# MULTI-SPAN STATISTICAL LANGUAGE MODELING FOR LARGE VOCABULARY SPEECH RECOGNITION

*Jerome R. Bellegarda*

Spoken Language Group
Apple Computer, Inc.
Cupertino, California 95014, USA

## ABSTRACT

The goal of multi-span language modeling is to integrate the various constraints, both local and global, that are present in the language. In this paper, local constraints are captured via the usual $n$-gram approach, while global constraints are taken into account through the use of latent semantic analysis. An integrative formulation is derived for the combination of these two paradigms, resulting in an entirely data-driven, multi-span framework for large vocabulary speech recognition. Because of the inherent complementarity in the two types of constraints, the performance of the integrated language model compares favorably with the corresponding $n$-gram performance. Both perplexity and average word error rate figures are reported and discussed.

## 1. INTRODUCTION

Stochastic language modeling plays a central role in large vocabulary speech recognition, where it is typically used to constrain the acoustic analysis, guide the search through various (partial) text hypotheses, and/or contribute to the determination of the final transcription. Over the past decade, the $n$-gram paradigm [1] has emerged as the formalism of choice for a wide range of domains. Concerns regarding parameter reliability, however, restrict current implementations to low values of $n$ (cf., e.g., [2]), which in turn imposes an artificially local horizon to the language model. As a result, $n$-grams are inherently unable to capture large-span relationships in the language.

Taking such global constraints into account has traditionally involved a paradigm shift toward parsing and rule-based grammars, such as are routinely and successfully employed in small vocabulary recognition applications. This approach solves the locality problem, since it typically operates at the level of an entire sentence. Unfortunately, it is not (yet) practical for large vocabulary recognition. This has motivated further investigation into alternative ways to extract suitable long distance information, other than resorting to a formal parsing mechanism.

One such attempt was based on the concept of word triggers [3]. Unfortunately, trigger pair selection is a complex issue: different pairs display markedly different behavior, which limits the potential of low frequency triggers [4]. Still, self-triggers seem to be particularly powerful and robust [3], which underscores the desirability of exploiting correlations between the current word and features of the document history.

This observation led the author to explore the use of latent semantic analysis (LSA) for such purpose [5] – [7]. In some respect, the LSA paradigm can be viewed as an extension of the word trigger concept, where a more systematic framework is used to handle the trigger pair selection. In [5], LSA was used for word clustering, and in [6], for language modeling. In both cases, it was found to be suitable to capture some of the global constraints present in the language. In fact, multi-span language models, constructed by embedding LSA into the standard $n$-gram formulation, were shown to result in a substantial reduction in perplexity [7].

In this paper, we are primarily interested in the behavior of such multi-span language modeling in actual recognition. The paper is organized as follows. In the next section, we review the salient properties of LSA-based statistical language modeling. In Section 3, we discuss the integration of this framework with conventional $n$-gram modeling. Section 4 addresses some of the implementation issues involved in using the resulting multi-span models for large vocabulary recognition. Finally, Section 5 illustrates some of the benefits associated with multi-span modeling on a subset of the Wall Street Journal task.

## 2. LSA LANGUAGE MODELING

Let $\mathcal{V}$, $|\mathcal{V}| = M$, be some vocabulary of interest and $\mathcal{T}$ a training text corpus, comprising $N$ articles (documents) from a variety of sources. (Note that this implies that the training data is tagged at the document level, i.e., there is a way to identify article boundaries. This is the case, for example, with the ARPA North American Business (NAB) News corpus [8].) Typically, $M$ and $N$ are on the order of ten thousand and hundred thousand, respectively; $\mathcal{T}$ might comprise a hundred million words or so.

The LSA approach defines a mapping between the sets $\mathcal{V}$, $\mathcal{T}$ and a vector space $\mathcal{S}$, whereby each word $w_i$ in $\mathcal{V}$ is represented by a vector $u_i$ in $\mathcal{S}$ and each document $d_j$

in $\mathcal{T}$ is represented by a vector $v_j$ in $\mathcal{S}$. For the sake of brevity, we refer the reader to [9] for further details on the mechanics of LSA and LSA language modeling, and just briefly summarize here.

