New telephone speech databases for French: a children database and an optimized adult corpus

Djamel MOSTEFA(1) **and Arnaud VALLEE**(2)

(1) Evaluations and Language resources Distribution Agency
55-57, rue Brillat Savarin
75013 Paris
http://www.elda.org

(2) Telisma
9, rue Blaise Pascal
22300 Lannion
http://www.telisma.com

Abstract

This paper presents the results of the NEOLOGOS project: a children database and an optimized adult database for the French language. A new approach was adopted for the collection of the adult database in order to enable the development of new algorithms in the field of speech processing (study of speaker characteristics, speakers similarity, speaker selection algorithms,...) The objective here was to define and to carry out a new methodology for collecting significant quantities of speaker dependent data, for a significant number of speakers, as was done for several databases oriented towards speaker verification, but with the additional constraint of maximising the coverage of the space of all speakers. The children database is made of 1,000 sessions recorded by children between 7 and 16 years old. Both speach databases are SpeehDat compliant meaning that they can be easily used for research and development in the field of speech technology.

1. Introduction

The *NEOLOGOS* project (Pinto et al., 2004) was a speech databases creation project for the French language, resulting from a collaboration in the speech recognition field between French universities and industrial companies, and subsidised by the French ministry for research. This paper presents the results of *NEOLOGOS*: first, a large telephone database, called *PAIDIALOGOS*, of children's voices following the SpeechDat guidelines while adapting them to the context of children speakers ; second, an extensive telephone database, called *IDIOLOGOS*, of grown up voices, with a special design in order to provide significantly more data than classical databases in order to develop systems making extensive use of information speakers characteristics.

Today, children are not well enough represented in existing publicly available speech databases, and this limits the ability to design speech bases products and services targeted to children. The *PAIDIALOGOS* speech database consists of the recordings of one thousand phone calls from different children using the PSTN telephone network. The selected sample of speakers is composed of girls and boys in equal proportion; it covers each all ages from seven through sixteen and it provides a relatively detailed coverage of the dialectal areas of France. The linguistic contents of the collected speech is composed of thirty-seven utterances including application words, sequences of digits and numbers, dates and time, spelled words and names, city names, and phonetically balanced sentences and words. In order to record children's speech with good quality, adaptations had to be brought to the standard procedures such as the ones used in the SpeechDat databases: the linguistic content had to be simplified and a recording mode with repeated speech was partially introduced, in addition to the classical read and spontaneous speech procedures. The *PAIDIALOGOS* database is one of the first children speech databases designed as such.

The *IDIOLOGOS* speech database, targeted for-detailed speaker dependent modeling of a significant set of speakers, is significantly different from the classical SpeechDat databases developed for many languages. The objective here was to define and to carry out a new methodology for collecting significant quantities of speaker dependent data, for a significant number of speakers, as was done for several databases oriented towards speaker verification, but with the additional constraint of maximising the coverage of the space of all speakers. The speakers recorded in the database were selected to serve as reference speakers for their voice characteristics.

We also call the reference speakers of the *IDIOLOGOS* database *eigenspeakers*, in a slightly improper way, because such data will be very useful to create *eigenvoice* models (Kuhn et al., 2000). The use of such data should help in improving the performance of Automatic Speech Recognition (ASR) systems.

2. The adult Idiologos database

2.1. General goals

Automatic speech processing techniques require large and therefore expensive speech databases. This is true for speech recognition, speaker identification or verification, text-to-speech synthesis,...When collecting a speech database, two criteria are usually considered. The first one defines constraints on the corpus to be recorded. A good linguistic coverage is achieved by building a text corpus with phonetically rich words and sentences. The second criteria is on speakers profiles. Usually constraints are defined in terms of sex, regional accent and age. But all these constraints are considered *a priori*. For example a speaker living in one region is considered as having a specific accent even if it is not true. But somehow this way

of collecting database mismatches the most recent development is speech processing. Automatic speech recognition algorithms tend to make use of specialized speaker models instead of a unique general model (Kuhn et al., 2000). Speaker models have been recently used in speaker recognition (Sturim et al., 2001; Collet et al., 2005). Text-to-speech synthesis require the availability of a wide range of speakers for improvements of the overall quality. These new techniques require a important quantity of speech material per speaker. We propose here a two steps data collection design in which, for the first step a important number of speakers are recorded, then a subset of representative speakers are selected *a posteriori* and asked to record much more speech material.
The *IDIOLOGOS* database was created in three successive steps:

