

SALA: Speechdat Across Latin America.

Results Of The First Phase

Asuncion Moreno(1), Robrecht Comeyne (2), Keith Haslam (3), Henk van den Heuvel (4), Harald Höge (5), Sabine Horbach (6), Giorgio Micca (7)

(1) UPC, Barcelona, Spain. (2) Lernout & Hauspie, Ieper, Belgium. (3) Vocalis, Cambridge, UK. (4) SPEX, Nijmegen, Netherlands, (5) Siemens AG, München, Germany, (6) Philips, Aachen, Germany, CSELT, Torino, Italy

(1) Universidad Politecnica de Catalunya
Jordi Girona 1-3
08034 Barcelona Spain
asuncion@gps.tsc.upc.es

Abstract

The objective of the SALA (SpeechDat across Latin America) project is to record large SpeechDat-like databases to train telephone speech recognisers for any country in Latin America. The SALA consortium is composed by several European companies, (CSELT, Italy; Lernout & Hauspie, Belgium; Philips, Germany; Siemens AG, Germany; Vocalis, U.K.) and Universities (UPC Spain, SPEX The Netherlands). This paper gives an overview of the project, introduces the definition of the databases, shows the dialectal distribution in the countries where recordings take place and gives information about validation issues, actual status and practical experiences in recruiting and annotating such large databases in Latin America

1. Introduction

SALA (SpeechDat Across Latin America) belongs to the family of SpeechDat projects¹ producing speech databases to train recognizers for telephone applications. In 1998 the project SALA was launched by the companies CSELT (Italy), Lernout & Hauspie (Belgium), Philips (Germany), Siemens (Germany), Vocalis (United Kingdom) and the universities UPC (Spain), SPEX (Netherlands) [Moreno (1998)] The specification of the databases is coherent with those of the SpeechDat II databases. Some adjustments and additions of the specifications have to be made reflecting the different needs for Latin America. Also the validation procedure of the databases is performed along the rules of SpeechDat II. For validation of the databases the institute SPEX – ELRA's validation center of spoken language resources – is responsible.

The goal of SALA is to produce speech databases in a way that the recognizers trained with these databases show good recognition performance for all different dialectal variants of Spanish and Portuguese as spoken in Latin America. Due to the lack of precise phonetic knowledge of the dialectal regions in Latin America the members of the SALA project decided to first collect sample databases recorded from 1000 – 2000 speakers for selected Latin American countries (phase I of the project). In a second phase of the SALA project the recordings should be enlarged for specific dialectal regions. The selection of these regions should be based on recognition experiments made with the databases

produced so far. The project SALA is still in phase I. The countries so far covered are:

Brazil: 2000 speakers

Mexico: 2000 speakers

Venezuela: 1000 speakers

Colombia: 1000 speakers.

Some other companies are in negotiation to join the consortium providing speech databases for Argentina. Still the dialects found in Chile or Peru are not yet covered.

Due to the experiences made so far it turned out that the production of speech databases in Latin America takes much longer time than expected. Main problems which have to be solved concern

- insufficient developed infrastructure of the telephone network
- long lasting contractual and legal issues
- delay through customs regulation.

Now phase I of the SALA project is in an advanced stage where the first databases are ready. So initial recognition experiments can be done at the end of year 2000 to lay the basis of phase II of the SALA project.

In the following sections we will describe the database design and recording platform (section 2), the results of the prevalidation of the first recordings (section 3), dialectal issues (section 4), recruitment of speakers (section 5) and issues on annotation (section 6).

2. Database design and recording platforms

Databases recorded in the SALA project follows the specifications of the SpeechDat II project [Höge (1999)] with minor deviations.

Databases of either 1000 or 2000 speakers are recorded.

¹ <http://www.speechdat.org/>

In each call a minimum of 43 read and spontaneous items are recorded comprising digits and number sequences; phonetically rich sentences and words; names of cities, companies and person names; spellings; keywords and sentences with embedded keywords; dates and times.

