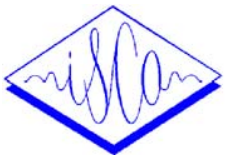


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DECISION TREES FOR INTER-WORD CONTEXT DEPENDENCIES IN SPANISH CONTINUOUS SPEECH RECOGNITION TASKS*

K. López de Ipiña, A. Varona, I. Torres L. J. Rodríguez.
Departamento de Electricidad y Electrónica. Universidad del País Vasco
Apartado 644. 48080 Bilbao. Spain. e-mail: karmele@we.lc.ehu.es

ABSTRACT

Context Dependent Units are broadly used in Continuous Speech Recognition (CSR) system, being decision trees a suitable clustering technique to obtain this kind of units. This work was aimed to extend the decision tree based clustering to model inter-word context dependencies in Spanish CSR tasks. We first used a set of previously defined context dependent units to model word boundaries. A decision tree derived pair grammar was then used at decoding time to prune each network connecting pairs of words. Then, specific sets of decision tree based inner context dependent units were obtained to model word boundaries. Both approaches were experimentally evaluated and compared to classical approaches over a Spanish CSR task. Experimental results showed the potential contribution of modelling between-word contexts to CSR systems. These units were selected by decision trees and provided full coverage while keeping a suitable computational cost.

1. INTRODUCTION

Context Dependent Units (CDUs), like triphones, diphones, demiphones, etc., are broadly used to get accurate acoustic models to be applied in Continuous Speech Recognition (CSR) tasks. These sets of units consider the influence of neighbouring phones in the acoustic model of each phone. Decision Tree based clustering is one of the most popular techniques to get an optimal set of trainable, discriminative and generalised context dependent units. This approach combines some phonetic knowledge of the language and some validation procedures based on the likelihood of the speech samples with regard to some stochastic models [2][3][4][5]. Moreover, sets of units obtained in such a way guarantee the full coverage of context dependencies.

When both left and right contexts are considered, words models do not usually exploit the full ability of CDUs to model context dependencies. Intraword context dependencies are fully defined and thus word models are simply obtained by concatenation of the corresponding CDUs. Nevertheless, a problem arises when looking for suitable CDUs for word boundaries, since outer contexts are not known. Thus, a network consisting of all the CDUs fitting the inner context should be considered, but this implies very high computational cost. Several proposals appear in the literature to deal with this problem [2][3][4]. They typically use specific sets of units at word boundaries: either context independent units, or one context (left or right) dependent units, or specifically trained units for word boundaries. However, these proposals do not fully exploit the ability of CDUs to model context dependencies. In previous works [1] some sets of decision tree based CDUs have been proposed for Spanish acoustic phonetic decoding. The aim of this work was to extend this methodology to model inter-word context dependencies in Spanish CSR tasks.

The paper is organised as follows. Section 2 presents a brief resume of the decision tree based clustering used to get CDUs. In Section 3, several proposals to deal with between-word context modelling are discussed. In Section 4, these proposals are experimentally evaluated over a Spanish CSR task. Finally, in Section 5 some concluding remarks are presented.

2. DECISION TREE-BASED CONTEXT DEPENDENT SUBLEXICAL UNITS

A set of decision trees associated to a previously established set of Context Independent phone like Units (CIU) were built. Each decision tree, associated to a given CIU, was built as follows. All the samples corresponding to that CIU were assigned to the root node. Then a set of binary questions, manually established by an expert phonetician, related to one or more left and right contexts, were made to classify the samples. Any given question Q divided the set of samples Y into two subsets, Y_l and Y_r. The resulting subsets were evaluated according to a quality measure, the so-called *Goodness of Split* (*GOS*) function, reflecting how much the likelihood of the samples increased with the split. Heuristic thresholds were applied to discard those questions yielding low likelihoods (*GOS* threshold) or unbalanced splits (trainability threshold). Among the remaining questions, the one giving the highest quality was chosen, thus appearing two new –left and right- nodes, being the samples distributed according to the answer (*YES/NO*) to that question. This procedure was iterated until no question exceeded the quality thresholds. As a consequence, a set of Decision Tree based Context

* Work partially supported by the Spanish CICYT under grant TIC99-0423-C06-03

Dependent Units (DT-DCU) was defined by the leaves associated to each CIU.

Each sample of the training corpus corresponding to a CIU consisted of a string of labels, obtained by a vector quantization procedure of the acoustic observation vectors. In fact, four different strings of labels were used simultaneously, each corresponding to a different acoustic observation vector quantization codebook. Following the classical scheme, a simple histogram was used to model acoustic events, each component of the histogram being modelled as a Poisson distribution. In fact, the model consisted of four different histograms, whose likelihoods were multiplied to yield the combined likelihood. To evaluate the quality of the splits the classical GOS function was applied [1]:

$$GOS(Y, M) = \log\left\{\frac{P(Y_l|M_l)\cdot P(Y_r|M_r)}{P(Y|M)}\right\}$$

where Y_l and Y_r stand for the sets of samples resulting of the split of set Y that were used to train models M_l, M_r and M respectively; $P(Y|M)$ is the joint likelihood of a set of samples Y with regard to a previously trained model M. This *GOS* function measures the likelihood improvement resulting from the split –i.e. from the question Q.

