WSJCAM0 Corpus and Recording Description

Jeroen Fransen, Dave Pye, Tony Robinson,
Phil Woodland and Steve Young

Cambridge University Engineering Department (CUED) Speech Group
Trumpington Street, Cambridge CB2 1PZ, UK

September 2, 1994

1 Introduction

The speech database described in this document is the UK English equivalent of a subset of the US American English WSJ0 database [1][1]. The name of the UK English version, WSJCAM0, represents the Wall Street Journal recorded at the University of CAMbridge (phase 0). It consists of speaker-independent (SI) read material, split into training, development test and evaluation test sets. There are 90 utterances from each of 92 speakers that are designated as training material for speech recognition algorithms. A further 48 speakers each read 40 sentences utterances containing only words from a fixed 5,000 word vocabulary of 40 sentences from the 64,000 word vocabulary, which will be used as testing material. Each of the total of 140 speakers also recorded a common set of 18 adaptation sentences. Recordings were made from two microphones: a far-field desk microphone and a head-mounted close-talking microphone.

All resulting waveforms will be distributed in compressed digitised form, accompanied by orthographic transcriptions and automatically generated phone and word alignments.

2 Recorded Material

All recorded sentences were taken from the Wall Street Journal (WSJ) text corpus. Since this text corpus had previously been recorded in American English for identical purposes, we could benefit from the materials used for that effort (see Paul & Baker, 1992; also ftp: `gov.nist.ncsl.jaguar`). Therefore we could make use of existing conventions, utilities, vocabularies, and large selections of processed texts from a real newspaper. The main problems with recording British English talkers reading WSJ prompts came as a consequence of the US origin of the text. This posed an extra pronunciation problem to some speakers, which compounded the difficulties with WSJ's financial jargon and written

[1]The WSJ0 corpus and associated wordlists and language models are available from the Linguistic Data Consortium, 441 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104-6305, USA. Phone: +1 (215) 898-0464, Fax: +1 (215) 573-2175, e-mail: ldc@unagi.cis.upenn.edu

style. Therefore a modified pronunciation dictionary had to be constructed that covered UK pronunciations for some US-specific words.

The recording text material was used as follows. A common set of 18 *adaptation* utterances were recorded at the start of the session for each speaker (see Appendix A for all sentences):

- one 3-second recording of background noise
- 2 phonetically balanced sentences
- the first 15 of the 40 sentences designated for adaptation in the original WSJ0 corpus

The *training* sentences were taken from the WSJ0 training subcorpus of about 10,000 sentences. Each training speaker read about 90 training sentences, selected randomly in paragraph units. It was found that this was the maximum number of sentences that could be recorded in a one hour session. The same sentences were allowed to occur in several speakers prompts, though never more than once per speaker.

Each *test* speaker read 80 sentences from the subcorpus originally designated for development testing in WSJ0, consisting of 40 sentences from the 5,000-word corpus (which contained a total of 2,000 sentences) and 40 sentences from the 64,000-word corpus (a total of 4,000 sentences). The test sentences were randomly selected and each test sentence was allowed to occur in only one speaker's prompt material. Since no sentence repetition between or within speakers was allowed for the testing portion of the corpus, this selection procedure exhausted the 5,000-word sentences almost completely (48 speakers by 40 sentences each). All sentences were taken from the non-verbalised pronunciation texts (i.e. there was no written punctuation words in the prompt material). All numerical data were written out in words in the prompts (i.e. so-called normalised texts were used).

3 Recording Room

All recordings were made in a quasi-soundproof room in the ECR Lab of the Engineering Department. This 'quiet room' measured five metres by five metres. The room was closed off by double doors during recordings and its windows are double-glazed. Fresh air was blown in through sound-trapped ventilation.

Speakers sat on a chair in front of a desk. A fan-free workstation monitor that displayed the prompt texts was on the desk, along with the workstation keyboard and mouse. The actual workstation base-unit was placed outside the room to eliminate fan noise. The far-field desk microphone and a pre-amplifier for the close-talking microphone were placed to the left of the speaker. The close-talking microphone was attached to a pair of headphones which were worn by the speaker. The entire interactive recording process was co-ordinated by clicking on buttons on the workstation screen. More details of this procedure can be found in the *recording session* section of this report.

4 Recording Equipment

Recordings were made with two microphones. The close-talking microphone (Sennheiser HMD414-6) fed directly into a separate pre-amplifier (Symetrix SX202) and then to the analogue line input of a workstation (Silicon Graphics Iris Indigo) which, in turn, was connected to the workstation's internal A/D converter. The desk microphone's signal (Canford C100PB) first fed into a built-in pre-amplifier and then into the analogue line input. The technical specifications of the recording equipment is given in Appendix B.

