VOCATEL and VOGATEL: Two Telephone Speech Databases of Spanish Minority Languages (Catalan and Galician)

L. Villarrubia*, P.León*, L.Hernández , J.M.Elvira*
C. Nadeu ,I.Esquerra**,J.Hernando**,**
C.García-Mateo*, L.Docío *****

(*) Telefónica Investigación y Desarrollo, Madrid, Spain
(**) Universitat Politècnica de Catalunya, Barcelona, Spain
(***) ETSI de Telecomunicación, Universidad de Vigo, Spain

luigi@craso.tid.es ,climent@gps.tsc.upc.es, carmen@tsc.uvigo.es

ABSTRACT

Castilian (usually known as Spanish) is the official language for all the regions of Spain, while Catalan, Galician and Basque are minority languages that are becoming more and more important due to their recognition as official languages for their respective regions.

These languages are each time in stronger position with official recognition and funding. According to the 1991 census the 93,76% / 68,34% of the Catalonia population understand/ speak Catalan and the same data in 1996 show an important increased to 94,97% / 75,30% respectively. One of the main reasons for this increment in the number of Catalan speakers is the enormous influence of the language education in the schools. Considering the Galician language, again according to the 1991 census, a 94% of the Galician population understand Galician and an 88% can speak it. Therefore, due to the increasing importance of these minority languages, Telefonica I+D decided to include these languages in its speech technology allowing the design of new voice response systems for them. To achieve this goal two speech telephone corpus in Galician (VOGATEL) and in Catalan (VOCATEL) were planned to be collected by the Speech Technology Division of Telefonica I+D with the collaboration of the University of Vigo and an association of the Polytechnical University of Catalonia and *Fundacio Catalana per la Recerca*. Also recently Telefonica I+D is being involved in the collection of a Basque speech telephone corpus with the collaboration of the University of Basque Country.

In this paper we describe the collection results of the VOCATEL and VOGATEL databases for Catalan and Galician languages.

1. COLLECTION SYSTEM

The collection of VOCATEL and VOGATEL databases was done following a methodology close to the one previously used to collect the VESTEL Castilian telephone corpus collected by Telefonica I+D [1] in 1993. The database collection methodology was based on a publicitary campaign (newspapers, radio, etc.) for the recruitment of callers and on the use of automatic recording systems connected to the real telephone network that prompt the caller through a recorded voice to say isolated digits, control words, natural numbers (including telephone numbers, ID number), phonetic balanced sentences, regions, names and surnames. The major difference between VESTEL and VOCATEL and VOGATEL databases is that VESTEL was designed principally for isolated-word speaker-independent automatic speech recognition while VOGATEL and VOCATEL were designed for natural numbers and subword speech recognition.

1.1. The Recording System

The equipment in which the recording system was implemented was a PC Pentium with DIALOGIC ™ boards with a capacity of 12 telephone lines. The system was programmed to answer the phone, play the pre-recorded messages and record the caller's response, asking one more time the questions if the caller speaks before or during the beep, or the system did not detect any response from the caller.

1.2. The Recording Protocol

The basic flow of the recording protocol consist on: a welcome message, explanation messages, questions and farewell messages. It is important to remark that a female voice was used for questions and a male voice was used for welcome, explanation and farewell messages to make more clear the dialogue. Besides, special care was taken over the protocol to achieve speaker independent corpus (i.e. we can not ask train and test sentences or numbers to same caller).

A new problem, compared to a previous methodology in VESTEL, derived for this recording system method is to consider that the caller has to be able to repeat the question and this is special problematic for sentences and numbers. Therefore the sentences corpus was designed considering a maximum length of 4 words for sentence. Large sentences were only allowed if they were easy to remember and to repeat. This problem does not appear in the collection of other king of databases based on read sentences.

1.3. Publicitary Campaign

The publicitary campaigns were based on a carefully designed set of advertisements in the main newspapers and radios of Galicia and Catalonia. People were recruited by offering a prize that was raffled among the callers and using a toll free telephone number to call the recording system without any charge.

	Newspaper advertisements	Radio advertisements	Diptych	POPULATION RANGE (%)	DURATION	AVERAGE RESPONSE [Minimum-Maximum]	PRIZE
VESTEL	20	-	-	10.2	4 MONTHS	0.37% [0.17-0.61]	$1500
VOGATEL	5	5	1000	57.2	4 MONTHS	0.27% [0.08-0.67]	$1500
VOCATEL	10	5	25000	70.0	3 MONTHS	0.33% [0.05-0.90]	PC Pentium

Table 1: Principal characteristic of the publicity campaign.

	RECEIVED CALLS			RECORDED QUESTIONS		QUESTIONS PER CALL	RESPONSES PER CALLS	PERCENTAGE OF CALLS + 5 REPETITIONS :	
	Number	Completed(%)	Empty (%)	Nº	%	Nº	Mean number	Silent	Anticipation
VESTEL	16400	77.0	8.7	331485	81.6	23	20.2	4.0	2.6
VOGATEL	10465	41.6	26.1	148884	56.9	25	14.2	7.0	2.7
VOCATEL	7990	78.0	8.1	163431	81.8	25	20.4	13.0	6.2

Table 2: Analysis of the received calls.

