A FAX Reader for the Blind

Allen E. Milewski
Henry S. Baird

AT&T Bell Laboratories

Crawford's Corner Road, Room 1J-337
Holmdel, New Jersey 07733

600 Mountain Avenue, Room 2C-557
Murray Hill, New Jersey 07974

Abstract

We describe an experiment in which a blind person places arbitrary printed pages in a FAX machine and soon afterwards hears the contents read out loud over the telephone by a synthesized voice. This study focuses on technical issues including the accuracy achievable by state-of-the-art optical character recognition operating on FAX images, methods to improve the intelligibility of synthesized speech for this application, and ease of interaction with the user. In a small-scale trial of the service under laboratory conditions, we have observed that accuracy and intelligibility on some commonly-occurring types of documents, including typewritten letters, is usefully high. We believe that technology developments in the near future will support services acceptable to a wide range of users.

1. Introduction

We describe a small-scale exploratory study of a new telecommunications service concept directed at visually-impaired populations, to read hard-copy printed text out loud over the telephone fully automatically. This requires integrating state-of-the-art technologies for voice, data, and image processing. The service accepts images of documents using facsimile transmission (FAX), translates the image to ASCII text using optical character recognition (OCR), and converts the text to synthesized speech (TTS).

Due to the ubiqitousness of TouchTone telephones and FAX machines, such a service could enable visually-impaired people to cope with paper copies of documents and letters in a convenient and inexpensive manner, both at work and at home. We have carried out a small-scale feasibility trial under laboratory conditions, focusing on technical issues including the accuracy achievable by state-of-the-art OCR operating on FAX images, methods to improve the intelligibility of TTS in this application, and ease of interaction with the user. The results of the study show that the service concept is both technically feasible and useful to the blind using currently-available technology. As a guide to future work, we propose a set of enhancements, principally in human-factors design and OCR accuracy, many of which should be straightforward to implement, and a few of which may require further research.

24th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, California, November 5-6, 1990.

2. Definition of the Service

The concept of a network-based text-reading service originated in discussions the first author had with numerous visually-impaired participants in the TRACE Institute Planning Workshop on Computer Access by Blind Individuals, October, 4-7, 1988. To discover even the most basic facts about paper documents, visually-impaired persons must do one of three things:

1. *Ask a colleague or family member to read to them.* While this appears to be common for reading short, urgent papers, the interpersonal burden of the visually-impaired frequently relying on their business colleagues is obvious. In a home environment, there is a problem when the visually-impaired person is alone.
2. *Hire a human reader.* This, too, is a common practice. The major problem with this procedure is that human readers are too expensive to hire for full-time service, especially when their services may only be required sporadically. Generally, they are hired for an hour or two at a time, and paper materials are saved for their arrival.
3. *Use OCR devices.* This option is open only to those blind with easy access to special OCR equipment. Commercially available stand-alone OCR devices for the visually-impaired are expensive, typically ranging from $8,000 to $20,000 per unit. There are less expensive OCR cards and scanners available for IBM-compatible and MacIntosh PCs. However, these cards tend to work well on only a limited set of machine-printed font styles. Further, using these cards requires an investment of money and training time in the use of PC equipment. While a growing set of visually-impaired persons are willing to make this investment, there are still many who do not have PCs.

Visually-impaired persons may find it useful to be able to submit one or more paper pages to a telephone-based service and hear, after a short delay, a translation into speech. The service proposed here is an automated FAX to synthetic speech facility available through any standard FAX machine and TouchTone telephone.

After receiving the FAX submission from the user, the service would carry out OCR on the image of each page. Depending on options selected by the caller, the text would either be read back to the user in synthetic speech or mailed electronically in ASCII form (*e.g.* via ATTmail). Following synthetic speech recitation, the service would offer users the option of transferring to a human reader for further clarification. Short (*e.g.* single-page) submissions could be read back to the caller immediately, while for longer, multi-page submissions, a ‘‘mailbox’’ facility would be offered so the user could call back at a later time to listen to the translated material.

Impaired populations represent a potentially large market for network services. It is estimated that there are 8.4 million individuals in the United States with some degree of visual impairment; 600,000 of these are legally blind. The ubiquitous nature of FAX and TouchTone telephones makes this FAX-to-speech translation service available to visually-impaired persons both in the home and business environment while minimizing the need to purchase expensive equipment.

