SpeechDat across all America: SALA II

Asunción Moreno[1], Oren Gedge[2], Henk van den Heuvel[3], Harald Höge[4], Sabine Horbach[5], Patricia Martin[6], Elisabeth Pinto[7], Antonio Rincón[8,] Franco Senia[9], Rafid Sukkar[10]

[1]UPC, [2]NSC, [3]SPEX, [4]Siemens AG, [5]Philips Speech Processing, [6]Microsoft Corp, [7]Telisma, [8]ATLAS, [9]Loquendo, [10]Lucent Technologies

Asunción Moreno, Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034 Barcelona Spain
asuncion@gps.tsc.upc.es
http://www.sala2.org

Abstract

SALA II is a project co-sponsored by several companies that focuses on collecting linguistic data dedicated for training speaker independent speech recognizers for mobile/cellular network telephone applications. The goal of the project is to produce *SpeechDat*-like databases in all the significant languages and dialects spoken across Latin America, US and Canada. Utterances will be recorded directly from calls made from cellular telephones and are composed by read text and answers to specific questions. The goal of the project should be reached within the year 2003.

1. Introduction

Within the past project "SpeechDat Across Latin America" (SALA) [Moreno 2000], a large number of databases were recorded in Latin America. Databases have 1000-2000 speakers each and were designed to have enough data to train ASR systems for fixed network telephone applications in any Latin American country. The quick development of mobile/cellular telephony makes necessary to extend the ASR systems to mobile networks. There is a lack of public domain databases recorded from mobile telephones in America. CTIMIT is the only one available in the LDC[1] and was obtained playing back TIMIT through cellular telephones. In Europe, a set of databases recorded in the framework of the SpeechDat project can be found in the ELRA[2] catalogue. The goal of the current project 'SpeechDat across all America' (SALA II) is the collection of a large number of databases in Latin America, US and Canada. The speech databases created in the project cover all dialectal regions of America representing the dialectal variants of English, French, Portuguese and Spanish languages. America is divided in large recording areas where a database is created. Further, a fine dialectal map is defined for each recording area. Each database will be composed by up to 4000 speakers. Utterances will be recorded directly from calls made from cellular telephones and are composed by read text and answers to specific questions. The goal of the project should be reached within the year 2003. The databases are dedicated for training speaker independent speech recognisers applicable for mobile/cellular network telephone applications. Most of the database will be public available through ELRA.

This paper describes the consortium, goals of the project and specifications of the databases. Section II describes the organisation, structure and partners of the SALA II consortium. Sections III and IV describes the recording areas and databases to be produced in the project. Section V describes the specifications and section VI deals with the validation procedure of the databases produced in the consortium.

2. Organization of the SALA II Project

The SALA II project is founded by an Industrial consortium. Members of the consortium are: Natural Speech Communication (Israel), Siemens AG (Germany), Philips Speech Processing Aachen (Germany), Microsoft Corp, (United States), Telisma (France), Applied Technologies on Language and Speech, S. L. "ATLAS" (Spain), Loquendo, (Italy) and Lucent Technologies (United States). All these members produce at least one database in the project. Additionally, two other members compose the consortium: Universitat Politècnica de Catalunya "UPC" (Spain), that is the co-ordinator of the project and Speech Processing Expertise Centre "SPEX" (Netherlands) that performs the validation of the produced databases.

All industrial members produce a database in Latin America and optionally, a database in US and Canada. The consortium is open to industrial or public members. All members that produce a database in Latin America have access to the databases produced in Latin America. The same criterion applies to the databases produced in US and Canada.

3. Recording areas and databases produced in the project

As mentioned above, the objective of the SALA II project is to create speech databases to train speech recognition systems for many telephone cellular oriented applications in any American country. The speech databases created in the project cover all dialectal regions of America representing the dialectal variants of the English, French, Portuguese and Spanish languages. For this need America is divided in large recording areas and a set of databases and their corresponding dialectal coverage have been defined and accepted by the consortium. The

[1] http://www.ldc.upenn.edu

[2] http://www.icp.grenet.fr/ELRA/home.html

list is split in two groups, the Latin American group and the US & Canada group.

The Latin America group includes eighth databases, 1000 speakers each, covering the most important Spanish and Portuguese dialectal variants. The definition and partner responsible of the collection is shown in Table 1:

Country	Partner
Brazil	Philips Speech.Proc.
Mexico	NSC
Caribbean: Venezuela	ATLAS
Central America: Costa Rica	Telisma
Colombia	Siemens A.G.
Peru	Microsoft Corp.
Chile	Loquendo
Argentina	Lucent Technologies

Table 1: *SALA II* databases produced in Latin America and partners

In the US & Canada group up to four databases have been defined to cope with English and Spanish as spoken in USA and French and English as spoken in Canada. Table 2 shows the databases to be produced in US and Canada, the number of speakers to be collected and the partners involved. At the moment of writing this paper, no company is willing to record Canadian English yet.