In table 1 we can see the principal characteristic of the campaign. This table includes the number of advertisements for radio and television. In some cases we also used information diptychs. It is important to point out that apart from the planned advertisement, an important number of interviews and articles appeared in different communication media. That increased the global impact of the campaign and resulted in an important free of charge publicity.

In table 1, the population range means the percentage of the total population reached by the campaigns calculated based on the audience of the newspapers and radios. It can be observed that the population range for VOGATEL and VOCATEL is higher, around the 60% that the population range for VESTEL. Obviously is easier to reach more population range for minority languages where the number of speakers is lower than for majority languages. Comparing the average response of the people that called the system can be concluded that it is similar for VESTEL and VOCATEL, and a little bit lower for VOGATEL. The prize for all the campaigns was similar but for VOCATEL people were enticed by a PC instead of a metallic prize.

A general analysis of the received calls is showed in table 2. It can be compared the number of calls received to the system, the percentage of calls where the caller answered all the questions -completed- and the percentage of calls where the caller did not answer any questions. The "recorded questions" column reflects the percentage of answered questions from the total number of questions done for all the calls, and the absolute number of questions recorded.

From these two last columns it can be observed, comparing VOCATEL and VOGATEL, that in spite of the higher number of calls received in VOGATEL the final number of responses recorded is higher in VOCATEL. This is due to the behaviour of the Catalan callers that respond the 81,82 % of the questions opposite to the 56,91% obtained for the Galician callers. This fact is also remarked by the much better values for completed and empty calls in VOCATEL compared to VOGATEL. Finally another derived data is the higher mean number of response per calls 14'2 vs. 20'4. Also, all the values corresponding to the received calls for VESTEL and VOCATEL are similar. However, it is important to remark that the questions to repeat in VESTEL are in general easier than the questions in the VOCATEL and VOGATEL recording.

Finally, it is also interesting to note that the percentages of the repetitions due to silent anticipation of the caller in VOCATEL are the highest but the percentage of recorded question is also the highest as it was concluded before.

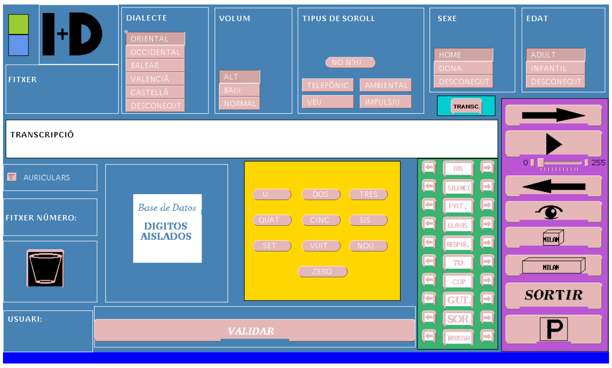
2. CALLS LABELING

2.1 General Labeling

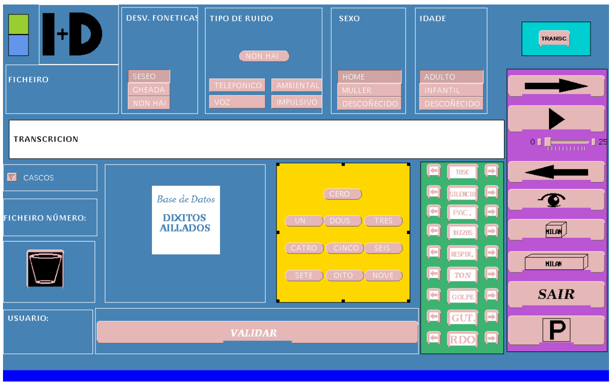
A modified version of a windows based tool, developed for VESTEL labelling was used. Different screens were development to take in account the different kind of utterances received (numbers, sentences, names, etc.). Figure 1 shows the labelling screen that was implemented to label isolated digits for Catalan (a) and Galician (b). The screen is divided in four parts:

- Top area for general event labels.
- Middle area for the transcription.
- Low area for vocabulary buttons.
- Right area for isolated noises labels and control panel buttons (play, view, etc.)

This program was developed considering the big amount of calls that was received by our capture method. Therefore, this labelling tool allows labeling each file by using both the keyboard and the mouse reducing the labelling time.



(a)



(b)

Figure 1 Isolated Digits Labeling Screens:

(a) Catalan, (b) Galician.

In our labelling methodology we encoded the following information:

a) Orthographic transcription of the utterance.

b) Isolated noises that appear with the utterance like breath, smacks and coughs. We also include different mark-ups if the noise overlap the voice of the caller or not.

c) General events that affect all the utterance: background noise, sex and age of the caller.

Two important facilities included for the VOCATEL and VOGATEL labelling were:

- The program automatically presents the expected transcription of the file to the labeled. This mechanism significantly reduces the time of labelling. Also, a button incorporated in the labelling screen allows to erase the hold transcription. It is useful when the utterance is completely different to the expected transcription. Besides, the labels corresponding to isolated noises can be located in the transcription using the mouse.