1. a preliminary collection of an initial set of one thousand different speakers (bootstrap database), selected on gender, regional and age characteristics, speaking a set of phonetically balanced sentences, carefully optimised to facilitate the comparison of speaker characteristics ;

2. the identification of a subset of 200 reference speakers through a comparison of the voice characteristics of the initial 1,000 speakers;

3. the recording of large quantities of speech data for the identified set of the selected two hundred reference speakers.

2.2. Preliminary collection of 1,000 speakers: the bootstrap database

This database is made of recordings of 1,000 speakers from the fixed telephone network. Each speaker recorded 50 items as depicted in table 1.

1 sequence of digits
1 telephone number
1 credit card number
2 sequences of spelled letters (natural and artificial)
45 phonetically rich sentences

Table 1: Recorded items for the Bootstrap database

While the sequence of digits and letters are different from one speaker to another in order to have enough samples of each digit or letter, the set of 45 phonetically rich sentences is identical for all speakers and optimized to facilitate the comparison of the speaker characteristics. The sentences were created from large publicly available newspaper by applying automatic corpora reduction methods in order to satisfy a criterion of minimal representation of all the phonemes and diphone classes (Francois and Boeffard, 2001). In order to produce consistent pronouncing, sentences are semantically natural and contain from 5 to 15 words.
For each speaker, the 50 items were recorded in one call. Since these 1,000 speakers were used to select 200 representative ones, it was important to have a good distribution in terms of sex, accent and age. France was divided in 12 regions, according to an expert in French language and its dialects. The database is balanced across gender, age and regional characteristics (accents). While elderly speakers (60 and more) are usually considered as optional, they were mandatory in this data collection and represents a proportion equal to that of the other age ranges (18-30, 30-45, 45-60).
The database is transcribed according to the SpeechDat conventions (Senia and van Helden, 1997). The transcription is orthographic and includes noise markers. Since all items are read, transcriptions were generated in a semi-automatic way. The expected transcription was automatically produced and the listener had to correct it if needed.

2.3. Identification of 200 reference speakers

The extraction or selection of the N (N=200) reference speakers from M (M=1000) initial speakers has been done through clustering techniques. The aim is to select a representative list L of N speakers which minimizes the total loss or distances with the initial speakers. If $ref(x_i)$ is the representative for speaker x_i, we want to minimize the following cost function:

$$Q(L) = \sum_{i=1}^{M} dist(x_i, ref(x_i))$$

where $dist(x_i, ref(x_i)$ is a distance between x_i and $ref(x_i)$.
The number of different lists of N speakers across M ones is $C_M^N = \frac{M!}{N!(M-N)!}$ which is a huge number. Therefore it's not possible to test each list and find the optimal list of speakers, but methods of Hierarchical Clustering or K-means can find a local optimal solutions.
The voice space had been partitioned into homogeneous subspaces and then a representative speaker was selected for each subspace. Four kind of distances were used for the partition: Canonical Vowels, Dynamic Time Warping, Gaussian Mixture Models and HMM affiliated phone models. The optimization was done in order to keep a diversity of voices while pruning the number of speakers. The method is detailed in (Charlet et al., 2005; Krstulovic et al., 2005; Krstulovic et al., 2006). The selection of the 200 reference speakers had to be updated when a selected speaker was not available anymore for making new recordings.
The 200 reference speakers are made of 103 females and 97 males. Figures 1 and 2 show the distribution in terms of age and dialect of the bootstrap database (1,000 speakers) and the 200 reference speakers. Thus, we can observe that the automatic selection algorithm has kept more or less the equal distribution over sex, age groups and dialectal regions.

2.4. Collection of large quantities of speech data for the selected reference speakers: the Idiologos eigenspeakers database

After the selection of 200 reference speakers was done, each selected speaker was asked to record 10 more sessions of 50 items each. Thus each speaker uttered the items listed in table 2.