Language	Speakers	Partners
US English	4000	Loquendo, Microsoft, NSC, Siemens
US Spanish	2000	Philips, ATLAS
Canadian French	1000	Telisma
Canadian English	1000	

Table 2: *SALA II* databases produced in US and Canada and partners

4. Dialectal distribution of each database

It' s necessary that each database produced in the consortium have a good representation of all the dialects spoken in the area is being recorded. In the former SALA project Latin America was divided into eighth large recording areas with similar broad dialect variation. Each area is composed by one or more countries. For each of the Spanish spoken Latin America databases to be produced in the consortium a fine dialectal division was defined [Moreno, 1997], [Moreno 2000]. In the current SALA II project, the same distribution is going to be used.

Concerning Brazil, US Spanish and US English, a similar study has been made. The final distribution for each database is shown below.

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 into five major dialectal regions (macro linguistics regions): South (Paranaense, Catarinense, Gaúcho), São Paulo (Grande São Paulo, Litoral paulista, Centro paulista, Oeste paulista), South-East (Carioca, Mineiro, Capixaba), North-East (Baiano, Pernambucano, Cearense), and North plus Centre-West (Centro-Norte, Amazonense, Centro Oeste). 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.

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 is the following:

Region	Population (%)	Target speakers
South	15%	200
São Paulo	23%	225
South-East	21%	225
North-East	25%	275
North & Centre-West	16%	75
Total	**100%**	**1000**

Table 3: Brazilian database: dialectal regions, population and target distribution of speakers to be recorded.

4.2. US Spanish

The total US population in the United States is around 240 millions inhabitants (1990). The total Hispanic population is 22 million inhabitants. Table 4 shows the distribution of the Hispanic US population in terms of their specific origin dialect. In this table, the origin dialects are grouped as defined in [Moreno 1997].

It becomes obvious that most of the Hispanic population is either Mexican 61%, or Caribbean (Puerto Rican, Cuban, Dominican) 19,26 %. Therefore, the recordings concentrate on the first two dialects (Mexican and Caribbean Spanish). The collection will be done where they appear most, that is: California, Texas, New York and Florida.

Dialect origin	Population %
Mexican	61.19
Caribbean	19.26
Central American	5.33
Colombian and Panamanian	2.14
Peruvian and Ecuadorian	1.5
Miscell. Central and South	1.6
Miscellaneous Other	8.98
Total Hispanic	**100**

Table 4: Relative distribution of Hispanic population in US in terms of their dialectal origin.

4.2.1. Bilingualism

For doing the US_Spanish data collection, the fact of bilingualism Spanish/English in the U.S. has to be considered. Especially younger people who have been growing up in the States will tend to use both languages together, therefore mixing US_Spanish with English and vice versa, also within an application. This means that a certain percentage of Hispano-English has to be covered as well, in order to be able to switch between the two languages while using the recognition system. So approximately 5% of the corpus should be made up by English words, especially names (company, brand, person and city names) and in the phonetically rich sentences.

With this approach, it should be possible to deliver a US_Spanish lexicon while using the US_Spanish phoneme inventory plus a separate Hispanic English lexicon with a US_English phoneme inventory.

4.3. US English

A database of 4000 speakers will be recorded in English as spoken in US. American English is characterized by phonetic variation based on the speakers' origin from geographical regions. In order to cover this variation adequately, the American English database consists of speakers collected from nine regions of the United States, based on the ' Phonological Atlas of North America' . These regions are as follows:

1. North Central: major parts of North and South Dakotas, Wisconsin, Upper Peninsula of Michigan (excluding the eastern tip), all of Minnesota, upper northern part of Iowa.
2. Inland North: almost all of mainland Michigan, eastern corner of Michigan Upper Peninsula, Southeast Wisconsin, Northeast Illinois and West New York state.
3. Eastern New England: Maine, New Hampshire, Northeastern Massachusetts.
4. New York City: the city of New York
5. Western New England: Vermont, west and south Massachusetts, all of Connecticut, Rhode Island, Eastern New York state (excluding New York city)
6. North Midland: Nebraska (excluding northwest corner), northeast Kansas, south Iowa, upper north Missouri, north half of Illinois (excluding northeastern corner), northern haves of Indiana and Ohio, parts of Pennsylvania, including Pittsburgh and St.-Louis.
7. South Midland: south and west Kansas, north Oklahoma, upper northwest of Texas, central Missouri, south halves of Illinois, Indian and Ohio, upper northern corner of West Virginia, northern half of Maryland, north and central Delaware, parts of Pennsylvania, including Philadelphia.
8. West: Washington, Oregon, Idaho, Montana, Wyoming, Colorado, New Mexico, Arizona, Utah, California, Nevada, northwestern Nebraska, west corner of Texas.
9. South: Texas (not west or northeastern corners), southern Oklahoma, Arkansas, Louisiana, south Missouri, Mississippi, Alabama, Georgia, Florida, North and South Carolinas, Tennessee, Kentucky, Virginia, West Virginia (excluding upper northern corner).

