

## SPEECH RECORDINGS FROM CHILE

|            |   |
|------------|---|
| Author(s): | Asunción Moreno                                 |
| Institute: | Universitat Politècnica de Catalunya            |
| Address:   | Jordi Girona 1-3 Mod D-5, 08034 Barcelona Spain |
| email:     | asuncion @gps.tsc.upc.es                        |
| Date:      | July, 15, 1998                                  |
| Version:   | V2  |

# CONTENTS

|  |           |
|--|-----------|
| <b>1. INTRODUCTION</b>                   | <b>3</b>  |
| 1.1 Speech file format                   | 4         |
| 1.2 File nomenclature                    | 4         |
| 1.3 Directory structure                  | 5         |
| 1.4 Label files                          | 6         |
| <b>2. DATABASE DESIGN AND COLLECTION</b> | <b>10</b> |
| 2.1 Recording site and platform          | 10        |
| 2.2 Speaker recruitment                  | 10        |
| 2.3 Transcription                        | 11        |
| <b>3. DATABASE CONTENTS DEFINITION</b>   | <b>12</b> |
| <b>4. BIBLIOGRAPHY</b>                   | <b>14</b> |

## 1. Introduction

This database contains speech collected from Chile. Collection was performed at the Department of Signal Theory and Communications of the Universitat Politècnica de Catalunya (UPC) (Spain).

This database comprises telephone recordings from 507 speakers recorded directly over the fixed PSTN using an E-1 interface.

The database is distributed in a ISO 9660 CD-ROM volume. The list of the directories are contained in the README.TXT file, stored on the CD-ROM. Further details regarding the database contents, files and directories are provided in the documentation files in the DOC directory, the speaker and lexicon tables in the TABLE directory and contents files in the INDEX directory.

File types are identified with the following extensions:

- \*.DOC - Microsoft Word V7.0document
- \*.LST - DOS text index file with ISO Latin 1 symbols
- \*.TBL - DOS text file with ISO Latin 1 symbols
- \*.TXT - DOS text file
- \*.ESO - SAM label file, text file with ISO Latin 1 symbols
- \*.ESA - A-law speech signal file
- \*.PS - Postscript file
- \*.EXE - Executable program

The CD-ROM has the following directory structure:

```
\:
README.TXT - describes files in the CD-ROM
CHILE7ES/ - data directory
  CHILE7ES \DOC:
    DESIGN.DOC - Chile database documentation file
    ISO88591.PS - ISO 8859_1 table
  CHILE7ES \INDEX:
    CONTENTS.LST - file/utterance/speaker index table
  CHILE7ES \TABLE:
    SPEAKER.TBL - speaker table
  CHILE7ES \OTHERS
    CONTEN00.DOC - contents table in MS-Word'97 format for the first
                  100 sessions of the database. It contains macros that
                  allow to hear the utterances
    CONTEN01.DOC - Idem for session 0100 to 0199
    ...
```

PLAYA.EXE - program used by the macros of CONTEN<sub>xx</sub>.DOC to play the signals. It runs on Windows NT and Windows 95

CHILE7ES\:  
 BLOCK00\  
 BLOCK01\  
 ...  
 CHILE7ES\BLOCK00:  
 SES0002\ - session directories for each telephone call  
 ...  
 CHILE7ES\BLOCK00\SES0002:  
 A00002A1.ESO - SAM label file  
 A00002A1.ESA - speech signal file  
 ...

This database was designed following the specifications given by Siemens. Speech file format and SAM label files follows the specifications given by the SpeechDat project. Most important points are copied or summarized in this document. A detailed information can be found in the Bibliography.

