ELSEVIER

Speech Communication 29 (1999) 65–76

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Improving continuous speech recognition in Spanish by phone-class semicontinuous HMMs with pausing and multiple pronunciations

Javier Ferreiros *, José M. Pardo

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

Received 21 October 1998; received in revised form 17 February 1999; accepted 22 February 1999

Abstract

This paper presents a comprehensive study of continuous speech recognition in Spanish. It shows the use and optimisation of several well-known techniques together with the application for the first time to Spanish of language specific knowledge to these systems, i.e. the careful selection of the phone inventory, the phone-classes used, and the selection of alternative pronunciation rules. We have developed a semicontinuous phone-class dependent contextual modelling. Using four phone-classes, we have obtained recognition error rate reductions roughly equivalent to the percentage increase of the number of parameters, compared to baseline semicontinuous contextual modelling. We also show that the use of pausing in the training system and multiple pronunciations in the vocabulary help to improve recognition rates significantly. The actual pausing of the training sentences and the application of assimilation effects improve the transcription into context-dependent units. Multiple pronunciation possibilities are generated using general rules that are easily applied to any Spanish vocabulary. With all these ideas we have reduced the recognition errors of the baseline system by more than 30% in a task parallel to DARPA-RM translated into Spanish with a vocabulary of 979 words. Our database contains four speakers with 600 training sentences and 100 testing sentences each. All experiments have been carried out with a perplexity of 979, and even slightly higher in the case of multiple pronunciations, to be able to study the acoustic modelling power of the systems with no grammar constraints. © 1999 Elsevier Science B.V. All rights reserved.

Zusammenfassung

In diesem Artikel wird eine umfassende Studie über das Erkennen von kontinuierlicher Sprache im Spanischen vorgestellt. Der Gebrauch und die Optimierung mehrerer bekannter Techniken werden gezeigt, zusammen mit der für die spanische Sprache erstmaligen Anwendung spezifischer Sprachkenntnisse auf diese Systeme. Dazu gehören die sorgfältige Auswahl des Lautinventars, die benutzten Lautklassen und die Auswahl alternativer Ausspracheregeln. Wir haben ein semi-kontinuierliches, von den Lautklassen abhängendes, kontextuelles Modell entwickelt. Durch das Benutzen von vier Lautklassen haben wir Reduzierungen der Erkennungsfehlerrate erreicht, die annähernd der prozentualen Zunahme der Parameteranzahl entspricht, verglichen mit dem semi-kontinuierlichen, kontextuellen Ausgangsmodell. Wir zeigen ebenfalls, dass der Gebrauch von Pausen im Trainingssystem und Mehrfach-Aussprachen im Vokabular dazu beitragen, die Erkennungsraten erheblich zu verbessern. Das derzeitig angewandte Markieren von

* Corresponding author. E-mail: jfl@die.upm.es

PII: S 0 1 6 7 - 6 3 9 3 (9 9) 0 0 0 1 3 - 8

Pausen in den Trainingssätzen und die Berücksichtigung von Assimilationseffekten verbessern das Transkribieren in kontextabhängige Einheiten. Aussprachevarianten werden erzeugt, indem man allgemeine Regeln benutzt, die man leicht auf beliebiges spanisches Vokabular anwenden kann. Mithilfe dieser Ideen haben wir die Erkennungsfehler des Ausgangssystems um mehr als 30% reduziert, bei einer Aufgabe, die parallel zu DARPA-RM – ins Spanische übersetzt – ist, mit einem Vokabular von 979 Wörtern. Unsere Datenbasis enthält vier Sprecher mit jeweils 600 Trainingssätzen, und 100 Testsätzen. Alle Experimente wurden mit einer Perplexität von 979 durchgeführt und sogar noch mit einer ein wenig höheren im Fall der Mehrfach-Aussprachen, um die Mächtigkeit der akustischen Modellierung des Systems ohne grammatische Einschränkungen zu untersuchen.

Résumé

Dans cet article, nous présentons une étude exhaustive de la reconnaissance de la parole continue en espagnol. On montre l'utilisation et l'optimisation de plusieurs techniques bien connues et en même temps l'application des connaissances spécifiques de la langue à ses systèmes qui se fait pour la première fois à l'espagnol. Ceci veut dire: la selection soigneuse de l'inventaire des phones, les classes des phones utilisées et la selection des règles d'autres prononciations. Nous avons dévéloppé un modèle semi-continu et contextuel dependant des classes des phones. En utilisant quatre phones, nous avons obtenu des réductions du taux des erreurs de reconnaissance qui sont approximativement équivalent à l'augmentation par pourcentage du nombre des paramètres, comparé avec le modèle semi-continu et contextuel qui sert de point de comparaison. Nous montrons également que l'utilisation de la pause dans le système d'entraînement et les prononciations multiples dans le vocabulaire contribuent à améliorer le taux de reconnaissance considérablement. Les pauses actuelles des phrases d'entraînement et le considération des effets d'assimilation améliorent la transcription dans des unitées qui dependent du contexte. Des multiples possibilités de prononciation sont générées en utilisant des règles générales qui s'appliquent facilement à n'importe quel vocabulaire espagnol. À l'aide de ces idées, nous avons réduit le taux des erreurs du système de base en plus du 30% dans un travail qui est parallèle à DARPA-RM, traduit à l'espagnol avec un vocabulaire de 979 mots. Notre base de données contient quatre locuteurs avec 600 phrases d'entraînement et 100 phrases de test par locuteur. Toutes les expériences ont été réalisées avec une perplexité de 979 et même avec une perplexité un peu plus haute dans le cas des prononciations multiples pour pouvoir étudier la force du modèle en acoustique des systèmes sans contraintes grammaticales.

