below, we list all known missing files for various speakers. The majority of these were the result of data missing from our original DAT copy of the corpus. Some files contained mispronunciations of items (common examples include digits being reversed in a four digit sequence, or an incorrect word being substituted). Files corresponding to mispronunciations were excluded from the corpus even though they contained valid speech data. They were retained, however, in the directory *INFO/doc/errata*.

There were a handful of utterances that contained some anomalous behavior. These are listed in Table 13. Most often, these were the result of a recording chopping the end of a word prematurely,

Subset	Partition	Size (Gbytes)	No. Files
control words (set A)	all	0.808	15,080
	c_0	0.417	7,800
	c_1	0.391	7,280
	males	0.401	7,540
	females	0.407	7,540
control words (set B)	all	1.624	27,840
	c_0	0.839	14,400
	c_1	0.784	13,440
	males	0.799	13,920
	females	0.825	13,920
control words (set C)	all	1.538	30,158
	c_0	0.794	15,599
	c_1	0.744	14,559
	males	0.761	15,080
	females	0.777	15,078
isolated digits	all	0.795	17,390
	c_0	0.411	8,995
	c_1	0.384	8,395
	males	0.400	8,700
	females	0.395	9,230
four digit sequences	all	3.111	40,586
	c_0	1.607	20,993
	c_1	1.504	19,593
	males	1.531	20,296
	females	1.580	20,290
city names	all	6.413	115,976
	c_0	3.313	59,987
	c_1	3.100	55,989
	males	3.159	57,992
	females	3.254	55,984
Subset	Partition	Size (Gbytes)	No. Files
monosyllables	all	5.654	127,579
	c_0	2.918	65,989
	c_1	2.736	61,590
	males	2.805	63,797
	females	2.849	63,782
TOTAL	all	19.943	374,609
	c_0	10.299	193,763
	c_1	9.644	180,846
	males	9.856	187,325
	females	10.087	187,284

Table 11. Sizes of various subsets of the JCSD Corpus.

Speaker No.	Subset	Utterance/ Repetition	Explanation
f1002	four digits	d025_r2 d026_r2 d027_r2	missing from DAT missing from DAT missing from DAT
f1012	monosyllables	m101_r4	missing from DAT
f1016	city names	n098_r2	missing from DAT
f1023	isolated digits	i011_r1 i012_r1 i013_r1 i014_r1 i015_r1	missing from DAT missing from DAT missing from DAT missing from DAT missing from DAT
f1034	control words c	c016_r2	mispronunciation
f1035	city names	c029_r2	mispronunciation
f1047	four digits city names monosyllables	d001_r1 n041_r4 n042_r4 n043_r4 m110_r1	missing from DAT missing from DAT missing from DAT missing from DAT missing from DAT
f1048	city names	n001_r2	missing from DAT
f1052	city names	n076_r2	mispronunciation
f1053	city names	n040_r4	missing from DAT
f1060	monosyllables	m104_r4 m105_r4 m106_r4 m107_r4 m108_r4 m109_r4 m110_r4	missing from DAT missing from DAT missing from DAT missing from DAT missing from DAT missing from DAT missing from DAT
f1069	four digits	d033_r1	transposition of words
m0001	city names	n099_r2	missing from DAT
m0007	city names monosyllables	n100_r1 m059_r4	missing from DAT missing from DAT
m0010	city names	n100_r4	missing from DAT
m0027	four digits	d016_r3	mispronunciation
m0032	city names	n001_r4	missing from DAT
m0035	four digits	d007_r3	mispronunciation
m0050	monosyllables	m110_r2	missing from DAT
m0053	city names	n021_r2	missing from DAT

Table 12. Documented missing files for the JCSD Corpus.

or an artifact on the tape causing some form of artificial distortion (for example, DATs tend to produce a white noise-type signal when there is a tape dropout). These files were included in the final corpus because the utterance is easily recognizable by humans. These utterances would be useful for some forms of robustness experiments, but not perhaps as useful for basic speech recognition system training. With the documentation below, they can be easily removed from the corpus by individual sites.

Speaker No.	Subset	Utterance/Repetition	Explanation
f1009	control words c	c026_r4	end of utterance truncated
f1014	four digits	d033_r4	beginning of utterance truncated
f1027	four digits	d008_r2	stutter in the middle of the utterance
f1029	city names	n087_r2	end of utterance truncated
f1033	isolated digits	i013_r1	"shichi" pronounced as "ichi"
f1045	city names	n059_r3	corrupted data (tape failure)
f1047	monosyllables	m109_r1	end of utterance truncated
f1048	isolated digits	i006_r1	end of utterance truncated
f1064	four digits	d004_r3	end of utterance truncated
m0023	control words b	b023_r1	end of utterance truncated
m0035	four digits	d035_r1	end of utterance truncated
m0047	control words a	a013_r4	end of utterance truncated
m0060	control words a control words b	a013_r4 b005_r1 b016_r1 b022_r1 b003_r2 b020_r2	end of utterance truncated corrupted data (tape failure) corrupted data (tape failure) corrupted data (tape failure) corrupted data (tape failure) corrupted data (tape failure)

Table 13. Miscellaneous anomalous files in the JCSD Corpus.

6. CONCLUSIONS

The JCSD Corpus is impressive if nothing else due to its size and scope. The overall quality of the data is high — only a small percentage of files (approximately 0.01% of the files) are defective. Even though speech recognition technology has moved beyond isolated phrase tasks, this corpus, due to its size and variation, still represents a useful corpus for bootstrapping Japanese language technology — particularly speaker-dependent technology.

It is interesting to note that preparation of this corpus, starting with the DAT tapes, required approximately 2080 hours of labor. This resulted in 86 hours of speech data (actually, only approximately half of that is useful speech data, the remainder is background noise), or a ratio of

0.04 hours of data per hour of labor — a reasonable ratio by today's standards for this type of corpus. Further, the nonrecurring cost of this corpus was approximately 0.3 minutes of speech per dollar of labor — again a fairly reasonable ratio in today's market. Of course, these numbers do not reflect the effort required to originally collect the data, which was easily the lion's share of the project. We believe our productivity was largely due to the highly efficient validation tools developed specifically for this project.