Decision trees were grown until any of the stopping criteria verified. Two thresholds were used, the first one establishing a minimum *GOS* value, and the second one giving the minimum number of training samples. After some preliminary experimentation, adequate values were heuristically fixed for these thresholds.

3. BUILDING WORD MODELS

The construction of word models can take a great advantage of the Decision Tree-based context dependent sublexical units (DT-CDUs). In the linear lexicon framework applied in this work, a more consistent word model results from the concatenation of context dependent units. Intraword contexts can be handled in a straightforward manner because left and right contexts are known, and DT-CDUs guarantee a full coverage of such contexts. A challenging problem arises when considering between-word contexts, i.e. the definition of border units, because outer contexts are not known. A lack of coverage is found for these situations: which contexts should be considered in word-boundaries?

A *brute force* approach would expand border units with all the context dependent units fitting the inner context. Full sets of DT-CDU according with the left context and with the right one should be considered at the end and at the beginning of each word, respectively. This leads to a nearly intractable combinatorial problem when dealing with a great search automaton. More usually, this problem is solved either by simply using context independent units, or by explicitly training border units [2] [3] [4]. However, these proposals only deal with context dependencies inside each word whereas those appearing between words -a great amount – are not considered. Thus, an appropriate balance need to be found between a good modelling of inter-word context dependencies and a suitable computational cost.

Three different approaches to represent inter word context dependencies were considered and tested in this

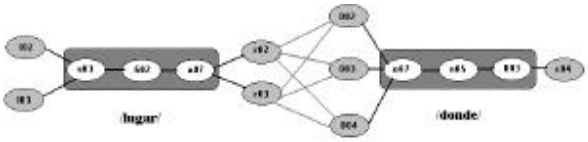


Figure I. Network connecting two DT-based context dependent units derived from phone /r/ at the end of the word /lugar/ to three units derived from /D/ at the beginning of the word /donde/.

work. DT-CDUs introduced in previous Section were used inside the words in any case.

3.1. Context independent units at word boundaries

Context Independent phone like Units (CIU) were used at word boundaries. As mentioned above, this approach involves a low computational cost but does not consider many acoustic influences of neighbouring phones.

3.2. Decision tree based pair grammar

This approach attempts to reduce the size of whole network obtained when full sets of DT-CDU according with the left context and with the right one are considered at the end and at the beginning of each word, respectively. Figure I shows such a network connecting two DT-CDUs derived from phone /r/ at the end of the word /lugar/ to three DT-CDUs derived from phone /D/ at the beginning of the word /donde/. Only the set of units derived from the tree of phone /r/ that could be preceded by phone /a/ are considered (unit labelled as r01 did not). In the same way, only /D/ derived units that could be followed by phone /o/ were considered (unit labelled as D01 did not). In any case, such a representation expands the transcription of the lexicon leading to a nearly intractable size of the search trellis at decoding time.

Nevertheless, the decision trees used to derive these context dependent units provide additional information that could be used to reduce the size of the word connection network. Each path connecting the root node of a phone decision tree to its leaves summarises the meaningful set of questions that classified each subset of samples appearing at the leaves. Thus, a careful review of these paths can establish the set of phones that could precede or follow each leaf of the tree. Figure II shows a decision tree of phone /D/. A simply analysis of the set of question labelling each split leads to conclude that only unit labelled by D04 can be preceded by phone /r/. Therefore, the network represented in Figure I can be strongly reduced: only arcs going to D04 need to be considered when connecting words /lugar/ and /donde/ (see Figure III). Figure II also shows that, as mentioned

above, unit labelled by D01 could not be followed by phone /o/.

Thus, a simply pair grammar has been derived from decision trees. The associated matrix represents the set of decision tree-derived context dependent units that can follow each phone. Such a matrix was then used to prune a great amount of arcs and thus reduce the computational cost involved to handle the whole network at decoding time (see Figure III).

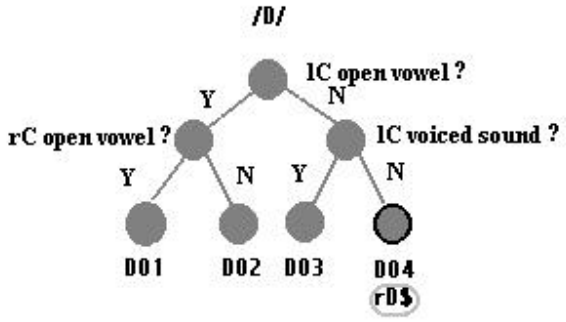


Figure II. A decision tree of phone /D/. Only unit labelled by D04 can be preceded by phone /r/. In the same way unit labelled as D01 can not be followed by phone /o/.

