Speech and Text-Image Processing in Documents

Marcia A. Bush

Xerox Palo Alto Research Center
3333 Coyote HIll Road
Palo Alto, CA 94304

ABSTRACT

Two themes have evolved in speech and text image processing work at Xerox PARC that expand and redefine the role of recognition technology in document-oriented applications. One is the development of systems that provide functionality similar to that of text processors but operate directly on audio and scanned image data. A second, related theme is the use of speech and text-image recognition to retrieve arbitrary, user-specified information from documents with signal content. This paper discusses three research initiatives at PARC that exemplify these themes: a text-image editor[1], a wordspotter for voice editing and indexing[12], and a decoding framework for scanned-document content retrieval[4]. [1] The discussion focuses on key concepts embodied in the research that enable novel signal-based document processing functionality.

1. INTRODUCTION

Research on application of spoken language processing to document creation and information retrieval has focused on the use of speech as an interface to systems that operate primarily on text-based material. Products of such work include commercially available voice transcription devices, as well as DARPA-sponsored systems developed to support speech recognition and database interaction tasks (e.g., the Resource Management[9] and Air Travel Information Systems[8] tasks, respectively). Similarly, work on text image processing has focused primarily on optical character recognition (OCR) as a means of transforming paper-based documents into manipulable (i.e., ASCII-based) electronic form. In both cases, the paradigm is one of format conversion, in which audio or image data are converted into symbolic representations that fully describe the content and structure of the associated document or query.

Over the past few years, two themes have evolved in speech and text image processing work at Xerox PARC that expand and redefine the role of recognition technology in document-oriented applications. The first of these is the development of systems that provide functionality similar to that of text processors but operate directly on audio and scanned image data. These systems represent alternatives to the traditional format conversion paradigm. They are based on a principle of *partial document modeling*, in which only enough signal analysis is performed to accomplish the user's goal. The systems are intended to facilitate authoring and editing of documents for which input and output medium are the same. Examples include authoring of voice mail and editing of facsimile-based document images.

A second, and related, research theme is the use of speech and text-image recognition to retrieve arbitrary, user-specified information from documents with signal content. The focus is again on partial document models that are defined in only enough detail to satisfy task requirements. Depending upon application, format conversion may or may not be a desired goal. For example, in retrieving relevant portions of a lengthy audio document via keyword spotting, a simple time index is sufficient. On the other hand, extracting numerical data from tabular images to facilitate on-line calculations requires at least partial transcription into symbolic form.

This paper discusses three research initiatives at PARC that exemplify these themes: a text-image editor[1], a wordspotter for voice editing and indexing[12], and a decoding framework for scanned-document content retrieval[4]. Overviews of the three systems are provided in Sections 2 through 4, respectively; concluding comments are contained in Section 5. The discussion focuses on key concepts embodied in the research that enable novel signal-based document processing functionality. Technical details of the individual efforts are described in the associated references.

2. TEXT IMAGE EDITING: IMAGE EMACS

Image EMACS is an editor for scanned documents in which the inputs and outputs are binary text images[1]. The primary document representation in Image EMACS is a set of image elements extracted from scanned text through simple geometrical analysis. These elements consist of groupings of connected components (i.e., connected regions of black pixels)[2] that correspond roughly to character images. Editing is performed via operations on the connected components, using editing commands patterned after the text editor EMACS[11].

Image EMACS supports two classes of editing operations.

[1] Thanks to S. Bagley, P. Chou, G. Kopec and L. Wilcox for agreeing to have their work discussed here. This work represents a sample, rather than a full survey, of relevant speech and text-image research at PARC.

The first class is based on viewing text as a linear sequence of characters, defined in terms of connected components. Traditional "cut and paste" functionality is enabled by a selection of insertion and deletion commands (e.g., delete-character, kill-line, yank region). As with text, editing is typically performed in the vicinity of a cursor, and operations to adjust cursor position are provided (e.g., forward-word, end-of-buffer). Characters can also be inserted by normal typing. This is accomplished by binding keyboard keys to character bitmaps from a stored font or, alternatively, to user-selected character images in a scanned document. Correlation-based matching of the character image bound to a given key against successive connected-component groupings allows for image-based character search.

