A NEW METHODOLOGY FOR SPEECH CORPORA DEFINITION FROM INTERNET DOCUMENTS

D. Vaufreydaz, C. Bergamini, J.F. Serignat, L. Besacier, M. Akbar

Laboratoire CLIPS-IMAG, équipe GEOD
Université Joseph Fourier, Campus scientifique
B.P. 53, 38041 Grenoble cedex 9, France
(Dominique.Vaufreydaz, Carole.Bergamini, Jean-Francois.Serignat, Laurent.Besacier, Mohammad.Akbar)@imag.fr

ABSTRACT

In this paper, a new methodology for speech corpora definition from internet documents is described, in order to record a large speech database, dedicated to the training and testing of acoustic models for speech recognition. In the first section, the Web robot which is in charge of collecting Web pages from Internet is presented, then the web text to French sentences filtering mechanism is explained. Some information about the corpus organization (90% for training and 10% for test) is given. In the third section, the phoneme distribution of the corpus is presented and comparison is made with others French language studies. Finally tools and planning for recording the speech database with more than one hundred speakers are described.

1. INTRODUCTION

Nowadays, many people can access the Internet, either from work, school or home. This growing population is not only passive by consulting existing documents on the Web pages, news servers and chat sessions ; they participate actively on the Internet by creating, publishing and/or synthesizing contents. Depending on the context (professional, personal, educational…) these documents are of very different nature. People generally use simplified vocabulary and ungrammatical expressions as they do in everyday life. This means that by using the documents publicly available on the Internet we can obtain a very large corpus that is a mixture of well-written text and of free text more representative of what can be said in spontaneous speech. These internet based corpora are very interesting to handle different tasks :

- train language models more appropriate in the context of dialog systems and/or spontaneous speech recognition.
- get a large amount of sentences for large speech database recording.

The first task related to Language Models (LMs) was discussed in a former paper presented in last ASRU conference (Vaufreydaz, 1999). It was shown how the use of Internet, to automatically prepare LMs adapted to a given task, can lead to a word accuracy up to 15% better than a system using LMs trained on written text.

This paper is rather dedicated to the second task since we used sentences captured on the Web to record a large speech database. In section 2, we describe how using our indexing engines and appropriate filters we collected a very huge set of French documents that could directly be used either in LM learning or for database recording. Treatments specific to data cleaning for large speech database recording are also described in this section. Section 3 is dedicated to evaluate the adequacy of the resulting corpus to represent phonemes distribution in the French language. Details on database recording environment are given in section 4. Finally, in section 5 we draw some conclusions.

2. DATA ANALYSIS

2.1 Gathering Internet documents

As we already mentioned, a former paper (Vaufreydaz, 1999) presents document collection methods from Internet. In this section, we only discuss important points in order to describe the source corpus of our work.

In collaboration with MRIM[1], another team of our laboratory, a Web robot which is in charge of collecting Web pages from the Internet has been developed. This bot is called *CLIPS-Index*. It is an RFC-2068 compliant robot that respects privacy of documents. It takes one or several starting points on the Web and finds all pages and text documents it can reach from there. It filters out documents according to their types (separating html and text documents from others like images, audio files, etc…) and/or the name of the Web server.

CLIPS-Index provides a good way to collect Web data quickly. During February 1999, we performed a first collection of Web data (HTML and text): *WebFr*. This corpus contains more than 1,550,000 different documents amassed in 84 hours. They were found on more than 20000 servers from the French domain '.fr'. *WebFr* is a collection of about 10 gigabytes of HTML and text documents.

2.2 Text corpus generation for speech database

Data extracted from the Web is not in a suitable textual form to be presented directly to a speaker for recording speech. Thus, we use a set of filters in order to get the text in the appropriate form. Moreover, to increase quality of the resulting speech corpora, we conduct

[1] see MRIM Web page at http://www-clips.imag.fr/mrim/

several analysis to check phonetic distribution in the corpora.

2.2.1 From Web to French sentences

In Web documents, headers, tags and other diacritics are superfluous and must be removed. Moreover, all the documents in *WebFr* are not in French language (documents can be English, German or in a multilingual form). However the selection of *.fr* domain helped us to reduce the chance of gathering texts written in other languages to a minimum. So the first step was to filter out the documents. Figure 1 presents the filter composition used in this first step.