### 1.1 Speech file format

Speech files are stored as sequences of 8-bit 8 Khz A-law uncompressed speech samples (CCITT G.711 recommendation). Each prompted utterance is stored within a separate file. Each speech file has accompanying ASCII SAM label file

### 1.2 File nomenclature

File names follow the ISO 9660 file name conventions (8+plus+3 characters) according to the main CD-ROM standard. The following template is used:

DD NNNN CC . LL F

where:

|      |   |
|------|---|
| DD   | Database identification code (00-ZZ)<br>A1=fixed network recordings; B1=mobile network recordings; C1=speaker verification database |
| NNNN | Recording session progressive number (0000-9999)  |
| CC   | Corpus code (A0-Z9)   |
| LL   | Two letter ISO 639 language code  |
| F    | File type code<br>O=Orthographic label file, A=A-law coded speech file  |

**Table 1 - Filename convention**

The following table shows the Corpus Code and Contens of the database. The last column contains the English word provided by Siemens. The Spanish translation is adapted to the meaning of the English word for the intended application.

| Corpus Code | Contens   | English word |
|-------------|-----------|--------------|
| A0          | borrar    | Delete       |
| A1          | aceptar   | Enter        |
| A2          | enviar    | Forward      |
| A3          | ayuda     | Help         |
| A4          | escuchar  | Listen       |
| A5          | parar     | Stop         |
| A6          | grabar    | Record       |
| A7          | número    | Pound (#)    |
| A8          | salvar    | Save         |
| A9          | saltar    | Skip         |
| B0          | asterisco | Star (*)     |
| B1          | entrar    | Enter        |
| Q0          | no        | No           |
| Q1          | si        | Yes          |
| I0          | cero      | Zero         |
| I1          | uno       | One          |
| I2          | dos       | Two          |
| I3          | tres      | Three        |
| I4          | cuatro    | Four         |
| I5          | cinco     | Five         |
| I6          | seis      | Six          |
| I7          | siete     | Seven        |
| I8          | ocho      | Eight        |
| I9          | nueve     | Nine         |

**Table 2 Contents and corpus codes of the Chile Spanish Database**

### 1.3 Directory structure

The directory structure uses a shallow directory nesting with contiguous numbers to identify the individual sub-directories and call directories. The following three-levels directory structure is defined:

\`<database>`\\`<block>`\\`<session>`

Where:

|                               |  |
|-------------------------------|--|
| <code>&lt;database&gt;</code> | Defined as: <code>&lt;name&gt;</code> <code>&lt;#&gt;</code> <code>&lt;language code&gt;</code> i.e.CHILE7ES |
|-------------------------------|--|

|            |   |
|------------|---|
|            | Where:<br><name> is CHILE<br><#> is 7 for Siemens<br>< language code > is the ISO 2-letters code: ES for Spanish  |
| <block >   | Defined as: BLOCK<nn><br>where <nn> is a progressive number from 00 to max. 99 These numbers are the same as the first 2 digits used in <nnnn> described below.   |
| <session > | Defined as: SES<nnnn><br>Where <nnnn> is a progressive number in the range 0000 to max. 9999, being the numeric call identification number also encoded in each filename. As there are no more than 50 utterances per call, the total number of speech files and associated transcription files does not exceed the CD-ROM recommended limit of approximately 100 files in a directory. |

**Table 3 - Directory structure**

Both signal files and label files are put in the same directory. All sessions have complete recordings for all items. In addition to the previous structure the following directories are used to store some other files:

|                    |                             |
|--------------------|-----------------------------|
| \\<database>\DOC   | documentation files         |
| \\<database>\INDEX | index files - contents file |

**Table 4 - Non-speech related directory structure**

Finally the root directory contains a 'README.TXT' ASCII file describing all files in the CD-ROM; signal and label files are reported by specifying their templates;

#### 1.4 Label files

Associated with each speech file is a ASCII SAM label file. Associated SAM label files are text files where each row can be up to 80 chars long and <CR><LF> ended (according to the DOS format). Rows are produced according to the main SAM paradigm:

ABC: x, y, z, ...