Keywords: Spanish continuous speech recognition; Semicontinuous HMMs phone-class dependent; Use of pausing information during training; Use of multiple pronunciations during recognition

Nomenclature

SCCN – Base-Line Semicontinuous HMMs, context dependent
SCPD – SCCN, Phone-class dependent
ScpdBas – SCPD, Basic strategy
ScpdPan – SCPD, Strategy that makes use of pausing information during training
ScpdMul – SCPD, Strategy that makes use of multiple pronunciation during recognition
ScpdPauMul – SCPD, Combined strategy

1. Introduction

Many aspects are helpful in obtaining successful continuous speech recognition systems. Of course, it is of primary importance to obtain the best possible models. The "best models" means the models with the highest recognition rates with independent data test sets and trained with the always-limited training set. Thus, we have a compromise between the acoustic definition and the trainability of the models with the available training set; a compromise that we have to evaluate through the recognition results of independent test sets. Other people have already proposed an increase in the definition of the models in the soft vector quantization stage of the semicontinuous modelling, making this stage dependent on the phone unit being processed (Hwang et al., 1994; Peinado et al., 1994). We present one way to gradually increase the acoustic definition of semicontinuous HMM models that obtains the best model from the training database available: the semicontinuous phone-class dependent modelling. In our proposal, not all the different phones have their own set of gaussians to perform the soft quantization procedure. Instead, the phones are grouped so that those that are similar share the same set of gaussians. In this sense, our proposal is closer to the one by Hwang et al. (1994). In (Peinado et al., 1994), they present results only for an isolated word task with 16 words (digits + 6 keywords) where each basic unit has its own set of gaussians. We have studied this idea with limited training data (600 sentences for each speaker) and found an optimum of four phone-classes for our continuous speech, a 1000-word vocabulary, recognition task in Spanish (Ferreiros and Pardo, 1995). We are presenting the methodology used to gradually increase the number of parameters of the model, using different levels of grouping the phones to obtain the number of classes that improve the recognition rates significantly without increasing the number of parameters too much.

Other aspects are related to the information required for the training process such as the transcription of the training phrases, specifically the construction of the unit's inventory. This aspect is not always well considered, for example in (Huerta et al., 1998) we can find the application of SPHINX III to Spanish with a poor basic unit's inventory that does not allow to describe relevant effects of Spanish. In this paper, we are also presenting experimental results that show that the use of pausing in the training system improves the transcription into basic units with significant consequences when we use context-dependent units. For context dependent units, we apply assimilation effects when no pause occurs to select the best triphone sequence. The implementation of the recogniser needs a special way to connect words through the different contexts considering the possibility of a pause between words. We propose a variant of the recognition procedure close to the one proposed by Hwang et al. (1989) for context-dependent recognisers.

Finally, we studied the effect of allowing multiple pronunciations of the words in the vocabulary used by the recognition algorithm as another significant aspect to be considered in our system to increase recognition accuracy. Many people have proposed similar ideas (Weintraub et al., 1989). For Spanish, we find in (Huerta et al., 1998) also the necessity to obtain multiple pronunciations. They obtained them through automatic learning from the text transcription and we believe that, due to their poor phone inventory, they were training more an error model than a really useful set of pronunciation alternatives. The particularity of our proposal is that we pursue pronunciation variations even inside the same dialect. These multiple possibilities stand for the different forms of pronunciation found in Spanish speakers. We have obtained simple transcription modification rules useful for any Spanish recogniser.

When we tried all techniques at the same time we obtained significant error reductions that prove the importance of these aspects when building a Spanish continuous speech recogniser.

The paper is organised as follows. Section 2 introduces the semicontinuous phone-class dependent HMM modelling. In Section 3, we study the benefits obtained after the use of pausing and assimilation processes on the training material. Section 4 presents the technique used to define multiple pronunciations for the words in the recognition vocabulary. The experimental work and the results obtained when we combined both the pausing and multiple pronunciations strategies are shown in Section 5. In Section 6 we discuss the characteristics of all strategies with the consideration of the 95% confidence recognition bands. Finally, in Section 7 we review the conclusions of this work.