The first step is the construction of a matrix ($W$) of co-occurences between words and documents. In marked contrast with $n$-gram modeling, word order is ignored: the matrix $W$ is accumulated from the available training data by simply keeping track of which word is found in what document. Among other possibilities, a suitable expression for the $(i,j)^{\text{th}}$ element of $W$ is given by (cf. [5]):

$$w_{i,j} = g_i \frac{c_{i,j}}{n_j}\,, \tag{1}$$

where $g_i$ is the normalized entropy complement of $w_i$ in the corpus $\mathcal{T}$, $c_{i,j}$ is the number of times $w_i$ occurs in document $d_j$, and $n_j$ is the total number of words present in document $d_j$.

The second step is to compute the singular value decomposition (SVD) of $W$ as:

$$W \approx \hat{W} = U\,S\,V^{T}\,, \tag{2}$$

where $U$ is the $(M \times R)$ matrix of left singular vectors $u_i$ $(1 \le i \le M)$, $S$ is the $(R \times R)$ diagonal matrix of singular values, $V$ is the $(N \times R)$ matrix of right singular vectors $v_j$ $(1 \le j \le N)$, $R \ll M (\ll N)$ is the order of the decomposition, and $^{T}$ denotes matrix transposition. The left singular vectors represent the words in the given vocabulary, and the right singular vectors represent the documents in the given corpus. Thus, the space $\mathcal{S}$ sought is the one spanned by $U$ and $V$. An important property of this space is that two words whose representations are "close" (in some suitable metric) tend to appear in the same kind of documents, whether or not they actually occur within identical word contexts in those documents. Conversely, two documents whose representations are "close" tend to convey the same semantic meaning, whether or not they contain the same word constructs. Thus, we can expect that the respective representations of words and documents that are semantically linked would also be "close" in the LSA space $\mathcal{S}$.

The third step is to leverage this property for language modeling purposes. Let $w_q$ denote the word about to be predicted, and $H_{q-1}$ the admissible history (context) for this particular word, i.e., the current document up to word $w_{q-1}$, denoted by $\tilde{d}_{q-1}$. Then the associated LSA language model probability is given by:

$$\Pr\left(w_q|H_{q-1},\mathcal{S}\right) = \Pr\left(w_q|\tilde{d}_{q-1}\right), \tag{3}$$

where the conditioning on $\mathcal{S}$ reflects the fact that the probability depends on the particular vector space arising from the SVD representation, and $\tilde{d}_{q-1}$ has a representation in the space $\mathcal{S}$ given by:

$$\tilde{v}_{q-1} = \tilde{d}_{q-1}^{T}\,U\,S^{-1}\,, \tag{4}$$

through a straightforward extension of (2). This vector representation for $\tilde{d}_{q-1}$ is adequate under some conditions on the general patterns of the domain considered; see [9] for a complete discussion.

In (3), $\Pr\left(w_q|\tilde{d}_{q-1}\right)$ reflects the "relevance" of word $w_q$ to the admissible history, also referred to as a pseudo-document [9]. As such, it will be highest for words whose meaning aligns most closely with the semantic fabric of $\tilde{d}_{q-1}$ (i.e., relevant "content" words), and lowest for words which do not convey any particular information about this fabric (e.g., "function" words like *the*). Since content words tend to be rare and function words tend to be frequent, this will translate into a relatively high perplexity value. Thus, the model (3), by itself, will likely exhibit a rather weak predictive power. Hence the need to integrate it as part of a multi-span formalism.

## 3. INTEGRATION WITH N-GRAMS

The LSA framework provides a way to handle some of the global constraints in the language. To obtain a multi-span language model, it remains to combine them with local constraints, such as provided by the $n$-gram paradigm. Obviously, the goal of the resulting integrated approach is to leverage the benefits of both.

The integration can occur in a number of ways, such as straightforward interpolation, or within the maximum entropy framework [4]. In the following, we develop an alternative formulation for the combination of the $n$-gram and LSA paradigms. The end result, in effect, is a modified $n$-gram language model incorporating large-span semantic information.

