FRESCO: THE FRENCH TELEPHONE SPEECH DATA COLLECTION - PART OF THE EUROPEAN SPEECHDAT(M) PROJECT

[1] *D. Langmann, R. Haeb-Umbach,*
[2] *L. Boves, E. den Os*

[1] Philips GmbH Forschungslaboratorien, Weißhausstr. 2, D-52066 Aachen, Germany
[2] University of Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands &
Speech Processing Expertise Center (SPEX), P.O. Box 421, 2260 AK Leidschendam, The Netherlands

ABSTRACT

This paper describes the design, collection and postprocessing of the French SpeechDat corpus FRESCO. Being a database of approximately 35,000 utterances recorded from 1000 callers over the terrestrial telephone network in France, it comprises immediately usable and relevant speech for the initial training and assessment of speaker-independent phoneme-model or word-model based speech recognizers, as they are employed in automated telephone services. FRESCO is one of the 1000-speaker telephone speech databases produced as "case studies" within the European project SpeechDat(M).

1. INTRODUCTION

The SpeechDat(M) project has been started in 1994 as a European initiative for providing spoken language resources for voice-interactive telephone services and was funded by the Commission of the EU (LRE 63314). It addresses the fields of production, standardization, evaluation and dissemination of Spoken Language Resources (SLR). As part of this project, the industrial partners have conducted a case study resulting in 1000-speaker telephone speech databases for Danish, English, French, Swiss French, German, Italian, Portuguese and Spanish. The recently launched European Language Resource Association (ELRA) will be in charge of the dissemination. The SpeechDat(M) server provides further information for the interested reader (http://www.icp.grenet.fr/SpeechDat/home.html).

This paper is concerned with 'FRESCO', the French SpeechDat corpus, which has been produced by Philips and SPEX. We describe the database design, the actual collection and the postprocessing of the data. Besides giving statistics of the database a main objective of this paper is to discuss design alternatives and to share with the reader some of the lessons learnt in the course of the database creation process, which is potentially useful information for anyone intending to carry out a similar task.

2. SPEECH DATABASE DESIGN

2.1. Goals and Requirements

The main goal of this speech data collection is to support initial setups of telephone-based speech recognition applications. Here 'initial' means that it is recommended to continue to collect data once the service is operational by recording real-life sessions of the automated telephone service in use. We observed that the data collected in a running service are quite different from what is obtained in 'offline' data collection sessions. Retraining the system with these data can improve success rate very much compared to the initial setup.

To be specific, the main purpose of FRESCO, the French SpeechDat database, is to support the initial development (training and assessment) of speaker-independent phoneme-model or word-model based speech recognition systems. Further, the database allows investigations about the influence of corded versus cordless telephones on the recognition accuracy, and it offers some insight in the influence of read versus spontaneous speech.

Concerning speaker selection, the following rules were set forth

- The speakers should be native French (desirable, but not mandatory)
- Balance with respect to gender
- A minimum of 20% of the calles in each of 3 age categories (16-30, 31-45, 46-60 years)
- The geographical distribution of the calls should roughly reflect the demographics of France.

We did not aim at any repesentative sampling of educational or socioeconomic status, since this was considered far too difficult to implement.

2.2. Database Contents

The design has been motivated by the COCOSDA Polyphone database initiative, which has so far resulted in databases recorded in American English ("Macrophone") [1] American Spanish [2] (both supported by the LDC) and Dutch [3].

The vocabulary of the database contains:

- isolated digits and strings of digits
- spelled words and names
- dates, times and money amounts
- application-oriented words
- sentences providing "phonetic coverage" of the language.

In total there were 37 items, resulting in a mean duration of a recording session well below 10 minutes.

Items	Text	Read	Spontaneous
1	isolated digit	1	
3	connected digits	3	
1	natural number		1
2	money amounts	1	1
3	dates	2	1
2	times of day	2	
4	'yes/no' answers		4
3	spelled words	3	
9	application words	9	
9	sentences for phonetic coverage	9	
37	Total	30	7

Table 1: Text specification.

