THE DESIGN OF A LARGE VOCABULARY SPEECH CORPUS FOR PORTUGUESE

João P. Neto, Ciro A. Martins, Hugo Meinedo and Luís B. Almeida

INESC - IST
R. Alves Redol, 9
1000 Lisboa - Portugal
E-Mails: jpn@inesc.pt, cam@inesc.pt, hdsm@inesc.pt, lba@inesc.pt

ABSTRACT

The last years show a great development of large vocabulary, speaker-independent continuous speech recognition systems and some research in multilingual aspects. To allow that development to also be extended to the European Portuguese language we decided to develop and collect a large database of continuous speech based on a large amount of text. In the development of this new Portuguese database our aim was to create a corpus equivalent in size to WSJ0. We selected the database texts from the PÚBLICO newspaper, which is characterized by a broad coverage of matters and different writing styles. The recording population was selected from a large engineering school, assuring a large variability of speakers. The recordings are being done as we write this paper and we expect to release the database in CD format in September 1997.

1. INTRODUCTION

The development of a large vocabulary, speaker-independent continuous speech recognition system needs an appropriate database of both text and speech. The situation of the European Portuguese language is different from those of the English[1, 2] and some other European languages[3, 4]. None of the available Portuguese speech databases has the adequate size and contents for training the kind of recognizers that we want to address.

From this point we decide to develop and collect a large database of continuous speech based on a large amount of text. The development of a new database of this kind is a time and manpower consuming task. For that reason it was necessary to bring together the appropriate funding resources[1].

[1]This work was partially funded by the EU through LTR project 20077 SPRACH, by the Portuguese PRAXIS XXI program through project 1654, and with the collaboration of INESC, IST and the PÚBLICO newspaper.

In the development of this new Portuguese database our aim was to create a corpus equivalent in size to the WSJ0 database[1]. We also chose as corpus a newspaper's text. Our choice was the Portuguese PÚBLICO newspaper. PÚBLICO is one of the best daily newspapers in the European Portuguese language, with a broad coverage of subjects and written by a excellent set of journalists and collaborators. The newspaper has its editions available on the WEB, through the "PÚBLICO ON-LINE" initiative (http://www.publico.pt/). This WEB version has all the text of the daily paper edition.

As recording population we selected the students from the Instituto Superior Técnico (IST), a large engineering school from the Technical University of Lisbon, with undergraduate and graduate students. This population presents, in terms of this task, some limitations and some advantages. The main limitation comes from the age range which is only between 19 and 28. However there are several advantages. Due to being one of the best and larger engineering schools of the country, we can find many students from different regions (from south to north, from the larger cities on the coastline to the small ones in the interior) and from different social levels, which gives us a large variability of speakers with different accents.

The recordings are taking place in a sound proof room at INESC (Lisbon), which is located in the neighborhood of IST. The recordings started in mid April 1997 and are being done as we write this paper. We expect to release the database in CD format in September 1997.

2. GOALS OF THIS DATABASE

In this database the speakers are asked to read a set of sentences extracted in paragraph blocks from the newspaper text. However with this database we do not want to restrict ourselves to dictation systems.

We have learned from dictation products that these systems are essentially characterized by cooperative speakers operating in speaker-dependent or speaker-adaptive modes, generating continuous speech in a careful fashion to facilitate accurate transcription.

In this database we are imposing a dictation task but in a speaker-independent mode, given the number of speakers and the quantity of data for each speaker. We know that this situation is not the best in terms of continuous speech recognition but it certainly will help us in the development of domain independent acoustic models, pronunciation dictionaries and language models and thereby domain independent continuous speech recognition systems.

With this system we will have the opportunity to explore several aspects of speech recognition:

- developing adaptation of the acoustic models through speaker-adaptation, as we previously did for English[5]
- developing adaptation of language models to specific tasks
- creating new language models, exploiting the regularity of the Portuguese language and reducing the vocabulary size and perplexity
- learning how to include the expected large variability of word pronunciations in a efficient way in the lexicons.

3. TEXT PREPROCESSING

The first phase of our work consisted on the collection of the text of the first 6 months of the newspaper that were available through WWW (from the beginning, on September 22, 1995 till March 31, 1996). Next these data were converted from *html* format to text format. Afterwards the files were cleaned up (removing *html* headers and some duplicated information) and a single file was created for each day. Each edition was labeled with a unique *id* made from the date, and each article also with a unique *id* based on the edition and on the original filename. This makes it very easy to locate any part of the text at any stage of the processing.

