

Detection of Recognition Errors and Out of the Spelling Dictionary Names in a Spelled Name Recognizer for Spanish

R. San-Segundo, J. Macías-Guarasa, J. Ferreiros, P. Martín, J.M. Pardo

Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica. UPM
E.T.S.I. Telecomunicación. Ciudad Universitaria s/n, 28040 Madrid, Spain

lapiz@die.upm.es

<http://www-gth.die.upm.es>

Abstract

This paper deals with improved confidence assessment for detecting recognition errors and out of dictionary names in a Spanish Recognizer of continuously spelled names over the telephone. We present a hypothesis-verification approach for spelled name recognition. We evaluate the system for several dictionaries, obtaining more than 90.0% recognition rate for a 10,000 name dictionary. For confidence scoring, we consider several features obtained from the different recognition stages. The paper investigates the ability of each feature set to detect recognition errors and names out of the spelling dictionary. We use a neural network to combine all the features in order to obtain the best confidence annotation. Using the data collected from 1,000 phone calls, it is shown that 57.9% incorrectly recognized names and 68.3% out of the spelling dictionary names are detected at a 5% false rejection rate.

1. Introduction

Accurate spelling recognition of telephone speech represents a challenging task very useful for many applications such as directory assistance [1], or identification of city names for travel services [2]. Natural spelling implies the recognition of connected letters. This is a difficult task, especially over the telephone, because of the confusable letters contained in the alphabet, the distortions introduced by the telephone channel and the variability due to an arbitrary telephone handset.

Currently, spelled name recognition systems are being widely used as a fall back strategy [3], and for detecting data out of the spelling dictionary [4]. In these situations a high level of accuracy is required. Because of this, approaches based on several recognition stages are usually used [5][6] and long range language models are incorporated [7].

In English, the main difficulty in the spelling recognition task lies in the recognition of the E-set = {B, C, D, E, G, P, T, V, Z}. In [8], we presented a detailed description of the spelling task for Spanish and a wide analysis of different recognition architectures. In this paper we present several improvements over the hypothesis-verification architecture and we introduce an approach for detecting recognition errors and names out of the spelling dictionary.

The paper is organised as follows. Section 2 introduces the recognition system architecture, describing the main modules (hypothesis and verification) and the final recognition rates obtained. In this section we also present the database used to train and evaluate our alphabet recognizer. Section 3 presents the features proposed for estimating the confidence measures and the method used to combine all features in order to get a unique confidence value. Section 4 shows the results for the experiments of recognition errors detection, evaluating each of

the proposed feature sets. In section 5, we describe the experiments for detecting names out of the spelling dictionary. Experiments for detecting simultaneously recognition errors and out of the dictionary names are shown in section 6. Finally, in Section 7 we review the conclusions of this work.

2. The Spelled Name Recognition System

The system proposed is based on a hypothesis-verification approach similar to [6]. In the hypothesis stage, we obtain N-best sequences of letters given acoustic HMMs of the letters and we compare these sequences with all the names in the dictionary using a dynamic programming algorithm to obtain the M-best similar names. These names are passed to the verification stage. In this stage, a dynamic grammar is built with the M-best names and the HMM recognizer is used again with this highly constrained grammar. In figure 1, we can see the block diagram of the system.

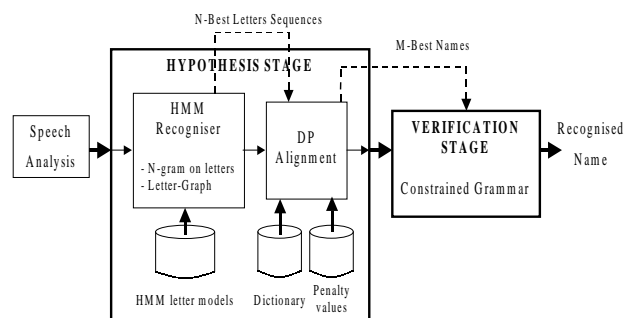


Figure 1. Block Diagram of the system.