The second class of operations supported by Image EMACS is based on viewing text as a two-dimensional arrangement of glyphs on an image plane[1]. These operations provide typographic functionality, such as horizontal and vertical character placement, interword spacing, vertical line spacing, indentation, centering and line justification. Placement of adjacent characters is accomplished using font metrics estimated for each character directly from the image[5]. These metrics allow for typographically acceptable character spacing, including character overlap where appropriate.

Taken together, Image EMACS commands are intended to convey the impression that the user is editing a text-based document. In actuality, the system is manipulating image components rather than character codes. Moreover, while the user is free to assign character labels to specific image components, editing, including both insertion and search, is accomplished *without* explicit knowledge of character identity. This approach enables interactive text-image editing and reproduction, independent of font or writing system.

Figures 1 and 2 show an example of a scanned multilingual document before and after editing with Image EMACS. The example demonstrates the results of image-based insertion, deletion, substitution and justification, as well as intermingling of text in several writing systems and languages (paragaphs 4 through 7).[2] Such capabilities are potentially achievable using a format-conversion paradigm; however, this would require more sophisticated OCR functionality than currently exists, as well as access to fonts and stylistic information used in rendering the original document.

3. AUDIO EDITING AND INDEXING

A second example of signal-based document processing is provided by a wordspotter developed to support editing and indexing of documents which originate and are intended to remain in audio form[12]. Examples include voice mail, dictated instructions and pre-recorded radio broadcasts or commentaries. The wordspotter can also be used to retrieve relevant portions of less structured audio, such as recorded lectures or telephone messages. In contrast with most previous wordspotting applications (e.g., [15, 10]), unconstrained keyword vocabularies are critical to such editing and indexing tasks.

The wordspotter is similar to Image Emacs in at least three ways: 1) it is based on partial modeling of signal content; 2) it requires user specification of keyword models; and 3) it makes no explicit use of linguistic knowledge during recognition, though users are free to assign interpretations to keywords. The wordspotter is also speaker or, more accurately, sound-source dependent. These constraints allow for vocabulary and language independence, as well as for the spotting of non-speech audio sounds.

The wordspotter is based on hidden Markov models (HMM's) and is trained in two stages[13]. The first is a *static* stage, in which a short segment of the user's speech (typically 1 minute or less) is used to create a background, or non-keyword, HMM. The second stage of training is *dynamic*, in that keyword models are created while the system is in use. Model specification requires only a single repetition of a keyword and, thus, to the system user, is indistinguishable from keyword spotting. Spotting is performed using a HMM network consisting of a parallel connection of the background model and the appropriate keyword model. A *forward-backward* search[13] is used to identify keyword start and end times, both of which are required to enable editing operations such as keyword insertion and deletion.

The audio editor is implemented on a Sun Microsystems Sparcstation and makes use of its standard audio hardware. A videotape demonstrating its use in several multilingual application scenarios is available from SIGGRAPH[14].

4. DOCUMENT IMAGE DECODING

Document image decoding (DID) is a framework for scanned document recognition which extends hidden Markov modeling concepts to two-dimensional image data[4]. In analogy with the HMM approach to speech recognition, the decoding framework assumes a communication theory model based on three elements: an image generator, a noisy channel and an image decoder. The image generator consists of a message source, which generates a symbol string containing the information to be communicated, and an imager, which formats or encodes the message into an ideal bitmap. The channel transforms this ideal image into a noisy observed image by introducing distortions associated with printing and scanning. The decoder estimates the message from the observed image using *maximum a posteriori* decoding and a Viterbi-like decoding algorithm[6].

[2]Syntax and semantics are not necessarily preserved in the example, thanks to the user's lack of familiarity with most of the languages involved.

A key objective being pursued within the DID framework is the automatic generation of optimized decoders from explicit models of message source, imager and channel[3]. The goal is to enable application-oriented users to specify such models without sophisticated knowledge of image analysis and recognition techniques. It is intended that both the format and type of information returned by the document decoder be under the user's control. The basic approach is to support declarative specification of a priori document infomation (e.g., page layout, font metrics) and task constraints via formal stochastic grammars.

