RECOGNITION OF ELDERLY SPEECH AND VOICE-DRIVEN DOCUMENT RETRIEVAL

Stephen Anderson, Natalie Liberman, Erica Bernstein, Stephen Foster, Erin Cate, Brenda Levin
Dragon Systems, Inc.
320 Nevada St., Newton MA 02460

Randy Hudson
Intermetrics, Inc.
23 Fourth Ave., Burlington MA 01803

ABSTRACT

We have collected a corpus of 78 hours of speech from 297 elderly speakers, with an average age of 79. We find that acoustic models built from elderly speech provide much better recognition than do non-elderly models (42.1 vs. 54.6% WER). We also find that elderly men have substantially higher word error rates than elderly women (typically 14% absolute). We report on other experiments with this corpus, dividing the speakers by age, by gender, and by regional accent.

Using the resulting "elderly acoustic model", we built a document-retrieval program that can be operated by voice or typing. After usability tests with 110 speakers, we tested the final system on 37 elderly speakers. Each retrieved 4 documents from a database of 86,190 Boston Globe articles, 2 by typing and 2 by speech. We measured how quickly they retrieved each article, and how much help they required. We find no difference between spoken and typed queries in either retrieval times or in amount of help required, regardless of age, gender, or computer experience. However, users *perceive* speech to be substantially faster, and overwhelmingly prefer speech to typing.

1. INTRODUCTION

Senior citizens are one of the last groups to benefit from access to computers. There are many reasons for this, from not having used computers at work, to a reluctance to use new technology, to physical difficulty in using the keyboard and mouse. Speech recognition can address this last concern; seniors with arthritis or poor typing skills can control their computers by voice, entirely bypassing the keyboard and mouse.

One of the most useful aspects of the internet is browsing for information. Since browsing does not require a great deal of typing, and is particularly robust to speech recognition errors [1], it seems like a natural first computer step for seniors who have difficulty with the keyboard. With this in mind, we built a speech-driven document retrieval engine, *GLOBE*, for a database of 86,190 articles from the *Boston Globe*. (Of course, it could function as a front end to any database, including the internet itself.)

Current speech recognition systems have difficulty working with elderly voices [2]. To address this problem, we recorded and transcribed 78 hours of elderly speech from 297 speakers. With this data, we built an acoustic model which substantially improved recognition accuracy on elderly voices. We report on experiments involving gender, age, and regional accent in section 2.

In field trials with the *GLOBE* program, we asked 37 speakers to retrieve two articles from our database of *Boston Globe* articles by voice and two by typing. We timed the searches, to determine whether spoken or typed queries are faster. We also noted how many times the user needed help ("cues") from the interviewer, in order to assess which interface was more difficult to use. We report these results in section 3.

We conducted a "usability" trial of the system prior to the field test. 30 in-house speakers and 80 elderly external speakers used five successive versions of the program to optimize the task of document retrieval by voice. In section 3.5 we report on the features that seniors found useful.

2. ELDERLY SPEECH RECOGNTION

2.1 Elderly Speech Collection

Before asking elderly users to use our voice-driven *GLOBE* program, we wanted to be sure that recognition accuracy would be as high as possible. Our first goal, therefore, was to build an acoustic model from elderly voices.

2.1.1 Training Set

We recorded 297 elderly speakers (186 females, 111 males) in senior centers in greater Boston and Boca Raton, Florida, for a total of 78 hours. The average speaker age was 79.0 (78.8 male, 79.2 female). Our distribution of speech by age and gender is shown in Figure 1.

Roughly 1/3 (32.7%,) of the speech was collected near Boston, where most speakers had Boston accents. The other two thirds of the speech was recorded in Boca Raton, Florida, where most of the speakers had New York accents.

A small fraction (8%) of the speech consisted of read Wall St. Journal articles, but most was "evoked monologues" in which the speakers would tell stories in response to questions. The transcripts included many word fragments, hesitation sounds, and breath/mouth noises.

2.1.2 Elderly Test Sets

A subset of 40 speakers (7.5 hours) was chosen as a "test set", with the remaining 257 speakers (71.5 hours) constituted the "training set". The sets were matched by gender, age, and Boston/Florida collection site. The first 2 minutes of each test

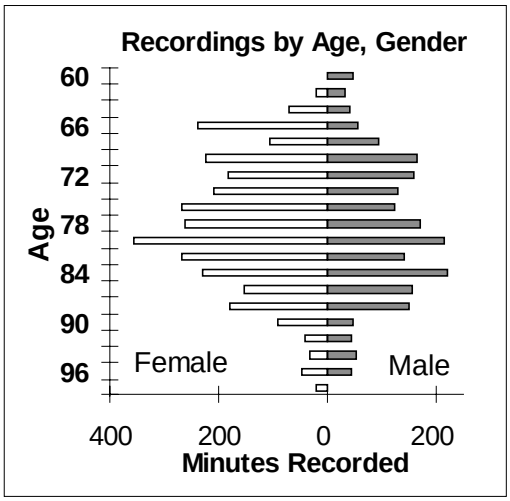

Figure 1: Elderly speech recorded by gender and age

speaker's data was used for vocal-tract length normalization, and the first 5 for speaker adaptation.

