Evaluation of the Telefónica I+D Natural Numbers Recognizer over different Dialects of Spanish from Spain and America

C. de la Torre, F. J. Caminero-Gil, J. Álvarez, C. Martín del Álamo and L. Hernández-Gómez[1]

Speech Technology Group
Telefónica Investigación y Desarrollo
Emilio Vargas 6, 28043 - Madrid, SPAIN
e-mail: {celinda, jcam, jorge, cma}@craso.tid.es, luish@gaps.ssr.upm.es

ABSTRACT

In this paper we present the results obtained when evaluating the Natural Numbers Recognizer of Telefónica I+D over some particular dialects of Spanish from Spain and America. The evaluation was made over two different data sets corresponding to two different situations. A first set includes dialects of Spanish from Spain, that were considered in the training and design of our baseline system, and a second set corresponds to Argentinian Spanish, that was not considered to train the original system. Just because we are interested in a system able to be used by a wide range of users, we tested the possibilities of MAP (Maximum-A-Priori techniques) to adapt the original HMMs in order to represent all the dialects. The experimental results show the capabilities of our recognizer to be used in applications spread over a great number of Spanish-speaking countries.

1. INTRODUCTION

The final goal of the Recognition Systems developed in Telefónica of Spain is to work with real customers. For this reason the main difficulty is to deal with all the possible speakers that have access to the system through the telephone line.

Like all the languages spoken and spread in many different and huge countries, the Spanish has a great number of dialectal variations. Sometimes this variations are noticeable from one country to another, and sometimes inside one country, from one region to another.

Our goal is to evaluate the ability of our recognizer system when managing these variations and which performance it provides in the different cases.

With dialectal variation we mean only differences in intonation and pronunciation, but always using the same basic vocabulary.

The studies presented in this paper consisted in two phases. In the first one we have evaluated the behavior of the system with the different dialects of Spain. With the criterion of dialectal similarities, Spain was divided in 7 different regions. The number of pronunciations used was chosen proportional to the population size of the region related to the whole population of Spain, with a total of 3000.

The evaluation over this first set of files has provided good results in all the regions, mostly important in some of them with specially serious variation of pronunciations, like the dialects from Andalusia and from the Canary Isles. It must be taken into account that the dialects evaluated in this first phase have representation in the training corpus.

The second phase, in which we have evaluated the Spanish dialect of Argentina, is a more difficult one. Some more American dialects of Spanish are planned to be evaluated in the near future.

The rest of the paper is organized as follows: Section 2 describes the baseline system. The dialectal division of Spain is presented in Section 3 and the evaluation of these dialects in Section 4. In Section 5, a first evaluation of the Spanish from Argentina is presented and in Section 6, we consider dialect adaptation techniques. Finally, in Sections 7 and 8 future works and conclusions are discussed.

2. BASELINE SYSTEM

The selected front-end was the one used by Telefónica I+D for some of its telephone applications [1] [2] [3]. The Speech Signal digitalized at 8 kHz, is pre-emphasized by a factor of alpha=0.97. The speech is then blocked into frames of 32ms every 16 ms. A total of 18 Mel cepstral parameters are extracted: 8 mel-cepstrums, 8 delta-mel and the energy and delta-energy.

[1] E.T.S.I. Telecomunicación. Universidad Politécnica Madrid.

An initial codebook was obtained by merging the mixtures from bootstrap continuous HMM. Three separated codebooks were trained for each different stream: one for the mel-cepstrums, one for the delta-mel cepstrums and another for the energy and delta-energy. The codebook sizes for the baseline configuration were 180/180/100 gaussians respectively.

The models of the baseline system had a tridiagonal transition matrix.

The grammar used in the recognition process has a language model with a perplexity of 43, because in most cases every word can be followed by any other word of the vocabulary. The application chosen for the experiment was the recognition of telephone numbers. The length of the pronunciations varies from 5 to 10 words, depending on the province to which the number correspond and on the use of area code, for this reason, the grammar has no information about the string length.

In table 1, word error rate and sentence error rate are showed, for the baseline system. These results are detailed for the 3-best candidates.

	1st cand.	2nd cand.	3rd cand.
Word Error Rate	2.1%	1.7%	1.4%
Sentence Error Rate	9.6%	5.7%	4.4%

Table 1: Baseline system.

3. DIALECTS FROM SPAIN: DATABASE

For the evaluation through Spain the database used was the VESTEL database [7]. VESTEL is a telephone speech corpus collected at the Speech Technology Division of Telefónica Investigación y Desarrollo. The database was designed to support research in speaker-independent automatic speech recognition (ASR) based on word and subword units. The database contains speakers throughout Spain, covering all dialects of Castilian Spanish.

For the spontaneously spoken natural connected numbers task, we used the recordings corresponding to telephone numbers. Each file contains a string of numbers. The string length is variable and mostly ranges from 5 to 10 words.

