A Multimodal Interface for the Immediate Transcription of Radiology Dictation

Wu-chun Feng
Los Alamos National Laboratory
University of California
Los Alamos, NM 87545
feng@lanl.gov

Abstract

We present the design and implementation of an integrated multimodal interface that delivers instant turnaround on transcribing a radiology dictation. This instant turnaround time virtually eliminates a hospital's liability with respect to improper transcriptions of oral dictations and all but eliminates the need for transcribers. The multimodal interface seamlessly integrates three modes of input – speech, handwriting, and written gestures – to provide an easy-to-use system for the radiologist.

1. Introduction

Although computers have quickly become an essential part of today's society, their ubiquity has been stymied because many still find the computer "unnatural" (and even difficult) to use. While scientists and engineers take their computer skills for granted, a large number of potential users still have limited experience in using a computer. To make computers (or products with embedded computers, e.g., an automobile) easier and more natural to use, manufacturers have proposed the use of a speech recognition system. Even for computer-savvy users, speech can be used to boost productivity because nearly everyone can talk faster than they can type, typically more than 200 words per minute (wpm) versus 50 to 75 wpm. However, speech recognition is never perfect; recognition errors are made. In order to correct these errors, the end user currently uses a keyboard and mouse.

Instead, we propose a system that seamlessly integrates speech, handwriting, and written gestures and provides a natural multimodal interface to the computer. To ensure that the interface is easier to use than a "keyboard-and-mouse" interface, the speech recognizer must have a high recognition rate, e.g., 95%, and the handwriting and gesture recognizers should provide nearly error-free recognition of stylus-inputted handwriting and gestures, respectively, to correct errors made by the speech recognizer. These corrections can then be applied to the speech recognizer itself to improve future recognition.

Initially, we target our integrated system at the radiologist. Today, a radiologist dictates x-ray diagnoses into a tape recorder because, as busy as his schedule is, he does not have the time to type each analysis as he is reciting it. He continues to record x-ray analyses onto a cassette tape until it is full and then gives the tape to a transcriber who then types the radiologist's analyses into hardcopy reports. Unfortunately, by the time a given report reaches the radiologist's desk for his signature, 24 to 48 hours (and potentially hundreds of other x-ray analyses) have passed. Because of the slow turnaround, the radiologist cannot be expected to remember exactly what he said and for which x-ray. Consequently, this increases the liability to the hospital because the radiologist must place blind faith in the transcriber's competence. So, if the transcriber inadvertently types "left ventricle" instead of "right ventricle" because the last several transcribed analyses referred to the *left* side of the body, the radiologist is held accountable because his signature appears on the report, not the transcriber's. Our proposed system (which seamlessly integrates different modes of input, i.e., speech, handwriting, and written gestures) eliminates the 24- to 48-hour turnaround time by allowing the radiologist to dictate his analysis into a speech recognizer and then quickly correct any errors, or even re-organize the report, using the handwriting and written-gesture recognizers.

2. Recognition systems

While there exist several commercial speech-recognition products (e.g., IBM's Via-Voice and Dragon System's Naturally Speaking) as well as gesture and handwriting-recognition products (e.g., Palm Pilot and Apple's now-defunct Newton), none have *seamlessly* integrated all three aspects, i.e., speech, handwriting, and written gestures, into a single package. The above products are standalone technologies that could benefit from appropriately leveraging the other technologies, as we show in this paper.

For instance, personal digital assistants (PDAs) like the Palm Pilot work wonderfully as daily organizers, but beyond that, they are not that natural or efficient to use, particularly for anything involved like radiology dictation or e-mail. A speech recognizer on such a device has the potential to greatly increase productivity. (Who really wants to use the error-prone Graffiti alphabet of the Palm Pilot to slowly write an email with a stylus when it can be quickly dictated with a speech recognizer?)

2.1. Speech recognition

To ensure accurate, but still high-speed, speech recognition, we use a more "primitive" recognizer based on discrete speech (i.e., talking like a robot) rather than one based on continuous speech. Though continuous speech is what we are used to, consciously placing a nearly imperceptible pause between spoken words can improve recognition accuracy by as much as 30%, thus reducing the need to make corrections after dictation. In addition, to allow the speech recognizer to operate in environments with significant ambient noise, e.g., the "garage-door" radiology room of the hospital that we tested this system, we use a directional microphone to minimize the pick-up of background noise and a bandwidth-pass filter in the speech recognizer itself to catch and eliminate any low- and high-frequency noise that manages to sneak-in through the microphone.