As can be seen, about 80% of the items are read. Being aware of the dangers of the artificial situation of reading text from a session sheet, this was the only way to assure a sufficient amount of training tokens for the most common words associated with dates, times, numbers, amounts, and spelled items.

A good example of the benefits of spontaneous utterances are the 'yes/no' answers. Questions are asked which the user should affirm or negate, resulting in spontaneous and natureal answers which reflect also the preference for affirmative/negative words (a potentially useful information for language model training). Spontaneous utterances are therefore desirable also for other items, e.g. application words. However, finding the right questions which result in the intended responses, is often very difficult. For the 'yes/no' answers we used the following questions:

Are you ready to start?
Do you use a cordless telephone?
Have you been living abroad for a longer time period?
Is your native language French?

The 3 digit strings were: a 6-digit string (prompting sheet identity number), an 8-digit string presented as a telephone number (with the typical 2-digit grouping of French telephone numbers) and a 16-digit string, introduced as a credit card number.

For each session sheet, the 9 application words had been selected out of a pool of 39. The application words are typical control keywords of automated telephone services.

Nine sentences for phonetic coverage of the language are read by each caller. For each session sheet the set of 9 sentences has been selected such that each speaker utters each (context-independent) phoneme of the language at least two times. There are only a few very rare phonemes for which this rule need not be met: [N], [J], and the undifferentiated phonems [E/], [&/] and [O/]. The sentences have been obtained from the "Le Monde" corpus [4]. Only those sentences had been selected which would fit in no more than two lines. In total, 8000 sentences have been obtained. IDIAP added another 9000 sentences to the pool. By this, we obtained about 1800 different 9-sentence-chunks which adhered to the above phonetic coverage constraint.

3. RECORDING EQUIPMENT

The recordings were done on the same recording platform as was used for the Dutch Polyphone corpus [3]. For the French recordings an international green number was hired. The lines entered the Netherlands digitally. The Dutch PSTN infrastructure guarantees that a speech signal with an ISDN connection as its destination remains in an A-law coded digital form after the first major network switch that it encounters.

The recordings were performed on an OS/2-based PC, type Compaq Prolinea 4/33i, 33 MHz, 250 Mb Hard Disk, 16 Mb RAM. The underlying hardware is a combination of a Rhetorex (RDSP16000) voice board and a ACULAB (1TR6 ISDN-30) telephone interface. Apart from this OS/2-based recording platform, a UNIX computer is used for permanent storage of sampled data via ethernet using NFS. This environment provides 16 independent input lines, from which calls can be recorded and stored.

Each utterance is stored as a headerless sampled data file. The sampled data files contain 8-bit A-law coded PCM signals, sampled at 8000 Hz.

4. SPEECH DATA COLLECTION

4.1. Speaker Recruitment

We recruited Philips' employees by sending personally addressed letters in bundles to contact persons at the different company locations in France. The bundles were sent according to a routing schedule by a Dutch direct mail company. Each letter contained a unique introduction letter and the individual session sheet. The collection was advertised by posters in the Philips outlets. No gift or other financial incentive was offered to the potential callers.

Approaching its employees seems very attractive for a large company. However one should not underestimate the time necessary to set up the contacts! Further, there will almost inevitably be problems with the geographic distribution of the callers, see Section 6.

4.2. Collection Phase

The collection was conducted within 3 months during summer '95 in France. Before the start of the collection, we asked some French partners to try the toll-free number in a one-week test phase. Their recommendations concerning acoustic prompts, the design of prompting sheets, and the adjustment of silence detection were very helpful for fixing last problems and polishing the recording session.

After this test phase, the mailing of the session sheets started. Figure 1 shows the distribution of outgoing letters and incoming calls for the whole period of the collection. The peak of incoming calls after about 1/3 of the collecting period is due to a second poster action, where we again asked people to call. 11 000 prompting sheets were mailed resulting in 1300 calls which is a response rate of 12%. This poor response rate may be typical of

Philips employees. More likely it is, however, due to the fact that no financial incentive whatsoever was offered. About 4% of calls were incomplete. A "complete call" is defined as a call containing at least 36 items.