The vernacular as spoken by African Americans, known as African American Vernacular English (AAVE) is not collected as a separate dialect, since judicious choice of speakers within each geographical region can cover these variations.

Table 5 shows the definition of each area and the target number of speakers to be recorded in each area.

Region	Target # of speakers
1. North Central	445
2. Inland North	445
3. Eastern New England	445
4. New York City	445
5. Western New England	445
6. North Midland	445
7. South Midland	445
8. West	445
9. South	440
Total	**4000**

Table 5: Target number of speakers to be recorded in each dialectal region in the US English database

5. Database specification

Within the SpeechDat [Höge, 1999] family projects, the speech databases are carefully specified[3] concerning content [Winsky 1997] (digits, numbers, application words and phrases, phonetic rich words and sentences,...), speaker coverage [Senia, 1997 b]: sex, age and dialectal regions, recording environments and recording devices (ISDN recording). Further, the annotation procedure [Senia, 1996]. and format of the database is defined [Senia, 1997a]. SALA II follows closely those specifications with minor modifications.

[3] http://www.speechdat.org

5.1. Corpus content

The corpus content of each SALA II database is summarised in Table 6

Corpus contents
6 application words
1 sequence of 10 isolated digits
1 sheet number (5+ digits) optional
1 telephone number (9-11 digits)
1 credit card number (14-16 digits)
1 PIN code (6 digits)
1 spontaneous date, e.g. birthday
1 prompted date, word style
1 relative and general date exp.
1 word spotting phrase using an application word (embedded)
2 isolated digits
1 spontaneous spelling, e.g. own forename
1 spelling of direct. city name
1 real/artificial spelling for coverage
1 currency money amount
1 natural number
1 spontaneous, e.g. own forename
1 city of birth / growing up (spont)
1 most frequent cities
1 most frequent company/agency
1 "forename surname"
1 predominantly "yes" question
1 predominantly "no" question
9 phonetically rich sentences
1 time of day (spontaneous)
1 time phrase (word style)
4 phonetically rich words

Table 6. SALA II corpus contents

5.2. Speakers sex and age

Each database produced in the consortium is composed by different speakers and contains the same number of male and female speakers. Speakers will be recruited in the following age categories:

Category 1: 16 - 30 years represented by at least 20% of calls
Category 2: 31 - 45 years represented by at least 20% of calls
Category 3: 46 - 60 years represented by at least 15% of calls

Callers younger than 16 or above 60 are optional.

5.3. Environments

The SALA II specifications for mobile/cellular recordings include calls made from both, hand held and hands free devices in cars in addition to the office, home, public places, and vehicle environments commonly used in the SpeechDat databases recorded from mobile telephones.

Each above mentioned environment has to be represented in the database. There are two constrains. One affects to the full database as a whole, and the other affects each dialectal region defined in the database. Table 7 shows the distribution of speakers by environment to be applied in each database. Each dialectal region requires at least 20% of speakers calling from noisy environments (labelled 1, 2, 3) and 20% of speakers calling from environment 4.

Environment	Full DB distribution	Each dialect region distribution
1. Car, train, bus	20 % ±5%	≥ 20%
2. Public place	25 % ±5%	
3. Street	25 % ±5%	
4. Home/Office	25 % ±5%	≥ 20%
5. Car kit (hand free mode)	5 % ± 1%	No restriction

Table 7 - SALA II environment distribution in the full database and in each dialectal region.

5.4. Transcription and annotation

Each signal file has an accompanying label file. Label file contains, among other, the orthographic transcription of what was really uttered by the speaker. The orthographic transcription is done manually by trained transcribers. Mispronunciations, truncations or unintelligible words are annotated with a symbolic convention. Additionally, some noise marks are added at transcription time: Speaker noise, background or stationary noise, intermittent noise and distortions due to mobile channel such as fading or loss of signal.

