

Toward Acoustic Models for Languages with Limited Linguistic Resources

Luis Villaseñor-Pineda¹, Viet Bac Le²,
Manuel Montes-y-Gómez^{1,3} & Manuel Pérez-Coutiño¹

¹ Laboratorio de Tecnologías del Lenguaje,
Instituto Nacional de Astrofísica, Óptica y Electrónica
Sta. María Tonantzintla, Puebla, México
{villasen,mmontesg,mapco}@inaoep.mx

² Laboratoire CLIPS-IMAG,
385, Avenue de la Bibliothèque, B.P. 53, 38041 Grenoble Cedex 9
Viet-Bac.Le@imag.fr

³ Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, España
mmontes@dsic.upv.es

Abstract. This paper discusses preliminary results on acoustic models creation through acoustic models already in existence for another language. In this work we show as case of study, the creation of acoustic models for Mexican Spanish, tagging automatically the training corpus with a recognition system for French. The resulting set of acoustic models for Mexican Spanish has gathered promising results at the phonetic level, reaching a recognition rate of 71.81%.

1 Introduction

A system for continuous speech recognition is formed by (i) a system, which using a set of acoustic models from the target language, builds a chain of symbols (usually phonemes) starting from acoustic boundaries extracted from the voice signal; and (ii) a system responsible for the reconstruction of words and sentences given a language model adapted to a language and, often adapted to the application domain of the recognition system [1]. Current statistical techniques used in the computation of acoustic models demands large volumes of data (oral and text corpus). Thus, specification, compilation and tagged of such data volumes are complex tasks and the human effort required is huge.

There are a diversity of initiatives in order to develop large acoustic data bases, like GlobalPhone data base [2], which has compiled data for Arab, Chinese, Croatian, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish and Turk languages. To date, the project has compiled 233 hours of speech from 1300 speakers approximately. Another effort is the SpeechDat project [3], currently with a total of 28 data bases for 11 European languages and some preponderant dialect variants and minority languages. These data bases have been compiled as basic elements for the