Composing Letters with a Simulated Listening Typewriter

JOHN D. GOULD, JOHN CONTI, and TODD HOVANYECZ IBM Research Center

ABSTRACT: With a listening typewriter, what an author says would be automatically recognized and displayed in front of him or her. However, speech recognition is not yet advanced enough to provide people with a reliable listening typewriter. An aim of our experiments was to determine if an imperfect listening typewriter would be useful for composing letters. Participants dictated letters, either in isolated words or in consecutive word speech. They did this with simulations of listening typewriters that recognized either a limited vocabulary (1000 or 5000 words) or an unlimited vocabulary. Results suggest that some versions, even upon first using them, could be at least as good as traditional methods of handwriting and dictating. Isolated word speech with large vocabularies may provide the basis for a useful listening typewriter.

INTRODUCTION

A "listening typewriter" is a potentially valuable aid in composing letters, memos, and documents. Indeed, it might be a revolutionary office tool, just as the typewriter, telephone, and computer have been. With a listening typewriter, an author could dictate a letter, memo, or report. What he or she says would be automatically recognized and displayed in front of him or her. A listening typewriter would combine the best features of dictating (e.g., rapid human output) and the best features of writing (e.g., visual record, easy editing). No human typist would be required, and no delay would occur between the time an author creates a letter and when he or she gets it back in typed form. This might lead to faster and better initial composition by the author, psychological closure because of no wait for (and uncertainty about) a typed copy, quicker and better communication, and displaceable typing and organizational costs.

The state of the art in automatic speech recognition today, however, is not advanced enough to make a reliable listening typewriter (see summaries in [9, 14]).

An increasingly important—and available—human factors tool is simulation of user interfaces before the interfaces are ever built. Such studies can guide and impact the development of technology when the most flexibility for change and improvement exists. The present experiments make use of simulation in studying the value of composing letters with a listening typewriter (since a real listening typewriter does not exist today).

This study compared people's performance on and feelings about a listening typewriter with their performance on and feelings about traditional methods of composing (writing and dictating). Of particular interest was whether an imperfect listening typewriter would be useful in composing letters.

Perhaps the most difficult technical problem for automatic speech recognition today is the problem of word segmentation. The boundary cues marking the end of one word and the beginning of the next word, although good enough for human perception, are generally not clear enough for automatic speech recognition. That is why most speech recognition devices commercially available

John D. Gould is an experimental psychologist who likes to apply experimental psychology to interesting problems in the use of computers. Todd Hovanyecz received his BA in Mathematics and Physics from Manhattanville College in 1981 and is now working as a systems engineer. John Conti is now a programmer.

Authors' Present Addresses: John D. Gould, IBM Research Center, Box 218, Yorktown Heights, New York 10598; Todd Hovanyecz, IBM, 590 Madison Ave, New York, New York; John Conti, Touchstone Applied Science Associates, Inc., 150 Clearbrook Rd, Elmsford, NY.

today require the user to speak in isolated words, that is, to put a clear pause of 100 msec. or longer between each word. Thus, one variable studied was speech mode. Participants composed some letters by speaking in isolated words and some letters with consecutive word speech. Would the necessity to pause between each word bother people, significantly interrupt their thought processes, or otherwise affect their composition behavior? If people compose letters with isolated words as well as they compose with consecutive word speech, then a useful listening typewriter would be much easier to make.

A second variable studied was the size of the vocabulary that the listening typewriter could recognize. Reviews of commercially available isolated word systems report that in actual applications they can recognize only about 10-30 words (at any word choice position) with an accuracy of 97-98 percent [2, 12]. Much better performance is reported in laboratory settings [14]; some of this gain in vocabulary size seems to be occurring now in applications. For example, Poock [13] recently showed (using a Threshold Technology T600 recognition system) that vocabularies of up to 240 words are recognized with 97-98 percent accuracy. Nippon Electric Company reports they are able to recognize a vocabulary size of 120 continuously spoken words, and estimates they can recognize 1000 isolated words [10]. Thus, progress is being made toward successful recognition of larger vocabulary sizes. Clearly, the particular words that comprise the set of words to be recognized significantly affect recognition accuracy. In practice, speech recognition devices attempt to recognize whatever utterance is said within a time window of a second or two, regardless of whether the utterance is a single word or short phrase. Thus, single isolated words could be considered as a special case of isolated phrase recognition.

In the present study, vocabulary sizes of 1000, 5000, and an unlimited number of words were used. Again, if people perform as well with a relatively small vocabulary as with the (theoretically unrealizable) unlimited one, then a useful listening typewriter could probably be designed sooner and the resulting system would be less costly.

Two additional variables, composition strategy and whether participants had experience at dictating, were also studied. In Experiment 1 participants composed half their letters using a draft strategy (as explained below) and half their letters using a first time final strategy. Pilot data suggested these strategies interacted with the system parameters, and therefore needed to be studied separately. In Experiment 1 participants had no experience at dictating, whereas in Experiment 2 participants were experienced dictators.

We had several hypotheses:

1. Participants would compose written letters somewhat faster than they would compose letters with the listening typewriter because of their relative unfamiliarity with the latter and because of the lack of easy editing with the listening typewriter simulated here.
2. Participants would compose letters in consecutive word speech faster than in isolated words because of the required pauses in the latter.
3. Quality of letters would be the same with all methods, based upon earlier findings about lack of quality differences among methods of composing [5-7].
4. In the subsequent proofediting stage, there would be more changes made and more time spent on letters composed with the listening typewriter than on letters composed with writing because of the limited editing capability provided with the former. Also, in this proofediting stage, there would be more changes made and more time spent on letters composed with a first time final strategy because in the latter some of this work was done while composing. Finally, there would be more changes made and more time spent on letters composed with consecutive word speech than on letters composed with isolated words because in the latter the pauses would serve to allow the author to be more careful about sentence construction and word selection.

GENERAL METHOD

General Procedure

Participants learned to use a listening typewriter by watching a 20 minute videotape that showed another author using it. The videotape made reference to one page of editing instructions that participants had in front of them (Appendix A). Participants then practiced composing two-four letters with different versions of the listening typewriter. No mistakes in using the listening typewriter were ever made after two letters.

The formal experiment then started. Participants were given the description of a letter to compose and the method to use to compose it. In each letter, participants tried to convince a recipient of something. These included trying to win a bid for paper supplies and recommending a favorite teacher for an annual award. Participants were allowed to make written notes if they chose. While composing, they were allowed to make any changes they wanted. After composing a letter, participants were given a typed version of it about 20 minutes later and allowed to proofedit it. This was called the Proofediting stage. There was only one redo, or proofediting, of the printed version. While composing, participants were videotaped. The amount of time they actually talked was automatically recorded [6]. When participants finished, they were briefly interviewed about their feelings for that method and they formally rated it.

Simulation of Listening Typewriter

Figure 1 depicts the simulation method. A typist, located in another room, listened to a participant dictate via a closed-circuit TV system and typed what was said. The

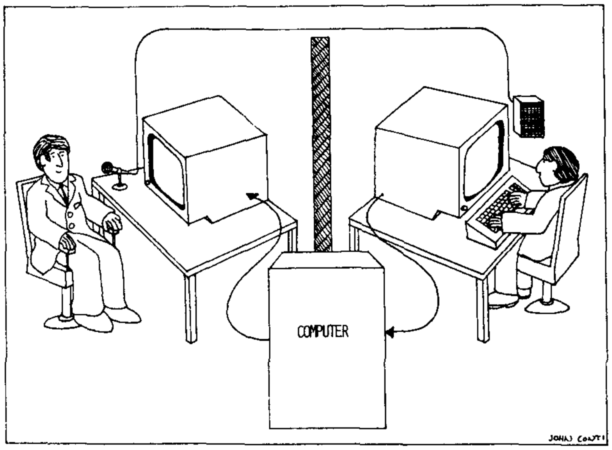

FIGURE 1. Schematic of the Experimental Setup.

information typed was not only displayed on the typist's computer-controlled cathode-ray tube (CRT) display terminal (IBM 3277), but also appeared on the participant's terminal (IBM 3277), which was yoked to the typist's through the IBM full screen support system. Specifically, the typist heard a word, typed it, and then hit the "Enter" key on the terminal. The computer then checked whether the word was in the dictionary being simulated. The dictionaries were taken from Kucera and Francis' norms [11] of the most frequently used English words, and were stored in the computer (IBM 168V). If the word was not in the dictionary, XXXX's were displayed on the participant's screen. If it was in the dictionary, the computer then checked whether it was a homophone (or homonym, as such words are usually referred to in the literature). If it was, then the most frequent version of the homophone was displayed, regardless of the version that had been typed. This was done because we wanted to simulate an "unintelligent" recognition system. These data could then be used as a baseline for comparison with recognition systems of various amounts of "intelligence."