Where:

- ABC is a three letter mnemonic followed by a colon: no spaces are allowed between them, so we can define as SAM-mnemonic the set "ABC:";
- after the mnemonic are all the defined items separated by commas;
- missing items are accepted and nothing needs to be put between commas to substitute them;
- spaces are not significant.

A label file begins with the mnemonic "LHD: " and end with "ELF: ". The mnemonic "LBD:" splits a label file into two sections: the LABEL FILE HEADER and the LABEL FILE BODY.

All mnemonics used in the Chile Spanish Database are listed below. For each one the explanation, the format and the domain accepted for the related items are given in the SpeechDat documentation. In this document, only the deviations from SpeechDat are considered

## Label file header

### Identification rows

| mnem. | item format   | example               | comments              |
|-------|---------------|-----------------------|-----------------------|
| LHD:  | "%s, %d.%02d" | SAM, 5.10             | format name + version |
| ELF:  | ""            |                       | end of label file     |
| CMT:  | "%.75s"       | This is a comment row | comment row           |

### Session rows

| Mnem. | Item format | example                         | comments           |
|-------|-------------|---------------------------------|--------------------|
| DBN:  | "%.75s"     | Chile_Spanish_Fixed_Netw<br>ork | database name      |
| VOL:  | "%.11s"     | CHILE7ES_01                     | database volume ID |
| SES:  | "%04d"      | 1000                            | session number     |

### File rows

| mnem. | item format         | example                       | comments                                 |
|-------|---------------------|-------------------------------|--|
| DIR:  | "\\%.8s\\...\\%.8s" | \CHILE7ES\<br>BLOCK10\SES1000 | signal file directory                    |
| SRC:  | "%8s.%3s"           | A11000A1.ESA                  | signal file name                         |
| CCD:  | "%.2s"              | A1                            | corpus code                              |
| CRP:  | "%.02d"             |                               | corpus repetition                        |
| REP:  | "%s, %s, %s"        | UPC,<br>BARCELONA,<br>SPAIN   | recording place: place, city,<br>country |
| RED:  | "%02d/%.3s/%4d"     | 16/Oct/1997                   | recording date                           |
| RET:  | "%02d:%02d:%02d"    | 20:20:48                      | recording time                           |
| BEG:  | "%lu"               | 0                             | labeled sequence start position          |
| END:  | "%lu"               | file length - 1               | labeled sequence end position            |

### Data file coding rows

| <b>mnem.</b> | <b>item format</b> | <b>example</b> | <b>comments</b>  |
|--------------|--------------------|----------------|--|
| SAM:         | "%d"               | 8000           | sampling frequency   |
| SNB:         | "%d"               | 1              | number of (8-bit) bytes per sample                                 |
| SBF:         | "%2s"              |                | sample byte order (meaningless with single byte samples, "SNB: 1") |
| SSB:         | "%d"               | 8              | number of significant bits per sample                              |
| QNT:         | "%.75s"            | A-LAW          | Quantisation   |

### Speaker rows

"ACC:" speaker accent, i.e. the regional/dialectical colouring factor;(void)

| <b>mnem.</b> | <b>item format</b> | <b>example</b> | <b>comments</b> |
|--------------|--------------------|----------------|-----------------|
| SCD:         | "%06d"             | 172000         | speaker code    |
| SEX:         | "%c"               | M              | speaker sex     |
| AGE:         | "%d"               | 35             | speaker age     |
| ACC:         | "%.75s"            |                | speaker accent  |

### Recording condition rows

"REG:" calling region,(void)

| <b>Mnem.</b> | <b>Item format</b> | <b>example</b> | <b>comments</b>     |
|--------------|--------------------|----------------|---------------------|
| REG:         | "%.75s"            |                | calling region      |
| ENV:         | "%.75s"            | HOME/OFFICE    | calling environment |
| NET:         | "%.75s"            | PSTN           | telephone network   |
| PHM          | " "                |                | hand-set type used  |