5 Speakers

Speakers were recruited by placing advertisements on computer bulletin boards, in the University student newspaper '*Varsity*', in the Engineering Department's newsletters, and on notice boards in common rooms throughout the University. The advertisement asked for native speakers of UK English who would like to contribute to speech research by reading out newspaper sentences for an hour. Each speaker was offered a reward of £5 for their efforts.

The age distribution of training speakers is displayed below. The age distribution of the test speaker population is similar. A concerted effort was made to attract as wide a range of regional accents as possible. However, we were obviously limited to those speakers who lived or worked in the Cambridge area.

Age Range Distribution of Training Speakers

Range	Female	Male
18-23	21	25
24-28	11	19
29-40	3	4
> 40	4	5

5.1 Speaker Partitions for Evaluation and Development Test

The test material was recorded by 48 test speakers without making an *a priori* division of speakers into the development test and evaluation test groups. Two evaluation sets and a development set of speakers were defined, with the aim of obtaining balanced speaker groups. The procedure that placed speakers in the different test sets aimed to balance the number of male and female speakers in each set, then to balance the age distribution, and finally a coarsely calculated speaking rate measure. Therefore, each of the 3 test groups contain roughly the same proportion of speakers of the same sex and from the same age range. The speaking rate was computed for each test speaker by dividing the total number of samples per speaker by the number of words in the transcription `.dot` files. The adaptation material was not included in this computation as it was missing for one speaker. Within each age/sex group individual speakers with comparable speaking rates were then equally distributed over the test groups.

The development test speaker group comprises 18 speakers and is distributed with the training material. The two evaluation test sets contain data from 14 speakers that will be released at a later stage.

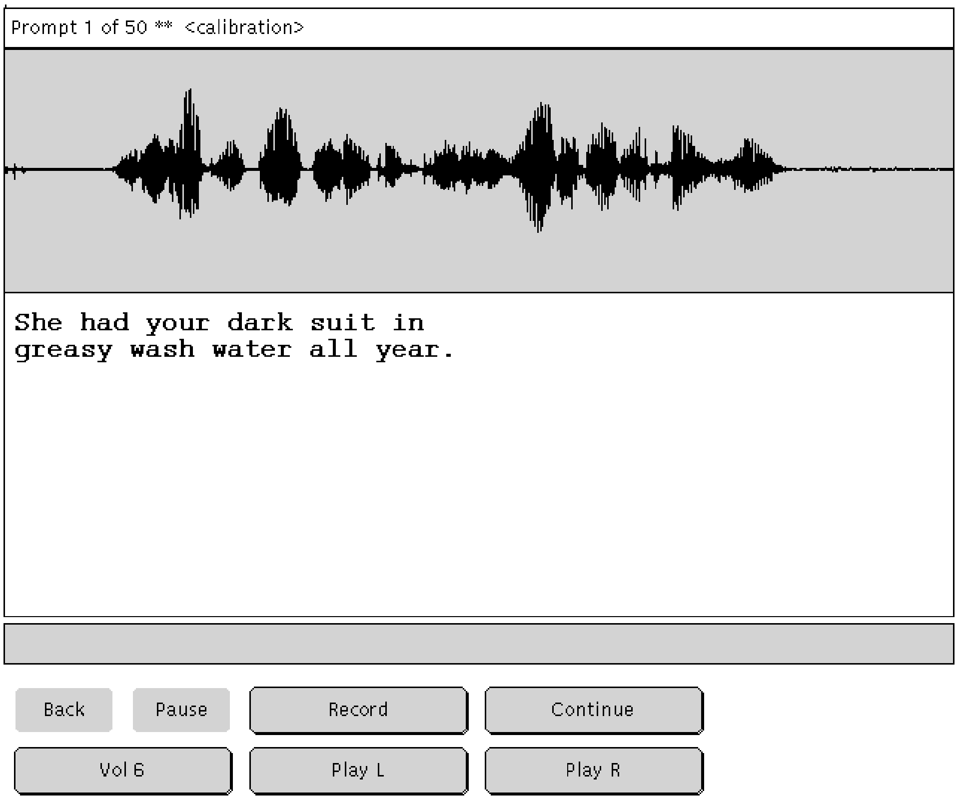

Figure 1: Computer display during a recording session

6 Recording Session

At the start of each recording session, the speaker was given some instructions to read. These instructions provided a general introduction to the project and a description of the task (reading US American English business newspaper sentences to a microphone). The speakers were asked to read each sentence through before recording it. They were also warned of the possibility of grammatical and spelling errors occurring, the financial business jargon and the written language style. The speakers were instructed to speak naturally and clearly in their usual accent at normal volume. Each speaker was given a short demonstration of how to use the recording setup.