In simulation of isolated word versions of the listening typewriter, the typist hit the "Enter" key after each word. Participants were instructed not to say the next word until the previous word had been displayed to them. A beep sounded when the word was displayed, so that participants did not always have to look at the screen. There was almost a two-second delay from the time participants began to say a word until they could begin to say the next word. This estimated delay was the sum of the following approximate values: (a) 0.3 sec for a participant to say the word; (b) 1.0 sec for the typist to type the word and press the Enter key; (c) 0.1 sec for the computer to process and display the word; (d) 0.2 sec reaction time for the participant to say the next word. Thus this isolated word simulator could achieve rates of about 30 wpm.

If a participant said the next word too quickly, an electronic monitoring device detected this and prevented the typist from hearing the next word. In practice this was not necessary, as it almost never happened.

In consecutive word speech, participants could talk as fast as they wanted. They did not have to wait between words. (One participant had to be told to slow down, however, because the typist could not keep up with him.) If a participant was speaking rapidly, the typist would type several words before hitting the Enter key. This had little or no effect on the computer processing time, allowed the typist to get up to speed, and thus allowed the simulator to keep up better with a participant. Presumably, the typist was typing these word strings at about 80 wpm. Thus, the consecutive word speech simulator could achieve rates of 50-60 wpm.

Participants' Editing Commands

Participants spoke each editing or formatting command. The typist typed an abbreviation of it, and its effect was shown on the participants' screen. Participants could use editing commands offering function equivalent to that contained in ordinary dictating equipment. We provided this limited function rather than a more elaborate facility because this was standard, easily describable, required less training of participants, less programming of the simulator, and did not force us to invent an editor for a composing method that did not exist anyway. In addition, the results could serve as a baseline against which to measure editors having more power.

The only way a participant could change what he or she had already said was, in effect, to record over the word. A participant would say "NUTS," which erased the last word shown, and then say another word. If a participant wanted to change the fifth to last word, he would say "NUTS 5," which erased the last five words, and then say the new word and the last four words over. If the wrong homophone (e.g., "in" rather than "inn") was displayed, a participant could say "NUTS" and repeat the homophone, which caused a different version of it to appear (regardless of which version the typist typed).

Participants could spell unrecognized words by saying "SPELLMODE," spelling the word, and then saying "ENDSPELLMODE." They could capitalize the first letter of a word ("CAPIT") or capitalize all letters of a word ("CAPALL"). They could cause numerals to be displayed ("NUMMODE" . . . "ENDNUMMODE") rather than the spelled out version of them. They could cause modifications in the formatting of a letter with four commands: "NEWPARAGRAPH," "NEWLINE," "INDENT N(spaces)," and "SPACE N(spaces)." (The simulator automatically started a new line once the previous line was filled; the NEWLINE command started a new line without the previous line being filled.)

The inside address and return address were supplied for participants on their CRT prior to the beginning of each letter. The CRT could display 20 lines at one time. Once participants had composed a longer letter than this, they could scroll through parts of their letter not displayed by saying "SCROLL-TOWARD-BEGINNING" and "SCROLL-TOWARD-END."

The Typist

The typist played a critical role in the success of these experiments. Our typist was selected because she typed 80 wpm, was excellent at following the rules of simulation, remained cool, did not provide a participant with any help, and was available for several months. In addition, she was a practicing stenotypist, which gave her experience with transcribing oral material in real time. She also knew shorthand, which was useful in Experiment 2. She practiced with the simulator for two-three weeks prior to the experiments. During this time, several human factors improvements were introduced to lighten her burden, speed overall performance, and make the simulation more compelling. She was particularly accurate at typing exactly what the participant said, including exclamations and parenthetical comments which a participant made to himself. She seemed effective in typing what she heard, even when context suggested that the participant had said a different word. However, she did misspell words occasionally.

EXPERIMENT 1

Method

Participants. Ten people with characteristics similar to many professional, managerial, and technical office workers, spent two days each composing letters. They ranged in age from about 25 to 70. Most had at least a bachelor's degree, four worked for IBM, four were female. The IBMers were volunteers, and the non-IBMers were obtained from a local temporary employment agency.

Letter-tasks. There were ten letter-tasks, or composing assignments. In each a participant composed a letter to convince a recipient of something. Tasks included applying for a job, applying for a grant of money for a favorite project, and recommending a relocation site for one's office.

Composing Methods and Design. Eight composing methods used the listening typewriter. These corresponded to the eight combinations of three variables: speech mode [isolated word (*I*) vs. consecutive word speech (*C*)], vocabulary size [1,000 words (*1*) vs. unlimited (*U*)], and composing strategy [draft (*D*) vs. first time final (*F*)]. Sometimes a method will be referred to by its initials, for example, *C1D* stands for a letter composed in consecutive word speech with the 1000 word vocabulary with the author using a Draft strategy. With a Draft strategy, participants were instructed to make a quick draft. They were told they could leave unrecognized words on the screen, and make any changes they wished in the subsequent proofediting stage. With a Final strategy, participants were instructed to make the listening typewriter version of their letter as close as possible to the final version of their letter. They were told to remove all unrecognized words by spelling them. They were told they could, however, make any changes they wanted to in the subsequent proofediting stage.

For control or comparison purposes, participants wrote two letters, one when they arrived for the experiment and the other later on during the experiment. Each participant wrote his/her first letter on a different letter-task. The order of the remaining nine composing methods, the order of the letter-tasks, and the combination of the composing methods and letter-tasks were varied from participant to participant, with a 9 × 9 greco-latin square. The tenth participant received another row from a different 9 × 9 square. The use of ten participants let each letter-task be completely balanced with the two written letters and the other eight composing methods.

Performance Evaluation. Participants were told their performance would be evaluated on the time to compose their letters and the resulting effectiveness of them, and that these two factors would be weighted equally. Composition time was measured from when the participant began reading about a letter-task until he or she was finished composing. Effectiveness was rated by three judges (one experimenter and two English teachers) who compared all ten letters written on a particular topic and rank ordered the best three. Since all letters were essentially requests or recommendations, a judge rated the letters' effectiveness according to how likely he or she would be to grant the request or follow the recommendation. The three judges worked independently.

A letter was assigned an effectiveness score of 10 if it received a first place vote from a judge (i.e., it was the best of 10 letters), 9 if it received a second place vote, and 8 if it received a third place vote. A score of 4 was assigned if it received no vote (4 is the average score if the remaining seven letters had been rank-ordered). The scores for each letter were summed over the three judges, and ranged from 30 for a letter with the three first place votes to 12 for a letter with no votes.

We attempted to weight effectiveness and time equally by transforming the effectiveness scores to have the same mean, same variance, and same range as the time scores, and to go in the same direction as the time scores (i.e., smaller values reflect better performance). We could not achieve all these goals with a single linear transformation. As a compromise, we transformed the effectiveness scores linearly to decrease with greater effectiveness, to have the same mean as the time scores, and to have the same interquartile range as the time scores. The effectiveness scores were calculated with $(64 - .5R)$, where R is the sum of the ratings of the three judges, which varied from 12 to 30.

Preference Rating Scale. After using a version of the listening typewriter, each participant compared it to writing using a seven-point scale, where 1 = significantly worse than writing, 3 = a little worse than writing, 4 = same as writing, 5 = a little better than writing, and 7 = significantly better than writing.

Results

Time and Effectiveness of Letters. Table I shows detailed results for the eight versions of the listening typewriter studied. As shown by the standard errors of the means,

TABLE I. Participants' Mean Composing Time (min.), Mean Effectiveness Arbitrary Units (with lower scores being more effective than higher scores), and Median Preferences (scale of 1-7) in Experiment 1. (*W1* = first written letter; *W2* = second written letter; *I* = isolated word speech; *C* = connected word speech; *1* = 1000 word vocabulary; *U* = unlimited word vocabulary; *D* = draft strategy; *F* = first time final strategy)

	W1	W2	I1D	I1F	IUD	IUF	C1D	C1F	CUD	CUF
Composition time	24.0	26.5	23.8	28.4	18.1	24.8	20.6	27.0	14.7	18.7
(St. Error)	(3.5)	(5.9)	(4.8)	(4.1)	(3.3)	(3.2)	(3.2)	(5.4)	(2.7)	(2.6)
Proofedit time	2.6	2.3	4.0	1.6	3.7	2.1	4.4	1.8	3.7	2.0
(St. Error)	(0.6)	(0.5)	(0.6)	(0.5)	(2.0)	(0.4)	(0.6)	(0.5)	(1.5)	(0.5)
TOTAL TIME	26.6	28.7	27.9	30.0	21.8	26.9	25.0	28.8	18.4	20.7
(St. Error)	(3.7)	(6.3)	(5.2)	(4.5)	(4.7)	(3.5)	(3.3)	(5.8)	(3.9)	(2.8)
Effectiveness	34.2	30.8	31.2	27.8	31.4	28.2	34.8	27.0	34.8	31.2
(St. Error)	(2.6)	(2.9)	(3.5)	(4.6)	(3.3)	(4.3)	(3.5)	(3.8)	(1.8)	(2.4)
Composition time + Effectiveness	29.1	28.6	27.5	28.1	24.7	26.5	27.7	27.0	24.8	25.0
(St. Error)	(2.0)	(2.2)	(2.4)	(2.3)	(1.5)	(2.0)	(2.6)	(2.6)	(1.6)	(1.1)
Total time + Effectiveness	30.4	29.8	29.5	28.9	26.6	27.5	29.9	27.9	26.6	26.0
(St. Error)	(1.9)	(2.4)	(2.5)	(2.3)	(2.1)	(2.0)	(2.6)	(2.7)	(2.1)	(1.1)
Preference Rating re Writing (on 7-pt scale)	—	—	5	6	5	6	5	4	6	6