3. The Exploratory Study

The goal of the study was to explore the technical feasibility and usefulness of the service concept by making a prototype service available to a user. In particular, we investigated the quality achievable with Group 3 facsimile images of unconstrained printed documents. The success of a network-based, automated text reading service depends on several issues.

1. *Is the OCR accuracy achievable via facsimile adequate?* FAX images are of low quality. This is due to the variable quality of scanners, noise in transmission, and, most importantly, coarse digitizing resolution. FAX Group 3 resolution is 200x100 pixels per inch

(ppi) in standard mode and 200x200 ppi in detail mode. Experiments [1] have shown that, at 200x200 ppi, the recognition rate of current OCR techniques declines markedly for printed text below 11 point size. To put this in perspective, 12 point is roughly typewriter pica size, and 10 point is typewriter elite size. Thus a large fraction of printed material, imaged via FAX, presents a considerable challenge to today's OCR technology. Another key difficulty is that the FAX protocol offers no convenient way for the user to specify any constraints that might be known such as limits on text size, alphabet, number of columns, etc., so that the OCR could exploit them for improved results. In any case, a blind user will often not know these in advance. Thus the OCR must be as versatile and autonomous as possible. A crucial difficulty is that no general and fully-automatic technique exists today to analyze complex page layouts such as multiple columns with headers, footnotes, and embedded line-art or photographs (this is discussed in detail in [2] and [3]).

2. *Is the current quality of synthetic speech acceptable?* Traditionally, synthetic speech technology has not been considered of high enough quality for use in network-based services such as operator assistance, etc. On the other hand, many blind persons regularly use PC-based text-to-speech products, often of fairly low quality. Since their alternatives are limited, the motivation of visually impaired persons to learn to understand synthetic speech is high. This study investigated whether there were serious problems in using current synthetic speech technologies for the service.
3. *Can specialized articulation rules enhance the comprehension of OCR output?* Intelligibility of spoken words may remain surprisingly high even though some letters are altered or omitted. We hypothesized that the comprehension of the OCR output might be higher in some cases when the material was pronounced (synthetically) than it would have been if displayed visually. Assuming that some OCR errors are inevitable, the study informally explored whether special policies of synthetic pronunciation might lessen their disruptive effect on intelligibility.
4. *What can be gleaned from imperfectly translated text?* Given the obstacles to providing error-free OCR and synthetic speech on FAX images of unconstrained text, we were interested in assessing whether users could nevertheless infer useful information from what we can provide. Would a user be willing to pay to learn only the general class of the document — for example, whether it is a legal notice or a piece of junk mail — while not being able to understand every detail?

All these issues were assessed during the trial.

4. The Trial

A blind technical supervisor at AT&T Bell Laboratories volunteered to serve as the subject in a trial lasting 6 weeks. He had extensive prior experience with a text-to-speech translation device of a different brand than that used in the study. He regularly used a Votrax synthetic speech system as a reader integrated with his computer terminal.

He had an AT&T model 3510D Fax machine at home and another at his office, and could therefore submit pages to the service and/or listen to speech translations at any time of the day or night. Shortly after listening to the synthesized speech for each submission, he either made written comments or telephoned verbal comments that we later summarized. He was asked to indicate, for each submission, whether he could discern what the page was about and who it concerned. In addition, he was asked to indicated any other comments he had regarding the service or specific submission.

Figure 1 shows the equipment and call-flow of the feasibility study.