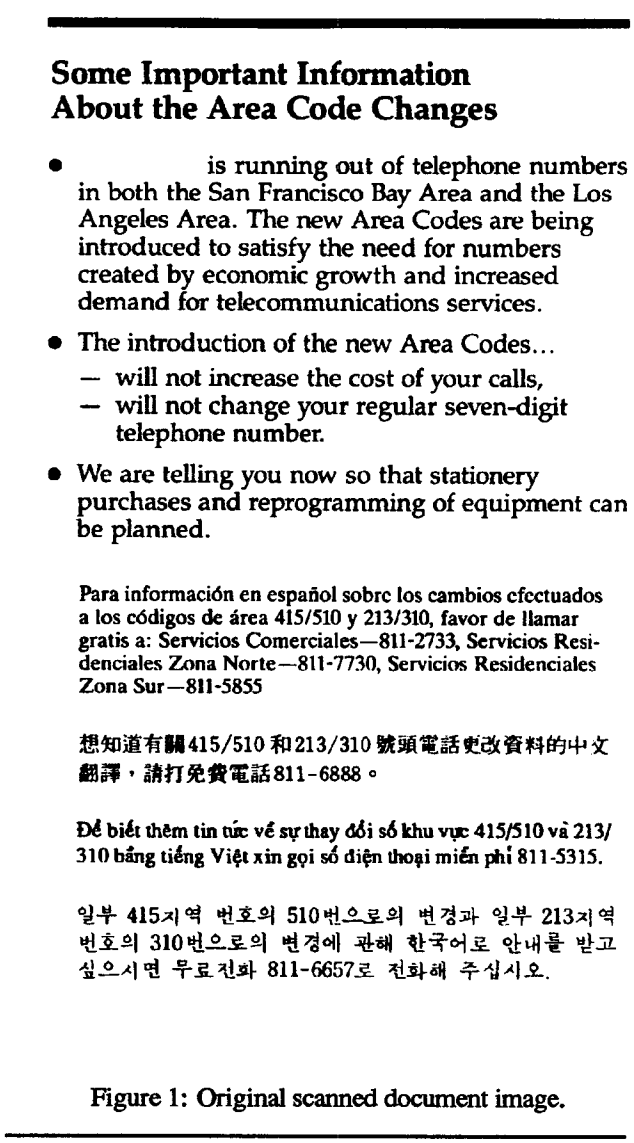
Figures 3 through 5 illustrate the application of DID to the extraction of subject headings, listing types, business names and telephone numbers from scanned yellow pages[7]. A slightly reduced version of a sample scanned yellow page column is shown in Figure 3 and a finite-state top-level source model in Figure 4. The yellow page column includes a subject heading and examples of several different listing types. These, in turn, are associated with branches of the source model. The full model contains more than 6000 branches and 1600 nodes. Figure 5 shows the result of using a decoder generated from the model to extract the desired information from the yellow page column. Automatically generated decoders have also been used to recgnize a variety of other document types, including dictionary entries, musical notation and baseball box scores.

5. SUMMARY

The speech and text-image recognition initiatives discussed in the preceding sections illustrate two research themes at Xerox PARC which expand and redefine the role of recognition technology in document-oriented applications. These include the development of editors which operate directly on audio and scanned image data, and the use of speech and text-image recognition to retrieve arbitrary information from documents with signal content. Key concepts embodied in these research efforts include partial document models, task-oriented document recognition, user specification and intepretation of recognition models, and automatic generation of recognizers from declarative models. These concepts enable the realization of a broad range of signal-based document processing operations, including font, vocabulary and language-independent editing and retrieval .