The actual recording process for each sentence consisted of five steps, each co-ordinated by clicking on screen buttons as shown in figure 1. First, the speaker silently read the sentence. When ready, the speaker would then click on **record** to start recording and then read the sentence aloud. The recording was terminated by clicking on the **stop** button. The user could check that the recording is satisfactory by using the **play** button to listen to what was recorded. If the speech was truncated or mispronounced it could be re-recorded by again using **record**. Speakers were advised to only utilise the play back and re-record facilities when not absolutely confident of the correctness of their utterance. When a speaker had finished with the current sentence, the **continue** button moved to the next sentence.

All speakers were first given the chance to practice on 4 sentences under supervision of the recording co-ordinator. Any problems detected in the process were corrected or explained. The practice phase was also used to set the recording levels. The results of the each practice phase were discarded.

In the adaptation phase all speakers first recorded 3 seconds of background noise. They then uttered 2 'phonetically rich' adaptation sentences, followed by the first 15 adaptation sentences from the WSJ0 adaptation corpus. During this phase the recording co-ordinator was always present and, where necessary, corrected speakers.

The main phase of each session consisted of the recording of either 90 training sentences or 80 test sentences. The recording co-ordinator was no longer present in the room during this period. Afterwards, a short anonymous questionnaire was completed by the speaker—the information from which can be found in the `.ifo` files.

7 Pronunciation Dictionary

In constructing a speech recognition system, pronunciation information must be provided for all words spoken in both test and training data. A British English Example Pronunciation Dictionary (BEEP) has recently been developed for large vocabulary speech recognition with the WSJCAM0 corpus in mind. Two thirds of this dictionary stem from a combination of two sources: the MRC Psycholinguistic Database [2] and CUVOALD [3] (the Computer Usable Oxford Advanced Learners Dictionary). These databases have been publicly available for a number of years. In addition, a considerable number of new pronunciations have been provided from sources in Durham and Cambridge Universities.

The set of new symbols specific to British English were defined as follows, /oh/ for the vowel in "pot", and /ia/, /ea/ and /ua/ for the diphthongs in "peer", "pair" and "poor" respectively. The complete phone set used is shown in Appendix C of this document.

The accumulation of these diverse sources into a single standardised format produced over 100,000 word pronunciations. However, this process failed to cover 2,700 words of the 20,000 word lexicon evaluation task. The process of constructing pronunciations for these additional words proceeded in a number of stages. First, a tool was written to automatically derive word pronunciations for inflected words, by looking-up their stem pronunciations and appending that of the correct inflection from morphological rules of English [4] . This functioned for words with the following suffixes:

[-es, -s, -'s, -', -ed, -d, -er, -r, -est, -ing]

This worked by considering each member of the 20,000 word list (having a suffix in the above set) and searching for their stems from the current pronunciation directory. As a result, several hundred new words were found. The remaining words needed for the 20k evaluation task (currently around 2,000) were taken from the CMU [5] pronunciation dictionary. This source was used as a last resort, since since they are US pronunciations. However, it is hoped to reduce the number of them in subsequent releases of the pronunciation dictionary. Finally, fifty seven words were not found in the CMU dictionary, and the pronunciations for these were manually entered into the dictionary.

The pronunciation dictionary on first release contains over 96,000 word definitions. Its format has been designed with machine readability a primary factor. It is thus unashamedly simple and an extract is included below.

OUIJA-BOARDS	w iy1 jh ax b ao d z
OUIJAS	w iy1 jh ax z
OUNCE	aw n s
OUNCES	aw1 n s ih z
OUR	aw1 ax
OUR	aw1 ax r
OURS	aw1 ax z
OURSELF	aw ax s eh l f
OURSELVES	aw ax s eh l v z

The dictionary is freely available for non-commercial use by Internet FTP from host `svr-ftp.eng.cam.ac.uk` with file name `/pub/comp.speech/data/beep-0.3.tar.Z`.

8 Data Formats

The primary file format for all waveforms is NIST's SPHERE format, the DOT file format is used for transcriptions and the original prompt files are in PTX format. In addition, an information file about each speaker (`.ifo`) is part of the distribution. The format specifications as described below are taken from the document `wsj-format-spec.doc` at the NIST ftp site: `gov.nist.ncsl.jaguar`.