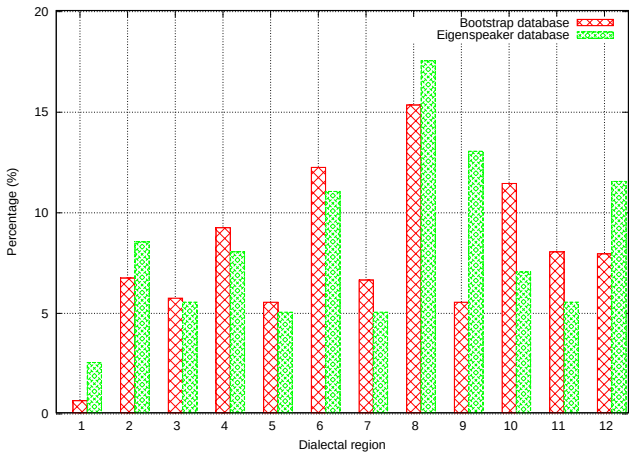

Figure 1: Distribution of speakers over dialectal regions for the 1,000 speakers database and the 200 reference speakers database

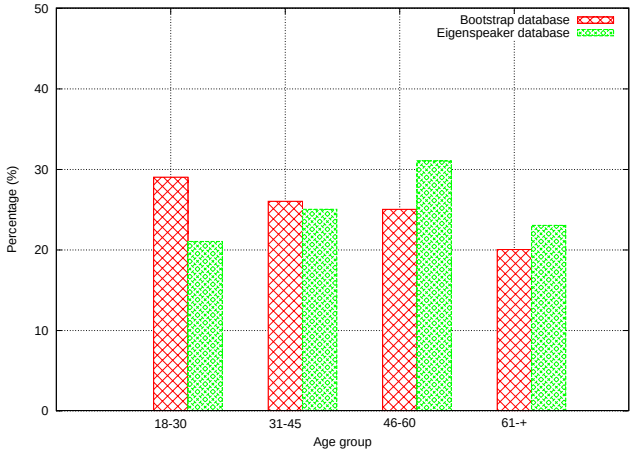

Figure 2: Distribution of speakers over age groups for the 1,000 speakers database and the 200 reference speakers database

In order to have enough material per speaker, specifications planed to have 500 sentences per speaker. The Eigenspeakers database contains 450 sentences for around 20,000 phones. In order to produce consistent pronouncing, sentences are semantically natural and contain from 5 to 15 words. Machine learning requirements result in a minimum of 50 representatives per phoneme in any configuration of pronouncing. Table 3 gives the number of occurrences per SAMPA phoneme for French in the whole database. As we can see, the database is very rich phonetically with a minimum of 2387 occurrences for the /N/ nasal phoneme found in loanwords (*camping* /ka∼piN/) and a maximum of 357173 occurrences for the liquid consonant /R/ phoneme

10 sequences of digits
10 telephone numbers
10 credit card numbers
20 sequences of letters
450 phonetically rich sentences

Table 2: Recorded items for the Eigenspeakers database

(*rond* /Ro∼/). Moreover the optimization was also made in terms of diphone classes and all the diphones are covered at least once.

French Sampa	Freq.	French Sampa	Freq.
2	33432	&	5945
9	29363	9∼	30865
@	96151	a∼	154431
a	336763	A	3395
b	53240	d	177842
e	177473	E	181435
E	63743	e∼	70955
f	66323	g	37346
H	28674	i	260798
J	7385	j	97754
k	174333	l	264864
m	123981	n	126077
N	2387	O	40468
O	62440	o	67429
o∼	79054	p	137414
R	357173	s	298585
S	30714	t	249021
u	69689	v	91568
w	43200	y	92818
Z	52792	z	61414

Table 3: Phoneme frequency for the eigenspeakers database

The database is transcribed with the same conventions and methodology as for the bootstrap database (see section 2.2.). The Eigenspeakers database has been recorded one year after the Bootstrap database and therefore can be used to analyse changes in voices over the time for a sufficient number of speakers.