7. ACKNOWLEDGEMENTS

First and foremost, we wish to thank Professor Shuichi Itahashi of the University of Tsukuba for his assistance in this project. Itahashi-sensei was instrumental in facilitating the acquisition of JCSD Corpus, and supplied us with extensive documentation on the content and organization of the DAT version of the corpus. Second, we wish to thank Dr. Jack Godfrey, without whom this project would have never happened. LDC's acquisition of the corpus is largely due to his vision of LDC as an international organization offering historically significant corpora in many languages. Third, we wish to thank Dr. Mark Liberman, Dave Graff, and Rebecca Finch at LDC for their continual support and guidance during this project.

Validating over 350,000 utterances, needless to say, takes it toll on human listeners. We would like to acknowledge several people (listed alphabetically) who worked as validators on this project: Sarah Kaye Blackwell, Michael Brunt, Suehvla El-Attar, Sean Lauderdale, Brian Molen, Mary E. Weber, Leigh A. Webster, Daniel Williams. Their productivity averaged approximately 550 utterances per hour, providing enough throughput to make it possible to do extensive quality control on the corpus.

8. REFERENCES

[1] S. Itahashi, "A Japanese Language Speech Database," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 321-324, Tokyo, Japan, April 1986.

[2] G.R. Doddington and T. B. Schalk, "Speech Recognition: Turning Theory Into Practice," IEEE Spectrum, pp. 26-32, January 1981.

[3] S. Itahashi, "Recent Speech Database Projects in Japan," in *Proceedings International Conference on Spoken Language Processing*, pp. 1081-1084, Kobe, Japan, November 1990.

[4] B. Wheatley, "Transcription Conventions For The CALL HOME Corpus," presented at the TEXLEX Workshop, Linguistic Data Consortium, University of Pennsylvania, Philadephia, Pennsylvania, USA, September 1994.

[5] A. Ganapathiraju, L. Webster, J. Trimble, K. Bush, and P. Kornman, "Comparison Of Energy-Based Endpoint Detectors," in *Proceedings IEEE Southeastcon,* pp. 500-503, Tampa, Florida, USA, April 1996.

APPENDIX A. PROMPTING TEXT

The corpus is divided into seven groups of utterances: Control Words A, Control Words B, Control Words C, Isolated Digits, Four Digit Sequences, City Names, and Monosyllables. The prompting text for each category is shown below, using a hiragana transcription system provided with the original corpus:

Control Words:

Index	Control Words A	Control Words B	Control Words C
1	syo-kai	zeNkaku	bi-
2	horyu-	so-nyu-	oN
3	hurikomi	kaigyo-	jaku
4	do-zo	haNkaku	e-
5	mo-ichido	buNmatsu	shita
6	zaNdaka	heNkaN	shi-
7	hai	ido-	dai
8	torikeshi	jipe-ji	kesu
9	owari	ke-seN	ko-tai
10	hajime	kaipe-ji	kyo-
11	i-e	katakana	shimeru
12	te-se-	seNtariNgu	suta-to
13	toritsugi	muheNkaN	hidari
14		zeNpe-ji	mae
15		taNgo to-roku	sutoppu
16		jikko-	sho-
17		kiNto-waritsuke	ushiro
18		supe-su	zeNshiN
19		tabu	ohu
20		shu-ryo-	appu
21		sakujo	chu-
22		aNda-raiN	ue
23		baikaku	dauN
24		wakuake	migi
25			tsukeru
26			akeru

Digit Sequences:

Index	Isolated Digits	Four Digit Sequences
1	zero	zero ni hachi nana
2	san	go nana san ni
3	ni	kyu roku zero ichi
4	rei	yon ichi go roku
5	nana	ichi ichi kyu kyu
6	yon	ichi san kyu hachi
7	go	roku hachi yon san
8	maru	zero nana ichi ni
9	shi	go ni roku nana
10	roku	roku roku san san
11	ku	ni yon zero kyu
12	hachi	nana kyu go yon
13	shichi	ichi hachi ni san
14	kyu	roku san nana hachi
15	ichi	hachi hachi nana nana
16		san go ichi zero
17		hachi zero roku go
18		ni kyu san yon
19		nana yon hachi kyu
20		ni ni yon yon
21		yon roku ni ichi
22		kyu ichi nana roku
23		san zero yon go
24		hachi go kyu zero
25		go go zero zero
26		roku kyu nana ni
27		go hachi roku ichi
28		san roku yon kyu
29		zero san ichi roku
30		nana zero hachi san
31		hachi ichi kyu yon
32		kyu ni zero go
33		ichi yon ni nana
34		ni go san hachi
35		yon nana go zero

City Names:

Index	City Names	Index	City Names	Index	City Names	Index	City Names
1	hachinohe	26	matto-	51	meguro	76	chiryu-
2	keseNnuma	27	o-bu	52	gushikawa	77	buNkyo-
3	yukuhashi	28	ko-be	53	numazu	78	o-muta
4	sapporo	29	gobo-	54	rikuzeNtakada	79	yu-ki
5	kitami	30	siNgu-	55	koganei	80	suzu
6	eniwa	31	kure	56	yonago	81	neyagawa
7	yokote	32	mine	57	sasebo	82	eNzaN
8	toride	33	buzeN	58	tochigi	83	hisai
9	warabi	34	beppu	59	kamo	84	muko-
10	asahi	35	nemuro	60	wako-	85	daito-
11	tsuruga	36	susono	61	bisai	86	kuroiso
12	takehu	37	gamago-ri	62	tsuru	87	ninohe
13	hekinaN	38	so-ja	63	ube	88	goteNba
14	yasugi	39	chiba	64	iyo	89	moriguchi
15	zentsu-ji	40	noda	65	nihoNmatsu	90	so-ka
16	rumoi	41	zushi	66	yame	91	seki
17	bibai	42	date	67	takeo	92	hoNjo-
18	seNdai	43	himi	68	hoNdo	93	saNjo-
19	teNdo-	44	gihu	69	mito	94	ebina
20	naNyo-	45	mutsu	70	naha	95	masuda
21	mo-ka	46	ageo	71	shiNjo-	96	ko-hu
22	kazo	47	ehime	72	kiryu-	97	bizeN
23	ni-za	48	hujiidera	73	hanyu-	98	huji
24	ho-ya	49	ibusuki	74	teNryu-	99	hagi
25	uozu	50	zama	75	huchu-	100	kiyose

