Design of a Phonetic Corpus for Speech Recognition in Catalan

Ignasi Esquerra[*], Climent Nadeu[*], Luis Villarrubia[+], Paloma León[+]
([*]) Universitat Politècnica de Catalunya, Barcelona, Spain
([+]) Telefónica Investigación y Desarrollo, Madrid, Spain
{ignasi,climent}@gps.tsc.upc.es, {luigi,paloma}@craso.tid.es

Abstract

In this paper, we present the design of a corpus for speech recognition to be used for the recording of a speech database in Catalan. A previous database in Spanish was the reference in setting the specifications about the characteristics of the sentences and in the minimum number of units required. An analysis of unit frequencies were carried out in order to know which units were relevant for training and to compare the results with the figures from the designed corpus. Three different sub-corpora were generated, one for training, the other for vocabulary-independent verification and the third for vocabulary-dependent verification. Short sentences were obtained that contained all phones and relevant diphones in a sufficient quantity. Evaluation of the corpus characteristics was performed using several parameters to validate database specifications. Using this corpus, a speech database was recorded over a telephone line and manually labelled, and it is currently used to train and test several speech recognition systems.

1. Introduction

Speech recognition systems require large databases for training the models. In particular, systems based in phonetic units as models need databases made of sentences from which models for all sounds of a language can be trained. The process of designing a corpus of sentences for such a database is described in this paper. This large telephone Catalan database has been developed in co-operation by the *Universitat Politècnica de Catalunya* and *Telefónica Investigación y Desarrollo* with the support of *Fundació Catalana per a la Recerca.*

The VOCATEL database consists of three different databases depending on the recognition task. The first database (BD1) is for isolated digits, commands and control words; the second database (BD2) will be used in number sequence recognition; and the last database (BD3) is for a phonetic vocabulary-independent recognition system.

A database of similar characteristics was already collected in Spanish [1]. In order to develop applications for other languages spoken in Spain, databases in Catalan and Galician have been collected by and *Telefónica Investigación y Desarrollo* [2]. The database acoustic specifications were taken basically from the experience of previous Spanish databases (VESTEL, Albayzín) [1,3].

Due to the fact that this is not a read database, the sentences had to be short and easy to remember, typically from two to five words, since they were acoustically prompted by the system. However, programs developed to design phonetic corpora for speech recognition databases at the UPC research group were mainly developed for long sentences [3,4]. Some modifications had to be made, as for instance, the constraint that all phones had to occur at least once within each set of nine sentences. Because of the larger number of Catalan phonemes and the short sentence length, the minimum number of unit repetitions per set of sentences had to be lowered.

As the phonetic database has to contain all Catalan sounds, and some combinations of them, it was first necessary to know which units to select and how many of them. For this reason, an analysis of phonetic frequencies was done to be used as reference in the corpus design as described in section 2. Afterwards, the corpus of sentences was created (section 3) and evaluated (section 4). The paper concludes with a summary of the characteristics of this phonetic corpus and the problems reported in its design.

2. Reference Corpus Analysis

In order to know how often the phonetic units appear in Catalan, a frequency analysis was performed over a text corpus of around 66000 words obtained mainly from an electronic newspaper on the Internet. Since the objective of this corpus is to analyse colloquial speech, opinion articles and interviews were mainly selected among all available texts. Hereafter, this corpus is called reference corpus because the unit frequencies obtained from its analysis will be used later in the design of the phonetic corpus for recognition.

2.1. Acquisition and Segmentation

First of all, texts were processed to put abbreviations, numbers and other non-readable symbols in its orthographic form. Then, the corpus was converted into phonemes using a transcription program developed at UPC for a text-to-speech system [5]. A set of 37 phonetic

symbols, including some allophonic variations and one special symbol for pauses, has been considered to represent the sounds of Catalan using SAMPA notation.

In the frequency analysis, not only single units were counted, but also combinations of two of them. Left and right contexts were considered together since they represent the same unit from a counting point of view. For example, the phonetic unit [a] followed by [n] (i.e. [a]+[n]) occurs the same number of times than [n] preceded of [a] (i.e. [a]-[n]). That is the reason why phones with context will be called diphones in the reference corpus analysis, following the name used in the synthesis field for units made of two phones.

Performance of the transcription tool is of key importance to this process since diphone frequencies strongly depend on the rules to generate these units [6]. The text-to-phoneme program was evaluated over a dictionary of 1400 words with their phonetic transcription achieving a high percentage of correct words. Differences between the reference and the automatic transcription are due mainly to problems with few specific phonemes that are difficult to solve only by means of rules. For instance, mid-vowels <e> and <o> in a stressed syllable with no orthographic accent, some times are pronounced as mid-open ([E]/[O]) while others as mid-close ([e]/[o]). A dictionary with the correct transcription is looked up for such words. As a default rule, the more common mid-close phone is transcribed. This results in a higher proportion of units containing these phones, while the mid-open vowels are less frequent than they should be actually.

