SALA II across the finish line: a large collection of mobile telephone speech databases from North and Latin America completed

Henk van den Heuvel[1], Phil Hall[2], Harald Höge[3], Asuncion Moreno[4], Antonio Rincon[5], Francesco Senia[6]

[1]SPEX, Nijmegen, Netherlands; [2]Appen Pty Ltd, Chatswood, Australia; [3]Siemens AG, Munich, Germany; [4]UPC, Barcelona, Spain; [5]S.L. "Atlas", Barcelona, Spain; [6]Loquendo Vocal Technology and Services, Turin, Italy

e-mail: henk@spex.nl

Abstract

The SALA II project comprises mobile telephone recordings according to the SpeechDat (II) paradigm for several languages in North and Latin America. Each database contains the recordings of 1000 speakers, with the exception of US Spanish (2000 speakers) and US English (4000 speakers). A quarter of the recordings of each database are made respectively in a quiet environment (home/office), in the street, in a public place, and in a moving vehicle. This paper presents an evaluation of the project. The paper details on experiences with respect to the implementation of design specifications, speaker recruitment, data recordings (on site), data processing, orthographic transcription and lexicon generation. Furthermore, the validation procedure and its results are documented. Finally, the availability and distribution of the databases are addressed.

1. Introduction

The goal of the project 'SpeechDat across all America' (SALA II) is the collection of several Spoken Language Resources (SLR) to train speech recognition systems for a wide range of cellular telephone applications in any American country. The SALA II project is funded by an Industrial consortium. Industrial members of the consortium are: Natural Speech Communication (Israel), Siemens AG (Germany), Microsoft Corp., (United States), Phonetic Systems (Israel), Telisma (France), Applied Technologies on Language and Speech, S. L. "ATLAS" (Spain), Loquendo (Italy), and Scansoft (Belgium). Two non-commercial members complete the consortium, viz. Universitat Politècnica de Catalunya "UPC" (Spain), the co-ordinator of the project, and the Speech Processing Expertise Centre "SPEX" (Netherlands) that performs the validation of the produced databases.

All industrial members produce one SLR of 1000 speakers in Latin America, and one SLR of 1000 speakers either in US or Canada. The consortium is open to industrial or public members. All members that produce a SLR in Latin America have access to the other SLR produced in that continent. The same criterion applies to the SLR produced in 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. For this purpose, America is divided in large recording areas where a database is created. Table 1 shows the produced SLR, the number of speakers in each database and the partners involved. The databases and their corresponding dialectal coverage have been agreed upon by the consortium. Figure 1 shows a map of America. Dots show places were recordings took place.

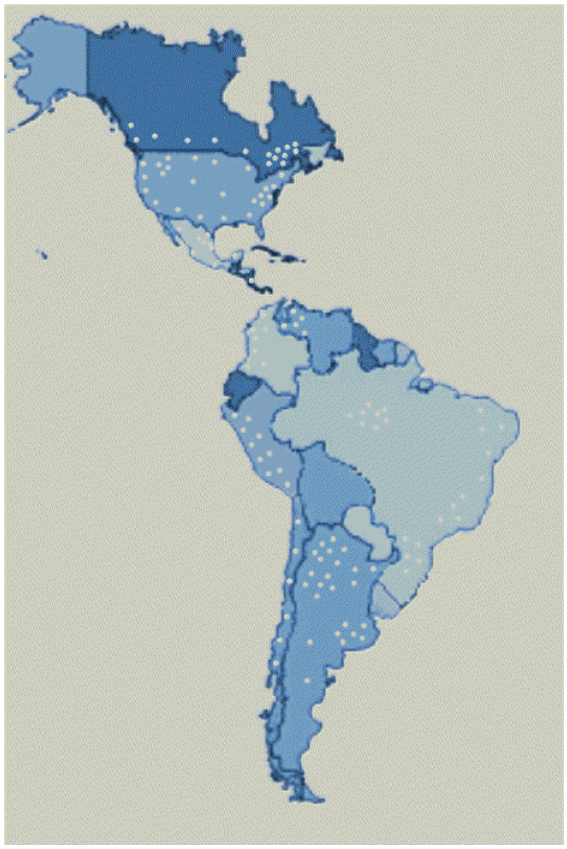

Figure 1. Area covered by the SALA II recordings. White dots indicate the recording places.