Table 1 shows the list of mandatory items to be recorded in each database

Corpus contents
Digits, strings of digits, numbers and amounts
1 sequence of 10 isolated digits
1 isolated digit
1 sheet number (5+ digits)
1 telephone number (9-11 digits)
1 credit card number (14-16 digits)
1 PIN code (6 digits)
1 currency money amount
1 natural number
Dates
1 spontaneous date, e.g. birthday
1 prompted date, word style
1 relative and general date exp.
Times
1 time of day (spontaneous)
1 time phrase (word style)
Application words and sentences
6 SALA application words (25 fixed, 5 free)
1 word spotting phrase using an application word
Spellings
1 spelling of surname (set of 500),
1 spelling of direct. city name
1 real/artificial for coverage
Directory assistance names
1 surname (set of 500)
1 city of birth / growing up (spontaneous)
1 most frequent cities
1 most frequent company/agency
1 "forename surname"
Questions, including "fuzzy" yes/no
1 predominantly "yes" question
1 predominantly "no" question
Phonetically balance
9 phonetically rich sentences
4 phonetically rich words

Table 1. Corpus contents of the SALA databases

A label file containing a manual orthographic transcription of the speech really uttered by the speaker accompanies each audio file. Each database comes with a lexicon containing the phonemic transcription in SAMPA of each word in the orthographic transcriptions

Recording platforms were installed in each country. The recordings were mainly obtained from the fixed telephone network although mobile calls are accepted. The different platforms are described below

2.1. Brazil

The recordings were done at CPqD, Campinas, São Paulo, Brazil. An E1 digital link was connected to the recording platform and a toll-free number was acquired to allow callers to access the system.

The recording platform was based on a Compaq Deskpro PC with a 166 MHz Intel Pentium processor, hard disk of 2.2GB and 64MB of RAM. It was used a Dialogic D/300SC-E1 card for digital trunk E1 with 8 time slots for inbound call, and a Dialogic D/41ESC card for analogue line with attendant. The operating system used was MS-Windows NT 4.0 for workgroups.

The software used in the collection was projected and developed at CPqD by the AudioVisual Digital Signal Processing team. A backup system and a mirror platform were available all the time.

All calls were originated from the PSTN network. In most of the country, the access to a PSTN is analogue, that is, the communication between the final subscriber (user) and the central switching is analogue. Among PSTNs the communication is digital. Usually, but not necessarily, the communication between PSTNs and some research institutes, Universities and companies is also digital. Between PSTN and CPqD the communication is digital. The acquisition platform of CPqD doesn't do the digitalization of the signal. The digitalization is done at PSTN and the acquisition platform simply captures the signal arriving from the central switching.

2.2. Colombia

Recordings took place at Siemens Bogota. The equipment was a Pentium PC at 120 MHz, with 32 MB RAM, 4 GBytes SCSI Hard disk, PCI Network card and Windows NT. The interface is ISDN basic access (BRI) and the board AVM-ISDN-A1. Two lines can be recorded simultaneously. A modem was installed to control the equipment and monitoring the status of the calls from UPC Spain.

The programming interface is a COMMON-ISDN-API Version 2.0 (CAPI 2.0) and the software used for this application is ADA [(Fonollosa 1998)].

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.

There were serious problems of telephone communication in some of the areas we intended to record. Specifically, we noticed problems from the dialectal area of the Choco and from the city Manizales. From Choco we received very few complete calls. From Manizales was imposible to record complete calls and all these recordings were discarded

2.3. Venezuela

A recording platform was installed in the University of Los Andes, in Mérida. The equipment was similar to the previously described except the telephone lines were

analogue and consequently, the interface board. Recordings are stored in μ law.

Low level signals were received from some places and communication failures occur very often. An additional problem is that the electricity power was interrupted very often and supervision of the equipment became necessary. Atmospheric problems damaged the equipment once, and, additionally, during one month the telephone communication was almost stopped because the lines were seriously damaged in the whole country

3. Validation

In order to guarantee quality and homogeneity among databases, a validation procedure is mandatory for each database. Validation, as we will use the term, refers to the quality evaluation of a database against a checklist of relevant criteria. The validation is carried out by SPEX, the Dutch validation center. The validation checks include:

- information in the documentation
- completeness of the recordings
- format of the files
- speaker characteristics
- recording conditions
- acoustic quality of the speech files
- transcription
- lexicon completeness and phone symbol set

The validation criteria are taken from the SpeechDat(II) project. A complete survey of these criteria is given in [Van den Heuvel (1996)]. Some minor adaptations of these criteria were judged appropriate for SALA. For instance, line lengths in the label files may exceed the SAM standard of 80 characters, and phonetically rich sentences are considered missing if all words are mispronounced, unintelligible or truncated.

