THE INFLUENCE OF BIGRAM CONSTRAINTS ON WORD RECOGNITION BY HUMANS: IMPLICATIONS FOR COMPUTER SPEECH RECOGNITION

Ronald A. Cole   *Yonghong Yan*   *Troy Bailey*

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
20000 N.W.Walker Road, Portland, OR 97291-1000
{cole, yan, bailey}@cse.ogi.edu

ABSTRACT

The gap between human and machine performance on speech recognition tasks is still very large. Recognition of words in telephone conversations is slightly better than 50%, based on results reported on the Switchboard corpus by leading researchers using state of the art HMM systems. We know from our own experience that human perception typically delivers much more accurate word recognition over the telephone. Why is there such a large gap between machine and human performance, and what can be done to close this gap? One way to address this question is to study the sources of linguistic information in the speech signal that are known to be important for word recognition, and measure how well machine systems utilize this information relative to humans. As an initial step in this direction, we measured word recognition performance of listeners presented with words from the Switchboard corpus. Stimuli consisted of actual utterances excised from the Switchboard corpus, high quality recordings of utterances that occurred in Switchboard conversations, and recordings of word sequences with zero, medium and high bigram probabilities based on a language model computed from transcriptions of the Switchboard corpus. The results show that human listeners are very good at recognizing words in the absence of word sequence constraints, and that statistical language models fail to capture much of the high level lingusitic information needed to recognize words in fluent speech. The results are discussed in terms of their implications to current approaches to acoustic and language modeling in computer speech recognition.

1. Introduction

Machine recognition performance in natural conversations, such as telephone conversations in the Switchboard corpus [1], depends heavily upon the benefits derived from statistical language modeling. When word sequence constraints provided by statistical language models are removed from the recognition system, word recognition performance degrades dramatically, by a factor of three to five; from about 50% word recognition accuracy on the Switchboard corpus, to about 10% with current HMM-based systems [2].

Current HMM approaches to transcription of words in natural continuous speech rely mainly on acoustic modeling and language modeling. In the former, the system uses acoustic data to model words as sequences of subword units, such as phonemes, which are matched to word models. In the latter, the system uses word sequence constraints to chose among word candidates. The two sources of information are then combined according to a weighting factor computed from training data. To gain insights into the relative contribution of acoustic and language models on word recognition in fluent speech, we performed perceptual experiments with human listeners.

2. Experiment 1: the Role of Language Modeling in Speech Perception

Experiment 1 was designed to examine the effect of word sequence constraints on word recognition performance by humans. We were especially interested in word recognition accuracy when these constraints are absent, given the high error rates observed for conversational speech when language models are not used.

We constructed word sequences with 3 levels of bigram probabilities computed from the Switchboard corpus: (1) zero, (2) medium, and (3) high probabilities. The resulting word sequences were recorded by a professional announcer, who was instructed to make the word sequences sound as much like a natural sentence as possible; i.e., with natural sentence-like intonation.

Examples of the three sequence types are:

1. Zero Bigram Probability
   - acid how or crazy lost
   - guns eat utmost grad you stint
2. Medium Bigram Probability
   - fine at most turn move
   - safe moment cook road but tough
3. High Bigram Probability
   - amazing how guns left it
   - hi today like what distance rod

The generation of word sequences consisted of two steps; generating a very large base of word strings (utterances), and then selecting a subset of these at three levels of bigram constraint that were balanced by word category and sequence length. The bigram model was estimated using approximately 200,000 utterances in the original Switchboard database.

Thirty million utterances were generated by permutating the words (with vocabulary size 4129) in the Switchboard database. Since longer utterances would run the risk of biasing performance due to limitations of human short-term memory, the utterances were all 5 to 7 words in length.

Based on the language scores, ninety utterances with various bigram probabilities were selected. Among them, 30 utterances had zero bigram probability, 30 utterances had medium bigram probabilities and 30 had high bigram probabilities. There were 10 five-word utterances, 10 six-word utterances and 10 seven-word utterances in each category (zero, medium and high).

The zero bigram probability sequences consisted of word squences in which adjacent words never occurred in the training database. The medium and high bigram probabilities were defined according to the histogram of bigram probabilities of the thirty million generated word sequences; medium bigram probabilities refer to sequences in which all word pairs occurred at the median of the histogram, and high bigram probability sequences consisted of word pairs that occurred with greatest frequency.

The 90 utterances were recorded using a Sennheiser HMD 414 close-talking microphone by a male native speaker of American English in a quiet office environment using the 8-bit D/A audio interface on a Sun SPARCclassic workstation. The speech data were sampled at 8 kHz. The recordings were then checked by another native speaker of American English for both recording quality and naturalness.

The short words (such as "it", "as" etc.) were balanced for each category. For words in each category, approximately 75% of them are single syllable words, which is consistent with natural spoken language (based on our analysis of the Switchboard corpus).

