

## Speech recordings from Colombia

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:	November, 20, 1998
Version:	V3

# CONTENTS

<b>1. Introduction</b>	<b>3</b>
<b>1.1 Speech file format</b>	<b>4</b>
<b>1.2 File nomenclature</b>	<b>4</b>
<b>1.3 Directory structure</b>	<b>5</b>
<b>1.4 Label files</b>	<b>6</b>
<b>2. Database design and collection</b>	<b>10</b>
<b>2.1 Recording site and platform</b>	<b>10</b>
<b>2.2 Speaker recruitment</b>	<b>10</b>
<b>2.3 Transcription</b>	<b>11</b>
<b>3. Database contents definition</b>	<b>13</b>
<b>Spontaneous answers</b>	<b>13</b>
<b>4. Bibliography</b>	<b>13</b>

## 1. Introduction

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

This database comprises telephone recordings from 1065 speakers recorded directly over the fixed telephone network using an E-1 interface.

The database is distributed in a ISO 9660 CD-ROM volume. The list of the directories is 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 V7.0 document
- \*.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  
COLOM7ES/  
    COLOM7ES \DOC:  
        DESIGN.DOC - Colombia database documentation file  
        ISO88591.PS - ISO 8859_1 table  
    COLOM7ES \INDEX:  
        CONTENTS.LST - file/utterance/speaker index table  
    COLOM7ES \TABLE:  
        SPEAKER.TBL - speaker table  
    COLOM7ES \OTHERS  
        CONTEN00.DOC - contents table in MS-Word'97 format for the first  
                    100 sessions of the database. It contains macros that  
                    allow hearing the utterances  
        CONTEN01.DOC - Idem for session 0100 to 0199  
    ...
```

PLAYA.EXE

- program used by the macros of CONTENxx.DOC to play the signals. It runs on Windows NT and Windows 95

COLOM7ES\:

- contains the data block directories

BLOCK00\  
BLOCK01\  
...

526.25 0 (526.2TD SES0002\)-15.75 Tw (direcES\:) T3346.25 0 3346.22801 sessi sins the data f  
C:\SPECH\DATA\CONTE\Nxx.DOC 3577625 T607a0c283 Tejt the IM sin important files are Wip

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

Corpus Code	Contents	English word
A0	borra	Delete
A1	español	Spanish
A2	fin	End
A3	ayuda	Help
A4	número	Pound (#)
A5	asterisco	Star (*)
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
D0	Spontaneous date, birthday	
O0	Spontaneous Department of growing up	
Q2	Spontaneous predominantly no question	

**Table** Erreur! Argument de commutateur inconnu. **Contents and corpus codes of the Colombia 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.COLOM7ES Where: <code>&lt;name&gt;</code> is COLOM from Colombia <code>&lt;#&gt;</code> is 7
-------------------------------	---

	< language code > is the ISO 2-letters code: ES for Spanish
<block >	Defined as: BLOCK<nn> 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> 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** Erreur! Argument de commutateur inconnu. - **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** Erreur! Argument de commutateur inconnu. - **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 Colombia 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"	Colombia_Spanish_Fixed_ Network	database name
VOL:	"%.11s"	COLOM7ES_01	database volume ID
SES:	"%04d"	500	session number

#### File rows

Mnem.	item format	example	Comments
DIR:	"\\%.8s\...\%.8s"	\COLOM7ES\ BLOCK10\SES1000	signal file directory

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)

Mnem.	item format	example	comments
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)

Mnem.	Item format	example	comments
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

mnem.	item format	example	Comments
EXP:	"%s"	Sergio Oller	labeling expert
SYS:	"%s, %d.%02d"	UPC_RevBD,1.00	labeling system
DAT:	"%02d/%.3s/%4d, %02d: %02d: %02d"	18/Jul/1998, 18:52:47	date of completion of labeling

### Label file body

mnem.	item format	example	Comments
LBD:	""		label body keyword

LBR:	"%lu, %lu, %d, %d, %d, \"%s\""	0, 12457, , , , "This is what the speaker should have uttered"	Labeling during recording: begin, end, gain, min, max, orthographic text prompt
EXT:	"%.75s"		line extension
LBO:	"%lu, %lu, %lu, %s"	2800, 4700, 6600, This is what the speaker actually said	Orthographic labeling: begin, center, end, orthographic transcription text

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

LHD: SAM, 5.10  
DBN: Colombia\_Spanish\_Fixed\_Network  
VOL: COLOM7ES\_01  
SES: 1000  
CMT: \*\*\* Speech file information \*\*\*  
DIR: \COLOM7ES\BLOCK10\SES1000  
SRC: A11000A1.ESA  
CCD: A1  
BEG: 0  
END: 19987  
REP: SIEMENS, BOGOTA, COLOMBIA  
RED: 16/Jul/1998  
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/Jul/1998, 18:52:47  
LBD:

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

## 2. Database design and collection

### 2.1 Recording site and platform

Recordings took place at Siemens Colombia. The main characteristics of the recording platform is:

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 NT.  
 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 Siemens personnel, students from several Universities around Colombia and their relatives. The following sex and age distribution has been obtained

Age groups	Number of speakers			Percentage of total %
	Male	Fem.	Total	
under 16	38	18	56	5.26
16-30	277	265	542	50.89

31-45	178	169	347	32.58
46-60	59	40	99	9.30
over 60	11	10	21	1.97
Total	563	502	1065	

The dialectal distribution of speakers is:

REGION	DIALECT	STATES AND AREAS	RECORDED SPEAKERS
Coast	Atlantic	Guajira, Cesar, Magdalena, Atlantico, Bolivar, Sucre, Cordoba, Antioquia, North Santander, Choco	380
Andes	Andes Oriental	Tolima, Cauca, Narin/o, Huila, Cundinamarca, Bocaya, Santander, Bogota.	481
	Andes Occidental	Valle, Antioquia, Caldas, Quindio, Risaralda,	85
Unknown			119

### 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 COLOM7ES\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 part of the prompted text, e.g. lip smack, cough, grunt, throat clear, tongue click, loud breath, laugh, loud sigh.

[sta]: Stationary noise. This category contains background noise that is not intermittent and has a more or less stable amplitude spectrum. Examples: car noise, road noise, channel noise, GSM noise, voice babble (cocktail-party noise), public place background noise, street noise.

[int]: Intermittent noise. This category contains noises of an intermittent nature. These noises typically occur only once (like a door slam), or have pauses between them (like phone

ringing), or change their color over time (like music). Examples: music, background speech, baby crying, phone ringing, door slam, door bell, paper rustle, cross talk.

The Database has been transcribed using the software tool UPCRevBD.v1, developed at UPC.

### 3. Database contents definition

Next table shows the list of words and the definition.

delete	delete an entry, message or list item
Spanish	Country language
end	End of a message
help	request information or menu options for the current dialogue node
no	no
pound	same as hash; telephone keyboard symbol #
star	star of some telephone keyboards *
yes	yes
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**

#### Spontaneous answers

In addition to the above listed set of words, three spontaneous answers were recorded:

Birthday: Recorded to know the age of the Speaker

Department of growing up: Recorded to know the dialectal of the speaker

Predominantly no question: Is the answer to the question: Do you call from a telephone booth?. This is the first question the speaker answer. It's a very easy question to familiarize the speaker with the recording system.

### 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.