Database	Speaker	Partners
US English	4000	Loquendo, Microsoft, NSC, Siemens
US Spanish	2000	Scansoft, ATLAS
Canadian French	1000	Telisma
Canadian English	1000	Phonetic Systems
Brazil	1000	Scansoft
Mexico	1000	NSC
Venezuela	1000	ATLAS
Costa Rica	1000	Telisma
Colombia	1000	Phonetic Systems
Peru	1000	Microsoft
Chile	1000	Loquendo
Argentina	1000	Siemens

Table 1. List of databases, number of speakers and producing partner. If no language is specified, the official language in that country is recorded.

Currently, all SLR with variants of Spanish and the US English SLR recorded in North America are completed. The other databases are expected during summer 2004.
This paper is structured as follows. Section 2 summarises the design specifications. Lessons learnt are addressed in section 3. Section 4 deals with the validation of the SLR and the results. Finally, section 5 presents information about the availability of the SLR.

2. Database Specifications

All SLR have the same corpus contents. The items to be read by each speaker are listed in Table 2.

Corpus contents
6 application keywords/keyphrases
1 sequence of 10 isolated digits in one utterance
1 sheet number (5+ digits) (optional)
1 telephone number (9-11 digits)
1 credit card number (14-16 digits)
1 PIN code (6 digits) (set of 150 codes)
1 spontaneous date, e.g. birthday
1 prompted date, word style i.e. not digital
1 relative and general date expression
1 word spotting phrase using embedded application words
2 isolated digits
1 spelling of proper name, spontaneous (e.g. own forename) or read speech (set of 500+)
1 spelling of directory assistance city name
1 real/artificial to maximise letter coverage
1 money amount in local currency, mixed size and units
1 natural number
1 proper name, spontaneous (e.g. own forename) or read speech (set of 500+)
1 city of birth / growing up (spontaneous)
1 most frequent cities (set of 500)
1 most frequent companies/agencies (set of 500)
1 "forename surname" (set of 150 "full" names)
1 predominantly "*yes*" question
1 predominantly "*no*" question
9 phonetically rich sentences
1 time of day (spontaneous)
1 prompted time phrase, word style i.e. not digital
4 phonetically rich words

Table 2. SALA II corpus contents per recording session. Spontaneous items are shaded.

The SALA II specifications for mobile/cellular recordings include calls made in five different environments. Table 3 shows the distribution of speakers over the environments per database. Each dialectal region requires at least 20% of speakers calling from the noisy environments (labelled 1, 2, 3) and 20% of speakers calling from environment 4.

Environment	Full database distribution	Each dialect region distribution
1 Moving vehicle	20 % ± 5%	≥ 20 %
2 Public place	25 % ± 5%	
3 Street	25 % ± 5%	
4 Home/office	25 % ± 5%	≥ 20 %
5 Car kit (hands free mode)	5 % ± 1%	No restriction

Table 3. SALA II environment distribution in the full database and in each dialectal region.

Each signal file has an accompanying label file. A label file contains, among other things, 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 the time of transcription.
Full specifications of the SALA II SLR are provided in Moreno (2002).

3. Lessons Learnt / Experiences

Per database this section presents relevant experiences with respect to the implementation of design specifications, speaker recruitment, data recordings (on site), data processing, orthographic transcription and lexicon generation. It focuses on lessons learnt and (resulting) deviations from the specifications. Reports are available for Peruvian Spanish (3.1), US English (3.2) and the other SLR in Spanish (3.3). The experiences for the Spanish databases recorded in Argentina, Chile, Costa Rica, Venezuela, Mexico and the USA are kept in one section, since all of them were recorded by the same producer (ATLAS).

3.1 Peruvian Spanish

Development of the SALA II Peruvian Spanish database was planned for a four and a half month period from July to November 2002. Completion of the database was delayed when a six week telecommunication strike prevented installation of the ISDN line required for data collection. The database was nonetheless developed by

