DECISION TREE-BASED CONTEXT DEPENDENT SUBLEXICAL UNITS FOR SPANISH CONTINUOUS SPEECH RECOGNITION TASKS

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

ABSTRACT

This paper presents a new methodology, based on the classical decision tree classification scheme proposed by Bahl [1], to get a suitable set of context dependent sublexical units in Spanish continuous speech recognition tasks. The original method was applied as a first baseline approach. Then two new features were added: a discriminative function to evaluate the quality of the splits and the use of discrete HMMs to compute the likelihoods. A second approach was explored, based on the fast and efficient Growing and Pruning algorithm fitting both the size and the acoustic modelling capability of the decision trees. In addition, the use of these units to build word models was addressed, considering only intraword contexts. The baseline approach gave recognition rates clearly outperforming those of context independent phone-like units. Then the two new features and the alternative methodology outlined above were evaluated. Recognition rates were similar to those of the baseline approach, being the discriminative function the most promising feature. Finally, modelling explicitly the between-word contexts appearing in the test database made a prospective attempt. This approach gave the best results, suggesting further work in pronunciation modelling using context dependent phone-like units[1].

Keywords: Sublexical Units, Decision Trees, Growing and Pruning Algorithm

1. INTRODUCTION

The choice of a suitable set of sublexical units is one of the most important issues in the development of a Continuous Speech Recognition (CSR) system. As shown in the literature, authors have proposed a wide range of them: diphones, triphones and other context dependent units [2] [3] [4] [5], transitional units [6] and, lately, the so-called demiphones [7]. Such a variety of approaches aims at accurately model the influence of contexts in the realisation of Context Independent Phone-Like Units (CI-PLUs). System efficiency can exploit the benefits of context modelling by using context dependent sublexical units to generate lexical baseforms, taking into account not only intraword but also between-word contexts, as we will see.

[1] This work has been partially supported by the Spanish CICYT, under project TIC95-0884-C04-03.

Decision Trees (DT) are one of the most common approaches to the problem of selecting a suitable set of context dependent sublexical units for speech recognition [1] [8] [9]. DT combine the advantages of applying some phonetic knowledge about how contexts affect the articulation of speech and a strictly quantitative validation procedure based on the likelihood of speech samples with regard to some probabilistic models.

A binary tree is built from the training samples corresponding to a given CI-PLU. Each DT node poses a question about the phonetic identity of one or more left and right contexts, so that each training sample is taken from the root node to one of the leaf nodes, which stands for a generalised context category depending on the questions made. Therefore, the leaf nodes behave as contextual allophones of the original CI-PLU. Then, each CI-PLU, with its phonetic context in a sentence, can be classified just by answering a number of DT questions before a leaf node is reached.

In this work DT have been used to model both intraword and between-word context dependencies. Starting from the classical scheme [1], some attempts have been made in order to improve the accuracy and the discriminative power of the models. An alternative methodology, the fast and efficient Growing and Pruning algorithm [10], has also been applied to build the decision trees.

The paper is organised as follows. Section 2 reviews the basic DT methodology, describing more carefully those points where major changes have been introduced. Section 3 presents the alternative DT methodology, based on the Growing and Pruning algorithm. In section 4, the issue of between-word context modelling is discussed and some solutions are proposed. Finally, in Section 5 DT-based Context Dependent Units (DT-CDUs) are applied to a Spanish CSR task, and experimental results are discussed.

2. THE BASELINE METHODOLOGY

Firstly, automatic segmentation of the training corpus was carried out to get the set of samples corresponding to each of the CI-PLUs, each sample consisting of a string of labels, obtained by vector quantization of the acoustic observation vectors. In fact, four different strings of labels were used simultaneously, each corresponding to a different acoustic observation VQ codebook.

Each DT, associated to a given CI-PLU, was built as follows. All the samples corresponding to that CI-PLU 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, a *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.

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:

$$GOS1(Q:\{Y,M\}\rightarrow\{Y_l,M_l\}\cup\{Y_r,M_r\})=\log\left\{\frac{P(Y_l|M_l)\cdot P(Y_r|M_r)}{P(Y|M)}\right\}$$

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.

Although -as we will see- context dependent units obtained using this configuration clearly outperformed the CI-PLUs, further improvements could be expected from two important changes: a) to define a discriminative GOS function, and b) to compute the acoustic likelihoods by using Multiple Codebook Discrete HMMs (MC-DHMMs).