After a careful labelling a total number of 7000 strings from the VESTEL Database were selected for our task. The whole set was split into disjoint training and testing databases: 4000 for training and 3000 for testing, both sets balanced respect to the number of pronunciations of all dialectal regions of Castilian Spanish. “Balanced” means in this case that the contribution to the final set (both for training and testing) was proportional to the population contribution of the corresponding region to the whole of Spain population.

The dialectal division of Spain was made based on phonetical and pronunciation similarities. With this criterion, Spain was divided in 7 different dialectal regions (see figure 1), which are the following:

(1) Castile.

(2) Catalonia, Valencia, Balearic Isles and Aragon.

(3) Extremadura.

(4) Basque Country and Navarre.

(5) Galicia and Asturias.

(6) Andalusia and Murcia.

(7) Canary Isles.

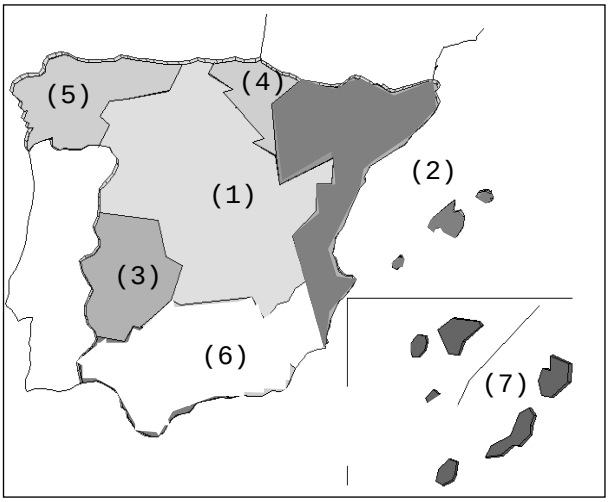

Figure 1: Dialectal division of Spain

The information about the dialect of each speaker was decided based on two questions:

1. Where were you born?
2. Where do you live?

to take a decision about the dialect of the speaker both answers must correspond to the same dialectal region. If not, the dialect of the speaker was labeled with: “Incomplete”. Only in a 73.5% of the calls, the dialect could be identified.

It is important to note that although from a more detailed linguistic point of view, there are more dialect than the ones we have considered, we have grouped some similar zones to achieve statistically significant results.

4. DIALECTS FROM SPAIN: EVALUATION

In table 2 are presented the results obtained separately for every dialect.

Region	1st cand.	2nd cand.	3rd cand.
Castile	9.8%	6.8%	5.5%
Catalonia-Valencia-Balearic Isles-Aragon	6.9%	3.5%	3.0%
Extremadura	11.1%	4.4%	2.2%
Basque Country-Navarre	10.6%	4.1%	3.3%
Galicia-Asturias	12.7%	7.9%	5.5%
Andalusia-Murcia	10.2%	5.1%	4.2%
Canary Isles	14.29%	9.52%	9.52%

Table 2: Results for the different dialectal regions.

It can be observed some deviations in the performances from one dialect to another. As reference we took the Central Castilian dialect, being the most extended and the one used in mass communication media. The principal deviations can be observed in those dialects with greater differences in terms of pronunciation (Andalusian and Canarian) and in those with less percentage of appearances in the training set (Extremaduran and Canarian).

The Canarian has some peculiarities closer to some speaking styles of South America than those from Spain. It has some exclusive sounds or phonemes, like the sound /jj/ , when pronounced the letter “ch”, that in the training set only appeared in the pronunciations corresponding to this region.

Some frequent problems occurred when dealing with andalusian speakers is the constant crossing and confusion between the sounds: /s/ and /θ/, corresponding to the letters: “s” and “z”, respectively. This produces alternative pronunciation for some of the words: “siete” and ″θiete″, for the digit “7”, ″θero″ and “sero” for the digit “0”.

Another important problem is the aspiration and disappearance of the final /s/ sound, like “tre” from “tres” (digit 3), “do” from “dos” (digit 2). Some great problems are inducted from this behavior, increasing the acoustic confusion between the HMMs. In this case, for example, the word “cientos” becomes identical to the word “ciento”.

5. ARGENTINIAN SPANISH

The database with speakers from Argentina was collected over the telephone line and it contains a total of 1181 pronunciations corresponding to natural pronunciations of telephone numbers.

The results obtained are presented in table 3. They must be compared with those of our baseline system (see table 1).

The more relevant and differential characteristics of argentinian speakers are the following:

- Intonation: they usually have a high decrease in the fundamental frecuency value at the endings of the pronunciations. Some time those decreases are so strong that the endpoint detector misses the endings of the pronunciation.
- Aspiration and omission of fricatives, like final /s/.
- Disappearance of sound /θ/ , substituted by /s/.

	1st cand.	2nd cand.	3rd cand.
Word Error Rate	2.7%	1.7%	1.5%
Sentence Error Rate	10.4%	5.2%	4.0%

Table 3: Results for Argentinian dialect with baseline system.

