DECISION TREE-BASED CONTEXT DEPENDENT SUBLEXICAL UNITS FOR CONTINUOUS SPEECH RECOGNITION OF BASQUE

K. López de Ipiña[1], M. Graña[2], N. Ezeiza[3], M. Hernández[2], E. Zulueta[1], A. Ezeiza[3]

[1]Sistemen Ingeniaritza eta Automatika Saila Gasteiz. email: {isplopek, iepzugee}@vc.ehu.es
[2] Konputazio Zientziak eta Adimen Artifiziala Saila, Donostia. email:{ccpgrrom, ccphegom}@si.ehu.es
[3]IXA group, Donostia. email: {aitzol, nerea}@si.ehu.es

University of the Basque Country

ABSTRACT

This paper presents a new methodology, based on the classical decision trees, to get a suitable set of context dependent sublexical units for Basque Continuous Speech Recognition (CSR). The original method proposed by Bahl [1] was applied as the benchmark. Then two new features were added: a data massaging to emphasise the data and a fast and efficient Growing and Pruning algorithm for DT construction. In addition, the use of the new context dependent units to build word models was addressed. The benchmark Bahl approach gave recognition rates clearly outperforming those of context independent phone-like units. Finally the new methodology improves over the benchmark DT approach.

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, transitional units or demiphones. Such a variety of approaches aim at the accurate model of the influence of contexts in the realisation of Phone Like Context Independent Sublexical Units (PL-CI-SLUs). 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.

Decision Trees (DT) are one of the most common approaches to the problem of selecting a suitable set of context dependent sublexical units (DT-CD-SLUs) for speech recognition [1][2][3].

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.

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 [4], 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 and on the data massaging. 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 and Semicontextual Units (DT-CD-SLUs and DT-SC-SLUs) are applied to a Basque CSR task, and experimental results are discussed. Conclusions are summarised in section 6.

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 PL-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 PL-CI-SLUs, was built as follows. All the samples corresponding to that PL-CI-SLUs 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 partitioned 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:

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

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.

3. METHODOLOGICAL IMPROVEMENTS

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 thresholds must be fixed for each training database.

An alternative methodology was designed to overcome this problem, based on the fast and efficient Growing and Pruning (G&P) algorithm [4].

The G&P algorithm divides the set of training samples corresponding to a given PL-CI-SLUs 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. 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 leaf nodes below a given threshold. It can be shown that the algorithm converges after a few steps [4]. Among the DT building methods, G&P provides a good balance between classification accuracy and computational cost, compared to other methods like CART [5]. Note, however, that we use an alternative to the classic G&P. A new 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.

Preprocessing the data (data massaging) may improve the performance of DT when databases are small. In this work we have computed the square to each histogram element to emphasise it, obtaining a better discrimination.

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

Type of units	Context window size	G&P	Preprocessing	#Units	% REC
CI-PLU	-	-	-	28	64.01
DT-std1	1	-	Standard	256	71.10
DT-std2	1	-	Standard	217	71.45
DT-g&p1	1	G&P	Standard	220	71.32
DT-g&p2	2	G&P	Standard	234	70.99
DT-g&p-mass	1	G&P	Data-massaging	215	**71.52**

4. THE WORD MODELS

The construction of word models can take a great advantage of the DT-CD-SLUs. In the linear lexicon framework applied in this work, a more consistent word model results from the concatenation of this kind of units. Intraword contexts are handled in a straightforward manner, because left and right contexts are known and DT-CD-SLUs 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 an intractable combinatorial problem when dealing with a large search automaton. Usually, this problem is solved either by simply using context independent units, or by explicitly training border units [1] [2] [3] [6].

Two different approaches to represent inter word context dependencies were considered and tested in this work. DT-CD-SLUs introduced in previous section were used inside the words in any case.

a) PL-CI-SLUs 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 Semicontextual sublexical (DT-SC-SLUs) units. Specific decision tree-based context dependent units were used at word boundaries [7]. These sets of units were specifically obtained to be insideword context dependent and outsideword context independent. These units were obtained using binary questions about either the left context or the right 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.

