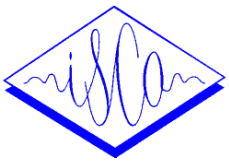

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

PRELIMINARY EXPERIMENTATION OF DIFFERENT METHODS FOR CONTINUOUS SPEECH RECOGNITION IN SPANISH

Javier Ferreiros, J.M. Pardo

Grupo de Tecnología del Habla
Departamento de Ingeniería Electrónica
Escuela Técnica Superior de Ingenieros de Telecomunicación
Universidad Politécnica de Madrid, Spain.
jfl@die.upm.es, pardo@die.upm.es

ABSTRACT

In this paper we make first a comparative study of the recognition performance of different HMM training algorithms in Spanish: Discrete context-independent phone models, Discrete context-dependent (aglomerative-clustered generalized triphone) models and Semicontinuous context-independent and context-dependent models.

We also propose two alternatives to improve the performance of the systems, the first one by using phone-class dependent modelling. The second one by preprocessing the training sentences in order to separate interword pauses and consequently train better contextual models.

Preliminary experiments on a speaker dependent database, 1000 words vocabulary show good improvement of both systems compared to the baseline system.

DATABASE

The experiments have been carried out on a continuous speech Spanish database. This database is an approximate translation of DARPA RM task into Spanish with a vocabulary of about 1000 words. We do not use any grammatical restriction in the experiments shown in this paper, so the perplexity is about 1000. We have recordings of 4 Spanish speakers. Each speaker recorded 1000 sentences. 700 of them have been used for the experiments on this paper and 300 have been kept for further experimentation. Out of these 700 sentences, we use 600 for training and 100 for recognition.

SPEECH ANALYSIS

The parametrization process produces 10 MFCC parameters and the log-Energy of every 10 msec. frame. It also calculates the first and second derivatives of each of these 11 parameters. Each set is vector-quantized with 256 codewords, so we can run experiments with one codebook that contains the information of the basic parameters or we can choose two codebooks including the first derivatives or three codebooks including also the second derivatives. In semicontinuous systems, only the parameter vectors are supplied and no VQ is needed. We use 256 gaussian pdfs for each parameter set.

BASIC MODELLING COMPARISON

Basic modelling experimentation was designed to observe the gain of recognition accuracy changing only the type of phone models, but with a strong simplification of the interword connections.

For basic and semicontinuous phone-class dependent (see next section) modelling experiments, we consider that the transcription of a sentence consists of a beginning silence followed by the words separated by an interword unit between them and an ending silence. Therefore, the transcription of a sentence is:

sentence: $< w_1 \,\&\, w_2 \,\&\, \ldots \,\&\, w_n >$

where: $<$ is the initial silence unit
w_i are the word models
& are the interword unit
$>$ is the final silence unit

Each word w_i is transcribed as the concatenation of its phone units:

$w_i : p_{i1}\ p_{i2} \ldots p_{im}$

These phone units are the basis of the acoustic modelling. We use Markov models of three states with three forward transitions from each state for each of these units.

The use of the interword unit simplifies the definition of the interword triphone units in contextual modelling. All of them will be defined by the last phone of the previous word followed by the interword unit and the first phone of the next word. For example, for a sentence piece like: "... w_i & w_j ...", the unit that appears between the two words defines the triphone "[p_{im} & p_{j1}]" meaning a "&"

in the context defined by a "p_{im}" left phone and a "p_{j1}" right phone.

We have obviously detected that most of the times, the interword unit is assigned only one frame (the minimum in our transition scheme) by the Viterbi alignment used in the training procedure, because most of the times there is no pause i.e. no need to use a special unit between two words. This fact led us to the study of the incorporation of pausing information into the training process that will be discussed in the last section of this paper.

The recognition is performed using a one-pass algorithm that concatenates the words in the vocabulary. In this vocabulary each word is transcribed as a concatenation of basic phones. When generalized triphones are used, the transcription of a word is supposed to have interword units before and after the word, because this is the way in which we have most often seen the word in the training material. For example, to transcribe the word "w_i" we will use the phone sequence for the recognition vocabulary:

w_i : & p_{i1} p_{i2} ... p_{im} &

Thus, the triphones for this word will be:

[& p_{i1} p_{i2}] [p_{i1} p_{i2} p_{i3}] ... [$p_{i(m-1)}$ p_{im} &]

We have carried out recognition experiments with discrete and semicontinuous context independent models for 56 Spanish units. These units compile the different phones where the vowels have different models for normal vowels, nasalized vowels, stressed vowels and both nasalized and stressed vowels. Plosive sounds are separated into the silence model plus the burst model and the Spanish sound "ñ" is also split in two segments.

Then, we extracted all triphones (2424) in our training database and trained them. With an entropy difference distance measure, we automatically clustered them to reduce the number of units to 350 generalized triphones. This set of generalized triphones is used to train context-dependent units (always keeping context-independent models trained in parallel, to be used if a triphone occurs in the recognition vocabulary but was not observed in the training material), both in discrete and semicontinuous versions.

