Development of new telephone speech databases for French : the NEOLOGOS Project

Elisabeth Pinto[1] ; Delphine Charlet[4] ; Hélène François[3] ; Djamel Mostefa[5] ; Olivier Boëffard[3] ; Dominique Fohr[6] ; Odile Mella[6] ; Frédéric Bimbot[2] ; Khalid Choukri[5]; Yann Philip[1] ; Francis Charpentier[1]

[1]TELISMA, 9 rue Blaise Pascal 22300 LANNION France epinto@telisma.com
[2]IRISA/CNRS METISS Pièce A 123 Campus Universitaire de Beaulieu 35042 RENNES cedex France
[3]IRISA /ENSSAT, Université de Rennes I, 6 rue de Kerampont, BP447 F-22305 Lannion cedex, France
[4]FTR&D, rue Pierre Marzin 22300 LANNION France
[5]ELDA, 55-57, rue Brillat-Savarin 75013 Paris France
[6]LORIA - Campus Scientifique BP 239 F54506 VANDOEUVRE Cedex, France

Abstract

The NEOLOGOS project is a speech databases creation project for the French language, resulting from a collaboration between French universities and industrial companies, and supported by the French Ministry for Research. The goal of NEOLOGOS is to create new kinds of speech databases: firstly, a 1000 speakers telephone database of children's voices, called PAIDIALOGOS, following the SpeechDat guidelines with some adaptations to the context of children speakers; secondly, a 200 speakers telephone database of adult voices, called IDIOLOGOS, with a new special design to provide adequate data for very fast adaptation techniques and for ASR systems making use of speakers characteristics.

1 The NEOLOGOS Project

The NEOLOGOS project is a speech databases creation project for the French language subsidized by the French ministry for research in the framework of the Technolangues program. Academic laboratories (LORIA and IRISA) and industrial companies (France Telecom, ELDA and TELISMA, coordinator of the project) are collaborating in the field of speech recognition for the creation of two new kinds of speech databases :

- a SpeechDat-like speech database for children's voices (PAIDIALOGOS sub-project);
- a speech database with a novel kind of structure for adult voices (IDIOLOGOS sub-project).

In both subprojects the goal is to bring to the research community new sources of telephone speech data likely to improve ASR performance : on one hand, to significantly improve speech recognition for children (with PAIDIALOGOS), on the other hand to provide speech data to support the development of advanced ASR techniques such as eigenvoices (with IDIOLOGOS). IDIOLOGOS should also provide the means of advanced studies on speakers characteristics, with a significant panel of reference speakers, including in the area of speech synthesis and speaker identification.

2 PAIDIALOGOS : a children 1000 speakers speech database

Today children voices are not represented well enough in publicly available speech databases. Consequently ASR systems perform significantly worse for children voices than for adult voices (Potiamos et al, 1997) (Stemmer et al, 2003). An early database, available through the LDC, was created for English for non telephone speech and for a small number of speakers in the context of the LISTEN project (Eskenazi et al, 1997). The largest SpeechDat databases, available through ELRA (www.elda.fr), contain only small proportions of speakers under 16 years old: about 200 for the French and German databases, about 150 for the Italian database, only 40 for the Spanish database and none for the British English database. Concerning non telephone speech, the recording of 50 children has been planned for each language of a large set of 20 languages in the SPEECON project (Iskra et al, 2002), and a limited corpus has also been recorded to study the recognition of children speech for Swedish in the PF-Star project (Blomberg and Elenius , 2003).

To overcome such a shortage of data in the case of French, our goal in the PAIDIALOGOS sub-project is simply to create a French SpeechDat-like database dedicated to children voices.

1000 children between 7 and 16 years old (included) are recorded over the fixed telephone network, in a relatively quiet environment. The database is evenly split between boys and girls and evenly balanced across twelve French regions. As for the balance of ages, the range between 7 and 11 is emphasized as can be seen in the following table which defines the minimum of number of speakers in a given age range, both for the whole database and for any of the twelve regions defined for regional accents:

Age range	Minimum number of speakers	Minimum number of speakers per region
7-11	500	30
12-14	250	15
15-16	150	9

Table 1. PAIDIALOGOS speakers age distribution constraints

The linguistic items will be recorded directly from calls made from fixed telephones, consisting of 37 items that are either read, or repeated or that correspond to spontaneous answers to specific questions.

