Integrating different acoustic and syntactic Language Models in a Continuous Speech Recognition System

A. Varona, I. Torres, K. López de Ipiña, L.J. Rodriguez

Dpto. Electricidad y Electrónica. Universidad del País Vasco

Apdo. 644 48080 Bilbao. SPAIN

E-mail (amparo@we.lc.ehu.es)

Abstract[1]

Continuous Speech Recognition (CSR) systems require acoustic models to represent the characteristics of the acoustic signal and Language Models (LM) to represent the syntactic constraints of the language. Both acoustic and LM probability distributions are usually independently obtained and evaluated. Then, the respective "best" models are selected to be integrated in the CSR systems. But, in this paper it was proved that the use of more accurate acoustic models (for example the use of semicontinuous models instead of discrete ones or the use of a bigger number of then representing a more complete set of sub-lexical units) didn't always mean a better performance of the integrated system because the acoustic improvements were softened when the LM probabilities were applied. This experimental evaluation was carried out over a Spanish speech application task.

1. Introduction

The final goal of a Continuous Speech Recognition (CSR) system is to obtain a sequence of linguistic units W, given a sequence of acoustic observations A. Thus, acoustic probabilities P(A/W) and Language Model (LM) probabilities P(W) must be integrated using the well-known Bayes´s formula. Acoustic and LM probability distributions are usually independently obtained and evaluated. Then, the respective "best" models (the set of acoustic models which give a better performance in a phonetic-acoustic decoder and the LM with a lower perplexity value) are chosen to be integrated in the CSR system.

In this work, the classical Hidden Markov Models (HMM) were evaluated. First, a basic set of context independent sub-lexical units was considered and a Discrete HMM was obtained for each unit. Then more accurate acoustic models were evaluated:

a) semicontinuous instead of discrete ones

b) a bigger number of then representing a more complete set of context dependent sub-lexical units.

The use of more accurate acoustic models usually gives a better performance when a phonetic-acoustic decodification is carried out but it doesn't always mean a better performance of the integrated system when the LM probabilities are also applied. In fact, in this work we showed that discrete HMM can outperform semicontinuous HMM with significant faster decoding speeds, which is also reported in other recent works [1].

Acoustic models were first evaluated over an acoustic phonetic decoding task showing better performances when more accurate models were considered. Then, the three sets of acoustic HMM were integrated and evaluated in the CSR system [2] over a task oriented speech corpus representing a set of queries to a Spanish geography database (1208 words).

The LM generation is usually based on statistical methods (N-grams). In this work, the k-Testable in the Strict Sense (k-TSS) LMs were used and evaluated with k=2..5. They are a sub-class of the regular grammars and a syntactic approach of the well-known N-grams [3].

It is well known that, to obtain the best performance of a CSR system is needed the modification of one of both P(W) or P(A/W) by introducing a balance parameter (α) [4]. This parameter seems to lessen the effects derived from the fact that, both P(A/W) and P(W) are independently obtained from speech and text training sets respectively. So that, several values of the balance parameter applied to the smoothed LM probabilities were also tested.

In Section 2, the acoustic-phonetic decoding evaluation of the three proposed sets of acoustic models was presented. Section 3, deals with the experimental evaluation of the integrated CSR system taking into account the effect of scaling the LM probabilities. Finally, some concluding remarks are presented in Section 4.

2.- Acoustic-Phonetic decoding.

The basic sub-lexical unit set consisted of 24 phone-like units. A Discrete left to right HMM with 3 looped states

[1] Work partially supported by the Spanish CICYT under grant TIC98-0423-C06-03

and 4 discrete observations per state was obtained for each unit. Emission and transition probabilities were estimated using both the Baum-Welch and Viterbi procedures, applying the Maximum Likelihood criterion.

But, to increase the accuracy of the acoustic models in the CSR system, two different ways were considered:

a) The use of a semicontinuous HMM for each phone-like unit instead of the discrete one. The joint optimization of the codebook together with the parameters of the model is one of the main advantages of this kind of models [5].

b) To increase the number of sub-lexical units keeping the use of discrete HMM. The well known technique of decision tree clustering [6] was applied to obtain an optimal set of 101 trainable, discriminative and generalized context dependent units. An additional set of border units was specifically trained to generate lexical baseforms, covering all possible intraword context and being context independent to the outside [7].

Acoustic models were first evaluated over an acoustic phonetic decoding task with a balanced phonetic corpus both for training and testing purposes. The training corpus was composed of 1529 sentences, involving around 60000 phones. The -speaker independent- test corpus was composed of 700 sentences and around 33500 phones.