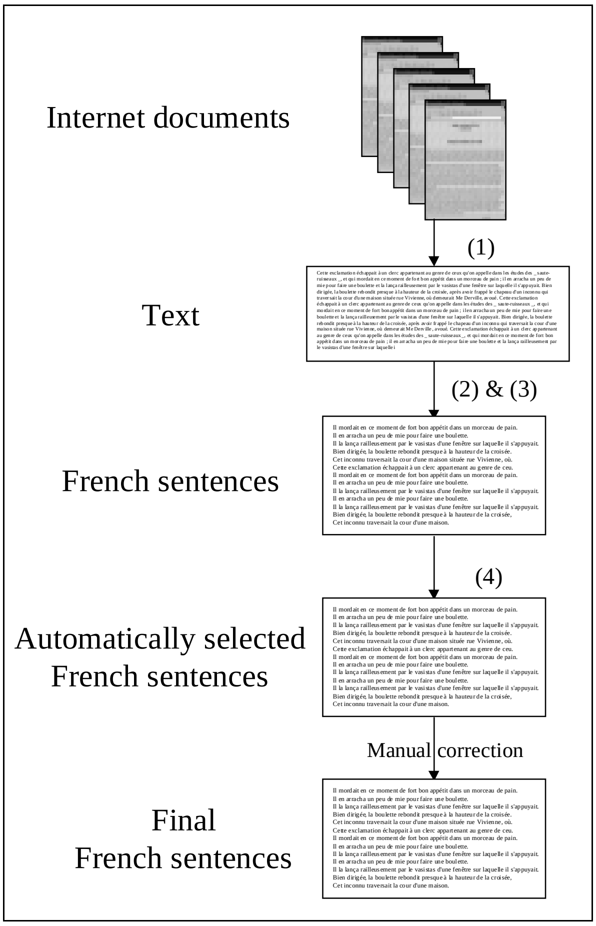

Figure 1: web text to French sentences filtering

The first filter takes Internet documents and produces text with inserted document separators (1). Next, we have used two lexicons to select sentences : BDLex (Pérennou, 1987), a dictionary with 245,000 entries, enlarged by ABU (Universal Bibliophiles' Association[2]) to about 400,000 lexical forms. The second filter produces all the sentences exclusively made with words of this vocabulary (2). It also transcribes numbers in context (date, money, etc.) to textual form in order to prevent various pronunciations between speakers. Thus, variability of signals labeling decreases and its reliability increases. Due to first names and non French vocabulary words found on the Web, we assume that the amount of complete French is not very high compared to the large amount of data we processed. After this stage, the amount of sentences we kept was 70759. Then, The third filter (3) deletes the following type of sentences:

- sentences made up of less than 15 words
- duplicate sentences
- sentences including more than one "."
- sentences including two identical consecutive words
- sentences containing spellings

At this point, a corpus of 47482 sentences was obtained. After that, in order to correct all the sentences, we need to make a first automatic pass. We used the treetagger (Schmid 1994) to transform every sentence in a list of grammatical tags. A language model built on a tagged version of *Grace*[3] corpus, is then applied to remove sentences corresponding to unknown trigrams (i.e. unknown grammatical form). After this automatic pass, about 10% of the sentences were removed and 43279 sentences remained in the corpus. Finally, the perplexity of the LM is calculated for each sentence in order to select only sentences for which perplexity is below a given threshold. The value of this threshold is used to tune two parameters : the number of sentences needed for recording and the accuracy of these sentences. Of course, there should be a compromise between these parameters (4). The speech database we want to record has to be 20 hours long (72000s). So, if we consider that sentences of 15 words (the shortest ones in our corpus !) have a mean duration of 6s ; we deduce that 12000 sentences are enough to build a 20 h speech database. By keeping sentences of perplexity less than 12, we managed to obtain 12239 sentences.

The last filter is a human one. A manual correction of sentences was performed, using a HTML interface in order to facilitate the collaborative work of readers. For technical reasons, the use of this interface allowed us to propose 11262 sentences to the readers. After this 10470 sentences were kept (the other 792 were judged impossible to correct, unusable, or were less than 15 words length after correction by human). The treatments

Filter	Documents type	Size
-	HTML + Text from Web (in Go)	10
(1)	French Text (in Go)	2
(2) & (3)	French sentences	47482
(4)	Automatically selected sentences	12239
Manual	Final French sentences	10470

