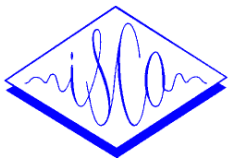

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

THE VESTEL TELEPHONE SPEECH DATABASE

D. Tapias, A. Acero (*), J. Esteve, J. C. Torrecilla

Telefónica Investigación y Desarrollo

C/ Emilio Vargas, 6

28043 - Madrid, SPAIN

ABSTRACT

VESTEL is a telephone speech corpus collected at the Speech Technology Division of Telefónica Investigación y Desarrollo. The data base was designed to support research in speaker-independent automatic speech recognition (ASR) based on word and subword units. Over sixteen thousand people called in response to newspaper advertisements. They were prompted by a recorded voice to say digits, numbers and commands, and to answer questions asking them the city where they lived and they were born, their name and surnames, a yes/no question and to spell some words. The utterances were spoken over commercial telephone lines, and each call was composed by twenty five separate utterances. Spain was divided into ten dialectical regions in order to take into account the main Spanish dialects of Castilian (usually known as "Spanish"). Each call was checked and transcribed by two people. In this report we describe the system implemented to record the data base, the publicity campaign, the recording protocol, the regions in which Spain was divided, and the statistical information of the tasks that were carried out.

I. THE VOCABULARY

The data base was designed to support research in speaker-independent ASR of isolated words of small, medium and large vocabularies, and also ASR of connected digits and numbers in order to be able to introduce new services in the Spanish Telephone Network, so that the data base was composed by:

- Isolated digits (10 words)
- Control words (22 words)
- Provinces, regions and cities of Spain (90 words)
- Names and surnames
- Spelled words
- Telephone numbers, constituted by 6 or 7 digits and driving licence numbers, constituted by 6, 7 or 8 digits, although they are not only pronounced in a digit by digit way, but also dividing numbers into groups of digits and pronouncing them like numbers, and
- Out of vocabulary words that were said accidentally.

II. PUBLICITY CAMPAIGN

There are two different approaches to obtain volunteer callers to record a data base: to make a previous selection of them in order to have a balanced data base from the dialect, sex, age, cultural level,... points of view or to use the brute force approach recording a big data base taking the assumption that everything is going to be enough represented. We chose the second approach because it is faster and cheaper than the first and it provided more speech, what is always interesting.

With this purpose, we launched a publicity campaign into the main newspapers of each region in Spain instead of using national newspapers because of several reasons: (a) To control the number of calls received from each region in order to undertake possible corrective strategies in the case of poor response to the advertisements. (b) To reach, with a high probability, people belonging to the region in which the campaign was being launched. (c) To avoid the saturation of the system we had available to record the data base.

People were enticed to call us by means of a modest prize ($1500) that was raffled among the callers at the end of the campaign, and a toll free number to make possible to call us without any charge.

It was estimated that the publicity campaign (constituted by 20 advertisements) reached 4 million people, that is 10.2% of the Spanish population, receiving 16,400 calls, what represents an average response rate of 0.4%. The response rate varied between 0.17% and 0.61% of the audience of the newspaper in which the advertisement was inserted (a single newspaper is read, in average, by 3.5 people, so the audience is approximately: number of newspapers sold times 3.5), what means that we received between 300 and 1800 calls per advertisement, so just one insertion of the advertisement in each of the newspapers selected was enough. The publicity campaign and the data base recording phase lasted one month and a half.

III. COLLECTION REGIONS

Spain was divided into ten collection regions in order to take into account the main Spanish dialects of Castilian. The criterion followed to make the division was to group

(*) Currently working at Microsoft Corporation (One Microsoft Way, 9/1166. Redmond, WA 98052-6399)

those accents of Castilian that were close enough and to put into separate dialectical regions the dialects of Castilian. We also tried to minimize the number of collection regions in order to make easy the use of the data base.

The collection regions obtained were:

1.- Galicia and Asturias.

2.- Cantabria, Castilla-León, Madrid, Castilla-La Mancha, La Rioja and the south of Aragón.

3.- Pais Vasco and Navarra.

4.- North of Aragon.

5.- Cataluña, Valencia and Baleares.

6.- Extremadura.

7.- Andalucía Occidental

8.- Andalucía Oriental.

