Three Probabilistic Language Models for a Large-Vocabulary Speech Recognizer[1]

P. Dumouchel, V. Gupta[2], M. Lennig[2], P. Mermelstein[2]
INRS-Télécommunications
3 Place du Commerce
Montreal, Quebec, Canada H3E 1H6

Abstract

This paper compares relative performance of three different language models applied to the linguistic decoding part of a 75,000-word speech recognizer. These models are the trigram model, the tri-POS model (POS stands for parts of speech), and a smoothed trigram model with tied distributions for words three or more syllables long. The full trigram model gives the best performance but is most expensive in terms of data and storage requirements. The smoothed trigram and tri-POS models yield equivalent performance. For general text entry tasks use of the tri-POS model is suggested since it is less sensitive to variations in the discourse domains. For applications specific to individual discourse domains, trigram models trained on data obtained from the target domain are recommended.

I. Introduction

The task of the INRS-Télécommunications 75,000-word recognizer is to automatically transcribe speech into text. The recognition task has been divided into two parts: acoustic recognition and linguistic decoding. For each word of spoken input, the acoustic recognizer generates a list of word hypotheses and their associated acoustic likelihoods. These word hypotheses are then input to the linguistic decoder which applies syntactic and semantic constraints to generate the a posteriori most likely word string. The linguistic constraints are generated by the language model. The focus of this paper is statistical language modeling. We compare three different statistical language models used by the linguistic decoder.

The trigram language model (Jelinek, 1985a) has been used successfully for both a 5000-word and a 20,000-word office correspondence task (Averbuch et al., 1987). To obtain reliable trigram statistics for our 75,000-word task recognizing sentences from a general discourse domain requires a large training text collected from different task domains. The storage required for the trigram probabilities also turns out to be prohibitive.

To reduce the storage and training text requirements, Derouault and Merialdo (1986) use a tri-POS model (where POS stands for parts of speech) for a 250,000-word recognizer. We have implemented both the trigram language model and the tri-POS language model. In addition, we have tested a smoothed trigram language model with tied distributions for words three or more syllables long.

The frequency of error for current speech recognizers using only acoustic information is much higher for shorter words than for longer words. We attempt to focus the power of the language model on shorter words by tying the distributions (Bahl et al., 1983) of words three or more syllables long in the trigram language model. We call the trigram language model with such tyings the *smoothed* trigram model.

In section II, we give an overview of our 75,000-word recognizer. Section III describes the text corpuses used for training the language models. Section IV specifies the 75,000 word lexicon. Section V describes the experiments with the trigram language model. Section VI gives the recognition results with the tri-POS language model. Section VII introduces the smoothed trigram language model and compares its performance with the trigram and the tri-POS language models.

II. Overview of the 75,000-word isolated word recognizer

We recognize words in two steps. The first step is the acoustic recognition of the spoken words. The input to the acoustic recognizer consists of words spoken with pauses of at least 150 ms between them. The spoken text consists of sentences read from text which was selected randomly from magazines, books, newspaper articles, and office correspondence. An endpoint detector segments the acoustic data A for the spoken word-string W into subsegments $A_1, \ldots, A_n$ corresponding to the words $w_1, w_2, \ldots, w_n$ in the word-string. For each segment A_i, the acoustic recognizer generates a list $w_i^1, w_i^2, \ldots, w_i^{N_i}$ of N_i most likely word choices, together with their likelihoods $(P(A_i|w_i^j), j = 1, N_i)$. These likelihoods are smoothed by taking their seventh root before being passed to the second step of recognition. The probabilities

[1] This work was supported by the National Sciences and Engineering Council of Canada

[2] also with Bell-Northern Research, Montreal, Canada

CH2561-9/88/0000-0513 $1.00 © 1988 IEEE

$$P(A|W) = P(A_1|w_1)P(A_2|w_2)\ldots P(A_n|w_n)$$

are used in the second step of recognition. The acoustic recognizer has been described in detail by Gupta, Lennig and Mermelstein (1987).