The following table summarizes the contents of each call:

Corpus contents
4 application words
3 sequences of 3 isolated digits
1 sheet number
1 telephone number (10 digits)
1 spontaneous date, e.g. birthday
1 prompted date, word style
1 relative and general date exp.
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 (spontaneous)
1 most frequent cities
1 "forename surname"
2 predominantly "yes" questions
2 predominantly "no" questions
6 short phonetically rich sentences (repeated)
1 time of day (spontaneous)
1 time phrase (word style)
2 phonetically rich words

Table 2. PAIDIALOGOS corpus contents

In order to obtain good quality recordings with children under ten, some adaptations are brought to the standard approach used for the SpeechDat databases. The linguistic content is simplified : sequence of numbers are shortened, and so are the phonetically rich sentences which are also chosen with meanings that are easy to grasp by the youngest (e.g. "il est assis par terre" for "he is sitting on the floor"). Also, the recording mode with prompted speech to be repeated (the "repetition mode") is introduced for the number sequences and the phonetically rich sentences. Consequently the number of phoneme occurrences should be smaller than for an adult SpeechDat database but we believe this is a necessary constraint so as to obtain good recordings of children.

3 IDIOLOGOS : a 200 adult reference speakers speech database

3.1 The reference speakers database approach

The IDIOLOGOS speech database must enable the accurate speaker-dependent modeling of a significant set of speakers, called reference speakers. Consequently its design is significantly different from the classical SpeechDat databases already developed for many languages. The objective is to collect significant quantities of "speaker-dependent data", for a significant number of speakers, as was done for several databases oriented towards speaker verification such as (Asham and Wheatley, 1999), but with the following differences:

- We need to maximize the coverage of the space of all speakers;
- The voice of any recorded reference speaker must vary as little as possible.

The IDIOLOGOS database is created in three successive steps :

- The collection of a "bootstrap database" : a first set of 1000 different speakers are recorded over the fixed telephone network; these "bootstrap speakers" record a set of phonetically balanced sentences identical for all speakers; such sentences are optimized to facilitate the comparison of speaker characteristics between the "bootstrap speakers";
- A subset of 200 reference speakers are selected through a comparison of the voice characteristics of the 1000 "bootstrap speakers";
- The final collection of "reference speakers" database : the 200 reference speakers are requested to read and pronounce a large corpus of 450 phonetically rich sentences, also identical for all speakers, in 10 successive telephone calls that must be completed in a short period of time to avoid shifts of the voice characteristics.

We also call the reference speakers database the "eigenspeakers database", in a slightly improper way because these speakers are real and not mathematical objects, but we indulge into this because such data will be very useful to create well trained eigenvoice models according to one of the leading techniques for very fast speaker adaptation (Kuhn et al, 2000). More generally, we expect the "eigenspeakers database" to provide very useful data for improving the performance of ASR systems through any of the very fast adaptation techniques.

3.2 Speakers distribution in the bootstrap database

As for SpeechDat databases, the bootstrap database is balanced across gender, regional and age characteristics. As for PAIDIALOGOS, twelve French geographic areas are used, corresponding to a finer representation than used in previous French databases. Also elderly speakers (60 and more) are better represented than in other databases, since we use in the same proportion of elderly people than for the three other age ranges, as can be seen in the following table :

Age range	Minimum number of speakers	Minimum number of speakers per gender
17-30	200	50
31-45	200	50
46-60	200	50
60 and more	200	50

Table 3. IDIOLOGOS speakers age distribution constraints

3.3 Corpus Design

Two text corpora were designed to meet the requirements of the IDIOLOGOS "bootstrap database" and "eigenspeakers database". Both corpora are essentially composed of sets of phonetically rich sentences, which are fixed sets and do not depend on the speakers ID (note that this is a major difference with SpeechDat corpora).

