A Brazilian Portuguese Language Corpus Development

Mauricio C. Schramm, Luis Felipe R. Freitas, Adriano Zanuz
Dante Barone

Universidade Federal do Rio Grande do Sul
Instituto de Informática

ABSTRACT

This article presents the techniques that are being used for the creation of a database related to the Brazilian Portuguese language. This database is composed of a collection of recorded voices, from different speakers and different regions of Brazil. The collected voices contain varied phonetic and phonologic information. The applications of this database are diverse, including synthesis and recognition systems and data for linguistic studies.

The corpus is composed of read sentences in Brazilian Portuguese, similar to sentences found in the TIMIT corpus, as well as answers to questions such as the speaker’s name, address, telephone number, ZIP code, and other information. The data were recorded at 44 kHz with a direct connection from the microphone to the sound card. The corpus contains information from about 200 speakers, although future development efforts will expand the corpus size to 1000 speakers. The paper covers in some detail the protocol used to design this corpus and the methods of data collection.

An HMM/ANN-hybrid continuous digits recognizer developed using a small subset of this corpus has 96.18% word-level accuracy and 78.95% sentence level accuracy. This recognizer was trained on 48 files, developed using 11 files, and tested on 19 files, with an average of 5 digits per file. A total of 103 context-dependent categories were used in training. A general-purpose recognizer capable of recognizing arbitrary words is currently under development.

This article is within the context of the Spoltech Project that is a project on computational linguistic research. It aims to create, develop and improve the technologies of speech synthesis and recognition. This interdisciplinary project is composed of researchers, teachers and students of several different institutions, they are:

- Instituto de Informática and Instituto de Letras (Language and Literature College) of the Universidade Federal do Rio Grande do Sul, Brazil;
- Departamento de Informática of the Universidade de Caxias do Sul, Brazil;
- CSLR/CU (University of Colorado, Boulder), USA;
- CSLU/OGI (Oregon Graduate Institute), USA.

A basic need to the project’s progress, is a database of linguistic information about the Brazilian Portuguese (BP) language. This database is the “Corpus”, and its creation process is the subject of this paper.

2. THE CORPUS

One of the main objectives of the Spoltech Project is to develop Brazilian Portuguese language corpora. Its immediate application is to use the available information to build up a speech recognizer for BP. Over and above, other applications of these corpora are possible, overall, on the linguistic field. The Spoltech has recorded voices from several speakers from different regions of the country. To collect voices, sentences for repetition were elaborated, and so were a questionnaire, to which the speakers answer things like digits, person’s names and place names.

2.1. Balanced Phrases

In order to build up a speech recognizer, it is profitable and indispensable to check out the existing Linguistic background at first. For a good functioning of phoneme based it is necessary to provide that with a wide covering of the sounds of the language. All the phonemes and allophones of the language are required, in several contexts of uttering. A vast corpus is needed because they can vary according to the phonemic boundary and even from one speaker to another.

As a strategy to cover all the phonemes, within different contexts, phonetically balanced phrases are composed, such like the sentences found in the TIMIT corpus [3]. After listing the phonemes and its allophones, they are grouped in currently
spoken sentences. These must, if possible, be meaningful and understandable to the speaker, so that he or she can easier utter them naturally. The voices are recorded to the database of the corpus. Later on, all this information is analyzed and labeled.

To set almost all the consonants on sentences is no problem. The same applies to the vowels, even if nasal. Although, Portuguese has some vowel conjunctions, whether diphthongs or tritongs (conjunction of two semivowels plus a vowel or two vowels plus a semivowel). These sequences of vowels can also be nasal, if in nasal contexts. These tritongs don’t occur regularly in the language, which makes harder its placement even on elaborated phrases [4].

Some sets of different sentences are needed in order to locate phonemes within different contexts. It is, one set of sentences must cover all the phonemes, the other one must cover all the phonemes again but now contrasting them with different phonemes, giving them varied boundaries.

To give one example of balanced sentence, we present:

“O presidente da república faz advertência ao ministro da justiça”

Which means: “the president of the republic warns the minister of justice”.

And whose phonetic transcription (supposing an average speaker of BP) is:

\[ /u p c p r e z i d e t S i d a x e p c p u b l i k k a f a z a d Z i v e r t e t e j s a w m i n i s t r u d a Z u s t S i s a / \]

2.2. Questions

The answering to the questions of the protocol is part of the corpus. These answers are also labeled. The purpose of getting specific words is indeed to create specific recognizers. These recognizers are composed with a limited lexicon that will be useful to calling services. Within the Spoltech’s corpus there are questions requiring straight answers such like names of people, places, food, numbers; yes and no; time references and so on. For answers to full name, there is a great rate of foreigners last names that may be considered out of Portuguese Phonology rules. In spite of that, they are also labeled acoustically. There are some questions in which the speakers are requested to answer their telephone and ZIP code digit per digit. The questions were also similar to those from CSLU protocol [5].

3. LABELING PORTUGUESE

The labeling process basically consists of a time-aligned transcription that, according to the CSLU Labeling Guide [6], must be phonetically and orthographically. As it is supposed to be the utterance acoustically, the phonetical transcription is a sequence of sounds, characteristics from the language. So the labeling process is done according to the acoustic analysis of the speech from spectrogram and waveform. The tool used is the Speech Viewer (figure 1), from the CSLU Toolkit, described below. The alphabet of transcription used is the Worldbet [7].

