Toward Acoustic Models for Languages with Limited Linguistic Resources

Luis Villaseñor-Pineda[1], Viet Bac Le[2],
Manuel Montes-y-Gómez[1,3] & Manuel Pérez-Coutiño[1]

[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 discuses 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 development of telephonic applications like information services, transactions and other voice-based services. It is evident that such initiatives operate with a huge amount of human and material resources. Given this context, the treatment of minority languages within a lack of resources becomes extremely difficult. The aim of the present work is to take up this problem proposing a methodology in order to bring down the amount of required data to model languages with few resources, reusing both data and models existing for languages with abounding resources.

The remaining sections present a case of study which reuses the acoustic models initially developed for French, to construct the acoustic models for the Mexican Spanish. The relevance in the definition of this methodology is the possibility of the automatic treatment (from systems for language identification to specific recognizers) of the indigenous languages spoken by 54 ethnic groups distributed along the territory of Mexico.

2 Proposed Methodology

The aim of this work does not strive in the development of a system for automatic recognition of multilingual speech with recognition capabilities for several languages with a recognition quality equivalent between such languages. This work attempts to develop a monolingual system for a specific language (target language) reusing data and acoustic models from another language (source language).

The main idea lies on the hypothesis that there are some mapping between the phonemes of both languages. Thus, the challenge consists in the definition of the most pertinent correlation, i.e. what phonemes in the base language are the nearest to those in the target language?. There are two approaches to answer this question: the knowledge-based methods and the data-based methods [4, 5]. The first try to determine the correlation through phonetic similarity of the data, this require an expert whose define similarities and determine the correlation. One disadvantage of this approach is that the determination of acoustic units is performed independently from acoustic data. Then, the quality of the acoustic models depends on the quantity of oral data for the target language. Automatic methods derive the acoustic units using few acoustic data from source language, this require either, confusion matrixes analysis or distance metrics (usually based on relative entropy) to determine what model of the base language is the closer to the model of the target language. One disadvantage of such approach is the presence of particular sounds in the target language, which are not present in the base language.

In this work we explore the knowledge-based approach and propose an *a priori* mapping established with the help of linguists and our own expertise. As second step, after establish the mapping, a suite of acoustic models adapted for the target language is set up. With these models we start an automatic alignment process on a set of recordings. Thus, the first *real* version of the acoustic models is computed starting from these recordings and their approximated alignment. Then, the corpus is realigned with the first version and a second version of the acoustic models is computed. The process of alignment and computation for the acoustic models are repeated until the difference in the recognition between versions is minimal.

2.1 Mapping Spanish–French

The correlation between Spanish and French phonemes is relatively simple, given that the majority of Spanish phonemes are present in French. Only three phonemes require a special treatment. In these cases it was needed to approximate the model for Mexican Spanish starting from two French models. In the French system, a phoneme is modeled by a three state Hidden Markov Model: initial, intermediate and final. The approximation consists in take initial and intermediate state of a French model and combine them with the final state of a second model. This is the case for phonemes fonemas /tS/, /x/, /ll/. Table 1 shows the approximations applied in this experiment.

Grapheme	Spanish	French	Approximations		
			initial	intermediate	final
ch	tS	t + S	**t**	**t**	**S**
j	x	k + h	**k**	**k**	**h**
ll	L	j + J	**j**	**j**	**NJ**

Table 1. Approximations for Mexican Spanish starting from French models[1].