2.1. Experimental Procedure

Thirteen high school graduates who were native speakers of American English were paid to participate in the experiment. The experiments were conducted in a quiet office. Speech utterances were played using the built-in audio device of a Sun workstation. Subjects listened through high quality Sony MDR-84 closed ear-cushion headphones. A simple GUI was designed to allow subjects to interact with computers, so that they might control the volume of each utterance, and proceed through the experiment at their own pace. Each subject had a short practice session before the experiment to become familiar with the test requirements. Subjects were taught to use an emacs editor to enter and alter their responses.

During the experiment, subjects listened to each utterance five times, in order to eliminate possible effects of short term memory. Subjects were instructed to type in as many words as they thought they understood after each presentation, and to edit their previous responses if they so desired. We saved and analyzed the subjects' responses after each presentation.

The experiment was divided into two sessions, each of which contained 45 utterances with an even number of high, mid, and low probability utterances on each session. Word length was also balanced across both sessions. In each session, utterances were presented in random order.

2.2. Results and Discussion

Before the data were subject to analysis of variance, the text files of the subjects' responses were spell checked. Subjects' responses were analyzed for word-level accuracy using a string-alignment program which makes use of dynamic programming to compare subjects' responses with the actual sequence of words in each utterance. Word accuracy scores were derived from the alignment algorithm.

Table 1 shows recognition scores (based on word-level accuracy) for 3 conditions: high probability, medium probability and zero probability (no structure). Analysis of variance showed that word recognition accuracy was greater for the high and medium probability utterances than for the zero probability utterances ($p < .01$), and that no difference in performance was observed between high and medium bigram constraint. (84.2% and 85.8%, respectively, on the fifth presentation).

Presentation No.	Bigram Probability		
	High	Medium	Zero
1	53.0	44.0	33.6
2	73.0	68.4	57.0
3	79.1	77.0	67.8
4	81.5	80.8	72.8
5	85.8	84.2	77.3

Table 1: Percentage of Words Correct for Three Probability Levels

One of the main findings of this experiment is that human listeners are able to do quite well recognizing words in sequences with minimal bigram constraint. Subjects in our experiment recognized 77% of the words in zero bigram sequences by the fifth repetition.

Repeating the word sequences had a large effect on recognition accuracy. In all three conditions, about twice as many words were recognized after 5 presentations than after a single presentation. Performance improved steadily over the five repetitions in all three conditions, including from the fourth to fifth presentation, suggesting that word recognition performance may not have asymptoted by the fifth repetition.

3. Experiment 2: Recognition Benchmarks for the Switchboard Corpus

The second experiment was motivated by both theoretical and practical issues in computer speech recognition. The goals of the experiment were (a) to provide a perceptual benchmark for recognition of Switchboard utterances; and (b) to compare word recognition by human listeners on read versus conversational speech.

3.1. Test Data Selection and Preparation

Sixty utterances were selected from the Switchboard Corpus. Thirty of these utterances were recorded by the speaker used in experiment 1. The other thirty were excised from the actual Switchboard corpus conversations. Each utterance contained 5 to 7 words. Utterances were selected to maintain the balance of short words to multisyllabic words across the two conditions.

Examples of the selected utterances include the following:

- you think it starts out well
- it is incumbent upon me
- put that on a card
- seldom, only in dire emergencies
- my youngest daughter is, and

The utterances in the "read speech" group were recorded by the same speaker as in the first experiment, with the same recording environment and equipment. As before, the sequences were produced as naturally as possible.

The speech intervals for the utterances in the "excised speech" group were extracted from the digitized files of the Switchboard corpus based on the time-aligned transcriptions. The utterances were double checked by two native American English speakers. Some alignment mistakes in the transcriptions were found and corrected.

The Signal Noise Ratio (SNR) was 34 db for the read speech utterances and 27 db for the original Switchboard wave files.

3.2. Experimental Procedures

Twelve subjects, different from those used in experiment 1, were paid to participate in a session lasting approximately one hour. The subjects were high school graduates, native speakers of American English and reported no known hearing loss. The equipment and experimental procedures were the same as those used in experiment 1, and the experiments were conducted in the same location. The subjects were given a short practice session to familiarize them with the equipment and procedures. During the test session, each utterance was played 5 times, with the pace controlled by the subjects.

3.2..1 Results

Subjects' responses were spell-checked as in the first experiment. Pauses and background comments were removed from the subjects' responses, and orthographic problems were resolved. Results of the experiment (including repetitions) are summarized in Table 2.

The results are interesting. Despite the semantically anomalous nature of some of the utterances, subjects were able to recognize 9 out of 10 words (89.7%) on utterances excised from Switchboard telephone conversations. When Switchboard utterances were presented as read speech, subjects identified 97.7% of the words.