In the end we obtained 188 files corresponding to the same number of editions of the newspaper. This represents approximately 220 Mb of text. In the next step we analyzed the texts in order to correct misspellings, and to convert numbers into ortographics. The process of selection and conversion was done automatically, and then manually verified and corrected. This step was very consuming in terms of time and manpower.

After all these steps we considered the texts ready for use both as selection material for the sentence prompts and as a basis for language model development.

4. TEXT STATISTICS

The next phase was the definition of the different training and test sets and the selection of the sentences to record. We started by computing the overall totals of these texts (Table 1).

Number of	
Editions	188
Articles	24,287
Paragraphs	148,657
Sentences	416,617
Words	10,976,009
Different words	155,867
Different words occurring more than twice	62,015

Table 1. Text totals.

Next we performed a statistical analysis of the texts to help us decide which should be the parameters to use in the selection of sentences. Those statistics led us to decide that our spoken paragraphs should have 2 to 4 sentences each, and each sentence should have between 6 and 39 words. We rejected paragraphs with just one sentence because we want to maintain coherent paragraph blocks of text which "provide semantically meaningful material, thereby facilitating the production of realistic speech prosodics"[1] and longer paragraphs (more than 4 sentences) which occur very infrequently and normally are harder to read. These limiting parameters and the restriction that the words should be among those that occur more than twice defined the set of paragraphs and sentences that were available for selection.

5. TEXTS SELECTION FOR RECORDINGS

The text was divided into three parts: training, development and evaluation. We used 80% of the text for training, 10% for development and 10% for evaluation. This selection was made in a random fashion having the paragraph as unit.

From the training part of the text we randomly selected paragraphs with a total of 10,000 sentences

to be used as recording material. In this training set we have a total of 21,025 different words.

For both the development and evaluation test sets we decided to have two vocabularies: a small one with no more than 5K words and a larger one with no more than 20K words. As in WSJ0 we would like to have 2,000 sentences for each of the 5K sets and 4,000 sentences for each of the 20K sets.

For the 20K development test set we picked sentences at random obeying to the restrictions defined in the previous section and an additional one of using no more than 20K different words. We selected 4,000 sentences with a final vocabulary of 13,070 words.

The 5K words development test set was harder to select. We followed the same procedure as for the 20K set and we obtained only 809 sentences. At that point we decided to allow the vocabulary size to increase until we got 2,000 sentences. The results are presented in Table 2.

Vocabulary dimension	Number of sentences
5,000	809
6,000	1,111
7,000	1,401
8,800	2,000

Table 2. Evolution of the number of sentences for the development test set obeying to the restriction of the vocabulary dimension.

All these sets are available but we used only the first one of 5K words with only 809 sentences. Since we want to select a total of 400 sentences (with repetition) for recording, the number of sentences that we got is still sufficient. It is important for us to maintain the total vocabulary words within 5K due to computational limitations. The same process was applied for both development and evaluation test sets.

Additionally, 15 speaker-adaptation and three calibration sentences were selected. These sentences were chosen to be phonetically rich. They were originally from the PÚBLICO texts but were modified by hand.

The overall selected sentences were individually examined, to eliminate those that were hard to read. Then they were converted into prompts to be used in the recording phase, and into standard SGML format to be used in the recognizer score.

6. RECORDING SETS

The next step was to define the various recording sets. We decided to have a large training set of 8,000 sentences from 100 speakers, and development and evaluation test sets of 400 sentences from 10 speakers each, for both 5K and 20K vocabularies.

The numbers of speakers and sentences are the following for each set:

- Training set with 100 speakers (50 male and 50 female). 80 sentences plus 3 calibration sentences for each speaker.
- 5K development test set with 10 speakers (5 male and 5 female). 40 sentences plus 15 speaker-adaptation sentences and 3 calibration sentences for each speaker.
- 5K evaluation test set with 10 speakers (5 male and 5 female). 40 sentences plus 15 speaker-adaptation sentences and 3 calibration sentences for each speaker.
- 20K development test set with 10 speakers (5 male and 5 female). 40 sentences plus 15 speaker-adaptation sentences and 3 calibration sentences for each speaker.
- 20K evaluation test set with 10 speakers (5 male and 5 female). 40 sentences plus 15 speaker-adaptation sentences and 3 calibration sentences for each speaker.