2.2. Handwriting and written-gesture recognition

Our handwriting and gesture recognizers provide real-time feedback to the radiologist, i.e., the writing is recognized as it is being written [1,4]. To enable real-time response, the system captures the temporal (or dynamic) information as the radiologist writes. Examples of temporal information include the *number* of strokes, the *order* of the strokes, the *direction* of the strokes, and the *speed* of the strokes (where a stroke is defined as the writing from the pen-down position to pen-up position). Although this information complicates the recognition process, it substantially improves recognition accuracy because a user's writing variations may not be apparent in static images of the writing. For example, the letter *M* may be written with one to four strokes in a variety of stroke orders or directions but appears the same when completed.

In our handwriting recognizer, handwriting strokes make up alphanumeric characters and symbols that are in a fixed orientation and represent that character or symbol. While technology currently exists for cursive-script recognition of alphanumeric characters and symbols; our integrated speech, handwriting, and gesture prototype uses a handwriting recognizer based on printing in order to provide faster and nearly error-free recognition.

The recognition of hand-generated symbols can be generalized to handle symbols other than handwritten ASCII characters. Examples include shorthand, editing symbols, and flow-chart symbols [2, 3]. With handwriting, the recognition of written ASCII characters depends on shape consistency (i.e., the difference between the same symbol produced at different times is less than the difference between different symbols). However, this approach does not necessarily work for written-gesture recognition because there are symbols that violate shape consistency yet are still recognizable as the same symbol by the human eye. For example, the human eye recognizes the commonly used arrow symbol, as shown in Figure 1, even though it may vary in size and orientation.

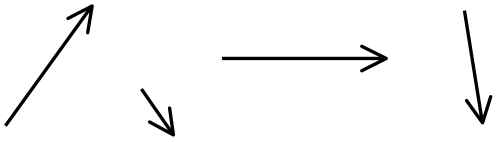


Figure 1. Sample arrows

Since the difference measures and feature analyses for handwriting recognition do not always work for gesture recognition, we introduce new measures and analyses to account for the following gestural variations: non-linear scaling, rotation, and direction reversal.

The vocabulary of our gesture recognizer consists of a *gesture alphabet* and *gesture commands*. Unlike the handwriting alphabet, the gesture alphabet has no fixed set of gestures defined and used by the general population. Based on a paper-and-pencil study of gestures [5], a representative set of gestures for an initial gesture alphabet is shown in Figure 2.

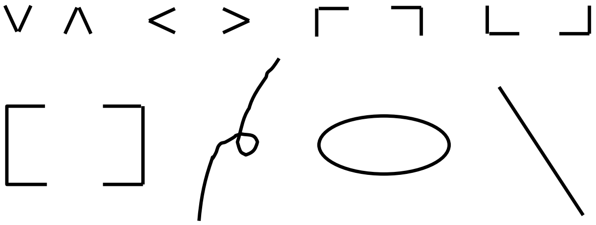
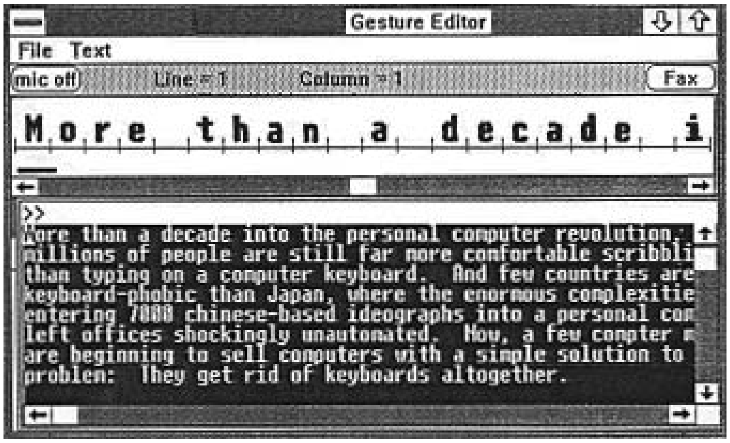


Figure 2. Basic gesture alphabet

Without any context, a gesture by itself does not necessarily have any meaning. In addition, several gestures may be combined in sequence, similar to the way that characters are put together to form a word. For instance, a circle gesture can be used to select a word, followed by a delete gesture to remove the selected word.

