DOUBLE TREE BEAM SEARCH USING HIERARCHICAL SUBWORD UNITS

Juan Carlos Torrecilla, Ismael Cortazar, Luis Hernandez
Speech Technology Group
Telefónica Investigación y Desarrollo
C/ Emilio Vargas, 6
28043 - Madrid, SPAIN
e-mail: jcarlos@craso.tid.es, ismael@craso.tid.es, luish@craso.tid.es

Abstract

In this paper we proposed an efficient beam search procedure that combines well-known search techniques as a lexicon organization using tree-structured grammars with a novel approach of using different types of subword units depending on the local scores of the active words. An efficient double-tree structure using phonemes and triphones is presented. Experimental results on an isolated word recognition systems reveals that the proposed strategy is suitable for important reductions in computational cost with only negligible increases in recognition errors. Tests over a vocabulary of 955 Spanish words presents a 0.5% of increase in error rate for a 32% reduction in the number of senones to be evaluated.

1.- Introduction

As the vocabulary size of an Automatic Speech Recognition (ASR) system grows, it becomes more important to develop improved beam search algorithms in order to maintain recognition complexity and system response delay under acceptable values.

In this paper we introduce a new beam-search technique for HMM based ASR systems. In the proposed technique the benefits of a double tree organization of the grammar are extended with the use of different types of subwords units depending on the frame-by-frame word recognition probabilities. That is, we combine two major computation reduction strategies:

- Tree organization: As the vocabulary size of an ASR system increases, most of the computation effort is concentrated on searching the initial units of each vocabulary word. Therefore, we use a tree organization of the lexicon which is a well-known strategy to importantly reduce computational cost.

- Different types of subword units: The performance of an ASR system increases as it uses more detailed sub-word units. However, more detailed acoustic units usually correspond to an increase in the number of subword units. And thus a high number of subword units results in an important increase in computational cost. For example, in Spanish there are 24 basic context-independent (CI) units [2], that could generate up to 576 left or right context-dependent units, or 13824 triphones. In our efficient search procedures different types of units are used: phonemes and triphones are hierarchically used depending on the frame-by-frame word probabilities. For each frame, word probabilities are computed and three levels of probabilities are established. Those words belonging to the first level (the highest probability level) use triphones, those belonging to the second (CI level) use context-independent units, and words in the third level are desactivated.

In the figure 1 it can be seen a word that is desactivated at the beginning of the recognition because it has a probability under the CI level. Later, its probability improves, and the word gets activated with the CI version. As the probability gets on improving, the word changes from its CI version to the triphone one and at the end of the utterance reaches the best recognition probability. The global process is like a softer pruning.

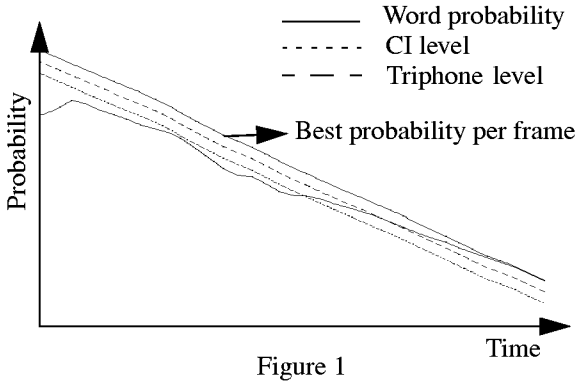

Figure 1

The rest of the paper is organized as follows: Section 2 presents the basic concepts of a hierarchical lexicon organization using a double-tree structure. Section 3 describes the behaviour of the beam search algoritm using the structure presented in Section 2. Experimental results and conclusions are given in Sections 4 and 5 respectively.