To achieve this goal, we need to compute:

$$\Pr\left(w_q|H_{q-1}\right) = \Pr\left(w_q|H_{q-1}^{(n)}, H_{q-1}^{(l)}\right), \tag{5}$$

where the history $H_{q-1}$ now comprises an $n$-gram component $(H_{q-1}^{(n)} = w_{q-1}w_{q-2}\ldots w_{q-n+1})$ as well as an LSA component $(H_{q-1}^{(l)} = \tilde{d}_{q-1})$. This expression can be rewritten as:

$$\Pr\left(w_q|H_{q-1}\right) = \frac{\Pr\left(w_q, H_{q-1}^{(l)}|H_{q-1}^{(n)}\right)}{\sum_{w_i \in \mathcal{V}} \Pr\left(w_i, H_{q-1}^{(l)}|H_{q-1}^{(n)}\right)}\,, \tag{6}$$

where the summation in the denominator extends over all words in $\mathcal{V}$. Expanding and re-arranging, the numerator of (6) is seen to be:

$$\begin{aligned}
&\Pr\left(w_q, H_{q-1}^{(l)}|H_{q-1}^{(n)}\right) \\
&\quad = \Pr\left(w_q|H_{q-1}^{(n)}\right)\Pr\left(H_{q-1}^{(l)}|w_q, H_{q-1}^{(n)}\right) \\
&\quad = \Pr\left(w_q|w_{q-1}w_{q-2}\ldots w_{q-n+1}\right) \\
&\qquad\qquad \cdot \Pr\left(\tilde{d}_{q-1}|w_q w_{q-1}w_{q-2}\ldots w_{q-n+1}\right).
\end{aligned} \tag{7}$$

Now we make the assumption that the probability of the document history given the current word is not affected by the immediate context preceding it. This reflects the fact that, for a given word, different syntactic constructs (immediate context) can be used to carry the same meaning (document history). This is obviously reasonable for content words, and probably does not matter very much for function words. As a result, the integrated probability

becomes:

$$\Pr(w_q|H_{q-1}) = \frac{\Pr(w_q|w_{q-1}w_{q-2}\ldots w_{q-n+1})\Pr(\tilde{d}_{q-1}|w_q)}{\sum_{w_i\in\mathcal{V}}\Pr(w_i|w_{q-1}w_{q-2}\ldots w_{q-n+1})\Pr(\tilde{d}_{q-1}|w_i)}. \quad (8)$$

Note that, if $\Pr(\tilde{d}_{q-1}|w_q)$ is viewed as a prior probability on the current document history, then (8) simply translates the classical Bayesian estimation of the $n$-gram (local) probability using a prior distribution obtained from (global) LSA.

## 4. IMPLEMENTATION ISSUES

There are two ways to take advantage of multi-span modeling for large vocabulary speech recognition. One is to rescore previously produced N-best lists using the integrated models. (This was the scenario implicitly assumed in [7] and [9].) The other is to use the multi-span models directly in the search itself. The latter is preferable, since it allows incremental pruning based on the best knowledge source available. Compared to N-best rescoring, however, using multi-span modeling directly in the search entails a much higher computational cost.

In a typical large vocabulary search performed on an average sentence of length, say, 10 seconds, several hundred to several thousand unique word contexts could be active at any given frame. Thus, the computational load is potentially several orders of magnitude greater than for simple post-search rescoring. One concern, in particular, is the calculation of each pseudo-document vector representation in (4), which requires $\mathcal{O}(MR)$ floating point operations.

Fortunately, we can exploit the sequential nature of pseudo-documents to reduce this computational cost. Clearly, as each word context is expanded, the document context remains largely unchanged, with only the most recent candidate word added. Assume further that the training corpus $\mathcal{T}$ is large enough, so that the normalized entropy complement does not change appreciably with the addition of each pseudo-document. Then it is possible to express the new pseudo-document vector directly in terms of the old pseudo-document vector, instead of each time re-computing the entire mapping from scratch.