Appen within 6 months, and was delivered to SPEX for validation in January 2003.
During the recording phase of the project, the only significant issues encountered were: incomplete calls due to mobile network failure, intermittent mobile signal dropout, and occasional items being read out of sequence. Incomplete calls were fully re-recorded since it was considered impractical for callers to begin a new call that would continue from the point at which the previous call had been lost. In some cases intermittent signal dropout made it difficult to associate calls with their respective scripts. The inclusion of additional non-mandatory items aided this process. The problem of items being read/recorded out of sequence was more serious, and was ultimately resolved by hand checking. Literacy of speakers did not prove to be problematic, not even in older age groups. Based on prior experience, it was decided that all calls should be made under supervision (scripts were not simply distributed to potential callers) – this strategy was an important factor in the successful completion of the data collection within a tight timeframe.
Deviation from SALA II specification was only by way of expansion of coverage. Each speaker recorded 49 items which, in addition to the prescribed SALA II script items, included a number of non-mandatory items. Speakers were asked to give the region in which they grew up, their gender, and the defined "scenario" for the call. Speakers also read the prompt sheet identification number which provided additional digit tokens. Steps were also taken to achieve more substantial coverage of equivocal or "fuzzy" responses. Two items were added to the script for this purpose: a confirmation question designed to elicit equivocal or "fuzzy" responses; and a "read" fuzzy item ("*posiblemente*", "*no creo*" etc).

3.2 US English

The US English speech database was recorded in 2003 by recruiting over 4000 American English speakers. The database is co-owned by Loquendo, Microsoft, NSC and Siemens and was collected by the OGI School of Science and Engineering at the Oregon Health & Science University; ELDA supervised the collection.
During recording, the recording script played the initial instruction, asking users to key in their prompt sheet number, gender, age and other information useful to fully characterise a call. It then went through all of the prompts. If a call was terminated half way through, the user could phone back and, after keying in their prompt sheet number, resume where they left off. As in the past for similar recordings, one of the most difficult tasks was the speaker recruitment. OGI started contacting non-profit organizations, private and public schools, as well as student organizations to contract regional recruiters. The regional recruiters had to pass out the prompt sheets to the members of their organizations, or the general public who were willing to participate. The organizations received a fee for each subject that they recruited.
The use of non-profits did not work as well as had been hoped; thus, roughly 3000 of the subjects were recruited by OGI personnel directly in the northeast, Columbus, Ohio and New Orleans.
The prompting material was generated with a large range of different values to ensure a lot of variability in the corpus, but in some cases, such as phonetically rich sentences and words, OGI used feedback to ensure a good coverage 'on the fly'. In fact OGI first created 20000 different prompt sheets and, after 3200 phone calls were recorded, they recreated 800 new prompt sheets designed to ensure coverage of the prompt items that had been recorded least often.
The database complies to the general design adopted by the SALA-II consortium with small variations to adapt to the specific characteristics of the USA. For example, it was not possible to ensure that 50% of the telephone numbers artificially generated were cellular phone numbers, because no standard numbering plan for cellular telephones exist in the US.
The United States was divided into nine dialectal regions, based on the Phonological Atlas of the University of Pennsylvania. Some 25 Spanish speaking persons were recorded (mainly from Florida) and marked as "foreign".
More than seven different network providers were involved (using four different mobile telephone networks technologies: AMPS, CDMA, TDMA, GSM and iDEN - Motorola proprietary), but most of the calls came from T-Mobile (GSM), Verizon (CDMA) and AT&T (TDMA).
A more detailed account of this database can be found in Heeman (2004).

3.3 Other Spanish SLR

The SLR containing Spanish as spoken in Argentina, Chile, Costa Rica, Venezuela, Mexico and the USA were produced by ATLAS.
The prompted material was designed to cover the dialectal differences in each country. In Argentina, Chile, Mexico and Venezuela, previous prompt texts were available from the fixed network telephone collection of the SALA I project; only minor modifications to adapt the prompted texts to the new specifications were required. For the collection in the USA, where the main target population are mainly Mexican and Caribbean speakers of Spanish, part of the phonetically rich sentences and words from the Mexican and Venezuelan collections was re-used to design the prompt texts. Other promps, typically names of persons, cities and companies, had to be localised for the US American setting. For Costa Rica, where no fixed network database from SALA I is available, a new text corpus was generated. All prompt texts were designed to allow double oversampling.
In Argentina, Chile, Mexico and Costa Rica the recording platforms run on Windows 2000, with an AVM ISDN board connected to one basic digital telephone line. ATLAS used ADA software (Rodríguez et al., 1998) to communicate with the ISDN cards using the CAPI protocol. The signals were stored directly in A-law, which is the standard for Latin American telephone networks. In the USA, Eicon hardware was used to connect to two basic digital lines and the signals were stored in Mu-law. In Venezuela, no digital lines are public available; recordings were performed using an analogic Dialogic card to communicate with the PSTN network. ADA software for Dialogic cards was used to control the recordings. The signals were converted and stored in A-law, to maintain the same signal coding as in other Latin American databases.
During recording, the running script played the initial instruction, asking users to say their prompt sheet number, gender, age and other information useful to fully

characterise a call. If problems were detected during the call, the system asked the speaker to restart the call, as the script took no more than 8 minutes. This was never problematic, with the exception of Venezuela, where cellular calls drop frequently due to the poor cellular coverage, and it was necessary to redo several calls to achieve a completed session. Approximately 6500 calls were required to achieve 1100 completed sessions.