9.- Murcia

10.- Canarias.

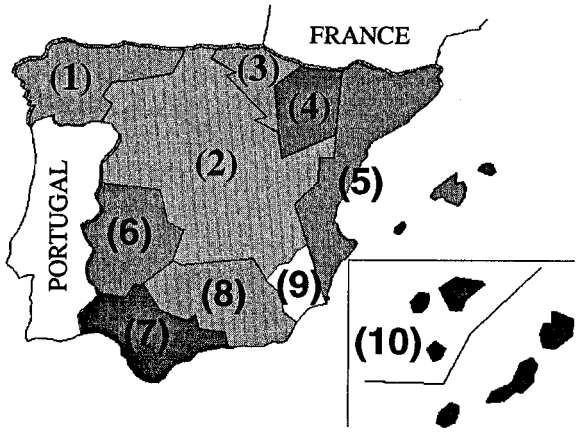

Figure 1: collection regions of Spain

IV. DATA ACQUISITION

1.-The Recording System

The recording system had twelve analog telephone lines because the VESTEL data base will be used to train ASR systems connected to analog lines and it is convenient to train the ASR in the same conditions in which it is going to work. The equipment in which the recording system was implemented was a PC/AT-486 at 66 MHz with Dialogic™ boards. The sampling frequency was 8 KHz and each of the samples was coded using mu-law. The device was programmed to answer the phone, play the prerecorded messages and digitize the caller's response for a fixed period of time, asking again the questions if the caller spoke before or during the beep, the response was not loud enough or the caller did not respond at all.

Our advertising company estimated the highest number of daily newspapers sold for the designated regions at 80000, the average audience per newspaper copy at 3.5, a response rate of 0.5%, a call duration of 3 minutes, the advertisement effectiveness duration of 1 day, and flat response between 11am and 8pm. With these estimates, we obtained a worst case traffic of about 8 Erlangs (80000 * 3.5 * 0.005 * 3) / (9 * 60). We selected a system capacity of 12 lines, for which the lost call rate was 10%. After analyzing the results we observed the worst case was 82 calls/hour, and the total percentage of lost calls was 0.13%, which was quite acceptable.

2.-The Recording Protocol

A well designed protocol is essential to correctly guide the callers along the call. Factors like the duration and the depth of the explanations have to be carefully devised. Equally important is the voice selected to record the protocol, in order to (a) make perfectly understandable the instructions, (b) correctly emphasize the sentences, stressing what is most important and (c) speak at the most suitable speed. For this reason a professional announcer recorded the protocol.

The protocol had four kind of messages:

- The welcome message
- The farewell message
- The questions and
- The explanation messages,

so the caller heard the following instructions:

"Thank you for calling to the Speech Technology Laboratory of Telefónica Investigación y Desarrollo. We remind you that this is a toll free call and lasts three minutes.

We are developing a system able to understand spoken commands in Castilian, and for that reason we need to record speech from all the regions, so we ask you to speak with naturalness and not to change your natural way of speaking.

We ask you to say some digits:

- Please, say "tres" (three).
- Say "cinco" (five).
- Say "seis" (six).
- Say "cero" (zero).
- Say "nueve" (nine).
- Say "cuatro" (four).
- Say "uno" (one).
- Say "siete" (seven).
- Say "dos" (two).
- Say "ocho" (eight).
- Repeat now the word "ayuda" (help).
- Repeat the word "siguiente" (next).

In order to classify your accent we need to know the province in which you were born and the province in which you live:

- Say your birth province.
- Please, spell the name of this province.
- Say the province in which you live.
- Say the name of your region.

Now think of two more Spanish provinces:

- Say the name of one of them.

- Say the name of the other.

To contact you if you win the prize that will be raffled before the notary public on November 29th, we need to know some private data:

- Please, say your last name.

- Say your name

- Spell your name.

- Say your telephone number.

In future applications of speech recognition the machines will need to understand driving licence numbers:

- Please, say a driving licence number, for instance, your number changing some digit.

Thank you for your call. If you wish to receive more information, you can write to the following address:

Telefónica Investigación y Desarrollo.

Emilio Vargas, 6

28043 - Madrid."

