Bounded discounting: a new smoothing technique for k-TSS language models.

A. Varona, I. Torres

Dpto. Electricidad y Electrónica. Universidad del País Vasco

Apdo. 644 48080 Bilbao. SPAIN

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

Abstract[1]

Continuous Speech Recognition systems require a Language Model (LM) to represent the syntactic constraints of the language. A sub-class of the regular languages, the k Testable in the Strict Sense (k-TSS) languages, has been used in this work to generate LMs. A smoothing technique needs to be applied when using a Language Model in a Continuous Speech Recognition (CSR) System to also consider events not represented in the training corpus. Two syntactic backing off smoothing approaches, the well-known Witten-Bell discounting and a new proposal, the Bounded discounting, were applied to several k-TSS language models (LM). The Bounded discounting deals with the Turing discounting problems while keeping the Katz' smoothing schema. The experimental evaluation carried out over a Spanish speech application task showed different LM behavior when perplexity and CSR system performance were used as evaluation measure. However, this behavior changed when the LM probabilities were modified by introducing a balance parameter in the Baye's rule. There is a strong dependence between the amount of probability reserved by the smoothing technique to be assigned to *unseen* events and the value of the balance parameter needed to get the best system performance.

1. Introduction

Continuous Speech Recognition (CSR) systems transcribe speech signals into sequences of linguistic units $\Omega \equiv w_1 w_2 .. w_{|\Omega|}$, usually words, from some finite vocabulary $\Sigma = \{w_j\}$, $j = 1 \ldots |\Sigma|$ previously established. CSR systems require a Language Model (LM) to integrate the syntactic and/or semantic constraints of the language. The generation of LM is a classical pattern recognition problem where both statistical (typically N-grams) and syntactic approaches have been extensively used. A syntactic approach based on regular grammars, the k-Testable in the Strict Sense (k-TSS) languages [1] has also been proposed in previous works [2] to generate LM. They are a subclass of regular languages and can be considered as a syntactic approach of the well-known N-grams models.

A major problem to be solved when using a Language Model (LM) is the estimation of the probabilities to be assigned to those events not represented in the training corpus, that is, *unseen* events. Thus a smoothing technique need to be applied when integrating a LM in a CSR system. In previous works [3] a syntactic backing-off smoothing was proposed and evaluated using k-TSS language models. The recursive schema required by the smoothing procedure has been well integrated in the finite state formalism and, thus, an efficient implementation of the backing-off mechanism is achieved [4].

Smoothing techniques are based on a discounting-distribution schema: a mass of probability needs to be

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

discounted from *seen* events to further be assigned to *unseen* events. In previous works, a Witten-Bell based discounting procedure was proposed and evaluated [3] [4]. In this approach the discounting factor was applied to the whole set of *seen* events in the training corpus [5]. As a consequence, the mass of probability to be assigned to *unseen* events could be overestimated, as it is shown in this work. Thus a new proposal, based on the well-known Turing discounting [6], the Bounded discounting, was developed and presented in this work. In this case, discounting factors were only applied to those events scarcely observed in the training corpus.

The must reliable way to evaluate the real performance of a LM is to measure the obtained Word Error Rates (%WER) after integrating it into a CSR system. However, the most common way to assess the goodness of different LMs is the evaluation of the test set perplexity, even if the relationship with the acoustic models is not considered. So that, some important points related to a LM generation (like smoothing techniques), could not be always well assessed by the perplexity [7]. Thus, in this work both, the test set perplexity and %WER, were considered and compared to evaluate the behavior of two smoothing approaches: Witten-Bell and Bounded discounting, developed under the *k*-TSS language modeling formalism. After this evaluation, the ability of the test set perplexity to predict the real behavior of a smoothing technique when working in a CSR system was questioned [7].

CSR systems are invariably based on the well-known Bayes' rule. Bayes' rule maximizes the product of the probability of a sequence of acoustic observations A given a sequence of words Ω, $P(A/\Omega)$, and the probability that the word sequence Ω will be uttered, $P(\Omega)$. However it is well known that the best performance of a CSR system is obtained when $\mathrm{P}(\Omega)$ is modified by introducing a balance parameter α [8] in the following way: $\mathrm{P}(\Omega)^{\alpha}$. The effect of scaling LM probabilities was finally evaluated in this work showing a strong relationship between the behavior of the smoothing technique and the value of the scaling factor required to obtain the best CSR system performance.