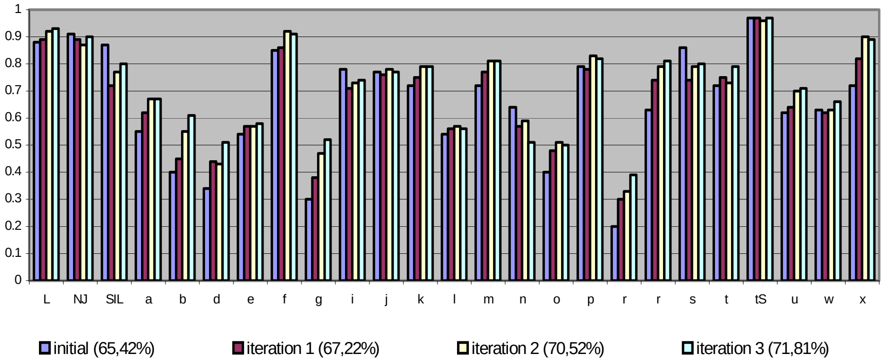

Figure 1. Results gathered at the phonetic level.

3 Preliminary Results

Starting from the 22 acoustic models adapted for Mexican Spanish, we made the first automatic alignment of recordings for a Mexican Spanish corpus [6]. Starting from this data set we compute the acoustic models for Mexican Spanish. For this step the corpus was divided. The training set is formed for 4694 sentences and test set for 1173 sentences. The size of the vocabulary is of 8754 different words. The corpus contains recordings from 100 speakers, for each of them, the corpus contains 60 recordings. Each recording was made with short sentences (3 seconds aprox.) for a total of 5 hours of recording.

[1] The notation used for the phonologic transcription is defined in SAMPA, http://www.phon.ucl.ac.uk/home/sampa

The computation of the models required 43 features: 13 MFCC coefficients (besides their first and second derived), zero crossing, and the energy (plus first and second derived); besides we used the 3 state HMMs. For the computation we use the JANUS toolkit [7].

After the computation of the first version of the models, the process of realignment and computation of new models was repeated 3 more times, gathering a recognition rate at phonetic level of 71.81%. It is important to note that for the case of French, the recognition rate at phonetic level is: 68%. Figure 1 shows the evolution of the results during the four iterations.

4 Conclusions

The treatment of languages in scenarios where resources are limited is extremely important, more over in the context of the Mexican reality. The proposed methodology shows promising results, at least, between Mexican Spanish and French. As further work, we envisage the combination of the knowledge and data based approaches, also the inclusion of other features; the later for both cases, the Mexican Spanish–French as well as Mexican indigenous languages. In the treatment of Mexican Spanish we will also perform some other experiments: (i) redefine the phonetic mapping of “r” and “rr” given that it was not satisfactory with the substitution of correlations r → R and rr → R for: r → l et rr → l; (ii) introduce phonologic variants in the dictionary, currently there is only one slot for each word; (iii) perform a full evaluation with the recognition system at the word level.

Acknowledgements. This work has been partly supported by the project “Man-Machine Spoken Interaction” LAFMI (Laboratorio Franco Mexicano de Informática).

References

1. Manning C. and Schütze, H. (2000). *Foundation of Statistical Natural Language Processing*. MIT Press.
2. T. Schultz, T. and Waibel A. (1998). “*Language independent and language adaptive large vocabulary speech recognition*” Int. Conf. on Spoken Language Processing, Australia.
3. *http://www.speechdat.org*
4. Beyerlein, P., Byrne, W., Huerta, J., Khudanpur, S., Marthi, B., Morgan, J., Peterek, N., Picone, J. and Wang, W. (1999) *Towards language independent acoustic modeling*, ASRU.
5. Wong, E., Martin, T., Svendsen, T. and Sridharan, S. (2003) *Multilingual Phone Clustering for Recognition of Spontaneous Indonesian Speech Utilising Pronunciation Modelling Techniques*, Eurospeech '03, Geneva (Switzerland), pp. 3133-3136.
6. Pineda, L., Cuétara, J., Castellanos, H., López, I., Villaseñor, L. (2004). *DIMEx100: A New Phonetic and Speech Corpus for Mexican Spanish*. IBERAMIA, pp 948-957. Lecture Notes in Artificial Intelligence. Springer-Verlag. (in Press).
7. Rogina, I. and Waibel A. (1995). *The Janus Speech Recognizer*, Proceedings of the ARPA SLT workshop.