For the bootstrap corpus, a small set of number-based and letter-based utterances (1 PIN-code, 1 Telephone number, 1 Credit card number, 2 spelled items) have been added

for the purpose of additional control tests to check the validity of the IDIOLOGOS approach. The rest of the corpus consist of a fixed set of 45 phonetically rich sentences, containing approximately 1800 phone occurrences. In order to produce a stable and consistent pronunciation, the sentences are semantically natural and they contain between 5 and 15 words each. This is the set sentences which will be used to extract from the 1000 bootstrap speakers the final panel of 200 reference speakers.

As mentioned above, the eigenspeakers database consists exclusively of a large fixed set of 450 phonetically rich sentences to be recorded by each speakers rapidly in a sequence of 10 calls of 45 utterances each.

Both corpora of phonetically rich sentences were constructed by processing and simplifying sentences from large publicly available newspaper corpora in French. Automatic corpora reduction methods such as the greedy algorithm reported in (François and Boeffard, 2002) were used to extract a subset of sentences meeting a criterion of minimal representation of all phonemes as well as a criterion of minimum representation of diphone classes. There were 99 diphone classes constructed from 10 broad phonetic classes including the silence.

3.4 Reference speaker selection

As mentioned above, the goal is to select 200 reference speakers among the 1000 bootstrap speakers. A particular attention is paid to the selection of such reference speakers, as at the end recognition models will only be built on these speakers. The question is: how can we select out of 1000 speakers the 200 that represent the best the 1000 speakers? This question can be divided into two sub-questions:

- What is the criterion to decide whether a speaker is a good reference speaker?
- For a given criterion, what is the method to select the reference speakers?

At the beginning of the project, it was clear that we can not decide of the criterion a priori. Hence, it was decided to elaborate a methodology that enables to evaluate and compare different criteria and the reference speakers they select. Then, the task of finding reference speakers was reformulated so as to be divided into two phases:

- Phase 1 : evaluate and compare various criteria for reference speakers selection
- Phase 2 : choose the two best criteria and make the speaker selection according to the two best criteria

We present in detail Phase 1, that concerns the methodology of selection of the reference speakers for a given criterion, and the evaluation of a set of reference speakers for a given criterion

3.4.1 Phase 1: Methodology of selection.

The requirements for a given criterion are the following:

- It should give the set of the selected reference speakers
- It should give a measure of the "quality" of a set of reference speakers selected by another criterion.

Hence, we need to define a measure of "quality" for a given criterion. With this measure of "quality", the selection of the reference speakers becomes an optimization problem, the function to be optimized being the measure of quality. Moreover, this measure of quality enables to compare different criteria, by answering the question: how good are the reference speakers selected according to criterion A when evaluated according to criterion B ?

As the reference speakers are supposed to be the most representative ones, whatever the criterion, our methodology is based on a formalism of a dissimilarity measure between a pair of speakers. Then, we set:

- $d_A(x_i, x_j)$: quantifies the loss of quality of the speaker selection when replacing model of speaker x_i by model of speaker x_j, according to criterion A. It can be seen as a non-symmetric dissimilarity between x_i and x_j
- M=1000 is the number of initial speakers, N=200 is the number of selected speakers.
- $\left\{L_j^A\right\}_{j=1,..,N}$ are the reference speakers selected according to criterion A.
- $ref_A(x_i / L^B) = \arg\min_{j=1,..,N} d_A(x_i, L_j^B)$ is the reference speaker for speaker x_i chosen according to criterion A , among the L^B possible reference speakers.

The quality of a set of reference speakers for a criterion A is then defined as:

$$Q_A(L^A) = \sum_{i=1}^{M} d_A(x_i, ref_A(x_i, L^A))$$

It corresponds to the "loss of quality" when we replace the M initial speakers with the N selected speakers. Finding the reference speakers is then an optimization problem to minimize the "loss of quality":

$$L^A = \arg\min Q_A(L)$$

We can use heuristic methods to determine the reference speakers, and in practice, we use for instance hierarchical clustering.

