Baseline Acoustic Models for Brazilian Portuguese Using CMU Sphinx Tools

Rafael Oliveira, Pedro Batista, Nelson Neto, and Aldebaro Klautau

Federal University of Pará, Signal Processing Laboratory,
Rua Augusto Correa 1, 660750110, Belém, PA, Brazil
{rafaelso,pedro,nelsonneto,aldebaro}@ufpa.br
http://www.laps.ufpa.br

Abstract. Advances in speech processing research rely on the availability of public resources such as corpora, statistical models and baseline systems. In contrast to languages such as English, there are few specific resources for Brazilian Portuguese. This work describes efforts aiming to decrease such gap. Baseline acoustic models for Brazilian Portuguese were built using the CMU Sphinx toolkit and public domain resources: speech corpora, phonetic dictionary and language model. Experiments were carried on for dictation and grammar tasks and the obtained results can be used to support further researches. Part of the trained acoustic models and a reference speech corpus were made publicly available.

Keywords: Speech recognition, Brazilian Portuguese, CMU Sphinx.

1 Introduction

Sphinx is a public domain software package maintained by the Carnegie Mellon University (CMU) for implementing automatic speech recognition (ASR) systems. Sphinx has been investigated in many works [1–3] and its current version has performance equivalent to other open source softwares widely used on the community, such as HTK [4]. But the HTK tools cannot be freely distributed, unlike Sphinx. In addition, free softwares compatible with the HTK acoustic models, like the Julius decoder, have not yet reached the same performance obtained with the HTK decoders [5].

Therefore, the development of specific resources for Sphinx has gained importance because there is no an eficiente HTK to Sphinx model converter and the PocketSphinx and the Sphinx3 decoders are interesting alternatives. Sphinx has been largely used for many languages [6–8]. However, to the best of the author's knowledge, there are no previous works using Sphinx for Brazilian Portuguese (BP). This work presents a first effort for developing baseline acoustic models for BP based on the Sphinx toolkit. In the sequel the resources used to build and evaluate this acoustic models are described.

H. Caseli et al. (Eds.): PROPOR 2012, LNAI 7243, pp. 375–380, 2012.

2 Resources Used

The phonetic dictionary used for developing the acoustic models is described in [9]. This dictionary has approximately 60 thousand words transcribed in the SAMPA alphabet. The speech corpora used to build the acoustic models were: (1) the West Point Brazilian Portuguese corpus suggested in [10]; (2) the LapsStory corpus [5] composed by audio files from audiobooks; and (3) a corpus collected by the Centro de Estudos em Telecomunicações (CETUC) [11]. The publicly available dictation test corpus LapsBenchmark [5] was adopted to evaluate the models. Table 1 summarizes the characteristics of these speech databases.

Table 1. Speech corpora used to build and evaluate the acoustic models

Database	Hours	Speakers	Words	Acoustic environment
West Point	8	128	484	not controlled
LapsStory	16.28	8	8257	controlled
CETUC	142.83	101	3528	not controlled
LapsBenchmark	0.96	35	2731	not controlled

The trigram language model (LM) was trained with 710,000 sentences extracted from CETENFolha [12] and LapsStory corpora as described in [5]. It has perplexity 170 measured with a test set of 10,000 sentences randomly selected from CETENFolha (unseen during the training stage).

In this work, the development of a free speech database named LapsMail was initiated. This corpus was designed to represent a basic set of commands to control a electronic mail application, aiming establish it as a reference benchmark corpus for this context. Actually, the LapsMail corpus consists of 86 BP sentences including 43 commands and 43 names spoken by 25 volunteers (21 male and 4 female) which corresponds to 84 minutes of audio. Its vocabulary has 95 different words. The following are three typical sentences in the corpus.

```
(1) <s> abrir caixa de entrada </s>
(2) <s> responder ao remetente </s>
(3) <s> ana carolina </s>
```

The LapsMail corpus was recorded using a high quality microphone (Shure PG30), sampled at 16 kHz and quantized with 16 bits. The acoustic environment was not controlled. The LapsMail corpus is publicly available [13].

3 Building Acoustic Models for BP

Both continuous and semi-continuous acoustic models were trained using the SphinxTrain package tools [14]. The acoustic waveforms from the training corpus were parametrized into 13-dimensional cepstrum. For continuous models, these