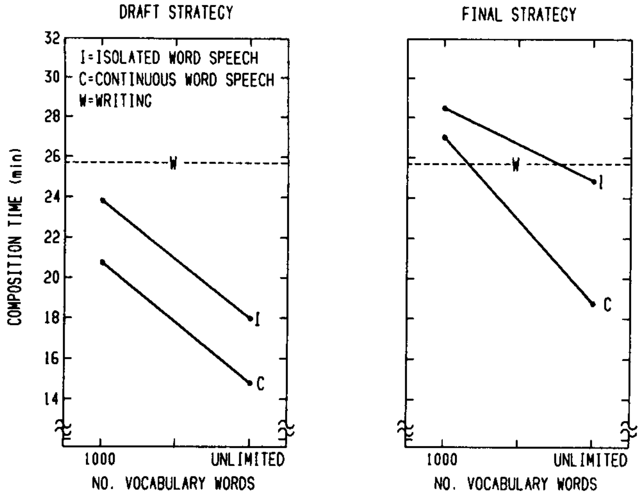

FIGURE 2. Mean Composition Times for the Composing Methods in Experiment 1.

composition times and total times were highly variable. Some of this was due to the letter-tasks themselves, which accounted for as much variance as did composition methods. Most individual time scores or effectiveness means in any one row in Table I were not significantly different from each other.

Figure 2 shows that composition times for letters composed with the listening typewriter under Draft instructions tend to be faster than composition times for written letters, whereas letters composed with the listening typewriter under Final instructions were closer to the times for Written letters. This is consistent with the fact that participants were instructed to compose Written letters with a first time final strategy.

With the listening typewriter, letters composed under Draft instructions were faster than letters composed under first time final instructions [analysis of variance; $F(1,9) = 9.28$; $p < .05$].[1] Letters composed in consecutive word speech were somewhat faster than letters composed in isolated word speech [($F(1,9) = 3.99$; $p < .10$]. Letters composed with an unlimited vocabulary were somewhat faster than letters composed with the 1,000 word vocabulary [$F(1,9) = 4.69$; $p < .10$]. When proofediting time was added to composition time, these trends remained about the same (Figure 3).

As shown in Table I, letters composed with a Final strategy were more effective (28.6) than letters composed with a Draft strategy [33.1; $F(1,9) = 5.23$; $p < .05$]. Effectiveness was not influenced by speech mode or vocabulary size [analyses of variances; $p > .10$]. Interjudge reliability on effectiveness of these letters was low.

Figure 4 shows combined performance scores, based upon equal weighting of effectiveness and total time. Combining time scores and effectiveness measures may seem curious since this is a little like mixing apples and oranges and secondly, they have the opposite polarity (until effectiveness scores are transformed, as explained in the Methods section). However, their combination gives a more complete picture of composing efficiency, and that is why we report the combined scores. Lower scores mean better performance. Although the trends are similar to those for time scores alone, none of the differences shown in Figure 4 are significant [analyses of variance; all $p > .10$]. The rank-order correlation, grouped across participants, between the effectiveness of a method and total time spent with that method was not significant ($R = -.54$; $p > .10$). (There was a significant negative rank-order correlation between effectiveness and composition time, however; $R = -.81$; $p < .01$, that is, methods which led to faster composition times also led to less effective letters.)

Composition Rate. The mean number of words in a letter did not differ significantly from method to method (Table II; mean = 179 words; analysis of variance; $p > .10$). (These word counts do not include the 30-word return address and inside address which were supplied to participants.) This implies that differences in composition rate (words/composition time) are mainly due to differences

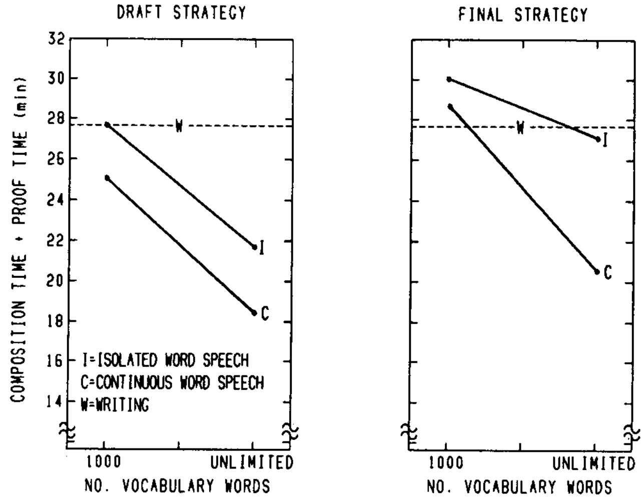

FIGURE 3. Mean Total Time (composition time plus proofediting time) for the Composing Methods in Experiment 1.

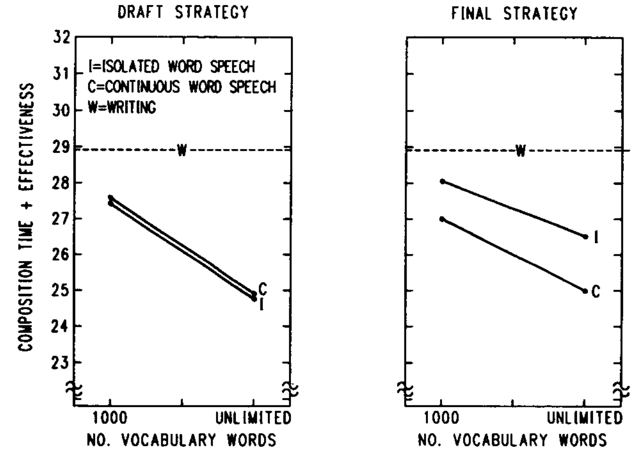

FIGURE 4. Mean Performance Scores, Based Upon Equal Weighting of Composition Time and Effectiveness for each Letter, for the Composing Methods of Experiment 1. The effectiveness scores have been transformed so that, like the time scores, smaller values are better.

[1] For the reader unfamiliar with statistical tests of significance, the p values given following the result of an F-test or a R-analysis estimate the probability that the difference reported is due to chance. For example, a $p < .05$ provides an estimate that the difference among the scores tested could occur by chance less than 1 out of 20 times.

in composition time only. Composition rate was faster for unlimited vocabulary than for limited vocabulary (11.5 vs. 7.1 wpm; $F(1,9) = 46.0$; $p < .001$). Composition rate was also faster for Draft strategy than for Final strategy (10.6 vs. 8.0 wpm; $F(1,9) = 5.13$; $p < .05$), but was only somewhat faster for consecutive word speech than for isolated word speech (10.3 vs. 8.3 wpm; $F(1,9) = 4.49$; $p < .10$).

Differences among Participants. The range of participants' composition time scores was 10.8–39.8 min. After adding proofediting times to these, the range was 12.2–44.2 min. The variance among participants' total time scores was about seven times greater than the variance among the ten composition methods. There was a significant negative rank-order correlation between the time a participant spent composing letters and the resulting effectiveness of them ($R = -.79$; $p < .01$), and between a particpant's total time and the resulting effectiveness of his or her letters($R = -\ .75$; $p < .01$). Participants, for example, who took relatively long to compose letters had more effective letters than did participants who were faster.

Proofediting. In the proofediting stage, participants were given a printout of their letter about 20 minutes after they had composed it. They spent 2.8 min. proofediting. The quality of their proofediting was poor. Forty-five of 100 letters were left with at least one spelling error. The average was about two spelling errors in these 45 letters. In addition, participants left homophone spelling errors that would rarely occur in more traditional forms of composition (e.g., "... body and sole"). Almost no major changes were made. As shown in Table III, only a few changes were made in each letter, except in *I1D* and *C1D* where participants spelled out over 20 XXXX's per letter. Most changes were minor rewordings, punctuation, and capitalization.

We expected that there would be more changes to Draft letters than to Final letters. This was not the case, however, as there were fewer changes made in each of nine categories in Table III with the Draft strategy! This may explain why Final letters were judged more effective than *D* letters. Letters composed with *C1D* and *CUD* were proofread especially poorly, and letters composed with these two methods were rated as least effective. In Written letters, participants made fewer punctuation and capitalization changes, and more minor rewordings, than with the listening typewriter.

Participants' Opinions. Participants compared each method with Writing, on a seven-point scale. Participants varied considerably among themselves on how well they liked individual versions of the listening typewriter. Four participants gave (almost) all versions a rating of 6 ("Better than writing") or 7 ("Significantly better than writing"). Two participants rated most versions less favorably than writing. The remaining participants differentiated more broadly among the methods. As shown in Table I, the median rating of all versions was either 4 ("Same as writing"), 5 ("A little better than writing"), or 6 ("Better than writing").