2. The semicontinuous phone-class dependent (SCPD) HMM model

2.1. The model

In a semicontinuous HMM model the acoustic information is modelled with a mixture of shared gaussian probability density functions stored in what we call codebooks for simplicity. The weights for the different states indicate the use of these gaussians. The probability density function for the observation vector x given the state s_t can be written as (Huang and Jack, 1989; Huang et al., 1989)

$$f(x|s_t) = \sum_{j=1}^{L} f\left(x|O_j, s_t\right) \Pr\left(O_j|s_t\right), \tag{1}$$

where L stands for the number of gaussian codewords in the codebook. Being a semicontinuous model, the emission probability density function conditioned on the codewords, $f\left(x|O_j, s_t\right)$, can be assumed to be independent of the Markov states s_t and Eq. (1) be rewritten as

$$\begin{aligned} f(x|s_t) &= \sum_{j=1}^{L} f\left(x|O_j\right) \Pr\left(O_j|s_t\right) \\ &= \sum_{j=1}^{L} f\left(x|O_j\right) b_{s_t}\left(O_j\right), \end{aligned} \tag{2}$$

where the probability for each state $\Pr\left(O_j|s_t\right)$ of using each codeword O_j is rewriten as $b_{s_t}\left(O_j\right)$, paralleling the B matrix of a discrete HMM modelling.

To obtain a phone-class dependent semicontinuous modelling, we generate sets of gaussian functions, each of them to be used by different classes of phones as proposed in (Hwang et al., 1994).

$$f(x|s_t) = \sum_{j=1}^{L} f\left(x|O_{\text{class}(s_t),j}\right) \cdot b_{s_t}\left(O_{\text{class}(s_t),j}\right). \tag{3}$$

There are many more weights $b_{s_t}\left(O_{\text{class}(s_t),j}\right)$ than parameters in the codewords $O_{\text{class}(s_t),j}$, allowing the definition of multiple classes as a refinement of the model without increasing the total number of parameters too much. The function class(s_t) specifies the class to which the phoneme that has the state s_t belongs. In CMU they found an optimum of 27 phone-classes for SPHINX-II on the 20,000-word open-vocabulary continuous speech speaker-independent WSJ task (Hwang et al., 1994). We have found an optimum of 4 phone-classes for our 1000-word continuous speech speaker-dependent task presented in Section 2.2.1 that has small training sets (600 sentences) for each speaker.

For our context-dependent generalised triphones modelling, where we have units like [**a**, **b**, **c**, **d**, **e**, **f**, **g**] meaning an **e** sound with left context **a**, **b**, **c** or **d** sound and right context **f** or **g** sound, we look at the central symbol **e** to decide which set of gaussians these units will use, taking into account which is the class this phone symbol belongs to.

Using context-independent discrete p.d.f. phone HMMs, we performed an automatic clustering of them to form the phone-classes for the phone-class dependent semicontinuous modelling. We have used the well-known clustering algorithm guided by entropy variation (Lee, 1988). We have kept the different phone-classes at each new step of the algorithm. From the initial 56 basic units, we created clusterings for every level, from 55 classes all the way down to 2 classes for each speaker. Then, we tried to use these different clusterings to generate the phone-class dependent gaussians and found an optimum with four classes to obtain significant improvements in the recognition rates.

Once we realised that four classes were enough for our task, we decided to combine the automatically obtained speaker-dependent classes for 4 speakers and extracted the common observed rules between speakers. In Table 1 you can see the set of four phone-classes we have finally used in the experiments shown in the paper for all speakers. Table 2 explains the meaning of the labels used for this set of phones.

These classes have been manually designed to follow the rules extracted from the speaker-dependent automatic clusters: Class 0 contains fricative-like sounds and bursts; Class 1 contains all nasal sounds plus the closure of the b and g sounds which certainly have a nasal-like sound before the burst in Spanish; Class 2 is made up of closures and silence models; Class 3 contains the rest, that is, the voiced sounds.

We have not found significative differences in the recognition performance when using automatic or manually derived phone-classes. The figures we will present in the experimental work section

Table 1
The four classes of phones

Class 0	p t k s f X T #T/ T/
Class 1	m n N a∼ e∼ i∼ o∼ u∼ ′a∼ ′e∼ ′i∼ ′o∼ ′u∼ #N∼ N∼ #b #g
Class 2	#p #t #k #d ⟨ ⟩
Class 3	a e i o u ′a ′e ′i ′o ′u j w l b d g B D G r R J L &

Table 2
Help for phone labels interpretation

- Plosive sounds (p,t,k,b,d,g) have two component models as in:
  Nápoles n 'a #**p** **p** o l e s
  #**p** is the closure model
  p is the burst model
- **X** is the Spanish linguovelar fricative as in:
  origen o r 'i **X** e n
- **T** is the interdental fricative as in:
  príncipe #p p r 'i n **T** i #p p e
- #**T/** **T/** are the two models for the linguopalatal africate as in:
  ochocientas o #**T/** **T/** o T j 'e n #t t a s
- **N** is the nasal sound **n** before a velar sound as in:
  Inglaterra i~ **N** #g g l a #t t 'e R a
- ~ is a nasalization mark
- ' is a stress mark
- #**N**~ **N**~ are the models for the Spanish linguopalatal nasal as in:
  Logroño l o G r 'o #**N**~ **N**~ o
- ⟨ and ⟩ are initial and final silence models
- **j** and **w** are the semi-vowel sounds as in:
  opciones o #p p T **j** 'o n e s
  prueba #p p r **w** 'e B a
- **B**, **D** and **G** are the fricative versions of **b**, **d** and **g** as in:
  reserva R e s 'e r **B** a
- **R** is the multiple vibrant Spanish sound as in:
  Navarra n a B 'a **R** a
- **J** is the voiced fricative as in:
  Pompeya #p p o m #p p 'e **J** a
- **L** is the Spanish lateral linguopalatal as in:
  rejilla R e X 'i **L** a
- **&** is an inter-word unit

correspond to the ones obtained with the manually derived ones. We think that these four classes of phones presented in this work have a very good potential to be used for other Spanish tasks. Anyway, if this is not the case for a particular application, the designer can always follow our automatic methodology to obtain a proper SCPD modelling with a different number of classes.