In order to attain the desired distributions of speakers and uttered text material one has in principle the option between a *'feedback'* and an *'oversampling'* strategy. In the *feedback* method the collection is closely watched. The distributions of the calls received sofar are evaluated and emerging unbalances are counteracted immediately by appropriate measures, e.g. by changing the speaker sampling methodology if a gender unbalance is observed.

We opted for the *oversampling* method. We watched, though, the number of incoming calls, the number of females and males, and, after transliteration, the number of valid utterances in order to sense severe problems at an early stage. However, expected misbalances, e.g. w.r.t. gender, had to be compensated for by collecting a sufficient amount of calls in order to attain the minimum requirements, e.g. the desired minimum number of female speakers., see Section 6.

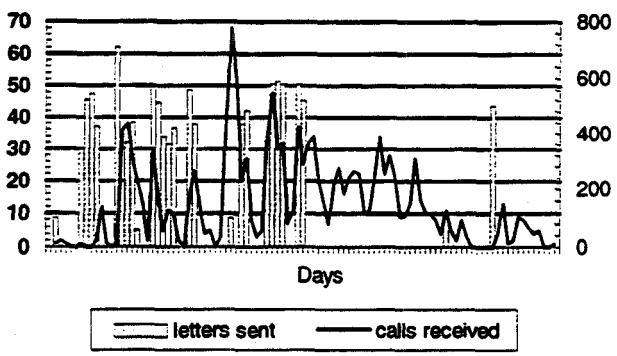

Figure 1: Distribution of mailings and recordings over time.

5. POSTPROCESSING

5.1. Annotation

Annotation consists of

1. Verbatim graphemic transliteration of the prompted speech as spoken indeed.
2. Transcription of extraneous sounds and noises, including their position relative to the words.
3. Assessment of the speech items in terms of signal-to-noise ratio, presence of additional noises, adherence to prompting text, etc.
4. A pronunciation lexicon with graphemic and phonemic forms of all words in the corpus also containing all lexicalised non-standard forms used in the transliteration.
5. A list with all symbols used to transcribe additional sounds/noise.

For transliterations to be useful for automatic training of a speech recogniser, reliable transcriptions are essential. We tried to attain a basic consistency of the transliterations by providing a booklet with transliteration guidelines and giving a short course to the transliterators.

According to SpeechDat(M), the transliterations were stored in files separate from the waveform files.

5.2. Validation

Here, validation is defined as ascertaining that the database meets a specified set of quality requirements in terms of accuracy, completeness and consistency. It has been agreed upon that all databases collected in the SpeechDat project have to undergo a validation stage.

Validation has been applied to the following entities: documentation, speech data files, annotation, pronunciation lexica, and CD-ROM printing and mastering. In most cases the validation procedures can be carried out automatically, i.e. without much human interaction. Those procedures that require a lot of manual work, e.g. listening to speech data files, were reduced to a minimum in order to limit the overall cost. For more information on postprocessing the reader is referred to the SpeechDat web page.

6. DATABASE CHARACTERISTICS

From a total of 1253 complete calls, 521 male speakers and all 479 female speakers were selected for the 1000-speaker database. 200 calls with cordless phones were included (see Table 2).

	Male	Female
Corded Telephone	440	360
Cordless Telephone	81	119

Table 2: Number of calls per gender and telephone type.

As can be seen, an "oversampling" of about 25% was necessary to obtain roughly the required male/female balance. In contrast, the intended age distribution was met without the need for further calls: 28.6% of the callers were between 16 and 30 years old, 35.3% between 31 and 45, and 25.5% between 46 and 60 years.

The geographical distribution of the callers over France is listed in Table 3. It shows a somewhat too large representation of the Paris area, whereas the south is underrepresented.

From the 253 208 spoken words a pronunciation lexicon of 13136 entries were built up. The distribution of the 911438 phonemes is shown in Figure 2. There are 3000 mispronunced words (refering to 1116 lexicon entries). 75 000 non-speech acoustic events have been labeled according to a predefined inventory of 25 events.