Monosyllables:

Index	Mono-syllables	Index	Mono-syllables	Index	Mono-syllables	Index	Mono-syllables
1	ha	29	nya	57	na	85	mu
2	hyo	30	shi	58	ku	86	so
3	a	31	ho	59	nu	87	sho
4	myu	32	chu	60	ji	88	ba
5	ga	33	byu	61	ma	89	zo
6	mo	34	ne	62	e	90	ni
7	cho	35	bi	63	za	91	da
8	be	36	bya	64	hya	92	myo
9	bo	37	byo	65	tsu	93	pe
10	nyo	38	ka	66	hu	94	sha
11	kya	39	po	67	go	95	chi
12	pi	40	ro	68	ta	96	ke
13	gu	41	i	69	ze	97	do
14	sa	42	me	70	rya	98	o
15	ya	43	pa	71	gyo	99	nyu
16	pyo	44	kyo	72	kyu	100	ko
17	mya	45	bu	73	u	101	n
18	se	46	gya	74	de	102	ti
19	gi	47	pu	75	ja	103	je
20	ge	48	su	76	ri	104	fa
21	zu	49	to	77	pya	105	fo
22	pyu	50	shu	78	yo	106	she
23	ra	51	hi	79	gyu	107	di
24	ju	52	yu	80	wa	108	fi
25	mi	53	ru	81	he	109	che
26	jo	54	re	82	cha	110	fe
27	te	55	no	83	ryu		
28	ryo	56	hyu	84	ki		

APPENDIX B. DAT INVENTORY

The source format for the JEIDA corpus was a set of 76 Digital Audio Tapes (DAT):

Tape No.	ID No.	Material	Duration (hours)	Speakers
1	1-1	Control Words A	1	m0001 — m0020
2	1-2	Control Words A	1	m0021 — m0025, m0076, m0027 — m0029, m0077, m0031 — m0040
3	2-1	Control Words A	1	m0041 — m0044, m0075, m0045 — m0059
4	2-2	Control Words A	1	m0060 — m0074
5	3-1	Control Words A	1	f1001 — f1020
6	3-2	Control Words A	1	f1021 — f1040
7	4-1	Control Words A	1	f1041 — f1060
8	4-2	Control Words A	1	f1061 — f0175
9	1-1	Control Words B	1	m0001 — m0015
10	1-2	Control Words B	1	m0016 — m0025, m0076 m0027 — m0029, m0077
11	2-1	Control Words B	1	m0031 — m0044, m0075
12	2-2	Control Words B	1	m0045 — m0059
13	3	Control Words B	1	m0060 — m0074
14	4-1	Control Words B	1	f0001 — f0015
15	4-2	Control Words B	1	f0016 — f0030
16	5-1	Control Words B	1	f0031 — f0045
17	5-2	Control Words B	1	f0046 — f0060
18	6	Control Words B	1	f0061 — f0075
19	1	Control Words C	2	m0001 — m0025
20	2	Control Words C	2	m0076, m0027 — m0029, m0077, m0031 — m0044, m0075, m0045 — m0049
21	3	Control Words C	2	m0050 — m0074
22	4	Control Words C	2	f1001 — f1025
23	5	Control Words C	2	f1026 — f1050
24	6	Control Words C	2	f1051 — f1075

Tape No.	ID No.	Material	Duration (hours)	Speakers
25	1-1	Isolated Digits	1	m0001 — m0020
26	1-2	Isolated Digits	1	m0021 — m0025, m0076, m0027 — m0029, m0077, m0031 — m0040,
27	2-1	Isolated Digits	1	m0041 — m0044, m0075, m0045 — m0059
28	2-2	Isolated Digits	1	m0060 — m0074
29	3-1	Isolated Digits	1	f1001 — f1020
30	3-2	Isolated Digits	1	f1021 — f1040
31	4-1	Isolated Digits	1	f1041 — f1060
32	4-2	Isolated Digits	1	f1061 — f0175
33	1	4-Digit Sequences	2	m0001 — m0017
34	2	4-Digit Sequences	2	m0018 — m0025, m0076, m0027 — m0029, m0077, m0031 — m0034
35	3	4-Digit Sequences	2	m0035 — m0044, m0075, m0045 — m0050
36	4	4-Digit Sequences	2	m0051 — m0067
37	5	4-Digit Sequences	2	m0068 — m0074
38	6	4-Digit Sequences	2	f1001 — f1017
39	7	4-Digit Sequences	2	f1018 — f1034
40	8	4-Digit Sequences	2	f1035 — f1051
41	9	4-Digit Sequences	2	f1052 — f1068
42	10	4-Digit Sequences	1	f1069 — f1075

Tape No.	ID No.	Material	Duration (hours)	Speakers
43	1	City Names	2	m0001 — m0005, m0009, m0007, m0008, m0010
44	2	City Names	2	m0006, m0011 — m0018
45	3	City Names	2	m0019 — m0025, m0076, m0027
46	4	City Names	2	m0028 — m0036
47	5	City Names	2	m0037 — m0044, m0075
48	6	City Names	2	m0045 — m0053
49	7	City Names	2	m0054 — m0062
50	8	City Names	2	m0063 — m0071
51	9	City Names	1	m0072 — m0074
52	10	City Names	2	f1001 — f1009
53	11	City Names	2	f1010 — f1018
54	12	City Names	2	f1019 — f1027
55	13	City Names	2	f1028 — f1036
56	14	City Names	2	f1037 — f1045
57	15	City Names	2	f1046 — f1054
58	16	City Names	2	f1055 — f1063
59	17	City Names	2	f1064 — f1072
60	18	City Names	1	f1073 — f1075

Tape No.	ID No.	Material	Duration (hours)	Speakers
61	1	Monosyllables	2	m0001 — m0010
62	2	Monosyllables	2	m0010 — m0020
63	3	Monosyllables	2	m0021 — m0025, m0076 m0027 — m0029, m0077
64	4	Monosyllables	2	m0031 — m0040
65	5	Monosyllables	2	m0041 — m0044, m0075, m0045 — m0049
66	6	Monosyllables	2	m0050 — m0059
67	7	Monosyllables	2	m0060 — m0069
68	8	Monosyllables	1	m0070 — m0074
69	9	Monosyllables	2	f1001 — f1010
70	10	Monosyllables	2	f1011 — f1020
71	11	Monosyllables	2	f1021 — f1030
72	12	Monosyllables	2	f1031 — f1040
73	13	Monosyllables	2	f1041 — f1050
74	14	Monosyllables	2	f1051 — f1060
75	15	Monosyllables	2	f1061 — f1070
76	16	Monosyllables	1	f1071 — f1075

APPENDIX C. SPEAKER SUMMARY

A listing of key demographic information for each speaker is given below.