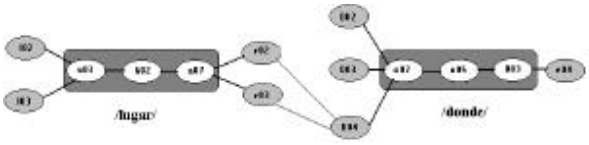


Figure III. Network represented in Figure I pruned by the knowledge derived from decision tree analysis. Decision tree of Figure II shows that only arcs connecting to D04 should be now considered.

3.3. Decision Tree based One context dependent units

Specific decision tree-based context dependent units were used at word boundaries. These sets of units were specifically obtained to be insideword context dependent and outsideword context independent, i.e. they were inner context dependent. Thus, two sets of Decision Tree based One Context Dependent Units needed to be established. To get the first one, the set of binary questions used to classify the samples at each node of the corresponding decision tree only applied to the right context (rC). Thus, these units were used to transcribe the first phone of words. In the same way, another set of units was obtained by only using binary questions about the left context (lC) context. This set was used to transcribe the last phone of each word. This procedure agrees with the classical decision tree methodology used to get context dependent units. Thus, full coverage of inner contexts is guaranteed while keeping outside context independence. On the other hand, the size of the lexicon as well as the computational cost of the search did not increase. Figure IV shows the decision tree for sound /D/ and right context dependency. The leaves of the tree were used to transcribe the sound /D/ when appearing at the beginning of a word. Figure V shows the transcription of word /donde/. The unit DO2 is the only leaf of the tree in Figure IV that agrees with sound /o/ as right context, being left context independent.

4. EXPERIMENTAL EVALUATION

An experimental evaluation of the proposed word models was carried out over a Spanish corpus.

The corpus used to obtain all the DT-derived context dependent units previously presented was composed of 1529 sentences, phonetically balanced and uttered by 47 speakers, involving around 60000 phones. These samples were then used to train the acoustic model of each DT-derived context dependent unit. Discrete HMMs with four observation codebooks were used as acoustic models in these experiments.

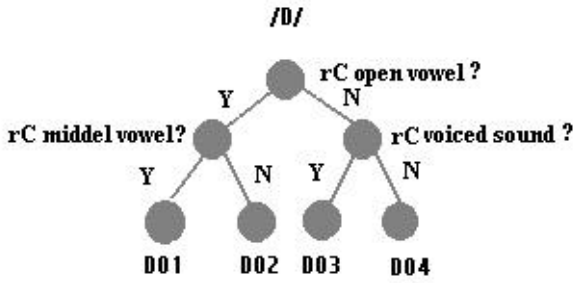
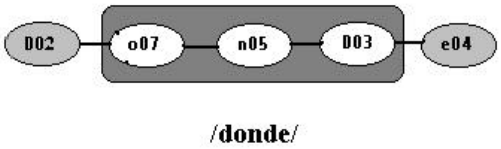


Figure IV. Decision Tree for sound /D/ when only right context is considered. The leaves of the tree will be used to transcribe sound /D/ when appearing at the beginning of a word.



/donde/

Figure V. Transcription of word /donde/ with an only right context dependent unit and an only left context dependent unit at the beginning and the end of the word, respectively. The unit DO2 is the only leaf of the tree in Figure IV that agrees with sound /o/ as right context, being left context independent.

The corpus used to obtain all the DT-derived context dependent units previously presented was composed of 1529 sentences, phonetically balanced and uttered by 47 speakers, involving around 60000 phones. These samples were then used to train the acoustic model of each DT-derived context dependent unit. Discrete HMMs with four observation codebooks were used as acoustic models in these experiments.

A task-oriented Spanish corpus (BDGEO) [6] consisting in 82,000 words and a vocabulary of 1,213 words was

used to evaluate word models. This corpus represents a set of queries to a Spanish geography database. For testing purposes, a subset of the test corpus consisting of 600 sentences and a vocabulary of 203 was used. No language model was applied in these experiments.

Three sets of sublexical units were used in these experiments:

- The first and simplest one consisted of 24 Context Independent phone like units (CIU) and it was used as a reference set.

- The second reference set represents the classical triphones. This set of context dependent units was simply obtained by selecting the more frequents in the training corpus (Freq-CDU). However, a lack of coverage is always found in the test corpus and, as a consequence, a mixture of triphones, diphones and monophones – selected in that order- needed to be used. Diphones and monophones were also used at word boundaries.