General Observations. Participants made few notes. They easily learned how to use the listening typewriter. None of them quit the experiment. They had no difficulty giving verbal commands to the simulator while composing. They made almost no mistakes in using it after their first practice letter. They reported feeling no pressure to talk—unlike novices upon first learning to use traditional dictating equipment (see [3]). They did not view this as dictating. Rather they said they could "see their writing." No one reported being frightened of the microphone.

Observations on Methods of Composition

Writing. Participants volunteered that editing, or making changes, was much easier with writing than with a listening typewriter.

Isolated Word Speech/1000 Word Vocabulary. Behaviorally, participants had little trouble speaking in isolated words. Typically, they would say a word while looking at the screen, wait for it to be displayed, and then say another word. They almost never spoke the next word too soon. Indeed, there often was a pause between when a participant could speak the next word and when he or she actually did speak it. Participants usually looked at the screen while composing, although they could look away and listen for a beep that signalled when they could say the next word.

Participants' comments were mainly negative, with more centering on the limited vocabulary size than on the restriction to speak in isolated words. Twenty-four percent of the words participants used were not recognized (Table II), and XXXX's were therefore displayed. When using a Final strategy, participants became aggravated at having to spell out about one of every four words. Use of SPELLMODE "... adds time," "... requires more concen-

TABLE II. Word Analyses Based upon Means of each Method in Experiment 1. (*W1* = first written letter; *W2* = second written letter; *I* = isolated word speech; *C* = connected word speech; *1* = 1000 word vocabulary; *U* = unlimited word vocabulary; *D* = draft strategy; *F* = first time final strategy.)

	W1	W2	I1D	I1F	IUD	IUF	C1D	C1F	CUD	CUF
Number of words in letter	155	200	162	152	174	198	157	186	204	201
Number unrecognized words	(41)	(54)	39	39	(44)	(54)	32	43	(51)	(48)
Number XXXX's left for proofediting	...	...	22.2	...	...	...	25.7	...	...	...
Percent words in 1000 vocabulary	(74)	(73)	76	74	(75)	(73)	80	77	(75)	(76)
Percent words if 5000 vocabulary	(91)	(91)	(90)	(91)	(92)	(89)	(93)	(91)	(91)	(92)
Composition rate (words per min.)	7.2	8.5	7.2	5.6	12.2	8.4	9.3	6.5	13.9	11.5

Note. The numbers in parentheses in lines 2 and 4 indicate the words that would have been affected if the 1000 word vocabulary limitation had been applied to those composition methods. The numbers in parentheses in line 5 indicate the percent of words that would have been affected if the 5000 word vocabulary limitation used in Experiment 2 had been applied to all those composition methods.

TABLE III. Total Number of Changes Made in the Proofediting Stage of the Ten Letters Composed with each Method in Experiment 1. (*W1* = first written letter; *W2* = second written letter; *I* = isolated word speech; *C* = connected word speech; *1* = 1000 word vocabulary; *U* = unlimited word vocabulary; *D* = draft strategy; *F* = first time final strategy.)

	W1	W2	I1D	IUD	I1F	IUF	C1D	CUD	C1F	CUF
Formatting; spacing	6	4	3	1	5	2	4	1	7	2
Spelling	1	4	4	0	0	3	2	1	7	0
Punctuation	2	4	3	3	6	1	2	5	10	6
Minor rewording	29	24	14	6	17	15	19	15	41	11
Major rewording	1	0	2	0	6	1	1	0	4	2
Defining XXXX's	0	0	196	0	0	0	249	0	0	0
Capitalizing	0	4	8	7	36	8	19	0	11	7
Homophones	2	1	4	0	2	1	4	2	7	3
Typos made by the simulator	3	4	1	0	2	1	0	1	1	0
System problems	0	1	2	0	0	2	0	3	6	1

tration," and "... is distracting." The limited vocabulary required spelling the more difficult words, and "I'm a poor speller," said one participant. With a Draft strategy, participants were often uncertain about how to handle XXXX's. They could either spell them or leave them to the proofediting stage. Some participants found that if they waited they forgot what a particular XXXX stood for. This was frustrating, as much of the meaning was in these words ("meat words," said one participant). On the average, 22.2 XXXX's were left in these Draft letters, which was about half of the 40 unrecognized words per letter with this version (Table II).

Isolated Word/Unlimited Vocabulary. Participants preferred this method to the 1,000 word vocabulary because there were no XXXX's to handle. Some participants criticized the isolated word restriction, saying that it was "... hard to wait for the beep," it "... interrupts flow," "... interrupts plans," "... doesn't allow cohesive thought." Others said it provided a "nice pace." I "... can choose my words," and "... don't have to think as fast."

The simulator, displaying one word at a time, could handle dictating rates of about 30 wpm. Participants dictated much slower than this (Table II) and did not complain about delays (although, as will be seen, participants in Experiment 2 did complain).

Consecutive Word Speech/1,000 Word Vocabulary. A significant finding was that participants were, in effect, compelled to dictate in isolated words when using this consecutive speech method with a Final strategy. In order to change a word participants had to erase all words said after the one to be changed. Thus, they typically would say a word, wait to see whether it was recognized, and go on to say the next word if it was. If it was not, they would spell the word before saying the next word.

Participants commented mainly on the disadvantages of the 1,000 word vocabulary. With a Draft strategy, they left 25.7 XXXX's, or 70 percent of the words that were not recognized (Table II). "XXXX's are more annoying in consecutive speech," said one participant. "You get ahead," said another, "then you must go back and spell." "I forgot a lot of phrases because of the need to spell," said another. "It's disconcerting not to know the 1,000 words." "It's no good with adjectives," and "... anything technical." "I have to look at the screen constantly to fill in words."

Consecutive Word Speech/Unlimited Vocabulary. This is the unachievable ultimate in speech recognition. Typically a participant said a phrase, read it, possibly edited it, paused, said another phrase, etc. Editing was very local, probably influenced by the primitive editing facilities. Participants dictated substantially faster with this version of the listening typewriter than with the others (Table II, mean = 12.7 wpm). They did not, however, dictate as fast as the simulator could go with this method (about 50–60 wpm), or as fast as the rates of 17–25 wpm found with standard dictating equipment [4].

Unlike with the other three versions, participants' comments centered on composing, not on the listening typewriter itself. "Final strategy is hard work," and "I must think more." "With Draft (as opposed to Final) I get my thoughts down quickly," "... could concentrate on content," and "... can talk without looking at screen." On the other hand, one participant said, "I didn't notice much difference between Draft and first time Final strategies." Participants mentioned that spelling out XXXX's was compelling even with Draft, that waiting for the machine was disconcerting, and that the conversational tone was more human than they imagined working with computers would be. They said that they were able to visualize their work better than with traditional Writing.

EXPERIMENT 2

This experiment evaluated the use of a listening typewriter by experienced dictators. It compared their performance and attitudes on listening typewriters with their performance and attitudes on dictating to a machine (*DMACH*) and dictating to a secretary taking shorthand (*DSEC*).

We hypothesized that their performance would be at least as good with the more efficient versions of the listening typewriter as with *DMACH*. This was based upon the finding that experienced dictators are 25 percent faster at dictating than at writing [4], and the finding that letters composed with several of the listening typewriters studied in Experiment 1 were composed more than 25 percent faster than were Written letters. We also hypothesized that they would like at least some listening typewriters better than *DMACH* and *DSEC* because they could see what they had said.

The importance of studying experienced dictators was to learn whether they displayed differential performance and attitudes, compared to nondictators, on various versions of the listening typewriter. For example, experience at dictating might enhance performance with a listening typewriter. Also, dictation experience might lead to more positive attitudes about using listening typewriters.

We were particularly interested in carefully assessing the opinions of these participants about different versions of the listening typewriter. We did this in three ways. Participants compared each version of the listening typewriter, just after using it, with their favorite method of composition. Second, after having used all versions of the listening typewriter, participants rank-ordered each version according to which one they would most likely use in real life. Third, at the end of the experiment, participants composed a letter of their own and had to choose a method of composition with which to do it.

We reduced the number of conditions studied because we thought that this might have contributed to the large variance in Experiment 1. Five listening typewriter versions were studied. Four of these were the same as in Experiment 1. Isolated word speech/1000 word vocabulary (*I1000*) was chosen because it was a minimal version, and consecutive word speech/unlimited word vocabulary (*CU*) was chosen as the ultimate, although unachievable, speech recognition system. The trade-off between vocabulary size and speech mode was assessed by studying *C1000*, *I5000*, and *IU*. Had a 5000 word vocabulary been used in Experiment 1, about 91 percent of words would have been recognized (Table II), which is what the Kucera and Francis [11] norms predict also.

Method

Participants. Eight IBM executives, all with considerable experience with dictation, spent one day composing eight letters. Most participants were in their thirties, and they ranged in age from 33-52. All had at least a bachelor's degree. They were professionals in marketing requirements for office products.

Composing Methods. Each participant composed a different letter with seven different methods: *I1000*, *I5000*, *IU*, *C1000*, *CU*, dictating to an IBM 6:5 dictating machine (*DMACH*), and dictating to a secretary taking shorthand (*DSEC*). Participants were allowed to use either a Draft or first time Final strategy.