Speaker No.	Information			
	Sex	Age	Address Under Age 12	Present Address
	Height	Noise	Recording Environment	
f1001	Female	10-19	Kanagawa, Kanto	-
	150cm	19dBA	Soundproof room	
f1002	Female	20-29	Tokyo, Kanto	Minato, Tokyo
	160cm	19dBA	Soundproof room	
f1003	Female	30-39	???	Suginami, Tokyo
	162cm	19dBA	Soundproof room	
f1004	Female	40-49	Sodegaura, Kimizu, Chiba, Kanto	Edogawa, Tokyo
	153cm	19dBA	Soundproof room	
f1005	Female	50-59	Tochigi, Tochigi, Kanto	Tochigi, Tochigi
	152cm	19dBA	Soundproof room	
f1006	Female	20-29	Yokosuka, Kanagawa, Kanto	Miura, Kanagawa
	152cm	45dBA	Soundproof room	
f1007	Female	20-29	Yokosuka, Kanagawa, Kanto	Yokosuka, Kanagawa
	153cm	45dBA	Soundproof room	
f1008	Female	20-29	Sagamihara, Kanagawa, Kanto	Sagamihara, Kanagawa
	157cm	45dBA	Soundproof room	
f1009	Female	40-49	Yokosuka, Kanagawa, Kanto	Yokosuka, Kanagawa
	151cm	45dBA	Soundproof room	
f1010	Female	50-59	Tokyo, Kanto	Musashino, Tokyo
	148cm	45dBA	Soundproof room[	
f1011	Female	20-29	Yuki, Ibaraki, Kanto	-
	153cm	-	Simple soundproof room	
f1012	Female	20-29	Yokohama, Kanagawa, Kanto	-
	155cm	< 30dBA	Simple soundproof room	
f1013	Female	30-39	Niigata, Niigata, Chubu	-
	158cm	< 30dBA	Simple soundproof room	
f1014	Female	40-49	Ota, Tokyo, Kanto	-
	153cm	< 30dBA	Simple soundproof room	
f1015	Female	50-59	Dairen, Kanton, Mansyu, Kanto	-
	156cm	< 30dBA	Simple soundproof room	
f1016	Female	20-29	Yokohama, Kanagawa, Kanto	Suginami, Tokyo
	156cm	< 30dBA	Soundproof room	
f1017	Female	20-29	Nerima, Tokyo, Kanto	Suginami, Tokyo
	161cm	< 30dBA	Soundproof room	
f1018	Female	30-39	Nakano, Tokyo, Kanto	Ichikawa, Chiba
	162cm	< 30dBA	Soundproof room	
f1019	Female	40-49	Shima, Mie, Kinki	Kamakura, Kanagawa
	158cm	< 30dBA	Soundproof room	

Speaker No.	Information			
	Sex	Age	Address Under Age 12	Present Address
	Height	Noise	Recording Environment	
f1020	Female	50-59	Mito, Ibaraki, Kanto	Meguro, Tokyo
	160cm	< 30dBA	Soundproof room	
f1021	Female	20-29	Okaya, Nagano, Chubu	Hachioji, Tokyo
	156cm	46dBA	Soundproof room	
f1022	Female	30-39	Oume, Tokyo, Kanto	Oume, Tokyo
	153cm	46dBA	Soundproof room	
f1023	Female	30-39	Nishitama, Tokyo, Kanto	Nishitama, Tokyo
	163cm	46dBA	Soundproof room	
f1024	Female	40-49	Tokyo, Kanto	Kokubunji, Tokyo
	149cm	46dBA	Soundproof room	
f1025	Female	50-59	Tokyo, Kanto	Kokubunji, Tokyo
	153cm	46dBA	Soundproof room	
f1026	Female	20-29	Fuji, Shizuoka, Chubu	Setagaya, Tokyo
	157cm	-	Meeting room	
f1027	Female	30-39	Sado, Niigata, Chubu	Komae, Tokyo
	157cm	47dBA	Meeting room	
f1028	Female	30-39	Tokyo, Kanto	Hachioji, Tokyo
	150cm	-	Meeting room	
f1029	Female	40-49	Tokyo, Kanto	Musashino, Tokyo
	158cm	-	Meeting room	
f1030	Female	40-49	Kokubunji, Tokyo, Kanto	Toshima, Tokyo
	150cm	-	-	
f1031	Female	20-29	Komae, Tokyo, Kanto	Yokohama, Kanagawa
	158cm	29dBA	Simple soundproof room	
f1032	Female	20-29	Kawasaki, Kanagawa, Kanto	Kawasaki, Kanagawa
	163cm	29dBA	Simple soundproof room	
f1033	Female	30-39	Higashiuwa, Ehime, Shikoku	Machida, Tokyo
	149cm	29dBA	Simple soundproof room	
f1034	Female	40-49	Tokyo, Kanto	Yokohama, Kanagawa
	150cm	29dBA	Simple soundproof room	
f1035	Female	50-59	Tokyo, Kanto	Yokohama, Kanagawa
	150cm	29dBA	Simple soundproof room	
f1036	Female	20-29	Nishinomiya, Hyogo, Kinki	Yokohama, Kanagawa
	163cm	34dBA	Simple soundproof room	
f1037	Female	20-29	Naka, Kanagawa, Kanto	Naka, Kanagawa
	153cm	34dBA	Simple soundproof room	
f1038	Female	30-39	Odawara, Kanagawa, Kanto	Odawara, Kanagawa
	155cm	34dBA	Simple soundproof room	
f1039	Female	40-49	Minamikorai, Nagasaki, Kyushu	Kamakura, Kanagawa
	158cm	34dBA	Simple soundproof room	
f1040	Female	50-59	Kamakura, Kanagawa, Kanto	Kamakura, Kanagawa
	158cm	34dBA	Simple soundproof room	