In this experiments, the phone recognition rates (%REC) was measured as %REC=c/(i+s+d+c)*100, where c accounts for the number of correct recognitions, and i, s and d were the number of insertions, substitutions and deletions, respectively. Table 1 shows the obtained results when discrete and semicontinuous HMM were evaluated over the set of 24 phone-like units. Table 2 shows the obtained results when the 24 context independent and the 101 context dependent sets of units were evaluated using discrete HMM.

Table 1. Phone recognition rates for a Spanish acoustic-phonetic decoding task, using discrete and semicontinuos HMM (24 phone-like units were represented)

Kind of model	% REC
discrete	63.82
semicontinuous	65.07

As it was expected, the use of more accurate models gave higher phone recognition rates. In fact, the use of a bigger number of discrete HMM representing context dependent sub-lexical units seems to be even more efficient than the use of semicontinuous HMM over the 24 context independent units.

Table 2. Phone recognition rates for a Spanish acoustic-phonetic decoding task, using phone-like units and context dependent units (discrete HMM were used).

Type of unit	# units	% REC
phone-like units	24	63.82
context dependent	101	66.44

In phonetic acoustic decodification , it was obvious that the use of more accurate models must be taking into account. In next section we could see if this behavior is kept when the evaluation is over the complete CSR system.

3.- The CSR System

Previously evaluated acoustic models were integrated with the k-TSS LM in the CSR system. In this section, first the syntactic LM is presented and then the experimental evaluation carried out.

3.1- The syntactic Language Model.

A syntactic approach of the well-known N-grams models, the k-Testable Language in the Strict Sense (k-TSS) has been used in this work to be integrated in a CSR system. The use of k-TSS regular grammars [8] allowed to obtain a deterministic Stochastic Finite State Automaton (SFSA) integrating K k-TSS models (with k=1, 2..K) into a self-contained model [3]. In such a model, each state of the automaton represents a string of words $w_{i-k}w_{i-(k-1)}...w_{i-1}$, k = 1...K-1, with a maximum length of K-1, where i stands for a generic index in any string $w_1...w_i...$ appearing in the training corpus. Such a state is labeled as w_{i-k}^{i-1}. Each transition represents a k-gram, k = 1...K; it is labeled by its last word w_i and connects two states labeled up to with K-1 words. As an example, transitions corresponding to strings of words of length K connecting states associated to string lengths K-1 are defined as:

$$\delta^K\left(w_{i-(K-1)}^{i-1}, w_i\right) = \left(w_{i-(K-1)+1}^{i}, P(w_i / w_{i-(K-1)}^{i-1})\right) \qquad (1)$$

The probability to be associated to each transition $\delta^K\left(w_{i-(K-1)}^{i-1}, w_i\right)$ can be estimated under a maximum likelihood criterion as:

$$P_{ML}(w_i / w_{i-(K-1)}^{i-1}) = \frac{N(w_i / w_{i-(K-1)}^{i-1})}{\sum_{\forall w_j \in \Sigma} N(w_j / w_{i-(K-1)}^{i-1})} \qquad (2)$$

where Σ is the vocabulary, that is, the set of words appearing in the training corpus, $N(w_j / w_{i-(K-1)}^{i-1})$ is the

number of times the word w_j appears at the end of the K-gram $w_{i-(K-1)}...w_{i-1}w_j$, that is the count associated to the transition labeled by w_j coming from state labeled as $w_{i-(K-1)}^{i-1}$.

The whole and detailed definition of the Automaton, i.e. initial and final states, unigram representation, etc., can be found in [3]. As an example, Figure 1 represents the K-grams $w_{i-(K-1)}^{i-1}$ and $w_{i-(K-1)+1}^{i}$ labeling two states of the automaton. When w_i is observed an outgoing transition from the first to the second state is set and labeled by w_i.

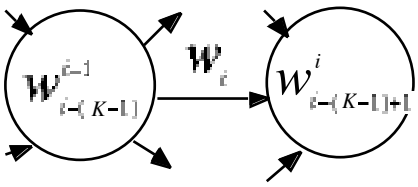

Figure 1: Two states of the K-TSS automaton labeled by K-grams $w_{i-(K-1)}w_{i-(K-1)+1}...w_{(i-1)}$ and $w_{i-(K-1)+1}...w_i$ labeling two states of the automaton. Transitions are labeled by words appearing in the training sample after K-grams labeling the source state.