The 35,933 words in the test set included 4,033 breath/mouth noises, 597 word fragments, and 342 out-of-vocabulary words.

2.2 Elderly Speech: Recognition Experiments

2.2.1 Effects of Age

Other authors [2] have found that elderly speech is poorly recognized with models trained from non-elderly voices. To test this, we built an "elderly acoustic" model from our training data and recognized the test speakers both with this model and with a "non-elderly" acoustic model. The "non-elderly" model was built from 80 hours of Wall Street Journal data and 30 hours of in-house read speech [3]. All acoustic models are vocal-tract normalized triphone models. There are a total of 6300 output distributions, each of which has up to 6 multivariate gaussian components.

The LM was built from general English [3], combined with the training speakers' text.

Table 1 shows speaker independent/dependent results for elderly test speakers using each model:

Training	Elderly Test Speakers		
	Female	Male	All Elderly
non-elderly	59.7 / 49.9	74.1 / 63.4	64.7 / 54.6
elderly	43.6 / 37.2	58.9 / 51.0	49.0 / 42.1

Table 1: WER (%) for elderly test speakers (unadapted/ adapted) using elderly and non-elderly acoustic models.

We see that elderly speakers are recognized much better with models built from elderly speech than from non-elderly (42.1% vs. 58.9% WER). (However, the elderly acoustic models were trained primarily on *spontaneous* speech, while the non-elderly models were trained from *read* speech. We investigate this further in section 2.2.4.) The striking result in Table 1 is that elderly men's speech has a much higher WER than elderly women's speech (14.2% worse using an elderly acoustic model, 13.5% worse with a non-elderly model). The speaker-independent (unadapted) results are similar.

To look at the effects of age *within* the elderly speakers, we divided both the training and test speakers by age; those above and below 79 years old (the average age). Table 2 shows that there is not a large age effect *within* the elderly population.

Training	Test Speakers		
	< 79 years	> 79 years	All
< 79 years	43.3	42.5	42.9
> 79 years	44.5	42.5	43.3
all elderly	42.8	41.3	42.1
non-elderly	53.6	55.7	54.6

Table 2: WER (%) for adapted "younger senior" (< 79 years old) and "older senior" (>79 years old) test and training.

2.2.2 Gender-Dependent Models

To try to improve the recognition for elderly men, we built gender-dependent models from the elderly training data. Table 3 shows the speaker-adapted results of using these models. (As in all our experiments, both test and training speakers are vocal-tract length normalized).

Training	Test Speakers		
	Female	Male	All
female	37.6	52.8	42.8
male	42.3	50.6	45.2

Table 3: WER (%) with gender-dependent elderly models.

Table 3 shows that gender-dependent models are not the solution to poor recognition of elderly male voices.

2.2.3 Effect of Regional Accent

To assess the importance of regional accent, we built "accent-dependent" models from the training data collected in Boston and Florida (NY accents). Table 4 shows the speaker-adapted results.

Training	Test Speakers		
	Boston	NY	All
Boston	41.5	48.6	45.4
NY	44.1	43.3	43.6
All	40.4	43.4	42.1

Table 4: WER (%) for Boston and NY accented models.

We see in Table 4 that regional accent is important, with mismatches causing a 3-5% degradation in recognition.

2.2.4 Relative Effects of Age and Speaking Style

In a final experiment, we try to assess the relative effects of age and speaking style. We construct a second test set, of "elderly read" speech, to compare to our "elderly spontaneous" set. We recognized this set with both our elderly (spontaneous) and non-elderly (read) speech models.

The test set of elderly *read* speech came from the enrollment data of the *GLOBE* study (see section 3 below). 82 speakers read 5 minutes each, for a total of 6.8 hours. With only 5 minutes/speaker, we could not do speaker-adapted recognition.

Training	Test Set	
	Eld. (Spont.)	Eld. (Read)
elderly (spont.)	49.0	42.0
non-elderly. (read)	64.7	46.7

Table 5: WER (%). Effects of age and speaking style (unadapted).