Speaker No.	Information			
	Sex	Age	Address Under Age 12	Present Address
	Height	Noise	Recording Environment	
f1041	Female	20-29	Yokohama, Kanagawa, Kanto	Minato, Tokyo
	159cm	30dBA	Soundproof room	
f1042	Female	20-29	Tanashi, Tokyo, Kanto	Yokohama, Kanagawa
	150cm	30dBA	Soundproof room	
f1043	Female	30-39	Kawasaki, Kanagawa, Kanto	Kawasaki, Kanagawa
	156cm	30dBA	Soundproof room	
f1044	Female	30-39	Ota, Tokyo, Kanto	Yokohama, Kanagawa
	149cm	30dBA	Soundproof room	
f1045	Female	40-49	Chuo, Tokyo, Kanto	Yokohama, Kanagawa
	150cm	30dBA	Soundproof room	
f1046	Female	20-29	Hachioji, Tokyo, Kanto	Setagaya, Tokyo
	155cm	30dBA	Soundless room	
f1047	Female	30-39	Kamakura, Kanagawa, Kanto	Setagaya, Tokyo
	154cm	30dBA	Soundless room	
f1048	Female	30-39	Takada, Niigata, Chubu	Sagamihara, Kanagawa
	151cm	30dBA	Soundless room	
f1049	Female	40-49	Bunkyo, Tokyo, Kanto	Bunkyo, Tokyo
	158cm	30dBA	Soundless room	
f1050	Female	40-49	Tokyo, Kanto	Suginami, Tokyo
	145cm	30dBA	Soundless room	
f1051	Female	20-29	Setagaya, Tokyo, Kanto	Setagaya, Tokyo
	160cm	37dBA	Soundproof room	
f1052	Female	30-39	Chuo, Tokyo, Kanto	Hachioji, Tokyo
	153cm	37dBA	Soundproof room	
f1053	Female	30-39	Hachioji, Tokyo, Kanto	Hachioji, Tokyo
	161cm	37dBA	Soundproof room	
f1054	Female	40-49	Tokyo, Kanto	Hino, Tokyo
	153cm	-	Soundproof room	
f1055	Female	50-59	Saeki, Hiroshima, Chugoku	Hachioji, Tokyo
	158cm	37dBA	Soundproof room	
f1056	Female	20-29	Toyonaka, Osaka, Kinki	Nara, Nara
	162cm	30dBA	Soundproof room	
f1057	Female	20-29	Kitauwa, Ehime, Shikoku	Tenri, Nara
	165cm	30dBA	Soundproof room	
f1058	Female	20-29	Aira, Kagoshima, Kyushu	Nara, Nara
	155cm	29dBA	Soundproof room	
f1059	Female	40-49	Higashiosaka, Osaka, Kinki	Nara, Nara
	158cm	32dBA	Soundproof room	
f1060	Female	50-59	Tenri, Nara, Kinki	Tenri, Nara
	152cm	32dBA	Soundproof room	
f1061	Female	20-29	Yokohama, Kanagawa, Kanto	Yokohama, Kanagawa
	168cm	40dBA	Simple soundproof room	

Speaker No.	Information			
	Sex	Age	Address Under Age 12	Present Address
	Height	Noise	Recording Environment	
f1062	Female	20-29	Fukushima, Tohoku	Kawasaki, Kanagawa
	154cm	40dBA	Simple soundproof room	
f1063	Female	30-39	Yokohama, Kanagawa, Kanto	Inagi, Tokyo
	158cm	40dBA	Simple soundproof room	
f1064	Female	40-49	Niigata, Kanto	Kawasaki, Kanagawa
	151cm	40dBA	Simple soundproof room	
f1065	Female	50-59	Fukuoka, Fukuoka, Kyushu	Kawasaki, Kanagawa
	160cm	40dBA	Simple soundproof room	
f1066	Female	20-29	Nanyo, Yamagata, Tohoku	Sakura, Niihari, Ibaraki
	159cm	< 25dBA	Simple soundless room	
f1067	Female	30-39	Iwaki, Fukushima, Tohoku	Sakura, Niihari, Ibaraki
	157cm	< 25dBA	Simple soundless room	
f1068	Female	30-39	Osaka, Kinki	Sakura, Niihari, Ibaraki
	155cm	< 25dBA	Simple soundless room	
f1069	Female	40-49	Tokyo, Kanto	Sakura, Niihari, Ibaraki
	162cm	< 25dBA	Simple soundless room	
f1070	Female	60-69	Ibusuki, Kagoshima, Kyushu	Sakura, Niihari, Ibaraki
	157cm	< 25dBA	Simple soundless room	
f1071	Female	20-29	Utsunomiya, Tochigi, Kanto	Shimotsuga, Tochigi
	156cm	25dBA	Soundproof room	
f1072	Female	30-39	Nihonmatsu, Fukushima, Tohoku	Utsunomiya, Tochigi
	157cm	25dBA	Soundproof room	
f1073	Female	30-39	Touhaku, Tottori, Chugoku	Utsunomiya, Tochigi
	159cm	25dBA	Soundproof room	
f1074	Female	40-49	Tochigi, Kanto	Utsunomiya, Tochigi
	156cm	25dBA	Soundproof room	
f1075	Female	50-59	Utsunomya, Tochigi, Kanto	Utsunomya, Tochigi
	156cm	25dBA	Soundproof room	
m0001	Male	20-29	Yokohama, Kanagawa, Kanto	Sagamihara, Kanagawa
	181cm	19dBA	Soundproof room	
m0002	Male	20-29	Tokyo, Kanto	Yokohama, Kanagawa
	171cm	19dBA	Soundproof room	
m0003	Male	30-39	Shinagawa, Tokyo, Kanto	Machida, Tokyo
	169cm	19dBA	Soundproof room	
m0004	Male	40-49	Fuchu, Hiroshima, Chugoku	Funabashi, Chiba
	164cm	19dBA	Soundproof room	
m0005	Male	50-59	Tokyo, Kanto	Shinjuku, Tokyo
	167cm	19dBA	Soundproof room	
m0006	Male	50-59	Tokyo, Kanto	Kokubunji, Tokyo
	161cm	45dBA	Soundproof room	
m0007	Male	30-39	Himeji, Hyogo, Kinki	Koganei, Tokyo
	173cm	45dBA	Soundproof room	