Letter-tasks. Seven of the ten letter-tasks used in Experiment 1 were used here.

Design. The order of the seven composing methods, the order of the seven letter-tasks, and the combination of the composing methods and letter-tasks were varied, from participant to participant, with a 7 × 7 greco-latin square. We were able to obtain an eighth participant, and he received a row from a different 7 × 7 square.

Performance Evaluation. Performance was evaluated just as in Experiment 1.

Preference Rating Scale. Prior to the experiment participants told us their favorite method of composing. Five preferred *DMACH* (oftentimes after writing an outline) and three preferred Writing. After using a version of the listening typewriter, each participant compared it to his or her favorite method of composing on a seven-point scale, where 1 = significantly worse than my favorite method, 3 = a little worse than my favorite method, 4 = same as my favorite method, 5 = a little better than my favorite method, and 7 = significantly better than my favorite method.

Results

Time and Effectiveness of Letters. Figure 5 shows that time scores for *DMACH* and *DSEC* were faster than several versions of the listening typewriter. They were significantly faster than all three isolated word versions [Table IV; F(6,42) = 19.34; $p < .001$; Duncan range test; $p < .05$]. *I1000* was significantly slower than all methods (Duncan range test; $p < .05$).

Figure 5 shows that *C1000* was relatively faster in this experiment than in the previous one. This was due, at least in part, to participants speaking fast, not spelling a large proportion of XXXX's (Table VI), and leaving them until the proofediting stage (see Table IX). Also, letters composed with *C1000* were the least effective (Table IV).

As shown in the right panel of Figure 5, proofediting time partially compensated for some of the differences in composition times among the methods. The result was that total times (composition time plus proof time) were about the same on four of the listening typewriter versions as on *DMACH* (Table IV). Only *I1000* differed significantly from both controls, as well as from all listening typewriter versions.

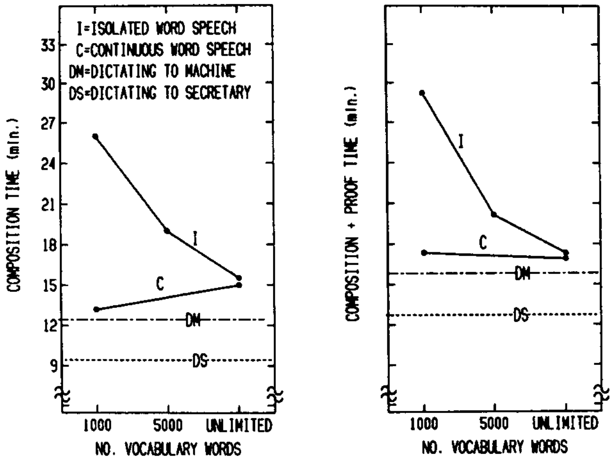

FIGURE 5. Mean Composition Times (left panel) and Mean Total Time (composition time plus proofediting time) for the Composing Methods in Experiment 2.

There were no significant differences in the effectiveness of letters composed with different methods of composition (Table IV), and (unlike in Experiment 1) there was no significant rank-order correlation between composition time and effectiveness (Table IV; $p > .10$). Figure 6 shows the combined score for total time and effectiveness. Effectiveness was calculated by (34.4 − 1.5R), where R varied from 12 for letters receiving no votes to 23 for a letter receiving two first place votes and one second place vote. The trends in Figure 6 for the various listening typewriter versions are similar to those of Experiment 1. *I1000*, *I5000*, and *C1000* all had significantly poorer scores than did *DMACH* and *DSEC* (see Table IV for Duncan range results). *I5000* was not significantly different from any listening typewriter version.

Composing With The Listening Typewriter. Table V shows how participants spent their time while composing with the listening typewriter. Two-thirds of their time was spent planning (range = 58 to 78 percent, depending upon method). Planning time was divided into three subtimes (Table V). Prior to actually dictating a letter, partici-

TABLE IV. Participants' Mean Composing Times (min.) and Mean Effectiveness (arbitrary units) in Experiment 2. (*I* = isolated word speech; *C* = connected word speech; *1* = 1000 word vocabulary; *5* = 5000 word vocabulary; *U* = unlimited word vocabulary; *DMACH* = dictated with dictating equipment; *DSEC* = dictated to a secretary.)

	I1000	I5000	C1000	CU	IU	DMACH	DSEC
Composition time	26.0	18.8	13.3	15.2	15.7	12.2	9.2
(St. Error)	(5.2)	(2.6)	(2.9)	(4.6)	(2.6)	(2.7)	(2.5)
Proofedit time	2.9	1.9	4.4	2.5	1.8	3.8	3.0
(St. Error)	(0.9)	(0.7)	(0.7)	(0.4)	(0.4)	(1.0)	(0.7)
TOTAL TIME	28.9	20.7	17.7	17.7	17.5	16.0	12.2
(St. Error)	(4.8)	(2.7)	(2.5)	(5.2)	(2.9)	(3.2)	(3.1)
Effectiveness	14.8	16.1	16.7	13.6	13.1	10.2	11.7
(St. Error)	(1.9)	(2.0)	(1.4)	(2.2)	(2.9)	(2.2)	(2.4)
Composition time + Effectiveness	20.4	17.5	15.0	14.4	14.4	11.2	10.4
(St. Error)	(2.0)	(0.9)	(1.6)	(2.1)	(1.2)	(1.3)	(1.0)
Total time + Effectiveness	21.8	18.4	17.2	15.7	15.3	13.0	11.9
(St. Error)	(1.9)	(0.9)	(1.5)	(2.3)	(1.3)	(1.4)	(1.2)

Note. Means underlined by the same line in the same row are not significantly different from each other. Means not underlined by the same line in the same row are significantly different from each other at the 0.05 significance level, as measured with Duncan's range test.

pants typically spent about three minutes reading over the description of the letter to be composed, thinking about what they would say, and making notes. While composing, the four participants who made notes spent at least one minute referring to them. The third subtime, other pauses, accounted for the majority of planning time (Table V).

Participants spent 10–18 percent of their composition times actually generating (dictating) their letters. This is a slight overestimate because it includes key words for the editor, for example, "PERIOD," "COMMA." Participants did not appear in the videotapes to spend much time reviewing what they had already said (i.e., reading the screen). These review times may be an underestimate, as we did not include times in this category unless we were certain about them.

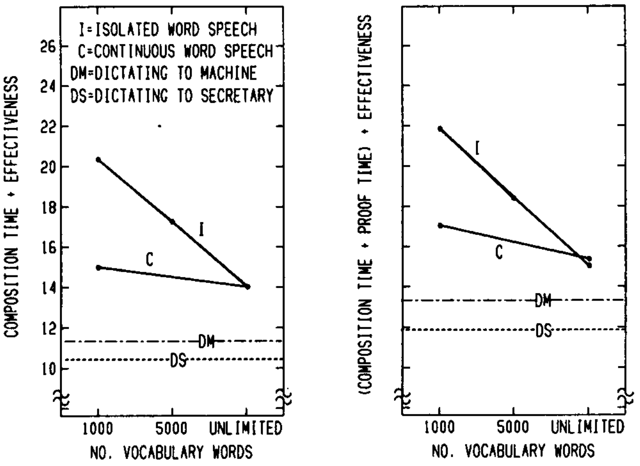

FIGURE 6. Mean Performance Scores, Based Upon Equal Weighting of Time and Effectiveness for each Letter, for the Composing Methods of Experiment 2. The effectiveness scores have been transformed so that, like the time scores, smaller values are better.

TABLE V. Mean Component Times (min.) in Experiment 2. (*I* = isolated word speech; *C* = connected word speech; *1* = 1000 word vocabulary; *5* = 5000 word vocabulary; *U* = unlimited word vocabulary.)

	I1000	I5000	IU	C1000	CU
Planning					
Getting started	3.1	3.7	3.1	2.1	2.4
Looking at notes	1.0	0.9	1.1	0.8	0.8
Pausing	12.0	6.4	8.0	4.1	7.7
Generating	3.3	3.4	1.6	2.1	1.5
Reviewing	0.8	0.4	0.4	0.4	0.4
Editing					
Looking at Editing Instr.	0.1	0.1	0.1	0.1	0.1
Modifying	5.7	3.9	1.4	3.7	2.3
COMPOSITION TIME	26.0	18.8	15.7	13.3	15.2

The remaining component time identified in Table V is editing or revising time. Editing time is the time taken to revise what has already been said. It is the time from when a participant began saying an editing command until the effect of that command was displayed, for example, "NUTS"; or until the participant was done with that command, for example, "SPELLMODE.... ENDSPELLMODE." About 21–28 percent of the time was spent editing with the limited vocabulary methods, and about half that was spent editing with the unlimited vocabulary. Editing, as measured here, is a behavioral activity and does not include the time for much of the cognitive activity (presumably included in planning time) during which a decision is reached whether and how to revise.