2.1. Speech Analysis

We use a 10 ms frame shift and a 25 ms analysis window. In the experiments, the static cepstral coefficients, the local energy, the first derivative of the energy and the first derivative of the static cepstral coefficients are considered to build-up the speech parametric representation (22 coefficients). We use the RASTA-PLP parameterisation as proposed in [6] where we can see a detailed Front-End analysis for the spelling task.

2.2. Hypothesis Stage

In the HMM recognizer we use letter continuous HMMs (C-HMMs) with a number of states proportional to the length of each letter (average number of frames). The shortest model has 9 states and it is associated to the vowel letter I. The longest one has 48 and is associated to letter W. The number of mixtures per state is proportional to the amount of data used for training. We consider a minimum number of three

mixtures and a maximum of nine. The HMM training has been carried out by a standard training procedure.

In this stage we consider a Letter Graph Generation based on the ideas proposed in [9] to incorporate the N-gram language model in the search space and to calculate N-best letter sequences with low computational demands.

Once obtained the N-best letter sequences, our target is to get the M-best names from the dictionary. The N letter sequences are compared against all the dictionary names using a Dynamic Programming (DP) algorithm. The DP algorithm applies different penalties for substitutions, deletions and insertions errors in the letter sequence. Given a letter sequence, the dictionary names with lowest alignment cost are selected.

2.3. Verification Stage

In this stage, we use the M-best candidate names to build a dynamic grammar and the HMM recognizer is invoked again with this constrained grammar. In this stage, the time consumption is low because the number of names considered is small and we use the same HMMs as in the hypothesis stage, so that the state distribution probabilities have already been stored to be used here. Of course, more detailed models or different recognition parameters can be used in the verification stage.

2.4. Recognition Experiments

The database used for the experiments is SpeechDat [10]. It was recorded from the Spanish telephone network using 1000 speakers. Each speaker was asked to spell a city name, a proper name and a random letter sequence, which guarantees a minimum number of training examples for each letter. From the city and proper names audio files we have randomly selected 600 files for two evaluation sets, the first one (300) for adapting the DP algorithm penalties and the second (300) for development. We also use 300 for final testing, leaving the rest (2100 audio files) for HMM training. We have repeated the evaluation six times providing a 6-Round Robin training to verify the results. In table 1, we present the average results of these experiments. The confidence interval of the results at 95% is $\pm 1.4\%$.

Size of the dictionary	Hypothesis Stage	Hypothesis and Verification Stages	M
1,000 (0.2)	94.2 %	96.3 %	10
5,000 (0.5)	88.7 %	92.8 %	20
10,000 (0.9)	86.2 %	90.3 %	50

Table 1: Recognition results considering only the hypothesis stage and the whole system.

We consider several dictionaries of different sizes (1,000, 5,000 and 10,000 city and proper names) obtained by random selection from the Spanish city and proper name directory. The average confusion (in parentheses) for the dictionaries is 0.2, 0.5 and 0.9 respectively. These values are a measure of dictionary confusion [6] and are calculated as the average number of name pairs from the dictionary that differ only by one letter substitution. In the third dictionary (10,000 names), there were 9,038 pairs of names that differ only by one letter substitution. This corresponds to an average of 0.9 confusions per name.

3. Features for Confidence Annotation

We consider features coming from both recognition stages: hypothesis and verification. In [11], we can see a detailed analysis of features for confidence annotation in hypothesis-verification recognition systems.

3.1. Hypothesis Stage

In this stage we propose features from the HMM recognizer and the DP alignment. From the HMM recognizer (F-1):

- **Best Score (BS-1)**: acoustic score of the 1st letter sequence divided by the number of frames.
- **Score Difference (SD-1)**: acoustic score difference between the 1st and 2nd letter sequences divided by the number of frames.

From the DP alignment we have considered (F-2):

- **Best Cost (BC-2)**: lowest alignment cost between the N-best letter sequences and the names of the dictionary divided by the length of the 1st letter sequence.
- **Cost Difference (CD-2)**: difference between the two best alignment costs divided by the length of the 1st letter sequence.
- **Cost Mean (CM-2)**: average cost along the 50 best alignment costs divided by the length of the 1st letter sequence.
- **Cost Variance (CV-2)**: cost variance along the 50 best alignment costs divided by the length of the 1st letter sequence.

