LILA: Cellular Telephone Speech Databases from Asia

Eric Sanders (1), Asuncion Moreno (2), Herbert Tropf (3), Lynette Melnar (4), Nurit Dekel (5), Breanna Gillies (6), Niklas Paulsson (7)
(1)SPEX, The Netherlands (2)UPC, Spain (3)Siemens AG, Germany (4)Motorola, USA (5)NSC, Israel (6)Appen, Australia (7)ELDA, France
E-mail: eric@spex.nl

Abstract

The goal of the LILA project was the collection of speech databases over cellular telephone networks of five languages in three Asian countries. Three languages were recorded in India: Hindi by first language speakers, Hindi by second language speakers and Indian English. Furthermore, Mandarin was recorded in China and Korean in South-Korea. The databases are part of the SpeechDat-family and follow the SpeechDat rules in many respects. All databases have been finished and have passed the validation tests. Both Hindi databases and the Korean database will be available to the public for sale.

1. Introduction

Booming economies in Asia have led to an increasing number of (cellular) telephone users followed by a demand for automated telephone applications. For the development of these applications, spoken language resources (SLR) are needed. To this purpose a consortium called LILA was established, consisting of 5 industrial partners: Siemens AG (Germany), Motorola (USA), Nuance (Belgium), Microsoft (USA) and NSC (Israel). Each partner produced one database following the scheme in Table 1. The creation of the databases was performed by two subcontractors, Appen (Australia) and ELDA (France). The consortium was coordinated by the Technical University of Catalonia, UPC (Spain), and the databases were validated by SPEX (the Netherlands).

Country	Language	Speakers	Producer
India	Hindi as 1st language (Hindi L1)	2000	Siemens
India	Hindi as 2nd language (Hindi L2)	1800	Motorola
India	Indian English	1800	Nuance
China	Mandarin	1800	Microsoft
South Korea	Korean	1000	NSC

Table 1: List of recorded databases.

This project largely followed the SpeechDat-family (www.speechdat.org) of SLR. All SpeechDat databases have a number of things in common, including e.g. database structure, type of recorded items, database exchange schedule, and independent validation procedure.
Since the databases are exchanged against each other, the overall cost per database should be approximately the same for each database. Cost parity is achieved by manipulating the total number of speakers per database.
This paper is a follow up of an LREC2004 paper (Moreno et al., 2004). The current paper focuses on the databases' specifications, language specific issues, and validation results. We conclude with information about availability and follow up plans.

2. Database Specifications

For all databases a fixed set of items was specified; this set can be found in Table 2.

6 application keywords/keyphrases
1 sequence of 10 isolated digits in one utterance
1 sheet number (5+ digits)
3 telephone number (9-11 digits) (2 read, 1 spontaneous)
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 spelling of real/artificial word 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
13 phonetically rich sentences
1 time of day (spontaneous)
1 prompted time phrase, word style i.e. not digital
4 phonetically rich words
1 “silence word” recording

Table 2: Overview of recorded items.

Note that there is considerable overlap with item sets from other SpeechDat (telephony) databases. New for the LILA databases are four more phonetically rich sentences and a "silence word". This silence word is a recording of only background noise for 10 seconds. Producers were allowed to add optional items to their database (and did so).

Care was taken to ensure that the databases were balanced in distribution of speakers. The percentage of speakers for each gender is in the 45-55% region. For age groups 16-30 and 31-45 the minimum percentage of required speakers is 20%; for age group 46-60 the minimum is 15%. Calls were made from five different environments, following the scheme in Table 3.

Environment	**Full database distribution (%)**	**Distribution per dialect region (%)**
Moving vehicle (car, train, bus)	15 ± 5	≥ 20
Public place (restaurant, airport hall)	15 ± 5	
Street	15 ± 5	
Car kit (hands free)	20 ± 1	
Quiet location (home, office)	35 ± 5	≥ 20

Table 3: Environment distribution.