### Information about the labeling session

| <b>mnem.</b> | <b>item format</b>                | <b>example</b>        | <b>comments</b>                |
|--------------|-----------------------------------|-----------------------|--------------------------------|
| EXP:         | "%s"                              | Sergio Oller          | labeling expert                |
| SYS:         | "%s, %d.%02d"                     | UPC_RevBD,1.00        | labeling system                |
| DAT:         | "%02d/%.3s/%4d, %02d: %02d: %02d" | 18/Oct/1997, 18:52:47 | date of completion of labeling |

### Label file body

| <b>mnem.</b> | <b>item format</b> | <b>example</b> | <b>comments</b> |
|--------------|--------------------|----------------|-----------------|
|--------------|--------------------|----------------|-----------------|

|      |                                   |  |  |
|------|-----------------------------------|--|--|
| LBD: | ""                                |  | label body keyword   |
| LBR: | "%lu, %lu, %d, %d,<br>%d, \"%s\"" | 0, 12457, , , , "This is<br>what the speaker should<br>have uttered" | labeling during recording:<br>begin, end, gain, min,<br>max, orthographic text<br>prompt |
| EXT: | "%.75s"                           |  | line extension   |
| LBO: | "%lu, %lu, %lu, %s"               | 2800, 4700, 6600, This<br>is what the speaker<br>actually said       | orthographic labeling:<br>begin, center, end,<br>orthographic<br>transcription text      |

Below is an example of a label file contained in the Database.

LHD: SAM, 5.10  
DBN: Chile\_Spanish\_Fixed\_Network  
VOL: CHILE7ES\_01  
SES: 1000  
CMT: \*\*\* Speech file information \*\*\*  
DIR: \CHILE7ES\BLOCK10\SES1000  
SRC: A11000A1.ESA  
CCD: A1  
BEG: 0  
END: 19987  
REP: UPC, BARCELONA, SPAIN  
RED: 16/Oct/1997  
RET: 20:20:48  
CMT: \*\*\* Speech data coding \*\*\*  
SAM: 8000  
SNB: 1  
SBF:  
SSB: 8  
QNT: A-LAW  
CMT: \*\*\* Speaker information \*\*\*  
SCD: 787900  
SEX: F  
AGE: 21  
ACC:  
CMT: \*\*\* Recording information \*\*\*  
REG:  
ENV: HOME/OFFICE  
NET: PSTN  
PHM:  
CMT: \*\*\* Labeling information \*\*\*  
EXP: Sergio Oller  
SYS: UPC\_RevBD, 1.00  
DAT: 18/Oct/1997, 18:52:47

LBD:  
CMT: \*\*\* Label file body \*\*\*  
LBR: 0, 19987, , -1632, 1056, "saltar"  
LBO: 2800, 4700, 6600, [sta] [int] saltar  
ELF:

## **2. Database design and collection**

### **2.1 Recording site and platform**

Recordings took place at UPC. Two recording platforms were used simultaneously. The main characteristics of each recording platform are:

|                        |  |
|------------------------|--|
| Interface:             | ISDN basic access (BRI)  |
| Board:                 | AVM-ISDN-A1.   |
| Computer:              | Pentium PC at 120 MHz, 32 MB RAM 4 GBytes SCSI Hard disk. PCI Network card |
| DOS:                   | Windows 95.  |
| Programming Interface: | COMMON-ISDN-API Version 2.0 (CAPI 2.0)                                     |
| Software:              | Application Software written in C (UPC ADA program)                        |
| Lines:                 | 2  |

#### **UPC ADA Call Server**

The software is based on the ISDN development software CAPI 2.0 to perform the recordings under Windows 95 or Windows NT.

The recording software includes a voice/silence detector. For each sentence to be recorded we can specify the minimum initial silence, maximum initial silence, final silence and maximum recording time. The terminating condition can then be used to request a repetition of the recording to meet the specifications or stop the call.

The ADA programmed can perform simultaneous recordings from the 2 lines of a ISDN-BRI interface and can be extended to support more lines from different BRI or a PRI.