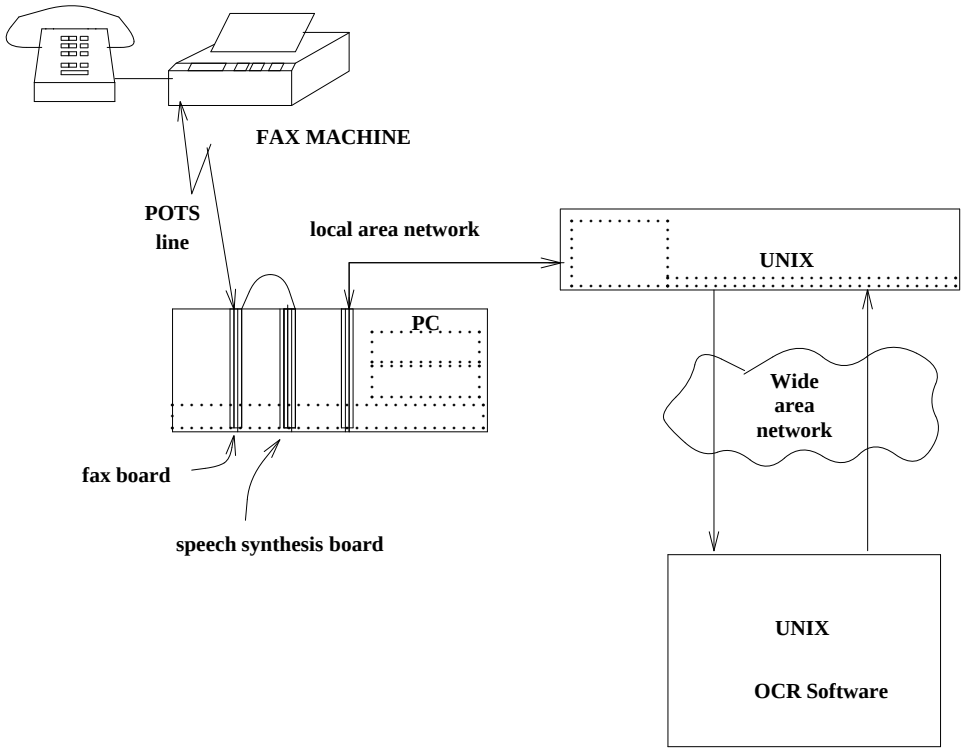


Figure 1. Equipment and Call Flow for Feasibility Study

For contingent technical reasons, the feasibility study was forced into a more cumbersome call-flow and user interface than is recommended for a real service. Several features did not work properly on the particular text-to-speech board used for the study: *e.g.* it did not provide the DTMF discrimination required to provide user-options for the service. Moreover, the method of communicating with the OCR software (uucp) did not permit a page to be submitted and listened to in the same call.

The following sequence describes the call-flow of the feasibility study. Later, we will say how the proposed service should deviate from the study.

1. The user called a telephone number from a TouchTone telephone connected as an attached telephone to a standard FAX (AT&T 3510D) machine. The user was told to place the fax machine in detail (high resolution) mode.
2. The incoming telephone line was connected to an AT&T FaxConnection board plugged into a PC6300. A SpeechPlus Corp., CallText 5000 board, also plugged into the PC was connected as an "attached telephone" to the FaxConnection board (set in high-resolution mode). The CallText board detected the call, went off hook and played a brief synthetic speech introductory announcement. The announcement greeted the user, and gave instructions.
3. After the instructions, the user pressed a TouchTone key to begin the procedure. The CallText board played a synthetic speech version of a previous page submitted by the user. In most cases, this was the most recent page submitted. However, since it took 10 minutes or more for communication and processing with the OCR machine, it was possible to call into the service before the most recent submission was completed. In these cases, the submission prior to that was heard.
4. The user could stop the synthetic speech and move to the next step by pressing a TouchTone key. Then, or after the entire page was read, the user was prompted to enter a TouchTone, place a new page in the FAX machine and press "start".
5. The FaxConnection board was put into receive mode. When the FaxConnection board received the carrier signal from the user's FAX machine, a standard Group 2 or 3 Facsimile connection was made between the user's FAX machine and the FaxConnection board.
6. Only single-page submissions were supported in the trial. If the PC determined that the submitted page was blank (*i.e.* the file was extraordinarily small), it prompted the user to retry. Otherwise, the user was presented a synthesized "thank-you" message, and the line was placed on-hook. If the user did not wish to submit a page, he could press a TouchTone and hangup.
7. The bitmapped file collected by the FaxConnection board was copied to a 3B-2, UNIX machine via STARLAN. The file was logged, and sent via uucp on the RADIAN network to a remote VAX for Optical Character Recognition (OCR).
8. The image file, compressed using CCITT Group 3 coding, was transmitted via the uucp network to a computer at Bell Labs' Murray Hill location, where it was processed automatically using an experimental network OCR service developed by the second author ([4] and [5]). This system copes automatically with many commonly-occurring variations, such as font styles, text sizes, and line-spacing. However, like all commercially-available OCR systems, it has some limitations. In particular, it assumed that each page consisted of one upright column of machine-printed or typewritten English text; as a result, any part of a page having multiple columns or handwriting was usually garbled. An English dictionary (essentially the UNIX spell checker) was used to improve recognition in ambiguous cases. The output of the procedure was a plain text file that preserved an approximation of the original page layout in the form of vertical and horizontal

spacing. More elaborate formatting (including text sizes and recognition quality) would have been easy to include, but we felt that this would complicate the text-to-speech processing unduly, at least within the constraints of this trial.