The second step of word recognition involves linguistic decoding. The linguistic decoder computes the most likely word-string $\widehat{W}$ using

$$\widehat{W} = \arg\max_{W} P(W)P(A|W).$$

The acoustic recognizer provides the probabilities $P(A|W)$, while the language model provides the probabilities $P(W)$. The search algorithm we have used to find $\widehat{W}$ is called the A^* algorithm (Nilsson 1980), and is also known as the *stack decoding* algorithm (Jelinek, 1976). The focus of this article is the language model which estimates the word-string probabilities $P(W)$.

III. Text Databases

Two databases were available to us to train the language models. One database is called the *Brown Corpus* (Francis and Kucera, 1979) and consists of 1.1 million words of text collected from different sources to represent written American English in 1960. The discourse domain for the Brown corpus is quite diverse. The second database is the proceedings of the Canadian Parliament in Ottawa (Hansard). The Hansard corpus comprises approximately 15 million words of text. Even though the discussions in the Hansard corpus cover a wide range of topics, the text is more homogeneous and this is reflected in a lower perplexity of the language model for the Hansard corpus than for the Brown Corpus (67 vs 329).

IV. 75,000-Word Lexicon

All recognition experiments reported here are based on a vocabulary consisting of 76,211 orthographically distinct words. These 76,211 words correspond to the contents of Merriam Webster's Seventh Collegiate Dictionary (henceforth W7) (1965), together with most of the words in the Brown Corpus (Francis and Kucera, 1979) not in W7, plus approximately 400 common first names of people. Words are distinguished orthographically. For example, *book* and *books* are considered two different words in the lexicon.

V. The Trigram Language Model

The trigram language model estimates the $P(W)$ for all possible word-strings $W = w_1 w_2 \ldots w_n$ in the language as

$$P(W) = P(w_1)P(w_2|w_1)\prod_{i=3}^{n} P(w_i|w_{i-2}, w_{i-1}).$$

SIZE	% OF TRIGRAM COVERAGE FOR	
(Million)	Brown Corpus	Hansard Corpus
1	24%	52%
4	32%	63%
8	36%	68%

Table 1 Size of the training database (Hansard) versus percent of trigram coverage in the test database.

In the trigram language model, the only parameters to be estimated are the conditional probabilities $P(w_3|w_1, w_2)$ for all possible words w_1, w_2, w_3 in the lexicon.

We have used the interpolated trigram model to estimate the conditional probabilities $P(w_3|w_1, w_2)$ (Jelinek 1985b) as

$$P(w_3|w_1, w_2) = q_3 f(w_3|w_1, w_2) + q_2 f(w_3|w_2) + q_1 f(w_3),$$

where q_i are the interpolation weights which sum to one and $f(w_3|w_1, w_2)$, $f(w_3|w_2)$ and $f(w_3)$ are the relative frequencies which are computed using a large training text corpus. The interpolation weights are computed using a forward-backward algorithm on a training text disjoint from the one used to compute the relative frequencies f (Jelinek 1985b).

Let us first estimate the amount of training text required to estimate the relative frequencies $f(w_3|w_1, w_2)$ for our 75,000-word recognizer. The training size required can be computed by estimating the percentage of trigrams in a disjoint test set which are covered by the training set. Generally, over 75% trigram coverage is required for adequate training of the trigram language model. For example, Jelinek (1985a) uses 30 million words of training text to estimate the trigram probabilities for a 5000-word vocabulary, resulting in 85% coverage of the trigrams in a new text.

To estimate the size of the training set required for 75% trigram coverage, we computed the percentage of trigrams in a test set of 400,000 words covered by the training set (Hansard corpus) as we increase the size of this set. The test set of 400,000 words was disjoint from the training text. The results are shown in Table 1.

The interesting result is that the trigram coverage for a test set from the Hansard corpus is significantly higher than for a test set from the Brown Corpus leading to the speculation that the discourse domain for the Canadian Parliament is much more homogeneous than for the Brown Corpus. A training corpus size of 8 million for the language model is probably sufficient to perform linguistic decoding for the word-strings taken from the Hansard corpus. To perform linguistic decoding of spoken word-strings taken from the Brown corpus, we probably require training of the language model with a much larger training text. For the 935 words of spoken text used to test our recognizer, only 26% of the trigrams are covered by the 8 million words of the Hansard corpus.