3. A multimodal interface for radiology transcription

A text editor serves as the integration point for our multimodal interface for transcription. The editor (via our speech recognizer) takes speech as its initial input. The speech recognizer then converts the spoken words into visible characters and words on a liquid-crystal display (LCD) or into editing commands (like those found in the Unix `vim` or `emacs` text editors). Handwriting, gestures, and even speech, can then be used to correct any speech-recognition errors that occurred during dictation.



```
M o r e   t h a n   a   d e c a d e   i
```

```
>>
More than a decade into the personal computer revolution,
millions of people are still far more comfortable scribbli
than typing on a computer keyboard.  And few countries are
keyboard-phobic than Japan, where the enormous complexitie
entering 7000 chinese-based ideographs into a personal com
left offices shockingly unautomated.  Now, a few compter m
are beginning to sell computers with a simple solution to
problem:  They get rid of keyboards altogether.
```

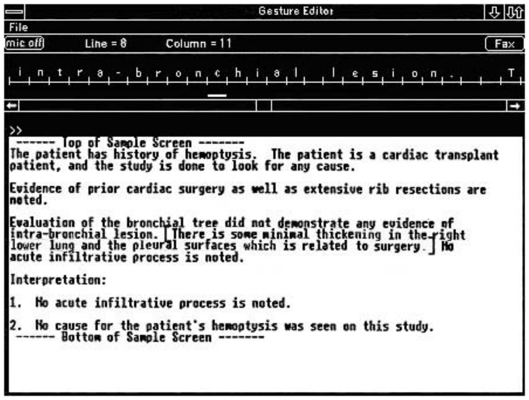
Figure 3. Multimodal interface for transcription

Figure 3 shows a screen dump of our multimodal interface, running on a low-resolution prototype of a tablet PC with a slow 100-MHz processor to demonstrate that the multimodal interface could eventually be run on a web pad or even a PDA. Starting from the top of the interface and working downward, the interface consists of a title bar, menu bar, status bar, handwriting/gesture window, and speech/gesture window. To begin dictation, the user must first turn on the microphone by tapping on the "mic off," thereby changing the status to "mic on." As the user dictates, the recognized words appear in the speech/gesture window. When the user wants to stop dictating, he taps on "mic on" to change the status back to "mic off."

The user may then correct speech-recognition errors using speech in the speech/gesture window, handwritten ASCII characters in the handwriting/gesture window, or the following "selection gestures" in either window — [,], ┌, ┘, ○ — coupled with one of the following "editing gestures": ɕ, ?, →, ╲.

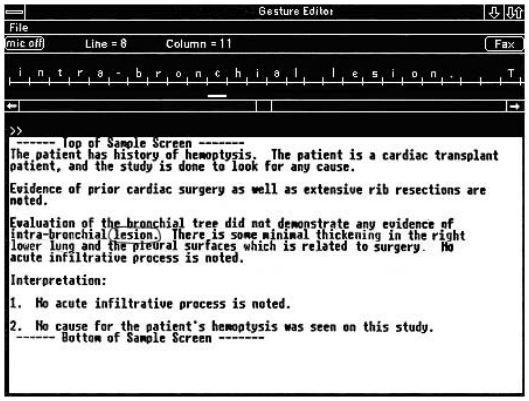
The selection gestures choose (i.e., "mark" in reverse video) a region of text. For example, Figure 4 shows how [and] are used to select a linear sequence of text, starting before the word "There" and ending after the word "surgery." (Figure 6 shows the resulting selected text in reverse video.) Similarly, the corner gestures, i.e., ┌ and ┘, select a linear sequence of complete lines, and the circle gesture quickly selects a word or set of words, as shown in Figure 5.

ɕ, ?, →, ╲ are gestures for editing regions of text that have been marked (in reverse video) by one of the selection gestures; hence, we denote them to be editing gestures. The delete (ɕ) gesture removes a region of selected text, e.g., the previously selected text from Figure 4 that now appears in reverse video in Figure 6. Using the question-mark (?) gesture on a given word pops up a drop-down menu of alternative words that the speech recognizer had considered before selecting the given word. When a word is misrecognized, oftentimes the correct word will be found in this alternative-word list. Figure 7 illustrates how the arrow (→) gesture is used to move selected text, e.g., the word that was selected in Figure 5. (It could have also been used to move, instead of delete, the selected text from Figure 4.) Alternatively, when only one word needs to be deleted, rather than selecting and then deleting the word, the user can simply point to the word and say, "Delete this." Finally, the line (╲) gesture can be used like the arrow gesture to move selected text, but it can also be used in a couple of other ways when text has not been selected. First, when drawn as a horizontal line over a word or sequence of words, the text editor deletes those words. Second, when drawn as a vertical line (typically between two words), the editor breaks the line of text at the vertical line and inserts a carriage return, thus splitting the line of text into two.