2.2. Experimental work with SCPDs

2.2.1. The Spanish continuous speech recognition task

We have designed a continuous speech recognition task in Spanish paralleling the DARPA-RM task (Price et al., 1988). We translated the 1000 DARPA-RM sentences into Spanish. English proper names were changed to Spanish names. The number of words in the vocabulary is close to DARPA-RM; we have 979 different words, although the average number of words per sentence is slightly higher in the Spanish version because more words are needed in Spanish to represent the same idea than in English. We have recorded 2 male and 2 female speakers in a quiet room at a 16 KHz sampling rate.

We have used two different sets of sentences: the training set with 600 sentences and the test set with 100 sentences for each speaker. We have not used any grammatical constraint to be able to analyse the pure acoustic power of our models. The recogniser is a one-pass algorithm (Bridle et al., 1982; Ney, 1984) with the special inter-word connection process, which we will introduce in Section 3, when we use triphones obtained after pausing.

The task uses an inventory of 56 basic context-independent allophones (Tables 1 and 2). It is a large inventory because plosive sounds use two allophones: the closure and the burst (for example #**p** **p**, where #**p** is the closure and **p** the burst). Vowels have also four different allophones: the basic one, the stressed, the nasal and the stressed nasal. The resolution in the basic allophone set allowed us to design very compact context-independent speech recognisers with enough accuracy for some applications (Hasan et al., 1989; Pardo and Hasan, 1989; Ferreiros, 1996). This inventory improves the acoustic definition of the system described in (Huerta et al., 1998) where we found that they do not have any of the fricatives **B**, **D** or **G**, any of the stressed and/or nasalised vowels, the semivowelsl **j** and **w** and even the sound **T** is assimilated as an **s**, while this is only nearly true for the Spanish spoken in Andalucía or Extremadura and Ibero-American Spanish realisations. We think that the right modelling of all basic allophones in Spanish is important because it allows us to better describe some effects (assimilation, proper b,d,g transcription, multiple pronunciation rules, distinction between verb tenses by different stressed vowels,...) from the acoustic level as will be discussed through the paper.

In this paper, we present results with a set of 347 generalised triphones obtained using automatic clustering (Lee, 1988). We keep context-independent models trained in parallel for those cases where the triphone needed in recognition has not been observed in the training material.

The front-end processing obtains an 11-dimensional vector (10 MFCC + Energy) every 10 ms frame and also other two vectors with the first and second derivative of the basic vector. For every set of vectors we have created a different codebook of gaussians. We have used both 2-codebook and 3-codebook models throughout this work. Thus, each gaussian probability function set has 11 components in each codebook. We have used 256 11-dimensional gaussians for each of the three codebooks.

2.2.2. The experimental results with SCPDs

First, we will show the comparison between the base-line and the phone-class dependent semicontinuous models in terms of number of parameters. If we count the number of parameters implied in the baseline context-dependent semicontinuous models we have to make room for 403 models (347 generalised triphones + 56 context independent models to be used when the triphone needed in recognition did not appear in the training).

We have used three-state HMM models with three transitions per state. For both types of modelling, the transition matrix needs only 3627 parameters, a small percentage of the total number of parameters.

The number of parameters of the base-line semicontinuous system is:

- The gaussians = 256 gaussians · 11 dimensions/gaussian · 2 parameters (mean and variance) · **CB** codebooks.
- The weights of these gaussians = 403 models · 3 states/model · 256 values/state · **CB** codebooks.

The number of parameters of the gaussian probability density functions is only the 1.8% of the total amount. This fact triggers the idea of increasing the resolution of the model just in this part, shared by all models, as this will create only a small increase in the total number of free parameters.

Thus, we have 633899 parameters for **CB** = 2 codebooks and 949035 for **CB** = 3 codebooks.

For the phone-class dependent modelling, all this is the same except that we have 4 sets of gaussians. This means 667691 parameters for 2-codebook models and 999723 for 3-codebook ones, that is and increase of 5.3% for both cases.

Table 3
Average word error rate (WER) and number of paramaters (PARS) comparison between base-line and phone-class dependent semicontinuous models

	sccn		scpd	
	WER	PARS	WER	PARS
2 CB	25.6	633899	24.0 (−6.3%)	667691 (+5.3%)
3 CB	21.1	949035	20.0 (−5.2%)	999723 (+5.3%)

Table 3 shows the average error rate in our 1000-perplexity task for these two systems over all speakers. **Sccn** stands for the base-line and **scpd** the corresponding phone-class dependent using **CB** codebooks.

As can be seen, there is a 6.3% decrease in the error rates for the 2-codebook models and 5.2% for 3-codebook ones, that roughly equals in percentage the increment of the number of parameters.