Whenever transcribing Brazilian Portuguese language, some things might be taken into account. The set of consonants is /p, b, t (allophone t$), d (allophone dZ), k, g, m, n, n~, f, v, s, z, S, Z, l, L, x (allophone rr) and t/. The vowels are i, e, E, a, u, o and >, being all nasalized, mainly before nasal consonants, although a few times distinctive from the non-nasals [4].

The vowels, diphthongs and tritongs are, in fact, very clear to acoustic analysis, with clear formants distinction, however sometimes close vowels like /E/ and /u/, for instance, that are distinctive in any context, may occur like allophones or like a “vowel at midway”. In such cases, it is the context that assures the meaning at colloquial speech itself, so the labeling goes according to this extra-acoustic information.

The consonants are usually well distinct, while categories such like stop, fricative, affricate, and nasals. Yet, within each category, they do not look so different. For example, in the phrase:

“O Zorro é outro dos filmes muito procurados nas locadoras atualmente”

Which means: “Zorro is another movie very required at video rent stores at present time”.

The sequence “dos filmes”, that could be transcribed for BP like /d u s f i w m i s/, has few, almost no one visible distinction at spectrographic analysis between the /s/ and /l/.

![Speech Viewer](image)

**Figure 1:** Speech Viewer

4. SPEECH COLLECTION METHODS

In order to make the process of collecting voices easier, it was necessary to develop a software to this task. This software was developed in Tcl/Tk, using several functions from the CSLU Toolkit [8] packages. With this tool it is possible to make the voice collection by microphone easy to the user, and by telephone totally automatic. The configuration is easy too; the user can change several recording parameters, like sample rate, audio level, maximum and minimum recording time and speech detection levels.

The voice files are stored in wave RIFF PCM (.wav) format, and the file naming is llc_sssss_uuu.wav, where ll stands for location (state, city, etc.), c for input type (microphone, telephone), sssss for speaker number and uuu for utterance type. There is also a directory structure in which each speaker’s utterances are stored inside a directory named llc_sssss, using the same naming method above. Inside of each one of this
directories are other three: `wav` with the wave files, `phn` with the label files and `text` with the text files.

### 4.1. Microphone

In this type of speech collection, a microphone is used, connected to a sound card, so the recording is made directly in the computer. The software creates the directories and file names automatically. The user has just to set the speaker number, read what utterance is being asked, press a button and talk. As soon as the button is pressed, the phrases are hidden so that the speaker utters them more naturally, rather than read speech. When the user finishes the utterance, he will hear what was recorded and check if everything is OK, and if its not, he can record it again. This process is repeated until all the utterances are recorded.

When recording trough a microphone, the audio will be stored in 44100 Hz, 16 bits, mono. Figure 2 shows the speech collection tool, main window.

![Figure 2: Speech collection tool.](image)

### 4.2. Telephone

When collecting voice trough telephone line a Dialogic D/21 board is used. When someone makes a call, at first a greeting message is heard and then a voice, that can be recorded voice or a synthesized voice, asks for several utterances, sequentially. This data recorded in this way, is stored 8000 Hz, 16 bits, mono. Figure 3 shows the speech collection tool, phone status window.

![Figure 3: Waiting for a phone call.](image)

### 5. PRACTICAL APPLICATIONS

Two HMM/ANN-hybrid speech recognizers were developed using the corpus presented in this paper, a digits recognizer and a general-purpose recognizer. The first has a vocabulary composed of the numbers in Portuguese. The second is a vocabulary independent recognizer.

#### 5.1. Digits Recognizer

To create the digits recognizer was used a small subset of the corpus having information of telephone and ZIP code numbers. This recognizer was trained on 48 speech files, developed using 11 files and tested on 19 files, with an average of 5 digits per file. A context-dependent modeling was used in training with a total of 103 categories. The results obtained had 97.71% word-level accuracy and 89.47% sentence level accuracy.

#### 5.2. General Purpose Recognizer

A general-purpose recognizer capable of recognizing arbitrary words is currently under development in Spoltech Project. This recognizer uses the whole of the developed corpus and the recognition is made in a phoneme level where all Portuguese phones and diphones are used to train the neural network. Until now, the results obtained to the general-purpose recognizer had 59.91% phoneme-level accuracy but with the insertion of more data in corpus (under collection) this results shall be better.

### 6. CONCLUSION

This paper discussed the work done by the Spoltech project to developing a Brazilian Portuguese language corpus so far, its contents and construction process. The applications to this material are several: voice synthesis and recognition systems and phonetic, phonologic and linguistic research. These applications could collaborate in the progress of man-machine communication and benefit handicapped people.

At the present, Spoltech project has a corpus of about 200 speakers. The available data is being used to train speech recognizers to the CSLU Toolkit. The results of this work will be available to the academic community, in order to collaborate with the advance in computational linguistics research, improving the quality of communication between men and machines.

### 7. ACKNOWLEDGEMENTS

This research was supported by a collaborative research grant from NSF (National Science Foundation, USA), and CNPq (Conselho Nacional de Pesquisa, Brazil). The authors would like to thank these institutions by the sponsoring.
8. REFERENCES