The probability associated to each transition representing *seen events* can be estimated under a maximum likelihood criterion (see Equation 2). However, a probability need also to be associated to those events not represented in the training corpus, i.e., *unseen events*. To deal with this problem, some probability mass should be discounted from observed events and then redistributed over *unseen* ones using some smoothing procedure. In previous works [9] Backing-off smoothing technique was chosen because the involved recursive scheme has been well integrated in the finite state formalism. Within the backing-off formalism the Witten-Bell discounting was applied [10].

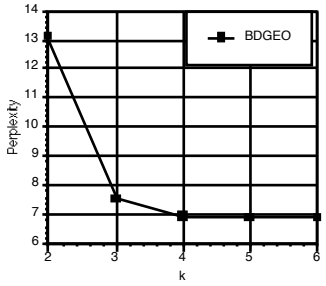

Figure 2. - Perplexity values obtained by several k-TSS LM after applying Witten-Bell discounting.

For the evaluation of the k-TSS LM, a task-oriented Spanish corpus [11], consisting of 82.000 words and a vocabulary of 1.208 words, was used. This corpus represents a set of queries to a Spanish geography database. The training corpus for the k-TSS language models consisted of 9150 sentences. For testing purposes, an independent but fully covered text containing 200 sentences was used. Figure 2 shows the obtained perplexity values. The values of the perplexity are almost constant for values of k higher than 3. This behavior is managed thank to the use of the back-off smoothing procedure.

3.2.- Experimental evaluation

The three sets of acoustic models were evaluated over a set of k-TSS (k=2..5) language models integrated in a CSR system [2]. This evaluation was carried out in terns of Word Error Rates (%WER) which was also calculated as %WER=(i+s+d)/(c+i+s+d) where c accounts for the number of correct recognized words, and i, s and d were the number of words inserted, substituted and deleted, respectively.

Each transition of the k-TSS automaton was replaced by a chain of Hidden Markov models, representing the acoustic model of each phonetic unit of the word. The time-synchronous Viterbi algorithm, along with a beam-search procedure reducing the involved computational cost, was used to decode uttered sentences. The beam-search factor (bf) was optimized in previous experiments [3]. Thus, the same fixed value, bf=0.7, was used for all the experiments in this work. LM probabilities were modified by introducing a balance parameter α [4] in the Bayes's rule: $P(w)^{\alpha}$ in order to obtain the maximum performance of the CSR system. The experiments were carried out by a Silicon Graphics O_2 with a R10000 processor.

For testing purposes, previously presented Spanish corpus was used. The 200 sentences were uttered by 12 speakers resulting in a total of 600 sentences and 5655 words.

Figure 3 showed the obtained results when discrete and semicontinuous HMM were integrated with several smoothed language models (k=2, 3, 4 and 5) and different values of the balance parameter around the optimum (α=4, 5, 6 and 7) were considered . The average number of active nodes in the trellis (acoustic and LM states) needed by every LM to decode a sentence is also represented.

Points at the bottom left corner of each plot represent the best system performance: the lowest %WER and the lowest average number of active nodes in the lattice. For any k-TSS model, an important increase in recognition rates (up to a maximum) along with a notable decrease in the average active nodes in the lattice can be observed (Figure 3) when the balance parameter α was increased.

When k=2 the use of semicontinuous models helped to obtain a better system performance but with higher values of k the differences around the optimum were not significative.

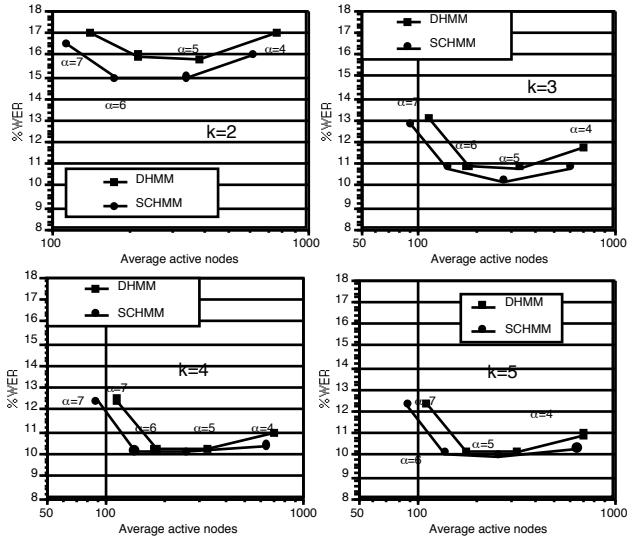

Figure 3. - %WER obtained by the Smoothed *k*-TSS LMs using Discrete HMM (DHMM) and semicontinuos (SCHMM) with different values of the α parameter.

Practically the same results can be found when discrete HMM of contextual sub-lexical units were considered (see Figure 4).