3.2. Verification Stage

In this stage we propose the following features (F-3):

- **Candidate Score (CS-3)**: acoustic score for the best candidate name obtained after the verification stage divided by the number of frames.
- **Candidate Score Difference (CSD-3)**: acoustic score difference between the two best candidates obtained in the verification stage divided by the number of frames.
- **Candidate Score Mean (CSM-3)**: average score along the 50 best candidate names divided by the number of frames.
- **Candidate Score Variance (CV-3)**: score variance along the 50 best candidate names divided by the number of frames.
- **Score Ratio (SR-3)**: score difference between the score of the 1st letter sequence (hypothesis stage) and the score of the best candidate name (verification stage) divided by the number of frames.

3.3. Feature combination

We have considered a Multi-Layer Perceptron (MLP) to combine the features in order to obtain a single confidence value. In this case we use the features directly as inputs to the MLP. A preprocessing has been used to limit the dynamic range of each measure to the (0,1) interval. Here, the normalization consists of scaling utilizing the minimum and maximum value obtained for each feature in the training set. The hidden layer consists of 10 units and one output node was used to model the name confidence. During weight estimation, a target value of 1 is assigned when the decoder correctly recognizes the name (or when the name is in the dictionary for experiments in section 5) and a value of 0, is assigned during incorrect recognition (or name out of the spelling dictionary respectively). The database used in the confidence experiments has been built with the results

obtained in the recognizer evaluation for the 10,000 name dictionary. In this case, we have 1,800 examples (300 from the testing set along six experiments), considering 1,200 for training the MLP, 300 as the evaluation set and 300 for testing. We have repeated it six times providing a 6-Round Robin training to verify the results and increase the statistical significance. In this paper we present the average results of these experiments.

4. Recognition Error Detection

In this section we investigate the ability of the features proposed in section 3 for detecting of recognition errors. Table 2 summarizes the correct detection rates of recognition errors at false rejection rates of 2.5% and 5.0%. It can be seen that features from the verification stage provide better indicators of confidence than hypothesis stage features.

Features		Correct Detection Rates		Minimum Classification Error Baseline 9.7%
		2.5% FR	5.0% FR	
Hypothesis	F-1	7.1%	12.9%	9.7%
	BC-2 CD-2	20.0%	26.5%	9.4%
	CM-2 CV-2	18.3%	26.0%	9.4%
	F-2	22.3%	29.5%	9.2%
Verification	CS-3 CSD-3	40.5%	54.3%	8.0%
	CSM-3 CSV-3	27.1%	38.2%	9.0%
	SR-3	30.1%	37.4%	9.0%
	F-3	46.7%	57.4%	7.6%
	F-2 and F-3	44.7%	57.9%	7.5%

Table 2. Correct detection of recognition errors at a 2.5% and 5.0% false rejection rate (FR) rate. Minimum classification error and Baseline Error are also shown.

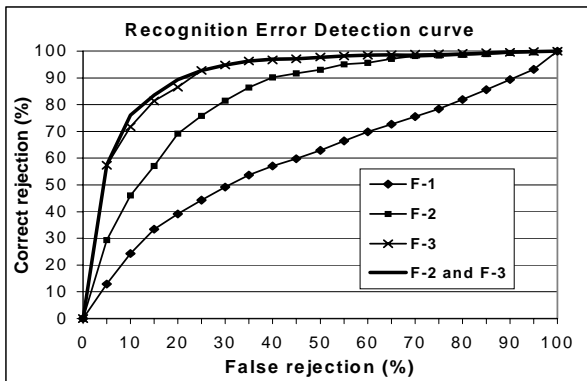


Figure 2. Correct Rejection vs. False Rejection of recognition errors for features from the hypothesis (F-1, F-2) and verification stages (F-3). Results combining F-2 and F-3 features are also shown

Considering all features from the verification stage, more than 57% of recognition errors are detected at 5% false rejection. Keeping in mind that the baseline error rate is 9.7% these features reduce the classification error 2.1 points. The best results are obtained combining features from the DP algorithm (F-2) and features from the verification stage (F-3). We report a 7.5% classification error for these case. We have not included F-1 features because of the low discrimination power. In Figure 2, we plotted the detection error trade-off curves.