A modified GOS function aiming at a higher discrimination between the two models arising from the split -instead of a higher internal likelihood of the samples belonging to each model- should increase the discriminative power of the set of context dependent units. The proposed function, defined as a direct to cross likelihood ratio weighting the original GOS function, tried to emphasise the differences between the two sets of samples resulting from the split:

$$GOS2(Q:\{Y,M\}\rightarrow\{Y_l,M_l\}\cup\{Y_r,M_r\})=\log\left\{\frac{P(Y_l|M_l)\cdot P(Y_r|M_r)}{P(Y|M)}\,\frac{P(Y_l|M_l)\cdot P(Y_r|M_r)}{P(Y_l|M_r)\cdot P(Y_r|M_l)}\right\}$$

On the other hand, the use of a more reliable model to compute the acoustic likelihoods should make more accurate the splitting procedure. At each DT node and for each possible question, three different MC-DHMMs were trained with: a) the whole set of samples belonging to that node, b) the subset of samples answering *YES* to the question, and c) the subset of samples answering *NO* to the question.

It seems clear that HMMs should fit the acoustic events more accurately than simple histogram models (Poisson distribution), which measure just an instantaneous acoustic likelihood and cannot model the time evolution of speech. At each discrete HMM state, a Gaussian distribution provides the acoustic likelihood of any given label (*emission probabilities*). Moreover, the HMM topology, which fixes the probability of any given sequence of states by applying the *transition probabilities*, allows to explicitly model the time characteristics of speech. At the final state of a discrete HMM, the joint emission and transition probability gives the likelihood of any given string of labels, at the expense of a high computational cost. Only experimental evaluation will prove the contribution of HMMs to this procedure.

3. THE GROWING AND PRUNING METHODOLOGY

As said above, DTs were grown until any of the stopping criteria verified. Two thresholds were used, the first one establishing a minimum GOS value, the second one giving the minimum number of training samples. After some preliminary experimentation, adequate values were heuristically fixed for these thresholds. This is a very simple but inconvenient way to stop the growing procedure, because for each training database some preliminary experimentation must be made to fix adequate values for the thresholds.

An alternative methodology was designed to overcome this problem, based on the fast and efficient Growing and Pruning (G&P) algorithm [8] [10]. The G&P algorithm divides the set of training samples corresponding to a given CI-PLU into two independent subsets. The tree is iteratively grown with one of the subsets, and pruned with the other, interchanging the roles of the two subsets in successive iterations.

The growing procedure was identical to that described in Section 2, but removing the GOS threshold. Only a minimum number of training samples was required for a node to be valid. As a second step, once a big DT was built, the pruning procedure applied a misclassification measure to discard those leaf nodes below a given threshold. It can be shown that the algorithm converges after a few steps [10].

Among the DT building methods, G&P provides a good balance between classification accuracy and computational cost, compared to other methods like CART [10] [11]. Note, however, that a threshold must be still heuristically fixed to control the size of the sample sets associated to the leaf nodes, because a minimum number of samples is necessary for the acoustic models to be trainable.

4. BUILDING WORD MODELS

The construction of word models can take a great advantage of the DT 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 are 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, and a lack of coverage is found for these situations.

Which contexts should be considered outside the edges of words? A *brute force* approach would expand these border units with all the context dependent units fitting the inner context. This leads to a nearly intractable combinatorial problem when dealing with a great search automaton. Usually, this problem is solved either by simply using context independent units, or by explicitly training border units [1] [8] [9] [12].

Three different approaches to represent inter word context dependencies were considered and tested in this work. DT-CDUs introduced in previous Section were used inside the words in any case.

a) Context Independent phone like Units 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.

b) Decision Tree based One context dependent units Specific decision tree-based context dependent units were used at word boundaries[13]. 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. 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. 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.

c) DT-CDUs also at the edges of words. As a preliminary -prospective- step, words appearing in the test database were explicitly transcribed using DT-CDUs according to their left and right contexts. In other words, DT-CDUs were used as border units, taking into account not only the inner context -which is always the same- but also the specific outer context -which depends on the adjacent word appearing in each particular test sentence. The experimental results obtained by this approach established an upper bound to the benefits attainable by using context dependent sublexical units to build word models.