The validation is performed in two steps, which may or may not be followed by a third step. In the first step, called "prevalidation", a complete mini-database of 10 speakers is checked at the start of the recordings. This prevalidation aims at avoiding design errors which may otherwise show up after completion of the database and are irremediable at that time. As the second step the final validation is done on the completed database. Third, if the database should be rejected after validation, a corrected version can be offered and accepted after a new validation (re-validation) has been carried out.

For all four language variants presently represented in the project (Mexican Spanish, Colombian Spanish, Venezuelan Spanish, Brazilian Portuguese) the first 10 speakers were recorded and stored in a mini-database. These databases were delivered to the validation center and prevalidated in the period July 1999 – March 2000. Some patterns became manifest in the results of these prevalidations:

The sound quality of the speech files is rather good; All databases have their speech files in A-law, except Venezuelan Spanish which has μ -law speech files; Spontaneously spelled items (as defined in SpeechDat(II)) are not included in the (Spanish)

databases, since people do not know how to spell in Spanish. Therefore, all spelled items are prompted; The pronunciation variant of Spanish or Portuguese should be reflected in the phone transcriptions in the lexicon. The lexicon should not be standard Spanish or Portuguese;

Various format errors in the label files and summary tables were observed;

Transcription markers for recording truncations and for filled pauses were missing in the orthographic transcriptions of some of the databases.

The final databases are expected to be submitted for validation during the Spring and the Summer of this year.

4. Dialectal areas

In a previous paper [Moreno (1998)] it was shown how Latin America was divided in eight wide recording areas. One or more countries make up each area. The criteria to define the recording areas include phonetic dialectal information and demographic information.

In the first phase of the project, four of the eight recording areas have been selected. In each area, one country has been chosen to start the collection. A finer dialectal division has been done in each of these countries. This section shows these final dialectal areas

4.1. Brazil

Brazil has a global population of about 159 million inhabitants (1997). Administratively it is divided into 26 states plus a Federal District (capital of the country). The distribution of the population is quite heterogeneous. Excepting the South-East region, that concentrates large population in its interior, people are concentrated in the eastern coast. Additionally, some metropolitan regions group significant part of country's inhabitants.

From a linguistic point of view, Brazilian Portuguese is more regular in pronunciation than Portuguese language. Up to now there is not sufficient scientific information available about the differences that separate the regional varieties existing in Brazil. Therefore, a detailed classification as for the Portuguese dialects is not possible. Still, it exists a proposal of classification based mainly in differences of pronunciation.

From this point of view, CPqD have decided to divide Brazil into five major dialectal regions (macro linguistics regions): South, São Paulo, South-East, North-East, and North plus Centre-West. It is possible to do a finer dialectal division inside each region, where it may be found various dialects spoken in. Table 2 shows the macro linguistics regions (column 1), the dialects spoken in each of them (column 2, finer dialectal division) and the areas and/or states where these dialects are spoken (column 3).

As there were three important migration flows in Brazil in the last 60, there are many dialects spoken in the large regional centres and in the north and Centre-West regions, which are not originals of these places. Any collect made in those regions will include their own dialects as well as dialects of other regions.