9. The text file was returned (via electronic mail) to the originating 3B-2. The file was logged and modified slightly to enhance the text-to-speech process. This modification included the omission of numerals and punctuation embedded within alphabetic strings, and the replacement of garbled lines (those containing only punctuation) by an "illegible passage" message.
10. The PC6300 polled the 3B-2 via STARLAN every 4 minutes. When new mail appeared from the OCR machine, it was copied to the PC6300, and stored in a file ready to be presented as synthesized speech for the next user call. In addition to being copied to the PC for speech synthesis, the ASCII file was mailed electronically to the user, who had his own synthetic speech reader system.

5. Trial Results

During the course of the study, twenty facsimile pages were submitted and listened to. These twenty pages were an assortment of postal mail and other documents received by the user. In this study, we have focused both on the recognition accuracy rates, and on the overall intelligibility of the text when presented as synthesized speech.

Recognition accuracy was measured on each page by counting the number of words that had no errors; that is, no missing characters, no extra characters and no misrecognized characters. The number of correct words was divided by the total number of words on the original page, which ranged from 64 to 661 with an average of 235. This word-level metric is a highly stringent measure of accuracy but was justified since the primary concern was intelligibility of submitted pages. According to this measure, recognition accuracy varied widely as a function of the material's type, layout and font-size. The median percent correct words, calculated across the twenty pages, was 85.5%, with a range of from 0% to 98%. The great majority of OCR failures were due to: (a) small text; (b) unusual fonts; (c) varied page layouts; and (d) non-textual clutter such as line drawings, photographs, logos, etc. One image was sprinkled with hundreds of tiny blemishes, possibly caused by electrical interference with the sending fax machine; the OCR software was easily modified to cope with these.

Of the pages on which recognition was worst, one was a multi-column table printed in landscape mode. One page featured a line-drawing map, which interfered with recognition of the surrounding text. A third was a page printed on a dot-matrix printer, that also appeared to have facsimile transmission errors causing shortened letters. Of the pages that had highest recognition accuracy, all were single-column documents printed cleanly in approximately 12 point pica text. The remaining pages had an assortment of layouts and fonts.

In general, the user found the service beneficial in that some information could be gotten from most of the pages submitted. The synthesized speech was found to be acceptable and the service was considered easy to use. The intelligibility ratings for submitted pages were, unsurprisingly, almost entirely a function of OCR accuracy, and are summarized below:

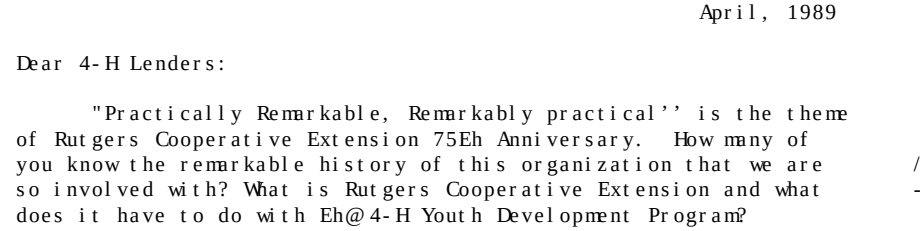
1. Ten of the pages were extremely good. The user judged these pages to be completely comprehensible. The user felt confident of the meaning of the pages. Nine of these pages were consecutive pages of the same document. After listening to the synthetic speech versions of these pages, the trial user received electronic mail copies and edited the text into a document he was preparing himself. The tenth page in this category was another typewritten letter that was judged 100% understood. The mean word-level OCR accuracy score for this group of pages was 93%.
2. The user judged the comprehension of an additional four pages to be good. While the

subject matter of the page could generally be determined, specific words could often not be made out. In some cases, sections of the page were perfectly understandable, but errors in other parts kept the user from trusting his overall understanding. In all these cases, comprehension was good enough so that the user knew what to do with the page (*e.g.* throw it away, get more information about it, etc). The mean OCR accuracy for this group of pages was 79%.