Repetition No.	Utterance Style	
	Dictation	Conversation
1	90.6	77.7
2	96.6	87.2
3	97.5	89.5
4	97.5	89.5
5	97.7	89.7

Table 2: Word Accuracy for Two different types of utterances

Subject No.	Utterance Style	
	Dictation	Conversation
bjd	97.8	88.4
j0s	99.4	90.1
jms	97.8	79.7
jrb	97.2	95.9
lew	96.7	93.2
mjn	98.9	94.2
mlc	97.2	84.3
p0h	98.3	94.8
prm	96.7	93.6
s0s	97.8	88.4
src	97.8	83.1
thc	96.7	91.3

Table 3: Percentage of Words Recognized for each subject on the last (the 5th) presentation

Examination of Table(2) shows that performance asymptoted after two to three presentations. For the read speech, performance asymptoted after two presentations; for the excised utterances, performance asymptoted between two and three presentations.

4. Discussion

The results of the two experiments can be summarized as follows: In experiment 2, subjects presented with utterances excised from the Switchboard corpus recognized words at about 90% accuracy after three presentations of the utterances. Subjects presented with utterances that occurred in

Switchboard, but were recorded over a high quality microphone, recognized about 98% of the words after two presentations. In experiment 1, after five presentations, subjects recognized words in recorded sequences with zero bigram frequency about 77% of the time, and in sequences with medium or high bigram about 85% of the time.

The following conclusions are suggested by the pattern of results observed in the two experiments:

1. Human listeners are very good at recognizing words from acoustic data in the absence of high level syntactic, semantic, discourse and other contextual constraints;
2. Word sequence constraints, such as those measured by bigram language models, help human listeners recognize words. The magnitude of this effect is small relative to the other results observed in our experiments, and appears limited to the range between no (zero) bigram constraint and moderate bigram constraint.
3. Bigram language models do not capture the constraints between words that are found in natural language. Our experiments showed a large difference in word recognition between recordings of word sequences that were spoken in Switchboard conversations and recording of word sequences with high bigram constraint. The magnitude of this effect is large– 98% (after 2 presentations) for recordings of natural utterances vs. 86% (after 5 presentations) for recordings of high bigram word sequences. This represents a 6- to 7-fold difference in the error rate, and suggests that bigram estimates do not capture many of the high level linguistic constraints used by humans. Apparently, high scores for bigram probabilities do not ensure language-like behavior. This should cause us to reflect on the nature of stochastic language modeling and not over-estimate its ability to replicate "natural language".
4. Recognition of words in actual telephone conversations is more difficult than recognition of words in recordings produced in the laboratory using a high quality microphone. Our results showed a large difference in recognition of words excised from Switchboard conversations compared to our speaker's recordings of word sequences that occurred in Switchboard (90% vs 98%, respectively).

   This large performance difference between read speech and conversational speech in experiment two is consistent with the results of perceptual experiments conducted by [3] on human listeners' ability to recognize read speech from the Wall Street Journal corpus. These researchers found error rates of about 1%; about an order of magnitude fewer errors than we observed for Switchboard utterances. It is clear that recognizing words in conversational speech is much more difficult than recognizing words in read speech, for both humans and machines.

The implications of these experiments are sobering. While it is widely accepted that human performance is substantially better than machine performance, the magnitude of this difference is very large. Our results, and those reported by others, show that human word recognition performance is five to ten times more accurate than machine performance. In difficult tasks, such as recognizing words in telephone conversations, word recognition performance without language modeling is very poor– about 10%. Yet our experiments show that statistical languge modeling techniques do not capture much of the important information that is used by humans to recognize words.

It is not surprising that current systems do not approach human performance. Studies of speech perception and acoustic phonetics have shown that there is a wealth of information in the speech signal that is used by human perceivers to discriminate among speech sounds and recognize words. Much of this information is not captured by current frame-based approaches to speech recognition. For example, current systems do a poor job of capturing discrete events, such as stop bursts and glottalization, of modeling the dynamic movements of the articulators, and of integrating prosodic information into the recognition process. Similarly, it is clear that syntactic and semantic information is often conveyed through complex word relationships that are not adequately modeled via bigram or trigram statistics.

The challenge in the future will be to create recognition architectures that are able to measure and integrate the information needed to scale to human performance. We suspect that this will be achieved by investigating new approaches to computer speech recognition rather than trying to modify current ones.

5. Acknowledgement

The authors wish to thank Dr. Chuck Wooters for generating the bigram table from switchboard for us, Dr. Mark Fanty for the comments and the member companies of CSLU for the partial support of this work.

6. References

[1] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research development," in *Proceedings 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, (San Francisco, USA), pp. I–517 – I–520, March 1992.

[2] J. Cohen, H. Gish, and J. Flanagan, "Switchboard—the second year," in *CAIP Summer Workshop in Speech Recognition: Frontiers in Speech Processing II*, 1994.

[3] W. Ebel and J. Picone, "Human speech recognition performance on the 1994 csr s10 corpus," in *the ARPA HLT meeting*, February 1995.