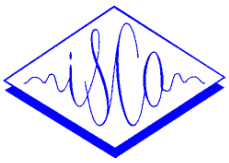

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

Recognition of Spontaneously Spoken Connected Numbers in Spanish over the Telephone Line

C. de la Torre, L. Hernández-Gómez (*), F. J. Caminero (*) and C. Martin del Alamo (*)

e-mail: celinda@craso.tid.es
Speech Technology Group
Telefónica Investigación y Desarrollo
Emilio Vargas 6, 28043-Madrid, Spain

Abstract

In this paper, we describe the results obtained with a pioneer application for spontaneously spoken Connected-Number recognition in Castilian Spanish over the telephone line. The results of applying well known techniques in other recognition tasks to our system showed improvements of nearly 68% (including N-Best techniques) comparing with our baseline SCHMM system, that represents a final Word Error Rate of 1.5%, which convert the system in a feasible one to be portable for commercial applications.

1. Introduction

Importance and popularity on Interactive Voice Response Systems (IVR) are daily increasing. They allow complicated transactions and information exchanges to be developed between customers and remote information systems by using simply a telephone and their voices. For many Applications of Speech Recognition over the telephone line, i.e. voice dailing, credit and account number based transactions, secret transactions based on a PIN or in a ID number; is essential to provide connected numbers recognition pronounced in a spontaneous way. The success of various connected digit recognition systems [1] [2] on non-telephone databases recorded under controlled conditions [3], encourages the extension of these systems and related techniques on telephone Databases [4] [5]. Recently, one of these related techniques, Semi-continuous Hidden Markov Modelling (SCHMM) [6], has been showed as a good alternative between the high-number of parameters and computational complexity of continuous HMM and the low accurate estimation of the output probabilities density functions and the sub-optimal parameter optimization in Discrete HMM.

We present the results of a spontaneously spoken connected-numbers task in Castilian Spanish over our SCHMM based connected speech recognizer.

The experiments were held on telephone recordings. We present a study of several techniques related to the String-error-rate (SER) and to the Word-Error-Rate (WER) reduction using SCHMM such as: spectral features, tied state models, the use of multiple candidates, models topology, spectral normalization, gender modelling and Noise Spotting.

Note: the WER is measured as the relation between the sum of total insertion, deletions and substitution of words over the total number of words in the evaluation set.

2. Telephone Database

Until now, there are no widely-used Spanish Telephone Databases, as the TI Digits Data Bases, to be used as standards to evaluated and compared different recognition algorithms. In the proposed contribution we will use as reference the VESTEL database [7].

VESTEL is a telephone speech corpus collected at the Speech Technology Division of Telefónica Investigación y Desarrollo. The data base was designed to support research in speaker-independent automatic speech recognition (ASR) based on word and subword units. It was designed to support research in speaker-independent ASR of isolated words of small, medium and large vocabularies, and also ASR of connected digits and numbers in order to be able to introduce new services in the Spanish Telephone Network. The database contains speakers throughout Spain, covering all dialects of Castilian Spanish.

For the spontaneously spoken connected numbers task, we used the recordings corresponding to telephone and driving license 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.

The vocabulary is constituted by the 43 words needed in Spanish to form all the natural numbers (see table 1) and includes many high confusible sets of words, i.e. the "-ce" set (see table 2), the "-ay" pairs set (see table 3), the "cien-" set (see table 4). Note that some pairs have very similar pronunciation, i.e. "veinte" and "veinti", "sesenta" and "setenta".

The appearance frecuency of the vocabulary words varies greatly from one to another. For example, digits can appear alone or completing the tens and the cents. For this

(*) E.T.S.I. Telecomunicación de Madrid, Spain

10.21437/Eurospeech.1995-508

reason, we would find in real applications, as our database showed, that there are approximately hundred appearances of digits for each cent.

cero	uno	dos	tres	cuatro
cinco	seis	siete	ocho	nueve
diez	dieci	once	doce	trece
catorce	quince	veinte	veinti	treinta
treintay	cuarenta	cuarentay	cincuenta	cincuentay
sesenta	sesentay	setenta	setentay	ochenta
ochentay	noventa	noventay	ciento	cientos
cien	nove	sete	quinientos	mil
millón	millones	silence		

Table 1: Vocabulary

once	doce	trece	catorce	quince

Table 2: The "-ce" set

treinta	cuarenta	cincuenta	sesenta
treintay	cuarentay	cincuentay	sesentay
setenta	ochenta	noventa	
setentay	ochentay	noventay	

Table 3: The "-ay" pairs set

cien	ciento	ciento

Table 4: The "cien-" set

Moreover, plenty of people, when asked to say their telephone number or their credit card number, use to say it by using only digits. After a statistical study of our recorded database, it showed that approximately the 37% of the speakers say their number by using only digits. Therefore, we consider that an evaluation of our system with a connected digits task could be also interesting for some specific application and because it show the performance of the system in a high percentage of the cases in a spontaneously spoken connected-numbers task. As an example we presented the results of our baseline system for this task (see table 5). In counterpart, some tasks need to have natural numbers recognition, i.e. prices or the amount of money in a credit card transaction. For this reason, the results presented in this paper are mainly focussed in the Connected-Number Task.