Speaker No.	Information			
	Sex	Age	Address Under Age 12	Present Address
	Height	Noise	Recording Environment	
m0008	Male	30-39	Kobe, Hyogo, Kinki	Yokosuka, Kanagawa
	168cm	45dBA	Soundproof room	
m0009	Male	20-29	Fukuoka, Fukuoka, Kyushu	Yokosuka, Kanagawa
	174cm	45dBA	Soundproof room	
m0010	Male	40-49	Kobe, Hyogo, Kinki	Yokosuka, Kanagawa
	169cm	45dBA	Soundproof room	
m0011	Male	20-29	Hanejima, Gifu, Chubu	Tsukuba, Ibaraki
	174cm	< 30dBA	Simple soundproof room	
m0012	Male	30-39	Toyama, Chubu	Niihari, Ibaraki
	176cm	< 30dBA	Simple soundproof room	
m0013	Male	30-39	Morioka, Iwate, Tohoku	Toride, Ibaraki
	170cm	< 30dBA	Simple soundproof room	
m0014	Male	40-49	Takayama, Gifu, Chubu	Ushiku, Ibaraki
	162cm	< 30dBA	Simple soundproof room	
m0015	Male	50-59	Ota, Tokyo, Kanto	Sakura, Niihari,Ibaraki
	167cm	< 30dBA	Simple soundproof room	
m0016	Male	20-29	Nakano, Tokyo, Kanto	-
	170cm	< 30dBA	Soundproof room	
m0017	Male	20-29	Minato, Tokyo, Kanto	-
	172cm	< 30dBA	Soundproof room	
m0018	Male	30-39	Saaebo, Nagasaki, Kyushu	-
	170cm	< 30dBA	Soundproof room	
m0019	Male	40-49	Hamakita, Shizuoka, Chubu	-
	167cm	< 30dBA	Soundproof room	
m0020	Male	60-69	Ueno, Mie, Kinki	-
	155cm	< 30dBA	Soundproof room	
m0021	Male	20-29	Toyoshima, Tokyo, Kanto	Kawasaki, Kanagawa
	176cm	46dBA	Soundproof room	
m0022	Male	20-29	Yokohama, Kanagawa, Kanto	Kokubunji, Tokyo
	170cm	46dBA	Soundproof room	
m0023	Male	30-39	Sugakawa, Fukushima, Tohoku	Tsukui, Kanagawa
	172cm	46dBA	Soundproof room	
m0024	Male	40-49	Suginami, Tokyo, Kanto	Musashino, Tokyo
	174cm	46dBA	Soundproof room	
m0025	Male	50-59	Sumida, Tokyo, Kanto	Tsukui, Kanagawa
	163cm	46dBA	Soundproof room	
m0027	Male	20-29	Chigasaki, Kanagawa, Kanto	Shibuya, Tokyo
	180cm	-	Meeting room	
m0028	Male	30-39	Meguro, Tokyo, Kanto	Yokohama, Kanagawa
	170cm	-	Meeting room	
m0029	Male	40-49	Mitaka, Tokyo, Kanto	Yokohama, Kanagawa
	175cm	-	Meeting room	

Speaker No.	Information			
	Sex	Age	Address Under Age 12	Present Address
	Height	Noise	Recording Environment	
m0031	Male	20-29	Tokyo, Kanto	Machida, Tokyo
	168cm	29dBA	Simple soundproof room	
m0032	Male	30-39	Numatsu, Shizuoka, Chubu	Sagamihara, Kanagawa
	166cm	29dBA	Simple soundproof room	
m0033	Male	30-39	Minato, Tokyo, Kanto	Minato, Tokyo
	165cm	29dBA	Simple soundproof room	
m0034	Male	40-49	Shirakawa, Fukushima, Tohoku	Yokohama, Kanagawa
	167cm	29dBA	Simple soundproof room	
m0035	Male	50-59	Omiya, Saitama, Kanto	Yokohama, Kanagawa
	166cm	29dBA	Simple soundproof room	
m0036	Male	20-29	Ota, Tokyo, Kanto	Yokohama, Kanagawa
	168cm	34dBA	Simple soundproof room	
m0037	Male	20-29	Yokohama, Kanagawa, Kanto	Fujisawa, Kanagawa
	175cm	34dBA	Simple soundproof room	
m0038	Male	30-39	Iida, Nagano, Chubu	Hiratsuka, Kanagawa
	171cm	34dBA	Simple soundproof room	
m0039	Male	40-49	Kyoto, Kyoto, Kinki	Kamakura, Kanagawa
	166cm	34dBA	Simple soundproof room	
m0040	Male	60-69	Tokyo, Kanto	Kamakura, Kanagawa
	163cm	34dBA	Simple soundproof room	
m0041	Male	20-29	Karatsu, Saga, Kyushu	Kawasaki, Kanagawa
	172cm	30dBA	Soundproof room	
m0042	Male	30-39	Musashino, Tokyo, Kanto	Musashino, Tokyo
	170cm	30dBA	Soundproof room	
m0043	Male	30-39	Shinjuku, Tokyo, Kanto	Kawasaki, Kanagawa
	170cm	30dBA	Soundproof room	
m0044	Male	40-49	Ota, Tokyo, Kanto	Machida, Tokyo
	163cm	30dBA	Soundproof room	
m0045	Male	20-29	Funabashi, Chiba, Kanto	Setagaya, Tokyo
	170	30dBA	Soundless room	
m0046	Male	20-29	Fuchu, Tokyo, Kanto	Komae, Tokyo
	165cm	30dBA	Soundless room	
m0047	Male	30-39	Urawa, Saitama, Kanto	Suginami, Tokyo
	165cm	30dBA	Soundless room	
m0048	Male	40-49	Nakauonuma, Niigata, Chubu	Minato, Tokyo
	168cm	30dBA	Soundless room	
m0049	Male	50-59	Tokyo, Kanto	Shibuya, Tokyo
	173cm	30dBA	Soundless room	
m0050	Male	20-29	Hitachiota, Ibaraki, Kanto	Kodaira, Tokyo
	166cm	37dBA	Soundproof room	
m0051	Male	30-39	Kitakyushu, Fukuoka, Kyushu	Sagamihara, Kanagawa
	168cm	37dBA	Soundproof room	