With these basic methods we obtain word accuracy recognition results of discrete context-independent models (average 56.3% for 2 codebooks (2cb) / 59.6% for 3 codebooks (3cb)), discrete 350 generalized triphone models (av. 66.0% (2cb)), semicontinuous context-independent models (av. 63.3 (2cb) / 66.6% (3cb)) and semicontinous 350 generalized triphones (av. 73.0% (2cb) / 77.6% (3cb)).

These results mean that by only changing the modelling from discrete to semicontinuous we obtain a 16% (2cb) and 17.3% (3cb) reduction of the recognition errors. Also the comparison between context independent and context dependent modelling shows 22.2% error reduction for discrete (2cb) and for semicontinuous modelling reductions of 26.4% (2cb) and 32.9% (3cb). The incorporation of the second order derivatives is reducing on its own 11.2% of the errors on average. If we consider the number of parameters of the system, we realize that the change between discrete and semicontinuous modelling is not so costly as the change between a context-independent model and a contextual one as shown in the following table:

SYSTEM	2 CB	3 CB
discrete context-independent	86.520	129.528
semicontinuous context-independent	97.784	146.424
semicontinuous contextual	633.899	949.035

***Table 1** - Number of parameters of the systems.*

You can have a look at *table 3* with a summary of all results described in the paper.

PHONE-CLASS DEPENDENT MODELLING

In this case the idea is to have more than one set of basic gaussians to be shared by different models as has been already presented in [1] and [2]. Our approach is to cluster the basic phones into four classes. All the triphones whose central phone belongs to the same class will share the same set of gaussians. The number four has been found to be an optimum for the speech material that we have.

We made two different groupings to define the central phone classes. For the first grouping, we have used an automatic aglomerative clustering of the discrete context-independent models of the phones to obtain four classes. This method is speaker-dependent because each speaker may have different class definitions. This modelling yields an average word accuracy of 74.3% (2cb) / 79.0% (3cb). Compared to the semicontinuous context-dependent, but without phone-class dependencies, we obtain an error reduction of 4.8% (2cb) and 6.3% (3cb), with a very small increment in the number of parameters (just the new set of gaussians for each class).

Let us, for example, calculate the number of parameters for both systems in the 2cb case:

Because we need both triphones and context-independent models, we will have 350 triphones + 56 context-independent phones - 3 units (the silences) that are only treated as context-independent units = 403 models.

For the case with no phone-dependent gaussians we have:

A matrix: 403 models x 3 states/model x 3 transitions/state = 3627 parameters.

Weights: 403 models x 3 states/model x 2 cb/state x 256 gaussian weights/cb = 619008 parameters.

Gaussians: 2 cb x 256 gaussians/cb x 11 parameters/cb x 2 parameters (mean and variance) / gaussian = 11264 parameters.

Adding up all these we get to 633899 total parameters.

For the case of phone dependency of the gaussians we have the same number of parameters for the A matrix and the Weights, but the gaussians are four times as many (one gaussian set per class). So the gaussian parameters are 11264 x 4 = 45056, and we get finally 667691 total parameters. You can also see in the next table the 3cb case:

SYSTEM	2 CB	3 CB
sc contextual	633.899	949.035
sc phone-dep. contextual	667.691	999.723

Table 2 - Number of parameters of the systems.

The results also show again an important (17.7% on average) reduction of the errors using the second order derivatives.

For the second grouping, once we had these class definitions for each speaker, we found common rules for all speakers such us nasals and fricatives in different classes, and decided to propose an unique phone-class definition for all speakers following the common rules.

The average word accuracy rises to 75.0% (2cb)/ 80.0% (3cb) . This class definition is convenient because it does not need automatic clustering and it is the same for all speakers.

The results show 7.4% (2cb) and 10.7% (3cb) reduction of the recognition errors compared to the no class-dependency system.

Of course, the increase in the number of parameters resembles that of the first grouping because only the way to define the classes has changed.

SPLITTING INTERWORD PAUSES

To further improve the modelling, we studied methods to incorporate automatic pausing detection to our training algorithms. In continuous speech, most of the words are directly concatenated without pauses and when the final phone of the previous word is the same as the first of the next word, they are often assimilated into only one.

The new training program decides whether an interword unit should be used or not automatically detecting pauses between words and making all possible phone assimilations. This algorithm will extract the actual transcription of the training sentences and lead us to new context-dependent generalized triphones. The resulting triphones are merged again into 350 clusters.

Using this new definition of the generalized triphones, we can not make any assumption about the first and final triphones of each word for the recognition vocabulary. We let the dynamic programming algorithm decide which is the best concatenating context depending on the previous word and calculate all possible contexts to concatenate to the following words.

Because of the use of generalized clustered triphones, we get an average of 10 different models to be used as the end of a word an to concatenate to following words. These different possibilities make the search algorithm to be a little more costly for these new units.