2.2. Frequency Results

The results of allophone frequencies show that the most common sound in Catalan is the schwa [@] [Table 1]. In the central dialect spoken by the greatest part of population in Catalonia, all non-stressed vowels <a> and <e> are pronounced with this sound. No distinction has been made between stressed and non-stressed vowels when they can occur in both positions; this is the case of [i] and [u].

The lowest frequencies correspond to affricated consonants. This fact makes valid the simplification of regarding these sounds as a combination of occlusive and fricative. It is interesting to note that only 6 allophones are required to achieve an accumulated frequency of 50%, and with half of the allophones is possible to transcribe 90% of a text. Some values are not very reliable as said before due to transcription errors, for instance [E]/[e] and [O]/[o] vowels.

With respect to diphones, 870 different units where found in the reference corpus. This figure represents a 63% over the maximum number of possible units (37x37=1369). Diphones were also counted according to their left and right context, so that relative frequencies per allophone were obtained. It is interesting to note that the most frequent 16 diphones represent 25% of accumulated frequency; 55 diphones are necessary to achieve the 50% of frequency, and with 523 diphones is covered the 99% of the units appearing in the reference corpus.

Allophone	# of units	Freq. (%)	Acum. (%)
@	57185	18.27	18.27
i	22877	7.31	25.58
s	20205	6.45	32.03
n	18677	5.97	38.00
l	17663	5.64	43.64
t	16672	5.33	48.97
u	15144	4.84	53.81
a	14058	4.49	58.30
k	13290	4.25	62.55
m	11374	3.63	66.18
e	11178	3.57	69.75
D	11109	3.55	73.30
r	10830	3.46	76.76
z	9245	2.95	79.71
p	9114	2.91	82.62
o	9039	2.89	85.51
/	8923	2.86	88.37
rr	7739	2.47	90.84
B	6750	2.16	93.00
f	3170	1.01	94.01
w	2633	0.84	94.85
G	2476	0.79	95.64
Z	1972	0.63	96.27
L	1732	0.55	96.82
d	1282	0.41	97.23
b	1263	0.40	97.63
E	1082	0.35	97.98
S	974	0.31	98.29
N	971	0.31	98.60
O	953	0.30	98.90
j	794	0.25	99.15
J	762	0.24	99.39
ts	693	0.22	99.61
g	561	0.18	99.79
dZ	292	0.09	99.88
dz	203	0.06	99.94
tS	184	0.06	100.00
	313069	100.00	

Table 1: Frequency results for the allophonic set of Catalan units

Allophone	# of left contexts	# of right contexts
i	37	36
e	37	35
E	26	29
a	37	34
o	34	34
O	23	24
u	37	35
@	37	37

Table 2: Number of contexts per vocalic allophone

Because it is impossible to list all diphones, only the number of left and right contexts for vocalic allophones is presented [Table 2]. It can be observed that only the schwa [@] exists in all possible contexts left and right contexts, while the other vocalic sounds present fewer possible combinations.

3. Corpus Design

The speech database has a triple objective: to train the phonetic models and to evaluate the speech recognition system, both in a vocabulary- independent and vocabulary-dependent tasks. For this reason, the database has been divided in three smaller databases: one for training (BD3-E), the second one for vocabulary-independent verification (BD3-IV) and the last one for vocabulary-dependent verification (BD3-DV).

Taking into account that it was expected to receive around 5000 calls, from which only half would be usable, and that each caller pronounced 9 sentences, this makes a database with approximately 22500 sentences.

Specifications required a minimum of repetitions for each allophone and relevant diphones in the BD3-E, BD3-IV and BD3-DV databases. The definition of relevant diphones will be presented later in this section.

In order to simplify the assignation of sentences, it was decided that three sub-corpora would be designed one for each database. The sub-corpora for BD3-E and BD3-IV have an equal number of sentences; a subset from BD3-E is used for the BD3-DV sub-corpus.

3.1. Sentence Generation

To generate the corpus of sentences, the following iterative method was used. From a large corpus of newspapers texts, sentences between 10 and 40 letters were selected, transcribed and sorted according to a phonetic probabilistic criteria [3]. The most "interesting" sentences, i.e. those having the less frequent allophones, were retained and units were counted to know whether they reach the minimum number of required repetitions; otherwise more sentences were taken and the process was done again. The probabilistic measure can be expressed as:

$$Prob(\text{sentence}) = \frac{1}{N}\sum_{i}^{N}\log_{10}(freq(i)/100)$$

where N is the number of phones in a sentence, and freq is the phone frequency (in %) obtained from the reference corpus analysis

A minimum number of unit repetitions in the corpus is set proportionally to the minimum specified in the collected database. For the majority of allophones it was easily reached in one or two iterations. However, some allophones were more difficult to obtain, especially those with a lower frequency, so that new sentences had to be written containing those allophones to finally get all required repetitions.

Relevant diphones are defined using frequency and phonetic selective criteria. Firstly, highly frequent units in the reference corpus were selected. In addition, units with more than 5% of relative frequency per allophone were also considered as relevant. The specification for relevant diphones was found very difficult to achieve because of the high number of diphones selected so far, so that a relaxed minimum of repetitions was set in order to obtain a reasonable amount of relevant diphones.