Speaker No.	Information			
	Sex	Age	Address Under Age 12	Present Address
	Height	Noise	Recording Environment	
m0052	Male	20-29	Mitaka, Tokyo, Kanto	Mitaka, Tokyo
	172cm	37dBA	Soundproof room	
m0053	Male	50-59	Sakura, Chiba, Kanto	Nakano, Tokyo
	164cm	37dBA	Soundproof room	
m0054	Male	40-49	Tokyo, Kanto	Hachioji, Tokyo
	165cm	37dBA	Soundproof room	
m0055	Male	20-29	Kure, Hiroshima, Chugoku	Tenri, Nara
	173cm	27dBA	Soundproof room	
m0056	Male	30-39	Nara, Nara, Kinki	Nara, Nara
	176cm	30dBA	Soundproof room	
m0057	Male	30-39	Nagoya, Aichi, Chubu	Nara, Nara
	164cm	30dBA	Soundproof room	
m0058	Male	40-49	Osaka, Osaka, Kinki	Nara, Nara
	174cm	32dBA	Soundproof room	
m0059	Male	50-59	Sakata, shiga, Kinki	Nara, Nara
	166cm	32dBA	Soundproof room	
m0060	Male	20-29	Fukuoka, Fukuoka, Kyushu	Ota, Tokyo
	170cm	40dBA	Simple soundproof room	
m0061	Male	30-39	Sapporo, Hokkaido, Hokkaido	Yokohama, Kanagawa
	166cm	40dBA	Simple soundproof room	
m0062	Male	30-39	Kobe, Hyogo, Kinki	Yokohama, Kanagawa
	170cm	40dBA	Simple soundproof room	
m0063	Male	40-49	Yokohama, Kanagawa, Kanto	Yokohama, Kanagawa
	168cm	40dBA	Simple soundproof room	
m0064	Male	60-69	Tokyo, Kanto	Shinagawa, Tokyo
	163cm	40dBA	Simple soundproof room	
m0065	Male	20-29	Naka, Kanagawa, Kanto	Sakura, Niihari, Ibaraki
	160cm	< 25dBA	Simple soundless room	
m0066	Male	20-29	Joetsu, Niigata, Chubu	Sakura, Niihari, Ibaraki
	173cm	< 25dBA	Simple soundless room	
m0067	Male	20-29	Fukushima, Fukushima, Tohoku	Sakura, Niihari, Ibaraki
	169cm	< 25dBA	Simple soundless room	
m0068	Male	40-49	Furukawa, Miyagi, Tohoku	Sakura, Niihari, Ibaraki
	161cm	< 25dBA	Simple soundless room	
m0069	Male	60-69	Hiroshima, Hiroshima, Chugoku	Sakura, Niihari, Ibaraki
	168cm	< 25dBA	Simple soundless room	
m0070	Male	20-29	Utsunomiya, Tochigi, Kanto	-
	172cm	25dBA	Soundproof room	
m0071	Male	20-29	Awa, Chiba, Kanto	-
	165cm	25dBA	Soundproof room	
m0072	Male	30-39	Utsunomiya, Tochigi, Kanto	-
	175cm	25dBA	Soundproof room	

<table>
<tr><th rowspan="3">Speaker No.</th><th colspan="4">Information</th></tr>
<tr><th>Sex</th><th>Age</th><th>Address Under Age 12</th><th>Present Address</th></tr>
<tr><th>Height</th><th>Noise</th><th colspan="2">Recording Environment</th></tr>
<tr><td rowspan="2">m0073</td><td>Male</td><td>40-49</td><td>Nishitagawa, Yamagata, Tohoku</td><td>Utsunomiya, Tochigi</td></tr>
<tr><td>171cm</td><td>25dBA</td><td colspan="2">Soundproof room</td></tr>
<tr><td rowspan="2">m0074</td><td>Male</td><td>50-59</td><td>Utsunomya, Tochigi, Kanto</td><td>-</td></tr>
<tr><td>165cm</td><td>25dBA</td><td colspan="2">Soundproof room</td></tr>
<tr><td rowspan="2">m0075</td><td>Male</td><td>50-59</td><td>Shinagawa, Tokyo, Kanto</td><td>Shinagawa, Tokyo</td></tr>
<tr><td>165cm</td><td>30dBA</td><td colspan="2">Soundproof room</td></tr>
<tr><td rowspan="2">m0076</td><td>Male</td><td>20-29</td><td>Kashihara, Nara, Kinki</td><td>Yokohama, Kanagawa</td></tr>
<tr><td>177cm</td><td>-</td><td colspan="2">Meeting room</td></tr>
<tr><td rowspan="2">m0077</td><td>Male</td><td>50-59</td><td>Nagoya, Aichi, Chubu</td><td>Kawasaki, Kanagawa</td></tr>
<tr><td>164cm</td><td>-</td><td colspan="2">Meeting room</td></tr>
</table>

A listing of the file count for each speaker in the corpus follows. Each entry contains two values: the number of files for channel 0 and channel 1. Discrepancies occur because some speakers were missing data for channel 1.Note that speakers f1006 through f1010, and m0006 through m0010 were missing channel 1.