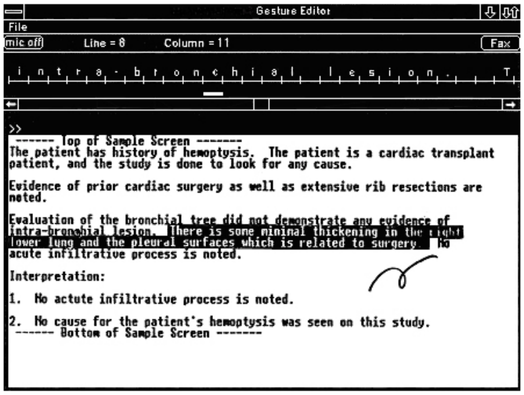
```
 ------ Top of Sample Screen -------
The patient has history of hemoptysis.  The patient is a cardiac transplant
patient, and the study is done to look for any cause.

Evidence of prior cardiac surgery as well as extensive rib resections are
noted.

Evaluation of the bronchial tree did not demonstrate any evidence of
intra-bronchial lesion. [There is some minimal thickening in the right
lower lung and the pleural surfaces which is related to surgery.] No
acute infiltrative process is noted.

Interpretation:

1.  No acute infiltrative process is noted.

2.  No cause for the patient's hemoptysis was seen on this study.
 ------ Bottom of Sample Screen -------
```

Figure 4. Bracket gestures:
Selecting a linear sequence of text



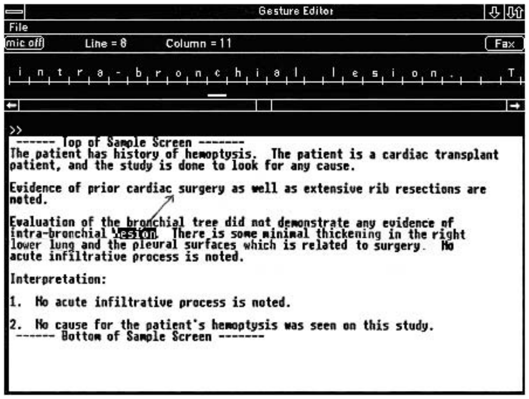
```
 ------ Top of Sample Screen -------
The patient has history of hemoptysis.  The patient is a cardiac transplant
patient, and the study is done to look for any cause.

Evidence of prior cardiac surgery as well as extensive rib resections are
noted.

Evaluation of the bronchial tree did not demonstrate any evidence of
intra-bronchial lesion.  There is some minimal thickening in the right
lower lung and the pleural surfaces which is related to surgery.  No
acute infiltrative process is noted.

Interpretation:

1.  No acute infiltrative process is noted.

2.  No cause for the patient's hemoptysis was seen on this study.
 ------ Bottom of Sample Screen -------
```

Figure 5. Circle gesture:
Selecting a region of word(s)



```
 ------ Top of Sample Screen -------
The patient has history of hemoptysis.  The patient is a cardiac transplant
patient, and the study is done to look for any cause.

Evidence of prior cardiac surgery as well as extensive rib resections are
noted.

Evaluation of the bronchial tree did not demonstrate any evidence of
intra-bronchial lesion.  There is some minimal thickening in the right
lower lung and the pleural surfaces which is related to surgery.  No
acute infiltrative process is noted.

Interpretation:

1.  No actute infiltrative process is noted.

2.  No cause for the patient's hemoptysis was seen on this study.
 ------ Bottom of Sample Screen -------
```

Figure 6. Delete gesture:
Removing the selected region of text



```
 ------ Top of Sample Screen -------
The patient has history of hemoptysis.  The patient is a cardiac transplant
patient, and the study is done to look for any cause.

Evidence of prior cardiac surgery as well as extensive rib resections are
noted.

Evaluation of the bronchial tree did not demonstrate any evidence of
intra-bronchial lesion.  There is some minimal thickening in the right
lower lung and the pleural surfaces which is related to surgery.  No
acute infiltrative process is noted.

Interpretation:

1.  No acute infiltrative process is noted.

2.  No cause for the patient's hemoptysis was seen on this study.
 ------ Bottom of Sample Screen -------
```

Figure 7. Arrow gesture:
Moving a selected region of text

Two other miscellaneous, but key, gestures are > and U. The > gesture is for the *playback* of the user's dictation. If > is invoked with a marked region, then the marked region is played back for the radiologist to listen to. Otherwise, the region around the current position of the cursor is played back for the radiologist. Lastly, the U gesture will *undo* the most recent action.