References

1. Bagley, S. and Kopec, G. "Editing images of text". Technical Report P92-000150, Xerox PARC, Palo Alto, CA, November, 1992.
2. Horn, B. *Robot Vision*, The MIT Press, Cambridge, MA, 1986.
3. Kopec, G. and Chou, P. "Automatic generation of custom document image decoders". Submitted to ICDAR'93: Second IAPR Conference on Document Analysis and Recognition, Tsukuba Science City, Japan, October, 1993.
4. Kopec, G. and Chou, P. "Document image decoding using Markov sources". To be presented at ICASSP-93, Minneapolis, MN, April 1993.
5. Kopec, G. "Least-squares font metric estimation from images, EDL Report EDL-92-008, Xerox PARC, Palo Alto, CA, July, 1992.
6. Kopec, G. "Row-major scheduling of image decoders". Submitted to *IEEE Trans. on Image Processing*, February, 1992.
7. Pacific Bell *Smart Yellow Pages, Palo Alto, Redwood City and Menlo Park*, 1992.
8. Price P., "Evaluation of Spoken Language Systems: The ATIS Domain," *Proc. Third DARPA Speech and Language Workshop*, P. Price (ed.), Morgan Kaufmann, June 1990.
9. Price, P., Fisher, W., Bernstein, J. and Pallett, D. "The DARPA 1000-Word Resource Management Database for Continous Speech Recognition". *Proceedings of ICASSP-88*, 1988, pp 651-654.
10. Rohlicek, R., Russel, W., Roukos, S. and Gish, H. "Continuous hidden Markov modeling for speaker-independent word spotting". *Proceedings ICASSP-89*, 1989, 627-630.
11. Stallman, R. *GNU Emacs Manual*, Free Software Foundation, Cambridge, MA, 1986.
12. Wilcox, L. and Bush, M. "HMM-based wordspotting for voice editing and audio indexing". *Proceedings of Eurospeech-91*, Genova, Italy, 1991, pp 25-28.
13. Wilcox, L. and Bush, M. "Training and search algorithms for an interactive wordspotting system". *Proceedings of ICASSP-92*, 1992, pp II-97 - II-100.
14. Wilcox, L., Smith, I. and Bush, M. "Wordspotting for voice editing and indexing". *Proceedings of CHI '92*, 1992, pp 655-656. Video on SIGGRAPH Video Review 76-77.
15. Wilpon, J., Miller, L., and Modi, P. "Improvements and applications for key word recognition using hidden Markov modeling techniques". *Proceedings of ICASSP-91*, 1991, pp 309-312.



Some Important Information About the Area Code Changes

- is running out of telephone numbers in both the San Francisco Bay Area and the Los Angeles Area. The new Area Codes are being introduced to satisfy the need for numbers created by economic growth and increased demand for telecommunications services.
- The introduction of the new Area Codes...
  - will not increase the cost of your calls,
  - will not change your regular seven-digit telephone number.
- We are telling you now so that stationery purchases and reprogramming of equipment can be planned.

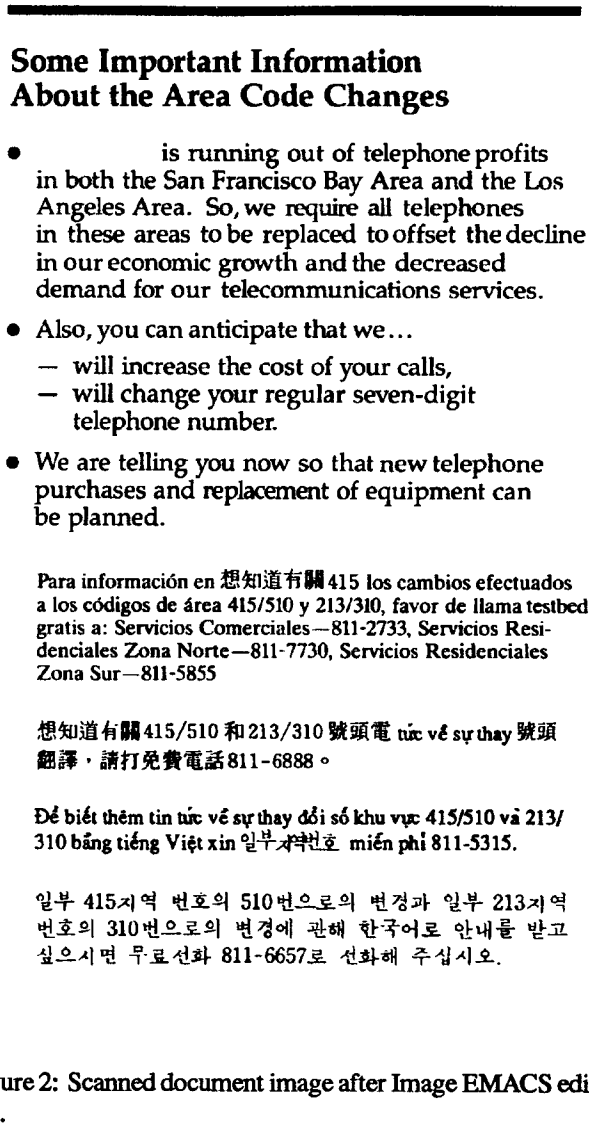
Para información en español sobre los cambios efectuados a los códigos de área 415/510 y 213/310, favor de llamar gratis a: Servicios Comerciales—811-2733, Servicios Residenciales Zona Norte—811-7730, Servicios Residenciales Zona Sur—811-5855