8.1 File Naming Formats

Data types are differentiated by unique filename extensions. All files associated with the same utterance have the same basename. All filenames are unique across all WSJCAM and ARPA-collected WSJ corpora. Utterance IDs (basenames) will not be re-used. The filename format is as follows:

```
<UTTERANCE-ID>.<XXX>
```

where,

```
UTTERANCE-ID ::= <SSS><T><EE><UU>
```

where,

SSS	::=	001 | ... | zzz (base-36 speaker ID)
T	::=	(speech type code)
		c (Common read no verbal punctuation)
		a (Adaptation read)
EE	::=	01 |...| zz (base-36 session ID;
		01 for all adaptation;
		02 for main session for training speakers and 5k for test speakers;
		03 for 20+k for test speakers)
UU	::=	01 |...| zz (base-36 within-session sequential speaker utterance code)

We were allocated the use of speaker IDs `c00-czz`. Speaker IDs `c00-c2z` were used for training speakers, speaker IDs `c30-c4z` were used for test speakers (both development and evaluation).

The file extensions are interpreted as follows:

```
XXX ::= (data type)
.wv1 (channel 1 - Sennheiser waveform)
.wv2 (channel 2 - Canford waveform)
.ptx (prompting text)
.dot (detailed orthographic transcription)
.ifo (information file about speaker)
.phn (TIMIT style phone alignments)
.wrd (TIMIT style word alignments)
```

8.2 The NIST file format: waveforms (.wv1, .wv2)

The waveforms are SPHERE-headered, digitised, zero-meaned (using NIST's bias), and compressed (using NIST's transparent shorten algorithm `w_encode`).

The filename extension for the waveforms will contain the characters, "wv", followed by a 1-character code to identify the channel. The 1024-byte NIST header for each waveform contains the following fields/types:

```
Field                   Type     Description - Probable defaults marked ()
speaker_id              string   3-char. speaker ID from filename
speaking_mode           string   speaking mode ("read-common",
                                 "read-adaptation")
recording_site          string   recording site
recording_date          string   beginning of recording date stamp of the
                                 form DD-MMM-YYYY.
recording_time -s11     string   beginning of recording time stamp of the
                                 form HH:MM:SS.HH.
recording_environment   string   text description of recording environment
microphone              string   microphone description
utterance_id            string   utterance ID from filename of the form
                                 SSSTEEUU as described in the filenames
                                 section above.
prompt_id               string   WSJ source sentence text ID - see .ptx
                                 description below for format
database_id             string   database (corpus) identifier
database_version        string   database (corpus) revision ("1.0")
channel_count           integer  number of channels in waveform ("1")
speaker_session_number  string   2-char. base-36 session ID from filename
sample_count            integer  number of samples in waveform
sample_max              integer  maximum sample value in waveform
sample_min              integer  minimum sample value in waveform
sample_rate             integer  waveform sampling rate ("16000")
sample_n_bytes          integer  number of bytes per sample ("2")
sample_byte_format      string   byte order (MSB/LSB -> "10",
   LSB/MSB -> "01")
sample_coding           string   waveform encoding
sample_checksum         integer  checksum obtained by the addition of all
                                 (uncompressed) samples into an unsigned
                                 16-bit (short) and discarding overflow.
sample_sig_bits         integer  number of significant bits in each sample
```

```
                                    ("16")
session_utterance_number string   2-char. base-36 utterance number within
                                    session from the filename
end_head                 none     end of header identifier
```

8.2.1 Example Header

```
NIST_1A
   1024
speaker_id -s3 c2e
speaking_mode -s11 read-common
recording_site -s4 CUED
recording_date -s11 09-Feb-1994
recording_time -s11 11:01:22.00
recording_environment -s14 ECR_Quiet_Room
microphone -s17 Sennheiser_HMD414
utterance_id -s8 c2ec0202
prompt_id -s22 87.034.861205-0013.1.2
database_id -s7 WSJCAM0
database_version -s3 1.0
channel_count -i 1
speaker_session_number -s2 02
sample_count -i 103500
sample_max -i 23478
sample_min -i -23889
sample_rate -i 16000
sample_n_bytes -i 2
sample_byte_format -s2 10
sample_sig_bits -i 16
session_utterance_number -s2 02
sample_coding -s26 pcm,embedded-shorten-v1.09
sample_checksum -i 34559
end_head
```