The structure of the database is as follows: each recorded item is in a separate speech file (A-law encoding). Each speech file is accompanied by a label file with the orthographic transcription and information about the speaker and the recording. All speech and label files associated with a particular recording are located in the same directory. Furthermore, each database comes with documentation, a phonemic lexicon and different meta files where the transcription, speaker and recording information of the individual recordings are assembled.

3. Language Specific Issues

In this section, we describe issues specific to particular databases or languages. Issues covered include the number of language speakers, dialect selection, recruitment strategies, deviations from the LILA specifications, and information about the orthographic and Romanised transcription.

3.1 Hindi L1

Hindi is spoken as first language by about 420 million people in India. Most native speakers reside in North-Central India, i.e. the states of Delhi, Uttar Pradesh, Haryana, Chattisgarh, Madhya Pradesh, Rajastan, Bihar, Himachal Pradesh and Uttranchal. The LILA Hindi L1 database comprises 2000 speakers from North-Central India.

There are about 228 million mobile phone subscribers in India, out of which about 80 million in North-Central India.

There are 5 main dialects of Hindi: Western Hindi, Eastern Hindi, Rajasthani, Bihari and the Pahari.

Supervisors for each region were hired to recruit speakers. A snowball effect was obtained as speakers were encouraged to ask friends and/or family members to also participate in the recordings.

In the LILA Hindi L1 database, spelling items are read out slowly instead of being spelled out as the notion of spelling does not apply to Hindi.

Also, the frequency of names for weekdays differs from the specifications since in Hindi there are a total of nine possible names for the seven weekdays. Orthographic transcriptions are written in Devnagari script. The Romanised transcription is in a modified form of INSROT where the vowels have been separated into inherent and non-inherent.

Also, a SAMPA set was developed for Hindi to provide the phonetic representations for the lexicon.

3.2 Hindi L2

Hindi is spoken as a second language in India by approximately 160 million speakers.

Hindi L2 speakers are classified according to their native language into ten dialects: Tamil, Guajarati, Telugu, Malayalam, Kannada, Bengali, Assamese, Punjabi, Marathi and Urdu. Speakers were selected based on Hindi L2 fluency, having studied Hindi as a language up to the secondary level.

An independent market research company was engaged to recruit speakers.

Deviations from LILA specifications are the same as for Hindi L1.

Motorola and Siemens agreed that the Hindi L1 and Hindi L2 databases should harmonize as much as possible; to achieve that objective, sub-contractors Appen and ELDA worked together to select or develop the same orthographic, Romanised, and phonetic transcription strategies.

It was decided not to include the phonemes only found in loan words (namely /s`/, /q/, /x/, /G/, /{/ and /Q/) in the main lexicon but to represent the foreign words with native sounds. These foreign phonemes were included in the two separate lexicons produced for the words of English and foreign (primarily Sanskrit and Arabic) origin. This method of representation provided more comprehensive information than grouping all phonemes together in the main phoneme set. The foreign lexicon for words of English origin used the same phoneme set as the LILA Indian English database.

Although Hindi as a second language has quite a regular spelling/pronunciation correspondence, there are some variations that occur as a result of the influence of the speakers' first language. The use of the nukta warrants some discussion on this point. (N.B. The nukta is a diacritic mark indicating that the character is of foreign origin.) Some of the nukta consonants are pronounced in one way only, e.g. ड़ and ढ़ will always be produced as /r`/ and /r`_h/, respectively. Others, such as क़ (k.), ख़ (kh.), and ग़ (g.) may vary in pronunciation depending on the first language of the speaker – those familiar with Arabic-influenced languages will produce /q/, /x/ and /G/,

while others not familiar with these languages will produce /k/, /K/ and /g/, hence the decision to put these in different lexicons (as discussed above). The nukta consonants फ़ (ph.) and ज़ (j.) are primarily pronounced as /f/ and /z/ by most speakers of Hindi as a second language and can be considered native sounds; however, there is also a proportion of speakers who use /p_h/ and /dZ/ respectively for these. Therefore these have been included as dispreferred variants in the main lexicon. There is little standardisation in the use of nuktas on these particular consonants, so where appropriate in the lexicon, □ (ph) and □ (j) which would normally have the pronunciation /p_h/ and /dZ/ respectively, may also have variation coded in the form of /f/ and /z/.
As there is little standardisation in the spelling of Hindi words within India, considerable work was done to ensure consistent and standard spellings. In particular, the use of diacritics, chandrabindu, anusvara, and word-final halant can be varied and internal rules were developed to ensure that spellings correctly reflecting the pronunciation were used.