### **2.2 Speaker recruitment**

The speakers from the database were mainly be recruited from students and their relatives from five Universities. These Universities are located around the country: Arica, Valparaiso Concepcion and Santiago de Chile. This method has access to a large quantity of people of several dialectal areas, sex and ages. In each University, a person preferably with a linguistic or speech processing background is in charge of recruit a previously determined number of speakers. Speakers are selected from students and their relatives to have a distribution in sex and age.

The following sex and age distribution has been obtained

| Age groups | Number of speakers |      |       | Percentage of total |
|------------|--------------------|------|-------|---------------------|
|            | Male               | Fem. | Total |                     |
| under 16   | 10                 | 23   | 33    | 6.5                 |
| 16-30      | 116                | 99   | 215   | 42.4                |
| 31-45      | 89                 | 118  | 207   | 40.8                |
| 46-60      | 20                 | 31   | 51    | 10                  |
| over 60    | 0                  | 1    | 1     | 0                   |
| Total      | 235                | 272  | 507   |                     |

### 2.3 Transcription

The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. SpeechDat conventions were used in this database. The extra marks contained in the transcription aid in interpreting the text form of the utterance. Transcriptions were made in two passes: one pass in which words are transcribed, and a second pass in which the additional details are added.

Transcriptions are CASE INSENSITIVE.

The character set to be used for the transcriptions is ISO-8859. The table used is printed onto the CDs in postscript. The directory used is CHILE7ES\DOC; the filename is ISO88591.PS.

Mispronounced words that are nevertheless intelligible will be marked with one star \* attached to the left of the word which is mispronounced e.g. \*transportation instead of the mispronounced transportetation

Words preceded by a star include mispronunciations such as words with extra or omitted syllables, but a star is not used to indicate pronunciations of words that represent normal dialectal or stylistic variations. In stretches of speech that are mispronounced, each mispronounced word is marked individually.

Words or stretches of speech that are completely unintelligible are denoted by a sequence of two asterisks: "\*\*". The "\*\*" marker is separated from neighboring words with spaces.

Word fragments, i.e. instances in which the speaker did not complete a word, are considered a mispronunciation. It is accordingly marked with a star attached to the left of the intended word. Usually, the full word appears, not a text fragment, as this can complicate the lexicon and create confusion if fragments are textually the same as valid words.

Truncations: If a speech signal file is truncated due to a recording error, the following notation is used:

|  |                 |
|--|-----------------|
| Beginning of utterance truncation:         | ~transcription  |
| End of utterance truncation:               | transcription~  |
| Beginning and end of utterance truncation: | ~transcription~ |

There is a difference between an utterance which is truncated and is now incomplete, but which has not damaged the initial or final words, and an utterance when word(s) have been damaged. The ~ indicates truncation of the word it is attached to. Otherwise truncated but good utterances will not be marked in any way. As with word fragments the full word should appear in the transcription, not a text fragment.

Non-Speech Acoustic Events have been arranged into 4 categories and transcribed. Events only are transcribed if they are clearly distinguishable. Very low-level, non-intrusive events are ignored. The event will be transcribed at the place of occurrence, using the defined symbols in square brackets. For noise events that occur over a span of one or more words, the transcription indicate the beginning of the noise, just before the first word it affects.

The first two categories of acoustic events originate from the speaker, and the other two categories originate from another source. Sounds originating from the speaker usually do not overlap with the target speech, sounds originating from other sources can of course occur simultaneously with the speech.

The 4 categories are:

[fil]: Filled pause. These sounds can well be modeled in a filled pause model in speech recognisers. Examples of filled pauses: uh, um, er, ah, mm.