8.3 Detailed Orthographic Transcription (.dot):

The transcriptions for all utterances in a session are concatenated into a single file of the form, `<SSS><T><EE>00.dot` and include the utterance-ID codes. The format for a single utterance transcription entry in this table is as follows:

`<TRANSCRIPTION-TEXT> (<UTTERANCE-ID>) <NEW-LINE>`

An example sentence illustrating this format is given below:

> The December contract rose one point oh seven cents a pound to sixty eight point six two cents at the Chicago Mercantile Exchange (c13c020l)

There is one .dot file for each speaker-session. It should be noted that the conventions used during transcription were slightly different from those used in WSJ0 and can be found in the section on 'WSJCAM0 Detailed Orthographic Transcription (.dot) Specification'.

8.4 Prompting Text (.ptx):

The prompting texts for all read Wall Street Journal utterances in a session including the utterances' utterance-IDs and prompt IDs have been concatenated into a single file of the form, `<SSS><T><EE>00.ptx`. The format for a single prompting text entry in the .ptx file is as follows:

```
<PROMPTING-TEXT> (<UTTERANCE-ID> <PROMPT-ID>)
```

The prompt ID is the sentence index designed by Doug Paul to reference the Wall Street Journal texts in the ACL/DCI CD-ROM. The format for this index is:

```
<YEAR>.<FILE-NUMBER>.<ARTICLE-NUMBER>
<PARAGRAPH-NUMBER>.<SENTENCE-NUMBER>
```

and an example sentence:

> The December contract rose one point oh seven cents a pound to sixty-eight point six two cents at the Chicago Mercantile Exchange (c13c020l 87.120.871013-0032.14.2)

The inclusion of both the utterance ID and prompt ID allows the utterance to be traced to its source sentence text and surrounding paragraph. There is one .ptx file for each speaker-session.

8.5 Phone and Word Alignments (.phn, .wrd)

An automatic alignment procedure was used to provide alignments of utterances at both the word and phone levels. Each alignment file is named in the following format `<SSS><T><EE><UU>.<XXX>`; where the suffix XXX is either `phn` for phone level alignments or `wrd` at the word level. The alignment files adopt the standard NIST format where each line (of a word file) has the following structure:

```
<START> <END> <WORD> <NEW-LINE>
```

Example:

8960	13568	today
13568	15616	the
15616	23040	chairmen
23040	24064	are
24064	27648	net
27648	36096	losers

One `.wrd` and a corresponding `.phn` file exist for each utterance.

9 Directory Structure

The following depicts the directory structure and default partitioning proposed for the subcorpora on different discs. Subcorpora categories are denoted by the directory names in level 2. Different subcorpora will reside on different discs unless specified otherwise below. Training and development test data will probably be distributed together on one series of discs.

top level:		
	`wsjcam0/`	corpus
2nd level:		
	`README`	a brief summary of the file structure
	`doc/`	on-line documentation
	`etc/`	official training and development test sets
	`si_tr/`	SI, training, 90-100 WSJ sentences
		SI, adaptation, 1 3-second recording of background noise, 2 'rich', 15 WSJ sentences
	`si_dt/`	SI, dev. test, 40 WSJ sentences, 20K voc.
		SI, adaptation, 1 3-second recording of background noise, 2 'rich', 15 WSJ sentences
		SI, dev. test, 40 WSJ sentences, 5K voc.
	`si_et_[12]/`	SI, eval. test, 40 WSJ sentences, 20K voc.
		SI, adaptation, 1 3-second recording of background noise, 2 'rich', 15 WSJ sentences
		SI, eval. test, 40 WSJ sentences, 5K voc.
3rd level:	`<XXX>/`	(speaker-ID, where XXX = "001" to "zzz", base 36)
4th level:	`<FILES>`	(corpora files, see previous chapters for format and types)

The different files for the 5k and 20+k test speakers can be distinguished between by means of the session number in the filename (see also section on File Naming Formats): 01 for all adaptation; 02 for the main session for training speakers; 02 for 5k for test speakers; 03 for 20+k for test speakers)

10 WSJCAM0 Detailed Orthographic Transcription (`.dot`) Specification

This is a modified version of the CSR WSJ0 Detailed Orthographic Transcription Specification as it was proposed by the CCCC Transcription Subcommittee (12/12/91) which was revised 01/05/93 by John Garofolo to relax rules requiring prosodic markings and capitalisation per the CCCC conference call 11/24/92.