The most frequent editing command was "NUTS," as shown in Table VI. More editing was done with limited vocabulary letters than with unlimited vocabulary letters (Table VI), which is consistent with the fact that about twice as much time was spent in editing limited vocabulary letters as unlimited vocabulary letters. Most of the "NUTS" commands were used to erase XXXX's, so that a participant could then spell an unrecognized word. The remainder were used prior to making word substitutions.

TABLE VI. Total Frequency of Use of Editing Commands in the 8 Letters Composed with each Method in Experiment 2. (*I* = isolated word speech; *C* = connected word speech; *1* = 1000 word vocabulary; *5* = 5000 word vocabulary; *U* = unlimited word vocabulary.)

	I1000	I5000	IU	C1000	CU
NUTS	289	87	77	144	54
NUTS 2	9	3	7	13	11
NUTS 3+	10	6	5	22	16
SPELLMODE	194	69	17	108	16
NUMMODE	8	2	3	4	1
CAPIT	8	1	5	1	—

Most often participants erased just one previous word, but occasionally they erased two or more previous words with a single NUTS command (Table VI). There was more erasing with the limited vocabulary methods than with the unlimited methods because participants would erase XXXX's and then spell the word.

When participants used SPELLMODE to spell a word, they usually did so after that word was not recognized and they had erased it with the NUTS command. Sometimes, however, they anticipated that a word would not be recognized, and spelled it in advance. Besides the editing commands shown in Table VI, participants also used the punctuation commands (period, comma, etc.) and formatting commands (INDENT, NEWLINE, NEWPARAGRAPH, etc.).

Composition Rate. As shown in Table VII, letters composed with the listening typewriter tended to have fewer words in them than letters composed with traditional dictation methods, particularly with *DMACH* [F(6,42)=3.84; p < .01; Duncan range test, p < .05]. The difference between *DSEC* and listening typewriters was not statistically significant. Composition rates with *DMACH* and *DSEC* were two-four times faster than composition rates with the listening typewriter [Table VII; *F*(6,42)=16.4; p < .001]. As shown in Table VII, the 15 wpm composition rates for *CU* and *C1000* were about twice as fast as the 8 wpm composition rates for *I1000* and *I5000*. Composition rate for *DMACH* was about that found in earlier experiments on dictation (25.2 wpm) [4].

Participants' Opinions. After using each method, participants compared that method with their favorite method of composition (specified before the experiment). Table VIII shows that *CU* was rated higher than participants' favorite method, whereas *I1000* and *C1000* were rated lower. Participants said they did not like these two methods because they were slow and distracting. *IU* and *I5000* were rated equivalent to their favorite method.

Although these ratings were favorable for three of the five listening typewriter versions, participants' remarks after rating each one were generally not enthusiastic—even for *CU* which received a rating of 6. The gist of most remarks was to point out weaknesses. Their views seemed stronger, more assured, and more authoritative than did those of the participants in Experiment 1. Two negative opinions were voiced strongly by nearly every participant: the discomfort in dealing with XXXX's and the perceived slowness of composing with listening typewriters, compared to *DMACH*. This apparent slowness was perceived to be affected by both (a) speaking in isolated words and (b) use of a limited vocabulary, which introduced uncertainty about whether a word would be recognized and sometimes required spelling it out. Participants reported that these problems sometimes led to losing their train of thought. Thus, these experienced dictators were more critical of the listening typewriters than were the inexperienced dictators.

The most frequently cited advantages of the listening typewriter were seeing what one said and lack of need to spell (recognized words) as one must with traditional handwriting. Some participants volunteered that they had for years secretly worried about spelling words correctly and a listening typewriter relieved them of this concern.

No mention was made about any lack of realism in the experimental environment, simulation, or letters. Although no reference was made directly to a slow response time computer, participants did say that the listening typewriters themselves were slow.

After the experiment, participants rank-ordered the five listening typewriter versions with respect to which "one you would most often use if it were conveniently available to you." Table VIII shows that, based upon the median ranks, the methods were ordered, from most likely to be used to least likely, *CU, IU, I5000, C1000*, and *I1000*. Four participants gave exactly this ranking, and three of the other four ranked the first three exactly this way.

Afterwards, when asked to compose a letter that they needed in their business or personal life, five participants chose to compose with CU, one chose *IU*, one chose *DMACH*, and the remaining participant did not have time to do this task. She indicated she would have chosen *W*, however.

Differences Among Participants. The variance among individual participants' time scores was about half that of Experiment 1. It was 2.5 times the variance among the seven methods of composition. The range of composition time scores was 6.6–34.1 min., and the range of the total

TABLE VII. Word Analyses Based upon Means of each Method in Experiment 2. (*I* = isolated word speech; *C* = connected word speech; *1* = 1000 word vocabulary; *5* = 5000 word vocabulary; *U* = unlimited word vocabulary; *DMACH* = dictated with dictating equipment; *DSEC* = dictated to a secretary.)

	I1000	I5000	C1000	CU	IU	DMACH	DSEC
Number of words in letter	155	146	143	168	152	250	206
Number unrecognized words	39	34	32	(41)	(38)	(63)	(52)
Number XXXX's left for proofediting	15.8	4.8	4.8	. . .	. . .	. . .	. . .
Percent words in 1000 vocabulary	74	(76)	77	(75)	(75)	(75)	(75)
Percent words in 5000 vocabulary	(90)	91	(92)	(90)	(92)	(91)	(91)
Composition Rate (words per min.)	7.5	8.4	14.1	15.8	10.2	24.8	30.1

Note. The numbers in parentheses in lines 2 and 4 indicate the words that would have been affected if the 1000 word vocabulary limitation had been applied to those composition methods. The numbers in parentheses in line 5 indicate the percent of words that would have been affected if the 5000 word vocabulary limitation had been applied to all those composition methods.

TABLE VIII. Participants' Mean Opinions about Different Versions of the Listening Typewriter Used in Experiment 2. (*I* = isolated word speech; *C* = connected word speech; *1* = 1000 word vocabulary; *5* = 5000 word vocabulary; *U* = unlimited word vocabulary; *DMACH* = dictated with dictating equipment; *DSEC* = dictated to a secretary.)

	I1000	I5000	C1000	CU	IU	DMACH	DSEC
Rating re previous favorite method	2.0	3.5	2.5	6.0	4.0	4.0	3.5
Median Rank	5th	3rd	4th	1st	2nd	. . .	. . .
No. participants choosing this method	. . .	. . .	. . .	5	1	1	. . .

Note. Preference rating is based upon participants comparing each method with their previous favorite composing method on a 7-point scale, where 1 = significantly worse than my favorite method and 7 = significantly better than my favorite method. Seven participants composed a letter with a method of their own choice, whereas one did not have time to do this. She indicated that she would have written her letter, however.

TABLE IX. Total Number of Changes Made During Proofediting Stage in Experiment 2 to the 8 Letters Composed with each Method. (*I* = isolated word speech; *C* = connected word speech; *1* = 1000 word vocabulary; *5* = 5000 word vocabulary; *U* = unlimited word vocabulary; *DMACH* = dictated with dictating equipment; *DSEC* = dictated to a secretary.)

	I1000	I5000	C1000	CU	IU	DMACH	DSEC
Formatting; spacing	0	1	1	4	0	6	3
Spelling	2	0	0	2	1	1	0
Punctuation	4	5	5	8	9	22	13
Minor rewording	17	7	28	18	17	44	33
Major rewording	0	1	1	2	1	2	5
Defining XXXX's	127	39	180	0	0	0	0
Capitalizing	2	8	19	31	22	12	9
Homophones	2	0	1	9	4	1	0
Typos made by the simulator	0	0	0	7	1	0	0
System problems	0	0	0	0	0	0	0

time scores was 9.6–38.0 min. As in Experiment 1, participants who took longer to compose their letters had more effective letters (Spearman rho rank-order correlation = −.90; $p < .01$). There was a similar negative rank-order correlation (rho = −.86; $p < .01$) between total time and effectiveness.

Proofediting. Participants spent 2.9 min. proofediting. They made few major changes. They detected fewer than half of the misspellings, leaving spelling errors in 22 of the 56 letters. In letters composed with the limited vocabulary listening typewriters, they had to define 24 XXXX's in *C1000*, 16 in *I1000*, and 5 in *I5000* (Table IX) in the Proofediting stage. Aside from this, they made about seven minor changes (punctuation, capitalization, formatting, homophones, minor rewordings) per letter composed with the listening typewriter, and about 10 per letter composed with *DMACH* and *DSEC* (Table IX).

DISCUSSION

In Experiment 1 participants' performance (combined time and effectiveness) with all versions of the listening typewriter was at least as good as with Writing (Figure 4). This was true even though participants (a) had no experience at using the listening typewriter; (b) had only primitive editing tools; and (c) had never dictated a letter before. In Experiment 2, the performance of experienced dictators was better (but not significantly so) than the performance of the inexperienced dictators in Experiment 1. The experienced dictators performed as well with some versions of the listening typewriter as with traditional dictation methods. The picture that emerges, then, is that all versions of the listening typewriter tested lie, for first-time users, somewhere between the traditional methods of writing and dictating.