想知道有關415/510和213/310號頭電話更改資料的中文翻譯，請打免費電話811-6888。

Để biết thêm tin tức về sự thay đổi số khu vực 415/510 và 213/310 bằng tiếng Việt xin gọi số điện thoại miễn phí 811-5315.

일부 415지역 번호의 510번으로의 변경과 일부 213지역 번호의 310번으로의 변경에 관해 한국어로 안내를 받고 싶으시면 무료전화 811-6657로 전화해 주십시오.

Figure 1: Original scanned document image.



Some Important Information About the Area Code Changes

- is running out of telephone profits in both the San Francisco Bay Area and the Los Angeles Area. So, we require all telephones in these areas to be replaced to offset the decline in our economic growth and the decreased demand for our telecommunications services.
- Also, you can anticipate that we...
  - will increase the cost of your calls,
  - will change your regular seven-digit telephone number.
- We are telling you now so that new telephone purchases and replacement of equipment can be planned.

Para información en 想知道有關415 los cambios efectuados a los códigos de área 415/510 y 213/310, favor de llama testbed gratis a: Servicios Comerciales—811-2733, Servicios Residenciales Zona Norte—811-7730, Servicios Residenciales Zona Sur—811-5855

想知道有關415/510和213/310號頭電 tức về sự thay 號頭翻譯，請打免費電話811-6888。

Để biết thêm tin tức về sự thay đổi số khu vực 415/510 và 213/310 bằng tiếng Việt xin 일부지역번호 miễn phí 811-5315.

일부 415지역 번호의 510번으로의 변경과 일부 213지역 번호의 310번으로의 변경에 관해 한국어로 안내를 받고 싶으시면 무료전화 811-6657로 전화해 주십시오.

Figure 2: Scanned document image after Image EMACS editing.

Telecommunications-Telephone Equipment & Systems-Service & Repair

APEX COMMUNICATIONS INC
254 San Geronimo
Sunnyvale 408 773 9600
Communications 2001
Voice Mail System
433 Airport Bl Burl 347 1500
Data-Tel SMto 349 6010
Davis Communication Services
683 Crane Av FosterCty 341 0214
G'raffe Communications Inc 594 9196
HBC

24 HOUR
PHONE SERVICE
REPAIR
SALES & SERVICE
JACK INST. - REWIRE
SERVICE CALLS
3628 El Camino Rl PA 493 3663

Integrated Technologies Inc 345 4484
Interconnect Services-Inc 591 3120
Interconnect Services Inc 851 9224
Mr Telephone Man 692 4128
Please See Advertisement This Page

MODERN APPLICATIONS GROUP
Telecommunications Professionals
"Quality Value Service"
.....961 8181

NEC BUSINESS COMMUNICATION SYSTEMS
6960 Koll Center Pkwy
Pleasanton 510 484 2010

NICKELL TEL COMM
4046 Mira Loma Wy
SanJose 408 629 1011

NORTHERN TELECOM INC
Worldwide Leader in Telecommunications Technology
nt northern telecom
- Meridian 1 and Norstar
- Business Telephone Systems
- Integrated Voice and Data Networks
- Private Communications Networks
- Telephone Industry Sales

"FOR INFORMATION CALL"
Northern Telecom Inc
685 E Middlefield Rd MtView .969 9170
PacTel Meridian Systems
333 Lakeside Dr FosterCity ...358 3300
PacTel Meridian Systems
2305 Mission College Bl
SantaClara 408 988 5550
Or
No Charge To Calling Party 800 667 8437