The three calibration sentences were the same for all the speakers and the 15 speaker-adaptation sentences were the same for the development and evaluation test sets speakers.

The allocation of the sentences to the speakers was random, with sentence replacement between speakers.

7. VOCABULARY DEVELOPMENT

After the allocation of the sentences to the speakers we compiled the resulting vocabulary for each set.

	Vocabulary size on	
Set	Total sentences	Selected sentences
Training	21,025	15,877
5K Development	5,000	2,528
5K Evaluation	5,000	2,543
20K Development	13,070	3,030
20K Evaluation	13,023	3,156

Table 3. Vocabulary size for the different sets.

In the end we compiled a list of the different words for all the sets (for the total sentences). That list has a total of 27,833 words. It was from this list that a

pronunciation dictionary was developed, as described in [6]. From this global dictionary we extracted the pronunciation dictionaries associated with each of the above sets.

8. LANGUAGE MODEL DEVELOPMENT

From the training part of the texts, four bigram back-off closed language models were computed (5K/20K words x development/evaluation test sets) using the *CMU-Cambridge SLM Toolkit*[2]. The perplexity results obtained for each set are presented in Table 4.

Set	Perplexity
5K Development	231
5K Evaluation	241
20K Development	261
20K Evaluation	262

Table 4. Language model perplexity for each set.

As we can observe the perplexity values associated to these tasks are large.

9. RECORDING PHASE

For the time being, we chose to record only the training set and the 5K development and evaluation test sets. We expect to record the 20K development and evaluation test sets in a later phase. It will be simple to create new sets from now on, because the recording conditions and the students will still be available.

The recordings are taking place at INESC, in a sound proof room. A desk mounted microphone is being used for the collection of the signal.

The action of database collection was advertised through out the IST *campus* and the students offer to participate in the project gracefully. As a compensation for their collaboration they received a *T-shirt* with the logo of the project. The recordings started in mid April 1997 and are proceeding as we write this paper. We expect to release the database in CD format in September 1997.

10. CONCLUSION

The database that we presented here is the result of a large and careful planning work. We expect that this database of both speech and text, and the supporting material, will be useful to the speech recognition research community to create and develop continuous speech recognition systems for European Portuguese. Preliminary recognition tests using a small part of this database are reported in [6].

[2]Available on `http://svr-www.eng.cam.ac.uk/~prc14/CMU-Camb_Toolkit_v2-BETA.4.tar.gz`

11. ACKNOWLEDGMENTS

This work was partially funded by the Long Term Research RTD project 20077 SPRACH and by PRAXIS XXI project 1654. Acknowledgments go to the PÚBLICO newspaper for making its texts available and to Philip Clarkson who made available the CMU-Cambridge SLM Toolkit. Among all the people that collaborated in this project we would like to thank our colleagues of the Neural Networks and Digital Signal Processing Group at INESC and also Prof. Isabel Trancoso and Prof. Luís C. Oliveira and the students that participate on the database collection.

REFERENCES

[1] D. Paul and J. Baker, *The Design for the Wall Street Journal-based CSR Corpus*, in Proceedings ICSLP 92, pp. 899-902, 1992.

[2] T. Robinson, J. Fransen, D. Pye, J. Foote and S. Renals, *WSJCAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition*, in Proceedings ICASSP 95, pp. 81-84, 1995.

[3] L. Lamel, J.-L. Gauvain and M. Exkénazi, *BREF, a Large Vocabulary Spoken Corpus for French*, in Proceedings EUROSPEECH 91, pp. 505-508, Genoa, Italy, 1991.

[4] L. Lamel, M. Adda-Decker and J. L. Gauvain, *Issues in Large Vocabulary, Multilingual Speech Recognition*, in Proceedings EUROSPEECH 95, pp. 185-188, Madrid, Spain, 1995.

[5] J. Neto, C. Martins and L. Almeida, *An Incremental Speaker-Adaptation Technique for Hybrid HMM-MLP Recognition*, in Proceedings ICSLP 96, Philadelphia, USA, pp. 1289-1292, 1996.

[6] J. Neto, C. Martins and L. Almeida, *The Development of a Speaker Independent Continuous Speech Recognizer for Portuguese* in Proceedings EUROSPEECH 97, Rhodes, 1997.