DEVELOPMENT OF SPANISH CORPORA FOR SPEECH RESEARCH (ALBAYZIN)[#]

F. Casacuberta[1], R. Garcia[2], J. Llisterri[3], C. Nadeu[4], J.M. Pardo[5] and A. Rubio[6] *

[1]Universitat Politècnica de València (UPV), Dept. DSIC
[2]Universidad Politécnica de Madrid (UPM), Dept. DSSR
[3]Universitat Autònoma de Barcelona (UAB), Dept. DFE
[4]Universitat Politècnica de Catalunya (UPC), Dept. DTSC
[5]Universidad Politécnica de Madrid (UPM), Dept. DIE
[6]Universidad de Granada (UG), Dept. DETC

ABSTRACT

The ALBAYZIN project attempts to overcome the lack of adequate Spanish speech corpora for the development of automatic speech recognition systems. Three different corpora will be designed and collected in this project: phonetically balanced sentences, sentences extracted from a geographical database inquiry task, and speech uttered in noisy environments. In this paper, both the general contents of the three corpora and the tasks of the project are described.

1. Introduction

Speech corpora are greatly needed in the development of automatic speech recognition systems. For a specific language, such as Spanish, and for a specific task, a speech corpus must cover the major sources of variability (acoustic, phonetic, intra and inter speaker) of the speech signal. On the other hand, a speech corpus must contain different types of information along with the signal, such as the corresponding phonetic transcription.

Work supported by the CICYT grant TIC91-1488-C06-04.

* The co-authors are listed in alphabetical order and each of them belongs to a different research group of the ALBAYZIN project. The coordinator of the project is the UPC's group. The other researchers working on that project are: J. Díaz and A. Peinado (UG); S. Aguilera and J. Menéndez-Pidal (UPM-DIE); J. Gómez and J. Santos (UPM-DSSR); N. Prieto, E. Sanchís, E. Segarra and E. Vidal (UPV); D. Poch (UAB); A. Bonafonte, E. Lleida, J.B. Mariño and A. Moreno (UPC).

The goal of a speech corpus is double. Firstly, actual speech recognition systems need a lot of speech data for training and the acquisition is an expensive task. Secondly, the assessment of those systems requires to have a standard corpus to compare their performances, and thus to compare the techniques used by them.

The great majority of currently available speech corpora were conceived under these considerations. Many references of this kind can be found [Fourcin et al., 89], [Garafolo et al., 89], [Kurematsu et al., 89], [Price et al., 89], [Shirai et al., 89], [Zue et al., 89].

It is essential to take into account the hardware and software features that the different speech corpus development projects finally provide to the users. The hardware support of currently available speech corpora is quite far from being standardized, and one can find a large variety of speech corpus media, ranging from traditional analog open-reel magnetic tapes to modern DATs and optical disks. In particular, it should be pointed out that most speech corpora of certain relevance are choosing CDROMs more and more as their end-user media, both for their high capacity and ease of use, as well as for the price, which gets rather low when large amounts of copies of the speech corpora have to be distributed to many users [Gaorofolo & Pallet, 89].

The software features are usually quite limited, though some recent efforts have in fact been devoted to achieve user-friendly data structuring and handling capabilities [Castagneri et al., 89], [Hendriks, 89].

In Spain, only partial efforts have been made up to now to develop speech corpora; neither they have attained a large-scale coverage of the most important sources of variability found in speech nor they allow to train and test a speech recognition system in a rather large task. This lack of adequate Spanish speech corpora for speech research, and particularly for the development of speech recognition systems, has motivated the formation of a consortium constituted by six speech research groups from Spain which are going to design and collect (in collaboration with a Spanish company) a set of speech corpora. The list of groups coincides with the list of affiliations of the co-authors given at the beginning of this paper. The project -named ALBAYZIN- has obtained financial support from the Spanish government and it will be launched in January 1992.

In the following sections, the description of characteristics of the three speech corpora and the planned tasks for their development are presented.