In Section 2, the *k*-TSS language model formalism is briefly described. Both smoothing approaches, Witten-Bell and Bounded discounting, are also presented in this Section. Section 3 deals with the experimental evaluation of several smoothed *k*-TSS LMs in terms of both, perplexity and %WER, along with the effect of scaling the LM probabilities. These experiments were carried out over a Spanish speech application task (1,208 words). Finally, some concluding remarks are presented in Section 4.

2. Smoothing *k*-TSS Language Models

The *k*-Testable in the Strict Sense (*k*-TSS) LM are a sub-class of the regular languages and can be considered as a syntactic approach of the well known *N*-grams models. The use of regular languages allowed us to obtain a deterministic *k*-TSS stochastic finite state automaton. In such a model, each state of the automaton represents a string of words $w_{i-(k-1)}w_{i-(k-1)}\dots w_{i-1} \equiv w^{i-1}_{i-(k-1)}$, where i stands for a generic index in any string $w_1 \dots w_i \dots$ appearing in the training corpus. Each transition represents a *k*-gram, it is labeled by its last word w_i and connects two states labeled up to with *k*-1 words. It is defined as:

$$\delta^k\left(w^{i-1}_{i-(k-1)}, w_i\right) = \left(w^{i}_{i-(k-1)+1}, P(w_i / w^{i-1}_{i-(k-1)})\right) \tag{1}$$

Figure 1 represents strings $w_{i-(k-1)}w_{i-(k-1)+1}...w_{(i-1)}$ and $w_{i-(k-1)+1}...w_i$ labeling two states of the automaton. When w_i is observed an outgoing transition from the first to the second state is set and labeled by w_i.

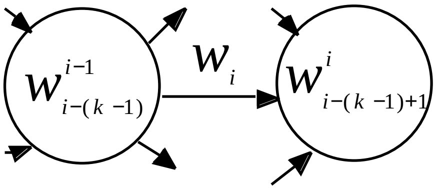

Figure 1: Two states of the k-TSS automaton labeled by k-grams wi-(k-1)wi-(k-1)+1...w(i-1) and wi-(k-1)+1...wi labeling two states of the automaton. Transitions are labeled by words appearing in the training sample after k-grams labeling the source state.

The probability associated to each transition representing *seen events* can be estimated under a maximum likelihood criterion. 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. Backing-off smoothing was chosen in previous works [3] because the involved recursive scheme has been well integrated in the finite state formalism [4]. The syntactic approach suggested a state-dependent estimation of the total discount and, consequently, the symmetry principle was locally applied [3]. Thus the modified probability $P(w/q)$ to be associated to a transition $\delta^k(q, w) = (q', P(w/q))$ is estimated according to:

$$P(w/q) = \begin{cases} [1-\lambda]\dfrac{N(w/q)}{N(q)} & w \quad \Sigma_q \\ \left(\displaystyle\sum_{w_i \quad \Sigma_q} \lambda \frac{N(w_i/q)}{N(q)}\right) \dfrac{P(w/b_q)}{1-\displaystyle\sum_{w_i \quad \Sigma_q} P(w_i/b_q)} & w \quad \Sigma - \Sigma_q \end{cases} \tag{2}$$

where: Σ_q is the vocabulary associated to state q and consists of the set of words appearing after the string labeling state q in the training corpus, i.e. words labeling the set of *seen* outgoing transitions from state q; $N(\mathrm{w}/q)$ is the number of times that word w_i appears after the string labeling state q; $N(q) = \sum_{w \quad \Sigma_q} P(w/q)$, and $P(w/b_q)$ is the estimated probability associated to the same event in the $(k$-1)-TSS model. A Witten-Bell discounting [5] was chosen in previous works [1]. Thus, $(1-\lambda)$ depends on $|\Sigma_q|$, the number of different events following the particular context in a state q:

$$1-\lambda = \frac{N(q)}{N(q)+|\Sigma_q|} \tag{3}$$

This proposal had been experimentally compared to other classical back-off methods leading to a significant decrease in test-set perplexity [3]. However, in Equation (2) the discounting factor was applied to the whole set of *seen* events, i e, $w \quad \Sigma_q$. As a consequence, the mass of probability to be assigned to *unseen* events could be overestimated when using a smoothed k-TSS in a CSR system. So that, in this paper a new proposal, the bounded discounting was proposed. Bounded discounting keeps the Turing discounting [6] philosophy discounting a mass of probability only from scarcely *seen* events.

Bounded discounting (Bd).