[spk]: Speaker noise. All kinds of sounds and noises made by the calling speaker that are not

Next table shows the list of words and the definition provided by Siemens.

|         |   |
|---------|---|
| delete  | delete an entry, message or list item                             |
| enter   | enter or add a new entry, message or list item                    |
| forward | forward (send) a call or a message to another subscriber          |
| help    | request information or menu options for the current dialogue node |
| listen  | listen to a message...  |
| no      | no  |
| pound   | same as hash; telephone keyboard symbol                           |
| record  | record a message or a voicemail greeting or an information file   |
| save    | save/archive current entry, message or list item                  |
| skip    | skip or omit a message, item or process                           |
| star    | star of some telephone keyboards                                  |
| stop    | stop current function, or proceed with next item or process       |
| yes     | yes   |
| oh      | 0   |
| zero    | 0   |
| one     | 1   |
| two     | 2   |
| three   | 3   |
| four    | 4   |
| five    | 5   |
| six     | 6   |
| seven   | 7   |
| eight   | 8   |
| nine    | 9   |

**Table 5 – Table of words and meaning provided by Siemens**

Next table shows the English word, the common Spanish translation in Spain and other alternatives with the same or similar meaning.

|           |                  |                |               |
|-----------|------------------|----------------|---------------|
| delete    | <b>borrar</b>    | suprimir       | eliminar      |
| enter     | <b>entrar</b>    | <b>aceptar</b> |               |
| forward   | <b>enviar</b>    |                |               |
| help      | <b>ayudar</b>    | help           |               |
| listen    | <b>escuchar</b>  | oir            | reproducir    |
| no        | <b>no</b>        |                |               |
| pound (#) | numeral          | almohadilla    | <b>número</b> |
| record    | <b>grabar</b>    | archivar       |               |
| save      | guardar          | <b>salvar</b>  |               |
| skip      | <b>saltar</b>    | siguiente      | omitir        |
| star (*)  | <b>asterisco</b> | estrella       |               |

|       |               |         |      |
|-------|---------------|---------|------|
| stop  | <b>parar</b>  | detener | stop |
| yes   | <b>si</b>     |         |      |
| oh    | <b>ceró</b>   |         |      |
| one   | <b>uno</b>    |         |      |
| two   | <b>dos</b>    |         |      |
| three | <b>tres</b>   |         |      |
| four  | <b>cuatro</b> |         |      |
| five  | <b>cinco</b>  |         |      |
| six   | <b>seis</b>   |         |      |
| seven | <b>siete</b>  |         |      |
| eight | <b>ocho</b>   |         |      |
| nine  | <b>nueve</b>  |         |      |

**Table 6 – English words and Spanish translations**

In order to find a set of words with the same meaning in a large number of countries, Tables 5 and 6 were distributed to people from eight different south american countries. Each person was asked to read the description of actions and mark the common word used in his country to describe the action. If the first column of table 6 describes the action, the first column must be chosen in order to maintain a common set.

The result is a set of Spanish words that describes the actions and digits in a large number of countries. The set is marked bold in Table 6. Word 'enter' was translated in all the countries as 'entrar' except in Mexico where it has a different meaning ('entrar' can be translated in Mexico as 'to call'). For this reason, two words were chosen 'aceptar' and 'entrar'.

#### **4. Bibliography**

- [1] Richard Winsky. "Definition of Corpus, scripts and standards for Fixed Networks", SpeechDat project, doc ref LE2-4001-SD1.1.3, 22 January 1997.
- [2] Franco Senia et al. "Environmental and speaker specific coverage for Fixed Networks", SpeechDat project, doc ref LE2-4001-SD1.2.1, 26 February 1997.
- [3] Franco Senia . "Specification of speech database interchange format", SpeechDat project, doc ref LE2-4001-SD1.3.1, 28 February 1997.
- [4] Henk Van den Heuvel, "Validation criteria", SpeechDat project, doc ref LE2-4001-SD1.3.3, 11 March 1997.
- [5] Franco Senia and J.G. Van Velden "Specification of orthographic transcription and lexicon conventions". SpeechDat project, doc ref LE2-4001-SD1.3.2, 1996.