Table I. Word recognition rates (%WR) obtained when testing several sets of sublexical units and the word boundaries transcriptions proposed in Section 3.

Sets of units	% WR
CIU	49.83
Freq-CDU	51.16
DT-CDU	49.44
DT-CDU (DT-pair grammar)	52.55
DT-CDU (CI at word boundaries)	52.86
DT-CDU DT-inner context word boundaries	53.26

- The third set of sublexical units was the DT-CDUs set obtained trough the methodology described in Section 2.

Different lexicon transcriptions were applied according to the approach used to model word boundaries (Section 3), while keeping DT-CDU inside words:

- CI units at word boundaries (see Section 3.1)

- DT-CDU used also at word boundaries. Last and first sounds of words expand all possible DT-CDU (see Figure I).

- DT-CDUs were also used at word boundaries but the DT-derived pair grammar was used to prune each network connecting words (see Section 3.2.).

- DT-based inner context dependent units at word boundaries (DT- inner CDU) (see Section 3.3).

Table I shows the word recognition rates obtained trough these experiments, when no language model was used.

Table I shows that the use of DT-context dependent units outperformed the reference sets CIU and Freq-CDU, even when CI were used at word boundaries. The use of DT-CDU at word boundaries led to similar recognition rates when the DT-pair grammar was applied. However, the involved computational cost is higher. The use of a DT-derived pair grammar prune a significant number of connections between words while increasing the word recognition rates. The best system performance was achieved when DT based inner context dependent units were applied at word boundaries, being the involved computational similar to the use of CI units.

Nevertheless, experiments shown in Table I were not fully comparable. The number of path hypothesis at each connection between pair of words when using DT-CDU at words boundaries, with or without pair-grammar, was clearly higher than in the other experiments. Thus, the number of insertion errors is particularly high, being important in any case. Only the use of a Language Model and adequate weights to balance Language and Acoustic Models could lead to comparable experiments.

As a preliminary –prospective- step, several word penalty were applied to some of the experiments reported in Table I. Table II shows the obtained word recognition rates for several word penalty values.

Table II. Word recognition rates (%WR) for some of the experiments in Table I when several word penalty values were applied: 1000, 5000, 8000 and 10000.

	Word penalty			
	1000	5000	8000	10000
CIU	52,02	54,31	54,76	55,01
DT-CDU (DT-pair grammar)	55,01	57,36	57,92	58,25
DT-CDU (CI at word boundaries)	54,43	56,22	56,64	56,73
DT-CDU DT-inner context word boundaries	56,28	58,78	59,35	59,50

Table II shows that the use of a word penalty factor increased the system performance in any case. The use of DT- CDU outperformed the reference set CIU as in previous experiment did. However, in this case word boundaries were clearly better represented by the set of DT-CDU pruned by the DT-pair grammar than by the set of CI units. The best system performance was also achieved in these experiments when DT based inner context dependent units were applied at word boundaries, being the involved computational similar to the use of CI units.

5. CONCLUDING REMARKS

This work was aimed to extend the decision tree based clustering to model inter-word context dependencies in Spanish CSR tasks. Several approaches to handle border units in the construction of word models were described and tested. Results showed the potential contribution of modelling between-word contexts to speech recognition. The best system performance was obtained when using specific sets of insideword context dependent units at word boundaries, being outside word context independent. These units were selected by decision trees and provided full coverage while keeping a suitable computational cost. In further works a language model need to be applied to confirm the approaches and results presented in this work.

6. REFERENCES

[1] López de Ipiña, K., L.J. Rodríguez,, A. Varona and I. Torres (1999), Decision Tree-Based Context Dependent Sublexical Units for Spanish Speech Recognition Task. *Proc. of the VIII Simposium Nacional de Reconocimiento de Formas y Análisis de Imágenes.*

[2] Bahl, L.R., .V.P. de Souza, P.S. Gopalakrishnan, D. Nahamoo and M.A. Picheny (1994), Decision Trees for Phonological Rules in Continuos Speech Recognition, *Proceedings of ICASSP-94*, pp.533-536.

[3] Kuhn, R., A. Lazarides and Y. Normandin (1995), Improving Decision Trees for Phonetic Modelling, *Proceedings of ICASSP-95*, pp, 552-555.

[4] Odell, J. (1995), The Use of Context in Large Vocabulary Speech Recognition. *Ph. Thesis*. Cambridge University.

[5] Young, S.J., J. Odell and P.. Woodland (1994), Tree-Based State Tying for High Accuracy Acoustic Modelling, *Proc. ARPA Workshop Human Language Technology*, pp. 286-291.

[6] Diaz, J.E., A.J. Rubio, A.M. Peinado, N. Prieto and F. Casacuberta (1993), Developmente of Task Oriented Spanish Speech Corpora, *Proceedings of Eurospeech-93*.