The scheme devised by Katz [6] combines Turing discounting with backing-off. According to this formalism the probability associated to events occurring more than a fixed number of times, say r times, are estimated under a maximum likelihood criterion whereas events occurring less than r times, $N(w_i/q)<r$, are discounted a certain mass of probability. Thus:

$$P(w/q)=\begin{cases}\dfrac{N(w/q)}{N(q)} & w\in\Sigma_q \quad N(w/q)>r\\[2ex] [1-\lambda]\dfrac{N(w/q)}{N(q)} & w\in\Sigma_q \quad 1\le N(w/q)\le r\\[2ex] \displaystyle\sum_{\substack{w_i\in\Sigma_q\\ 1\le N(w_i/q)\le r}}\left[\lambda\frac{N(w_i/q)}{N(q)}\right]\frac{P(w/b_q)}{1-\displaystyle\sum_{w_i\in\Sigma_q}P(w_i/b_q)} & w\in\left(\Sigma-\Sigma_q\right)\end{cases}\qquad(4)$$

In Turing discounting the discounted mass of probability depends on $n_1,\ n_2,\ \ldots,n_{r+1}$, (being n_i the number of events which occur i times). The lower the count $N(w/q)$ is, the bigger discounting is applied, because higher counts are supposed to be better estimated. This approach puts some constrains to the relative values of $n_1,\ n_2,\ \ldots,n_{r+1}$, which are not always satisfied by k-grams models with medium and high values of k, due to the lack of an adequate distribution of the samples.

To avoid the Katz discounting problems, we proposed the Bounded discounting. As in the Katz model, the discounting operation was limited to low counts, i.e., $N(w/q)\le r$ in the following way:

$$1-\lambda=d-\tau(r-N(w/q))\qquad \tau,d<1\quad \tau<<<d\qquad(5)$$

Discounting depends on d and τ parameters' values, which must be minor than one. The bigger the count was ($N(w/q)\le r$) the lower discounting was applied. When $N(w/q)=r$, the discounting was the minimum (only depends on d parameter), and when $N(w/q)=1$, the discounting applied was the maximum (d-τ(r-1)).

Another problem to be addressed when using Katz' discounting is that additional checks are required for those states for which all the events are *seen* more then r times. The remedy used in the CMU toolkit [5] to solve this problem is to increase the count at eat state $N(q)$ by one using the gained probability mass $1/(N(q)+1)$ to be redistributed over *unseen* events. However, the discount is applied to all the *seen* events' probabilities, which does not agree with Katz´s discounting philosophy. In our proposal, only the minimum counts were decremented (discounting only depends on d parameter) in the following way:

$$P(w/q)=\begin{cases}\dfrac{N(w/q)}{N(q)} & w\in\Sigma_q\quad N(w/q)>min(N(w/q))\\[2ex] d\dfrac{N(w/q)}{N(q)} & w\in\Sigma_q\quad N(w/q)=min(N(w/q))\\[2ex] \left(\displaystyle\sum_{\substack{w_i\in\Sigma_q\\ 1\le N(w_i/q)\le r}}\left[[1-d]\frac{N(w_i/q)}{N(q)}\right]\right)\dfrac{P(w_i/b_q)}{1-\displaystyle\sum_{w_i\in\Sigma_q}P(w_i/b_q)} & w\in\left(\Sigma-\Sigma_q\right)\end{cases}\qquad(6)$$

As a consequence, events occurring a high number of times are estimated under a maximum likelihood criterion as in Katz proposal [6].

3.- Experimental evaluation.

Both backing-off smoothing schema, Witten-Bell and Bounded discounting were evaluated over a set of *k*-TSS language models integrated in a CSR system [9]. This evaluation was carried out in terms of both, the test set perplexity and the Word Error Rates (%WER) obtained by the CSR system.

The test set Perplexity (PP) is based on the mean log probability that a LM assigns to some test set w_1^L of size *L*. So that, it is based exclusively on the probability of words which actually occur in the test as follows.

$$PP = P(w_1^L)^{-1/L} = e^{-\frac{1}{L}\sum_{i=1}^{L} \log(P(w_i / w_1^{i-1}))} \qquad (7)$$

The test set perplexity measures the branching factor associated to a task, which depends on the number of different words in the text. Low perplexity values are obtained when high probabilities are assigned to the test set events by the LM being evaluated.

For these experiments a task-oriented Spanish speech corpus [10], consisting in 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 used to obtain the *k*-TSS models, consisted in 9150 sentences. The text test set consisted in 200 different sentences. These sentences were then uttered by 12 speakers resulting in a total of 600 sentences that composed the speech test set. Uttered sentences were decoded by the time-synchronous Viterbi algorithm with a fixed beam-search to reduce the computational cost. A chain of Hidden Markov models representing the acoustic model of the word phonetic chain replaced each transition of the k-TSS automaton.