The current revision is by Jeroen Fransen. It includes minor adaptations that were needed to deal with the specific characteristics of the WSJCAM0 situation. These consist of additions to the types of non-speech events that occur. Also, the first adaptation sentence for each speaker will be a waveform containing background noise only. The transcription for this will be defined as a space followed by the usual utterance id.

The Detailed Orthographic Transcription (`.dot`) file will contain a case-sensitive transcription consisting of markings for an utterance's orthography, some prosodics, disfluencies and non-speech events.

10.1 Orthography

The lexical tokens in the transcription were generated without special regard to case and capitalisation. Appropriate capitalisation was used most of the time but was not strictly adhered to. Grammatical punctuation has been excluded except for periods (`.`), used specifically in abbreviations, and apostrophes (`'`). Non alpha- numeric characters such as these, which are part of a lexical item, have been prefaced by the escape character, (\).

Normal lexical items were represented as they are in the prompt text used to elicit the speech.

10.1.1 Hyphenated words

Hyphens were removed in the transcription. The vocabulary file wfl-64 was checked to see the if word occurred without the hyphen. Otherwise it was broken up into separate words. e.g.:

```
compound in wfl-64:
        Nonstop -> Non-stop
compound not in wfl-64:
        hard-headed -> hard headed
```

10.1.2 Misquotations

Where an utterance differs from its prompt text yet remains fluent and linguistically valid, the DOT transcription represents the actual spoken utterance. This utterance text may consist of deletions, transpositions, substitutions and (theoretically unlikely) insertions of words from the original prompt text. This approach increases the quantity of testing and training data at the expense of losing consistency with the newspaper text.

10.1.3 Mispronunciations

Mispronounced but intelligible words in utterances not satisfying the misquotation criteria have been delimited with a "*". If the prompt read, "He grew up in Belair." and the subject said, "He grew up in Blair." then the utterance was transcribed: "He grew up in *Belair*"

10.1.4 False Starts and Spoken Word Fragments

Incompletely spoken words were transcribed using the following notation:

- Beginning of word truncation: -(missing_fragment)spoken_fragment
- End of word truncation: spoken_fragment(missing_fragment)-

10.1.5 Prosodic Markings

Pauses

Only conspicuous pauses were marked with a single "." indicating the location of the pause.

Emphatic Stress

Emphatic stress is indicated by prepending a "!" to the word or syllable which was stressed. This only includes stress which would not normally occur due to lexical and syntactic factors.

Lengthening

Lengthening is transcribed by appending a ":" to the lengthened sound. This only includes lengthening which would not normally occur due to lexical and syntactic factors.

10.1.6 Descriptive Markings of Speech and Non-Speech Events

Non-speech Events

Non-speech events are indicated by a descriptor enclosed in square brackets "[]". The descriptor contains only alphabetic characters and underscores and was drawn from the following list:

`AH`	`CHAIR_SQUEAK`
`COUGH`	`CROSS_TALK`
`DOOR_SLAM`	`ER`
`GRUNT`	`LAUGHTER`
`LIP_SMACK`	`LOUD_BREATH`
`MM`	`PAPER_RUSTLE`
`PHONE_RING`	`SIGH`
`THROAT_CLEAR`	`TONGUE_CLICK`
`UH`	`UNINTELLIGIBLE`
`UM`	
`MOUSE_CLICK` (specific to WSJCAM)	
`MIKE_OVERLOAD` (specific to WSJCAM)	

An example file:

The doctor said [throat_clear] open wide

10.1.7 Descriptor Placement and Concurrent Events

A descriptor was placed in the orthography at the point at which the non-speech event occurs. If a non-speech event overlaps with a spoken lexical item, the descriptor was placed next to the lexical item it co-occurred with and the character, ">" or "<" was appended or prepended to the descriptor depending on whether it is placed to the left or right of the co-occurring lexical item:

The escaped convict [< door_slam] ran for his life

and

The escaped [door_slam >] convict ran for his life

are roughly equivalent.

If a phenomenon was noted throughout, or co-occurred with, more than one lexical item, then the phenomenon's descriptor is used in the following notation to bound the lexical items it spans:

`[DESCRIPTOR/] WORD WORD ... WORD [/DESCRIPTOR]`

Example: [cross_talk/] The plane narrowly escaped disaster [/cross_talk] as it took off

10.1.8 Bad Recording

If the recording quality of an utterance was so bad that it defies transcription, then the flag, "[bad_recording]", was substituted for the transcription in the .dot file and the utterance will be viewed as unusable.