Why was this so? A main reason that several versions of the listening typewriter were faster than Written letters was because participants used a Draft strategy with them and used a Final strategy with Written letters (Figure 2). Two reasons why some versions of the listening typewriter were slower than *DMACH* and *DSEC* was because participants often waited with the slower versions of the listening typewriter and spent more time changing what they had already said. Measured in several different ways, participants preferred most versions of the listening typewriter, particularly those with unlimited vocabulary, to traditional methods of composing.

The gap between Written letter times in Experiment 1 (25.3 min.) and *DMACH* times in Experiment 2 (12.2 min.) was greater than expected on the basis of earlier work [4]. That study showed, for experienced dictators, that *DMACH* required about 80 percent of the time that Written letters did (rather than 50 percent as found here).

Why this larger than expected gap? One possibility is that the present study is correct in showing that *DMACH* is twice as fast as *W*—at least for letters of this type. This is certainly understandable given the 500 percent difference in potential output rates (see [4]). Consistent with this notion is that *DMACH* letters were even longer than Written letters.

A second possibility is that the gap between Written and *DMACH* letters is narrower than what the present experiments showed. Perhaps participants in Experiment 1 were simply slower composers than participants in Experiment 2. Consistent with this is that participants in Experiment 1 (mean=22.0 min.) were somewhat slower, on the average, at composing than were participants in Experiment 2 (mean=17.6 min.) on the four versions of the listening typewriter that both groups used in common, although this difference was not significant [$F(1,16)=1.02$;

p>.10]. A similar difference was found for wpm (Tables II and IV). In the earlier work a "within subject" design was used which prevented intergroup differences from arising. If the gap is, in fact, narrower, then either Written letter composition time is overestimated, *DMACH* time is underestimated, or both. Written letter times may be overestimated here since wpm for Written letters (7.9) were less than those found previously for written letters of this type (12.8 wpm in [6]; and 13.0 wpm in [7]). It is also possible, but less likely, that *DMACH* composition time is underestimated here. Participants in the earlier experiments [4] used a Final strategy with both Written and *DMACH* letters. Here they used a Final strategy with Written letters (as instructed), but apparently used a Draft strategy with *DMACH* (where they had a choice of strategies). This is supported from their comments and from the fact that they made about 10 changes per letter in the Proofediting stage with *DMACH* letters (Table IX), which is more than in the earlier experiments on *DMACH* letters [4]. However, this argument of explaining some of the gap between Written and *DMACH* letters in terms of different strategies is weakened by the fact that the composition rate for *DMACH* letters (24.8 wpm) is not much higher than that found previously for *DMACH* letters when participants were using a Final strategy [4].

At any rate, these speculations are somewhat idle because we are talking about composition method means that have large standard errors—much greater than in our previous studies of writing, dictating, speaking, and text-editing.

Knowing how to dictate did not enhance performance with the listening typewriter. We found no qualitative differences in the performances of the two groups of participants on the versions of the listening typewriter which they used in common, and their composing times did not differ significantly [$F(1,16)=1.02$; $p > .10$]. The experienced dictators were much more aware of and concerned about sometimes being slowed down, however.

Wpm composition rate was affected by all three variables studied. Summed over results in both experiments, composition rate was 40 percent faster for unlimited vocabulary than for limited vocabularies (12.0 vs. 8.4 wpm) primarily because participants had to spell 25 percent of the words in the former. It was also faster for consecutive word speech than for isolated word speech (11.8 vs. 8.5 wpm) primarily because participants had to wait between words with the latter. It was also faster for Draft than for Final (Experiment 1; 10.7 vs 8.0 wpm) because participants could leave unspelled words in Draft letters and because they spent less time planning (presumably) their word selection.

The finding that letters composed with a Final strategy take more time than letters composed with a Draft strategy suggests that more time is spent planning them, which is exactly what occurred in Experiment 1. Further, participants spelled all unrecognized words when composing letters with a Final strategy but spelled only some of them when composing letters with a Draft strategy. In addition, the proofedited versions of the letters composed with a Final strategy were judged to be more effective than the proofediting versions of letters composed with a Draft strategy.

The lack of differences in the percent of recognized words with the limited vocabulary listening typewriter, compared to *W*, *DMACH*, *DSEC*, and the unlimited vocabulary listening typewriters (Tables II, V), suggests participants did not (successfully) use a different vocabulary to avoid XXXX's—even though some suggested they were trying to do so.

One might think that people's performance would improve with additional experience in using the listening typewriter and with additional editing capability. Perhaps the limited editing capability had more serious consequences for the versions of the listening typewriter which fared relatively poorly than for those which fared better, since the letters composed with the latter probably required less editing.

Of theoretical interest, this experiment demonstrates what we found in earlier studies of dictation, namely, that people can rapidly learn to use oral language (required in using a listening typewriter) to mimic their written language (produced to be read). This finding shows the adaptability of human language, and indicates that differences between written and oral language are differences in practice and not in principle.

Differences Between The Two Groups of Participants

The experienced dictators, who in real life have responsibility for recommending the characteristics that office products should have, were more critical of the listening typewriter than were the other participants. They felt strongly about the need for a faster system, whereas participants in Experiment 2 said little about this. Both groups disliked the XXXX's. In addition, the experienced dictators viewed using a listening typewriter as dictating, whereas the nondictators of Experiment 1 thought of it as being able to see their writing.

Observations About This Simulation Technique

This simulation was extremely compelling. Once participants began the experiment, they seemed to think about and refer to what they were working with as a real system. No references were ever made to a typist, but rather to "it," or "the system," or "the computer." Midway through the experiment we reminded participants about the simulation and, in case they had not already surmised, revealed the facts about a human typist being involved. Most were surprised; some even tried to explain why this could not be the case. The fact that the typist and the computer program followed consistent rules, especially with the use of homophones, contributed to this feeling. We have simulated user interfaces prior to their being built before (e.g., [8, 15]). When done while a language or a technology is still evolving, such human factors efforts can be supportive to and guide the direction of development. They can provide critical contributions, perhaps gained no other way, to the design of tools for people that will be useful and usable.

Implications for Speech Recognition Systems

The present simulation serves to organize and structure the human factors issues for speech recognition research aimed at developing a listening typewriter. People will probably be able to compose letters with listening typewriters at least as efficiently as with traditional methods. Even with the little editing capability used here, they are preferred to traditional methods. A listening typewriter can lead to productivity increases in the office with a clear displaceable cost because no typing may be required. In addition, it would lead to productivity increases in the organization because faster distribution of documents is possible, due to the document being in machine readable form immediately upon creation. Further, there is undoubtedly some value to an author in having a final typed version as soon as he or she is done composing.

Participants felt *CU* and *IU* were the versions they would most like to have themselves. But unlimited vocabulary is theoretically impossible in a real system. Participants were clear in their dislike for *I1000* and *C1000* for composing letters. Participants felt vocabulary size was more important than speech mode. They felt that *I5000* would be a better system to compose with than *C1000*. The implications of the present experiments, in the absence of data on *C5000*, is that an *I5000* listening typewriter, which participants rate as almost equivalent to their favorite composing method today, would be a good target system to try to build. However, as participants clearly stated, they should be able to dictate more rapidly than they could with the isolated word speech system stimulated in this study.

A note of caution is necessary here. There was a lack of enthusiasm in participants' comments in Experiment 2 about using a listening typewriter. Similarly, while the experiments were going on, no one familiar with them tried to convince us to let them use the system to compose letters of their own. Thus, the evidence presented here should not be taken as convincing that a *I5000* system would be in high demand if it were available. Work needs to be done to determine if a faster *I5000* system would be more in demand.

It appeared from experienced dictators' comments that they saw a major difference between a hit rate of 91 percent and a hit rate of 100 percent. An increase in system speed might reduce some of this attitudinal difference, since speed was a big concern of these experienced dictators. Perhaps the vocabulary could be customized to reduce unrecognized words. Or, perhaps an author's spoken version of each unrecognized word could be stored and replayed upon demand, which would eliminate participants concern about not remembering what they had said. Welch (as cited by [1]) is evaluating image interpreters using a combination of speech recognition and typing (unrecognized words) for vocabularies of up to 1000 words. We suspect that an improved editing system would not have helped participants' performance much, but would have made them feel even more positive toward a listening typewriter.

Some limitations of these experiments are that it was possibly a slow system, had limited editing capability, and studied only one type of letter. Participants did not compose letters in the course of their own work, and we were unable to assess the value to them of having a typed (or final) copy when they completed composing.

This simulation differed in several ways from what a potential listening typewriter might likely be. The required pauses in isolated word speech were longer, by perhaps a factor of 5 or 10, than what they may have to be. In isolated word speech, each word was displayed just after it was said, whereas some possible language models might want to postpone deciding upon what the word just said is until the speaker says a few more words. There were no "false recognitions" (except for typos and for homophones). A word was either "recognized" and printed correctly or XXXXs were displayed. The "processing delay" (the typing time in this simulation) might be different. Editing capability will almost certainly be more powerful with an actual listening typewriter.