The performance decrease in the evaluation of this dialect is not worse than those corresponding to dialects from Spain. Special attention must be put in the particular Word Error Rate of some words of the vocabulary that have suffered an increase over the mean. Such an example is the “cero”, which in Argentina is pronounced as “sero” all over the country, since in Spain the percentage corresponding to this pronunciation is 20%. In the case of Argentina, an effort in training a particular “sero” would be interesting.

Also, the substitution of sound /θ/ by /s/ produces a considerable decrease of discrimination between some models. For example the difficulty when discriminating between the words “dos” and “doce” is considerably increased when the latter is pronounced like “dose”.

6. RETRAINING AND ADAPTATION TECHNIQUES

Two techniques were considered in order to adapt the original models to the new group of speakers: retraining with Maximum Likelihood and MAP.

To test this techniques, we have perform a recognition test with a simpler grammar, only formed by the ten basic digit. In this task, the global recognition rate do not worsen for the argentinian dialect, but the detached word rates are quite higher for some words. To test these adaptative techniques, we choose the case of the word “cero”, in order to outperform our system.

- ML: Maximum Likelihood

In this case the model are re-trained using only the new argentinian database collected with the particular group of speaker we want to adapt the models to.

The results obtained (see table 4) show that it produces a considerable improvement in performance with the new speakers group, while the generality of the original models has been lost. The performances with the initial castilian recognition set worsened considerably.

- MAP: The use of Adaptive Techniques [4] [5] [6] produced a greater improvement in the WER of word “cero” for the argentinian recognition set, while the castilian recognition set do not worsen substantially respect to the baseline system. Besides, the MAP techniques need much fewer training utterances than the ML re-estimation. In our MAP re-estimation, we employ only 70 training utterances.

In Table 4, we present the global results, in terms of Word Error Rate (WER) and Sentence Error Rate (SER), from applying these techniques to the model of the word “cero”, and in Table 5 the detached results for this word are showed.

System	Argentinian		Castilian	
	WER	SER	WER	SER
Baseline	0.99%	5.16%	0.95%	5.48%
ML	0.83%	4.10%	2.50%	14.8%
MAP	0.87%	4.20%	1.08%	6.66%

Table 4: Results from applying MAP and ML techniques to re-estimate the model of the word “cero”.

WER for “cero”	Argentinian	Castilian
Baseline	4.9%	0.6%
ML	1.8%	19.0%
MAP	0.6%	2.3%

Table 5: Detached WER results for the adapted word “cero”.

7. FUTURE WORK

The major research lines we plan to further improve our system are the following:

- Generalization of the MAP training to the whole model set.
- Smoothing techniques will be applied to the original and new models obtained applying MAP.
- Applying MAP techniques to the dialect of the Spanish from Spain, specially from the Canary Isles, Extremadura and Galicia regions.

8. CONCLUSION

The results, in terms of performance, of the Natural Numbers Recognizer of Telefónica I+D when evaluated over several dialects from Spain and America, have shown the viability of using this recognizer in services that would work in many different places and countries.

The use of some re-training techniques shows a quick way to adapt the models to new dialects in order to improve those particular behaviors which have worsened severely.

As major result we found that in most of the dialects, the recognition performance is close to the one obtained for the Castilian Spanish, which is being the base of several real Telefónica I+D applications already in use. Therefore, we conclude that similar applications will be successfully mounted in other Spanish-spoken zones.

REFERENCES

[1] C. de la Torre, L. Hernández-Gómez, F.J. Caminero-Gil and C. Martín-del Álamo. "Recognition of Spontaneously Spoken Connected Numbers in Spanish over the Telephone Line", EUROSPEECH-95, Madrid, pp. 2123-2126.

[2] F.J. Caminero, C. de la Torre, L. Hernández and C. Martín. “New N-Best based Rejection Techniques for Improving a Real-Time Telephonic Connected Word Recognition System”, EUROSPEECH-95, Madrid, pp. 2099-2102.

[3] C. de la Torre, F.J. Caminero-Gil, L. Hernández-Gómez and C. Martín del Álamo. "On-line Garbage Modeling for Word and Utterance Verification in Natural Numbers Recognition", ICASSP-96, Atlanta, pp. 845-848.

[4] C. H. Lee and J. L. Gauvain. "Speaker Adaptation based on MAP Estimation of HMM Parameters", ICASSP-93, Minneapolis, pp. 558-561.

[5] C. H. Lin, P.C. Chang and C. H. Wu. "An Initial Study on Speaker Adaptation for Mandarin Syllable Recognition with Minimum Error Discriminative Training", ICSLP-94, Yokohama, pp. 307-310.

[6] T. Matsui and S. Furui. "A Study of Speaker Adaptation based on Minimum Classification Error Training", EUROSPEECH-95, Madrid, pp. 81-84.

[7] D. Tapias, A. Acero, J. Esteve and J.C. Torrecilla. “The VESTEL Telephone Speech Database”, ICSLP-94, Yokohama, pp. 1811-1814.