As before, new sentences were added to the corpus, which was phonetically transcribed and sorted again. Since the new sentences replaced the ones at the bottom, hopefully the minimum number of allophone repetitions was preserved. Finally, all diphones were grouped into 286 classes following a phonetic clustering criteria provided by expert phoneticians from *Universitat Autònoma de Barcelona* (UAB), in order to verify that all allophones were present in the most relevant phonetic generic contexts a sufficient number of times. After successive additions of new sentences and frequency recalculation, a final corpus was achieved.

4. Evaluation Measures

Several measures about the corpus contents were performed. Apart from counting how many times the different types of units occur, other measures, like frequency distance to the reference corpus, have been computed to verify the corpus specifications [7].

The minimum number of repetitions was achieved for all phonetic units except for [E], [O]. However, this low number is due to a shortcoming in the transcription tool, as commented before. A manual verification of the transcription showed that a third of those vowels were erroneously transcribed as closed [e]/[o]. The implication of this is the fact that the number of [E]/[O] actually present in the corpora is higher, being enough to achieve the minimum number of required repetitions. Almost in all three sub-corpora, allophone frequency order is the same than in the reference corpus.

With respect to frequency distance, the most frequent allophones are close to the reference corpus values. The case of the "allophonic" unit representing pauses should be left apart since it depends very much on the sentence length. On the other hand, the less frequent allophones present a greater distance to the reference corpus due to the lower number of occurrences.

In reference to diphone results, not all diphones in the reference corpus were taken into account, only the most relevant ones. Two measures of phonetic coverage have been considered to validate similarity between reference corpus and the designed corpus.

The diphone coverage (DC) is defined as the percentage of different diphones in a sub-corpus (C1) with respect to another one (C2), usually the reference corpus. For the acoustic coverage (AC) the percentage is taken over the total number of repetitions. Among other, coverage measures have been computed between the three BD3-E, BD3-IV and BD3-DV sub-corpora and the reference corpus, and the verification sub-corpora and the training corpus [Table 3].

C1	C2	DC	AC
BD3-E	REF	81,03%	99,74%
BD3-IV	REF	81,95%	99,72%
BD3-DV	REF	69,08%	98,87%
BD3-IV	BD3-E	93,81%	99,69%
BD3-DV	BD3-E	84,53%	99,00%

Table 3: Sub-corpora coverage measures

From the previous results it can be seen that the acoustic coverage is very high; in particular, is 0.4% higher than it was specified in the design protocol. However, the diphone coverage is relatively lower compared to the 88% expected. The main reason for this figure is that fewer diphones had been regarded as relevant because of the large number of possible diphones in Catalan. Therefore, the selection of relevant diphones had to be more restrictive.

5. Conclusions

The process of designing and assessing a phonetic corpus for speech recognition has been presented. As a first step, a text corpus was transcribed and segmented to count the number of occurrences for each type of unit (phones, allophones and diphones). The analysis of frequencies was used later to decide which units would be considered in the corpus. To create the sentences which made up the corpus, an iterative methodology was employed that basically consists in getting sentences with an appropriate length, selecting the phonetically richest sentences, counting whether all the required units are present and adding more sentences with the missing units. The process is finished when enough sentences are found that contain at least the required number of unit repetitions.

Some problems regarding the adaptation of database specifications and methodology developed for another language have been reported. In particular, this has been made for a Catalan database, taking as a reference previously designed Spanish databases. A speech database has been collected using this phonetic corpus and it is currently being used in several research and development speech recognition projects.

Acknowledgements

We want to thank researchers from the Phonetic Laboratory of *Universitat Autònoma de Barcelona* (UAB) for their collaboration in this project, which consisted mainly in the selection of relevant contexts, in verifying corpus sentences correctness, and for providing part of the reference text corpus. Likewise, we want to express our gratitude to L. Hernández from *Telefónica Investigación y Desarrollo*.

References

[1] D. Tapias et al., "The VESTEL Telephone Database", Proc. ICSLP'94, pp 1811-1814

[2] L. Villarrubia et al., "VOCATEL and VOGATEL: Two Telephone Speech Databases of Spanish Minority Languages (Catalan and Galician)", (in this workshop)

[3] A. Moreno et al., "Albayzin Speech Database: Design of the Phonetic Corpus", Proc. EUROSPEECH'93, pp

[4] A. Moreno, "EUROM.1 Spanish Database", ESPRIT Technology Assessment in Multilingual Applications, ESPRIT project 6919, report D6

[5] A. Pujol, I. Esquerra, "Regles de transcripció fonètica del català", Internal Research Report UPC, October 1996

[6] I. Esquerra, "Avaluació del transcriptor en català", Internal Research Report UPC, July 1997

[7] I. Esquerra, C. Nadeu, A. Pujol, "Diseño de la base de datos fonética para el reconocimiento en catalán", Final Project Report, March 1997