[2] http://cedric.cnam.fr/ABU/index.html

[3] http://www.limsi.fr/TLP/grace/index.html

Table 1 : summary of the sentences filtering results

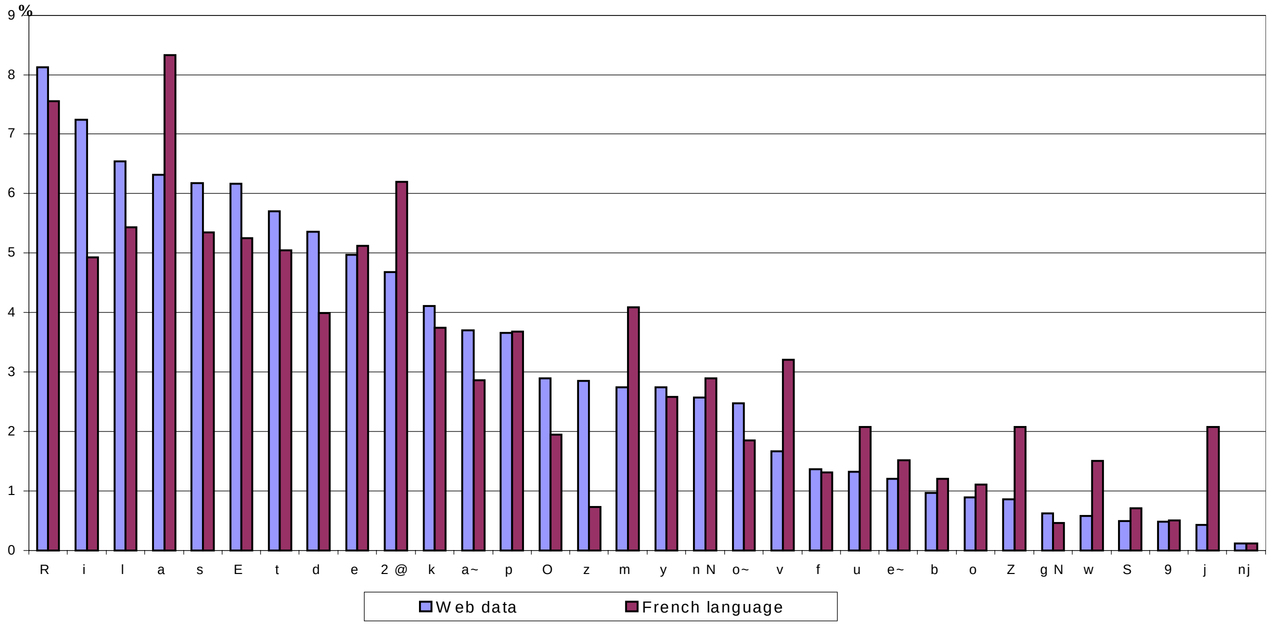

Figure 2 : adequacy of web data phoneme distribution to the French language

described in this section are summarized in Table 1.

Inside these 10470 correct sentences, 74% were kept without any change (77% if we do not take into account pure punctuation changes like comas), whereas the other ones were manually corrected. This high percentage of correct sentences shows that adequate filters applied to Internet data allows to automatically capture a large amount of clean and ready-to-use sentences ! We can see however that refining options alters drastically the size of the output corpus. Indeed, increasing constraints reduces number of words by a factor of 8000. Filtering parameters must be therefore carefully chosen.

2.2.2 Corpus organization

Our 10470 sentences corpus needs to be divided in a training corpus (90% of the total) and a test corpus (10% of the total). Moreover, phoneme distribution should be equivalent in each sub-corpus.

Thus, sentences were phonetized using the *text2phone*[4] script and 2 phonetically balanced corpus were obtained (training : 9444 sentences and test : 1026 sentences).

Since 100 speakers are planned to record this database, each sub-corpus was finally split and we obtained :

- a training corpus of 90 speakers (104 or 105 sentences/speaker – average of 26 to 28 words/sentence for each speaker)
- a test corpus of 10 speakers (102 or 103 sentences/speaker – average of 26 or 27 words/sentence for each speaker)

Moreover, a group of sentences, common for each speaker in both corpora, is added as a single recorded paragraph. This passage is extracted from “*La Science et L’hypothèse*” written by Henry Poincaré. So, one may be able to use these data in order to perform text dependent speaker identification experimentation, for example.

3. CORPUS EVALUATION

Since this speech database is dedicated to the training of acoustic models for speech recognition, it is necessary to know if the distribution of French phonemes is correct in our corpus. For this, we conducted an experiment in which we tried to measure how the Web could contribute to model and represent phonemes distribution in the French language.

For this, we compared our phoneme distribution to a French phoneme distribution found in the literature (Combescure, 1981). The histogram of both distributions is presented in Figure 2.

We see that the distribution of phonemes that we observe on sentences captured on the web, corresponds to a well-known phoneme distribution of French language. The correlation between both distributions was evaluated and we found a coefficient of 0.89, which confirms our previous conclusions.