3.3 Indian English

India's population is about 1.2 billion people. Of these approximately 90 million speak English, most as a second or third language.
Ten dialect regions were identified based on the native language of the speaker: Hindi/Urdu, Tamil, Guajarati, Telugu, Malayalam, Kannada, Bengali, Assamese, Punjabi and Marathi.
An independent market research company was engaged to recruit speakers. Due to the fact that most speakers of English in India speak English as a second language, the usual practice of recruiting only native speakers was altered to allow speakers of English as a second language to participate.
There were a number of words in the database from Hindi and related languages (or, in the case of names of the month, English words which have become nativised into these languages). To accurately represent the pronunciation of these words, a foreign lexicon was used. The phone set for the foreign lexicon was similar to the Hindi phone set.
Transcribers from across India and from a range of native language backgrounds were employed to transcribe the data, ensuring that any dialect-specific pronunciations were transcribed correctly.

3.4 Korean

Korean is the official language of both North and South Korea. In total about 78 million people speak Korean worldwide, there are 49 million speakers in South Korea and 23 million in North Korea.
The LILA Korean database comprises 1000 speakers from South Korea.
There are about 43.5 million mobile phone subscribers in South Korea, constituting 88% of the population of Korean speakers. Current mobile phone services are CDMA and WCDMA. The Korean database was recorded from the mobile phones network and over an ISDN BRI line.
There are five major dialects in South Korea: Seoul, Chungchong, Kyungsang, Cholla, and Jeju.
The number of speakers recorded per region was proportional to the total population of each region with the exception of Jeju Island, where a minimum number of 100 speakers was recorded.
Supervisors for each region were hired to recruit speakers, distribute prompt sheets, and ensure the recordings, together with the person in charge of the recorded material.
A snowball effect was obtained as speakers were encouraged to ask friends and/or family members to also participate in the recordings.
There are two counting systems in Korean: with Sino-Korean digits and with pure Korean digits. Therefore each of the digit items was divided into two sub-items, to represent the two counting systems used in South Korea.
Orthographic transcriptions are in Hangul. The Yale romanisation system was used and the only modification is that /ng/ is preserved when the character occurs in syllable-initial position.
A SAMPA set was developed to provide the phonetic representation of the phonemes in Korean.

3.5 Chinese Mandarin

Mandarin is spoken in China by approximately 867 million people. Four dialect regions were defined, conforming to the Speecon (Iskra et al., 2002) Mandarin model: Beijing region, Shanghai region, Chongqing region and the North-East provinces region.
An independent market research company was engaged to recruit speakers.
Spelling is not commonly used for Chinese symbols. Chinese speakers spell Chinese names or other characters by describing parts of which a character is composed of. In line with this language specific characteristic, two spelling items deviate from LILA S\specifications. Instead of asking speakers to spell a city or a prompted word, they were asked to read artificial Roman letter sequences.The Roman letters are spelled in the English way. Chinese speakers who don't speak English use Chinese syllables to approximate the English pronunciation.
The standard Romanisation scheme for Mandarin in China is Pinyin and this scheme was used for this collection. Pinyin and its basic tones have been applied in both orthographic transcriptions and the lexicon. In Chinese Mandarin speech, basic tones of some words can however change when they are combined with other words. This phenomenon is called tone sandhi. The most commonly cited example is where tone 3 changes to tone 2 when followed by another tone 3 as given in this example: 你好
(Chinese character set), ni3hao3 (Pinyin), ni2hao3 (Pinyin with tone sandhi).
Therefore in the lexicon an additional column called Harmonized Pinyin has been added, reflecting tones as occurring in speech including tone sandhi. The SAMPA in the lexicon is based on Harmonized Pinyin and shows corresponding tones.