3. Pausing the training sentences

3.1. The use of pausing in the training process

To obtain the transcription of the sentences into basic units for the training process, we used in the base-line a special unit between words, the inter-word unit **&**. This unit is useful when a real pause exists but in other cases it produces a waste of training material (Ferreiros et al., 1995). In this section, we introduce a pausing procedure that eliminates the inter-word unit from the transcription most of the times and uses it only when a real pause exists. To perform this pausing we generate an initial transcription based on a network where a transition skipping the inter-word unit is considered. Then, a Viterbi algorithm decides which inter-word units remain or which of them are deleted from the transcription. When the unit has to be deleted, we try to assimilate the last phone of the previous word to the first one of the following and then a second and definitive Viterbi alignment is performed with the improved transcription for model re-estimation. The triphones inventory is also extracted from the refined transcription obtained from the final training iteration that is kept as the best transcription of the training sentences.

This pausing procedure applied in the training process produces three main benefits compared to the old strategy. First, we obtain better-trained models because those frames that were lost when mistakenly chosen to train the inter-word unit, are used now to train the proper models.

Second, the modelling and subsequently the meaning of the inter-word unit have changed and we can now speak of a real pause unit. The old unit caused recognition errors substituting other short words, because its rather flat model accepted almost anything. This flatness was due to the different sounds between words that were pooled to train the inter-word unit in the old system. Now the unit is trained only with observations automatically labelled as silence by the pausing procedure and serves in the recognition process as a model for the actual pauses in the recognition sentences not substituting so easily other words.

Third, when we work with context-dependent models, the units sequence is different if we set a pause between two words or not. Besides, we obtain different context-dependent unit sequences if we can assimilate or not the last phone of the previous word with the first one of the following word. Then, the inventory of basic units may change according to the observed pauses.

We need a special process in the one-pass recognition algorithm to properly connect words through this new inventory of units. This modified recognition algorithm is based on similar ideas to those proposed in (Hwang et al., 1989). Instead of having only one probability of exiting the last state of the model of a word, we have to consider theoretically 56 different probabilities depending on the first phone of the following word because the triphones of a word will be:

$$w_i\colon [\mathbf{X}p_{i1}p_{i2}][p_{i1}p_{i2}p_{i3}]\cdots[p_{i(m-1)}p_{im}\mathbf{Y}]$$

where **X** and **Y** depend on the contexts needed by the surrounding words to w_{i}. Our new algorithm has to recover a word sequence that holds the following right connections:

$$\text{sentence: } [< p_{11}p_{12}][p_{11}p_{12}p_{13}]\cdots[p_{1(m1-1)}p_{1(m1)}\mathbf{Y}_1]$$
$$[(\mathbf{X}_2 = \mathbf{Y}_1)p_{21}p_{22}][p_{21}p_{22}p_{23}]\cdots[p_{2(m2-1)}p_{2(m2)}\mathbf{Y}_2]\ldots$$
$$[(\mathbf{X}_n = \mathbf{Y}_{(n-1))}p_{n1}p_{n2}][p_{n1}p_{n2}p_{n3}]\cdots[p_{n(mn-1)}p_{n(mn)}\mathbf{Y}_n],$$

where $\mathbf{Y}_n$ can be any phone, but usually our > ending silence model is chosen.

Our algorithm starts by assuming that all **X** contexts for all the words is a model < that we have for the beginning silence. Then for all the words, Viterbi decoding calculations are performed following the internal triphones of each word. When arriving at the last state, P different exiting probabilities are calculated hypothesising the P different **Y** contexts, P being the number of basic phones. Actually only $Q < P$ different probabilities need to be calculated because some of the $[p_{i(m-1)}p_{im}\mathbf{Y}]$ triphones are clustered in the same generalised triphone. In our experiments $P = 56$ and the average $Q = 6$ for 347 generalised triphones.

In theory, we should select from the previous frame the 56 best ending words depending on the 56 different possible connections for the 56 different **X** contexts, [1] but we made a reduction using only the best ending word for all connections. With this reduction, we have not found significant differences in the recognition accuracy and, besides, the algorithm takes up less CPU time.

The assimilation effect is not directly considered in this recognition algorithm, although there is one characteristic of our implementation that helps in those cases where an assimilation is needed. This characteristic is that we have always three transitions from each HMM, the auto-transition, the transition to the next state and the transition to the following, i.e. skipping the next state. This means also that we have two output transitions from each word. Thus, when the following word begins with the same phone than the previous word ending phone, the recogniser may choose the one state skipping transitions to hurry up trough the two identical concatenated models. This is an effect

[1] We are speaking about the no-grammar case that is the one we have tested in the experiments, but this is also true at each grammatical node if a grammar is used.

Table 4
Average recognition error rates for the strategy of pausing in the training

	Basic strategy	Pausing
2 CB	24.0	19.6 (−18.3%)
3 CB	20.0	18.2 (−9.0%)

similar to the assimilation of both units in the recognition process.

3.2. Experimental results

We now compare the recognition accuracy obtained using semicontinuous phone-class dependent modelling with traditional training and with the modified training procedure that includes pausing information.