3. Six of the twenty pages were judged of poor quality by the user. While some number of words could be understood, the user did not feel confident that the meaning or nature of the page was understood. The mean OCR accuracy for these pages was 25%. Of these six pages, three were printed using a coarse dot-matrix font that the system does not handle well. Two of the six pages were printed in a multi-column format that cannot yet be automatically analyzed.

Figure 2 shows a FAX image and its ASCII translation.

FAX image input:



ASCII text output:

```
                                                    April, 1989

Dear 4-H Lenders:

      "Practically Remarkable, Remarkably practical'' is the theme
of Rutgers Cooperative Extension 75Eh Anniversary.  How many of
you know the remarkable history of this organization that we are      /
so involved with? What is Rutgers Cooperative Extension and what      -
does it have to do with Eh@4-H Youth Development Program?
```

Figure 2. An excerpt from a FAX image is shown at the top, reduced to 0.7× original size. Note the non-zero skew angle (1.6°) and dirt fragments scattered in white space and clustered at the right margin. The ASCII text (output by the OCR) is shown below. It has three wrong words (four bad characters), for an OCR accuracy of 94%, typical of the performance we observed on detail FAXes of pica typewritten original copy.

6. Discussion of Results

The results of this exploratory study demonstrate the technical feasibility of a basic facsimile-based automated text reading service for the blind. For pages whose print quality and layout complexity do not pose too great a challenge to the OCR, high levels of comprehension have been achieved. Due to the small scale of the trial, we cannot estimate what fraction of documents of interest to blind users fall in this class: but it includes at a minimum all cleanly-printed typewritten letters using pica-size text, which are a large and interesting class. Even where the text is garbled to the point of unintelligibility, often the general type of the document can be confidently determined. Comprehension of borderline cases is helped by several factors, including the effect of pronunciation, and the human capacity to fill-in the sense from context.

The study also suggests that a significant amount of information can be obtained from a facsimile-based text reading service using technology available today. Fourteen of twenty pages were of sufficient quality to give the user a basic understanding of the page's subject matter. In a network service, this level of quality, together with an option to transfer to a human reader for those cases in which the automated OCR fails, is already sufficient to benefit blind individuals.

Several procedural changes should be planned for future versions of the service.

1. *Multi-page input.* The user often had multi-page documents, but the limited service-flow possible during the trial forced him to submit each page separately. This limitation should be relaxed.
2. *Immediate response vs mail-box.* One use of the service is merely to determine the subject matter of a particular page. For this purpose, it is most convenient to have the page read immediately, during the same call. Alternatively, blind users may want to make a submission and listen to the results later. For this purpose, it would be most convenient to have the speech stored in a personalized mail-box that can be accessed by the user at any time. The mailbox is especially important for multi-page submissions for which the user may not wish to wait. It also permits the user to retain certain submissions to be listened to repeatedly.
3. *Electronic-mail option.* The trial user found it very useful to have the ASCII text output by OCR mailed to him electronically, allowing him to retain and manage materials himself. Perhaps this could be provided as a user-profile feature selected at subscription time or selected via TouchTone at each use.
4. *Control of reading order.* The trial user would have found listening to pages much more convenient and useful if he had had a series of commands to control the synthetic speech output. TouchTone entries should be used to control the output. A useful set of reading-order commands might include: (a) speed up synthetic speech; (b) slow down synthetic speech; (c) repeat previous sentence; (d) repeat previous word; (e) repeat the page from the beginning; (f) enter/exit spelling-mode (in which words are spelled-out rather than pronounced); and (g) skip the next sentence.

This exploratory study points to further research areas that could increase the success rate of the automated text reading service significantly.

1. *Complex page layouts.* While single-column machine-printed or typewritten pages are likely to continue to make up a significant fraction of the input, there is an urgent need for fully-automatic analysis of more varied layouts. Research directed at this problem is underway [3]. Also, users may insert pages upside down, and pages may be printed in landscape format rather than the more common portrait format. At the moment, we envisage a crude but effective solution: simply try all four orientations and pick the best (*e.g.* the one with the largest number of correctly-spelled words). This brute-force

technique was not attempted during this trial; perhaps something better could be devised.