Jelinek (1985b) has defined the measure of quality as the perplexity of the language model. The perplexity is computed by taking a reasonably large text distinct from the training set with words $w_1 w_2 \ldots w_n$. Perplexity is then computed as

$$Perplexity = P(w_1 w_2 \ldots w_n)^{-1/n}.$$

Perplexity can be equated to the average branching factor (Jelinek 1985b) or to the average size of the set of words between which the recognizer must decide when transcribing a word of the spoken text. The perplexity of the trigram language model with increasing training text is given in Table 2. As is evident from Tables 1 and 2, the trigram coverage of a text and its perplexity are directly related.

The perplexity of the language model also depends on the discourse domain. The perplexity of the trigram language model for the Brown Corpus is significantly higher than that for the Hansard corpus as is evident from Table 2.

Size	Perplexity	
(Million)	Brown Corpus	Hansard Corpus
1	1130	172
4	495	86
8	329	66

Table 2 Size of the training data (Hansard corpus) for the trigram language model versus the perplexity of the test data.

We store the relative frequencies $f(w_3)$, $f(w_3|w_2)$, and $f(w_3|w_1, w_2)$ which actually occur in the training corpus together with the interpolation weights. The 14 million words of the Hansard corpus contains 75,000 distinct monograms, 1.1 million bigrams and 4.1 million trigrams requiring in excess of 100 megabytes of storage. We have, therefore, experimented with only 8 million words of text for training the trigram language model parameters requiring approximately 40 megabytes of storage.

We tested the trigram language model on 560 words of text spoken by speaker ML and 375 words of text spoken by speaker JM. The acoustic recognizer gave a recognition rate[3] of 80% on speaker JM and 83% on speaker ML. After linguistic decoding using the interpolated trigram model, we obtain 90% and 89% recognition rate on speakers JM and ML respectively. The trigram language model eliminated 43% of the errors observed after acoustic recognition.

[3] The recognition rate of the acoustic recognizer is given by the percent of words for which the top word choice is correct. Homophone confusions are not counted as errors. For recognition rates using the language model, orthographic differences are counted as errors.

VI. The tri-POS language model

The tri-POS language model (Derouault and Merialdo, 1986) computes the probability of a word-string W as

$$P(W) = \prod_{i=1}^{n} P(w_i|g_i)P(g_i|g_{i-2}, g_{i-1}),$$

where g_i is the part of speech for word w_i. In our tri-POS model, we have used a total of 21 parts of speech. The number of parameters to be estimated for the tri-POS language model is less than 100,000: these are the 21^3 parameters for the conditional probabilities $P(g_3|g_1, g_2)$, and less than 90,000 parameters for the conditional probabilities $P(w_i|g_i)$. The 21 parts of speech correspond primarily to those used in W7. The training and memory requirements are significantly reduced for the tri-POS model. The memory required for parameter storage is approximately 1 megabytes.

The tri-POS language model is trained using the 1.1 million words of Brown Corpus. The perplexity of the tri-POS language model turns out to be 1250 for a text from the Hansard Corpus, twenty times that for the trigram language model trained with 8 million words of text. However, for the 935 words of text used for testing the recognizer, the perplexity of the tri-POS is 1535, compared to 865 for the trigram language model.

After linguistic decoding, we obtain 86% recognition rate for speaker JM and 90% for speaker ML. The tri-POS language model eliminated 35% of the errors observed after acoustic recognition as compared to 43% for the trigram language model.