People said their name and surname without any problem, but some problems arose when we also tried to collect driving licence numbers, so we were forced to change the way in which we asked for them, asking for any driving licence number. The system asked callers to utter two different control words in each call, so one repetition of each control word was recorded every ten calls.

V. TRANSCRIPTION AND EVALUATION OF THE CALLS

Due to the big amount of calls we expected to receive, it was studied a procedure to make the transcription and evaluation of the calls faster than we did with previous data bases. As a result of this study a window based program was design and implemented. This program allows to label each file by using both the keyboard or the mouse, reducing the labelling time of each isolated word file to 12 seconds (on average).

The transcription and evaluation process consist of (a) listening of each utterance and transcription of what was said, and (b) evaluation of several factors of interest.

The transcription field can be filled with vocabulary words or non vocabulary words and labels corresponding to different events [1][2]. Words and labels are written with lowercase letters. The transcription conventions are as follows:

- coughs are transcribed as [tos].
- silences are transcribed as [sil].
- lip smacks are transcribed as [lab].
- breath noise is transcribed as [resp].
- telephone tones are transcribed as [tono].
- guttural sounds are transcribed as [gut].
- clicks and hits are transcribed as [golpe].
- non vocabulary words are written between brackets.

The factors that were evaluated by the listeners are:

- type of background noise

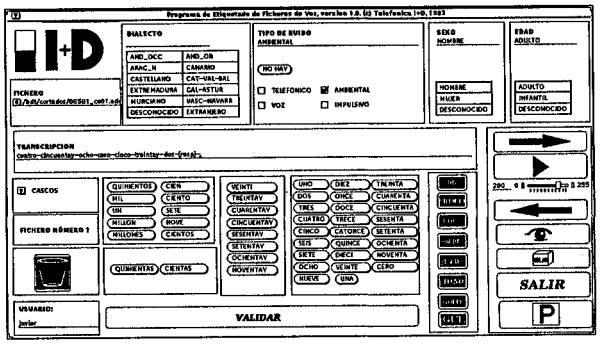

Figure 2: User interface to label connected digits and numbers. Each data base had its own labelling window.

- accent or dialect.
- Sex of the speaker (male, female or unknown)
- Age of the speaker (adult, child, or unknown).

Each utterance was heard, transcribed and evaluated separately by two different people with no previous experience on this task. After the transcription and evaluation process, a comparison between the two versions of the labelling was made automatically, and when some difference was detected the utterance was checked by two other people.

Figure 2 shows the user interface that was implemented to label connected digits and numbers. It is divided into four parts: (a) Fields to be filled by the listener (accent or dialect, background noise, sex, age and transcription), (b) control panel (that allows to play the file, go to the next or previous file,...), (c) events panel (to label coughs, lip smacks, breath noise,...) and (d) the vocabulary panel (to select the vocabulary word uttered).

As the utterances corresponding to each call belong to 6 different data bases (isolated digits, control words, province names, names and surnames, numbers and spelled words), the transcription and evaluation process was carried out data base by data base in the order in which the calls were received, removing automatically the non-speech intervals longer than 0.5 seconds before the transcription and evaluation step.

Finally, the signal to noise ratio was estimated for each of the files recorded.

Table 1: Labelling statistics per data base

DATA BASE	Labelling time (in seconds)	Standard deviation	Files recorded
Isolated digits	11.4	10.6	139,618
Control words	11.0	9.4	40,008
Provinces	13.2	12.6	65,212
Names	14.1	11.7	36,800
Spelled words	32.8	20.7	25,416
Numbers	28.1	21.4	24,431

Table 1 shows the average labelling time of a file, its standard deviation and the number of files recorded for each data base

VI. STATISTICS

Probably the most interesting result is the low cost of each call received, that has been estimated in 2 dollars including the cost of the telephone call (each call lasted three minutes), the publicity campaign, the professional announcer that recorded the prompts of the dialogue and the prize that was raffled among the callers. The cost of the evaluation and labeling process has been estimated in 4 dollars per call.

The calls were made from all the regions of Spain, so we were able to evaluate the SNR of the Spanish telephone network. Figure 3 shows its probability density function.