[bad_recording] (c05c021b)

10.1.9 Waveform Truncation

If a waveform file was truncated due to a recording error by the system or by the failure of the speaker to press the record button at the proper times, the following notation in the corresponding transcription file was used:

- Beginning of utterance truncation:

  `~ TRANSCRIPTION`

- End of utterance truncation:

  `TRANSCRIPTION ~`

- Beginning and end of utterance truncation:

  `~ TRANSCRIPTION ~`

- Null waveform

  `~~`

In the final corpus, null waveforms have been discarded.

10.1.10 Utterance Identification

The 8-character utterance ID from the filename (minus extension) was placed at the end of each transcription string in parentheses immediately followed by a new-line character. The parenthesised utterance ID was separated from the transcription string by one space character.

`TEXT TEXT TEXT (UTTERANCE-ID) <NEW-LINE>`

Example:

Los Angeles based Government Funding is used to picking up where banks leave off (c04c0202)

References

[1] D B. Paul & J. M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proc. Fifth DARPA Speech and Natural Language Workshop*, pages 357–362. DARPA, Morgan Kaufmann Publishers, Inc., 1992.

[2] M. Wilson. M.R.C. Psycholinguistic Database: Machine Usable Dictionary, Version 2. Informatics Division, SERC, Rutherford Appleton Laboratory. Available by Internet as `ftp://black.ox.ac.uk/pub/ota/dicts/1054/mrc2.doc`.

[3] R. Mitton. A Description of a Computer-Usable Dictionary File Based on the Oxford Advanced Learners Dictionary of Current English. Birkbeck College, University of London. Available by Internet as `ftp://black.ox.ac.uk/pub/ota/dicts/710/710.doc`.

[4] A.P. Cowie, Editor. *Oxford Advanced Learners Dictionary of Current English.* Oxford University Press, 1989.

[5] B. Weide. The Carnegie Mellon Pronouncing Dictionary. Department of Computer Science, Carnegie Mellon University. Available by Internet as `ftp://ftp.cs.cmu.edu/project/fgdata/dict`.

Acknowledgements

This work was funded by LRE Project 62-058 '*Speech recogniser Quality Assessment for Linguistic Engineering*' (SQALE) and by the Linguistic Data Consortium. Additionally, Tony Robinson was supported by an EPSRC Advanced Research Fellowship.

Many thanks are due to the people who helped us collect the WSJCAM0 Speech Database. Firstly, to Doug Paul, without whose text utilities, scripts (ftp site gov.nist. ncsl.jaguar) and his order in the WSJ data we wouldn't have been able to record this database. Secondly to Jonathan Foote, whose data collection software we used and for setting up the recording equipment. Thirdly, we owe a lot to all at NIST for providing the SPHERE utilities and the WSJ0 texts. Finally, a big thank you goes to all the people in the SVR lab, who were patient throughout this resource intensive project.

Appendix A: Adaptation Sentences for Speaker c3j

[Recording of Background Noise] (c3ja0101 calibration.01)
She had your dark suit in greasy wash water all year. (c3ja0102 calibration.02)
Don't ask me to carry an oily rag like that. (c3ja0103 calibration.03)
The female produces a litter of two to four young in November and December. (c3ja0104 adapt.01)
Numerous works of art are based on the story of the sacrifice of Isaac. (c3ja0105 adapt.02)
Their solution requires development of the human capacity for social interest. (c3ja0106 adapt.03)
His most significant scientific publications were studies of birds and animals. (c3ja0107 adapt.04)
In recent years she has primarily appeared in television films such as Little Gloria. (c3ja0108 adapt.05)
The process by which the lens focuses on external objects is called accommodation. (c3ja0109 adapt.06)
Two narrow gauge railroads from China enter the city from the northeast and northwest. (c3ja010a adapt.07)
Some maps use bands of color to indicate different intervals of value. (c3ja010b adapt.08)
Origins or causes of spontaneous mutation are not yet completely clear. (c3ja010c adapt.09)
Unusually high levels of radiation were detected in many European countries. (c3ja010d adapt.10)
Both petroleum and natural gas deposits are scattered through eastern Ohio. (c3ja010e adapt.11)
For the first time in years the Republicans also captured both houses of Congress. (c3ja010f adapt.12)
The South Carolina educational radio network has won national broadcasting awards. (c3ja010g adapt.13)
A tanker is a ship designed to carry large volumes of oil or other liquid cargo. (c3ja010h adapt.14)
The enormous amounts of carbon dioxide in the atmosphere cause this high pressure. (c3ja010i adapt.15)

Appendix B: Technical Specifications of Recording Equipment

The Far-field Desk Microphone: Canford C100PB Condenser Gooseneck

Output Impedance:	2k$\Omega \pm$ 20% @ 1kHz
Sensitivity:	-64dBV $\pm$ 3dB (0dB -1V/μ bar @ 1 kHz)
Polar Response Cardioid:	Front to back rejection 10dB approx. (1kHz)

The signal from the Canford microphone fed into its internal pre-amp, which fed into the analogue line input. Average measured desk microphone SNR in speech after digitisation (SNR computed using the NIST SPHERE utilities 'speech' and 'segsnr') was: 20-25dB.