- Buttons corresponding to frequent names and localities were incorporated to the respective labelling screens.

2.2 Particular Labeling

Particular labels were used for VOCATEL and VOGATEL databases transcription.

In VOCATEL three general events were included:

- *Volume* (normal, high and low). This can detect files with an analogic saturation produced in the telephone microphone and utterances abnormally low. These files sometimes can not be detected with the SNR parameter.

- *Language* (Catalan or Castilian). Using this label utterances that were pronounced in Castilian instead of Catalan were marked.

- *Dialect*. It was especially difficult to determinate the dialect due to phonetic-inexperience of the labelers and the impossibility of using lexical information or upper levels information.

In VOGATEL one general event named *Phonetic Deviations* was included. This label can take the values “Gheada”, “Seseo” or “There isn’t”. “Gheada” is a phonetic deviation that consist on the pronunciation of the phoneme /g/ (occlusive, velar, voiced) as [x] (fricative, velar, unvoiced). For example /gato/ as [xato]. “Seseo” consists in pronounce the phoneme /θ/ (fricative, interdental, unvoiced) as [s] /fricative, alveolar, unvoiced).

2.3 Labeling Problems

A general problem related to our recording method, detected in the labeling process, was that the caller forgets the question to repeat or repeats the question hesitantly. Another point to consider is that answers to spontaneous questions included word-spotting effect. It is also interesting to point out that the surnames labelling is the most complicated task (in special in Catalan) due that two different orthographic transcriptions can be phonetically correct.

Regarding to the minority language characteristic of VOCATEL and VOGATEL there was found a problem in the labelling process for proper names, surnames and localities. A lot of them, surnames in special, were pronounced in Castilian (Table 3 shows this for Catalan). To solve this problem, in VOCATEL was incorporated the language label explained before. Also, for Catalan and Galician when a Castilian pronounced word appear, this word was specially transcribed (between square brackets and preceded by the symbol %).

	LANGUAGE	
	Catalan (%)	**Castilian (%)**
Locality	99,8	0,2
Name	93,1	6,9
Surname	80,1	19,9

Table 3: Mark-up Language results for Catalan.

	VESTEL		VOCATEL		VOGATEL	
	Mean labelling time (seconds)	**Deviations**	**Mean labelling time (seconds)**	**Deviations**	**Mean labelling time (seconds)**	**Deviations**
Isolated digits	11,4	10,6	11,5	8,39	11,4	9,05
Control words	11	9,4	11,5	8,71	10,5	8,52
Localities	13,2	12,6	14,2	10,73	12,7	9,05
Names *	14,1	11,7	13,2	9,3	11,13	9,00
Numbers **	28,1	21,4	18,01	30,59	20,6	15,11
Sentences	-	-	11,64	10,00	15,4	11,53

** Names and surnames*

*** Numbers, telephone numbers and ID number*

Table 4: Time labelling per database.

2.4 Labeling Statistics

Some important results can be obtained from the statistics information files generated by the labeling tool. This statistics represents the average labeling time of different labellers for different labeling tasks.

The minimum labeling time is 21 seconds for numbers as shows table 4. The labeling time results are very similar for the three databases except for numbers. In VOCATEL and VOGATEL the labeling time for numbers is much lower than in VESTEL due to the automatic expected transcription method. Besides, the labeling time for the numbers are higher than the labeling time for the sentences because numbers incorporate spontaneous questions (ID number and personal telephone)

A noise analysis was also obtained from the labeling (Table 5). The mean SNR is approximately 30dB as in VESTEL. The more frequent isolated noises are lips smacks and breathing for Catalan and Galician databases. Considering the isolated and background noise, the values recorded are double for VOGATEL comparing VOCATEL. Therefore, the callers environment is cleaner for VOCATEL.

The sex is balance as Table 6 indicates for VOCATEL and VOGATEL. The age of the callers is principally adult but for VOGATEL there is an important percentage of child calls.

REFERENCES

(1) D. Tapias, A. Acero, J. Esteve and J.C. Torrecilla. “The VESTEL Telephone Speech Database”. ICSLP 94, pp. 1811-1814.

	SNR (dB)		BACKGROUND NOISE (%)	ISOLATED NOISES (%)			
	Mean	**Deviation**	**With**	**Lip smacks**	**Breathings**	**Rest**	**Total**
VOCATEL	30,6	6,3	21,8	8,3	5,2	3,3	16,8
VOGATEL	29,5	7,1	40,5	4,5	6,1	21,1	31,7

Table 5: Background and isolated noises labeling results.

	SEX			AGE		
	Male (%)	**Female (%)**	**Unknown (%)**	**Adult (%)**	**Child (%)**	**Unknown (%)**
VOCATEL	44,2	55,4	0,4	96,4	3,5	0,1
VOGATEL	48,7	50,8	0,5	78,9	20,7	0,4

Table 6: Sex and age labeling results.