In Table 4 we see stronger error reductions for the 2-codebook case. Our hypothesis is that these weaker models take more advantage of this improved training procedure.

4. Multiple pronunciations in the recognition vocabulary

4.1. The idea

Our recognition algorithm cannot assume that all Spanish speakers will pronounce a word in the same way. Some words have many pronunciation variants, and we have to deal with this fact. This effect is observed even inside the same dialect community. Using this language specific knowledge, we have defined different transcriptions for the words in the recognition vocabulary and, although the perplexity increases, we have observed significant improvements in the recognition rates. Similar ideas to those proposed here were presented in (Weintraub et al., 1989) for English systems. For Spanish we can also see this idea in (Huerta et al., 1998), but with automatic learned alternatives and with a poor phone inventory that produces more an error model than pronunciations alternatives because they do not have enough definition to describe some Spanish multiple pronunciations effects.

Our implementation is based on the use of a set of rules that describe how a Spanish speaker may utter a particular sequence of phones (Ferreiros et al., 1998). We have in Spanish the good luck of having a very linear transcription from the graphemes to the phone sequences that eases the rule-based definition of such different forms in a very general way. We have created a reduced set of all the rules by selecting the ones which occur most frequently. The aim of this approach is to obtain significant recognition improvements without increasing the acoustic perplexity of the vocabulary too much.

Another problem addressed with this technique is that, in Spanish, there are cases where the starting allophone of a word in continuous speech depends on the last allophone of the previous word. These cases occur when the word begins with the grapheme b, d or g. If the previous word ends with a nasal sound or a pause, the correct transcription of the grapheme is a plosive allophone. For the rest of the cases, the correct transcription of the grapheme is a fricative allophone.

With our multiple pronunciation strategy we have a natural way to solve this problem allowing both pronunciations and giving the recogniser the option of choosing the best match.We have used the set of rules in Table 5 that were run on each word of the vocabulary to generate the corresponding set of different allowed pronunciations. Of course, it may be the case where several rules apply for the same word.

For the experiments we present, the initial vocabulary of 979 words was expanded to 1211 entries (an increment of 23.7%) when the rules of multiple pronunciations are applied. Although the perplexity is increased by the same amount because we do not use any grammar constraint in this work, we obtain significant better recognition results.

4.2. Experimental work

Now we compare the baseline semicontinuous phone-class dependent modelling using only one or multiple pronunciations in the recognition vocabulary. We see significant error reductions both for 2-codebook and 3-codebook systems.

Table 5
List of multiple pronunciation rules applied in this work

RULE	EXAMPLE: Grapheme	EXAMPLE: Phone sequences
Initial ‘b’, ‘d’ and ‘g’ both as plosives and fricatives	Badajoz	#b b a D a X ′o Z B a D X ′o Z
‘x’ sound as ‘g s’, ‘k s’ and even only ‘s’	explorador	e #g g s #p p l o r a D ′o r e #k k s #p p l o r a D ′o r e s #p p l o r a D ′o r
‘#k k #t t’ sound as ‘T #t t’	activa	a #k k #t t ′i B a a T #t t ′i B a
‘#k k T’ sound as only ‘T’	proyección	#p p r o J e #k k T j ′o n #p p r o J e T j ′o n
‘#p p #t t’ as ‘B #t t’	concepto	#k k o n T ′e #p p #t t o #k k o n T ′e B #t t o
‘h’, without sound in Spanish, as a ‘G’	Huelva	w ′e l B a G w ′e l B a
‘D’ endings also as ‘T’ endings	capacidad	#k k a #p p a T i D ′a D #k k a #p p a T i D ′a T

Table 6
Recognition results for the strategy of multiple pronunciations

	Basic strategy	Multiple pronunciations
2 CB	24.0	21.1 (−12.1%)
3 CB	20.0	16.1 (−19.5%)

Table 6 shows that it is the 3-codebook system that takes better advantage of having multiple pronunciations in the recognition dictionary. We think that it corresponds to the higher modelling power of the 3-codebook models that are capable of better discrimination between the different pronunciation alternatives, without being so affected by the perplexity increase.

5. Combining the strategies

We have tested the strategies of pausing in the training process and multiple pronunciations at the same time with the phone-class dependent model and found the results in Table 7 which completes the preliminary work presented in (Ferreiros and Pardo, 1995).

Table 7
Recongition results for the combined strategy of pausing in the training and multiple pronunciations

	Basic strategy	Pausing + Multiple pronunciations
2 CB	24.0	15.4 (−35.8%)
3 CB	20.0	14.4 (−28.0%)

These results show that not only each of the strategies produce significant improvements, but the mixture of both strategies collaborates to produce a significant reduction in errors (Ferreiros, 1996).

In Fig. 1 we show the error rate evolution we have observed through the different systems presented in the paper.

Comparing the results for the phone-class dependent modelling using pausing in the training or the multiple pronunciations strategy, we have observed larger error reductions with multiple pronunciation for the 3-codebook models and larger error reductions with the pausing strategy for the 2-codebook models. Our explanation of this effect is that 3-codebook models are good enough as acoustic models and the improvement of the

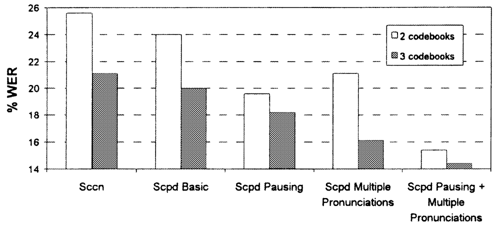

Fig. 1. Error rate evolution through the experiments.