4. Validation

To ensure the quality of the databases, each database went through a validation procedure. Just as the Dutch-based institute SPEX was responsible for the validation of all prior SpeechDat-family projects, they also performed validation of the five LILA databases.

Validation corresponds to three distinct phases of the project. The first validation (prevalidation) occurs after the specifications are completed and actual recording can start. The purpose of this validation is to detect possible errors that cannot be easily repaired once all recordings have started in earnest. In this phase, the prompt sheets are checked against the specifications, the phoneme transcriptions in the lexicon are verified, and a 10 session mini-database is validated.

The second part of validation (full validation) takes place after the complete database has been produced. The following aspects of the database are checked:

- documentation
- formal structure and file names
- corpus design
- quality of speech signals
- phoneme lexicon
- orthographic transcription (by a native speaker)
- speaker distributions

SPEX produces a validation report for the database. If there are deviations from the specifications in the database, the consortium votes whether the database is acceptable for exchange. Errors that are only cosmetic (and hence can be repaired with relative ease) are fixed by the producer. If a database is not accepted, the producer amends the database in such a way that it is acceptable for the consortium and a revalidation takes place.

In the final part of validation (prerelease validation), the database is checked to determine whether the repairable errors are fixed and whether the database is fit for exchange.

All five LILA databases have gone through the validation phases. The two subcontractors employed by LILA partners, Appen and ELDA, have a long history in creating SpeechDat-like databases. This experience leads to few errors in the LILA databases. Only one database contained a shortage of read digits; this was compensated by adding 100 speakers. All databases have been accepted by the consortium.

Language	**Prompts + Lexicon Validation**	**Preval.**	**Full Val.**	**Pre-Release Validation**
Hindi L1	May 06	May 06	Feb 07	Jul 07
Hindi L2	Feb 06	Feb 06	Aug 06	Dec 06
Indian English	Dec 05	Dec 05	Jun 06	Jul 06
Mandarin	Apr 06	Apr 06	Nov 06	Jan 07
Korean	Aug 07	Oct 07	Jan 08	Apr 08

Table 4: Realised time schedule of the different stages for validation.

Table 4 gives an overview of the production time line for the databases. The dates in the table correspond to the dates that each database was received by the validation institute. The validation institute returns their results within a month of delivery. The two major factors contributing to the difference in project start times are defining the design issues and determining the subcontractor to make the recordings. Except for the Korean database, the prompts/lexicon and prevalidation database were delivered together. The time lapse between the prevalidation database and the full database varies between 3 and 9 months with 6 months being fairly typical. The pre-release database follows in a couple of months.

5. Availability + Follow Up

The producing partners exchange their databases only after their individual databases have been accepted by the consortium on the basis of the validation report.

The Hindi L1 and the Korean databases will be available through ELRA. The Hindi L2 database will also become available, but distribution has not yet been decided. For Indian English and Mandarin there are no plans to make these available.

A follow up project, LILA-2, is being set up by Microsoft, Motorola, Nuance, and Siemens AG with Appen as the common subcontractor. In LILA-2, speech databases will be collected in India and other Asian-Pacific areas. In general, these databases will be conform to the LILA specifications. Principal differences in the LILA-2 specifications are the requirements for more spontaneous utterances and recordings made via both the fixed telephone and mobile telephone networks. The first languages to be covered are: Marathi, Kannada, Urdu and Bengali. This consortium is open for additional partners and languages.

6. References

www.speechdat.org

A. Moreno, K. Choukri, P. Hall, H.v.d. Heuvel, E. Sanders, F. Senia, H. Tropf (2004). Collection of SLR in the Asian-Pacific area. In *Proceedings of LREC2004.* Lisbon, pp. 101-104

J. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl (2002). SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation. In *Proceedings LREC2002*. Las Palmas, pp. 329-333

O. Vikas (2005) Multilingualism for Cultural Diversity and Universal Access in Cyberspace: an Asian Perspective. From *www.unesco.org*

INSROT – Indian Script to Roman Transliteration. From *tdil.mit.gov.in/insrot.pdf*

Romanization of Korean. From *www.korean.go.kr*