3. Baseline System

The selected front-end was the one used by Telefónica I+D for some of its telephone applications. The Speech Signal digitized 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.

An initial codebook was obtained by merging the mixtures from bootstrapped 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 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 this 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.

4. Experiments and Results

The performance of the baseline system is given in Table 5.

	Word Error Rate	String Error Rate
Connected-Number Task	3.82	14.9%
Connected-Digit Task	1.6%	9.5%

Table 5: Baseline System Performance

The better results of the connected-digits application are dued to the specific acoustic characteristics of the digits set, they constitute a less confusible group comparing with others, i.e. the "-ay" group (see table 3). Furthermore, detailed studies showed also that digits are usually confused with other digits rather than with the rest of the vocabulary, due to their shorter lenght.

Following, we will describe the most notable techniques added to our baseline system and the improvements obtained for the Spontaneously Spoken Connected-Number Task.

1. Spectral Features

A first improvement of performance of our baseline system were obtained by increasing the acoustic resolution. We increase the feature vector with 8 additional second order derivatives of the mel-cepstrum. The use of a codebook size of 180 for the new stream decreases the word error rate by 9.5%. The final goal of our system is to be a commercial product that must run in real time. Implementation restrictions forced us to

abandon this way of experiments, so we must renounce to some performance improvements due to implementation issues. For this reason, error rate improvements produced by the use of second derivatives has not been included in the evolution of the final system.

The same kind of limitations forced the vector codebook size to be fixed at 180/180/100 gaussians for the final system.

2. Tied State Models

To further improve the recognition results we need to exploit the special characteristics of our vocabulary. To increase the discriminative behaviour of our models we used Tied States Models, to share common roots and endings of vocabulary words as proposed in [5], an example of this are the ending "enta" or "entay", common to some of the tens, for example "cuare-enta", "cincue-enta", etc. See table 3.

Tied states decreased the WER by nearly 19% and the SER by 7% for the Connected-Numbers Task. Results can be seen in table 6.

WER	SER
3.1%	13.9%

Table 6: Performance of the System using Tied States Models for the Connected-Number Task.

3. Multiple Candidates

We have implemented a Lattice N-best algorithm adapted to our recognition network. We have supposed that a dialogue with the user in the final application would let us use the 3-Best Candidates obtaining in the recognition process. With this algorithm and using three candidates the SER reduction with the second candidate was of 38% and of 53% with the third. For the WER they were 15.6% and 29% respectively. Table 7 shows the performance of the system.

	WER	SER
1st Candidate	3.1%	13.9%
2nd Candidate	2.6%	8.6%
3rd Candidate	2.2%	6.5%

Table 7: Performance of the System with 3-Best and Tied States Models for the Connected-Number Task.

4. Models Topology

Unlike the results presented in [4] we found that skip transitions in our models were important to face the high coarticulation effects between words. We obtained an increase in the word error rate when skip transitions were removed, although the number of insertions were reduced. This effect can be explained with the fact that models without skip transitions force a longer duration of the pronunciations.

5. Spectral Normalization

Spectral Normalization has been demonstrated to be a very simple and effective way to provide robust speech recognition. Initially to have a reference, a non-real time CMN algorithm was used with a decrease of nearly 18% in the WER and of 12% in the SER. This improvements have been calculated using 3-Best candidates. Secondly a on-line CMN (RT-CMN) feasible for real-time applications was implemented for the final system with lower but reasonable improvements. In this case the initial mean cepstral is adapted with each new frame of speech. The improvements are of 9% in the WER and 4% in the SER. The results with N-Best are shown in table 8.

	CMN		RT-CMN	
	WER	SER	WER	SER
1st Candidate	2.5%	12.2%	2.8%	12.7%
2nd Candidate	2.1%	6.8%	2.2%	7.0%
3rd Candidate	1.8%	5.7%	2.0%	6.2%

Table 8: Compared Performances of the System when using different CMN Techniques for the Connected-Number Task.

6. Gender Models

Our aim was to obtain an improvement by using separate models for male and female voices but without an important increase in complexity. Computational restrictions led us to experiment with several automatic male/female discriminators. The best results were obtained with a discriminator based on the senon output probabilities for the speech frames, which permits to take a decision after de first 600 msec of speech (normal telephone number pronunciations are longer than 2.5 sec.) with a 92% of correct classification. After 600 msec. half of the models, those corresponding to the gender that has not been detected by the discriminator are pruned till the end of the pronunciation, so that the computational cost of the recognition process is half reduced. Main advantages of this method are that it doesn't need extra computational cost and that the discrimination criterion is the same that the one used for recognition, that are the output probabilities given by the HMMs. To compare with this result, we made another experiment without using the classification module, at the expense of doubling the number of models. The results show that gender classification errors produce only a low increase of 15.7% in WER and of 8.9% in SER respect to the use of gender models without the discriminative module. The compared results can be shown in table 9.

The final implemented system with male/female discriminator has improvements of 20% for WER and 22.6% for SER respect to the previous system.