To evaluate how a set of reference speakers selected according to criterion A is a good set of reference speakers according to criterion B, we measure :

$$Q_B(L^A) = \sum_{i=1}^{M} d_B(x_i, ref_B(x_i, L^A))$$

As the database is mainly speech recognition oriented, the criteria will be also mainly speech recognition oriented.

Several relevant criteria will be studied including the following :

- a criterion based on a distance between GMM of speakers
- a criterion based on the likelihood of HMM phone models for the speakers
- a criterion based on DTW between the speakers (that have pronounced the same items)

3.4.2 Phase 2: Final speaker selection

Having in the first phase selected and compared the reference speakers according to the various criteria, we will in the second phase choose the two criteria that appear to be the most consistent with each other. Then, we will select the 200 final reference speakers according to a mixture of the chosen criteria, and then contact the selected speakers to launch the recording of the "eigenspeakers database". Should any one reference speaker not be capable of completing his "eigenspeaker" recordings, a second best reference speaker would be chosen to replace him by using the chosen mixture of criteria.

4 Standards and quality

The PAIDIALOGOS and IDIOLOGOS databases will be produced in the classical SAM format as previously used for the SpeechDat projects family (SpeechDat II, SALA and SALA2, SpeeCon and OrienTel).

The validation forms an integral part of the production of a language resource, to ensure of the quality of the resource. Spoken language resource validation refers to the quality evaluation of a database against a checklist of relevant criteria. The validation criteria are similar to those used in the SpeechDat projects family : they cover documentation, formal and technical criteria, completeness of the database, file formats, signal quality, transcription quality, lexicon and speaker distribution. Some specific criteria corresponds to the specificities of the databases (simplified linguistic content for PAIDIALOGOS, development of speaker selection criteria for IDIOLOGOS).

5 Concluding remarks

Both the PAIDIALOGOS and IDIOLOGOS databases are planned to be completed by the end of 2004. As soon as they are complete, both PAIDIALOGOS and IDIOLOGOS database will be used for testing and evaluating the ASR performance gain, respectively for French children's telephone speech and for French adults' telephone speech.

Eventually, both databases will be made publicly available through the ELRA distribution channel.

6 References

Asham S.R. and Wheatley S.J., "SpeechDat British English database database for speaker verification". SpeechDat(II) British English database SDB-2400, ELRA Catalogue, corpus n° S0098, Database description document. www.elda.fr,

Blomberg M. and Elenius D., (2003) Collection and recognition of children's speech in the PF-Star project, PHONUM 9 pp.81-84 [5]

Eskenazi M. et al, The CMU Kids corpus, corpus n° LDC97S63, Linguistic Data Consortium Catalog, www.ldc.upenn.edu/Catalog

François H. and Boëffard O (2002). The greedy algorithm and its application to the construction of a continuous speech database. Proc. of LREC 2002, paper n° 265, pp.1420-1426.

Iskra D. et al (2002). SPEECON – Speech databases for consumer devices : database specification and validation. Proc. of LREC 2002 , paper n° 177, pp.329-333.

Kuhn R, Junqua J.-C., Nguyen P. and Niedzielski (2000), Rapid speaker adaptation in eigenvoice space, IEEE Trans. Speech Audio Processing, Vol.8, N° 6, pp.695-707, Nov. 2000

Potiamos A. et al, (1997) Automatic Speech Recognition for children, Proc. EUROSPEECH 1997, paper n°183

Stemmer, G. et al, (2003) Acoustic Normalization of Children's Speech, Proc. of EUROSPEECH 2003, Conference paper n°1161, pp.1313-1316

ELRA Catalogue. Large telephone databases : French SpeechDat(II) FDB-5000, German SpeechDat(II) FDB-4000, Spanish SpeechDat(II) FDB-4000, Italian SpeechDat (II) FDB-3000, British English SpeechDat (II) FDB-4000, www.elda.fr,