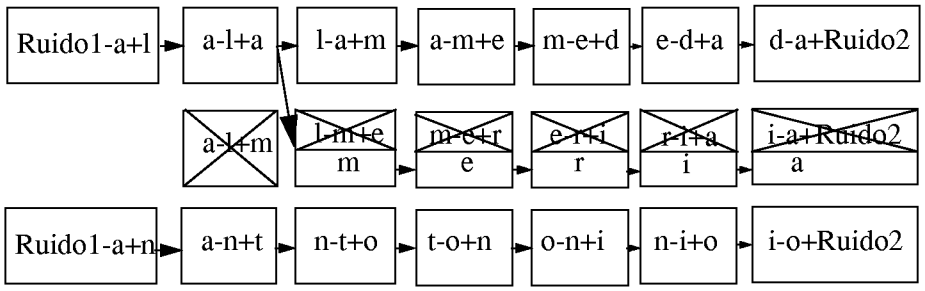

Figure 2: Changing from triphones to context-independent units

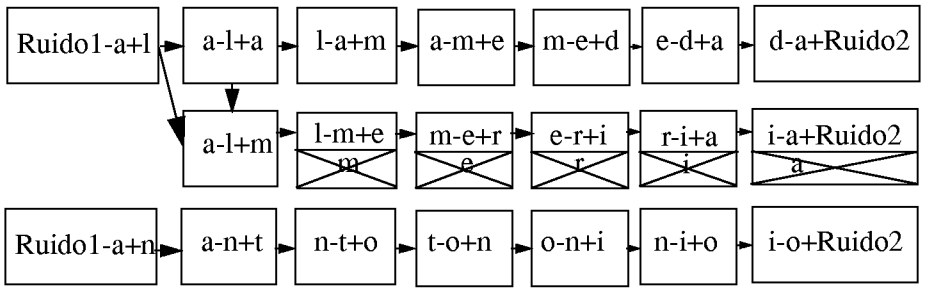

Figure 3: Changing from context-independent to triphones units

2.- The Double Tree

In this work we only consider a two-level hierarchical search using triphones as highly detailed Context Dependent (CD) units and phonemes for less-detailed CI units. Thus during decoding and for both triphones and context-independent units a tree based grammar is used: we call it a double tree. It must be noted that two different trees are needed, because in the triphone case fewer units are shared at the beginning of the words. For example, the spanish words "alameda" and "almeria" share the CI units "a" and "l", but using triphones we only share triphone "Ruido1-a+l". (The label Ruido1 references the silence or noise at the beginning of the word)

Figure 2 shows a case in which the probability of the word "almeria" gets worse, and then it begins to use its context-independent version. In this case our approach is to use the two first CD units for "almeria" shared with its root word "alameda" and the rest of the units modeled as CI. The rationality of this is double: with our approach we mantain both a high-degree of sharing for the tree-grammar organization and the precission in the acoustic representation of the initial states of words using CI models.

Using the same approach, in those cases where the probability of a word improves and changes to be represented using triphones two steps are necessary. First it is necessary to activate the probability of the second CD unit of the word (because as it was seen in the example of Figure 2 in the CI representation this unit was represented by the CD of its root word), and this is done by copying the score of the second unit of the root word into the second unit of the CD activated word, the triphone a-l+m in Figure 3. The second step is to intialize all the other CD units with the corresponding scores of the CI units.

To have a first look on the benefits of using our double-tree structure we compared it with a single-tree one using triphones. Figure 4 represents a typical evolution of the number of active subword units during the recognition of an utterance for both cases. As it can be seen in the figure the double tree case presents an evolution in number of active subword units that decreases along time. This is because as the probability of the worst words decrease they use their CI version, that are composed of fewer subword units.

We must also point out that the number levels in the tree could be higher, including different types of subword units (i.e. phonemes, biphones and triphones). The algorithm

could be easily modified to deal with different numbers of hierarchical levels that could be use to have different tunes between computational cost reduction and precission in the acoustic representation.

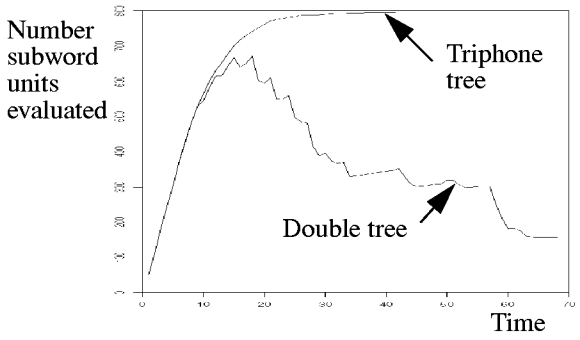

Figure 4

3.- Beam Search Strategy

As it is obvious our double tree beam seach algorithm requires the description of each word in the vocabulary through CD triphones and CI phonemes. However the recognition network does not require two parallel structures for the CD and CI representations. We only use a single tree-structure and a control strategy to evaluate the scores for the different units representing active words from their corresponding CD or CI models. This organization allow us to have the same memory requirements for our double tree than for a single triphone tree.

During recognition our beam search strategy starts with all the words activated with its triphone version. As their probability get worse they use their context independent version. At every frame the probability of the best state for each word is compared with the probability of the best state of all the words. If it is worse in more than a *changing threshold* all the states of that word are changed from triphones to context independent units. Also, the probability of every state of every word is compared with the probability of the best state and if it is worse than a pruning threshold that state is desactivated. If the *changing threshold* is greater than the pruning threshold all the words always use their triphone version. On the other hand, when the *changing threshold* is zero all the words are using their context-independent version. These are the two extreme points of performance.

Figure 5 shows an exception in the rule of changing from one version to the other. The words that are root of other words only can change from triphones to context-independent units if their pending words also have context-independent units. The figure shows an example that if the word "almeria" change from triphones to context-independent units, the word "almorzar" cannot take the probabilities from the root state, that have been desactivated. The rationality of this is the impossibility, with our control strategy, to mantain a proper propagation of initial states of those words in the tree depending on root words using CI units.

4.- Experimental results

Experimental results have been obtained for an isolated word recognition system that uses two sets of subword units presented in other works [7]. These are context independent phonemes and triphones, both of them are represented through speaker independent CHMM (Continuos Hidden Markov Models) trained using the VESTEL database [8].