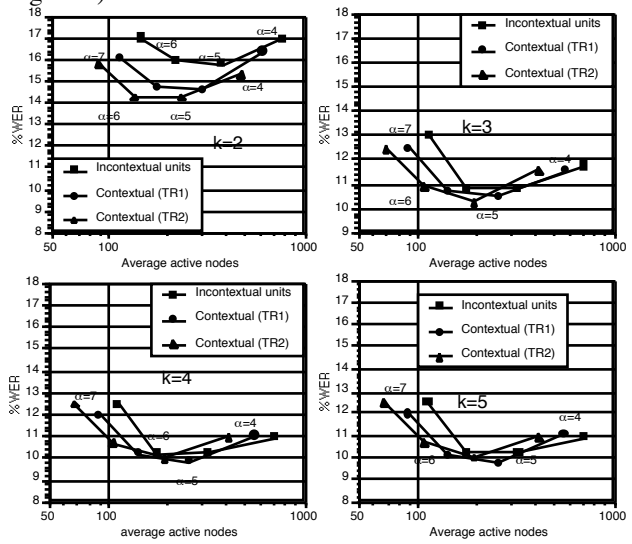

Figure 4. - %WER obtained by the Smoothed *k*-TSS LMs using discrete HMM of incontextual (24 units) and contextual (101 units) with different values of the α parameter.

The construction of linear models presented a problem when between-words contexts had to be considered because the outer context were not known. Two different procedures were used to obtain lexical baseforms, both considering words as isolated. In the first one, called TR1, border units were selected to be context independent both sides. In the second one, called TR2, border units were selected to be context dependent in the word side and context independent in the outside [7].

Figure 4 showed that the use of contextual units, especially TR2, helped to the recognition with k=2 and k=3, but there were not differences around the optimum with higher values of k.

4.- Concluding remarks.

Acoustic and LM probabilities are usually independently obtained and evaluated. Then, the respective best models are chosen to be integrated in the CSR systems. But, in this paper it was proved that the use of more accurate acoustic models doesn't always mean a better performance of the integrated system because the acoustic improvements usually are softened when the LM probabilities are applied.

5.- References

[1] Digalakis, V. Tsakalidis, S. Neumeyer, L. (1999). "Reviving discrete HMMS. The myth about the superiority of continuous HMMs". *Procc of EUROSPEECH´99*. Vol 6. pp 2463-2466.

[2] Rodriguez, L.J. Torres, I. Alcaide, J.M. Varona, A. López de Ipiña, K. Peñagarikano. M. Bordel, G. (1999). "An Integrated System for Spanish CSR Tasks". Procc of EUROSPEECH.99. Vol 2. pp 951-954

[3] Varona A. and Torres I. (1999) "Using Smoothed K-TSS Language Models in Continuous Speech Recognition". Procc. IEEE Int. Conf. Acoust, Speech, Signal Processing. Vol II pp. 729-732.

[4] Jelinek, F. (1996): "Five speculations (and a divertimento) on the themes of H. Bourlard, H. Hermansky and N. Morgan". Speech Communication 18, pp 242-246.

[5] Torres, I. Casacuberta, F. (1993): "Spanish phone recognition using semicontinuous Hidden Markov Models". Procc. IEEE ICASSP'93. pp. 515-518

[6] Bahl, L.R. Souza, V.P. Gopalakrishnan, P.S. Nahamoo,D. Picheny M.A. (1994) "Decision Trees for Phonological Rules in Continuous Speech Recognition". *Procc IEEE ICASSP´94*, pp 533-536

[7] López de Ipiña, K. Varona, A. Torres, I. Rodriguez, L.J. (1999)" Decision Trees for Inter-Word Context Dependencies in Spanish Continuous Speech Recognition Tasks". *Procc of EUROSPEECH, 99.* pp 899-902

[8] García, P. and Vidal, E. (1990): "Inference of k-testable languages in the strict sense and application to syntactic pattern recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 12, nº 9, pp. 920-925.

[9] Bordel, G. Torres, I. and Vidal, E. (1994): "Back-off smoothing in a syntactic approach to Language Modeling". Procc. of ICSLP-94, pp. 851-854.

[10] Clarkson, P. Rosenfeld, D. "Statistical language modeling using the CMU-CAMBRIDGE toolkit", (1997) Proceedings of EUROSPEECH 97 pp- 2707-2710.

[11] Diaz, J.E. Rubio, A.J. Peinado,. M. Segarra, E. Prieto N, and Casacuberta, F. (1993); "Development of Task Oriented Spanish Speech Corpora," Proceedings of EUROSPEECH 93