To see that, consider $\tilde{d}_q$, and assume, without loss of generality, that word $w_i$ is observed at time $q$. Then, from (1), we will have for $1 \le k \le M$, $k \ne i$:

$$w_{k,q} = w_{k,q-1}\,, \quad (9)$$

while, for $k = i$:

$$w_{i,q} = g_i\,\frac{c_{i,q-1}+1}{n_q} = \frac{n_q-1}{n_q}\,w_{i,q-1} \;+\; \frac{g_i}{n_q}\,. \quad (10)$$

Hence, with the shorthand notation $\gamma_{i,q} = g_i/n_q$, we can express $\tilde{d}_q$ as:

$$\tilde{d}_q = \frac{n_q-1}{n_q}\,\tilde{d}_{q-1} \;+\; [0\ldots\gamma_{i,q}\ldots 0]^T\,, \quad (11)$$

which is turn implies, from (4):

$$\tilde{v}_q = \frac{n_q-1}{n_q}\,\tilde{v}_{q-1} \;+\; \gamma_{i,q}\,u_i\,S^{-1}\,. \quad (12)$$

It is easily verified that (12) requires only $\mathcal{O}(R)$ floating point operations. Thus, we can update the pseudo-document vector directly in the LSA space at a fraction of the cost previously required to map the sparse representation to the space $\mathcal{S}$.

Note that the other potential bottleneck, the computation of the integrated probability (8), can also be alleviated through appropriate caching of the LSA probabilities. The above fast pseudo-document update therefore allows multi-span language modeling to be exploited in early stages of the search, if desired.

## 5. RECOGNITION RESULTS

As in [9], we have trained the LSA framework on the WSJ0 part of the NAB News corpus. This was convenient for comparison purposes since conventional $n$-gram language models are readily available, trained on exactly the same data [8]. The training text corpus $\mathcal{T}$ was composed of about $N = 87,000$ documents spanning the years 1987 to 1989, comprising approximately 42 million words. The vocabulary $\mathcal{V}$ was constructed by taking the 20,000 most frequent words of the NAB News corpus, augmented by some words from an earlier release of the Wall Street Journal corpus, for a total of $M = 23,000$ words.

We performed the singular value decomposition of the matrix of co-occurrences between words and documents using the single vector Lanczos method [10]. Over the course of this decomposition, we experimented with different numbers of singular values retained, and found that $R = 125$ seemed to achieve an adequate balance between reconstruction error (as measured by Frobenius norm differences) and noise suppression (as measured by trace ratios). Using the resulting vector space $\mathcal{S}$ of dimension 125, we constructed the LSA model (3) and combined it with the standard bigram, as in (8).

The resulting multi-span language model, dubbed bi-LSA model, was then used in lieu of the standard WSJ0 bigram model in a series of speaker-independent, continuous speech recognition experiments. These experiments were conducted on a subset of the Wall Street Journal 20,000 word-vocabulary task. The acoustic training corpus consisted of 7,200 sentences of data uttered by 84 different native speakers of English (WSJ0 SI-84). The test corpus consisted of 496 sentences uttered by 12 additional native speakers of English.

It is important to note that the task chosen represents a severe test of the LSA component implemented above. By design, no more than 3 or 4 consecutive sentences are related to a single article. As a result, the test corpus comprises 140 distinct document fragments, which means that each speaker speaks, on the average, about 12 different "documents." This prevents the multi-span model from

| Speaker | Reduction in Perplexity | Reduction in Word Error Rate |
|---|---|---|
| 001 | 22.8 % | 8.4 % |
| 002 | 28.5 % | 21.5 % |
| 00a | 30.6 % | 17.5 % |
| 00b | 27.4 % | 10.1 % |
| 00c | 33.6 % | 10.0 % |
| 00d | 26.2 % | 17.3 % |
| 00f | 33.3 % | 11.5 % |
| 203 | 35.3 % | 16.1 % |
| 400 | 15.4 % | 14.8 % |
| 430 | 19.7 % | 19.3 % |
| 431 | 20.0 % | 12.2 % |
| 432 | 24.7 % | 7.8 % |
| Overall | 24.7 % | 13.7 % |

**Table 1.** Performance Improvement Using Bi-LSA Language Modeling.

building a very accurate pseudo-document representation, since the context effectively changes every 60 words or so. (In situations like these, it is beneficial to implement a mechanism to consistently forget the context, to avoid relying on an obsolete representation; details will be presented in [11].)

