Experiments on the Construction of a Phonetically Balanced Corpus from the Web

L. Villaseñor-Pineda‡, M. Montes-y-Gómez‡, D. Vaufreydaz* and **J-F. Serignat***

‡ Laboratorio de Tecnologías del Lenguaje, INAOE, México
{villasen, mmontesg}@inaoep.mx
* Laboratoire CLIPS/IMAG, France
{Dominique.Vaufreydaz, Jean-Francois.Serignat}@imag.fr

Abstract. The construction of a speech recognition system requires a recorded set of phrases to compute the pertinent acoustic models. This set of phrases must be phonetically rich and balanced in order to obtain a robust recognizer. By tradition, this set is defined manually implicating a great human effort. In this paper we propose an automated method for assembling a phonetically balanced corpus (set of phrases) from the Web. The proposed method was used to construct a phonetically balanced corpus for the Mexican Spanish language.

1 Introduction

The construction of a speech recognition system requires a set of recordings to obtain the pertinent acoustic models. These recordings must consider several aspects in order to produce a robust recognizer. For instance, (i) the spoken corpus must be *rich*, i.e., it must contain all the phonemes of the language, and (ii) it must be *balanced*, i.e., it must preserve the phonetic distribution of the language.

The construction of a phonetically rich and balanced corpus is based on the selection of a set of *phrases* that will be recorded. Traditionally, this selection involves a great human effort. First, it is necessary to select a set of words phonetically rich, and join them to form the desired phrases. Later on, it is necessary to verify the phonetic distribution of the constructed phrases, and if required, add and delete some phrases. Certainly, these changes affect the overall phonetic distribution, and thus, the process must be repeated until an adequate distribution is reached.

In this paper, we propose a straightforward method for selecting a set of phrases to be recorded. This method is entirely different from the traditional process. It is supported on the hypothesis that the Web, for its huge size, is already a phonetically rich and balanced source, and thus, taking a subset of it is enough to assemble a phonetically rich and balanced corpus.

The following sections describe the proposed method, and illustrate the construction of a phonetically rich and balanced corpus for the *Mexican Spanish language*.

2 Collecting documents from the Web

In order to assemble the desired corpus, we first need to collect a set of documents from the web (a broad exposition on this problem was presented in [2, 3]). For this