4. DATABASE RECORDING

4.1 Database and recording tool

For the acquisition, and managing of speech signals during recording, we use the EMACOP system,

[4] http://tcts.fpms.ac.be/synthesis/mbrola.html

developed at CLIPS (Vaufreydaz, 1998). EMACOP is a Multimedia Environment for Acquiring and Managing Speech Corpora, running under Windows 9x and windows NT. EMACOP meets SAM (Tomlison 1991) specifications in input and in output.

Text files, representing SAM description of each speaker session, are prepared for EMACOP. For the training part, 990 files of ten sentences are created, 110 for the test part. One file is added for the single common paragraph, for each speaker. After this step, EMACOP integrates these files and create an internal representation in order to manage all these data and to control the recording phase. This tool also provides verification facilities, through an integrated interface, to check signal quality and consistency between resulting signals and what was presented to the speaker during recording.

4.2 Recording phase

Recording will take place in a quiet environment with a SENNHEISER HMD 410-6 head microphone and a microphone pre-amplifier PREFER MB-7. The sampling frequency will be 16 kHz, but we will also record the database on a DAT audio tape at 48 kHz which will allow us to provide different sampling rates. The frequencies below 60 Hz are removed.

The EMACOP system will be use in network mode. A server controls many client applications and collects data. First, the client registers information about the speaker, the recording date, and other useful information. Then, a calibration of the audio level is performed before recording. An operator will stay with the speakers during the whole recording session in order to ensure that sentences are correctly pronounced. No listening or visualization of a pronounced sentence is allowed to speakers but they can repeat a sentence until they are satisfied with the result.

4.3 Final corpora

Almost all the signals will be correct due to the mastering of speakers, but we will need to check them. That's why, after recording, a verification-by-listening of all utterances will be done with EMACOP, sentence by sentence, to correct, if needed, the labeling of signals. For example, if a speaker has inverted or has forgotten some words, it is possible to detect and to change it before exporting the final corpora.

Next, EMACOP will produce SAM data in output. That means that each corpus, training and test one, will contain the following items:

- a list of all speaker characteristics (code of speaker, mother tong, age, sex, smoker or not, etc.)
- a description of all sessions and how items were presented to speakers
- one directory by speaker
- in each speaker directory, a signal file and an associated description file (with speakers, recording and labeling information) for each session

The total size expected of speech recordings, for one hundred speakers, is about 3 gigabytes with a sampling frequency of 16000Hz, corresponding to 28 hours of recorded speech.

5. CONCLUSIONS

We have shown in this paper that Internet documents can be a very rich source of information for database recording. However to be useful, one has to clean and filter out the extracted documents based on the task in which data will be involved : in this example, speech corpora recording.

At the end, we will obtain two distinct corpora, phonetically balanced and representative of French. Moreover, neither speaker nor sentence will be present in both corpora, except the common passage. So, we can assume that these data will be useful to train and verify different acoustic features.

Because this acoustic modeling was not possible for us till now due to the lack of speech data to reliably train triphone models, these recorded corpora will allow us to build context dependent acoustic models for our spontaneous speech recognition module called RAPHAEL (Akbar & Caelen, 1998).

We are planning the recording session with more than one hundred speakers by the middle of April, 2000.

6. REFERENCES

Akbar M., Caelen J. (1998) “Parole et traduction automatique: le module de reconnaissance RAPHAEL”, COLLING-ACL’98, pp. 36-40, Montreal (Quebec), August 1998.

Combescure, P., (1981) “20 listes de dix phrases phonétiquement équilibrées”, Revue d’Acoustique n°56, pp 34-38, 1981.

Pérennou G., De Calmès M. (1987) “BDLEX lexical data and knowledge base of spoken and written French”, European conference on Speech Technology, pp 393-396, Edinburgh (Scotland), September 1987.

Schmid, H. (1994) "Probabilistic Part-of-Speech Tagging Using Decision Trees", International Conference on New Methods in Language Processing, September 1994.

Tomlison M.J. (1991) “Guide to Database Generation – Recording Protocol, Final Version”. SAM-RSRE-015, Marlvern, England.

Vaufreydaz, D., Akbar, M., Rouillard, J., (1999) “Internet documents : a rich source for spoken language modeling”, ASRU 99 Conference, pp 277-280, Keystone (USA).

Vaufreydaz, D., Akbar, M., Caelen, J., Serignat, J.F., (1998) “EMACOP : Environnement Multimédia pour l’Acquisition et la gestion de Corpus Parole”, Journées d’Etude sur la Parole, pp 175-178, Martigny (Switzerland), June 1998.