Region	Callers
Lyonnais, (Forez)	46
Gascon	32
Massif Central	5
Champagne, Brie	37
Alsacien	15
Bourbonnais, Berry	19
Poiton, Aunis, Angounnois, Saintonge	31
Dauphiné, Savoie	29
Franche-Comté	5
Île-de-France ("Parisien")	435
Bourgogne	55
Lorraine germanophone	15
Limousin	18
Breton	127
Provençal	33
Picardie	73
Languedoc-Méditerranéen	13
Languedoc-Occidental	9
Normandie	139
Lorraine-Romane	33
Corse	1
Other regions	59
Unknown	24

Table 3: Geographical distribution of calls (Total: 1253 calls).

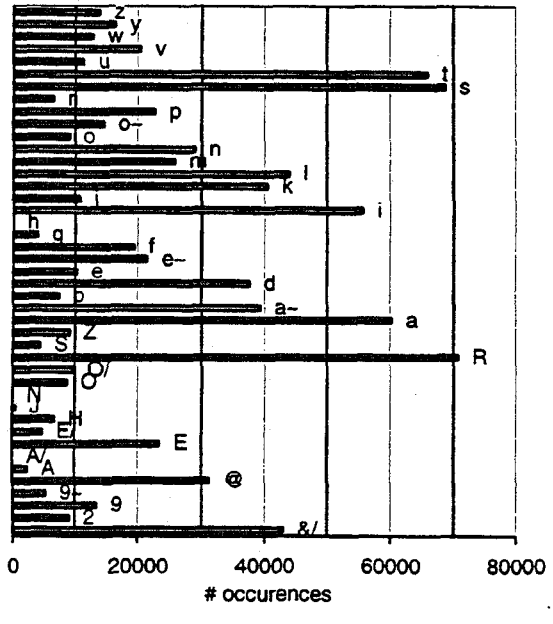

Figure 2: Phoneme distribution of FRESCO (phonemes annotated in SAMPA [5]).

7. OUTLOOK

With the experience of this "case study" we and the other SpeechDat partners are well prepared to approach larger and more complex speech data collections. As a successor of SpeechDat(M) a new SpeechDat project has started in March '96 (LE Reference LE2-4001 10373/0). It aims at more speakers (5000), more languages (all major languages of the EU), and more applications (terrestrial and mobile telephone network, speaker verification). Visit the following web page for more information: http://www.phonetik.uni-muenchen.de/Speechdat.html.

8. ACKNOWLEDGEMENTS

The authors would like to thank D. A. Constantinescu, C. Jaboulet, and G. Chollet from the Institute Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP), Martigny, Switzerland, for their support for the generation of the phonetically rich text material and for providing us with grapheme-to-phoneme conversions.

The various discussions among the SpeechDat(M) project partners were very helpful and are very much appreciated.

The data collection project has been funded by the commission of the European Union under the contract number LRE-6331410072.

8. REFERENCES

1. Bernstein, J.,Taussig, K., Godfrey, J. "MACROPHONE: An American English Telephone Speech Corpus for the Polyphone Project," *in Proc. ICASSP'94*, pp. I81- I84, Adelaide, Australia, 1994.

2. Muthsumay, Y., Holliman, E., and Wheatley, B. "Voice Across Hispanic America: A Telephone Speech Corpus of American Spanish ", *in Proc. ICASSP '95, Detroit, May 1995, pp 85-88.*

3. Damhuis, M., Boogaart, T., in 't Veld, C., Versteijlen, M., Schelvis, W. Bos, L., and Boves, L., "Creation and Analysis of the Dutch Polyphone Corpus," *in Proc. ICSLP'94, Yokohama, Japan, Sept. 1994, pp 1803-1806.*

4. Lamel, L.F., Gauvain, J.L., Eskenazi, M. "BREF, a Large Vocabulary Spoken Corpus for French," *Proc. of Eurospeech, pp. 505-508, Genova, Italy, 1991.*

5. Fourcin, A., Harland, G., Barry, W., Hazan, V. "Speech Input and Output Assessment. Multilingual Methods and Standards," John Wiley & Sons, New York, 1989.