5. EXPERIMENTAL EVALUATION

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) [14] 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.

5.1. Acoustic-phonetic decoding experiments

Three groups of sublexical units were used in these experiments:

- The first and simplest one consisted of 24 Context Independent phone like units (CIU-PLU) 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). A mixture of 103 triphones, diphones and monophones, was obtained –in that order- just by selecting those whose appearing counts in the training database exceeded a heuristically fixed threshold.

- The third group of sublexical units was the DT-CDUs set obtained trough the methodology described in Section 2. Both the standard approach -with and without the new features described above- and the G&P approach, were used to generate the corresponding DT-CDUs. The standard approach, using a set of phonetic questions about one left and one right contexts and three different thresholds controlling the size of the training sets, was applied to get the sets DT-std1, DT-std2 and DT-std3. The first one was considered optimal, so that equivalent thresholds to that of DT-std1 were applied in later experiments. A set of phonetic questions about two left and two right contexts was applied to obtain the set DT-std4. One left and one right contexts were explored in later experiments, unless another context window size was explicitly noted. The standard approach, but replacing the original GOS function with the discriminative GOS function, both defined in Section 3, was used to obtain the set DT-dis. The sets DT-hmm1 and DT-hmm2 were obtained by applying the standard approach but replacing the simple histogram models with MC-DHMMs to compute the likelihoods, and using sets of phonetic questions about one and five contexts, respectively. Finally, the G&P approach was applied to obtain the set DT-g&p. Results are shown in Table 1.

Table 1. Recognition rates for various sets of sublexical units in a speaker independent acoustic-phonetic decoding task.

Type of units	Context window size	Acoustic Models	GOS function	#Units	% REC
CI-PLU	-	-	-	24	63.97
Freq-CDU	-	-	-	103	65.90
DT-std1	1	Histograms	Standard	101	66.44
DT-std2	1	Histograms	Standard	178	66.30
DT-std3	1	Histograms	Standard	359	65.75
DT-std4	2	Histograms	Standard	102	66.48
DT-dis	1	Histograms	Discriminative	133	66.27
DT-hmm1	1	Discrete HMMs	Standard	129	66.51
DT-hmm2	5	Discrete HMMs	Standard	131	66.63
DT-p&g	1	Histograms	Standard	146	66.23

From these results we conclude that DT-CDUs outperform the reference sets CI-PLU and Freq-CDU. However, the two new features added to the standard DT methodology did not improve the performance. In fact, the best result (66.63%), obtained for DT-hmm2, is only slightly better than that obtained for DT-std1 (66.44%).

The alternative G&P methodology did not improve the performance either (66.23%). The small size of the training database and the limited efficiency of discrete HMMs could explain this poor result. Only 60000 samples were used to train around 150 sublexical units, which gives an average of 400 samples per unit. In fact, this could also explain the trend observed in the performances of DT-std1, DT-std2 and DT-std3. On the other hand, the G&P methodology performed faster than the standard. Most times the procedure did converge in two steps, each step involving half the samples of the standard methodology, thus providing considerable timesavings.

5.2. Word-level experiments

This second series of experiments was aimed to evaluate the proposed DT-CDU when used to build word models. Different lexicon transcriptions were applied according to the approach used to model word boundaries (Section 4), while keeping DT-CDU inside words: Context Independent Phone Like Units (CI-PLU), Decision Tree based inner context dependent units and general context dependent units (DT-CDU). For comparison purposes, CI-PLU's and the previously defined mixture (Freq-CDU) of triphones, diphones and monophones (diphones and monophones at word boundaries) were also used to build word models. Experimental results are shown in Table 2.

Table 2. Word recognition rates in a continuous speech recognition task, without language model, for various sets of sublexical units and three different approaches to the definition of border units.

		units used at word boundaries		
		CI_PLU	**DT-inner context**	**DT-CDU**
CI_PLU	49.83			
Freq-CDU	51.16			
DT-std1		52.86	53.26	58.01
DT-std3		36.19	40.62	56.03
DT-dis		51.11	52.86	58.38
DT-hmm1		51.20	52.00	57.03
DT-g&p		47.00	48.30	56.45

DT-PLUs outperformed the reference sets CI-PLU and Freq-CDU in most cases. Only DT-std3 and DT-g&p showed a clearly worse performance, maybe due to a lack of training samples. As expected, the use of DT-CDU at word boundaries led to the best results, establishing an upper bound to the benefits attainable by using context dependent sublexical units to build word models. This reveals the contribution of modelling between-word context to the speech recognition, and suggest further work in that line [13].