The Head-mounted Close-talking Microphone: Sennheiser HMD 414-6

Frequency Response:	50Hz-12kHz
Mode of Operation:	Pressure gradient transducer for close talking
Directional Characteristic:	super-cardioid
Rejection at 120 and 1000Hz:	20dB-2dB
Impedance at 1000Hz:	200Ω
Sensitivity:	1μV/50mG $\approx$ 1 μV/5 μ T

The Head-mounted Microphone Pre-amplifier: Symetrix SX202 Dual Mic Preamp

Frequency Response:	20Hz-20kHz, +0dB, -1dB
SNR:	95dB (-50dBV, 150Ω)
Max. Gain/Min Gain:	60/20dB

The signal from the Sennheiser fed into the Symetrix and then into the analogue input. Average measured close-talking microphone SNR in speech after digitisation (SNR computed using the NIST SPHERE utilities 'speech' and 'segsnr') was: 35-45dB.

Silicon Graphics IRIS Indigo's Stereo Line-Level Analogue Input

Nominal Input Impedance:	5kΩ
Input Signal:	
Max. Amplitude:	10Vpp
Minimum Level:	1Vpp (for full-scale input)

Silicon Graphics IRIS Indigo's A/D Converter

Resolution:	Stereo 16-bit
Modulation:	delta-sigma
Sampling Rate Used:	16kHz
Over sampling:	64x
Official SNR at 48kHz:	>80dB (20Hz-20kHz)

Recordings were made directly onto the IRIS Indigo's hard disk. This allowed the use of an easy interactive recording process for the speakers, as well as immediate identification tagging for further use.

Appendix C: The Phone Set

In this appendix, the extended version of the ARPAbet phone set used in the phonetic dictionary and in the phone level transcriptions is featured. The corresponding symbols used in schemes for British English have also been incorporated to allow comparison.

ARPAbet	MRPA	Edin.	Alvey	Example	Relative frequency
p	p	p	p	put	3.1%
b	b	b	b	but	2.3%
t	t	t	t	ten	6.8%
d	d	d	d	den	4.1%
k	k	k	k	can	4.7%
m	m	m	m	man	3.1%
n	n	n	n	not	6.5%
l	l	l	l	like	5.5%
r	r	r	r	run	5.4%
f	f	f	f	full	1.8%
v	v	v	v	very	1.2%
s	s	s	s	some	6.6%
z	z	z	z	zeal	3.6%
hh	h	h	h	hat	0.8%
w	w	w	w	went	0.9%
g	g	g	g	game	1.3%
ch	ch	ch	tS	chain	0.5%
jh	jh	j	dZ	Jane	0.8%
ng	ng	ng	9	long	1.6%
th	th	th	T	thin	0.3%
dh	dh	dh	D	then	12.2%
sh	sh	sh	S	ship	1.2%
zh	zh	zh	Z	measure	0.1%
y	y	y	j	yes	0.8%
iy	ii	ee	i	bean	1.4%
aa	aa	ar	A	barn	0.9%
ao	oo	aw	O	born	1.0%
uw	uu	uu	u	boon	1.0%
er	@@	er	3	burn	0.7%
ih	i	i	I	pit	10.0%
eh	e	e	e	pet	2.4%
ae	a	aa	&	pat	2.5%
ah	uh	u	V	putt	1.5%
oh	o	o	0	pot	1.6%
uh	u	oo	U	good	0.4%
ax	@	a	@	about	7.2%
ey	ei	ai	eI	bay	2.0%
ay	ai	ie	aI	buy	1.6%
oy	oi	oi	oI	boy	0.2%
ow	ou	oa	@U	no	1.5%
aw	au	ou	aU	now	0.4%
ia	i@	eer	I@	peer	0.7%
ea	e@	air	e@	pair	0.2%
ua	u@	oor	U@	poor	0.2%

See also the files available from the Oxford Text Archive by FTP from black.ox.ac.uk in directories ota/dicts/710 and ota/dicts/1054.