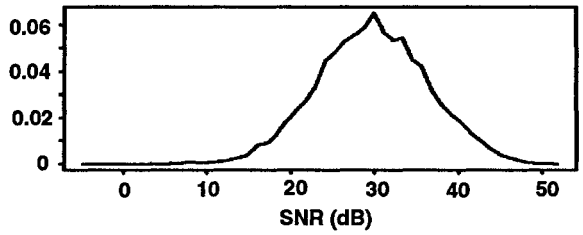

Figure 3: Probability density function of the SNR in the Spanish Telephone Network.

Concerning the recording phase of the data base: figure 4 shows the percentage of calls received each hour of the day, showing that almost 90% of the calls were made between 10 a.m. and 10 p.m. Figure 5 represents the number of questions per call that were repeated when the caller spoke before or during the beep. It can be seen that in 7% of the calls the system had to repeat more than 3 questions. Finally, figure 6 represents the number of questions per call that were repeated when the answer was not loud enough or the caller did not respond at all. It shows that 8.5% of the calls needed more than 3 repetitions of some question.

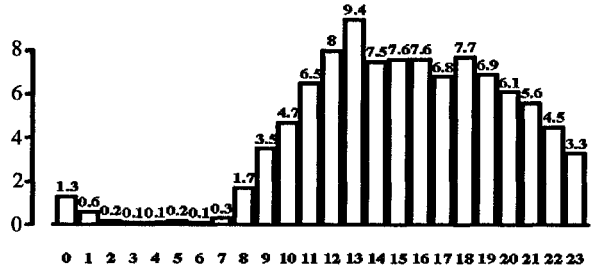

Figure 4: percentage of calls that were received each hour during the 24 hours of the day.

It is interesting to point out that 77% of people that called to our system answered all the questions (25) and that 9% of people did not answer any question. The rest of the people answered from 1 to 24 questions, due to the lack of interest of the caller or to low quality telephone lines that made almost impossible to communicate with the recording system.

Concerning the transcription process, 91.4% of the voice files were well transcribed the first time they were labelled while 5.7% were well transcribed the second time.

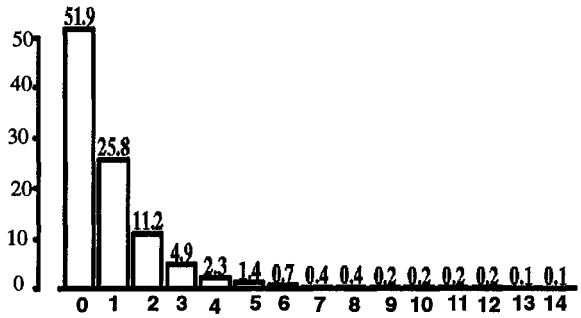

Figure 5: percentage of the number of questions per call that were repeated when the caller spoke before or during the beep.

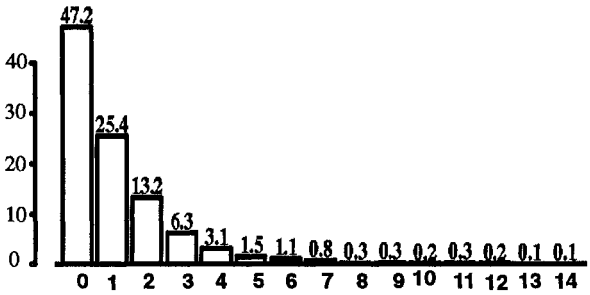

Figure 6: percentage of the number of questions per call that were repeated when the answer was not loud enough or the caller did not respond at all.

Summarizing, on average, each caller answered 21 questions per call, 6% of the questions were repeated because the caller spoke before or during the beep, 5% of the questions were repeated because the answer was not loud enough, what means that the strategy of repeating questions avoided to record 11% of the voice files that were not suitab

VII. ACKNOWLEDGEMENTS

The authors thank the Human Factors group of Telefónica Investigación y Desarrollo for their support in the design of the user interface implemented to transcribe and evaluate the telephone calls.

VIII. REFERENCES

[1] R. Cole, K. Roginski and M. Fanty, "A Telephone Speech Data Base of Spelled and Spoken Names", Proc. of ICSLP'92, vol. 3, pp. 891-893, October 1992.

[2] Y.K. Muthusamy, R. A. Cole and B. T. Oshika, "The OGI Multi-language Telephone Speech Corpus", Proc. of ICSLP'92, vol. 3, pp. 895-898, October 1992.