REGION	DIALECT	AREAS AND STATES
	Paranaense	Paraná
	Catarinense	Florianópolis , Vale do Itajaí and interior of Santa Catarina
	Gaúcho	Rio Grande do Sul
2. São Paulo	Grande São Paulo	Metropolitan region of São Paulo
	Litoral paulista	Coast region
	Centro paulista	Campinas , Piracicaba , center-north of São Paulo
	Oeste paulista	West of São Paulo , Mato Grosso do Sul, Triângulo Mineiro, north-west of Paraná
3. South-East	Carioca	Rio de Janeiro
	Mineiro	Center of Minas Gerais
	Capixaba	Espírito Santo
4. North-East	Baiano	Bahia, Sergipe and north-east of Minas Gerais
	Pernambucano	Paraíba , Pernambuco and Alagoas
	Cearense	Rio Grande do Norte, Ceará) and Piauí
5. North and Centre-West	Centro-Norte	Pará, Amapá and Maranhão
	Amazonense	Amazonas , Roraima and Acre
	Centro Oeste	Rondônia, Mato Grosso, Tocantins, Goiás and Distrito Federal

Table 2: Dialectal regions in Brazil

REGION	DIALECT	STATE
Central	Central	Guerreo, Morelos, Puebla, Tlaxcala, Distrito Federal, Estado de México, Hidalgo, Querétaro de Arteaga, Guanajuato, Jalisco, Aguascalientes, Michoacán, Nayarit
	North west	Baja California, Sur, Sonora, Sinaloa, Chihuahua, oeste de Durango
	North	Coahuila, oeste de Nuevo León, este de Durango, Zacatecas, San Luis Potosí
	North East	Tamaulipas, este de Nuevo León,
	Oaxaca	Oaxaca,
	Pacific coast	costa de Oaxaca, costa de Guerrero, costa de Michoacán, Colima, costa de Jalisco
	Tabasco	Tabasco
	Veracruz	Veracruz
	Yucatán	Yucatán, Campeche, Quintana Roo.
	Chiapas	Chiapas

Table 3: Dialectal regions in Mexico

4.2. Mexico

Mexico has a total population of about 91 million inhabitants, 50% of whom live in the central region. Mexico City alone has approximately 22 million inhabitants. Spanish is Mexico's only official language. The dialectal variation throughout Mexico can be divided into 5 general regions, 4 of which can be further sub-divided into smaller dialectal divisions. Table 3 depicts the dialectal categories together with the estimated population distribution.

4.3. Colombia

Spanish spoken in Colombia can be roughly divided in low lands (Coast) and high lands (Andes). The low land dialect is very similar to Caribbean dialect and can be divided in Pacific and Atlantic. We do not consider the Amazonian area. Table 4 shows the dialectal areas defined in Colombia

4.4. Venezuela

Venezuela can be divided in five dialectal areas that can be grouped in two main dialects: Coast and Andes. Amazonian is a big area in Venezuela but has been discarded because there are very few Spanish speakers there Table 5 shows the dialectal regions defined in

REGION	DIALECT	STATES AND AREAS
Coast	Atlantic	Guajira, Cesar, Magdalena, Atlantico, Bolivar, Sucre, Cordoba, Antioquia (north), Norte de Santander.
	Pacific	Choco
Andes	Andes Oriental	Tolima, Cauca, Narin/o, Huila, Cundinamarca, Bocaya, Santander, Bogota.
	Andes Occidental	Valle, Antioquia, Caldas, Quindio, Risaralda,

Table 4. Dialectal regions in Colombia

Venezuela.

4.5. SAMPA Symbols

Some phonetic studies were carried out to extend the already accepted Spanish SAMPA (Speech Assessment Methods Phonetic Alphabet)² symbols to the dialectal variety of Latin American. A final table for Spanish including Castillian (Spain) and Latin American variants was accepted by the SALA consortium.

A similar study was carried out by CpQd to have an extended SAMPA symbols for Brazilian Portuguese. The SAMPA symbols used for Brazilian Portuguese language are shown in the tables at the end of this paper. We have added two fricatives (**dZ** and **tS**) and two vowels (**I** and **U**) to the Portuguese set, to complete the sounds of Brazilian Portuguese. Additionally we have classified the liquid **R** of the Portuguese as fricative, and the semi-vowels **j** and **w** as approximants.

5. Speaker Recruitment

Between 1000 and 2000 calls are recorded from different speakers in each area. Speakers are recruited with the goal to obtain a homogeneous dialectal distribution within each area and a good balance of age and sex.