2. General contents of the corpora

The Albayzin database will be composed of three different corpora:

1) The first corpus will consist of a set of 200 phonetically balanced sentences which will be designed to cover a wide range of phonetic variability. Among the factors of variability that are being considered are:

a) speaker-dependent factors as sex, dialect and speaking rate;

b) phonetic factors as preceeding and following allophone, position in the syllable, degree of stress or position in the word.

2) Secondly, an application dependent corpus will be formed with sentences extracted from a geographical database inquiry task. The sentences will be semantically and syntactically constrained and they will include about 1000 different words.

3) The third corpus will consist of a set of frequently used word as well as a number of sentences. Utterances produced by few speakers will be recorded in clean and noisy environments including the Lombard effect.

About 100 speakers will be used for each of the two first corpora. They will be representative of the main geographical and social varieties of Spanish, and will be equally distributed with respect to sex. The final recordings will consist of 5000 sentences, 3 or 4 seconds long, for each of both corpora. The text of these sentences will be selected from spontaneous speech samples.

All three speech corpora will be recorded in CDROM.

3. Description of the tasks

To carry out the ALBAYZIN project, a set of tasks must be defined. The first tasks are oriented to each particular corpus, and correspond to the design stage:

1) Phonetically balanced corpus (Partners: UAB,UPC):

a) allophone selection and statistical study of frequency of occurrence of each allophone;

b) study of factors of phonetic variability; and
c) design of the set of sentences.

2) Task-dependent corpus (Partners: UG, UPV):
a) design of the task;
b) collection of spontaneous sentences of the task through interviews to potential users;
c) analysis of the collected sentences; and
d) analysis of the underlying syntactical structure.

3) Speech corpus recorded in a noisy environment (Partners: UPM-DIE, UPM-DSSR)
a) selection of words and sentences; and
b) choice of the noisy environment.

After this first stage of design, there will be a second stage of collecting speech data including the following steps which are common to all corpora:

a) selection and training of the speakers;
b) acquisition of speech data;
c) checking and endpointing of the acquired speech;
d) phonetic transcription of the uttered sentences (along with ortographic and lexical transcription for the second corpora); and
e) labeling of each utterance.

4. Final remarks

A general description of the ALBAYZIN project has been presented, as well as the corpora that will be developed by it. These corpora can be greatly useful to researchers working in automatic speech recognition in Spanish.

Once this project will be carried out, other projects must follow it in order to get an adequate coverage of the various languages existing in Spain.

References

- Castagneri, G. Vacchetta, L., Di Carlo, A.: "An application of relational database to recognizer testing workstation". *Workshop on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, 20-30 September 1989.

- Eskenazi, M.: "On coordinated assessment efforts in France". *Workshop on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, 20-30 September 1989.

- Fourcin, A.S. and SAM partnership: "Progress overview of the SAM Project". *EUROSPEECH'89*, pp. 308, Paris, 1989.

- Garafolo, J.S., Pallet, R.S.: "Use of CD-ROM for Speech Database Storage and Exchange".*EUROSPEECH'89*, pp. 309-312, Paris, 1989.

- Hendrikd, Jan P.M.: "An acoustic-phonetic formalism for database access". *Workshop on Speech Input/Output Assessment and Speech Databases.* Noordwijkerhout, The Netherlands, 20-30 September 1989.

- Kurematsu, A., Takeda, K., Kuwabara, H., Shikano, K.: "ATR Japansese speech database of speech recognition and synthesis". *Workshop on Speech Input/Output Assessment and Speech Databases.* Noordwijkerhout, The Netherlands, 20-30 September 1989.

- Price, P., Fisher, W., Bernstein, J., Pallet, D.: "The DARPA 1000-word resource management database for continouous Speech Recognition". *Proc. ICASSP'88,* pp. 651-654, 1988.

- Shirai, K., Fujisaki, H., Itahashi, S.: "Speech database projects in Japan -present and future". *Workshop on Speech Input/Output Assessment and Speech Databases,* Noordwijkerhout, The Netherlands, 20-30 September 1989.

- Zue, V., Seneff, S., Glass, J.: "Speech database development: TIMIT and beyond". *Workshop on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, 20-30 September 1989.