PANASONIC AUTHORIZED SERVICENTERS-PARTS DISTRIBUTORS—
PARTS & SERVICE
INFORMATION SERVICES Peoriall
No Charge To Calling Party 800 545 2672
FACTORY SERVICE CENTERS
MATSUSHITA SERVICES COMPANY
800 Dubuque Av SSF 871 6373

PRECISION PHONE 728 3623
Please See Advertisement This Page

PRECISION PHONE BUSINESS SYSTEMS

COMPLETE SYSTEMS PLUS

- REASONABLE REPAIRS -
- ADDITIONS OR MOVES OF EXISTING SYSTEMS -

NEW OR REBUILT

ITT • TIE • COMDIAL ALL BELL PRODUCTS TOSHIBA & OTHERS

728-3623

CALL FOR FREE ESTIMATES

Figure 3: Scanned yellow page column, slightly reduced.

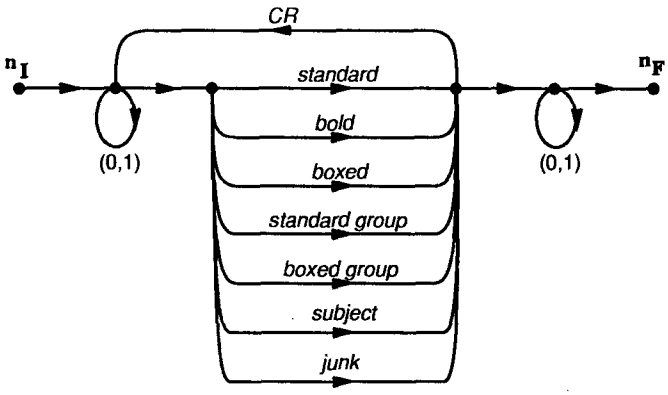


Figure 4: Top-level source model of yellow page column. Transition probabilities omitted for simplicity.

```
\subject{Telecommunications-\\Telephone Equipment\\& Systems-Service &\\Repair}
\bold{\name{APEX COMMUNICATIONS INC} \number{408 773 9600}}
\standard{\name{Coinmuhications ZOOI} \number{347 1500}}
\standard{\name{Data-Tel 'SMto} \number{349 6010}}
\standard{\name{Davis Coinmuhication Services} \number{341 0214}}
\standard{\name{G'raffe Communicatlons Inc} \number{594 9196}}
\boxed{\name{HBC} \number{493 3663}}
\standard{\name{Integrated Tecnnologies Inc} \number{345 4484}}
\standard{\name{Interconnect Services-Inc} \number{591 3120}}
\standard{\name{Interconnect Services Inc} \number{851 9224}}
\standard{\name{Mr Telephone Man} \number{692 4128}}
\bold{\name{MODERN APPLICATIONS GROUP} \number{961 8181}}
\bold{\name{NEC BUSINESS COMMUNICATION\\SYSTEMS} \number{510 484 2010}}
\bold{\name{NICKELL TEL COMM} \number{408 629 1011}}
\boxedgroup{\name{NORTHERN TELECOM INC}
   \subgroup{\name{''FOR INFORMATION CALL''}
      \standard{\name{Northern Telecorn Inc} \number{969 9170}}
      \standard{\name{PacTel Meridian Systems} \number{358 3300}}
      \standard{\name{PacTel Meridian Systems} \number{408 988 5550}}
      \standard{\name{' Qr} \number{0 667 8437}}}}
\standardgroup{\name{PANASONIC AUTHORIZED\\SERVICENTERS-PARTS\\DISTRIBUTORS}
   \subgroup{\name{PARTS & SERVICE}
      \stdgrpmem{\name{INFORMATION SERVICES 'P-J--JH} \number{545 2672}}}
   \subgroup{\name{FACTORY SERVICE CENTERS}
      \stdgrpmem{\name{MATSUSHITA SERVICES COMPANY} \number{871 6373}}}}
\bold{\name{PRECISION PHONE} \number{728 3623}}
\boxed{\name{NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNI} \number{8888881}}
```

Figure 5: Decoder output for yellow-page column in fig 3.