VII. The smoothed trigram model with tied distributions for words three or more syllables long

A precise probability distribution in the language model for words three or more syllables long may not be needed for two reasons. First, the error rate for the acoustic recognizer for words three or more syllables long is quite low. Secondly, very few words three or more syllables long have homophones. This suggests that we may be able to smooth the probability distributions for words three or more syllables long without affecting the performance of the recognizer. The smoothing is done by tying the distributions (Bahl et al. 1983) for all words three or more syllables long. We will call the trigram language model obtained after such smoothing as the *smoothed trigram model.*

The effect of such smoothing is to reduce the vocabulary size considerably. The number of words with distinct probabilities is reduced to 23,800 words from the original 75,000. The memory requirement is reduced by a factor of 4 compared to the trigram language model. Such smoothing introduces a certain information loss. When two of the words in $P(w_3|w_1, w_2)$ are three or more syllables long, then the discriminating power in the language model corresponds

to a monogram language model. Only when all the three words are two or less syllables long that the discriminating power of the language model corresponds to a trigram language model.

The error rate for speaker JM using the smoothed trigram model trained on 14 million words from the Hansard corpus is 86%, while for speaker ML it is 88%. The smoothed trigram language model results in 31% reduction in errors after acoustic recognition, compared to 43% for the trigram language model and 35% for the tri-POS model.

The perplexity for the three different language models is compared in Table 3, while their recognition performance is compared in Table 4. The perplexity of the trigram language model for the 965 word test set is the lowest and hence gives the best performance (43% reduction in error rate), while the perplexity of the smoothed trigram model is the highest and performs the worst (31% reduction in error rate).

Language model (training size in million)	Perplexity for		
	Brown Corpus	Hansard Corpus	965-word test set
Trigram (8)	329	67	865
tri-POS (1)		1250	1535
Smoothed Trigram (8)	1000	516	
Smoothed Trigram (14)	782	436	3197

Table 3 Comparison of perplexity for three different language models for three different task domains.

The performance of the language model is directly related to the perplexity of the task. The perplexity of the test set chosen from various magazines, books and newspapers varies between 865 and 3200 depending on the language model used. Were the test sentences chosen from the Hansard corpus, then one may expect significantly improved performance for the trigram model as compared to the tri-POS model and the smoothed trigram model.

Speaker	Language model			
	none	trigram	tri-POS	smoothed trigram
ML	83%	89%	90%	88%
JM	80%	90%	86%	86%

Table 4 Recognition results for three different language models. The test sets consist of 560 and 375 words of text for speakers ML and JM respectively.

VIII. Conclusions

The memory and training text required for tri-POS language model is significantly smaller than that for the trigram language model while the perplexity of the two language models for general discourse domains varies only by a factor of two due to insufficient training for the trigram language model. For a general discourse domain, the tri-POS language model is recommended as it results in a large saving in memory requirements.

Training for a trigram language model for a restricted discourse domain is quite feasible. One example of a restricted discourse domain is the Hansard corpus. For the Hansard corpus the perplexity is only 67 for the trigram model, 436 for the smoothed trigram model, and 1250 for the tri-POS model. For such discourse domains, trigram language model will lead to significant improvement in recognition performance over the smoothed trigram model and the tri-POS model.

REFERENCES

Averbuch, A. et al. (1987). "Experiments with the Tangora 20,000 word speech recognizer," *Proc. 1987 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 701-704.

Bahl, L.R., Jelinek, F. and Mercer, R.L. (1983). "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, no. 2, pp. 179–190.

Derouault, A.-M. and Merialdo, B. (1986). "Natural language modeling for phoneme-to-text transcription," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 742–749.

Francis, W.N. and Kucera, H. (1979). *Manual of information to accompany a standard sample of present-day edited American English, for use with digital computers*, Department of Linguistics, Brown University.

Gupta, V., Lennig, M., and Mermelstein, P. (1987). "Integration of acoustic information in a large vocabulary word recognizer," *Proc. 1987 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 697–700.

Jelinek, F. (1976). "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556.

Jelinek, F. (1985a). "The Development of an Experimental Discrete Dictation Recognizer," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1616–1624.

Jelinek F. (1985b). "Markov Modeling of Text Generation," *The Impact of Processing Techniques on Communications Series E: Applied Sciences*, no. 91, J.K. Skwirzynski, editor, Kluwer Academic, Netherlands.

G. & C. Merriam Co. (1965). *Webster's Seventh New Collegiate Dictionary.*

Nilsson, N.J. (1980). *Principles of Artificial Intelligence*, Tioga Publishing Co., Palo Alto, CA.