2. *Small and noisy images.* The low digitizing resolution of FAX strains conventional recognition techniques. A mathematical model of many distortions occurring during printing and imaging is under development [6] and will be used for ‘‘compensatory training’’ to make the classifier less vulnerable to noise.
3. *Font styles.* The experimental OCR system has been carefully trained and tested on over 100 font styles; and, as the trial has shown, it performs well on other fonts that happen to be similar to these. However, there is room for improvement, particularly for extreme variations such as bold, expanded, and dot-matrix fonts.
4. *Speed.* The principal speed bottleneck at present is within the OCR software, specifically during the classification of isolated character images. Anticipated CPU improvements may relieve this bottleneck soon; also, straightforward application of special hardware such as digital signal processors (DSP) and forseeable developments in neural-network VLSI hardware [7] may offer substantial speedups in critical inner loops. Delays owing to communication are only a small fraction of the total time, and in principle can be reduced to zero by running the entire system on a single computer.
5. *Graceful failure.* For the foreseeable future, certain types of documents (especially non-text documents such as maps) will always be garbled by OCR. Rather than confuse the user with poor output, it would be nice to simply report that the page as a whole is illegible. At the moment we aren’t sure how to recognize all these cases reliably. Also, we need to work on ways of presenting this to the user.
6. *More intelligible synthesized speech.* Improvements in TTS in recent years ([8] and [9]) promise higher intellibility on unconstrained text, including proper names.

7. Conclusions

A small-scale experiment has shown that a fully-automatic FAX reading service for the blind is technically feasible today. With technology currently running in the laboratory, the accuracy and intelligibility of the service on some commonly-occurring types of documents, including typewritten letters, is usefully high. We believe that, in the near future, improvements resulting from a combination of engineering development and basic research will permit a service of acceptable quality for a wide range of users. The ubiquitousness of FAX and TouchTone telephones will then enable visually-impaired people to cope with paper documents in a convenient and inexpensive manner both at work and at home.

8. Acknowledgements

The paper has benefited from careful readings by Dennis Ritchie, Doug McIlroy, and Tim Thompson.

9. References

[1] Lam, S., and H. S. Baird, ‘‘Performance Testing of Mixed-Font, Variable-Size Character Recognizers,’’ *Proceedings, 5th Scandinavian Conference on Image Analysis,* Stockholm, SWEDEN, June 2-5, 1987.

[2] Baird, H. S., ‘‘Global-to-Local Layout Analysis,’’ *Proceedings, IAPR Workshop on Syntactic and Structural Pattern Recognition,* Pont-à-Mousson, France, 12-14 September, 1988

[3] Baird, H. S., S. E. Jones, and S. J. Fortune, ‘‘Image Segmentation using Shape-Directed Covers,’’ *Proceedings, IAPR 10th Int’l Conf. on Pattern Recognition,* Atlantic City, NJ, 17-21 June, 1990.

[4] Baird, H. S., S. Kahan, and T. Pavlidis, ‘‘Components of an Omnifont Page Reader,’’ *Proceedings of the 8th Int’l Conf. on Pattern Recognition*, Paris, France, October 27-31, 1986.

[5] Baird, H. S., and K. Thompson, ‘‘Reading Chess,’’ *Proceedings, IEEE Computer Society Workshop on Computer Vision*, Miami, Florida, 30 November-2 December, 1987.

[6] Baird, H. S., ‘‘Document Image Defect Models,’’ *Proceedings, IAPR 1990 Workshop on SSPR*, Murray Hill, NJ, June 13-15, 1990.

[7] LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, and H. S. Baird, ‘‘Constrained Neural Network for Unconstrained Handwritten Digit Recognition,’’ *Proceedings, Intl’l Workshop on Frontiers in Handwriting Recognition*, Montreal, 2-3 April, 1990.

[8] Hirschberg, J., S. A. Riederer, J. E. Rowley, and A. Syrdal., ‘‘Voice Response Systems: Technologies and Applications,’’ *AT&T Technical Journal*, [to appear] November, 1990.

[9] Liberman, M. Y., K. W. Church, ‘‘Text Analysis and Word Pronunciation in Text-to-speech Synthesis,’’ *Advances in Speech Signal Processing*, S. Furui and M. Sondhi, eds. Marcel Dekker: [to appear] June 1991.