	No Gender Discriminator		With Gender Discriminator	
	WER	SER	WER	SER
1st Candidate	2.1%	10.0%	2.4%	10.9%
2nd Candidate	1.7%	5.8%	1.9%	6.3%
3rd Candidate	1.4%	4.5%	1.6%	4.8%

Table 9: Compared Performances of the System with Gender Models with and without gender Discriminator. Connected-Number Task.

7. Noise Spotting

The system must work with real users over the telephone line. In this situation the pronunciation will come not alone but embedded with noises, some of them from the environment and some produced by the user (lip noises, smacks, etc.). For this reason we trained a HMM to characterize them. More details of this technique can be obtained in [9]. The final performance of the system are presented in table 10. For the final system results for the Connected-Digit Task are also included.

The improvement of Noise Spotting over clean pronunciations is low, but it increases considerably the robustness of the system when dealing with noisy pronunciations and environments [9].

	Word Error Rate	String Error Rate
Connected-Number Task	1.5%	4.7%
Connected-Digit Task	1.0 %	4.0 %

Table 10: Final System Performance

5. Summary

Table 11 shows a summary of the results we have obtained with the different techniques considered, it contains the Error Rates (WER and SER) and the Relative Improvements (RI) of this rates.

New Feature Added	WORD		STRING	
	WER	WER-RI	SER	SER-RI
Baseline System	3.8%	---	14.9%	---
Tied State Models	3.1%	18.4%	13.9%	7.2%
Three Best Candidates	2.2%	29.0%	6.5%	53.2%
Real time CMN	2%	9%	6.2%	4.6%
Gender Models	1.6%	20%	4.8%	22.6%
Noise Spotting	1.5%	6.2%	4.7%	2.1%

Table 11: Summary of results in terms of Word Error Rate, String Error Rate and the Relative Improvements of this rates in the Connected-Number Task

6. Future Work

After this improvements, it still happens that most errors appeared among sub-sets of highly confusible words, i. e. the "-ce"set. To increase the discriminative behaviour of our models we have started some experiments with Discriminative Training Techniques. In particular we implemented the N-best based procedure proposed in [8] adapted to the codebook and stream weights in SCHMM [10]. With the early experiments a slight improvement of 7% of word error rate reduction has been obtained.

7. Conclusion

We have described several techniques to improve a spontaneously spoken Connected-Number Recognition System for Castilian Spanish over the telephone line. From the different techniques we have tested, it was found important increases in performance mainly with Tied State Modelling, Multiple Candidates, CMN, Gender modelling and Noise Spotting. The overall improvement of the system is of nearly 68% in the String Error Rate (using N-Best techniques) comparing with our baseline SCHMM system, that represents a final Word Error Rate of 1.5%. The resulted improved system is a feasible one to be portable for commercial applications.

References

[1] R. Cardin, Y. Normandin, E. Millien, "Inter-Word Coarticulation Modelling and MMIE Training for Improved Connected Digit Recognition", ICASSP-93 Mineapolis, MN, pp. 243-246, April 1993.

[2] R. Haeb-Umbach, D. Geller, H. Ney, "Improvements in Connected Digit Recognition Using Linear Discriminant Analysis and Mixture Densities", ICASSP-93 Mineapolis, MN, pp. 239-242, April 1993.

[3] R. Leonard, G. Doddington, "A Database for Speaker-Independent Digit Recognition", ICASSP-84, paper 42.11.

[4] E. R. Buhrke, R. Cardin, Y. Normandin, M. Rahim, J. Wilpon, "Application of Vector Quantized Hidden Markov Modelling to Telephone Network Based Connected Digit Recognition", ICASSP-94, Australia, pp. I-105 - I-108.

[5] P. Ramesh, J. Wilpon, M. McGee, D. Roe, C. Lee and L. Rabiner, "Speaker independent recognition of spontaneously spoken connected digits", Speech Communication 11 (1992), pp. 229-235.

[6] X. Huang, F. Alleva, H. Hon, M. Hwang, K. Lee, R. Rosenfeld, "The SPHINX-II speech recognition system: an overview", Computer Speech and Language, vol. 2, 1993, pp. 137-148.

[7] D. Tapias, A. Acero, J. Esteve, J. Torrecilla, "The VESTEL telephone speech database", ICSLP-94, Japan. Sept. 1994.

[8] J. Chen, F. Soong, "N-best Candidates-Based Discriminative Training for Speech Recognition Applications" IEEE Speech and Audio processing, vol.2, no.1, Jan. 1994, pp.206-216.

[9] F.J. Caminero, C. de la Torre, L.A. Hernández-Gómez and C. Martín, "New N-Best based Rejection Techniques for improving a Real-Time Telephonic Connected Word Recognition System". EUROSPEECH-95, Madrid, Sept. 1995.

[10] C. Martín del Álamo, F.J. Caminero, C. de la Torre and L.A. Hernández-Gómez, "Codebook Weights Adaptation for Discriminative Training of SCHMM-Based Speech Recognition Systems". EUROSPEECH-95, Madrid, Sept. 1995.