5. EXPERIMENTAL EVALUATION

The corpus used to obtain all the DT-CD-SLUs previously presented was composed of 10000 sentences, phonetically balanced and uttered by 40 speakers, involving around 200000 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 has been created for this purpose. The Miniature Language Acquisition (MLA) [8] in Basque has 15,000 sentences with about 150,000 words, being 47 the vocabulary size. It has very low perplexity and very restrictive vocabulary size. It was created for preliminary experiments of CSR. Then, the task underwent an automatic morphological segmentation and we created two sets of lexical units as alternative to the words. We considered these new lexical units because Basque is an aglutinative language [9]. Thus, MLA task reduces the vocabulary size to 35 pseudo-morphemes (PS-MORPHS). Finally, N-WORDS acoustically more robust units [9] were obtained resulting in, 40. The sentences of MLA task were divided into 14,500 sentences for training and 500 for test. 20 speakers, 10 males and 10 females, recorded the task, obtaining 400 sentences.

5.1. Acoustic-phonetic decoding experiments

Two groups of sublexical units were used in these experiments:

- The first and simplest one consisted of 24 PL-CI-SLU and it was used as a reference set.

- The second group of sublexical units was the DT-CD-SLUs 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-CD-SLUs. The standard approach, using a set of phonetic questions about left and one right contexts and two different thresholds controlling the size of the training sets, was applied to get the sets **DT-stdN**. The standard approach, but replacing the standard data by the massaging data defined in section 3, was used to obtain the set **DT-mass**. Finally, the G&P approach was applied to obtain the sets **DT-g&p**. Results are shown in table 1.

From these results we conclude that DT-CD-SLUs outperform the reference sets CI-PLU and Freq-CDU. The two new features added to the standard DT methodology improve the performance. In fact, the best result (71.52%) obtained for **DT-g&p-mass**, integrate two methods only slightly better than the obtained for **DT-std1** (71.45%) but improve the result obtained for G&P in [7] for Spanish. 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.

Table 2. Word recognition rates in a Basque CSR task (MLA), without language model, for various sets of sublexical units and three different approaches to the definition of border units by using three different sets of lexical units: WORDS, PS-MORPHS and N-WORDS

	units used at word boundaries					
	WORDS		**PS-MORPHS**		**N-WORDS**	
	PL-CI-SLU	**DT-SC-SLU**	**PL-CI-SLU**	**DT-SC-SLU**	**PL-CI-SLU**	**DT-SC-SLU**
CI-PLU	80.61	-	-	-	-	-
DT-std1	86.73	87.68	84.03	83.68	69.20	72.13
DT-g&p2	86.20	87.41	83.12	83.44	69.54	72.52
DT-g&p-mass	86.43	**90.75**	83.96	**84.19**	69.68	**73.33**

5.2. Lexical unit-level experiments

This second series of experiments was aimed to evaluate the proposed DT-CD-SLUs 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-CD-SLUs inside words: PL-CD-SLUs and DT-SC-SLUs.

The experiments have been carried out without grammar an in the case of morphemes and N-WORDS the output was aligned to words to compare the results appropriately. Experimental results are shown in table 2. DT-CD-SLUs outperformed the reference sets PL-CI-SLUs in all cases. As expected, the use of DT-SD-SLUs 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 suggests further work in that line.

DT-g&p-mass gave the best recognition rates, being the best choice when handling isolated lexical units both with PL-CI-SLUs (86,43% for words, 83,96 for N-WORDS and 69,68 for PS-MORPHS) or DT-SC-SLUs SLUs (90,75% for words, 84,19 for N-WORDS and 73,33 for PS-MORPHS). Finally, the G&P methodology has a performance similar to standard methodology, with a very low computational cost.

6. CONCLUDING REMARKS

The classical decision tree classification methodology was improved to obtain a suitable set of context dependent sublexical units for Basque CSR tasks. A data massing methodology was used to emphasising differences among the samples. 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, where our methodology outperforms the classical one proposed by Bahl. Two 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.

ACKNOWLEDGEMENTS

The authors would like to thank all the volunteer speakers that has collaborated recording the databases. We thank also all people have collaborated in the development of this work. This work has been partially supported by the University of the Basque Country, under project UPV00147.345-E-14895/2002.

REFERENCES

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

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

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

[4] Gelfand S.B., C.S. Ravishankar and 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.

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

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

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

[8] Feldman J.A., A. Lakoff, A. Stolcke and S.H. Weber, *"Miniature language Acquisition: a touch-stone for cognitive science"*. Technical Report, Tr-90-009. ICSI, Berkeley, California. Abril 1990.

[9] Lopez de Ipiña K., I. Torres, L. Oñederra, M. Hernandez, M. Peñagarikano, N.Ezeiza, A.Varona and L.J. Rodríguez. *"First Selection of Lexical Units for Continuos Speech Recognition of Basque"*, Proceedings of ICSLP. Beijing 2000, vol II, pg. 531-535