Speaker No.	Chan	Ctrl Words A (a)	Ctrl Words B (b)	Ctrl Words C (c)	Isolated Digits (i)	Four Digits (d)	City Names (n)	Mono-syllables (m)
f1001	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1002	c0	52	96	104	60	137	400	440
	c1	52	96	104	60	137	400	440
f1003	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1004	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1005	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1006	c0	52	96	104	60	140	400	440
	c1	0	0	0	0	0	0	0
f1007	c0	52	96	104	60	140	400	440
	c1	0	0	0	0	0	0	0
f1008	c0	52	96	104	60	140	400	440
	c1	0	0	0	0	0	0	0
f1009	c0	52	96	104	60	140	400	440
	c1	0	0	0	0	0	0	0
f1010	c0	52	96	104	60	140	400	440
	c1	0	0	0	0	0	0	0
f1011	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440

Speaker No.	Chan	Ctrl Words A (a)	Ctrl Words B (b)	Ctrl Words C (c)	Isolated Digits (i)	Four Digits (d)	City Names (n)	Mono-syllables (m)
f1012	c0	52	96	104	60	140	400	439
	c1	52	96	104	60	140	400	439
f1013	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1014	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1015	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1016	c0	52	96	104	60	140	399	440
	c1	52	96	104	60	140	399	440
f1017	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1018	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1019	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1020	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1021	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1022	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1023	c0	52	96	104	55	140	400	440
	c1	52	96	104	55	140	400	440
f1024	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1025	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1026	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1027	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1028	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1029	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1030	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1031	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1032	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440

Speaker No.	Chan	Ctrl Words A (a)	Ctrl Words B (b)	Ctrl Words C (c)	Isolated Digits (i)	Four Digits (d)	City Names (n)	Mono-syllables (m)
f1033	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1034	c0	52	96	103	60	140	400	440
	c1	52	96	103	60	140	400	440
f1035	c0	52	96	104	60	140	399	440
	c1	52	96	104	60	140	399	440
f1036	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1037	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1038	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1039	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1040	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1041	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1042	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1043	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1044	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1045	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1046	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1047	c0	52	96	104	60	139	397	439
	c1	52	96	104	60	139	397	439
f1048	c0	52	96	104	60	140	399	440
	c1	52	96	104	60	140	399	440
f1049	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1050	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1051	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1052	c0	52	96	104	60	140	399	440
	c1	52	96	104	60	140	399	440
f1053	c0	52	96	104	60	140	399	440
	c1	52	96	104	60	140	399	440

Speaker No.	Chan	Ctrl Words A (a)	Ctrl Words B (b)	Ctrl Words C (c)	Isolated Digits (i)	Four Digits (d)	City Names (n)	Mono-syllables (m)
f1054	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1055	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1056	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1057	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1058	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1059	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1060	c0	52	96	104	60	140	400	433
	c1	52	96	104	60	140	400	433
f1061	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1062	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1063	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1064	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1065	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1066	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1067	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1068	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1069	c0	52	96	104	60	139	400	440
	c1	52	96	104	60	139	400	440
f1070	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1071	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1072	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1073	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
f1074	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440

Speaker No.	Chan	Ctrl Words A (a)	Ctrl Words B (b)	Ctrl Words C (c)	Isolated Digits (i)	Four Digits (d)	City Names (n)	Mono-syllables (m)
f1075	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0001	c0	52	96	104	60	140	399	440
	c1	52	96	104	60	140	399	440
m0002	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0003	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0004	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0005	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0006	c0	52	96	104	60	140	400	440
	c1	0	0	0	0	0	0	0
m0007	c0	52	96	104	60	140	399	439
	c1	0	0	0	0	0	0	0
m0008	c0	52	96	104	60	140	400	440
	c1	0	0	0	0	0	0	0
m0009	c0	52	96	104	60	140	400	440
	c1	0	0	0	0	0	0	0
m0010	c0	52	96	104	60	140	399	440
	c1	0	0	0	0	0	0	0
m0011	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0012	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0013	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0014	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0015	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0016	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0017	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0018	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0019	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0020	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440

Speaker No.	Chan	Ctrl Words A (a)	Ctrl Words B (b)	Ctrl Words C (c)	Isolated Digits (i)	Four Digits (d)	City Names (n)	Mono-syllables (m)
m0021	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0022	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0023	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0024	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0025	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0027	c0	52	96	104	60	139	400	440
	c1	52	96	104	60	139	400	440
m0028	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0029	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0031	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0032	c0	52	96	104	60	140	399	440
	c1	52	96	104	60	140	399	440
m0033	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0034	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0035	c0	52	96	104	60	139	400	440
	c1	52	96	104	60	139	400	440
m0036	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0037	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0038	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0039	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0040	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0041	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0042	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0043	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440

Speaker No.	Chan	Ctrl Words A (a)	Ctrl Words B (b)	Ctrl Words C (c)	Isolated Digits (i)	Four Digits (d)	City Names (n)	Mono-syllables (m)
m0044	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0045	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0046	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0047	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0048	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0049	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0050	c0	52	96	104	60	140	400	439
	c1	52	96	104	60	140	400	439
m0051	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0052	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0053	c0	52	96	104	60	140	399	440
	c1	52	96	104	60	140	399	440
m0054	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0055	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0056	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0057	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0058	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0059	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0060	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0061	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0062	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0063	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0064	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440

Speaker No.	Chan	Ctrl Words A (a)	Ctrl Words B (b)	Ctrl Words C (c)	Isolated Digits (i)	Four Digits (d)	City Names (n)	Mono-syllables (m)
m0065	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0066	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0067	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0068	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0069	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0070	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0071	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0072	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0073	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0074	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0075	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0076	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440
m0077	c0	52	96	104	60	140	400	440
	c1	52	96	104	60	140	400	440

APPENDIX D. TAPE ARCHIVE DESCRIPTION

The corpus has been archived on a set of four tapes using the following tools:

- Sun Sparcstation 5
- Solaris 2.4
- Exabyte 8505XL 8mm tape drive
- GNU Tar 1.11.2

The tapes provided were 120m 8mm tapes. The tapes were created using the highest compression mode available. The command used to master these tapes was:

```
tar cvhf /dev/rmt/1cn -C corpus control_words
```

This is a fairly standard command with the notable exception that “1cn” instructs Solaris to use the highest compression mode for tape device no. 1, and to not rewind the drive when finished (thereby allowing multiple tar files to be written to a single tape). Each tape contains multiple tar files (except for tape no. 2). The contents to the tapes are described in Table 8.

To restore these tapes, the command for the above environment is:

```
tar xvf /dev/rmt/1cn
```

If a tape contains more than one file, this command must be run multiple times, once for each tar file. The contents of a tape can be listed using the option “t” (for table of contents) instead of “x” (for extract). Tar is a fairly common command across many platforms, particularly Unix platforms.

The tapes were written with multiple tar files to allow the data to be restored to different physical disks. We recommend building the corpus from a common mount point, and using links to reach the physical disks containing the data.