REGION	DIALECT	STATES AND AREAS
Coast	Central	Distrito Federal, Miranda, Carabobo, Aragua, Lara, Yaracuy, Falcón.
	Zuliana	Zulia.
	Llanos	Portuguesa, Guárico, Cojedes, Apure, Barinas
	South -oriental	Sucre, Nueva Esparta, Monagas, Anzoátegui, Delta Amacuro, Bolívar y Amazonas
Andes	Andes	Tachira, Merida, Trujillo

Table 5. Dialectal regions in Venezuela

Balance of sex is defined as 50 % male and 50% female with a deviation of 5%.

Balance of age means that the distribution of speakers should fulfil the following specifications:

Number of speakers between 16 and 30 years old $\geq 20\%$
 Number of speakers between 16 and 30 years old $\geq 20\%$
 Number of speakers between 16 and 30 years old $\geq 15\%$

And it's recommended to have a 5% of speakers less than 16 years old.

Different strategies of recruiting have been used in each country. This section explains the strategies and results obtained

5.1. Brazil

To reach a total of 2000 valid calls, a 10% over sampling strategy was used in the speech collection. So, about 2200 speakers were contacted from all over the country (1100 males and 1100 females).

Based in the economical importance of each region (measured in this case by its population and telephonic density), and by considering the mix of dialects present in some of them, the target population chosen to represent each region was the following:

REGION	POPULATION %	TARGET
1. South	15%	400
2. São Paulo	23%	450
3. South-East	21%	450
4. North-East	25%	550
5. North and Centre-West	16%	150
Total	100%	2000

Table 6: Target population of each dialectal region in Brazil

The target population of the fifth region was small because its strong migration and low demographic density.

In each region, it was planned to accomplish the age and sex distribution that is mandatory for the all database

The recruitment of speakers was subcontracted to a company and some humanitarian associations with offices along all over the country and covering the specified regions. They were controlled by CPqD to ensure that the designed percentages were complied with.

Each call was pre-tested to accept it or to reject it. This pre-test consisted in to listen to about 14 speech files of each call (typically, the phonetically rich words, the application words, one natural number, one date and 1 or 2 sentences). If these items were OK, the call was accepted; if there were 2 or more of these items with problems, the call was rejected

When a call was rejected, the speaker recruitment company was informed to provide us with a new call using the same script

5.2. Mexico

Speakers are being recruited via a large Mexican Telephone company – their employees are being contacted throughout Mexico and sent a prompt sheet together with an instruction sheet. Friends and family of the employees are also contacted and asked to participate. Also a couple of Universities have been asked to participate in the collection. A monetary incentive (ie chance to win a prize at the end of the data collection) is proffered to encourage potential speakers.

REGION	DIALECT	POPULATION %	TARGET COVERAGE %
Central	Central	48	50
North	North west	12	20
	North	10	
	North East	4	
Pacific	Oaxaca	4	10
	Pacific coast	6	
Atlantic	Tabasco	2	10
	Veracruz	7	
Yucatan	Yucatán	4	10
	Chiapas	4	

Table 7 Target distribution of speakers among dialects in the Mexican database

5.3. Colombia

Table shows the final distribution of speakers recorded in Colombia. Recruitment was carried out by students from several Columbian Universities

REGION	POPULATION	RECORDINGS
Coast_Atlantic	10,4	382
Coast_Pacific	0,3	14
Andes_East	14,2	428
Andes_West	10,2	169
Void		7

Table 8 Dialectal regions, population in millions of persons, and calls received from each region.

5.4. Venezuela

Speakers from Venezuela have been recruited via students, professors, friends and relatives from several Universities. A lottery prize incentive speakers to call. The distribution of recorded speakers among dialects is shown in Table 9

REGION	DIALECT	POPULATION	RECORDED
Coast	Central :	7,2	204
	Zuliana :	1,7	211
	Llanos	3,0	191
	South -oriental	2,6	198
Andes	Andes	1,6	216

Table 9 Recorded distribution of speakers among dialects in the Venezuelan database

6. Annotation

Annotation of the SALA databases is done at an orthographic level. An orthographic transcription of what was really uttered is done manually. Extra marks point to mispronunciations, truncations, unintelligible words and noises either coming from the speaker, the environment or the channel. For most of the databases, the annotation procedure is in an advanced stage.