5. Detection of names Out of the Spelling Dictionary

In this section we investigate the ability of the features proposed in section 3 for detecting names out of the spelling dictionary. To simulate names out of the spelling dictionary we have randomly removed names from the spelling dictionary. The names removed constitute around the 20% of the testing cases. In these experiments, we have tried to detect the cases where the name was removed from the dictionary. Table 3 summarizes the correct detection rates at false rejection rates of 2.5% and 5.0%.

Features		Correct Detection Rates		Minimum Classification Error Baseline 21.5%
		2.5% FR	5.0% FR	
Hypothesis	F-1	2.9%	5.7%	21.5%
	BC-2 CD-2	17.6%	33.4%	17.7%
	CM-2 CV-2	3.0%	5.3%	21.5%
	F-2	17.5%	34.5%	17.7%
Verification	CS-3 CSD-3	9.3%	15.5%	21.5%
	CSM-3 CSV-3	3.0%	6.3%	21.5%
	SR-3	53.0%	66.3%	11.2%
	F-3	53.5%	67.9%	10.9%
	F-2 and F-3	56.2%	68.3%	10.9%

Table 3. Correct detection of names out of the spelling dictionary at a 2.5% and 5.0% false rejection rate (FR). Classification error and Baseline Error are also shown.

The Score Ratio is the best feature for detecting out of the dictionary names. With only this feature more than 67% of recognition errors are detected at 5% false rejection. Considering that the baseline error rate is 21.5% this feature reduces the classification error in 10.6 points (50.7% relative classification error rate reduction). The best results are obtained combining features from the DP algorithm (in the hypothesis stage) and features from the verification stage. We report a 68.3% correct rejection rate at 5% false rejection rate. We have not included F-1 features because of the low discrimination power. In Figure 3, we plotted the detection error trade-off curves.

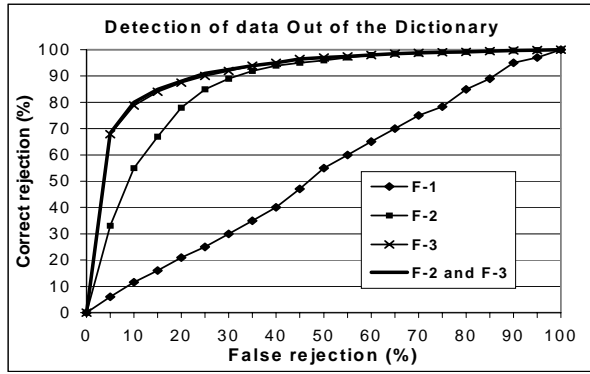


Figure 3. Correct Rejection vs. False Rejection curve for out of the dictionary name detection considering features from the hypothesis (F-1, F-2) and verification stage (F-3). Results combining F-2 and F-3 features are also shown.

6. Detection of Recognition Errors and Names Out of the Spelling Dictionary simultaneously

In this section we present the results for two new experiments. In the first one, we consider F-2 and F-3 sets of features for detecting recognition errors or names out of the spelling dictionary.

Features	Correct Detection Rates		Classification Error Baseline 29.2%
	2.5% FR	5.0% FR	
F-2 and F-3 sets	54.8%	65.8%	13.1%

Table 4. Correct detection of recognition errors or names out of the dictionary and Minimum classification error

In this case the relative classification error rate reduction is 55.1% higher than those obtained in previous sections. In the second experiment, we try to discriminate between correct, incorrect and out of the dictionary name using F-2 and F-3 sets of features. We consider a Multi-Layer Perceptron with three outputs. During weight estimation, a target value of 1 is assigned to the first output when the decoder correctly recognizes the name and 0 for the rest of cases, a target value of 1 is assigned for the second and third outputs when the name is incorrectly recognized or the name is out of the spelling dictionary respectively. In Table 5, we show the confusion matrix obtained for this experiment.