3. The children Paidialogos database

This database is a standard SpeechDat-like database, except the fact that only children voices are included in the corpus. The database contains 37 364 utterances from 1 010 different children, who are between 7 and 16 years old. The number of items present in each prompt sheet is of 37 items. Prompt sheets were generated before the collect phase in a first stage and generated to cover the missing items in a second stage. To minimize confusion and allow for a smooth recording process, instructions, questions and read items were grouped in different sections. Two questions requiring responses of similar types can be found in a row. List effect have not been taken into account: prompt items are grouped by type (numbers, words, spelling. . .)
Each child recorded 37 items as depicted in table 4.
Of course the corpus was adapted to the public of children. The sentences were shortened and carefully selected to make sure that a 7 years old child can read it without difficulty. 683 sentences were taken from a corpus given by France Telecom R&D for the project. Sentences were repeated by the speaker. As the sentence are pronounced by children, small sentences were used. As a result, minimum sentence length is two words, and maximum sentence length is five words.

4 application words
3 sequences of digits
1 PIN code number
1 telephone number
1 spontaneous date (birthday)
1 prompted date
word style
1 relative and general date expression
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 4: Recorded items for the children Paidialogos speech corpus

A phonetic lexicon of the words occurring in the corpus was prepared, and this was used to calculate the frequency with which each phoneme was occurring. Sentences were edited/modified to include additional tokens of phonemes identified as less frequently occurring.
No sentence was repeated more than 13 times.
The database is balanced across sex, age and dialectal regions. As for the adult database, we used 12 dialectal regions and recruited kids in each region. Three age classes were considered: 7-11, 12-14, 15-16.
The database is orthographically transcribed and packaged according to SpeechDat specifications.

4. Conclusions

This paper presents the results of the NEOLOGOS project: a children database and an optimized adult database for the French language. Both speech databases are SpeehDat compliant meaning that they include recordings with the orthographic transcription, phonetic lexicon, statistics,.... These databases can be easily used for research and development in the field of speech technology. Moreover, the adult database has been optimized in terms of corpus and speaker selection and can be used for the study of speaker characteristics, speakers similarity, speaker selection algorithms,... The databases have been successfully validated by an external validation center and are publicly available through ELRA's catalog of language resources [1]. These database collections have been sponsored by the French Ministry of Research under the TECHNOLANGUE action [2].

[1] http://catalag.elra.info

[2] http://www.technolangue.net

5. References

D. Charlet, S. Krstulovic, F. Bimbot, O. Boeffard, D. Fohr andO. Mella, F. Korkmazsky, D. Mostefa, K. Choukri, and A. Vallée. 2005. Neologos: an optimized database for the development of new speech processing algorithms. In *EuroSpeech'05*, Lisboa.

M. Collet, Y. Mami, D. Charlet, and F. Bimbot. 2005. Probabilistic anchor models approach for speaker verification. In *EuroSpeech'05*, Lisboa.

H. Francois and O. Boeffard. 2001. Design of an optimal continuous speech database for text-to-speech synthesis considered as set covering problem. In *EuroSpeech'01*.

S. Krstulovic, F. Bimbot, D. Charlet, and O. Boeffard. 2005. Focal speakers: a speaker selection method able to deal with heterogeneous similarity criteria. In *EuroSpeech'05*, Lisboa.

S. Krstulovic, F. Bimbot, O. Boeffard, D. Charlet, D. Fohr, and O. Mella. 2006. Optimizing the coverage of a speech database through a selection of representative speaker recordings. *Speech Communications*, 48:1319–1348.

R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski. 2000. Rapid speaker adaptation in eigenvoice space. *Speech Audio Processing*, 8 (6):695–707.

E. Pinto, D. Charlet, H. Francois, D. Mostefa, O. Boeffard, D. Fohr, O. Mella, F. Bimbot, K. Choukri, Y. Philip, and F. Charpentier. 2004. Development of new telephone speech databases for french : the neologos project. In *Language Resources Evaluation Conference*, Lisboa.

F. Senia and J.G van Helden. 1997. Specification of orthographic transcription and lexicon conventions speechdat technical report sd1.3.2. Technical report, CSELT.

D. Sturim, D. Reynolds, E. Singer, and J. Campbell. 2001. Speaker indexing in large audio databases using anchor models. In *ICASSP'01*, Salt Lake City.