In a first series of experiments, both Witten-Bell (WBd) and Bounded (Bd) discounting methods were applied to several smoothed *k*-TSS language models, with *k*=2,...,6. Different values of the *d* parameter in Equation 5 were tested (*d*=0.99, 0.80, 0.70 and 0.60) while keeping fixed values for parameter τ (τ=0.01). The minimum number of times *r* required for a maximum likelihood estimation of event probabilities (Equation 4) was also set to *r*≈7. Figure 2 shows the results of the first series of experiments in terms of the test set perplexity (PP).

Figure 2 shows that the more mass of probability (<<d) was assigned to the unseen events by the Bd procedure (up to a maximum *d*=0.70), the better value of perplexity was observed. WBd obtained the lower perplexity values for all *k*-TSS models (see Figure 2b)). Thus, the best perplexity results were obtained when the higher mass of probability was reserved to be applied to *unseen* events, even if the differences were not very meaningful.

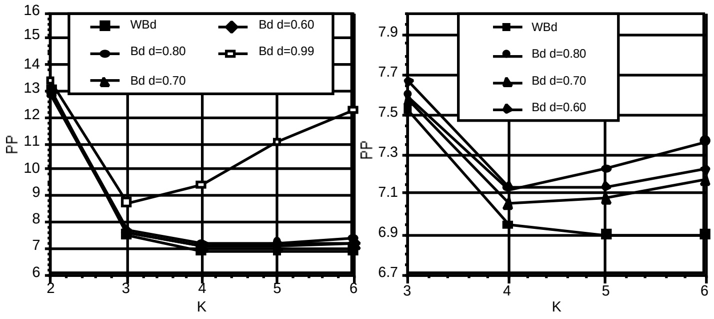

Figure 2. - a) PP obtained by several Smoothed k-TSS LM using Witten-Bell (WBd) and Bounded (Bd) discounting methods. b) A detail of Figure 2a for PP values around 7.

Figure 3 shows the %WER obtained by the same smoothed LM in Figure 2, when integrated in the CSR system. The average number of active nodes in the trellis (acoustic and LM states) needed by every LM to decode a sentence is also represented in Figure 3.

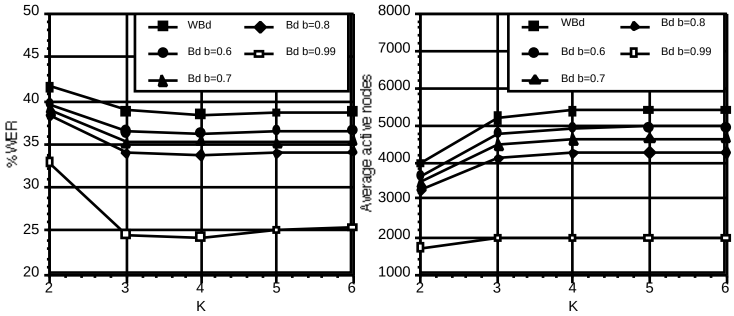

Figure 3.- a) %WER obtained by the Smoothed *k*-TSS LMs in Figure 2. b) Average number of active nodes at decoding time.

Figure 3 shows that the more mass of probability (<<d) was assigned to *unseen* events by the smoothing procedure, the worse (higher) %WER was observed and a high average active nodes was needed. Moreover, WBd achieved the worst system performance. Thus, the best system performance in terms of both, the %WER and average number of active nodes in the lattice were obtained when the lower mass of probability was reserved to be applied to *unseen* events. These results strongly contrast with the perplexity behavior shown in Figure 2. As a consequence, can a perplexity-based evaluation predict the real behavior of smoothed language models when integrated in CSR systems? [7].

In a second series of experiments the LM probabilities were modified by introducing a balance parameter α [8] in the Bayes's rule: $P(w)^{\alpha}$. Figure 4 showed the results obtained by several smoothed language models (*k*=2, 3, 4 and 5) used in the first series of experiments (Figures 2 and 3) when different values of the balance parameter (α=4, 5 and 6) were considered. Points at the bottom left corner of each plot represent the best 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 4) when the balance parameter α was increased.