Confusion Matrix			
	CRN	IRN	ODN
CRN	94.9% (1,213)	0.9% (13)	4.2% (52)
IRN	49.7% (68)	18.0% (25)	32.3% (44)
ODN	24.0% (92)	3.6% (14)	72.4% (279)

Table 5. Confusion matrix for name classification as Correctly Recognized Name (CRN), Incorrectly Recognized Name (IRN) or Out of Dictionary Name (ODN).

As we can see, the main problems appear in the incorrectly recognized name detection. In this case the MLP typically classifies the name as correct or as out of the dictionary. The main error in out of dictionary name detection is to consider the name a correctly recognized

7. Conclusions

In this paper we present a new version for the Spanish Spelled Name Recognizer over the telephone introduced in [8]. We evaluate the system for several dictionaries, obtaining more than 90.0% recognition rate for a 10,000 name dictionary. In this system we analyse different feature sets for confidence annotation. The confidence scoring is used for detecting recognition errors and names out of the spelling dictionary. We detect 57.9% of incorrectly recognized names and 68.3% of names out of the spelling dictionary at a 5% false rejection rate. For incorrect names detection, the best feature was the difference between the scores of the first and second candidates (CSD-3). For detecting out of dictionary names the best feature was the Score Ratio (SR-3), obtaining with it similar results than considering F-2 and F-3 feature set simultaneously.

Trying to discriminate between a recognition error and a out of dictionary name is a difficult task because of the great similarity in both behaviors.

8. Acknowledgments

The authors are very grateful to Rubén Pacheco for his help implementing some of the algorithms.

9. References

- [1] Lamel, L., Rosset, S., Gauvain, J.L., Bennacef, S., Garnier-Rizet, H., Prouts, B., "The LIMSI ARISE system". Speech Communications. Vol 31, No 4 pp 339-355. 2000
- [2] Schrâmm, H., Rueber, B., and Kellner, A., "Strategies for name recognition in automatic directory assistance systems". Speech Comm. Vol 31, No 4 pp. 329-338.2000
- [3] Bauer, J.G., Junkawitsch, J., "Accurate recognition of city names with spelling as a fall back strategy". In Proc. EUROSPEECH. pp. 263-266. 1999.
- [4] Jouvét, D., Monné, J., "Recognition of spelled names over the telephone and rejection of data out of the spelling lexicon". EUROSPEECH. pp. 283-286. 1999.
- [5] Mitchell, G., Setlur, A.R., "Improved spelling recognition using a tree-based fast lexical match". In Proc. ICASSP, pp. 597-600. 1999.
- [6] Junqua, J.C., "SmarTspell™ : A Multipass Recognition System for Name Retrieval over the Telephone". IEEE Trans. Speech and Audio Proc, Vol. 5, No. 2, Mar 1997.
- [7] Thiele, F., Rueber, B., Klakow, D., "Long range language models for free spelling recognition". In Proc. ICASSP, pp. 1715-1718. 2000.
- [8] San-Segundo, R., Colás, J., Ferreiros, J., Macías-Guarasa, J., Pardo, J.M., "Spanish Recogniser of continuously spelled names over the telephone". ICSLP 2000. 863-866
- [9] Ney, H., and Ortmanns, S., "Dynamic programming search for continuous speech recognition". IEEE Signal Processing Magazine, Vol 16 No 5 pp 64-83. 1999.
- [10] Moreno, A. *SpeechDat* [cd-rom]. Ver. 1.0. [Barcelona]: Universitat Politècnica de Catalunya <<http://www.upc.es/castella/recerca/recerca.htm>>, 1997. 4 cd-roms. (Spanish Fixed Network Speech Corpus).
- [11] J. Macías-Guarasa, J. Ferreiros, R. San-Segundo, J. M. Montero y J. M. Pardo. "Acoustical and lexical based confidence measures for a very large vocabulary telephone speech hypothesis-verification system" In Proc ICSLP, Beijing, China. 2000.