6.1. Brazil

In order to view and listen to the speech files, we used Cool Edit Pro V 1.2 (from Syntrillium Software Corporation)

To do the transcriptions of the accepted calls, various software tools were created. The procedure we used was the following:

1. First of all the transcriber has to listen to the spontaneous questions of the call and fill out an electronic file with this information. A program picks up this file and creates a "basic label file". A "basic label file" is a file with the information that is common to all the label files of the call

2. Another program takes the script file and creates a "basic transcription file". The "basic transcription file" contains automatic transcription of script items. This transcription is made following the SALA conventions. It is over this file that transcribers work to include transcription symbols, modify or correct transcriptions, etc

3. A third program takes the script file, the transcription file and the "basic label file" created in 1, generating all the label files of the call

There are some other software tools we create to manipulate the label and speech files and to build the table and index files of the database. The table and index files are manually revised to detect possible errors made in the label files. Also, there are some calls randomly chosen to be processed with VOX software to check their correctness.

About the most relevant problems we encountered, we can mention the low signal to noise ratio of some calls from Belém (North region). In some of these calls there are a constant background noise of high level. This is typical of this region because the low quality of its telephone net and because most of their calls reach São Paulo through a satellital link. Another problem we have encountered is the high number of calls we have had to reject, mainly, because the people spoke before the beep signal

6.2. Mexico

The transcription is being undertaken by the University of Puebla in Mexico and the transcription tool being used is CSELT's licensed software package Vox! The transcribers have been suitably instructed to use this tool and have been supplied with the SALA transcription rules. Transcriptions are case sensitive and are made using the ISO-8859 -1 character set. No problems have emerged to date.

6.3. Colombia

Annotation of the Colombian database was taken by UPC, Spain. Experienced people who had already transcribed SpeechDat II were selected to do this task. The Spanish Colombian Database has been transcribed using the software tool UPCRevBD.v1, developed at UPC.

6.4. Venezuela

The Venezuelan database is being annotated by UPC, Spain. Experienced people who had already transcribed SpeechDat II were selected to do this task. More than 500 speakers have been transcribed up to the date of writing this paper. The Spanish Venezuelan Database is being transcribed using the software tool UPCRevBD.v1, developed at UPC.

7. Current Status

At present, the collection of calls in Brazil has terminated. Since CPqD will have to provide 2,000 valid calls, the number of collected calls should be in the range 2200-2400

The current status of the Mexican collection is that the database has been prevalidated. The main recording has only just begun after prevalidation and approximately 50 speakers have now been recorded but none yet transcribed.

The Database recorded in Colombia is completely finished and at the moment of writing this paper is being validated by SPEX

The Venezuelan collection has also finished and more than a half of the files have been annotated.

8. References

- Draxler C., van den Heuvel H., S.Tropf H.S. (1998) SpeechDat Experiences in creating Large Multilingual Speech Databases for Teleservices. Proceedings LREC98, 28-30 May 1998, Granada, Spain, Vol. I, pp 361-366.
- Hoeghe, H., Draxler, C., Van den Heuvel, H., Johansen, F.T., Sanders, E., Tropf, H.S. (1999): Speechdat multilingual speech databases for teleservices: across the finish line. Proceedings EUROSPEECH99, Budapest, Hungary, 5-9 Sep. 1999, Vol. 6, pp. 2699-2702
- Höge, H., Tropf, H., Winski, R., Van den Heuvel, H., Haeb-Umbach, R. & Choukri, K. (1997) "European speech databases for telephone applications". Proceedings ICASSP97. Vol. III, pp. 1771-1774. Munich Germany
- Moreno, A., Höge, H., Koehler J., José B. Mariño J.B. (1998) "SpeechDat Across Latin America. Project SALA" LREC'98. Granada. Spain
- Van den Heuvel, H. (1996): Validation criteria. SpeechDat Technical Report SD1.3.3., 1996.
- Fonollosa J.A. R., A. Moreno A.(1998) Automatic Database Acquisition Software for ISDN PC Cards and Analogue Boards LREC'98, Granada (Spain). 28-30 May 1998.