5.5. Specification of the lexicon

A lexicon containing all the words really uttered is included in the database. The lexicon contains the list of words and their phonetic transcription into SAMPA symbols.

5.6. Recording platform

All databases are recorded from calls made from mobile or cellular hand sets. A recording platform is connected directly over the fixed PSTN using two ISDN lines [Rodríguez Fonollosa, 1998]

Eventually, in some countries ISDN lines are not available. In such case, analogue lines will be used instead.

6. Validation

Validation, as the term is used in the area of speech database collection, refers to the quality evaluation of a database against a checklist of relevant criteria. These criteria typically consist of the specifications of the databases, as defined at the start of the project, supplemented by a number of tolerance margins for these specifications (Van den Heuvel, 2000). In a project such as SALA II, it is of utmost importance that the produced collections are of equal quality. This warrants that the exchange of the databases between consortium partners is fair. As in SALA, the Speech Processing Expertise Centre, SPEX, was chosen as the validation centre for the project (Moreno et al., 2000).

6.1. Validation procedure

Validation is performed in two steps. In the first step, called "pre-validation", a complete mini-database of 10 speakers is compiled for each language and checked before substantial parts of the database are recorded. This pre-validation aims at avoiding design errors that may otherwise show up after completion of the database and are irremediable at that time. The second step is final validation that is performed on the completed database. A third step can be necessary, if the database should be rejected after validation. In that case, a corrected version can be offered for re-validation.

The pre-validations of the databases recorded in SALA II are estimated to take place after the summer of this year, whereas the validations of the completed databases are scheduled for year 2003.

6.2. Validation criteria

The validation checks pertain to the following features of the databases:

- information in the documentation files
- 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, more specifically those defined for the mobile recordings (Van den Heuvel, 1997). Typically new for SALA II are:

- Each application word should appear at least 125 times at orthographic transcription level (= #speakers/8);
- Each phoneme in the phonetically rich sentences should appear at least 100 times at orthographic transcription level (= #speakers/10);
- Idem for phonetically rich words;
- A maximum of 5% of the speakers may call twice;
- Each dialect in a database should be represented by at least (50 x size of the database/1000) speakers;
- In each dialect at least 20% of the speakers are recorded in the vehicle, public place and street environments;
- In each dialect at least 20% of the speakers are recorded in the home/office environment

7. References

Höge, H., C. Draxler, H. van den Heuvel, F.T. Johansen, E. Sanders, H. Tropf (1999). Speechdat multilingual speech databases for teleservices: Across the finish line. In *Proceedings of EUROSPEECH '99*, vol. 6 (pp. 2699–2702). Budapest: ESCA.

Moreno, A. (1997) *Dialectal areas in Latin America for speech recognition applications*. Deliverable for the SALA project. November 1997. http://www.sala2.org

Moreno, A. R. Comeyne, K. Haslam, H. v. d. Heuvel, H. Höge, S. Horbach, G. Micca (2000). SALA: SpeechDat across Latin America. Results of the first phase. In *Proceedings of the Second International Conference on Language Resources and Evaluation vol. II* (pp. 877–882). Athens. Greece.

Rodríguez Fonollosa, JA, A. Moreno (1998) Automatic Database Acquisition Software for ISDN PC Cards and Analogic Boards.Proceedings of First International Conference on Language Resources. Granada. Spain, 1998.vol II. Pp 1325-1329. Editores A. Rubio, N. Gallardo, R. Castro, A. Tejada. Copyrights ELRA 1998

Senia, F. and J.G. Van Velden (1996) *"Specification of orthographic transcription and lexicon conventions"*. SpeechDat project, doc ref LE2-4001-SD1.3.2, 1996.

Senia, F. (1997 a) *"Specification of speech database interchange format"*, SpeechDat project, doc ref LE2-4001-SD1.3.1, 28 February 1997.

Senia, F. et al. (1997 b) *"Environmental and speaker specific coverage for Fixed Networks"*, SpeechDat project, doc ref LE2-4001-SD1.2.1, 26 February 1997.

Van den Heuvel, H (1997) *Validation criteria for databases*. SpeechDat Technical Report, SD1.3.3. http://www.speechdat.org/SpeechDat.html

Van den Heuvel, H. (1997) *"Validation criteria"*, SpeechDat project, doc ref LE2-4001-SD1.3.3, 11 March 1997.

Van den Heuvel, H. (2000) The art of validation. In: *The ELRA Newsletter*, Vol. 5(4), pp. 4-6.

Winsky, R. (1997) *"Definition of Corpus, scripts and standards for Fixed Networks"*, SpeechDat project, doc ref LE2-4001-SD1.1.3, 22 January 1997.