4. Case study

To assess the usability of our multimodal interface, we deployed and tested our integrated system at a leading academic research hospital. For the period of a week, five radiologists graciously served as our test subjects. All the test subjects were middle-aged. Four were effectively native English speakers; one had a very thick European accent.

Prior to introducing the test subjects to our system, each subject went through 30-45 minutes of speaking reference text (using discrete speech) into our speech recognizer. This exercise served to enhance speech recognition (i.e., minimize recognition errors) by training the speech recognizer to learn the speech patterns of each test subject, hence tailoring the system for each test subject. In addition, we loaded the speech database with a radiology-specific vocabulary to significantly improve recognition speed.

We also had to instruct the subjects on how to use the speech recognition system. First, to ease the learning curve, we told them to complete their oral dictation *before* going back to correct errors. Second, punctuation marks had to be explicitly noted, i.e., the radiologist has to *say* the following sentence – "Evidence of prior cardiac surgery as well as extensive rib resections are noted … period … new paragraph …" – in order to get the text to appear as shown in Figures 4-7. Third, special lists had to be explicitly noted, e.g., "new bulleted list … end bulleted list" or "new numbered list … end numbered list."

With the radiology dictations occurring in rooms where entry was via a loud "garage door," the overall speech-recognition rates for the radiology dictations were surprisingly high – 97%, 97%, 95%, 94%, and 91%, respectively, for the five test subjects. After oral dictation, the radiologists would then spend a couple of minutes to proofread and correct errors. Towards the end of the week, radiologists became proficient enough that they started to use the system in a freeform manner, e.g., dictate one paragraph, correct errors (if necessary), dictate a second paragraph, correct errors (if necessary), and so on.

On a qualitative note, all the radiologists thought that the system had immediate impact (even the European whose disposition was clearly "anti-technology") and were generally pleased with its usability. The hospital administrator, on-hand for several of the tests, was ecstatic with the *immediate turnaround time*. As soon as the radiologist completed dictation and corrected any errors, he printed the report and signed off on it. However, the radiologists also brought up a number of issues, the most important of which were (1) the hassle of having to remember to load one's speech profile before starting, (2) the need to speak like a robot, and (3) consciously having to remember to turn "on" and "off" the microphone. The first two issues could be immediately addressed by eliminating the speakers' profiles and using continuous speech input, but it would reduce the speech recognition rate considerably (by as much as 20%-30%).

5. Conclusion

We presented an integrated system for the immediate transcription of radiology dictation. This system seamlessly integrates speech, handwriting, and gesture recognition systems to provide a natural multimodal interface to the computer. Future work includes enhancing the usability of the multimodal interface and integrating still images of x-rays into the text editor and then using the gesture recognizer to annotate the images. Gestures could also be linked to other tasks such as creating a WWW link from a word or phrase in the document to a still image that resides out in the Internet or on the local machine.

6. References

[1] J. Kim, "On-Line Gesture Recognition by Feature Analysis," *Proc. of Vision Interface*, 1988.

[2] C. G. Leedham and A. C. Downton, "On-Line Recognition of Shortforms in Pitman's Handwritten Shorthand," *Proc. of the 7th Int'l Conf. on Pattern Recognition*, July 1984.

[3] T. Sakai, K. Odaka, and T. Toiba, "Several Approaches to Development of On-Line Handwritten Character Input Equipment," *Proc. of the 7th Int'l Conf. on Pattern Recognition*, July 1984.

[4] C. C. Tappert, C. Y. Suen, and T. Wakahara, "On-Line Handwriting Recognition – A Survey," *Proc. of the 9th Int'l Conf. on Pattern Recognition*, November 1988.

[5] C. G. Wolf, "Can People Use Gesture Commands?" *ACM SIGCHI Bulletin*, October 1986.