One of the most difficult tasks was the speaker recruitment. ATLAS contacted local Universities to perform the recruitment in each country. This strategy has been used in similar collections with excellent results. In the USA, however, this strategy was not successful, as universities, non-profit organizations and Spanish associations experienced enormous difficulties to convince speakers to collaborate in the collection. The problems were mainly attributable to Spanish being a second language, and the large number of Spanish speakers in the USA who deny to provide any personal data as place and date of birth. As a result, the recruitment in the USA was performed by a market research company.

Orthographic transcription is performed in ATLAS offices by welltrained supervisors using RevBD software, which as been probed as an excellent and very efficient tool to transcribe SpeechDat-like databases (Nogeiras, Moreno, 1998).

4. Validation

All SLR produced in SALA II are thoroughly validated as is usual for SLR in collections of the SpeechDat-family. The validation checks address the following aspects of the database:

- documentation
- formal structure and file names
- corpus design
- quality of speech files
- the phonemic lexicon
- orthographic transcription
- speaker distributions
- distribution of recording environments

As compared to the SpeechDat(II) project a small number of extra validation checks were included which are summarised in Moreno et al. (2002).

Validation of each SLR is implemented as a two-stage procedure:

A. Pre-validation of a mini-database of 10 speakers before the main part of the recordings is made. The main objective of pre-validation is to trace and correct design errors at a stage where rectifications can be easily included in the recording scenario.
B. Full validation of the completed SLR. This is the final quality assessment of the SLR in order to ascertain if specifications and other quality criteria were met.

At present all SLR (except for the Canadian ones) have been pre-validated, and the full validations are expected to be completed in May 2004. By the end of February six SLR were successfully validated.

So far, the validation center has not found typical deviations that recur in multiple databases. The only exception is that a number of SLR contains too many tokens of a (small) subset of the phonetically rich word and/or sentences (max. 5 realisations of each word are allowed, not more). In most cases this correlated with the recording of more speakers than required, and thus the deviation could be positively interpreted as provision of additional speech material.

In conclusion, the overall amount and seriousness of the errors was minor. Two reasons can be given for this: first, the specifications in SALA II were built on years of SpeechDat experience and are very clear; second, the production of the SLR is subcontracted to a limited number of companies by the consortium partners, which substantially reduces the probability of making the same error twice.

5. Availability and Outlook

The distribution among the partners follows the general SpeechDat family rule, that each partner has access to the databases of the other partners as soon as his database is accepted by the partners based on the validation report of SPEX. As SALA II is a project that has no public funding, no obligations exist to distribute the databases externally. Some partners of the SALA II consortium will make their databases available for general distribution via ELRA/ELDA.

The SALA II databases (together with the SALA I databases) allow users to train recognisers for narrow band applications covering most language variants over the American continent. As SALA II provides many language variants concerning Spanish as spoken in South and North America, it would be interesting to investigate how the optimal speaker selection should be in these countries in order to achieve equivalent recognition rates for all Spanish dialectal regions. This knowledge could be used if a follow-up SALA II project for creating broadband applications as in SpeeCon were created.

6. References

Heeman, P.A. (2004). The American English SALA-II Data Collection. Proceedings LREC' 2004, Lisbon.

Moreno, A. (2002). The complete SALA II project specifications. Report obtainable via: http://www.sala2.org

Moreno, A., Gedge, O., Heuvel, H. van den, Höge, H., Horbach, S., Martin, P., Pinto, E., Rincon, A., Senia, F., Sukkar, R. (2002). SpeechDat across all America: SALA II Proceedings LREC' 2002, Las Palmas, pp. 1620.

Moreno, A. (1997) Dialectal areas in Latin America for speech recognition applications. SALA technical report. 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 LREC2000, Athens, Greece, pp. 877–882.

Rodríguez Fonollosa, JA, A. Moreno (1998) Automatic Database Acquisition Software for ISDN PC Cards and Analogic Boards. Proceedings LREC1998, Granada, Spain, pp. 1325-1329.