Table 8
Global error rate reduction through the ideas presented in the article

	sccn	scpdPanMul
2 CB	25.6	15.4 (−39.8%)
3 CB	21.1	14.4 (−31.8%)

training process is less important than using the correct variant pronunciation of the word. Instead, 2-codebook models also need the help of an improved training procedure.

Comparing the error rates of the phone-class dependent system trained with pausing and recognised with multiple pronunciations to those of the base-line semicontinuous system (Table 8), we observe a reduction of 39.8% in the errors for the 2-codebook system and 31.8% in the errors in the 3-codebook case. Throughout this work, the error reduction in general is higher for 2-codebook systems because 3-codebook systems have so many parameters that suffer from a lack of training data (Ferreiros, 1996).

6. Discussion

We have calculated the bands where, with a confidence of 95%, the rates we have obtained in this work may actually be. We have used the following formula for the confidence interval for a population proportion p (Weiss and Hasset, 1993, pp. 407, 408):

$$\frac{\text{band}}{2} = z_{\alpha/2}\sqrt{\frac{p(100-p)}{n}}, \qquad (4)$$

where p is the percentage sample proportion (number of successes divided by sample size × 100), n the sample size and $z_{\alpha/2} = 1.96$ for a 95% confidence interval. For this formula we have used $n = 4280$ words because each of the four speakers has 1070 words in the 100 reference sentences and all the figures presented here were obtained by calculating the average for the recognition rates for the four speakers. Thus, any recognition or error rate is translated into the band [p − band/2, p + band/2] with a confidence of 95%.

With these bands, we can analyse the significance of the results obtained and discuss the effects detected. In Fig. 2 we can see these bands for the results obtained with 2-codebook systems for all strategies.

The first thing we can see is that there is an overlap between the traditional semicontinuous modelling and the phone-class dependent one, although the limits of each modelling do not include the average rate of the other modelling. We do not observe any overlap between the basic strategies and the Pausing or Multiple pronunciation strategies, while the Pausing and Multiple pronunciation strategies do overlap themselves. The conclusion is that a system may be improved by these two ways: improving the modelling through the pausing strategy or improving the vocabulary with alternate pronunciations.

What happens with the combination of both strategies? It may be the case that both of them are solving the same kind of errors, so that the combination will not significantly improve the system. Or it may be that each strategy is solving a different kind of error and then the combination would the advisable. Both the intuition and the experimental results aim at this second possibility. It seems that

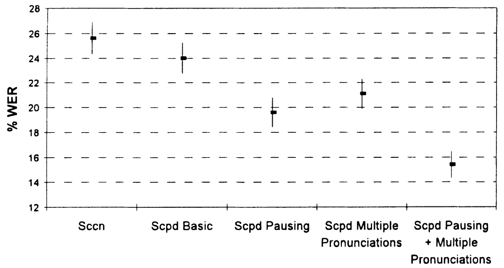

Fig. 2. Word error rate bands for 2- codebook experiments.

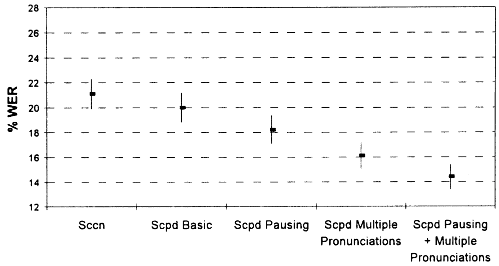

Fig. 3. Error rate bands for 3-codebook experiments.

the pausing strategy is improving the acoustic definition of the models while the multiple pronunciations strategy is adding a new capability of recognising alternate pronunciations of the words.

In fact, the combined strategy does not overlap with any other system and is providing the best error reduction while taking advantage of both strategies.

Let us have a look now at the 3-codebook experiments (Fig. 3). The conclusions are similar to those for 2 codebooks, although we can see now the following particularities.

The pausing strategy overlaps a little with the basic strategy. We think that this means that these 3-codebook models that have intrinsically more modelling power are not improved so much by the pausing strategy. We can see that the multiple pronunciation strategy is also helping these systems, because the improvement is not in the modelling itself but rather in the system's ability to cognize alternate pronunciations.

There is also a similar reason for the slight overlap of the combined strategy with that of multiple pronunciations. The difference between the two strategies is the pausing improvement of the training that does not help the 3-codebook systems to the same extent as the 2-codebook systems.

7. Conclusions

We have shown how language specific knowledge is applied to the careful selection of the phone inventory, the creation of phone-classes, and the selection of alternative pronunciation rules. The rich unit's inventory allows us to better describe some effects of Spanish (assimilation, proper b,d,g transcription, multiple pronunciation rules, distinction between verb tenses by different stressed vowels, ...) from the acoustic level that produces improvements in our recognition systems. We think that the data presented could be of interest to other researchers when facing the design of a Spanish recogniser.