The performance achieved using the bi-LSA language model were compared to that achieved using the baseline bigram, as measured by both test data perplexity and actual word error rate. Table 1 summarizes the results obtained, in terms of perplexity reduction (first column) and word error rate reduction (second column). It can be seen that overall we observed a reduction in perplexity of about 25%, and a reduction in average error rate on the order of 15%.

As usual, the reduction in average error rate is less than the corresponding reduction in perplexity, due to the influence of the acoustic component in actual recognition, and the resulting "ripple effect" of each recognition error. Note that in the case of $n$-LSA language modeling, this effect can be expected to be more pronounced than in the standard $n$-gram case. This is because recognition errors are potentially able to affect the LSA context well into the future, through the estimation of a flawed representation of the pseudo-document in the LSA space. This lingering behavior, which can obviously degrade the effectiveness of the LSA component, is an unfortunate by-product of large-span modeling. Clearly, the more accurate the recognition system, the less problematic this unsupervised context construction becomes.

## 6. CONCLUSION

We have described a data-driven framework for the integration of the various constraints, both local and global, that are present in the language. This approach exploits the complementarity between the $n$-gram formalism, which inherently relies on syntactically-oriented, short-span relationships, and latent semantic analysis, which tends to capture semantically oriented, large-span relationships between words. This synergy can be harnessed through an integrative formulation which combines the two paradigms.

The resulting multi-span language model was shown to outperform the associated standard $n$-gram on a subset of the Wall Street Journal speaker-independent, 20,000-word vocabulary, continuous speech task. Specifically, we observed a reduction in perplexity of about 25%, and a reduction in average error rate of about 15%.

## 7. REFERENCES


[1] L.R. Bahl, F. Jelinek, and R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Anal. Mach. Intel.*, Vol. PAMI–5, No. 2, pp. 179–190, March 1983.

[2] T. Niesler and P. Woodland, "A Variable–Length Category–Based N–Gram Language Model," in *Proc.* 1996 *Int. Conf. Acoust., Speech, Sig. Proc.*, Atlanta, GA, pp. I164–I167, May 1996.

[3] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger–Based Language Models: A Maximum Entropy Approach," in *Proc.* 1993 *Int. Conf. Acoust., Speech, Sig. Proc.*, Minneapolis, MN, pp. II45–48, March 1993.

[4] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer Speech and Language*, Vol. 10, London: Academic Press, pp. 187–228, July 1996.

[5] J.R. Bellegarda *et al.*, "A Novel Word Clustering Algorithm Based on Latent Semantic Analysis," in *Proc.* 1996 *Int. Conf. Acoust., Speech, Sig. Proc.*, Atlanta, GA, pp. I172–I175, May 1996.

[6] J.R. Bellegarda, "A Latent Semantic Analysis for Large-Span Language Modeling," in *Proc. EuroSpeech*'97, Rhodes, Greece, Vol. 3, pp. 1451–1454, September 1997.

[7] J.R. Bellegarda, "Exploiting Both Local and Global Constraints for Multi-Span Statistical Language Modeling," in *Proc.* 1998 *Int. Conf. Acoust., Speech, Sig. Proc.*, Seattle, WA, Vol. 2, pp. 677–680, May 1998.

[8] F. Kubala *et al.*, "The Hub and Spoke Paradigm for CSR Evaluation", in *Proc.* ARPA *Speech and Natural Language Workshop*, Morgan Kaufmann, pp. 40–44, March 1994.

[9] J.R. Bellegarda, "A Multi-Span Language Modeling Framework for Large Vocabulary Speech Recognition," *IEEE Trans. Speech Audio Proc.*, Vol. 6, No. 5, September 1998.

[10] M.W. Berry, "Large–Scale Sparse Singular Value Computations," *Int. J. Supercomp. Appl.*, Vol. 6, No. 1, pp. 13–49, 1992.

[11] J.R. Bellegarda, "Large Vocabulary Speech Recognition With Multi-Span Statistical Language Models," *IEEE Trans. Speech Audio Proc.*, in preparation.