The first column of Table 5 indicates that the *combined* effects of age and speaking style are large (15.7%), but we cannot conclude that age *by itself* is important. However, column 2 shows that elderly read speech is better recognized with an elderly spontaneous model than with a non-elderly read model, indicating that age is indeed important.

3. *GLOBE* FIELD STUDY

3.1 Participants

To test *GLOBE*, our voice-driven document retrieval program, we contacted Boston-area senior citizen centers with active computer clubs for potential subjects over a 3 month period. 37 seniors participated. The distribution of speakers by sex and age group is given in Table 6:

M/F	56-60	61-65	66-70	71-75	76-80	81+	Tot
M	0	5	5	6	0	0	16
F	1	6	4	7	1	2	21
Tot	1	11	9	13	1	2	37

Table 6: *GLOBE* study participant age and gender.

The average computer experience was in the 1-12 month range (24% < 1 month, 24% 1-12 months, 52% >12 months). The most popular computer uses were word processing, financial software, games, and the internet.

3.2 Document Retrieval Test Design

Participants began by completing a short questionnaire about their previous computer experience. They were then asked to read a 5-minute text to adapt the acoustic model to their voice. After enrollment, the interviewer instructed them in the use of the *GLOBE* retrieval program.

The user then practiced by retrieving two articles, one by speech and one by typing. The lead interviewer then handed them a series of 4 printed articles, and ask them to retrieve each from the database. Each speaker retrieved the same 4 articles, with even-numbered speakers retrieving #1 and #3 by speech and #2 and #4 by typing, and odd-numbered speakers doing the reverse. This balanced the differences in article retrieval difficulty, and the effects of learning.

The entire test took roughly 1 hour to complete. Participants filled out a post-test questionnaire assessing their perceptions of using speech *vs.* typing to achieve the same tasks. Which mode did they think was faster? Which did they prefer?

3.3 The GLOBE Program

The *GLOBE* program allows users to retrieve articles from a database consisting of 21 months (January 1996-September 1997) of *Boston Globe* articles (86,190 articles). All commands can be issued by voice or typing.

The screen has 5 sections:

1. All currently available commands are displayed.
2. Article titles are displayed. When a particular article is selected, the full text is displayed.
3. The recognition "*n-best*" results are displayed.
4. A space for two icons: "Mic On" and "Retrieving Articles".
5. A VU-meter, indicating mic volume and on/off status.

To use commands by voice, a user simply says the command displayed in section (1). To execute a command by typing, the user clicks on the command name. In the case of the "*Find Articles About*" and "*Display Article n*" commands, which require arguments, appropriate dialog boxes appear.

3.4 *GLOBE* Document Retrieval Results

3.4.1 Retrieval Times

The most striking result is that there is no significant difference in retrieval times using speech or typing. A comparison of average retrieval times is shown in Figure 2.

To test for a significant difference between typing and speech, we introduce the variables T=average retrieval time, C = average

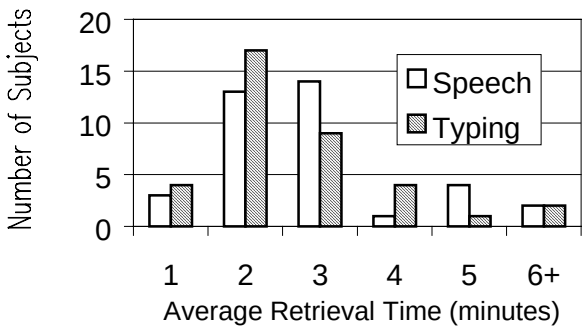

Figure 2: Retrieval Times for Speaking and Typing.

number of cues, $\Delta T = T_{speech} - T_{typing}$, $\Delta C = C_{speech} - C_{typing}$.

To determine whether speech or typing is faster, we test $\Delta T = 0$. We find no significant difference in times (t-test: P=0.626). Using a non-parametric Wilcoxon test also yields no significant difference (P=0.402). Similar results hold for the amount of help (number of cues) users required with both modes. Testing $\Delta C = 0$, there was no significant difference between typing and speech (t-test: P= 0.711, Wilcoxon test: P= 0.654).

Do factors such as age, gender, and previous computer experience predict whether speech will be faster or easier than typing? Doing a linear regression with these three variables yielded no significant correlations. More computer experience led to shorter retrieval times by voice *and* typing, but not to a difference *between* speech and typing.

3.4.2 Subjective Assessments

Participants overwhelmingly reported the perception that speech

is faster than typing for retrieving articles, and strongly preferred speech. Figure 3 presents a histogram of the post-test questionnaire.