CONCLUSIONS

Isolated word speech with large vocabularies may be nearly as good as connected speech systems for a listening typewriter. An imperfect listening typewriter is a potentially useful composition tool. With respect to the hypothesis stated in the Introduction, participants did not compose written letters faster than they composed letters with most versions of the listening typewriter. Participants generally composed letters faster with connected speech than with isolated word speech, and this difference was accounted for mainly by the pauses required with isolated word speech. Quality of the letters were generally the same with all letters, except that letters composed with a first time final strategy were of somewhat higher quality than those composed with a draft strategy. The participants were careless proofediting, regardless of composing method. There were only a few changes in each letter during proofediting except in those letters which contained words that were not recognized during composition. Years of experience at dictating did not lead to significantly faster composition with the listening typewriter, in part perhaps because the experienced dictators sometimes reported being frustrated by being slowed down by listening typewriters.

Acknowledgments. Some of the concepts for simulating a listening typewriter were arrived at in conversations with Evon Greanias. Eric Goldwasser's enthusiasm for studying the value of a listening typewriter and his demonstration of the feasibility of doing simulation experiments on an IBM VM168 computer using the full screen support system influenced our work. He allowed us to borrow freely from his computer simulation program. The work benefited from conversations with Clayton Lewis, and this manuscript is improved over earlier versions because he and Bob Mack provided us with helpful comments. Susan Wandzilak served as typist, displaying much more charm and warmth to us than we could expect from a real system. Jennifer Stolarz assisted with some tedious videotape analyses.

APPENDIX. INSTRUCTIONS FOR USING A LISTENING TYPEWRITER

I. How do I start?
SAY "START" AND BEGIN TALKING

II. How do I change what I said?
ERASE and DICTATE
Say: "Nuts 3" to erase last three words on the screen
Re-dictate

III. How do I punctuate?
USE PUNCTUATION KEYWORDS
Apostrophe
Period
Comma
Semicolon
Colon
Quotation Mark
Question Mark
Right Parentheses
Left Parentheses
Exclamation Point
Hyphen
Number Sign
Percent Sign
Dollar Sign

IV. What if the word I say is not recognized?
ERASE and SPELL THE WORD
Say: "Nuts" to erase that word
Say: "Spellmode" to enter spelling mode
Say: The characters you want, for example, "i" "n" "k" for "ink"
(You do not need to pause between letters when you spell.)
Say: "Endspellmode" to leave spelling mode

V. What if the sound I say is shown by the wrong spelling?
(for example, "sell" rather than "cell")
ERASE and SAY THE WORD AGAIN
Say: "Nuts" to erase that word
Say: The same sound again

VI. How do I capitalize?
USE CAPIT AND SAY THE WORD
Say: "Capit" To capitalize the next word
Say: The word you wish capitalized or
Say: "Capall" to capitalize the entire word

VII. How do I format?
USE FORMATTING KEYWORDS

New Paragraph	Begin a new paragraph
Newline n	Begin a new line n lines down
Indent n	Indent n spaces from the left
Space n	Leave n spaces (plus normal spacing)

VIII. What if I want numbers?
USE NUMBER MODE
Say: "Nummode" to enter number mode
Say: The number you want, for example, "1" "." "9" "5" for 1.95"
Say: "Endnummode" to leave the number mode

IX. What if I want to review part of my letter not shown on the screen?
SAY "SCROLL-TOWARD-BEGINNING" or "SCROLL-TOWARD-END"

REFERENCES

1. Beek, B., Cupples, Ferrante, J., Nelson, J., Woodard, J., and Vonusa, R. Trends and application of automatic speech technology. In Harris, S. (Ed.) *Proceedings of Symposium on Voice-Interactive Systems: Applications and Payoffs.* Dallas, Texas, 1980, 63–72.
2. Doddington, G. R. and Schalk, T. B. Speech recognition: Turning theory to practice. *IEEE Spectrum, 18,* 9 (Sept. 1981) 26–32.
3. Gould, J. D. An experimental study of writing, dictating, and speaking. In Requin, J (Ed.) *Attention and Performance VII.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1978, 299–319.
4. Gould, J. D. How experts dictate. *Journal of Experimental Psychology: Human Perception and Performance. 4,* 4 (1978) 648–661.
5. Gould, J. D. Experiments on composing letters: Some facts, some myths, and some observations. In Gregg, L. and Steinberg, I. (Eds.) *Cognitive Processes in Writing.* Erlbaum and Associates, Hillsdale, N.J., 1980, 98–127.
6. Gould, J. D. Writing and speaking letters and messages. *International Journal Man-Machine Studies. 16,* (1982), 147–171.
7. Gould, J. D. Composing letters with computer-based text editors. *Human Factors. 23,* (1981) 593–606.
8. Gould, J. D. and Boies, S. J. Human factors challenges in creating a principal support office system—The speech filing system approach. IBM Research Report, RC-9768, 1982.
9. Harris, S. (Ed.) *Proceedings of Symposium on Voice-Interactive Systems: Applications and Payoffs.* Dallas, Texas, 1980.
10. Kato, Y. NEC connected speech recognition system. Unpublished manuscript, 1980.
11. Kucera, H. and Francis, W. N. *Computational Analysis of Present-Day American English.* Brown University Press, Providence, Rhode Island, 1967.
12. Moshier, S. L., Osborn, R. R., Baker, J. M. and Baker, J. K. Dialog Systems automatic speech recognition capabilities—Present and future. In Harris, S. (Ed.) *Proceedings of Symposium on Voice-Interactive Systems: Applications and Payoffs.* Dallas, Texas, 1980, 163–187.
13. Poock, G. K. A longitudinal study of computer voice recognition performance and vocabulary size. Naval Postgraduate School Report NPS55-81-013, Monterey, California 93940, 1981.
14. Robinson, A. L. More people are talking to computers as speech recognition enters the real world. *Science, 203,* 1979, 634–638.
15. Thomas, J. C. and Gould, J. D. A psychological study of query-by-example. IBM Research Report, RC-5124, 1974.

CR Categories and Subject Descriptors: H.1. 2 [**Models and Principles**]: User/Machine Systems—*human factors*; H. 4. 1 [**Information Systems Applications**]: Office Automation; I. 2. 7 [**Artificial Intelligence**]: Natural Language Processing—*speech recognition and understanding*
General Terms: Human factors
Additional Key Words and Phrases: listening typewriter

Received 3/82; revised and accepted 12/82

Abstracts from Other ACM Publications

Journal of the ACM April Issue

Analysis of the Search Performance of Coalesced Hashing

Jeffrey Scott Vitter

An analysis is presented of the *coalesced hashing* method, in which a portion of memory (called the *address region*) serves as the range of the hash function while the rest of memory (called the *cellar*) is devoted solely to storing records that collide when inserted. If the cellar should get full, subsequent colliders must be stored in empty slots in the address region and thus may cause later collisions. Varying the relative size of the cellar affects search performance.

The main result of this paper expresses the average search times as a function of the number of records and the cellar size, solving a long-standing open problem. These formulas are used to pick the cellar size that leads to *optimum* search performance, and it is shown that this "tuned" method outperforms several well-known hashing schemes. A discussion of past work on coalesced hashing and a generalization of the method to nonuniform hash functions conclude the paper.

For Correspondence: J. S. Vitter, Dept. of Computer Science, Brown University, Providence, RI 02912.

The Complexity of LALR(k) Testing

Seppo Sippu, Eljas Soisalon-Soininen, and Esko Ukkonen

The problem of testing whether or not a context-free grammar possesses the LALR(k) property is studied. For each fixed integer $k \geq 1$ (i.e., only the subject grammar is a problem parameter) the problem is shown to be complete for polynomial space (PSPACE). For free k (i.e., both the grammar and the integer k are problem parameters) the problem is shown to be PSPACE-complete when k is expressed in unary and complete for nondeterministic one-level exponential time (NE) when k is expressed in binary. The PSPACE-hardness results are obtained by a reduction from the finite automaton nonuniversality problem, whereas the upper bound results are obtained by an economic nondeterministic algorithm that uses only linear space when k is fixed and quadratic space when k is in unary.

The lower bound result for fixed $k \geq 1$ is in contrast with the complexity of testing the membership in several other easily parsed classes of grammars, such as LR(k), SLR(k), LC(k), LL(k), and strong LL(k) grammars, for which deterministic polynomial-time tests are known. The upper-bound results for free k in turn demonstrate how the complexity of the membership testing problems is dominated by k: for k in unary LALR(k) testing is no harder (with respect to polynomial-time reductions) than LALR(1) testing, and for k in binary no harder than, for example, strong LL(k) testing (which is known to be NE-complete).

For Correspondence: S. Sippu, Dept. of Computer Science, University of Helsinki, Tukholmankatu 2, SF-00250 Helsinki 25, Finland.

Computable Error Bounds for Aggregated Markov Chains

G. W. Stewart

A method is described for computing the steady-state probability vector of a nearly completely decomposable Markov chain. The method is closely related to one proposed by Simon and Ando and developed by Courtois. However, the method described here does not require the determination of a completely decomposable stochastic approximation to the transition matrix, and hence it is