To train the system we first use a discrete HMM algorithm modified to produce automatic pausing detection. The criterium to allow a pause to appear between two words is the following: on each sentence we run first a Viterbi alignment of the states of the sentence against the observation frames. We then explore the frames that Viterbi assigns to each interword unit. If more than 3 of the frames (30 msec.) assigned to an interword unit, have energies below the minimum energy of the sentence file plus 20 dB, the interword unit is kept as a pause; otherwise it is deleted from the transcription of the sentence. A second Viterbi run is used to train the models with the new transcription.

The pause processing will change the triphones definition. For example a sentence that would have been transcribed so far as:

$< w_1 \,\&\, w_2 \,\&\, w_3 \,\&\, w_4 \,\&\, w_5 >$

Will now be transcribed as:

$< w_1\ w_2\ w_3 \,\&\, w_4\ w_5 >$

if only the third interword unit is validated. Looking in detail to the meaning of the deletion of the rest of interword units, it could have two cases:

1) if for example, in the sentence piece "... w_1 w_2 ..." the ending phone "p_{1m}" in word "w_1" is different from the beginning phone "p_{21}" in word "w_2", then the triphones in this piece of sentence will be:

... $[p_{1(m-2)}\ p_{1(m-1)}\ p_{1m}]\ [p_{1(m-1)}\ p_{1m}\ p_{21}]\ [p_{1m}\ p_{21}\ p_{22}]\ [p_{21}\ p_{22}\ p_{23}]$ ...

2) but, if for example, in the sentence piece "... $w_2\ w_3$ ..." the ending phone "p_{2m}" in word "w_2" is the same as the first phone "p_{31}" in the word "w_3", then we will asume an *assimilation effect*, because in continuous speech this phone will most of the times be uttered as only one assimilated phone. This means that now the transcription of this example will be:

... $[p_{2(m-2)}\ p_{2(m-1)}\ p_{assim}]\ [p_{2(m-1)}\ p_{assim}\ p_{32}]\ [p_{assim}\ p_{32}\ p_{33}]$ ...

being $p_{assim} = p_{2m} = p_{31}$ using the assimilation effect.

We will use this transcription procedure during the training process in all iterations and will save the transcription of the last iteration as the final transcription to be used to generate the triphones to be used in the rest of the system.

On average, across the 4 speakers, this new definition of the triphones produces 3066 triphones (we had 2424 with no pause processing), which proves a more rich contextual information and better tuned to the actual training material. Of course, these triphones are clustered again into only 350 different generalized units to assure they are trainable with the speech training material.

Thus, from this modified training algorithm we obtain both new phone models and a new transcription of the sentences with interword units only where the pauses have been detected in the last iteration.

Using the new discrete context-independent models generated by the new training algorithm, the average recognition is 59.1% (2cb) / 61.6% (3cb). Compared to the discrete models without pausing analysis, the reduction of errors is 6.4% (2cb) and 5% (3cb). It is shown the benefit of assigning the interword frames to the correct states. Using the new transcription to train semicontinuous context-independent models, the average recognition performs 67.5% (2cb) / 68.9% (3cb). Compared to the semicontinuous models without pausing information, we get an error reduction of 11.4% (2cb) and 6.9% (3cb).

Preliminary results of this method using discrete context-dependent generalized triphones give for one speaker 78.1% word accuracy (2cb) compared to 74.3% word accuracy (2cb) for the baseline system on the same speaker, i.e. a 14.8% reduction of the recognition errors.

SUMMARY OF RESULTS

SYSTEM	2 CB	3 CB
discrete context-independent (CI)	56.3	59.6
discrete CI with pausing	59.1	61.6
semicontinuous CI	63.3	66.6
semicontinuous CI with pausing	67.5	68.9
dicrete context-dependent (CD)	66.0	-
semicontinuous CD	73.0	77.6
semicontinuous CD phone-class dependent automatic clustering	74.3	79.0
semicontinuous CD phone-class dependent common rules	75.0	80.0

***Table 3** - Results for experiments on 4 speakers.*

SYSTEM	2 CB
discrete CD	74.3
discrete CD with pausing	78.1

***Table 4** - Preliminary result on one speaker.*

REFERENCES

[1] Hwang et al. "Improving Speech Recognition Performance via Phone-Dependent VQ Codebooks and Adaptive Language Models in SPHINX-II", IEEE International Conference on Acoustics, Speech and Signal Processing 94.

[2] A.M. Peinado, J.C. Segura, A.J. Rubio, M.C. Benitez "Using multiple vector quantization and semicontinuous hidden Markov models for speech recognition", IEEE International Conference on Acoustics, Speech and Signal Processing 94.

[3] J. Ferreiros, R. de Córdoba, M.H. Savoji, J.M. Pardo "Continuous speech HMM training system: Applications to speech recognition and phonetic label alignment", Granada, July 1993, en NATO ASI WORKSHOP "New advances and trends in speech recognition and coding".

[4] L. Devillers, C. Dougast, J. Ferreiros, R. de Córdoba, M.H. Savoji, J.M. Pardo, " Continuous speech recognition in Polyglot Project", ESPRIT SPEECH PROJECTS, Springer Verlag.