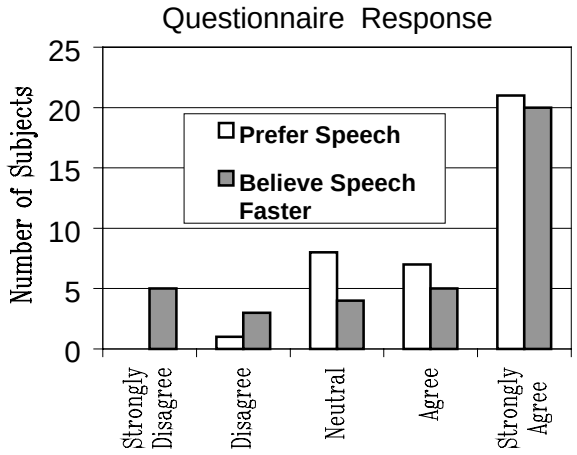

Figure 3: User Questionnaire Response.

3.5 Usability Issues

Before our final set of 37 speakers, we redesigned the *GLOBE* retrieval program five times based on feedback and observations of 30 in-house speakers and 80 elderly speakers. We observed that if subjects suffered less frustration and fatigue with the interface, they had higher recognition accuracy. Early usability testers faced a confusing interface, and blamed their difficulties on poor recognition. Later users, *using the same recognition engine*, enjoyed the program and were particularly complimentary about the recognition accuracy.

We discovered the following usability issues for seniors, listed according to the categories of Dumas & Redish [4]:

Speaking the Users' Language

Users issued spoken commands in their own words, even though the required syntax was always displayed on the screen. To accommodate this, we loosened the command grammar. For example, "*Show me the next page*" and "*Next page, please"* became synonyms for "*Show next page"*. Similarly, error messages had to be rewritten in everyday English.

Providing Clear Feedback

Many subjects commented on the need for clearer indications of what the program was doing. We therefore:

1. Added large icons to indicate "mic on" and "retrieving articles".
2. Used "earcons" (distinctive sounds) to indicate mic turning on/off, and to announce error messages.

Measures to Prevent User Mistakes

We needed to increase the end-of-utterance pause length to 1.5 sec to allow for some "thinking time" while voicing a command. We introduced oversize fonts, extra-large buttons, and minimized the amount of text displayed. Finally, we added a "push-to-talk" button to allow speakers to intersperse conversation with commands.

4. CONCLUSIONS

We have collected a new corpus of 78 hours of elderly speech from 297 speakers. We find that elderly speech is better recognized with "elderly" acoustic models. Within the elderly population, however, we see only a weak age effect; users 65-79 and 79+ could use models trained on the other group with little loss of accuracy. We find that elderly men have much worse (typically 14% absolute) recognition than elderly women, and that building gender-dependent models does not overcome this difficulty. We find that regional accent (Boston *vs.* New York) is an important effect.

We tested a voice-driven document retrieval program, *GLOBE*, among elderly users. 37 study participants were asked to retrieve database articles by voice and by typing. Interestingly, *all* speakers were able to retrieve documents by voice, while some could not type well enough to succeed using the keyboard. Anecdotally, we observed that command-and-control recognition accuracy improved noticeably for the 37 *GLOBE* test speakers, who used the "elderly" model, compared to the 80 "usability" speakers who used the "non-elderly" model.

There was no significant difference in the amount of time required to retrieve articles by speech or typing, and no difference in the amount of help needed. Users overwhelmingly preferred speech as a medium, and (contrary to objective measurements) perceived speech to be faster. We conclude that using speech for information retrieval is as fast as typing, and preferable for many users.

5. ACKNOWLEDGMENT

We wish to thank NIST-ATP for supporting this work through grant 70NANB5H1181.

6. REFERENCES

[1] Barnett, J. et al. "Experiments in Spoken Queries for Document Retrieval", *Proceedings of the 5th European Conference on Speech Communication and Technology*, Rhodes, Greece, Sept. 1997, p. 1323.

[2] Wilpon, J. G., and Jacobsen, C. N., "A Study of Speech Recognition for Children and the Elderly", *IEEE International Conference on Acoustics, Speech, and Signal Processing.* Atlanta, May 1996, p. 349.

[3] Roth, R. et al. "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer". *Proc. ARPA Spoken Language Systems Tech. Workshop*, Austin, 1995, p. 116.

[4] Dumas, J.S. and Redish, J.C. *A practical guide to usability testing*. Norwood, NJ: Ablex Publishing Corporation, 1994.