DT-std1 gave the best recognition rates, being the best choice when handling isolated words (around 53%), and the second one when handling connected words (58.01%), only slightly worse than DT-dis (58.38%). This later result suggests that the discriminative GOS function proposed in this paper, maybe with further refinements, could be a good alternative to the standard GOS function. On the other hand, the choice of discrete HMMs –instead of simple histograms- to compute the likelihoods did not improve the performance, but increased the computation time. Finally, the G&P methodology performed clearly worse than the standard methodology, given a fixed –relatively small- number of training samples. A bigger database should be used to validate this result.

6. CONCLUDING REMARKS

The classical decision tree classification scheme was used and adapted to obtain a suitable set of context dependent sublexical units for Spanish CSR tasks. Two different GOS functions were used: the standard and a new discriminative function emphasising differences between two sets of samples. Alternatively, discrete HMMs, instead of simple histograms, were used to compute likelihoods during the DT building procedure. An alternative methodology, based on the fast and efficient G&P algorithm, was also proposed. Various sets of DT-based context dependent sublexical units were tested in a first series of speaker independent acoustic-phonetic decoding experiments, outperforming two previously defined reference sets. Three different strategies to handle border units in the construction of word models were described and tested in a second series of experiments. Results showed the potential contribution of modelling between-word contexts to speech recognition, and suggest further work in that line.

REFERENCES

[1] L.R.Bahl, .V.P. de Souza, P.S. Gopalakrishnan, D. Nahamoo, M.A. Picheny. Decision Trees for Phonological Rules in Continuos Speech Recognition, Proc. IEEE ICASSP-94, pp.533-536.

[2] C.H. Lee, L.R. Rabiner, R.Pieraccini and J.G. Wilpon. Acoustic Modelling for Large Vocabulary Speech Recognition. Computer Speech and Language, Vol. 4, pp. 125-165, 1990.

[3] K.F. Lee. Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuos Speech Recognition. IEEE Trans. on ASSP, Vol. 38, pp. 599-609. 1990.

[4] H. Ney and R. Billi.Prototype Systems for Large Vocabulary Speech Recognition: Polyglot and Spicos. Proc. Eurospeech-91, pp. 193-200.

[5] L. Frissore, E. Giachin, P. Laface, G. Micca. Selection of Speech Units for a Speaker- Independent CSR Task. Proc. Eurospeech-91, pp. 1389-1392.

[6] A. Varona, I. Torres, F. Casacuberta. Discriminative-Transitional Units for Spanish Continuous Speech Recognition. Proc. Eurospeech-95, pp. 1200-1203.

[7] J.B. Mariño, A. Nogueiras, A. Bonafonte. The demiphone: an efficient subword unit for continuous speech recognition. Proc. Eurospeech-97, pp. 1215-1218.

[8] R. Kuhn, A. Lazarides, Y. Normandin. Improving Decision Trees for Phonetic Modelling. Proc. IEEE ICASSP-95, pp, 552-555.

[9] J. Odell. The Use of Context in Large Vocabulary Speech Recognition. Ph. Thesis. Cambridge University. March 1995.

[10] S.B. Gelfand, C.S. Ravishankar, E.J. Delp. An Iterative Growing and Pruning Algorithm for Classification Tree Design. IEEE Trans. on PAMI, Vol. 13, No. 2, pp. 163-174. 1991.

[11] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone. Classification and Regression Trees. Wadsworth&Brooks, 1984.

[12] S.J. Young, J. Odell, P. Woodland. Tree-Based State Tying for High Accuracy Acoustic Modelling. ARPA Workshop on Human Language Technology, pp. 286-291, March 1994.

[13] K. López de Ipiña, A. Varona, I. Torres L. J. Rodríguez, Decision Trees for Inter-Word Context Dependencies in Spanish Continuous Speech Recognition Tasks. EUROSPEECH99. Budapest.

[14] J.E. Diaz, A.J. Rubio, A.M. Peinado, E. Segarra, N. Prieto and F. Casacuberta. Development of Task Oriented Spanish Speech Corpora. Proc. EUROSPEECH-93. 1993, included in addendum.