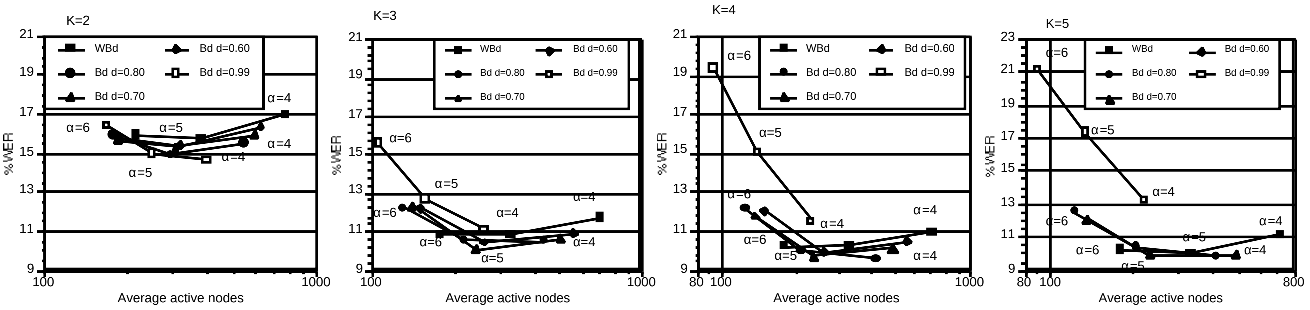

Figure 4. - %WER obtained by the Smoothed *k*-TSS LMs in Figure 2 and 3 using different values of the α parameter.

The use of a balance factor α can be understood as a new smoothing (redistribution) of the LM probabilities. As a consequence, there is a strong dependence between the smoothing technique and the value of the scaling parameter α needed to get the best performance of the system (which is in many cases perplexity independent). At the end of each word at the Viterbi trellis there is an accumulated probability. The gap between these accumulated probabilities is usually bigger than the gap between the LM probabilities due to the acoustic probabilities (which values are smaller and are applied much more times in the Viterbi trellis). The immediate consequence is that the values of LM probabilities are irrelevant in most part of the situations to decide the best way to follow. However, when the LM probabilities are raised to a power α: $(P(w))^{\alpha}$, all of them are attenuated, but this attenuation is higher for lower probability values. So that, the gap between the high and low probabilities is also bigger and then the LM probabilities are more and more competitive with the increase of α values, up to a maximum where LM probabilities are overvalued.

On the other hand each smoothing technique reserves a different mass of probability to be assigned to *unseen* events. When this mass of probability is higher (WBd), the probability assigned to seen events is lower (flat probability distribution). In such a case, a bigger value of the parameter α was needed to get the best performance (see Figure 4).

The best performance was reached by WBd with α=6 and Bd (*d*=0.60, 0.70 and 0.80) with α=5. The differences among the %WER values around the optimum are not significant. However, the system performance decreased for Bd with *d*=0.99, mainly for high values of *k*. This model also obtained the higher PP values (Figure 2). Thus, when significant differences in perplexity values were found in perplexity results, the behavior of LM´s models in a CSR system could be predicted. However, when differences are small, perplexity values have not relevance to predict the optimum performance of the CSR system, which could depend on an adequate selection of the balance parameter.

4.-Concluding remarks.

Two syntactic backing-off smoothing approaches, the well-known Witten-Bell discounting and the proposed Bounded discounting, have been tested and evaluated using *k*-TSS language models. The Bounded discounting deals with the Turing discounting problems while keeping the Katz' smoothing schema. So that, only probability assigned to scarcely observed events has been modified and, as a consequence, the overestimation of the probabilities to be assigned to *unseen* events found in the Witten-Bell discounting has been reduced.

These approaches has been evaluated using both, the test set perplexity and the CSR system performance (%WER). Both evaluation measures did not agree in this case since the best system performance was obtained when the highest perplexity smoothing technique was used.

However, this behavior changed when the LM probabilities were modified by introducing a balance parameter in the Bayes´ rule. These experiments showed that small perplexity differences could not predict the behavior of these LMs in the CSR system. In fact, the optimum performance was achieved when the adequate α valued was applied, whatever the value of perplexity was.

Finally, there is a strong dependence between the amount of probability reserved by the smoothing technique to be assigned to *unseen* events and the value of the balance parameter α needed to get the best system performance.

6. References

[1] P. García, and E. Vidal, (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.

[2] G. Bordel, A.Varona, and I.Torres (1997): "*k*-TLSS(S) Language Models for Speech Recognition". *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* vol 2, pp 819-822.

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

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

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

[6] S. M.Katz. (1987). "Estimation of Probabilities from Sparce Data for The Language Model Component of a Speech Recognizer". *IEEE Trans. on Acoustics, Speech and Signal Processing,.* vol. ASSP-35, n 3, pp. 400-401.

[7] P. Clarkson, T. Robinson (1999). "Towards improved language model evaluation measures". *Procc of EUROSPEECH.99.* Vol 5. pp 1927-1930

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

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

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