Experimental results were done using 1553 files not used for training with a vocabulary of 955 Spanish words in an isolated word recognition system. Therefore the recognition grammar is the complete set of vocabulary words in paralell, and no language model is used. In such a way, the perplexity is the greatest in a medium vocabulary recognition system, and the error rates are greater rather than in continuos speech recognizer.

The experiments have been made with different levels of pruning and *changing thresholds*. The goal of our beam serach procedure is to keep the error rate as low as possible while reducing the computational requirements during the search. Therefore we measured the error rate, the average number of senones calculated for every frame and file, and the mean number of states evaluated for every frame and file. The rejection of out-of-vocabulary words has not been considered in order to have an easy comparison of the different test conditions.

Table 1 shows the results with a pruning threshold equal to 200. All the words use triphones with a changing threshold equal to 200, and all use context-independent units with *changing threshold* 0. It can be seen that as the *changing threshold* decreases the error rate is mantained until we use very low *changing thresholds*, for witch almost all the words have context-independent units.

TABLE 1.

Changing threshold	Error Rate	Average Number of States	Average Number of senones
200	9.68	2838	585
150	10.32	2920	553
100	9.55	3198	489
50	10.19	3484	404
25	12.08	3589	366
0	17.34	5181	82

The number of active states increases as the *changing threshold* decreases, this is because more words uses their context-independent units and the pruning is worse. The

acoustic representation of the vocabulary words is obviously less detailed using context-independent units rather than with triphones, and the probabilities of the different words are closer.

Finally the number of senones decreases as the *changing threshold* decreases because more context-independent units are used and they are composed of fewer senones.

Table 2 shows the results with a pruning threshold of 1000.

TABLE 2.

Changing threshold	Error Rate	Average Number of States	Average Number of senones
1000	9.55	12869	844
500	9.68	12783	841
200	9.81	11731	739
100	9.94	11004	576
50	10.00	10734	434
0	17.66	10462	84

In this case all the words use their triphone version with a *changing threshold* equal to 1000, and all use their context-independent version with *changing threshold* equal to 0.

The pruning is lower in this case. Now the error rate increases as the *changing threshold* decreases, because more context-independent units are used, and the error rate is lower with triphone units. The number of states decreases as the *changing threshold* decreases because the words share more subword units with context-independent units than with triphones. This reduction is more important as more words are in the vocabulary.

The number of senones decreases dramatically. With a changing threshold equal to 100 the error rate increases less than 0.5%, but the number of senones calculated decreases in 32%.

The double tree perform like a softer pruning, and the most important reduction is the number of senones calculated. The reduction will be more important as more triphone units are used.

5.- Conclusions

Experimental results have been obtained for a medium-size vocabulary isolated word recognition system. According to our results the proposed strategy achieves an important saving in computation cost, reducing the average number of states to update and the total number of gaussian mixtures to evaluate. This computational cost reduction is obtained only at expenses of a negligible reduction in recognition performance.

The main advantages of the proposed approach are:

- Standard pruning is combined with tree organization of the lexicon and with the use of context-independent and context-dependent acoustic units.
- Using tree based grammar reduces the computational cost at the first frames, that is just when it is highest [4], because almost all the words have similar probabilities and the pruning techniques desactivate few words.
- Using triphones the error rate is lower than using CI phonemes, but the number of subword units used grows, and therefore the computational cost. This problem is solved with the use of the double tree, using different subword units for each word. Also, using CI phonemes, more subword units are shared in the tree than in the triphone tree, so an additional computational saving is got at the beginning of the words with the double tree.

This technique allows large vocabulary recognition with high recognition rate and low computational cost.

References

[1] H. Ney, R. Haeb-Umbach, B.H. Tran, M. Order. "Improvements in Beam Search for 10000 word continues speech recognition". ICASSP 92. Volumen I. pp 9-12.

[2] I. Torres, F. Casacuberta. "Spanish Phone Recognition Using Semicontinuos Hidden Markov Models". ICASSP 93. Volumen II. pp 515-518.

[3] Fil Alleva, Xuedong Huang, Mei-Yuh Hwang. "Improvements on the Pronunciation Prefix Tree Search Organization". ICASSP 96, pp 133-136.

[4] M. Philips, D. Goddeau. "Fast Match for Segment-Based Large Vocabulary Continuous Speech Recognition". ICSLP 94. pp 1359-1362

[5] Hermann Ney. "Architecture and Search Strategies for Large-Vocabulary Continuous-Speech Recognition". NATO-ASI BUBIÓN 93.

[6] Hugo Van hamme, Filip Van Aelten. "An Adaptive-Beam Pruning Technique for Continuous Speech Recognition". ICSLP 96. pp 2083-2086.

[7] L. Villarrubia, L.H. Gómez, J.M. Elvira, J.C. Torrecilla, "Context-Dependent Units for Vocabulary-Independent Spanish Speech Recognition", ICASSP'96, pp 451-454.

[8] D. Tapias, A. Acero, J. Estevez, J.C. Torrecilla, "The VESTEL Telephone Speech Database", ICSLP'94, Japan, pp 1811-1814.