From a baseline semicontinuous HMM system we have evolved to a phone-class dependent semicontinuous system, using automatic pausing and assimilation in training and with a vocabulary with multiple pronunciations that reduces the original recognition errors by more than 30%.

We use four speaker-independent classes sufficiently general in their definition to be used to improve a semicontinuous modelling without increasing the number of parameters of the system too much.

The use of pausing and assimilation in the training process allows us to obtain better models and at the same time to define more realistic triphone units. We have developed a special process in the one-pass recognition algorithm to properly concatenate words using these units.

The multiple pronunciation strategy in the recogniser vocabulary improves recognition accuracy by using some very general rules of the way the Spanish speakers talk.

Acknowledgements

This work has been realised inside the framework of the National Project CICYT TIC 94-0119. The authors also wish to thank in general to all GTH Speech Technology Group at Madrid University of Technology for their contribution to this work.

References

Bridle, J.S., Brown, M.D., Chamberlain, R.M., 1982. An algorithm for connected word recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 3–5 May 1982, Paris, France, Vol. 2, pp. 899–902.

Ferreiros, J., 1996. Contribution to markov models training methods for continuous speech recognition. Ph. D. Thesis, Universidad Politécnica de Madrid.

Ferreiros, J., Pardo, J.M., 1995. Preliminary experimentation of different methods for continuous speech recognition in Spanish. In: Proceedings of Eurospeech, 18–21 September 1995, Madrid, Spain, Vol. 2, pp. 1507–1510.

Ferreiros, J., de Córdoba, R., Savoji, M.H., Pardo, J.M., 1995. Continuous speech Hmm training system: applications to speech recognition and phonetic label alignment. In: Rubio, A.J., López, J.M. (Eds.), Speech Recognition and Coding: New Advances and Trends. Springer, Berlin, pp. 68–71.

Ferreiros, J., Macías-Guarasa, J., Pardo, J.M., Villarrubia, L., 1998. Introducing multiple pronunciations in Spanish speech recognition systems. In: Proceedings of ESCA Tutorial and Research Workshop Modeling Pronunciation Variation for Automatic Speech Recognition, 4–6 May 1998, Rolduc, The Netherlands, pp. 29–34.

Hasan, H., Pardo, J.M., Alexandres, S., Casado, C., 1989. Phonetic properties of a large Spanish lexicon and its implications for large vocabulary speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 23–26 May 1989, Glasgow, Scotland, Vol. 1, pp. 342–344.

Huang, X.D., Jack, M.A., 1989. Unified techniques for vector quantisation and hidden Markov modeling using semi-continuous models. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 23–26 May 1989, Glasgow, Scotland, Vol. 1, pp. 639–642.

Huang, X.D., Hon, H.W., Lee, K.F., 1989. Large-vocabulary speaker independent continuous speech recognition with semi-continuous hidden Markov models. In: Proceedings of Eurospeech, September 1989, Paris, France, Vol. 1, pp. 163–166.

Huerta, J.M., Thayer, E., Ravishankar, M., Stern, R.M., 1998. The development of the 1997 CMU Spanish broadcast news transcription system. In: Proceedings of DARPA BN Transcription Understanding Workshop, 8–11 February 1998, Lansdowne, Virginia, USA.

Hwang, M.Y., Hon, H.W., Lee, K.F., 1989. Modeling between-word coarticulation in continuous speech recognition. In: Proceedings of Eurospeech, September 1989, Paris, France, Vol. 1, pp. 5–8.

Hwang, M., Rosenfeld, R., Thayer, E., Mosur, R., Chase, L., Weide, R., Huang, X., Alleva, F., 1994. Improving speech recognition performance via phone-dependent Vq codebooks and adaptive language models in Sphinx-II. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 19–22 April 1994, Adelaide, South Australia, Vol. 1, pp. 549–552.

Lee, K.F., 1988. Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system. Ph. D. Thesis, CMU.

Ney, H., 1984. The use of a one-stage dynamic programming algorithm for connected word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-32 (2), 263–271.

Pardo, J.M., Hasan, H., 1989. Large vocabulary speaker independent isolated word speech recognition using hidden Markov models: Status report and planned research. In: Proceedings of Eurospeech, September 1989, Paris, France, Vol. 2, pp. 146–149.

Peinado, A.M., Segura, J.C., Rubio, A.J., Benitez, M.C., 1994. Using multiple vector quantization and semicontinuous hidden Markov models for speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 19–22 April 1994, Adelaide, South Australia, Vol. 1, pp. 61–64.

Price, P., Fisher, W.M., Bernstein, J., Pallet, D.S., 1988. The Darpa 1000 word resource management database for continuous speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 11–14 April 1988, New York, USA, Vol. 1, pp. 651–654.

Weintraub, M., Murveit, H., Cohen, M., Price, P., Bernstein, J., Baldwin, G., Bell, D., 1989. Linguistic constraints in hidden Markov model based speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 23–26 May 1989, Glasgow, Scotland, Vol. 2, pp. 699–702.

Weiss, N.A., Hasset, M.J., 1993. Introductory Statistics, 3rd ed., Addison-Wesley, Reading, MA, pp. 407–408.